Fluid Limit of A Many-Server Queueing Network with Abandonment

Weining Kang[*] Guodong Pang[†]

October 22, 2013

Abstract

This paper studies a non-Markovian many-server queueing network with abandonment, where externally arrived and internally routed customers are served under the non-idling First-Come-First-Serve (FCFS) discipline at each station of many parallel servers. Externally arrived and internally routed customers in each queue may have different patience time distributions. The system dynamics is described by the total number of customers in each queue (both waiting and receiving service) and a triplet of measure-valued processes, tracking the amount of service time each customer in service has received, the waiting times of externally arrived customers and the waiting times of internally routed customers in queue. We show a functional strong law of large numbers in the many-server regime for these processes, where the limit processes are the unique solution to a system of deterministic measure-valued integral equations, and characterize the invariant states of the fluid limit.

1 Introduction

We study a non-Markovian many-server queueing network model with customer abandonment and Markov routing. 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. Upon service completion, a customer is immediately routed to one of the service stations or leaves the system following a Markovian routing mechanism, independent of other customers. Our network model allows for customer feedback. Externally arrived and internally routed customers at each service station are served in the non-idling, First-Come-First-Serve (FCFS) discipline. Customers can be out of patience and leave the system (without reentry) when they are waiting in the queue before receiving service. The patience-time distributions of externally arrived and internally routed customers may be different. The service and patience time distributions of customers depend on the service station. See §2.1 for a more detailed description of the model.

This model has many interesting applications in customer contact centers and patient flow analysis. In particular, recent empirical analysis shows that the patience times of customers' first

[*] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (wkang@umbc.edu)

[†] The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802 (gup3@psu.edu)

visit and reentrants can be very different, see [18] and [38]. Thus, it is important to understand the system dynamics of this network model with this feature of 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 asymptotically. In the many-server regime, even for the Markovian model with exponential service and patience times at each station, the conventional approach in [29] does not work. Here we adapt the measure-valued stochastic-process framework developed by Kaspi and Ramanan [16] and Kang and Ramanan [15, 14] to obtain fluid approximations of the system dynamics in the many-server regime.

In particular, we use a triplet of measure-valued stochastic processes together with the process counting the total number of customers at each service station to provide a Markovian representation of the system evolution dynamics. Here, one measure-valued process keeps track of the amount of service time each 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, because a single measure-valued process cannot simultaneously characterize the different waiting (impatient) behaviors of externally arrived and internally routed customers. See §2.2 for a more detailed presentation of these state descriptors and dynamical equations.

We investigate the approximate system dynamics in the many-server heavy-traffic regime where arrival rates in all service stations and the numbers of servers in those service stations go to infinity together in an appropriate manner while the service and patience time distributions are fixed; specifically, we assume that 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 main results of this paper include a functional strong law of large numbers (FSLLN, fluid limit) (Theorem 6.1) for the Markovian state descriptor, i.e., the triplet of measure-valued processes and the process counting the number of customers at each queue, uniqueness of a solution to the limit (Theorems 3.4) and the characterization of its invariant states (Theorem 3.9). We also discuss the two special cases, a single station with immediate feedback and a tandem network of two many-server queues with abandonment.

The proof of the convergence for the FSLLN is based on the arguments in [16, 15, 14], by constructing martingales for the departure processes at each service station. However, characterizations of the limit processes are more involved than the single service station setting [15]. 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, patience-time differentiation of externally arrived and internally routed customers adds more complication because the two measure-valued processes describing the dynamics of customers in a single FCFS queue at each service station are only linked by the waiting-time process of the head-of-line customer, who can be either an externally arrived or internally routed customer. 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 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 patience-time differentiation. These methods may turn out to be useful for the study of uniqueness of fluid model solutions in service networks with other network structures, e.g., skill-based routing in [10, 35] and customers switching queues in [30].

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], [27], [12], [19], [1] and [34] 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], [28], [1] and [34]. 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 [36, 16, 14, 15, 39, 20, 21, 22, 23, 13] for fluid models using measure-valued and two-parameter processes tracking elapsed or residual times and [17, 32, 6, 7, 26, 8] for approximations in the Halfin-Whitt regime. (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 [16, 14] to study a multiclass non-Markovian many-server model with abandonment, in which customers are served according to a non-preemptive priority policy. Reed and Shaki [33] consider the $G/GI/N$ queue with multiple server pools and pool-dependent general service time distributions in the Halfin-Whitt regime, where customers are routed to the server pool with the longest weighted cumulative idle time in order to achieve fairness in the workload among the server pools. Liu and Whitt [24, 25] 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 [20] to this time-varying fluid network model. Their model is closest to ours, but they do not consider the differentiation of patience times of external and internal customers.

1.1 Organization of the Paper

The notation used in this paper is the same as in [16, 15, 14], so we give a brief description of notation in §1.2. We describe the model precisely and present the measure-valued stochastic-process descriptor for the system dynamics in §2. We present the fluid model in §3.1, and characterize its invariant state in §3.2. Two special cases are discussed in §4.1 and §4.2. We prove the uniqueness of solutions to the fluid equations in §5, and show the convergence of the fluid-scaled processes to the fluid limit in §6. Some additional proofs are collected in the Appendix.

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, $1\!\!1_B$ 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 $\text{supp}(\varphi)$ be the support of φ. For $H \le \infty$, 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 [31, 37], 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 $x \in [0,\infty)$.

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 [31],[37] so that they are Polish. Let $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ be the subset of non-decreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0) = 0$. 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 $X_n \Rightarrow X$.

2 Descriptions of Model and System Dynamics

2.1 Model Description 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 here we do 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 $N \to \infty$.

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 $t > 0$, $E_k^{(N)}(t)$ 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 $E_k^{(N)}$ is a non-decreasing, pure jump process with $E_k^{(N)}(0) = 0$ and a.s., for each $t \in [0,\infty)$, $E_k^{(N)}(t) < \infty$ and $E_k^{(N)}(t) - E_k^{(N)}(t-) \in \{0,1\}$. Also, for each $k \in \mathcal{K}$, let $\mathcal{E}_k^{(N)}$ 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 $\{\mathcal{E}_k^{(N)}, k \in \mathcal{K}\}$ are only used for bookkeeping purposes to keep track of the indices of customers.

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 i.i.d. sequences of i.i.d. random variables $\{v_i^{1,k},\ i \in \mathbb{Z}\}$ and $\{v_i^{2,k},\ i \in \mathbb{Z}\}$, with common cumulative distribution function G_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_k^s has density g_k^s. Let

$$C_k^s \doteq \sup\{x \in [0, \infty) : G_k^s(x) = 0\} \text{ and } H_k^s \doteq \sup\{x \in [0, \infty) : G_k^s(x) < 1\}.$$

Then C_k^s and H_k^s denote, respectively, the left end and the right end of the support of g_k^s. We assume that the service time distribution G_k^s has positive finite mean, that is,

$$m_k^s \doteq \int_{[0,H_k^s)} (1 - G_k^s(x))dx \in (0, \infty). \tag{2.1}$$

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 i.i.d. 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 cumulative distribution function $G_{1,k}^r$ 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 cumulative distribution function $G_{2,k}^r$ 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 $C^r_{j,k} \doteq \sup\{x \in [0, \infty) : G^r_{j,k}(x) = 0\}$ and $H^r_{j,k} \doteq \sup\{x \in [0, \infty) : G^r_{j,k}(x) < 1\}$ denote, respectively, the left end and the right end of the support of $g^r_{j,k}$. For each $j = 1, 2$, the mean of the patience time distribution $G^r_{j,k}$ is denoted by

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

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 System Dynamics

For each $k \in \mathcal{K}$, we shall use two measure-valued processes to describe the queue dynamics for external and internal customers of class k, respectively.

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

$$\eta^{(N),1,k}_t = \sum_{j=-\mathcal{E}^{(N)}_k+1}^{E^{(N)}_k(t)} \delta_{w^{(N),1,k}_j(t)} 1\!\!1_{\{w^{(N),1,k}_j(t) < r^{1,k}_j\}}, \tag{2.3}$$

where $w^{(N),1,k}_j(t) = \left[t - \zeta^{(N),1,k}_j\right] \vee 0 \wedge r^{1,k}_j$ 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 $\eta^{(N),2,k}_t$ to describe the potential queue dynamics for internal customers of class k. Specifically, let $\mathcal{C}^{(N)}_k$ 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 $\zeta^{(N),2,k}_j$ denote the time at which internal customer j of class k reenters the system upon service completion. Moreover, let $I^{(N)}_k(t)$ 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 $G^s_k,\ k \in \mathcal{K}$, 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, $I^{(N)}_k(t) - I^{(N)}_k(t-) \in \{0, 1\}$ for each $t \geq 0$ with probability one. Then $\eta^{(N),2,k}_t$ on $[0, H^r_{2,k})$ can be defined as

$$\eta^{(N),2,k}_t = \sum_{j=-\mathcal{C}^{(N)}_k+1}^{I^{(N)}_k(t)} \delta_{w^{(N),2,k}_j(t)} 1\!\!1_{\{w^{(N),2,k}_j(t) < r^{2,k}_j\}}, \tag{2.4}$$

where $w^{(N),2,k}_j(t) = \left[t - \zeta^{(N),2,k}_j\right] \vee 0 \wedge r^{2,k}_j$ 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

$$\eta_t^{(N),k} = \eta_t^{(N),1,k} + \eta_t^{(N),2,k}. \tag{2.5}$$

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 $X_k^{(N)}(t)$ be the total number of customers of class k (including external customers and internal customers) in the system and $Q_k^{(N)}(t)$ be the number of customers of class k waiting in queue at time t. Due to the non-idling condition, the queue length process $Q_k^{(N)}$ of customers of class k is then given by

$$Q_k^{(N)}(t) = [X_k^{(N)}(t) - N_k]^+. \tag{2.6}$$

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

$$\chi_k^{(N)}(t) \doteq \inf \left\{ x > 0 : \; \eta_t^{(N),k}[0,x] \geq Q_k^{(N)}(t) \right\} \tag{2.7}$$

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 $Q_k^{(N)}$ admits the following alternative representation in terms of $\chi_k^{(N)}$ and $\eta^{(N),k}$:

$$Q_k^{(N)}(t) \;\; = \;\; \eta_t^{(N),k}[0, \chi_k^{(N)}(t)]. \tag{2.8}$$

For $t \in [0,\infty)$, in a fashion analogous to (2.3) and (2.4), we can also define $\nu_t^{(N),k}$ to be a discrete non-negative Borel measure on $[0, H_k^s)$ that has a unit mass at the amount of time each of the customers (either external or internal) has spent in service by time t. For each $j \in \mathbb{Z}$, let $\varsigma_j^{(N),1,k}$ denote the time external customer j of class k enters service and $\varsigma_j^{(N),2,k}$ denote the time 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, the measure $\nu_t^{(N),k}$ is defined as

$$\nu_t^{(N),k} \;\; = \;\; \nu_t^{(N),1,k} + \nu_t^{(N),2,k}, \tag{2.9}$$

where

$$\nu_t^{(N),1,k} = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \delta_{a_j^{(N),1,k}(t)} \mathbb{1}_{\left\{a_j^{(N),1,k} < v_j^{1,k}\right\}} \tag{2.10}$$

and

$$\nu_t^{(N),2,k} = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \delta_{a_j^{(N),2,k}(t)} \mathbb{1}_{\left\{a_j^{(N),2,k} < v_j^{2,k}\right\}}. \tag{2.11}$$

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 $D_{kl}^{(N)}$, $l \in \mathcal{K} \cup \{0\}$, denote the cumulative routing processes, where for each $l \in \mathcal{K}$, $D_{kl}^{(N)}(t)$ is the cumulative number of customers of class k that have completed the service and joined class l in the time interval $[0,t]$, and $D_{k0}^{(N)}(t)$ is the cumulative number of customers of class k that have completed the service and left the system in the interval $[0,t]$. Then $D_{kl}^{(N)}(t)$ has the representation

$$\begin{aligned} D_{kl}^{(N)}(t) &= \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} \mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}} \\ &\quad + \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{\phi^{2,k}(j)=e_l\}} \mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}}. \end{aligned} \tag{2.12}$$

It is obvious that

$$I_k^{(N)}(t) = \sum_{l\in\mathcal{K}} D_{lk}^{(N)}(t). \tag{2.13}$$

In addition, the departure process $D_k^{(N)}$, where $D_k^{(N)}(t)$ represents the cumulative number of 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 $D_{kl}^{(N)}$, $l \in \mathcal{K} \cup \{0\}$, as

$$D_k^{(N)}(t) = \sum_{l\in\mathcal{K}\cup\{0\}} D_{kl}^{(N)}(t). \tag{2.14}$$

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

$$\begin{aligned} S_k^{(N)}(t) &= \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}} \\ &\quad + \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),2,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),2,k}}{dt}(s+)=0\right\}}. \end{aligned} \tag{2.15}$$

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

$$\begin{aligned} R_k^{(N)}(t) &= \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j^{(N),1,k}(s)\le\chi_k^{(N)}(s-),\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}} \\ &\quad + \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j^{(N),2,k}(s)\le\chi_k^{(N)}(s-),\frac{dw_j^{(N),2,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),2,k}}{dt}(s+)=0\right\}}, \end{aligned} \tag{2.16}$$

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 $k \in \mathcal{K}$, mass balances on the total number of customers of class k in the system, the number of customers of class k waiting in the "potential queue", and the number of customers of class k in service, show that

$$X_k^{(N)}(0) + E_k^{(N)} + I_k^{(N)} = X_k^{(N)} + R_k^{(N)} + D_k^{(N)}, \tag{2.17}$$

$$\langle \mathbf{1}, \eta_0^{(N),k} \rangle + E_k^{(N)} + I_k^{(N)} = \langle \mathbf{1}, \eta^{(N),k} \rangle + S_k^{(N)}, \tag{2.18}$$

and

$$\langle \mathbf{1}, \nu_0^{(N),k} \rangle + L_k^{(N)} = \langle \mathbf{1}, \nu^{(N),k} \rangle + D_k^{(N)}, \tag{2.19}$$

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

$$X_k^{(N)} = \langle \mathbf{1}, \nu^{(N),k} \rangle + Q_k^{(N)}. \tag{2.20}$$

Combining (2.17), (2.19) and (2.20), we obtain the following mass balance equation for the number of customers in queue:

$$Q_k^{(N)}(0) + E_k^{(N)} + I_k^{(N)} = Q_k^{(N)} + R_k^{(N)} + L_k^{(N)}. \tag{2.21}$$

Furthermore, the non-idling condition takes the form $N_k - \langle \mathbf{1}, \nu^{(N),k} \rangle = [N_k - X_k^{(N)}]^+$. Note that if $N_k = \infty$, then the above non-idling condition holds automatically.

3 The Fluid Model

In this section, we present the fluid model that asymptotically describes the system dynamics in §3.1, and characterize the invariant states for the fluid model in §3.2.

3.1 Fluid Model Equations

Recall that $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ denote the subset of non-decreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0)=0$. Define

$$\mathcal{S}_0 \doteq \left\{ \begin{array}{r} (e, x, \nu, \eta^1, \eta^2) \in \mathcal{I}_{\mathbb{R}_+}[0,\infty)^K \times \mathbb{R}_+^K \times \Pi_{k\in\mathcal{K}} \mathcal{M}_F[0, H_k^s) \times \\ \Pi_{k\in\mathcal{K}} \mathcal{M}_+[0, H_{1,k}^r) \times \Pi_{k\in\mathcal{K}} \mathcal{M}_+[0, H_{2,k}^r) : \\ s_k - \langle \mathbf{1}, \nu^k \rangle = [s_k - x_k]^+, \ [x_k - s_k]^+ \le \langle \mathbf{1}, \eta_0^{1,k} \rangle + \langle \mathbf{1}, \eta_0^{2,k} \rangle \end{array} \right\}, \tag{3.1}$$

where $\mathcal{M}_+[0, H_{1,k}^r)$ and $\mathcal{M}_+[0, H_{2,k}^r)$ represent the set of positive, locally finite measures on $[0, H_{1,k}^r)$ and $[0, H_{2,k}^r)$, respectively. 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 $G_{j,k}^r$ and G_k^s in the usual manner:

$$h_{j,k}^r(x) \ \doteq \ \frac{g_{j,k}^r(x)}{1 - G_{j,k}^r(x)}, \qquad x \in [0, H_{j,k}^r), \tag{3.2}$$

$$h^s_k(x) \doteq \frac{g^s_k(x)}{1 - G^s_k(x)}, \qquad x \in [0, H^s_k). \tag{3.3}$$

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

Definition 3.1 *The càdlàg function $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ defined on $\mathbb{R}_+$ such that $\overline{X} = (\overline{X}_k, k \in \mathcal{K}) \in \mathbb{R}^K_+$, $\overline{\nu} = (\overline{\nu}^k, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^s_k)$, $\overline{\eta}^1 = (\overline{\eta}^{1,k}, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_+[0, H^r_{1,k})$, 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 data $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$ and the hazard rate functions $h^r_{j,k}$ and h^s_k, $j = 1, 2$ and $k \in \mathcal{K}$, if and only if for every $t \in [0, \infty)$, $j = 1, 2$ and $k \in \mathcal{K}$,*

$$\int_0^t \langle h^r_{j,k}, \overline{\eta}^{j,k}_s \rangle \, ds < \infty, \qquad \int_0^t \langle h^s_k, \overline{\nu}^k_u \rangle \, du < \infty, \tag{3.4}$$

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

$$\begin{aligned}\int_{[0,H^s_k)} f(x)\, \overline{\nu}^k_t(dx) &= \int_{[0,H^s_k)} f(x+t) \frac{1 - G^s_k(x+t)}{1 - G^s_k(x)} \, \overline{\nu}^k_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)(1 - G^s_k(t-s)) \, d\overline{L}_k(s),\end{aligned} \tag{3.5}$$

where

$$\overline{L}_k(t) = \langle \mathbf{1}, \overline{\nu}^k_t \rangle - \langle \mathbf{1}, \overline{\nu}^k_0 \rangle + \int_0^t \langle h^s_k, \overline{\nu}^k_u \rangle \, du; \tag{3.6}$$

$$\begin{aligned}\int_{[0,H^r_{1,k})} f(x)\, \overline{\eta}^{1,k}_t(dx) &= \int_{[0,H^r_{1,k})} f(x+t) \frac{1 - G^r_{1,k}(x+t)}{1 - G^r_{1,k}(x)} \, \overline{\eta}^{1,k}_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)(1 - G^r_{1,k}(t-s)) \, d\overline{E}_k(s);\end{aligned} \tag{3.7}$$

$$\begin{aligned}\int_{[0,H^r_{2,k})} f(x)\, \overline{\eta}^{2,k}_t(dx) &= \int_{[0,H^r_{2,k})} f(x+t) \frac{1 - G^r_{2,k}(x+t)}{1 - G^r_{2,k}(x)} \, \overline{\eta}^{2,k}_0(dx) \\ &\quad + \int_0^t f(t-s)(1 - G^r_{2,k}(t-s)) \, d\overline{I}_k(s),\end{aligned} \tag{3.8}$$

where

$$\overline{I}_k(t) = \sum_{l\in\mathcal{K}} P_{lk} \int_0^t \langle h^s_l, \overline{\nu}^l_u \rangle \, du; \tag{3.9}$$

$$\overline{Q}_k(t) = \overline{X}_k(t) - \langle \mathbf{1}, \overline{\nu}^k_t \rangle; \tag{3.10}$$

$$\overline{Q}_k(t) \leq \langle \mathbf{1}, \overline{\eta}^{1,k}_t \rangle + \langle \mathbf{1}, \overline{\eta}^{2,k}_t \rangle; \tag{3.11}$$

$$\overline{R}_k(t) = \sum_{j=1}^2 \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds, \tag{3.12}$$

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

$$\overline{X}_k(t) = \overline{X}_k(0) + \overline{E}_k(t) + \overline{I}_k(t) - \int_0^t \langle h^s_k, \overline{\nu}^k_u \rangle \, du - \overline{R}_k(t); \tag{3.13}$$

and

$$s_k - \langle \mathbf{1}, \overline{\nu}^k_t \rangle = [s_k - \overline{X}_k(t)]^+. \tag{3.14}$$

It immediately follows from (3.10) and (3.14) that for each $t \in [0, \infty)$,

$$\overline{Q}_k(t) = [\overline{X}_k(t) - s_k]^+. \tag{3.15}$$

For future use, we also observe that (3.6), (3.10) and (3.13), when combined, show that for every $t \in [0, \infty)$ and $k \in \mathcal{K}$,

$$\overline{Q}_k(0) + \overline{E}_k(t) + \overline{I}_k(t) = \overline{Q}_k(t) + \overline{L}_k(t) + \overline{R}_k(t). \tag{3.16}$$

Remark 3.2 *It follows from (3.9) and (3.12) that for each $k \in \mathcal{K}$, processes $\overline{I}_k$ and $\overline{R}_k$ are continuous. If the process $\overline{E}_k$ is also continuous for each $k \in \mathcal{K}$, the relation in (3.13) implies that the process $\overline{X}_k$ and then $\overline{Q}_k$ by (3.10) are continuous. Moreover, for each $k \in \mathcal{K}$, the continuity of $\overline{E}_k$ also implies that $\overline{L}_k$ is continuous by (3.16).*

Remark 3.3 *For each $k \in \mathcal{K}$, if $s_k = \infty$, the non-idling condition (3.14) holds automatically and in this case, $\overline{Q}_k(t) = \overline{\chi}_k(t) = \overline{R}_k(t) = 0$ for all $t \geq 0$ by (3.15).*

We now state the uniqueness result for the solutions to the fluid model equations. Its proof is deferred to §5.

Theorem 3.4 *Suppose that for each $k \in \mathcal{K}$, $g^r_{2,k}$ is continuously differentiable on its support $[C^r_{2,k}, H^r_{2,k})$ and $h^r_{1,k}$ is locally bounded. Given any $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$ such that $\overline{E}$ is continuous and $\overline{\eta}^1_0, \overline{\eta}^2_0$ are continuous measures, there exists at most one solution $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ to the associated fluid equations* (3.4) – (3.14).

3.2 Invariant States

Given a positive constant vector $\lambda = (\lambda_k : k \in \mathcal{K})$, a state $(x_0, \nu_0, \eta^1_0, \eta^2_0)$ such that $(e_\lambda, x_0, \nu_0, \eta^1_0, \eta^2_0) \in \mathcal{S}_0$ and η^1_0, η^2_0 are continuous on $\mathbb{R}_+$ is said to be an *invariant state* for the fluid model equations in Definition 3.1 if there is a solution $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ to the fluid model equations associated with the initial data $(e_\lambda, x_0, \nu_0, \eta^1_0, \eta^2_0)$ satisfies $(\overline{X}(t), \overline{\nu}_t, \overline{\eta}^1_t, \overline{\eta}^2_t) = (x_0, \nu_0, \eta^1_0, \eta^2_0)$ for all $t \geq 0$, where $e_\lambda(t) = \lambda t$ for each $t \geq 0$.

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

$$\nu^k_*[0, x) = \int_0^x (1 - G^s_k(y))dy, \quad x \in [0, H^s_k), \tag{3.17}$$

$$\eta^{j,k}_*[0, x) = \int_0^x (1 - G^r_{j,k}(y))dy, \quad x \in [0, H^r_{j,k}). \tag{3.18}$$

Note that (2.1) implies that ν^k_* is actually a finite measure. Define the effective/overall arrival rate of customers of class k, $\bar{\lambda}_k$, as the kth entry of the vector $\bar{\lambda}$ defined by

$$\bar{\lambda} = (I - P')^{-1}\lambda = H\lambda. \tag{3.19}$$

Define $\mathcal{K}_1 := \{k \in \mathcal{K} : \ \bar{\lambda}_k m^s_k \geq s_k\}$, the set of potentially critically loaded or overloaded service stations in the absence of impatience. Note that $k \notin \mathcal{K}_1$ if $s_k = \infty$, which says that any service station with infinite servers can not be potentially critically loaded or overloaded. Let z and χ be K-dimensional vectors satisfying the equation

$$z = (\lambda - g(\chi)) + (I - G(\chi))P'z, \tag{3.20}$$

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

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

$$\mathcal{I}_\lambda = \{(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*),\ z^* \in \mathcal{Z},\ w^* = P'z^*,\ x^* \in \mathcal{X}_{z^*}\}, \tag{3.21}$$

where

$$\mathcal{Z} = \left\{ z \in \mathbb{R}^K_+ : \quad \begin{array}{c} \exists\ \mathcal{J} \subseteq \mathcal{K}_1 \text{ such that } z_k m^s_k = s_k \text{ for } k \in \mathcal{J}, \\ z_k m^s_k < s_k \text{ for } k \in \mathcal{K} \setminus \mathcal{J} \text{ and there exists } \chi \in \mathbb{R}^K_+ \text{ such that} \\ \chi_k = 0 \text{ for each } k \in \mathcal{K} \setminus \mathcal{J} \text{ and } (z, \chi) \text{ satisfies (3.20)} \end{array} \right\}, \tag{3.22}$$

and for each $z \in \mathcal{Z}$,

$$\mathcal{X}_z = \left\{ x \in \mathbb{R}^K_+ : \quad \begin{array}{c} z_k = \lambda_k(1 - G^r_{1,k}((F^{\eta^k})^{-1}((x_k - s_k)^+))) \\ \quad +w_k(1 - G^r_{2,k}((F^{\eta^k})^{-1}((x_k - s_k)^+))) \text{ for } k \in \mathcal{J}; \\ x_k = z_k m^s_k \text{ for } k \in \mathcal{K} \setminus \mathcal{J}, w = P'z,\ \eta = \lambda\eta^1_* + w\eta^2_* \end{array} \right\}. \tag{3.23}$$

It is clear that, if $\mathcal{K}_1 = \emptyset$, then $\mathcal{Z}$ contains only one element $z = \bar{\lambda}$. On the other hand, if $\mathcal{K}_1 \neq \emptyset$, the set $\mathcal{J}$ in (3.22) should be non-empty. Note that for each $k \in \mathcal{K}_1$, 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 (3.22) could be a strict subset of $\mathcal{K}_1$. In general, the set $\mathcal{Z}$ may contain several elements depending on the choices of $\mathcal{J}$.

Let $\mathcal{J}$ be a non-empty subset of $\mathcal{K}_1$. Define $z^{\mathcal{J}}$ and $\chi^{\mathcal{J}}$ as follows. Let $z^{\mathcal{J}}_j = s_j/m^s_j$ for each $j \in \mathcal{J}$ and $\chi^{\mathcal{J}}_j = 0$ for each $j \notin \mathcal{J}$. For each $j \notin \mathcal{J}$, we have from (3.20) that

$$z^{\mathcal{J}}_j = \lambda_j + \sum_{l \in \mathcal{J}} P_{lj} s_l/m^s_l + \sum_{l \notin \mathcal{J}} P_{lj} z^{\mathcal{J}}_l.$$

Let $P_{\mathcal{J}^c}$ be the principal submatrix of P by removing rows and columns of P with indexes in $\mathcal{J}$. It is clear from the properties of P and Lemma 7.5 of [5] that $I - P'_{\mathcal{J}^c}$ is also invertible and its inverse $H_{\mathcal{J}^c} \doteq (I - P'_{\mathcal{J}^c})^{-1}$ has the representation

$$H_{\mathcal{J}^c} = I + P'_{\mathcal{J}^c} + (P'_{\mathcal{J}^c})^2 + (P'_{\mathcal{J}^c})^3 + \cdots.$$

Let $z^{\mathcal{J}}_{\mathcal{J}^c}$ and $\lambda_{\mathcal{J}^c}$ be the vectors obtained from $z^{\mathcal{J}}$ and λ by removing entries with indexes in $\mathcal{J}$, respectively. Let $P_{\mathcal{J}\mathcal{J}^c}$ be the submatrix of P by removing rows of P with indexes not in $\mathcal{J}$ and columns of P with indexes in $\mathcal{J}$ and let $s_{\mathcal{J}}/m^s_{\mathcal{J}}$ be a vector whose entries are $\{s_l/m^s_l, l \in \mathcal{J}\}$. Then

$$z^{\mathcal{J}}_{\mathcal{J}^c} = H_{\mathcal{J}^c}(\lambda_{\mathcal{J}^c} + (P_{\mathcal{J}\mathcal{J}^c})' s_{\mathcal{J}}/m^s_{\mathcal{J}}).$$

Given the $z^{\mathcal{J}}$ defined above, for each $j \in \mathcal{J}$, by (3.20), we have

$$z^{\mathcal{J}}_j = \lambda_j(1 - G^r_{1,j}(\chi_j)) + (1 - G^r_{2,j}(\chi_j))(P'z^{\mathcal{J}})_j,$$

and let $\chi^{\mathcal{J}}_j$ be any solution to the above equation. This completes the definition of $\chi^{\mathcal{J}}$. So it is clear that the $z^{\mathcal{J}}$ defined above is in $\mathcal{Z}$ if $z^{\mathcal{J}}_{\mathcal{J}^c} < s_{\mathcal{J}}/m^s_{\mathcal{J}}$.

For example, consider $\mathcal{Z}$ when $K = 2$, $m^s_1 = m^s_2 = 1$,

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix}, \quad \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} 1/6 \\ 5/6 \end{bmatrix}, \text{ and } P = \begin{bmatrix} 1/2 & 1/4 \\ 1/4 & 1/2 \end{bmatrix}.$$

It is easy to check that

$$\left[\begin{array}{c}\bar{\lambda}_1 m_1^s\\ \bar{\lambda}_2 m_2^s\end{array}\right]=\left[\begin{array}{c}1\\ 1\end{array}\right]>\left[\begin{array}{c}s_1\\ s_2\end{array}\right].$$

Thus, $\mathcal{K}_1=\{1,2\}$. Choose $\mathcal{J}=\mathcal{K}_1$. Let $z_1=1/6$ and $z_2=5/6$. Obviously, $z_1 m_1^s \ge s_1$, $z_2 m_2^s \ge s_2$ and (3.20) has a solution for χ_1 and χ_2. Then $z=(1/6,5/6)' \in \mathcal{Z}$. Next, choose $\mathcal{J}=\{1\}$. Let $z_1=1/6$ and $\chi_2=0$. Then we can solve (3.20) for z_2 and χ_1 to get $z_2=7/12$ and χ_1 is a solution to the equation $\frac{11}{48}G_{2,1}^r(\chi_1)+\frac{1}{4}G_{1,1}^r(\chi_1)=\frac{5}{16}$. Note that $z_2 m_2^s=7/12<5/6$. Thus, $z=(1/6,7/12)' \in \mathcal{Z}$. At last, choose $\mathcal{J}=\{2\}$. Let $z_2=5/6$ and $\chi_1=0$. Then we can solve (3.20) for z_1 and χ_2 to get $z_1=11/12$ and χ_2 is a solution to the equation $12G_{1,2}^r(\chi_2)+31G_{2,2}^r(\chi_2)=3$. Note that $z_1 m_1^s=11/12>1/6$. Thus, $z=(11/12,5/6)' \notin \mathcal{Z}$. Therefore, in this example, $\mathcal{Z}$ has two elements.

Remark 3.5 *There are two possible sources for non-uniqueness of the invariant states in* (3.21)-(3.23) *associated with the fluid equations. One source comes from the fact that $\mathcal{Z}$ may have several elements, as shown in the above example. The other comes from* (3.23)*, that is, for a given $z\in\mathcal{Z}$, $\mathcal{X}_z$ may have more than one element due to the fact that the patience time distributions $G_{1,k}^r$ and $G_{2,k}^r$, $k\in\mathcal{K}$, may not be strictly increasing. This second source is the same for the single-station model in [15] where non-uniqueness is solely due to that the single patience distribution is not strictly increasing.*

Remark 3.6 *When customers in the system have infinite patience, that is, $G_{j,k}^r(x)=0$ for all $x\in[0,\infty)$, $k\in\mathcal{K}$ and $j=1,2$, and $\mathcal{K}_1=\mathcal{K}=\{k\in\mathcal{K}:\ \bar{\lambda}_k m_k^s>s_k\}$, 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 $z\in\mathcal{Z}$, there exists $\chi\in\mathbb{R}_+^K$ such that (z,χ) satisfies (3.20). Note that, in this case, (3.20) is reduced to the equation $z=\lambda+P'z$ and this implies that $z=\bar{\lambda}$. Since $\bar{\lambda}_k m_k^s>s_k$ for each $k\in\mathcal{K}$, $\mathcal{J}$ in (3.22) does not exist. This is a contradiction to the fact that $z\in\mathcal{Z}$. Thus, when all stations are overloaded, the set $\mathcal{Z}$ in (3.22) is an empty set, and so is $\mathcal{I}_\lambda$.*

Remark 3.7 *If the system has only one critically loaded or overloaded service station, that is, $\mathcal{K}_1$ has only one element, or every service station in the system is underloaded, that is, $\mathcal{K}_1=\emptyset$, then the set $\mathcal{Z}$ can only have a single element. If, in addition, the patience time distributions $G_{1,k}^r$ and $G_{2,k}^r$, $k\in\mathcal{K}$, are strictly increasing, then the system has only one invariant state.*

Remark 3.8 *If the system has feed-forward routing, that is, the routing matrix P has the property that $P_{ij}=0$ for each $i\ge j$, then the set $\mathcal{Z}$ also has a unique element and hence the system has unique invariant state if the patience time distributions are all strictly increasing. In fact, if $\mathcal{K}_1\neq\emptyset$, let $\mathcal{K}_1$ be the increasingly ordered set $\{k_1,k_2,\cdots,k_n\}$ and $\mathcal{J}$ be a non-empty subset of $\mathcal{K}_1$, which is associated with an invariant state. Note that all service stations with indices less than k_1 are underloaded, then $z_i=\bar{\lambda}_i$ for each $i<k_1$. Since service station k_1 is the first critically loaded or overloaded station, then $z_{k_1}=s_{k_1}/m_{k_1}^s$ and $k_1\in\mathcal{J}$. For any service station $k_1<i<k_2$, it will remain underloaded since the input rate from service station k_1 to service station i is less than the effective input rate due to abandonment at service station k_1 and then $z_i=\lambda_i+\sum_{j<i,j\neq k_1}H_{ij}\lambda_j+H_{ik_1}s_{k_1}/m_{k_1}^s$. For service station k_2, we know that the effective arrival rate is $\lambda_{k_2}+\sum_{j<k_2}H_{ik_2j}\lambda_j\ge s_{k_2}/m_{k_2}^s$ and the actual arrival rate is $\lambda_{k_2}+\sum_{j<k_2,j\neq k_1}H_{ij}\lambda_j+H_{k_2k_1}s_{k_1}/m_{k_1}^s$. So if $\lambda_{k_2}+\sum_{j<k_2,j\neq k_1}H_{ij}\lambda_j+H_{k_2k_1}s_{k_1}/m_{k_1}^s\ge s_{k_2}/m_{k_2}^s$, then $k_2\in\mathcal{J}$; otherwise, $k_2\notin\mathcal{J}$. Note that whether k_2 is in $\mathcal{J}$ or not depends only on λ_j,s_j,m_j^s for $j\in\mathcal{K}$ and P. By a similar argument for the rest of the stations in $\mathcal{K}_1$, we can see that the choice of $\mathcal{J}$ is unique based on λ_j,s_j,m_j^s for*

$j \in \mathcal{K}$ and P. Thus, there is a unique element in $\mathcal{Z}$. If $\mathcal{X}_z$ also has a unique element, for example, when all patience time distributions are strictly increasing, then the system has a unique invariant state.

Theorem 3.9 (*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 (3.4) − (3.14).*

Proof. Fix the arrival rate vector $\lambda = (\lambda_k : k \in \mathcal{K})$ and let $e_\lambda(t) = \lambda t$ for each $t \geq 0$. We break the argument into the following two claims.
Claim 1. The invariant state is a subset of $\mathcal{I}_\lambda$. Let $(x_0, \nu_0, \eta_0^1, \eta_0^2)$ be an invariant state and $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ be the solution to the fluid equations associated with the initial data $(e_\lambda, x_0, \nu_0, \eta_0^1, \eta_0^2)$ that satisfies $(\overline{X}_t, \overline{\nu}_t, \overline{\eta}_t^1, \overline{\eta}_t^2) = (x_0, \nu_0, \eta_0^1, \eta_0^2)$ for all $t \geq 0$. We will show that
(i) $\eta_0^1 = \lambda \eta_*^1$,
(ii) $\nu_0 = z^* \nu_*$, $\eta_*^2 = w^* \eta_*^2$, and $x_0 = x^*$ for $z^* \in \mathcal{Z}$, $w^* = P' z^*$ and $x^* \in \mathcal{X}_{z^*}$.

To establish (i), fix $k \in \mathcal{K}$. Since $\overline{\eta}_t^{1,k} = \eta_0^{1,k}$ for each $t \geq 0$, by (3.7), we see that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\begin{aligned}\int_{[0,H_{1,k}^r)} f(x)\, \overline{\eta}_0^{1,k}(dx) \quad &= \quad \int_{[0,H_{1,k}^r)} f(x+t) \frac{1 - G_{1,k}^r(x+t)}{1 - G_{1,k}^r(x)}\, \overline{\eta}_0^{1,k}(dx) \\ &\qquad + \lambda_k \int_0^t f(t-s)(1 - G_{1,k}^r(t-s))\, ds.\end{aligned}$$

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

$$\int_{[0,H_{1,k}^r)} f(x)\, \overline{\eta}_0^{1,k}(dx) \quad = \quad \lambda_k \int_{[0,H_{1,k}^r)} f(s)(1 - G_{1,k}^r(s))\, ds,$$

which implies that

$$\overline{\eta}_0^{1,k}(dx) = \lambda_k (1 - G_{1,k}^r(x))\, dx = \lambda_k \eta_*^{1,k}(dx). \tag{3.24}$$

This establishes (i).

Next, we focus on establishing (ii). Fix $k \in \mathcal{K}$. Since $\bar{X}_k(t) = x_0^k$ for each $t \geq 0$, by (3.15), we have $\overline{Q}_k(t) = (x_0^k - s_k)^+$ for each $t \geq 0$. Since $\bar{\eta}_t^{j,k} = \eta_0^{j,k}$ for each $t \geq 0$ and $j = 1, 2$, we also have $\bar{\eta}_t^k = \bar{\eta}_t^{1,k} + \bar{\eta}_t^{2,k} = \eta_0^{1,k} + \eta_0^{2,k} = \eta_0^k$ for each $t \geq 0$. This implies, in particular, that for each $t \geq 0$,

$$\bar{\chi}_k(t) = (F^{\bar{\eta}_t^k})^{-1}(\overline{Q}_k(t)) = (F^{\eta_0^k})^{-1}((x_0^k - s_k)^+) = \overline{\chi}_k^0. \tag{3.25}$$

Note that $\bar{\chi}_k$ is a constant function, and thus by (3.12), we can express $\overline{R}_k(t) = (c_{1,k} + c_{2,k})t$, for $t \geq 0$, where by (3.24),

$$c_{1,k} = \int_{[0,H_{1,k}^r)} 1\!\!1_{[0,\overline{\chi}_k^0]}(u) h_{1,k}^r(u) \eta_0^{1,k}(du) = \lambda_k G_{1,k}^r(\overline{\chi}_k^0),$$

and

$$c_{2,k} = \int_{[0,H_{2,k}^r)} 1\!\!1_{[0,\overline{\chi}_k^0]}(u) h_{2,k}^r(u) \eta_0^{2,k}(du). \tag{3.26}$$

Let

$$w_k = \sum_{l \in \mathcal{K}} P_{lk} \langle h_l^s, \nu_0^l \rangle. \tag{3.27}$$

Then it follows from (3.9) that

$$\overline{I}_k(t) = w_k t, \qquad t \geq 0. \tag{3.28}$$

Thus, by (3.16), we obtain that $\bar{L}_k(t) = (\lambda_k + w_k - c_{1,k} - c_{2,k})t$, $t \geq 0$. Let $z_k = \lambda_k + w_k - c_{1,k} - c_{2,k}$. Now by letting $\bar{\nu}^k_t = \nu^k_0$ for each $t \geq 0$ in (3.5), we see that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\int_{[0,H^s_k)} f(x)\nu^k_0(dx) = \int_{[0,H^s_k)} f(x+t)\frac{1-G^s_k(x+t)}{1-G^s_k(x)}\nu^k_0(dx) + z_k \int_0^t f(u)(1-G^s_k(u))\,du. \tag{3.29}$$

Again, letting $t \to \infty$ on both sides of (3.29) and using the fact that f has compact support, we obtain

$$\int_{[0,H^s_k)} f(x)\nu^k_0(dx) = z_k \int_{[0,H^s_k)} f(u)(1-G^s_k(u))\,du. \tag{3.30}$$

This implies that

$$\nu^k_0(dx) = z_k(1-G^s_k(x))dx = z_k\nu^k_*(dx). \tag{3.31}$$

Since $\langle \mathbf{1}, \nu^k_0 \rangle \leq s_k$, we then have

$$z_k m^s_k \leq s_k. \tag{3.32}$$

It follows from (3.31) that for each $l \in \mathcal{K}$, $\langle h^s_l, \nu^l_0 \rangle = z_l$. Thus, plugging into (3.27), we have

$$w_k = \sum_{l \in \mathcal{K}} P_{lk} z_l. \tag{3.33}$$

Next, by letting $\overline{\eta}^{2,k}_t = \eta^{2,k}_0$ for each $t \geq 0$ in (3.8) and using (3.28), we see that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\int_{[0,H^r_{2,k})} f(x)\eta^{2,k}_0(dx) = \int_{[0,H^r_{2,k})} f(x+t)\frac{1-G^r_{2,k}(x+t)}{1-G^r_{2,k}(x)}\eta^{2,k}_0(dx) + w_k \int_0^t f(s)(1-G^r_{2,k}(s))\,ds, \tag{3.34}$$

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

$$\int_{[0,H^r_{2,k})} f(x)\eta^{2,k}_0(dx) = w_k \int_{[0,H^r_{2,k})} f(s)(1-G^r_{2,k}(s))\,ds.$$

This implies that

$$\eta^{2,k}_0(dx) = w_k(1-G^r_{2,k}(x))dx = w_k\eta^{2,k}_*(dx). \tag{3.35}$$

Combining the above with (3.26), we have $c_{2,k} = w_k G^r_{2,k}(\overline{\chi}^0_k)$. It follows that

$$z_k = \lambda_k(1-G^r_{1,k}(\overline{\chi}^0_k)) + w_k(1-G^r_{2,k}(\overline{\chi}^0_k)). \tag{3.36}$$

Thus, from (3.36) and (3.33), we see that z and $\overline{\chi}^0$ satisfy (3.20), that is, $z = (\lambda - g(\overline{\chi}^0)) + (I - G(\overline{\chi}^0))P'z$. As a consequence, we have

$$(I - (I - G(\overline{\chi}^0))P')z = \lambda - g(\overline{\chi}^0) \tag{3.37}$$

and

$$z = (\lambda - g(\overline{\chi}^0)) + (I - G(\overline{\chi}^0))P'z \leq \lambda + P'z. \tag{3.38}$$

From (3.38), it is easy to see that

$$z \leq (I - P')^{-1}\lambda = \bar{\lambda}. \tag{3.39}$$

For each $k \notin \mathcal{K}_1$, the above implies that $z_k m_k^s \leq \bar{\lambda}_k m_k^s < s_k$. Now, let $\mathcal{J} = \{k \in \mathcal{K}_1 : \ z_k m_k^s = s_k\}$. Thus for each $k \in \mathcal{K} \setminus \mathcal{J}$, by (3.32), $z_k m_k^s < s_k$, then $\langle \mathbf{1}, \nu_0^k \rangle < s_k$ and hence $\overline{\chi}_k^0 = 0$. Therefore, $z \in \mathcal{Z}$, which is defined in (3.22). Note that for each $z \in \mathcal{Z}$, we obtain $w = P'z$ by (3.33), and $x_0 \in \mathcal{X}_z$ by (3.36) and (3.25).

Claim 2. The set $\mathcal{I}_\lambda$ is a subset of the invariant state. Fix $z^* \in \mathcal{Z}$, $w^* = P'z^*$ and $x^* \in \mathcal{X}_{z^*}$. Choose a process $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ to be such that $(\overline{X}(t), \overline{\nu}_t, \overline{\eta}_t^1, \overline{\eta}_t^2) = (x^*, z^*\nu_*, \lambda\eta_*^1, w^*\eta_*^2)$ for each $t \geq 0$. We now show that $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ is a solution to the fluid equations associated with the initial data $(e_\lambda, x^*, z^*\nu_*, \lambda\eta_*^1, w^*\eta_*^2)$. It is evident to see that (3.4) holds for $z^*\nu_*$, $\lambda\eta_*^1$ and $w^*\eta_*^2$. For each $k \in \mathcal{K}$ and $t \geq 0$, let

$$\bar{I}_k(t) \doteq \sum_{l \in \mathcal{K}} P_{lk} \int_0^t \langle h_l^s, \overline{\nu}_u^l \rangle \, du = \sum_{l \in \mathcal{K}} P_{lk} \langle h_l^s, z_l^* \nu_*^l \rangle t = (P'z^*)_k t = w_k^* t. \tag{3.40}$$

and

$$\bar{L}_k(t) \doteq \int_0^t \langle h_k^s, \overline{\nu}_u^k \rangle \, du = \langle h_k^s, z_k^* \nu_*^k \rangle t = z_k^* t. \tag{3.41}$$

Thus, (3.6) and (3.9) are satisfied by $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$. Since $z^* \in \mathcal{Z}$, there exists $\mathcal{J}^* \subseteq \mathcal{K}_1$ and $\chi^* \in \mathbb{R}_+^K$ such that $z_k^* m_k^s = s_k$ for each $k \in \mathcal{J}^*$, $z_k^* m_k^s < s_k$ and $\chi_k^* = 0$ for each $k \in \mathcal{K} \setminus \mathcal{J}^*$ and (z^*, χ^*) satisfies (3.20). Since $x^* \in \mathcal{X}_{z^*}$, we have that for each $k \in \mathcal{K} \setminus \mathcal{J}^*$, $x_k^* = z_k^* m_k^s$ and for each $k \in \mathcal{J}^*$,

$$z_k^* = \lambda_k (1 - G_{1,k}^r((F^{\eta^k})^{-1}((x_k^* - s_k)^+))) + w_k^* (1 - G_{2,k}^r((F^{\eta^k})^{-1}((x_k^* - s_k)^+))), \tag{3.42}$$

where $\eta = \lambda\eta_*^1 + w^*\eta_*^2$. For each $k \in \mathcal{K} \setminus \mathcal{J}^*$ and $t \geq 0$, let $\overline{Q}_k(t) = \overline{R}_k(t) \doteq 0$. Since $\langle \mathbf{1}, \overline{\nu}_t^k \rangle = \langle \mathbf{1}, z_k^* \nu_*^k \rangle = z_k^* m_k^s$ for each $t \geq 0$, we have that (3.10)–(3.12) and (3.14) hold. Notice that for each $t \geq 0$,

$$(e_\lambda)_k(t) + \bar{I}_k(t) - \int_0^t \langle h_k^s, \overline{\nu}_u^k \rangle \, du = (\lambda_k + w_k^* - z_k^*) t = 0,$$

where the last equality holds due to the fact that (z^*, χ^*) satisfies (3.20). This shows that (3.13) holds. Next for each $k \in \mathcal{J}^*$ and $t \geq 0$, let $\overline{Q}_k(t) = x_k^* - s_k \geq 0$. Since (z^*, χ^*) satisfies (3.20), we have that

$$z_k^* = \lambda_k (1 - G_{1,k}^r(\chi_k^*)) + (1 - G_{2,k}^r(\chi_k^*)) w_k^*.$$

This and (3.42) imply that $\chi_k^* = (F^{\eta^k})^{-1}(x_k^* - s_k)$. For each $t \geq 0$, let

$$\overline{R}_k(t) \doteq (\lambda_k G_{1,k}^r(\chi_k^*) + w_k^* G_{2,k}^r(\chi_k^*)) t.$$

Hence, we have checked that (3.10)–(3.14) hold.

It remains to show that (3.5), (3.7) and (3.8) hold. But this can be readily verified using the expressions of $(\overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$, $\bar{I}$ in (3.40) and $\overline{K}$ in (3.41). Hence $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ is a solution to the fluid equations associated with the initial data $(e_\lambda, x^*, z^*\nu_*, \lambda\eta_*^1, w^*\eta_*^2)$. Then we have that $(x^*, z^*\nu_*, \lambda\eta_*^1, w^*\eta_*^2)$ is an invariant state. This established Claim 2 and completes the proof of Theorem 3.9. □

4 Two Special Cases

4.1 A Single Station with Immediate Feedback

In this section, we consider the special case of our model where there is a single station of finitely many servers with immediate feedback. The subscript or superscript k is omitted to simplify the notation. The Markovian routing probabilities $P_{10} = p$ and $P_{11} = 1 - p$, that is, after each service completion, with probability p the customer will leave the system and with probability $1 - p$ the customer will reenter the system. The effective arrival rate $\bar{\lambda} = (1 - P_{11})^{-1}\lambda = \lambda/p$. Without loss of generality, let $s_1 = 1$. The underloaded, critically loaded and overloaded regimes are determined by $\bar{\lambda} m^s < 1$, $\bar{\lambda} m^s = 1$, and $\bar{\lambda} m^s > 1$, respectively. As a special case of Theorem 3.9, we obtain the following theorem for the invariant states associated with the fluid equations for this model of a single station with immediate feedback.

Theorem 4.1 (*Invariant States for A Single Station with Immediate Feedback*)

(i) If the system is underloaded, $\lambda m^s < p$, or critically loaded, $\lambda m^s = p$, then the invariant state has a unique element $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ given by

$$x^* = \lambda m^s/p, \quad z^* = \lambda/p, \quad w^* = (1-p)z^* = (1-p)\lambda/p, \tag{4.1}$$

and the head-of-line waiting time $\bar{\chi}^* = 0$.

(ii) If the system is overloaded, $\lambda m^s > p$, then any invariant state $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ has the representation:

$$z^* = 1/m^s, \quad w^* = (1-p)z^* = (1-p)/m^s, \tag{4.2}$$

and

$$x^* \ = \ 1 + \lambda \int_0^{\bar{\chi}^*} (1 - G^r_1(u))du + \frac{1-p}{m^s} \int_0^{\bar{\chi}^*} (1 - G^r_2(u))du, \tag{4.3}$$

where the head-of-line waiting time $\bar{\chi}^*$ is a solution to the equation

$$\lambda m^s (1 - G^r_1(\bar{\chi}^*)) = 1 - (1-p)(1 - G^r_2(\bar{\chi}^*)). \tag{4.4}$$

(iii) In (ii), if, in addition, $G^r_1 = G^r_2 = G^r$ and $m^r_1 = m^r_2 = m^r$, then any invariant state has z^* and w^* in (4.2)*, and*

$$x^* = 1 + (\lambda + (1-p)/m^s) \int_0^{\bar{\chi}^*} (1 - G^r(u))du, \tag{4.5}$$

where the head-of-line waiting time $\bar{\chi}^*$ is a solution to the equation

$$G^r(\bar{\chi}^*) = 1 - (\lambda m^s + (1-p))^{-1}. \tag{4.6}$$

Note that the equation (4.6) may not have unique solution for $\bar{\chi}^*$. When the service and patience time distributions are all exponential, we obtain the following corollary and as can be easily seen, the invariant states are unique.

Corollary 4.2 *Assume that $G^s(x) = 1 - e^{-\mu x}$, and $G^r_j(x) = 1 - e^{-\theta_j x}$, $j = 1, 2$, with $\mu, \theta_j \in (0, \infty)$.*

(i) If the system is underloaded $\lambda/(p\mu) < 1$ or critically loaded $\lambda/(p\mu) = 1$, then the invariant state has a unique element $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ given by

$$x^* = \lambda/(p\mu), \quad z^* = \lambda/p, \quad w^* = (1-p)z^* = (1-p)\lambda/p, \tag{4.7}$$

and the head-of-line waiting time $\bar{\chi}^* = 0$.

(ii) If the system is overloaded $\lambda/(p\mu) > 1$, then the invariant state has a unique element:

$$z^* = \mu, \quad w^* = (1-p)\mu, \quad x^* = 1 + \frac{\lambda}{\theta_1}(1 - e^{-\theta_1 \bar{\chi}^*}) + \frac{(1-p)\mu}{\theta_2}(1 - e^{-\theta_2 \bar{\chi}^*}), \tag{4.8}$$

where the head-of-line waiting time $\bar{\chi}^*$ is the unique solution to the equation

$$(1-(1-p)e^{-\theta_2\bar{\chi}^*})\mu = \lambda e^{-\theta_1\bar{\chi}^*}. \tag{4.9}$$

(iii) In (ii), if, in addition, $\theta_1 = \theta_2 = \theta$, then the invariant state has a unique element: z^* and w^* are in (4.8), and

$$x^* = 1 + \frac{\lambda - p\mu}{\theta}. \tag{4.10}$$

The head-of-line waiting time

$$\bar{\chi}^* = \theta^{-1}\log(\lambda/\mu + (1-p)). \tag{4.11}$$

We give a numerical example in Table 1 to illustrate our results in the model with one service station. In the first two models, the interarrival, service and patience times are all exponential with parameter values $\lambda = 100$, $\mu = 1$, $p = 0.4$, $n = 200$, and with identical abandonment rate $\theta_1 = \theta_2 = 0.5$ in the first model and different abandonment rates $\theta_1 = 0.5$ and $\theta_2 = 1$ in the second model. In the third model, the arrival process is Poisson with rate $\lambda = 100$, the service time distribution is H_2 with density $f(x) = e^{-2x} + 3^{-1}e^{-2x/3}$ and mean 1, the patience time distribution of new customers is H_2 with density $g_1(x) = 0.5e^{-x} + 6^{-1}e^{-x/3}$ and mean 2, and the patience time distribution of feedback customers is E_2 with mean 2. In all these models, the systems are in the overloaded regime. We conducted simulations to validate the heavy-traffic approximations in Theorem 4.1 and Corollary 4.2, in particular, we compare the heavy-traffic approximations and the simulated steady-state values of the average queue sizes Q_1 and Q_2 of new customers and feedback customers, respectively, and the average waiting-time χ of the customer at the head-of-the-line. We see that the approximations match very well with the simulation results, and, more importantly, the impact of differentiated patience time distributions of new and feedback customers is evident by the results in the first two models. The values below the simulation values are the halfwidth for the 95% confidence interval. The simulation results are the estimates from one sample path in the time interval $[20, 50]$, and the halfwidths of the 95% confidence interval are obtained by running four more independent simulations and using Student t-distribution with three degrees of freedom for each model.

Table 1: Comparison of the fluid approximations and simulations in many-server models of a single station with immediate feedback and differentiated patience-time distributions.

Model	Q_1		Q_2		χ	
	Sim.	Approx.	Sim.	Approx.	Sim.	Approx.
$M/M/n + M/M$	18.2226	18.1818	21.5017	21.8182	0.1922	0.1906
$\theta_1 = \theta_2$	± 0.0624		± 0.0603		± 0.0054	
$M/M/n + M/M$	12.1571	12.0197	13.7560	13.9902	0.1259	0.1240
$\theta_1 \neq \theta_2$	± 0.0790		± 0.0691		± 0.0064	
$M/H_2/n + H_2/E_2$	24.9345	24.9607	32.1587	32.3598	0.2748	0.2726
	± 0.0581		± 0.0766		± 0.0065	

4.2 A Tandem Network of Two Many-Server Queues with Abandonment

In this section, we consider a second special case: a tandem network of two service stations of finitely many servers with abandonment, where customers at each service station have patience time distribution of positive finite mean. Here $K = 2$ and the Markovian routing probabilities

$P_{11} = 0$, $P_{12} = p$, $P_{10} = 1 - p$, $P_{20} = 1$, $P_{21} = 0$ and $P_{22} = 0$, that is, after service completion at the first station, customers move to the second queue with probability p and leave the system with probability $1 - p$, and after service completion at the second station, customers must leave the system. We assume that $p \in (0, 1]$. When $p = 1$, all customers completing service from the first station will join the second station. Let $m^r_{j,k}$ and $G^r_{j,k,e}$ be the mean and stationary excess distribution function associated with the distribution $G^r_{j,2}$ for $j = 1, 2$ and $k = 1, 2$. Recall that $j = 1$ for external arrivals and $j = 2$ for internal customers. As a special case of Theorem 3.9, we give the invariant states associated with the fluid equations for this tandem network model below. As in Corollary 4.2, it is also easy to verify that the invariant state is unique when the service and patience times are exponential.

Theorem 4.3 (*Invariant States for a tandem network of two many-server queues with abandonment*)

(*i*) *If* $\lambda_1 m^s_1 \le s_1$ *and* $(\lambda_1 p + \lambda_2)m^s_2 \le s_2$, *then the invariant state has a unique element* $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ *given by*

$$x^* = (x^*_1, x^*_2)' = (\lambda_1 m^s_1, (\lambda_1 p + \lambda_2)m^s_2)', \quad z^* = (z^*_1, z^*_2)' = (\lambda_1, \lambda_1 p + \lambda_2)', \tag{4.12}$$

$$w^* = (w^*_1, w^*_2)' = (0, \lambda_1 p)', \quad \bar{\chi}^* = (\bar{\chi}^*_1, \bar{\chi}^*_2)' = (0, 0)'. \tag{4.13}$$

(*ii*) *If* $\lambda_1 m^s_1 \le s_1$ *and* $(\lambda_1 p + \lambda_2)m^s_2 > s_2$, *then any invariant state* $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ *has the representation:*

$$z^* = (z^*_1, z^*_2)' = (\lambda_1, s_2/m^s_2)', \quad w^* = (w^*_1, w^*_2)' = (0, \lambda_1 p)', \quad x^*_1 = \lambda_1 m^s_1, \quad \bar{\chi}^*_1 = 0, \tag{4.14}$$

$$x^*_2 = s_2 + \lambda_1 p m^r_{2,2} G^r_{2,2,e}(\bar{\chi}^*_2) + \lambda_2 m^r_{1,2} G^r_{1,2,e}(\bar{\chi}^*_2), \tag{4.15}$$

where $\bar{\chi}^*_2$ *is a solution to the equation*

$$\lambda_1 p(1 - G^r_{2,2}(\bar{\chi}^*_2)) + \lambda_2(1 - G^r_{1,2}(\bar{\chi}^*_2)) = s_2/m^s_2. \tag{4.16}$$

(*iii*) *If* $\lambda_1 m^s_1 > s_1$ *and* $(s_1 p/m^s_1 + \lambda_2)m^s_2 \le s_2$, *then any invariant state* $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ *has the representation:*

$$z^* = (z^*_1, z^*_2)' = (s_1/m^s_1, s_1 p/m^s_1 + \lambda_2)', \quad w^* = (w^*_1, w^*_2)' = (0, s_1 p/m^s_1)', \tag{4.17}$$

$$x^*_1 = s_1 + \lambda_1 m^r_{1,1} G^r_{1,1,e}(\bar{\chi}^*_1), \quad x^*_2 = (s_1 p/m^s_1 + \lambda_2)m^s_2, \quad \bar{\chi}^*_2 = 0, \tag{4.18}$$

where $\bar{\chi}^*_1$ *is a solution to the equation*

$$\lambda_1(1 - G^r_{1,1}(\bar{\chi}^*_1)) = s_1/m^s_1. \tag{4.19}$$

(*iv*) *If* $\lambda_1 m^s_1 > s_1$ *and* $(s_1 p/m^s_1 + \lambda_2)m^s_2 > s_2$, *then any invariant state* $(x^*, z^*\nu_*, \lambda\eta^1_*, w^*\eta^2_*)$ *has the representation:*

$$z^* = (z^*_1, z^*_2)' = (s_1/m^s_1, s_2/m^s_2)', \quad w^* = (w^*_1, w^*_2)' = (0, s_1 p/m^s_1)', \tag{4.20}$$

x^*_1 *has the same expression in* (4.18) *with* $\bar{\chi}^*_1$ *being a solution to the equation* (4.19), *and* x^*_2 *is*

$$x^*_2 = s_2 + s_1 p m^r_{2,2} G^r_{2,2,e}(\bar{\chi}^*_2)/m^s_1 + \lambda_2 m^r_{1,2} G^r_{1,2,e}(\bar{\chi}^*_2), \tag{4.21}$$

where $\bar{\chi}^*_2$ *is a solution to the following equation*

$$s_1 p(1 - G^r_{2,2}(\bar{\chi}^*_2))/m^s_1 + \lambda_2(1 - G^r_{1,2}(\bar{\chi}^*_2)) = s_2/m^s_2. \tag{4.22}$$

5 Uniqueness of Solutions to the Fluid Model Equations

In this section, we prove the uniqueness of solutions to the fluid model equations, Theorem 3.4. The complexity of the fluid equations in the general network setting with internal routing that allows for feedback makes the argument more involved than the single station case without feedback in [14]. The main challenge here to establish the uniqueness of solutions to the fluid equations is to verify the uniqueness of $\overline{L}$, which uniquely determines $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$.

Let $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ be a solution to the fluid model equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$, where $\overline{E}$ is continuous and $\overline{\eta}^1_0, \overline{\eta}^2_0$ are continuous measures. Recall the definitions of $\overline{Q}_k$ and $\overline{R}_k$ that are given in (3.10) and (3.12). It follows from (3.7), (3.8), the continuity of $\overline{E}$, $\overline{\eta}^1_0$ and $\overline{\eta}^2_0$, and the continuity of $\overline{I}$ by Remark 3.2 that $\overline{\eta}^1_t$ and $\overline{\eta}^2_t$ are continuous measures on $\mathbb{R}_+$ for each $t \geq 0$. By Remark 3.2, we also have that the processes $\overline{I}, \overline{X}, \overline{Q}$ and $\overline{L}$ are all continuous. As an immediate consequence of (3.12), we have the following elementary property.

Lemma 5.1 *For each $k \in \mathcal{K}$ and any $0 \leq a \leq b < \infty$, if $\overline{Q}_k(t) = 0$ for all $t \in [a, b]$, then $\overline{R}_k(b) - \overline{R}_k(a) = 0$.*

Next, we establish the intuitive result that the process $\overline{L}$ that represents the cumulative entry of "fluid" into service is non-decreasing.

Lemma 5.2 *The function $\overline{L}_k$ is non-decreasing for each $k \in \mathcal{K}$.*

Proof. The proof of this lemma is essentially identical to Lemma 4.5 of [14] with s_k instead of 1 and the following estimate on $\overline{L}_k$.

$$\begin{aligned} \overline{L}_k(t) - \overline{L}_k(u) &= \overline{L}_k(t) - \overline{L}_k(u) + \overline{R}_k(t) - \overline{R}_k(u) + \overline{Q}_k(t) - \overline{Q}_k(u) \\ &= \overline{E}_k(t) - \overline{E}_k(u) + \overline{I}_k(t) - \overline{I}_k(u) \geq 0. \end{aligned}$$

■

Lemma 5.3 *Let $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ and $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ be two solutions to the fluid model equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$. For $\overline{L}$ and $\overline{L}^*$ given by (3.6) with $\overline{\nu}$ and $\overline{\nu}^*$, respectively, and $k \in \mathcal{K}$, define*

$$\tau_k = \inf\{t \geq 0 : \ \overline{L}_k(t) \neq \overline{L}^*_k(t)\}.$$

Let $\tau = \min_{k \in \mathcal{K}} \tau_k$. Then, $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ and $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ agree on $[0, \tau)$.

The proof of this lemma is similar to that of Theorem 4.6 in [14] with slight changes in handling customer feedback, and is included in the Appendix for completeness.

Lemma 5.4 *Under the suppositions of Lemma 5.3 and assumptions in Theorem 3.4, $\tau_k = \tau$ for each $k \in \mathcal{K}$.*

Proof. If $\tau = \infty$, the result is trivial. Henceforth, we assume that $\tau < \infty$. Let $\overline{Q}, \overline{L}, \overline{I}, \overline{R}, \overline{\eta}$ and $\overline{Q}^*, \overline{L}^*, \overline{I}^*, \overline{R}^*, \overline{\eta}^*$ be the processes described in Definition 3.1 that are associated with the solutions $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ and $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ to the fluid equations for $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$, respectively. Let $\triangle A$ denote $A^* - A$ for $A = \overline{Q}, \overline{L}, \overline{I}, \overline{R}$. For each $k \in \mathcal{K}$, $t \geq 0$ and $j = 1, 2$, let $\triangle \overline{\nu}^k_t$ and $\triangle \overline{\eta}^{j,k}_t$ be the measures that satisfy $\triangle \overline{\nu}^k_t(\Xi) = \overline{\nu}^{*,k}_t(\Xi) - \overline{\nu}^k_t(\Xi)$ and $\triangle \overline{\eta}^{j,k}_t(\Xi) = \overline{\eta}^{j,*,k}_t(\Xi) - \overline{\eta}^{j,k}_t(\Xi)$, respectively, for every measurable set $\Xi \subset [0, \infty)$.

We prove by contradiction that $\tau_k = \tau$ for each $k \in \mathcal{K}$. Suppose that there exists some $k \in \mathcal{K}$ such that $\tau < \tau_k$. Let $k_1 \in \mathcal{K}$ be an index such that $\tau < \tau_{k_1}$ and for all $k \in \mathcal{K}$ such that $\tau < \tau_k$, we have $\tau_{k_1} \leq \tau_k$. Let $\mathcal{K}_0 = \{k \in \mathcal{K} : \ \tau_k = \tau\}$. By Lemma 5.3 and an action of time-shifts on solutions to the fluid equations that is similar to Lemma 3.4 of [14], we may assume, without loss of generality, that $\tau = 0$. Note that for each $t \in [0, \tau_{k_1}]$, $\overline{L}_k(t) = \overline{L}^*_k(t)$ for each $k \notin \mathcal{K}_0$. It follows directly from (3.5) that for each $k \notin \mathcal{K}_0$ and $t \in [0, \tau_{k_1}]$, $\overline{\nu}^k_t = \overline{\nu}^{*,k}_t$. Now, we choose $\delta > 0$ (whose value will be determined later) and define

$$\sigma_k(\delta) \doteq \inf\{t \geq 0 : \ |\triangle\overline{L}_k(t)| \geq \delta\} \text{ for each } k \in \mathcal{K}_0, \text{ and } \sigma = \sigma(\delta) \doteq \min_{k \in \mathcal{K}_0} \sigma_k(\delta).$$

Note that $\sigma(\delta) \to 0$ as $\delta \to 0$ by the continuity of $\triangle\overline{L}_k$ and $\triangle\overline{L}_k(0) = 0$ for each $k \in \mathcal{K}_0$ and the definition of τ. We argue by contradiction to show that $\sigma \geq \tau_{k_1}$. Suppose that $\sigma < \tau_{k_1}$. Choose a $k_0 \in \mathcal{K}_0$ such that $\sigma_{k_0}(\delta) = \sigma$. Note that such k_0 may not be unique. The continuity of $\triangle\overline{L}_{k_0}$ implies that $|\triangle\overline{L}_{k_0}(\sigma)| \geq \delta$. Thus, we have either $\triangle\overline{L}_{k_0}(\sigma) \geq \delta$ or $\triangle\overline{L}_{k_0}(\sigma) \leq -\delta$. We shall just consider the case when $\triangle\overline{L}_{k_0}(\sigma) \geq \delta$, since the other case can be treated in the same way by considering $-\triangle\overline{L}_{k_0}$ in stead of $\triangle\overline{L}_{k_0}$. We show that $\sigma < \tau_{k_1}$ will lead us to a contradiction with $\triangle\overline{L}_{k_0}(\sigma) \geq \delta$.

In order to show that, we define

$$r \doteq \sup\Big\{t < \sigma : \ \overline{Q}^*_{k_0}(t) < \overline{Q}_{k_0}(t)\Big\} \vee 0.$$

We show the contradiction by the following three steps.

Step 1. Show that $\overline{X}_{k_0}(0) \leq s_{k_0}$ and $\overline{Q}_{k_0}(0) = 0$, and $\overline{X}_{k_0}(\sigma) = \langle \mathbf{1}, \overline{\nu}^{k_0}_\sigma\rangle < s_{k_0}$ and $\overline{Q}_{k_0}(\sigma) = 0$. We can write

$$\triangle\overline{L}_{k_0}(\sigma) = \triangle\overline{L}_{k_0}(r) + \left(\triangle\overline{L}_{k_0}(\sigma) - \triangle\overline{L}_{k_0}(r)\right). \tag{5.1}$$

Step 2. Show that

$$\triangle\overline{L}_{k_0}(r) \leq \delta G^s_{k_0}(r) \leq \delta G^s_{k_0}(\sigma). \tag{5.2}$$

Step 3. Show that

$$\triangle\overline{L}_{k_0}(\sigma) - \triangle\overline{L}_{k_0}(r) \leq \delta\Big((2 + 2C_\sigma + D_\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0} G^s_l(\sigma) + C_\sigma G^r_{2,k_0}(\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0}\Big), \tag{5.3}$$

where $C_\sigma = \sup_{0 \leq s \leq \sigma} h^r_{1,k_0}(s)$ and $D_\sigma = 2\sup_{0 \leq s \leq \sigma} g^r_{2,k_0}(s) + \int_{C^r_{2,k_0}}^{\sigma \wedge C^r_{2,k_0}} |(g^r_{2,k_0})'(v)| dv$. Thus, with (5.1)–(5.3), we have

$$\begin{aligned}
\triangle\overline{L}_{k_0}(\sigma) \quad &\leq \quad \delta G^s_{k_0}(\sigma) + \delta\Big((2 + 2C_\sigma + D_\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0} G^s_l(\sigma) + C_\sigma G^r_{2,k_0}(\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0}\Big) \\
&\leq \quad \delta\Big(G^s_{k_0}(\sigma) + (2 + 2C_\sigma + D_\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0} G^s_l(\sigma) + C_\sigma G^r_{2,k_0}(\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0}\Big).
\end{aligned}$$

By choosing δ small enough, we may assume, without loss of generality, that σ is small enough such that $\sigma < 1$ and

$$G^s_{k_0}(\sigma) + (2 + 2C_\sigma + D_\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0} G^s_l(\sigma) + C_\sigma G^r_{2,k_0}(\sigma)\sum_{l \in \mathcal{K}_0} P_{lk_0} \leq 1/2, \tag{5.4}$$

and therefore, $\triangle\overline{L}_{k_0}(\sigma) \leq \delta/2$, which contradicts $\triangle\overline{L}_{k_0}(\sigma) \geq \delta$. Thus, with these three steps, we can prove that $\sigma \geq \tau_{k_1}$. By letting $\delta \to 0$, we have $\overline{L}_{k_0}(t) = \overline{L}^*_{k_0}(t)$ for all $t \in [0, \tau_{k_1}]$. But this contradicts the definition of τ_{k_0}. Thus $\tau_k = \tau$ for all $k \in \mathcal{K}$.

The remaining of the proof focuses on establishing these three steps.

Establishment of Step 1. We first show by contradiction that $\overline{X}_{k_0}(0) \le s_{k_0}$ and hence $\overline{Q}_{k_0}(0) = 0$. Suppose that $\overline{X}_{k_0}(0) > s_{k_0}$. Since $\overline{X}_{k_0}(0) = \overline{X}_{k_0^*}(0)$, by the continuity of $\overline{X}_{k_0}$ and $\overline{X}^*_{k_0}$, we have $\langle \mathbf{1}, \overline{\nu}^{k_0}_t \rangle = \langle \mathbf{1}, \overline{\nu}^{*,k_0}_t \rangle = s_{k_0}$ for all t in a neighborhood of 0. It follows from (3.6) with k_0 in place of k and (3.5) with $h^s_{k_0}$ in place of f, we have that

$$\begin{aligned} \overline{L}_{k_0}(t) &= \int_0^t \langle h^s_k, \overline{\nu}^k_u \rangle \, du \qquad (5.5) \\ &= \int_0^t \int_{[0,H^s_{k_0})} \frac{g^s_{k_0}(x+s)}{1-G^s_{k_0}(x)} \, \overline{\nu}^{k_0}_0(dx) ds + \int_0^t \int_0^s g^s_{k_0}(s-u) \, d\overline{L}_{k_0}(u) ds \\ &= \int_{[0,H^s_{k_0})} \frac{G^s_{k_0}(x+t) - G^s_{k_0}(x)}{1-G^s_{k_0}(x)} \, \overline{\nu}^{k_0}_0(dx) + \int_0^t g^s_{k_0}(t-u) \overline{L}_{k_0}(u) \, du. \end{aligned}$$

The last equality in (5.5) follows from an exchange of order of integration and an application of integration by parts. The same argument shows that (5.5) holds for $\overline{L}^*_{k_0}$, then we have that

$$\triangle \overline{L}_{k_0}(t) = \int_0^t g^s_{k_0}(t-u) \triangle \overline{L}_{k_0}(u) \, du \text{ for all } t \text{ in a neighborhood of } 0.$$

Then by using the key renewal theorem (cf. Theorem 4.7 of [2]), we have that $\triangle \overline{L}_{k_0}(t) = 0$ for all t in a neighborhood of 0. Thus, $\tau_{k_0} > 0$, which is a contradiction. Since $\overline{X}_{k_0}(0) \le s_{k_0}$, it follows that $\overline{\chi}_k(0) = 0$. Notice that only the mass before $\overline{\chi}_k(0)$ in $\overline{\eta}^{1,k_0}_0$ and $\overline{\eta}^{2,k_0}_0$ affects the evolution of the fluid equations. Thus, without loss of generality, we may assume that

$$\overline{\eta}^{1,k_0}_0 = \overline{\eta}^{2,k_0}_0 = \mathbf{0}. \qquad (5.6)$$

By the definition of σ, we have that for each $t \in [0, \sigma]$, $|\triangle \overline{L}_{k_0}(t)| \le \delta$. We then claim that for each $t \in [0, \sigma]$,

$$\triangle \overline{L}_{k_0}(t) \le \delta G^s_k(\sigma) \text{ if } \langle \mathbf{1}, \overline{\nu}^{k_0}_t \rangle = s_{k_0}. \qquad (5.7)$$

To see why this is true, suppose that $\langle \mathbf{1}, \overline{\nu}^{k_0}_t \rangle = s_{k_0}$ for some $t \in [0, \sigma]$. Since $\langle \mathbf{1}, \overline{\nu}^{*,k_0}_t \rangle \le s_{k_0}$, we have $\langle \mathbf{1}, \triangle \overline{\nu}^{k_0}_t \rangle \le 0$. When combined with (5.5) and the identity $\triangle \overline{\nu}^{k_0}_0 = 0$, this shows that

$$\begin{aligned} \triangle \overline{L}_{k_0}(t) &= \langle \mathbf{1}, \triangle \overline{\nu}^{k_0}_t \rangle + \int_0^t g^s_{k_0}(t-s) \triangle \overline{L}_{k_0}(s) \, ds \qquad (5.8) \\ &\le \int_0^t g^s_{k_0}(t-s) |\triangle \overline{L}_{k_0}(s)| \, ds \le \delta G^s_{k_0}(t) \le \delta G^s_{k_0}(\sigma). \end{aligned}$$

Thus (5.7) follows. Then, suppose that $\langle \mathbf{1}, \overline{\nu}^{k_0}_\sigma \rangle = s_{k_0}$, (5.7) implies that $\triangle \overline{L}_{k_0}(\sigma) \le \delta G^s_k(\sigma) < \delta$. This is a contradiction. Combining this with (3.10) and (3.14), we have

$$\overline{X}_{k_0}(\sigma) = \langle \mathbf{1}, \overline{\nu}^{k_0}_\sigma \rangle < s_{k_0} \quad \text{ and } \quad \overline{Q}_{k_0}(\sigma) = 0. \qquad (5.9)$$

Establishment of Step 2. By the definition of r, for every $t \in [r, \sigma]$, $\overline{Q}^*_{k_0}(t) \ge \overline{Q}_{k_0}(t)$. If $r = 0$, then $\triangle \overline{L}_{k_0}(r) = \triangle \overline{L}_{k_0}(0) = 0 < \delta$. On the other hand, if $r > 0$, there exists a sequence of $\{t_n\}_{n=1}^\infty$ such that $t_n < r$ and $t_n \to r$ as $n \to \infty$ and $0 \le \overline{Q}^*_{k_0}(t_n) < \overline{Q}_{k_0}(t_n)$ for each $n \in \mathbb{N}$. Since $\overline{Q}^*_{k_0}$ and $\overline{Q}_{k_0}$ are continuous, this implies that

$$\overline{Q}^*_{k_0}(r) \le \overline{Q}_{k_0}(r). \qquad (5.10)$$

Further, since $\overline{Q}_{k_0}(t_n) > 0$ for all $n \in \mathbb{N}$, due to (3.10) and (3.14), it also follows that $\overline{X}_{k_0}(t_n) > \langle \mathbf{1}, \overline{\nu}^{k_0}_{t_n} \rangle = s_{k_0}$ for every $n \in \mathbb{N}$. Since $t_n < r \le \sigma$ and $\langle \mathbf{1}, \overline{\nu}^{k_0}_{t_n} \rangle = s_{k_0}$ for every $n \in \mathbb{N}$, (5.8) and the continuity of $\overline{L}_{k_0}$ and $\overline{L}^*_{k_0}$ imply (5.2).

Establishment of Step 3. Since (3.16) is satisfied with $(\overline{L}_k, \overline{R}_k, \overline{Q}_k, \overline{I}_k)$ replaced by $(\overline{L}_{k_0}, \overline{R}_{k_0}, \overline{Q}_{k_0}, \overline{I}_{k_0})$ and $(\overline{L}^*_{k_0}, \overline{R}^*_{k_0}, \overline{Q}^*_{k_0}, \overline{I}^*_{k_0})$, respectively, it follows that

$$\begin{aligned} 0 &= \triangle\overline{L}_{k_0}(\sigma) + \triangle\overline{R}_{k_0}(\sigma) + \triangle\overline{Q}_{k_0}(\sigma) - \triangle\overline{I}_{k_0}(\sigma) \\ &= \triangle\overline{L}_{k_0}(r) + \triangle\overline{R}_{k_0}(r) + \triangle\overline{Q}_{k_0}(r) - \triangle\overline{I}_{k_0}(r). \end{aligned}$$

Hence,

$$\triangle\overline{L}_{k_0}(\sigma) - \triangle\overline{L}_{k_0}(r) = -(\triangle\overline{R}_{k_0}(\sigma) - \triangle\overline{R}_{k_0}(r)) + (\triangle\overline{I}_{k_0}(\sigma) - \triangle\overline{I}_{k_0}(r)) - \triangle\overline{Q}_{k_0}(\sigma) + \triangle\overline{Q}_{k_0}(r).$$

Since $-\triangle\overline{Q}_{k_0}(\sigma) = \overline{Q}_{k_0}(\sigma) - \overline{Q}^*_{k_0}(\sigma) = -\overline{Q}^*_{k_0}(\sigma) \le 0$ due to (5.9) and $\triangle\overline{Q}_{k_0}(r) \le 0$ by (5.10), we obtain that

$$\triangle\overline{L}_{k_0}(\sigma) - \triangle\overline{L}_{k_0}(r) \le -(\triangle\overline{R}_{k_0}(\sigma) - \triangle\overline{R}_{k_0}(r)) + (\triangle\overline{I}_{k_0}(\sigma) - \triangle\overline{I}_{k_0}(r)). \tag{5.11}$$

We will next obtain upper bounds for the two terms in the right-hand side of (5.11). For the second term, first, it follows from (3.9) that for each $s \in [0, \sigma]$,

$$\triangle\overline{I}_{k_0}(s) = \sum_{l \in \mathcal{K}} P_{lk_0} \int_0^s \langle h^s_l, \triangle\overline{\nu}^l_u \rangle \, du = \sum_{l \in \mathcal{K}_0} P_{lk_0} \int_0^s \langle h^s_l, \triangle\overline{\nu}^l_u \rangle \, du. \tag{5.12}$$

The last equality in the above display is due to the fact that for each $k \notin \mathcal{K}_0$ and $u \in [0, \tau_{k_1}]$, $\overline{\nu}^k_u = \overline{\nu}^{*,k}_u$. It follows from (3.5), an application of change of order of integrations and an application of integration by parts that for each $s \in [0, \sigma]$ and $l \in \mathcal{K}_0$,

$$\begin{aligned} \left| \int_0^s \langle h^s_l, \triangle\overline{\nu}^l_u \rangle \, du \right| &= \left| \int_0^s \int_0^u g^s_l(u - v) \, d\triangle\overline{L}_l(v) \, du \right| \\ = \quad \left| \int_0^s g^s_l(s - u) \triangle\overline{L}_l(u) \, du \right| &\le \int_0^s g^s_l(s - u) \left| \triangle\overline{L}_l(u) \right| \, du \le \delta G^s_l(s). \end{aligned} \tag{5.13}$$

Then we have that

$$\triangle\overline{I}_{k_0}(\sigma) - \triangle\overline{I}_{k_0}(r) = \sum_{l \in \mathcal{K}_0} P_{lk_0} \int_r^\sigma \langle h^s_l, \triangle\overline{\nu}^l_s \rangle \, ds \le \delta \sum_{l \in \mathcal{K}_0} P_{lk_0} (G^s_l(\sigma) + G^s_l(r)) \le 2\delta \sum_{l \in \mathcal{K}_0} P_{lk_0} G^s_l(\sigma). \tag{5.14}$$

We next focus on the first term in the right-hand side of (5.11). It follows from (3.12) that

$$\begin{aligned} \triangle\overline{R}_{k_0}(\sigma) - \triangle\overline{R}_{k_0}(r) = \sum_{j=1}^2 \int_r^\sigma & \left(\int_{[0, H^r_{j,k_0})} \mathbb{1}_{[0, \overline{\chi}^*_{k_0}(s)]}(u) h^r_{j,k_0}(u) \overline{\eta}^{j,*,k_0}_s(du) \right) ds \\ - \sum_{j=1}^2 \int_r^\sigma & \left(\int_{[0, H^r_{j,k_0})} \mathbb{1}_{[0, \overline{\chi}_{k_0}(s)]}(u) h^r_{j,k_0}(u) \overline{\eta}^{j,k_0}_s(du) \right) ds. \end{aligned} \tag{5.15}$$

By (5.6), the support of $\overline{\eta}^{j,k_0}_s$ and $\overline{\eta}^{j,*,k_0}_s$ is contained in $[0, s]$ for all $s \ge 0$ and $j = 1, 2$. Then, for each $s \in [r, \sigma]$, we have that $\overline{\chi}^*_{k_0}(s) \le s$ and $\overline{\chi}_{k_0}(s) \le s$, and we will consider two cases: $\overline{\chi}^*_{k_0}(s) \ge \overline{\chi}_{k_0}(s)$

and $\overline{\chi}^*_{k_0}(s) < \overline{\chi}_{k_0}(s)$. We first give expressions for the integrands in (5.15), following from (3.7), (3.8) and (5.6):

$$\int_{[0,H^r_{1,k_0})} \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{1,k_0}(u)\overline{\eta}^{1,*,k_0}_s(du) = \int_0^s \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\, d\overline{E}_{k_0}(u), \tag{5.16}$$

$$\int_{[0,H^r_{1,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{1,k_0}(u)\overline{\eta}^{1,k_0}_s(du) = \int_0^s \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\, d\overline{E}_{k_0}(u), \tag{5.17}$$

$$\int_{[0,H^r_{2,k_0})} \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{2,k_0}(u)\overline{\eta}^{2,*,k_0}_s(du) = \int_0^s \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)g^r_{2,k_0}(s-u)\, d\overline{I}^*_{k_0}(u), \tag{5.18}$$

and

$$\int_{[0,H^r_{2,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{2,k_0}(u)\overline{\eta}^{2,k_0}_s(du) = \int_0^s \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(s-u)g^r_{2,k_0}(s-u)\, d\overline{I}_{k_0}(u). \tag{5.19}$$

Recalling that $\overline{\eta}^1 = \overline{\eta}^{1,*}$ and for each $l \notin \mathcal{K}_0$ and $t \in [0, \tau_{k_1}]$, $\overline{\nu}^l_t = \overline{\nu}^{*,l}_t$.
Case 1: $\overline{\chi}^*_{k_0}(s) \geq \overline{\chi}_{k_0}(s)$. By (5.18) and (3.9), we have that

$$\begin{aligned}
&\sum_{j=1}^2 \int_{[0,H^r_{j,k_0})} \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,*,k_0}_s(du) \\
\geq\ & \int_{[0,H^r_{1,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{1,k_0}(u)\overline{\eta}^{1,k_0}_s(du) + \int_{[0,H^r_{2,k_0})} \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{2,k_0}(u)\overline{\eta}^{2,*,k_0}_s(du) \\
\geq\ & \int_{[0,H^r_{1,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{1,k_0}(u)\overline{\eta}^{1,k_0}_s(du) + \sum_{l\notin\mathcal{K}_0} P_{lk_0} \int_0^s \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(s-u)g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^l_u\rangle\, du \\
& + \sum_{l\in\mathcal{K}_0} P_{lk_0} \int_0^s \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^{*,l}_u\rangle\, du \\
\geq\ & \sum_{j=1}^2 \int_{[0,H^r_{j,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,k_0}_s(du) + \sum_{l\in\mathcal{K}_0} P_{lk_0} \int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l, \triangle\overline{\nu}^l_u\rangle\, du.
\end{aligned}$$

Case 2: $\overline{\chi}^*_{k_0}(s) < \overline{\chi}_{k_0}(s)$. By (5.16)–(5.19) and (3.9), we have that

$$\begin{aligned}
&\sum_{j=1}^2 \int_{[0,H^r_{j,k_0})} \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,*,k_0}_s(du) - \sum_{j=1}^2 \int_{[0,H^r_{j,k_0})} \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,k_0}_s(du) \\
=\ & \int_0^s \mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\, d\overline{E}_{k_0}(u) - \int_0^s \mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\, d\overline{E}_{k_0}(u) \\
& + \sum_{l\notin\mathcal{K}_0} P_{lk_0} \left(\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^l_u\rangle\, du - \int_{s-\overline{\chi}_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^l_u\rangle\, du \right) \\
& + \sum_{l\in\mathcal{K}_0} P_{lk_0} \left(\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^{*,l}_u\rangle\, du - \int_{s-\overline{\chi}_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^l_u\rangle\, du \right) \\
=\ & -\int_0^s \mathbb{1}_{(\overline{\chi}^*_{k_0}(s),\overline{\chi}_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\, d\overline{E}_{k_0}(u) - \sum_{l\in\mathcal{K}} P_{lk_0} \int_{s-\overline{\chi}_{k_0}(s)}^{s-\overline{\chi}^*_{k_0}(s)} g^r_{2,k_0}(s-u)\langle h^s_l, \overline{\nu}^l_u\rangle\, du
\end{aligned}$$

$$+\sum_{l\in\mathcal{K}_0} P_{lk_0}\int_{s-\overline{\chi}^*_{k_0}(s)}^{s} g^r_{2,k_0}(s-u)\langle h^s_l, \triangle\overline{\nu}^l_u\rangle\, du.$$

Since $s\in[r,\sigma]$, $\overline{Q}^*_{k_0}(s)\geq\overline{Q}_{k_0}(s)$, where $\overline{Q}^*_{k_0}(s)=\overline{\eta}^{1,*,k_0}_s[0,\overline{\chi}^*_{k_0}(s)]+\overline{\eta}^{2,*,k_0}_s[0,\overline{\chi}^*_{k_0}(s)]$ and $\overline{Q}_{k_0}(s)=\overline{\eta}^{1,k_0}_s[0,\overline{\chi}_{k_0}(s)]+\overline{\eta}^{2,k_0}_s[0,\overline{\chi}_{k_0}(s)]$. Note that for each $s\in[r,\sigma]$, it follows from (3.7), (3.8) and (5.6) that

$$\begin{aligned}\overline{\eta}^{1,*,k_0}_s[0,\overline{\chi}^*_{k_0}(s)]+\overline{\eta}^{2,*,k_0}_s[0,\overline{\chi}^*_{k_0}(s)] \;&=\; \int_0^s 1\!\!1_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)(1-G^r_{1,k_0}(s-u))\,d\overline{E}_{k_0}(u)\\ &\quad+\int_0^s 1\!\!1_{[0,\overline{\chi}^*_{k_0}(s)]}(s-u)(1-G^r_{2,k_0}(s-u))\,d\overline{I}^*_{k_0}(u)\end{aligned}$$

and

$$\begin{aligned}\overline{\eta}^{1,k_0}_s[0,\overline{\chi}_{k_0}(s)]+\overline{\eta}^{2,k_0}_s[0,\overline{\chi}_{k_0}(s)] \;&=\; \int_0^s 1\!\!1_{[0,\overline{\chi}_{k_0}(s)]}(s-u)(1-G^r_{1,k_0}(s-u))\,d\overline{E}_{k_0}(u)\\ &\quad+\int_0^s 1\!\!1_{[0,\overline{\chi}_{k_0}(s)]}(s-u)(1-G^r_{2,k_0}(s-u))\,d\overline{I}_{k_0}(u)\end{aligned}$$

This, (3.9) and the fact that $\overline{Q}^*_{k_0}(s)\geq\overline{Q}_{k_0}(s)$ for each $s\in[r,\sigma]$ imply that

$$\begin{aligned}&\int_0^s 1\!\!1_{(\overline{\chi}^*_{k_0}(s),\overline{\chi}_{k_0}(s)]}(s-u)(1-G^r_{1,k_0}(s-u))\,d\overline{E}_{k_0}(u)\\ &\qquad+\sum_{l\in\mathcal{K}} P_{lk_0}\int_{s-\overline{\chi}_{k_0}(s)}^{s-\overline{\chi}^*_{k_0}(s)}(1-G^r_{2,k_0}(s-u))\langle h^s_l,\overline{\nu}^l_u\rangle\,du\\ \leq\;&\sum_{l\in\mathcal{K}_0} P_{lk_0}\int_{s-\overline{\chi}^*_{k_0}(s)}^{s}(1-G^r_{2,k_0}(s-u))\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du.\end{aligned}$$

Combining (5.13) with an application of integration by parts, we have that for each $s\in[r,\sigma]$ and $l\in\mathcal{K}_0$,

$$\begin{aligned}&\int_{s-\overline{\chi}^*_{k_0}(s)}^{s}(1-G^r_{2,k_0}(s-u))\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\\ =\;&\Bigg(\int_0^s\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du-(1-G^r_{2,k_0}(\overline{\chi}^*_{k_0}(s)))\int_0^{s-\overline{\chi}^*_{k_0}(s)}\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\\ &\qquad-\int_{s-\overline{\chi}^*_{k_0}(s)}^{s} g^r_{2,k_0}(s-u)\Big(\int_0^u\langle h^s_l,\triangle\overline{\nu}^l_v\rangle\,dv\Big)du\Bigg)\\ \leq\;&\delta\left(2G^s_l(s)+G^r_{2,k_0}(s)\right).\end{aligned}$$

Recall that $C_\sigma=\sup_{0\leq u\leq\sigma}(h^r_{1,k_0}(u)+h^r_{2,k_0}(u))<\infty$ by Assumption 6.2, then we have from the above estimation that

$$\begin{aligned}&\int_0^s 1\!\!1_{(\overline{\chi}^*_{k_0}(s),\overline{\chi}_{k_0}(s)]}(s-u)g^r_{1,k_0}(s-u)\,d\overline{E}_{k_0}(u)+\sum_{l\in\mathcal{K}} P_{lk_0}\int_{s-\overline{\chi}_{k_0}(s)}^{s-\overline{\chi}^*_{k_0}(s)} g^r_{2,k_0}(s-u)\langle h^s_l,\overline{\nu}^l_u\rangle\,du\\ \leq\;& C_\sigma\int_0^s 1\!\!1_{(\overline{\chi}^*_{k_0}(s),\overline{\chi}_{k_0}(s)]}(s-u)(1-G^r_{1,k_0}(s-u))\,d\overline{E}_{k_0}(u)\end{aligned}$$

$$\begin{aligned}&+\sum_{l\in\mathcal{K}} C_\sigma P_{lk_0}\int_{s-\overline{\chi}_{k_0}(s)}^{s-\overline{\chi}^*_{k_0}(s)}(1-G^r_{2,k_0}(s-u))\langle h^s_l,\overline{\nu}^l_u\rangle\,du\\ \leq\ & C_\sigma\delta\sum_{l\in\mathcal{K}_0}P_{lk_0}\left(2G^s_l(s)+G^r_{2,k_0}(s)\right)\end{aligned}$$

Thus, it follows that

$$\begin{aligned}&\sum_{j=1}^2\int_{[0,H^r_{j,k_0})}\mathbb{1}_{[0,\overline{\chi}^*_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,*,k_0}_s(du)-\sum_{j=1}^2\int_{[0,H^r_{j,k_0})}\mathbb{1}_{[0,\overline{\chi}_{k_0}(s)]}(u)h^r_{j,k_0}(u)\overline{\eta}^{j,k_0}_s(du)\\ \geq\ & -C_\sigma\delta\sum_{l\in\mathcal{K}_0}P_{lk_0}\left(2G^s_l(s)+G^r_{2,k_0}(s)\right)+\sum_{l\in\mathcal{K}_0}P_{lk_0}\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du.\end{aligned}$$

By combining the two cases, we have from (5.15) that

$$\begin{aligned}\triangle\overline{R}_{k_0}(\sigma)-\triangle\overline{R}_{k_0}(r)\geq -C_\sigma\delta\sum_{l\in\mathcal{K}_0}P_{lk_0}\left(2G^s_l(\sigma)+G^r_{2,k_0}(\sigma)\right)(\sigma-r)\\ +\sum_{l\in\mathcal{K}_0}P_{lk_0}\int_r^\sigma\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\,ds.\end{aligned}$$

Since g^r_{2,k_0} is continuously differentiable on $[C^r_{2,k_0},H^r_{2,k_0})$ and vanishes outside of $[C^r_{2,k_0},H^r_{2,k_0})$, we have, by an application of integration by parts, that for each $l\in\mathcal{K}_0$,

$$\begin{aligned}&\int_r^\sigma\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\,ds\\ =\ &\int_r^\sigma\int_{(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s}}^{\bar{s}} g^r_{2,k_0}(s-u)\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\,ds\\ =\ &\int_r^\sigma\left(g^r_{2,k_0}(s-\bar{s})\int_0^{\bar{s}}\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du-g^r_{2,k_0}(s-(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s})\int_0^{(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s}}\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\right)ds\\ &+\int_r^\sigma\int_{(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s}}^{\bar{s}}\left(g^r_{2,k_0}\right)'(s-u)\int_0^u\langle h^s_l,\triangle\overline{\nu}^l_v\rangle\,dv\,du,\end{aligned}$$

where $\bar{s}=(s-C^r_{2,k_0})\vee 0$. Thus, by (5.13), we have that for each $l\in\mathcal{K}_0$,

$$\begin{aligned}&\left|\int_r^\sigma\int_{s-\overline{\chi}^*_{k_0}(s)}^s g^r_{2,k_0}(s-u)\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\,ds\right|\\ \leq\ &\int_r^\sigma\left(\sup_{0\leq s\leq\sigma}g^r_{2,k_0}(s)\left|\int_0^{\bar{s}}\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\right|+\sup_{0\leq s\leq\sigma}g^r_{2,k_0}(s)\left|\int_0^{(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s}}\langle h^s_l,\triangle\overline{\nu}^l_u\rangle\,du\right|\right)ds\\ &+\int_r^\sigma\int_{(s-\overline{\chi}^*_{k_0}(s))\wedge\bar{s}}^{\bar{s}}\left|\left(g^r_{2,k_0}\right)'(s-u)\right|\left|\int_0^u\langle h^s_l,\triangle\overline{\nu}^l_v\rangle\,dv\right|\,du\\ \leq\ &\delta G^s_l(\sigma)(\sigma-r)\left(2\sup_{0\leq s\leq\sigma}g^r_{2,k_0}(s)+\int_{C^r_{2,k_0}}^{\sigma\wedge C^r_{2,k_0}}\left|\left(g^r_{2,k_0}\right)'(v)\right|dv\right).\end{aligned}$$

Recall that $D_\sigma=2\sup_{0\leq s\leq\sigma}g^r_{2,k_0}(s)+\int_0^\sigma|(g^r_{2,k_0})'(v)|dv$ and $\sigma<1$. It then follows from (5.15) that

$$-(\triangle\overline{R}_{k_0}(\sigma)-\triangle\overline{R}_{k_0}(r))\leq C_\sigma\delta\sum_{l\in\mathcal{K}_0}P_{lk_0}\left(2G^s_l(\sigma)+G^r_{2,k_0}(\sigma)\right)+\delta D_\sigma\sum_{l\in\mathcal{K}_0}P_{lk_0}G^s_l(\sigma).$$

Combining this with (5.11) and (5.14), we have (5.3). This completes the proof. ■

Proof of Theorem 3.4. Let $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ and $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ be two solutions to the fluid equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0^1, \overline{\eta}_0^2) \in \mathcal{S}_0$. If $\tau = \infty$, the proof is complete by Lemma 5.3. On the other hand, suppose that $\tau < \infty$. Note that by Lemmas 5.3 and 5.4 we have $\tau_k = \tau$ for each $k \in \mathcal{K}$ and $(\overline{X}_k, \overline{\nu}_k, \overline{\eta}_k^1, \overline{\eta}_k^2)$ agrees with $(\overline{X}_k^*, \overline{\nu}_k^*, \overline{\eta}_k^{1,*}, \overline{\eta}_k^{2,*})$ on $[0, \tau]$. By an action of time-shifts on solutions to the fluid equations that is similar to Lemma 3.4 of [14], we may assume, without loss of generality, that $\tau = 0$. Choose $\delta > 0$ and define

$$\xi_k \doteq \inf\{t \geq 0 : \ |\triangle \overline{L}_k(t)| \geq \delta\} \text{ for each } k \in \mathcal{K} \text{ and } \xi = \min_{k \in \mathcal{K}} \xi_k.$$

The same contradiction argument used in the proof of Lemma 5.4 can be adapted to show that $\xi = \infty$. As a consequence, $\overline{L} = \overline{L}^*$ and then the argument used in the proof of Lemma 5.3 can also be adapted to show that $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ agrees with $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ on $[0, \infty)$. In the rest of the proof, we focus on proving that $\xi = \infty$ with emphasis on the argument that is different with the one used in Lemma 5.4.

Suppose that $\xi < \infty$. Let k_0 be the smallest index in $\mathcal{K}$ such that $\xi_{k_0} = \xi$. By choosing δ small enough, we may assume that ξ is small enough such that $\xi < 1$ and

$$G_{k_0}^s(\xi) + (2 + 2C_\xi + D_\xi) \sum_{l \in \mathcal{K}} P_{lk_0} G_l^s(\xi) + C_\xi G_{2,k_0}^r(\xi) \sum_{l \in \mathcal{K}} P_{lk_0} \leq 1/2, \tag{5.20}$$

where C_ξ and D_ξ is the same as C_σ and D_σ with ξ in placed of σ. Then $|\triangle \overline{L}_{k_0}(\xi)| \geq \delta$ and $|\triangle \overline{L}_k(t)| < \delta$ for each $t \in [0, \xi)$ and $k \in \mathcal{K}$. Thus, we have either $\triangle \overline{L}_{k_0}(\xi) \geq \delta$ or $\triangle \overline{L}_{k_0}(\xi) \leq -\delta$. We shall just consider the case when $\triangle \overline{L}_{k_0}(\xi) \geq \delta$, since the other case can be treated in the same way. The contradiction can be reached by following exactly the same three steps in Lemma 5.4 with $\mathcal{K}$ in place of $\mathcal{K}_0$. Thus, we have the desired result. ■

6 Convergence

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

$$\begin{aligned} \overline{E}^{(N)}(t) &\doteq \frac{E^{(N)}(t)}{N}, \quad \overline{X}^{(N)}(t) \doteq \frac{X^{(N)}(t)}{N}, \quad \overline{\nu}_t^{(N)}(B) \doteq \frac{\nu_t^{(N)}(B)}{N}, \\ \overline{\eta}_t^{(N),1}(B) &\doteq \frac{\eta_t^{(N),1}(B)}{N}, \qquad \overline{\eta}_t^{(N),2}(B) \doteq \frac{\eta_t^{(N),2}(B)}{N} \end{aligned} \tag{6.1}$$

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

$$\overline{A}^{(N)} \doteq \frac{A^{(N)}}{N} \text{ for } A = D, L, Q, R, S, I. \tag{6.2}$$

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

Assumption 6.1 (Initial conditions) *There exists an $\mathcal{S}_0$-valued random variable $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0^1, \overline{\eta}_0^2)$ 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) $\overline{X}^{(N)}(0) \to \overline{X}(0)$ *in* $\mathbb{R}_+^K$*;*

(iii) $\overline{\nu}_0^{(N)} \xrightarrow{w} \overline{\nu}_0$ *in* $\Pi_{k\in\mathcal{K}}\mathcal{M}_F[0,H_k^s)$*;*

(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 6.2 *For each* $k \in \mathcal{K}$*, there exists* $L_k^s < H_k^s$ *such that* h_k^s *is either bounded or lower-semicontinuous on* (L_k^s, H_k^s)*,* $g_{2,k}^r$ *is continuously differentiable on its support* $[C_{2,k}^r, H_{2,k}^r)$ *and either one of the following holds for* $h_{1,k}^r$*:*

(i) $h_{1,k}^r$ *is bounded;*

(ii) $h_{1,k}^r$ *is locally bounded and there exists* $L_{1,k}^r < H_{1,k}^r$ *such that* $h_{1,k}^r$ *is lower-semicontinuous on* $(L_{1,k}^r, H_{1,k}^r)$*.*

Theorem 6.1 *Suppose that Assumptions 6.1 and 6.2 hold. Let* $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0^1, \overline{\eta}_0^2) \in \mathcal{S}_0$ *be the limiting initial condition. Then there exists a unique solution* $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ *to the associated fluid equations* (3.4) − (3.14)*, and*

$$(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N),1}, \overline{\eta}^{(N),2}) \Rightarrow (\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2) \quad \text{as} \quad N \to \infty. \tag{6.3}$$

The proof of this theorem can be carried out in two steps as the proof of Theorem 3.6 of [14]. The first step is to show the fluid-scaled processes are tight and the second step is to show that every limit of any subsequence of the fluid-scaled processes solves the fluid equations. For tightness, we follow closely the steps in [14] 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 [14].

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 $[0, H_k^s) \times \mathbb{R}_+$, consider the process $D_\varphi^{(N),k}$ that takes values in $\mathbb{R}$, and is given by

$$\begin{aligned} D_\varphi^{(N),k}(t) \doteq & \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}} \varphi(a_j^{(N),1,k}(s), s) \\ & + \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}} \varphi(a_j^{(N),2,k}(s), s), \end{aligned} \tag{6.4}$$

for $t \in [0,\infty)$. It is clear that when φ is the constant function $\mathbf{1}$,

$$D_{\mathbf{1}}^{(N),k} = D_k^{(N)}. \tag{6.5}$$

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

$$S_\varphi^{(N),1,k}(t) \doteq \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}} \varphi(w_j^{(N),1,k}(s), s), \tag{6.6}$$

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

$$S^{(N),2,k}_{\varphi}(t) \doteq \sum_{j=-\mathcal{C}^{(N)}_k+1}^{I^{(N)}_k(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw^{(N),2,k}_j}{dt}(s-)>0,\ \frac{dw^{(N),2,k}_j}{dt}(s+)=0\right\}} \varphi(w^{(N),2,k}_j(s), s), \tag{6.7}$$

for $t \in [0,\infty)$. Clearly, $S^{(N),1,k}_{\mathbf{1}} + S^{(N),2,k}_{\mathbf{1}}$ equals to the cumulative potential reneging process $S^{(N)}_k$, that is,

$$S^{(N),1,k}_{\mathbf{1}} + S^{(N),2,k}_{\mathbf{1}} = S^{(N)}_k. \tag{6.8}$$

Next, comparing (2.16) with (6.6) and (6.7), it is clear that for each $k \in \mathcal{K}$, the cumulative reneging process $R^{(N)}_k$ satisfies

$$R^{(N)}_k(t) = S^{(N),1,k}_{\theta^{(N)}_k}(t) + S^{(N),2,k}_{\theta^{(N)}_k}(t), \qquad t \geq 0, \tag{6.9}$$

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

$$\theta^{(N)}_k(x,s) = \mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(x), \qquad x \in \mathbb{R},\ s \geq 0. \tag{6.10}$$

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

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

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 [14], we can see that all the processes $E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N),1}, \eta^{(N),2}$ and the auxiliary processes are $\{\mathcal{F}^{(N)}_t\}$-adapted. It follows immediately from (6.4), (6.6), (6.7) and the right continuity of the filtration $\{\mathcal{F}^{(N)}_t\}$ that for each $k \in \mathcal{K}$, $D^{(N),k}_{\varphi}$, $S^{(N),1,k}_{\varphi}$ and $S^{(N),2,k}_{\varphi}$ are $\{\mathcal{F}^{(N)}_t\}$-adapted.

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

$$A^{(N),k}_{\varphi,\nu}(t) \doteq \int_0^t \left(\int_{[0,H^s_k)} \varphi(x,s) h^s_k(x)\, \nu^{(N),k}_s(dx) \right) ds, \qquad t \in [0,\infty). \tag{6.11}$$

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

$$A^{(N),j,k}_{\varphi,\eta}(t) \doteq \int_0^t \left(\int_{[0,H^r_{j,k})} \varphi(x,s) h^r_{j,k}(x)\, \eta^{(N),j,k}_s(dx) \right) ds, \qquad t \in [0,\infty), \tag{6.12}$$

and

$$A^{(N),j,k}_{\theta^{(N)}_k,\eta}(t) \doteq \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(x) h^r_{j,k}(x)\, \eta^{(N),j,k}_s(dx) \right) ds, \qquad t \in [0,\infty), \tag{6.13}$$

where $\theta_k^{(N)}$ is defined in (6.10). A similar argument as Proposition 5.1 and Lemma 5.4 of [14] shows that $A_{\varphi,\nu}^{(N),k}$ (respectively, $A_{\varphi,\eta}^{(N),j,k}$, $j=1,2$, and $A_{\theta_k^{(N)},\eta}^{(N),j,k}$) is the $\mathcal{F}_t^{(N)}$-compensator of process $D_\varphi^{(N),k}$ (respectively, $S_\varphi^{(N),j,k}$, $j=1,2$, and $R_k^{(N)}$). That is, for each $k \in \mathcal{K}$, and for every bounded measurable function φ on $[0,H_k^s) \times \mathbb{R}_+$ 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 $M_{\varphi,\nu}^{(N),k}$ defined by

$$M_{\varphi,\nu}^{(N),k} \doteq D_\varphi^{(N),k} - A_{\varphi,\nu}^{(N),k} \tag{6.14}$$

is a local $\mathcal{F}_t^{(N)}$-martingale. Moreover, for every $N \in \mathbb{N}$, $t \in [0,\infty)$ and $m \in [0,H_k^s)$,

$$|A_{\varphi,\nu}^{(N),k}(t)| \le \|\varphi\|_\infty \left(X_k^{(N)}(0) + E_k^{(N)}(t) + I_k^{(N)}(t)\right)\left(\int_0^m h_k^s(x)\,dx\right) < \infty \tag{6.15}$$

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

$$\lim_{N\to\infty} \mathbb{E}\left[\langle \overline{M}_{\varphi,\nu}^{(N),k} \rangle(t)\right] = 0; \qquad\qquad \overline{M}_{\varphi,\nu}^{(N),k} \Rightarrow \mathbf{0} \text{ as } N \to \infty. \tag{6.16}$$

Furthermore, properties (6.14)–(6.16) also hold with $D^{(N),k}, A^{(N),k}, M^{(N),k}, a, \nu, H_k^s$ and h_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 $j=1,2$. For each $j=1,2$, the process $M_{\theta_k^{(N)},\eta}^{(N),k}$ defined by

$$M_{\theta_k^{(N)},\eta}^{(N),k} \doteq R_k^{(N)} - A_{\theta_k^{(N)},\eta}^{(N),1,k} - A_{\theta_k^{(N)},\eta}^{(N),2,k} \tag{6.17}$$

is a local $\mathcal{F}_t^{(N)}$-martingale. In addition, as $N \to \infty$,

$$\lim_{N\to\infty} \mathbb{E}\left[\langle \overline{M}_{\theta_k^{(N)},\eta}^{(N),k} \rangle(t)\right] = 0 \quad \text{ and } \quad \overline{M}_{\theta_k^{(N)},\eta}^{(N),k} \Rightarrow \mathbf{0}. \tag{6.18}$$

Notice that by (2.12) and (2.13), we have for each $k \in \mathcal{K}$, $l \in \mathcal{K} \cup \{0\}$ and $t \ge 0$,

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

and

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

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

$$\mathbb{E}\left[I_k^{(N)}(t)\right] \leq \sum_{l \in \mathcal{K}} (HP')_{kl} \mathbb{E}\left[E_l^{(N)}(t) + X_l^{(N)}(0)\right] < \infty, \tag{6.19}$$

where the last inequality holds due to Assumption 6.1. In addition, using (6.5), (2.17), (2.20), (6.19) and the non-negativity of $Q_k^{(N)}$, $R_k^{(N)}$ and $\langle \mathbf{1}, \nu^{(N),k} \rangle$, it follows from Assumption 6.1 that for any $t \in [0, \infty)$ and bounded, measurable φ,

$$\mathbb{E}\left[\left|D_\varphi^{(N),k}(t)\right|\right] \leq \|\varphi\|_\infty \, \mathbb{E}\left[X_k^{(N)}(0) + E_k^{(N)}(t) + I_k^{(N)}(t)\right] < \infty \tag{6.20}$$

and likewise, for each $t \in [0, \infty)$ and bounded measurable φ and ψ, (2.18) shows that

$$\mathbb{E}\left[\left|S_\varphi^{(N),1,k}(t)\right| + \left|S_\psi^{(N),2,k}(t)\right|\right] \leq (\|\psi\|_\infty + \|\varphi\|_\infty) \mathbb{E}\left[\langle \mathbf{1}, \eta_0^{(N)} \rangle + E_k^{(N)}(t) + I_k^{(N)}(t)\right] < \infty. \tag{6.21}$$

From (6.20) and (6.15) it is clear that for every t, the linear functionals $\overline{D}_\cdot^{(N),k}(t) : \varphi \mapsto \overline{D}_\varphi^{(N),k}(t)$ and $\overline{A}_{\cdot,\nu}^{(N),k}(t) : \varphi \mapsto \overline{A}_{\varphi,\nu}^{(N),k}(t)$ are finite Radon measures on $[0, H_k^s) \times \mathbb{R}_+$. Likewise, from (6.21) and the fact that (6.15) holds with $\nu^{(N),k}$, h_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 $\{\overline{D}_\cdot^{(N),k}(t) : \; t \in [0, \infty)\}$ and $\{\overline{A}_{\cdot,\nu}^{(N),k}(t) : \; t \in [0, \infty)\}$ can be viewed as $\mathcal{M}_F([0, H_k^s) \times \mathbb{R}_+)$-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

$$\begin{aligned}\overline{Z}_k^{(N)} \quad &\doteq \quad \Big(\overline{X}_k^{(N)}(0), \overline{E}_k^{(N)}, \overline{X}_k^{(N)}, \overline{R}_k^{(N)}, \overline{I}_k^{(N)}, \{\overline{D}_{kl}^{(N)}, l \in \mathcal{K} \cup \{0\}\}, \overline{\nu}_0^{(N),k}, \overline{\nu}^{(N),k}, \\ &\qquad \overline{\eta}_0^{(N),1,k}, \overline{\eta}^{(N),1,k}, \overline{\eta}_0^{(N),2,k}, \overline{\eta}^{(N),2,k}, \overline{A}_{\cdot,\nu}^{(N),k}, \overline{D}_\cdot^{(N),k}, \overline{A}_{\cdot,\eta}^{(N),1,k}, \overline{S}_\cdot^{(N),1,k}, \overline{A}_{\cdot,\eta}^{(N),2,k}, \overline{S}_\cdot^{(N),2,k}\Big).\end{aligned} \tag{6.22}$$

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

$$\begin{aligned}\mathcal{Y}_k \quad &\doteq \quad \mathbb{R}_+ \times (\mathcal{D}_{\mathbb{R}_+}[0, \infty))^{K+5} \times \mathcal{M}_F[0, H_k^s) \times \mathcal{D}_{\mathcal{M}_F[0, H_k^s)}[0, \infty) \times \mathcal{M}_F[0, H_{1,k}^r) \\ &\qquad \times \mathcal{D}_{\mathcal{M}_F[0, H_{1,k}^r)}[0, \infty) \times \mathcal{M}_F[0, H_{2,k}^r) \times \mathcal{D}_{\mathcal{M}_F[0, H_{2,k}^r)}[0, \infty) \times (\mathcal{D}_{\mathcal{M}_F([0, H_k^s) \times \mathbb{R}_+)}[0, \infty))^2 \\ &\qquad \times (\mathcal{D}_{\mathcal{M}_F([0, H_{1,k}^r) \times \mathbb{R}_+)}[0, \infty))^2 \times (\mathcal{D}_{\mathcal{M}_F([0, H_{2,k}^r) \times \mathbb{R}_+)}[0, \infty))^2\end{aligned}$$

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

$$\overline{Z}^{(N)} = (\overline{Z}_k^{(N)}, k \in \mathcal{K}). \tag{6.23}$$

Then, by applying Kurtz' criteria (see Theorem 3.8.6 of [9] for details) and a similar argument for Lemma 5.8(2) in [16] together with the bounds in (6.20) and (6.21), we can show that under Assumption 6.1, for each $k \in \mathcal{K}$, the sequences $\{\overline{X}_k^{(N)}\}$, $\{\overline{I}_k^{(N)}\}$, $\{\overline{L}_k^{(N)}\}$, $\{\overline{R}_k^{(N)}\}$, $\{\langle \mathbf{1}, \overline{\nu}^{(N),k} \rangle\}$, $\{\langle \mathbf{1}, \overline{\eta}^{(N),1,k} \rangle\}$, $\{\langle \mathbf{1}, \overline{\eta}^{(N),2,k} \rangle\}$, the sequences $\{\overline{D}_\varphi^{(N),k}\}$, $\{\overline{A}_{\varphi,\nu}^{(N),k}\}$, for every $\varphi \in \mathcal{C}_b([0, H_k^s) \times \mathbb{R}_+)$, and the sequences $\{\overline{S}_\varphi^{(N),j,k}\}$, $\{\overline{A}_{\varphi,\eta}^{(N),j,k}\}$, for every $j = 1, 2$ and $\varphi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+)$, are relatively compact. By a similar argument of Lemma 6.4 of [14], together with the fact that $\mathbb{E}\left[X_k^{(N)}(0) + E_k^{(N)}(t) + I_k^{(N)}\right] < \infty$, we can show that under Assumption 6.1, for every $f \in \mathcal{C}_c^1(\mathbb{R}_+)$ and $k \in \mathcal{K}$, the sequences $\{\langle f, \overline{\nu}^{(N),k}\rangle\}$ 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 [14]), in the same way as the proofs of Lemmas 6.6 and 6.7 of [14], we can show that under Assumption 6.1, for each $k \in \mathcal{K}$, the sequences $\{\overline{\nu}^{(N),k}\}$ and $\{\overline{\eta}^{(N),j,k}\}$, $j = 1, 2$, are relatively compact and the sequences $\{\overline{D}_{\cdot}^{(N),k}\}$ and $\{\overline{A}_{\cdot,\nu}^{(N),k}\}$ are relatively compact in $\mathcal{D}_{\mathcal{M}_F([0,H_k^s)\times\mathbb{R}_+)}[0, \infty)$. 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 $\{\overline{Z}^{(N)}\}$, which is summarized in the following theorem.

Theorem 6.2 *Suppose Assumption 6.1 is satisfied. Then the sequence $\{\overline{Z}^{(N)}\}$ defined in (6.23) is relatively compact in the Polish space $\Pi_{k\in\mathcal{K}}\mathcal{Y}_k$, and is therefore tight.*

We next focus on establishing the existence of a solution to the fluid equations, and thus, the convergence of the fluid-scaled measure-valued processes and auxiliary processes follows by the tightness and uniqueness of such a solution, so Theorem 6.1 is proved. The rest of this section is devoted to the proof of the following theorem. We note that Theorem 6.2 only requires Assumption 6.1 while Theorem 6.3 requires both Assumptions 6.1–6.2.

Theorem 6.3 *Suppose that Assumptions 6.1–6.2 hold. Let $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ be the limit of any subsequence of $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N),1}, \overline{\eta}^{(N),2})\}$. Then $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ solves the fluid equations.*

We first establish two supporting lemmas.

Lemma 6.4 *For each $k \in \mathcal{K}$ and $l \in \mathcal{K} \cup \{0\}$, $\overline{D}_{kl}^{(N)} - P_{kl}\overline{D}_{\mathbf{1}}^{(N),k} \Rightarrow 0$ as $N \to \infty$.*

Proof. Fix $k \in \mathcal{K}$ and $l \in \mathcal{K} \cup \{0\}$. It follows from (2.12) and (6.4) that

$$\begin{aligned}
&\overline{D}_{kl}^{(N)}(t) - P_{kl}\overline{D}_{\mathbf{1}}^{(N),k}(t) \qquad (6.24)\\
= \;& \frac{1}{N}\sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} (\mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} - P_{kl})\mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}}\\
&+\frac{1}{N}\sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} (\mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} - P_{kl})\mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}}.
\end{aligned}$$

Since the service distribution G_k^s has density, then with probability 1, any two customers will not finish service at the same time, thus for each $T > 0$,

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

By Assumption 6.1 and (6.19),

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

and hence the lemma is proved. ■

Lemma 6.5 *For each $k \in \mathcal{K}$, $j = 1, 2$ and $T \in [0,\infty)$, as $N \to \infty$,*

$$\mathbb{E}\left[\sup_{t\in[0,T]} \left|\overline{A}^{(N),j,k}_{\theta^{(N)}_k,\eta}(t) - \int_0^t \left(\int_{[0,H^r_{j,k})} 1\!\!1_{[0,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right) ds\right|\right] \to 0. \tag{6.25}$$

Moreover, almost surely,

$$\overline{R}_k(t) = \sum_{j=1}^{2} \int_0^t \left(\int_{[0,H^r_{j,k})} 1\!\!1_{[0,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right) ds,\ t \in [0,\infty). \tag{6.26}$$

The proof of this lemma is in the appendix.

Let $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0)$ be the $\mathcal{S}_0$-valued random variable that satisfies Assumption 6.1, and let $\{\overline{Z}^{(N)}\}_{N\in\mathbb{N}}$ be the sequence of processes defined in (6.23). Then, by Assumption 6.1, Theorem 6.2, Lemma 6.4 and the limits $\overline{M}^{(N),k}_{\cdot,\nu} = \overline{D}^{(N),k}_{\cdot} - \overline{A}^{(N),k}_{\cdot,\nu} \Rightarrow 0$ and $\overline{M}^{(N),j,k}_{\cdot,\eta} = \overline{S}^{(N),j,k}_{\cdot} - \overline{A}^{(N),j,k}_{\cdot,\eta} \Rightarrow 0$, there exist processes $\overline{X} \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)^K$, $\overline{R} \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)^K$, $\overline{I} \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)^K$, $\overline{\nu} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F[0,H^s_k)}[0,\infty)$,$\overline{\eta}^1 \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F[0,H^r_{1,k})}[0,\infty)$, $\overline{\eta}^2 \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F[0,H^r_{2,k})}[0,\infty)$,
$\overline{A}_{\cdot,\nu} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^s_k)\times\mathbb{R}_+)}[0,\infty)$, $\overline{D}_{\cdot} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^s_k)\times\mathbb{R}_+)}[0,\infty)$, $\overline{A}^1_{\cdot,\eta} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^r_{1,k})\times\mathbb{R}_+)}[0,\infty)$,
$\overline{S}^1_{\cdot} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^r_{1,k})\times\mathbb{R}_+)}[0,\infty)$, $\overline{A}^2_{\cdot,\eta} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^r_{2,k})\times\mathbb{R}_+)}[0,\infty)$, and
$\overline{S}^2_{\cdot} \in \Pi_{k\in\mathcal{K}}\mathcal{D}_{\mathcal{M}_F([0,H^r_{2,k})\times\mathbb{R}_+)}[0,\infty)$ such that $\overline{Z}^{(N)}$ converges weakly (along a suitable subsequence) to $\overline{Z}$, where for each $k \in \mathcal{K}$,

$$\begin{aligned}\overline{Z}_k \ &\doteq\ \Big(\overline{X}_k(0), \overline{E}_k, \overline{X}_k, \overline{R}_k, \overline{I}_k, \{P_{kl}\overline{A}^k_{\mathbf{1},\nu}, l \in \mathcal{K}\cup\{0\}\}, \overline{\nu}^k_0, \overline{\nu}^k,\\ &\qquad \overline{\eta}^{1,k}_0, \overline{\eta}^{1,k}, \overline{\eta}^{2,k}_0, \overline{\eta}^{2,k}, \overline{A}^k_{\cdot,\nu}, \overline{A}^k_{\cdot,\nu}, \overline{A}^{1,k}_{\cdot,\eta}, \overline{A}^{1,k}_{\cdot,\eta}, \overline{A}^{2,k}_{\cdot,\eta}, \overline{A}^{2,k}_{\cdot,\eta}\Big) \in \mathcal{Y}_k.\end{aligned}$$

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.

Proof of Theorem 6.3. Let $Y^{(N)} = (E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N),1}, \eta^{(N),2})$. Since $\overline{Z}^{(N)} \to \overline{Z}$ as $N \to \infty$, then it follows that, as $N \to \infty$, $(\overline{Y}^{(N)}_k, \{\overline{D}^{(N)}_{kl}, l \in \mathcal{K}\cup\{0\}\}) \to (\overline{Y}_k, \{P_{kl}\overline{A}^k_{\mathbf{1},\nu}, l \in \mathcal{K}\cup\{0\}\})$, where $\overline{Y} = (\overline{E}, \overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$. Together with (2.17) and the fact that $\sum_{l\in\mathcal{K}\cup\{0\}} P_{kl} = 1$, this implies that

$$\overline{X}_k = \overline{X}_k(0) + \overline{E}_k + \overline{I}_k - \overline{R}_k - \overline{A}^k_{\mathbf{1},\nu}. \tag{6.27}$$

Moreover, by the same argument in getting (7.2) of [14], we have that

$$\overline{A}^k_{\varphi,\nu} = \int_0^{\cdot} \langle \psi(\cdot,s)h^s_k(\cdot,s), \overline{\nu}^k_s\rangle ds. \tag{6.28}$$

On substituting (6.28) into (6.27), we see that the fluid equation (3.13) is satisfied. Next, Lemma 6.5 establishes the representation (3.12) of $\overline{R}$ given in the fluid equations.

Fix $t \in [0,\infty)$ and $j = 1,2$ such that for each $k \in \mathcal{K}$, $\overline{\nu}_t^{(N),k} \overset{w}{\to} \overline{\nu}_t^k$, $\overline{\eta}_t^{(N),j,k} \overset{w}{\to} \overline{\eta}_t^{j,k}$, $\overline{E}_k^{(N)}(t) \to \overline{E}_k(t)$, $\overline{X}_k^{(N)}(t) \to \overline{X}_k(t)$, $\overline{R}_k^{(N)}(t) \to \overline{R}_k(t)$, $\overline{I}_k^{(N)}(t) \to \overline{I}_k(t)$, $\overline{A}_{\cdot,\nu}^{(N),k}(t) \overset{w}{\to} \overline{A}_{\cdot,\nu}^k(t)$, $\overline{D}_{\cdot}^{(N),k}(t) \overset{w}{\to} \overline{A}_{\cdot,\nu}^k(t)$, $\overline{A}_{\cdot,\eta}^{(N),j,k}(t) \overset{w}{\to} \overline{A}_{\cdot,\eta}^{j,k}(t)$, $\overline{S}_{\cdot}^{(N),j,k}(t) \overset{w}{\to} \overline{A}_{\cdot,\eta}^{j,k}(t)$ as $N \to \infty$. Since $\overline{Z}^{(N)} \to \overline{Z}$ a.s., this occurs for t outside a countable set. By (6.28), this implies that for each $k \in \mathcal{K}$, as $N \to \infty$,

$$\overline{D}_{\varphi}^{(N),k}(t) \to \overline{A}_{\varphi,\nu}^k(t) = \int_0^t \langle \varphi(\cdot,s) h_k^s(\cdot,s), \overline{\nu}_s^k \rangle \, ds, \qquad \varphi \in \mathcal{C}_b([0,H^s) \times \mathbb{R}_+). \tag{6.29}$$

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

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

In particular, when $\varphi = \psi = \mathbf{1}$, the above two displays imply that (3.4) holds. Also, we immediately obtain that for each $k \in \mathcal{K}$, as $N \to \infty$, $\langle \mathbf{1}, \overline{\nu}_t^{(N),k} \rangle \to \langle \mathbf{1}, \overline{\nu}_t^k \rangle$ and $\langle \mathbf{1}, \overline{\eta}_t^{(N),j,k} \rangle \to \langle \mathbf{1}, \overline{\eta}_t^{j,k} \rangle$. When combining with (2.20), (2.19), (6.5), (2.17), (2.8), (6.26), (2.13), Lemma 6.4 and the non-idling condition, this implies that all the equations in Definition 3.1 are satisfied at time t except (3.5), (3.7) and (3.8).

It only remains to show that (3.5), (3.7) and (3.8) are also satisfied at time t. We shall just prove (3.5). The same argument will also show that (3.7) and (3.8) hold. By a similar argument as the proof of Theorem 2.1 in [14], in the fluid scale, we have

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

Since $\overline{\nu}_0^{(N),k} \overset{w}{\to} \overline{\nu}_0^k$ by Assumption 6.1(3), $\overline{\nu}_s^{(N),k} \overset{w}{\to} \overline{\nu}_s^k$ for a.e. $s \in [0,t]$, $\overline{\nu}_t^{(N),k} \overset{w}{\to} \overline{\nu}_t^k$ by our choice of t and $\varphi(\cdot,t)$ and $\varphi_x(\cdot,s) + \varphi_s(\cdot,s)$, $s \in [0,t]$, are bounded and continuous, as $N \to \infty$, we have

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

and, by the bounded convergence theorem,

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

On the other hand, using an integration-by-parts argument, the facts that $\overline{L}_k^{(N)}(0) = 0$, $\overline{L}_k^{(N)} \to \overline{L}_k$, $\overline{L}_k$ 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 $N \to \infty$, $\int_{[0,t]} \varphi(0,s) d\overline{L}_k^{(N)}(s) \to \int_{[0,t]} \varphi(0,s) d\overline{L}_k(s)$. Combining the last four displays with (6.29), it follows that (3.5) 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 $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ satisfies the fluid equations. ■

Acknowledgements

The authors thank the anonymous referees for their helpful comments on our paper, which have led to a better presentation of our results.

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. Submitted. 2010.

[2] S. Asmussen, *Applied probability and queues*, 2nd edition ed., Springer-Verlag, New York, 2003.

[3] R. Atar, H. Kaspi and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. Preprint. 2012.

[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*, JASA **100**, No. 469, 36–50, 2005.

[5] H. Chen and D. Yao, Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, (Springer, New-York, 2001).

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

[7] 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.

[8] D. Gamarnik and D.A. Goldberg. Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Annals of Applied Probability.* Forthcoming. 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 Oper. Management* Vol. 5, 79–141, 2003.

[11] O. Garnett, A. Mandelbaum and M.I. Reiman. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* Vol. 4, 3:208–227, 2002.

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

[13] W. Kang and G. Pang. Equivalence of fluid models for $G_t/GI/N + GI$ queues. *Submitted.* 2013.

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

[15] W. Kang and K. Ramanan. Asymptotic approximations for the stationary distributions of many-server queues with abandonment. *Annals of Applied Probability* . Vol. 22, 477–521. 2012.

[16] 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.

[17] H. Kaspi and K. Ramanan, SPDE limits for many-server queues. *preprint*, 2011.

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

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

[20] 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.

[21] 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.* Forthcoming. 2012.

[22] 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.

[23] Y. Liu and W. Whitt. Many-Server Heavy-Traffic Limits for Queues with Time-Varying Parameters. Submitted. 2012.

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

[25] Y. Liu and W. Whitt. Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing.* Forthcoming. 2012.

[26] A. Mandelbaum and P. Momcilovic, Queues with many servers and impatient customers. to appear in *Mathematics of Operations Research.* 2012.

[27] 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.

[28] A. Mandelbaum and S. Zeltyn. Data stories about (im)patient customers in tele-queues. Submitted. 2012.

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

[30] G. Pang and D.D. Yao. Heavy-traffic limits for a multiclass many-server queueing network with switchovers. *Advances in Applied Probability.* Forthcoming. 2013.

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

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

[33] J. Reed and Y. Y. Shaki. A fair policy for the $G/GI/N$ queue with multiple server pools. Preprint. 2012.

[34] P. Shi, M. Chou, J.G. Dai, D. Ding, and J. Sim. Hospital inpatient operations: mathematical models and managerial insights. Submitted. 2012.

[35] W. Whitt. Stochastic models for the design and management of customer contact centers: some research directions. *working paper*, 2002.

[36] W. Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. Vol. 54, No. 1, 37–54, 2006.

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

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

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

7 Appendix

Proof of Lemma 5.3. By the definition of τ, we have $\overline{L}_k$ agrees with $\overline{L}^*_k$ on $[0,\tau)$ for each $k \in \mathcal{K}$. Then the proof follows essentially the same argument in the proof of Theorem 4.6 starting from (4.14) in [14] with slight changes in handling feedback. Let $\overline{Q}, \overline{L}, \overline{I}, \overline{R}, \overline{\eta}$ and $\overline{Q}^*, \overline{L}^*, \overline{I}^*, \overline{R}^*, \overline{\eta}^*$ be the processes described in Definition 3.1 that are associated with the solutions $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ and $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ to the fluid equations for $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$, respectively. It follows directly from (3.7) that

$$\overline{\eta}^1 = \overline{\eta}^{1,*} \tag{7.1}$$

and from (3.5) and the continuity of $\overline{L}$, for each $k \in \mathcal{K}$ and $t \in [0,\tau]$, $\overline{\nu}^k_t = \overline{\nu}^{*,k}_t$. Hence, by (3.9) and (3.8), we have for each $k \in \mathcal{K}$ and $t \in [0,\tau]$, $\overline{I}_k(t) = \overline{I}^*_k(t)$ and $\overline{\eta}^{2,k}_t = \overline{\eta}^{2,*,k}_t$. As a consequence, by (3.16), we obtain that for each $t \in [0,\tau]$ and $k \in \mathcal{K}$,

$$\overline{R}_k(t) + \overline{Q}_k(t) = \overline{R}^*_k(t) + \overline{Q}^*_k(t). \tag{7.2}$$

We now show that, in fact $\overline{Q}_k = \overline{Q}^*_k$ and $\overline{R}_k = \overline{R}^*_k$ on $[0,\tau]$ for each $k \in \mathcal{K}$. Fix $k \in \mathcal{K}$. If there exists $t \in (0,\tau]$ such that $\overline{Q}_k(t) > \overline{Q}^*_k(t)$, let $s \doteq \sup\{v < t : \overline{Q}_k(v) \le \overline{Q}^*_k(v)\} \vee 0$. Then $\overline{Q}_k(s) \le \overline{Q}^*_k(s)$ and $\overline{Q}_k(v) > \overline{Q}^*_k(v)$ for each $v \in (s,t]$. Due to the fact that $\overline{\eta}^1 = \overline{\eta}^{1,*}$, $\overline{\eta}^{2,k}_l = \overline{\eta}^{2,*,k}_l$ for each $l \in [0,\tau]$, we have that $\overline{\eta}^k_l = \overline{\eta}^{*,k}_l$ for all $l \in [0,\tau]$. Then for each $l \in (s,t]$,

$$\overline{\chi}_k(l) = (F^{\overline{\eta}^k_l})^{-1}(\overline{Q}_k(l)) = (F^{\overline{\eta}^{*,k}_l})^{-1}(\overline{Q}_k(l)) \ge (F^{\overline{\eta}^{*,k}_l})^{-1}(\overline{Q}^*_k(l)) = \overline{\chi}^*_k(l),$$

and then by (3.12),

$$\begin{aligned}
\overline{R}_k(t) - \overline{R}_k(s) &= \sum_{j=1}^{2} \int_s^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(l)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_l(du) \right) dl \\
&\ge \sum_{j=1}^{2} \int_s^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}^*_k(l)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,*,k}_l(du) \right) dl \\
&= \overline{R}^*_k(t) - \overline{R}^*_k(s).
\end{aligned}$$

From (7.2) and the continuity of $\overline{R}_k$ and $\overline{R}^*_k$, we deduce that $\overline{Q}_k(t) - \overline{Q}_k(s) \le \overline{Q}^*_k(t) - \overline{Q}^*_k(s)$. Combining this with the inequality $\overline{Q}_k(s) \le \overline{Q}^*_k(s)$ proved above, we obtain $\overline{Q}_k(t) \le \overline{Q}^*_k(t)$, which leads to a contradiction. Hence $\overline{Q}_k(v) \le \overline{Q}^*_k(v)$ for all $v \in [0,\tau]$. By symmetry, we can also argue that $\overline{Q}_k(v) \ge \overline{Q}^*_k(v)$ for all $v \in [0,\tau]$. This shows that $\overline{Q}_k = \overline{Q}^*_k$ and, hence, $\overline{R}_k = \overline{R}^*_k$ on $[0,\tau]$. Lastly, by (3.10), we have $\overline{X}_k = \overline{X}^*_k$ on $[0,\tau]$. Thus, $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ agrees with $(\overline{X}^*, \overline{\nu}^*, \overline{\eta}^{1,*}, \overline{\eta}^{2,*})$ on $[0,\tau]$. ■

The following proposition, independent of Assumptions 6.1 and 6.2, holds by using the same argument as the proof of Proposition 4.1 of [14]. It will be used in the proof of Lemma 6.5.

Proposition 7.1 *If $(\overline{X}, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ is a solution to the fluid equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$, then for every $k \in \mathcal{K}$ and $f \in \mathcal{C}_b(\mathbb{R}_+)$,*

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

$$
\begin{aligned}
&\qquad\qquad + \int_{[0,t]} f(t-s)(1-G^r_{1,k}(t-s))\, d\overline{E}_k(s), &(7.3)\\
\int_{[0,H^r_{2,k})} f(x)\,\overline{\eta}^{2,k}_t(dx) \;&=\; \int_{[0,H^r_{2,k})} f(x+t)\frac{1-G^r_{2,k}(x+t)}{1-G^r_{2,k}(x)}\,\overline{\eta}^{2,k}_0(dx)\\
&\qquad\qquad + \int_0^t f(t-s)(1-G^r_{2,k}(t-s))\, d\overline{I}_k(s), &(7.4)\\
\int_{[0,H^s_k)} f(x)\,\overline{\nu}^k_t(dx) \;&=\; \int_{[0,H^s_k)} f(x+t)\frac{1-G^s_k(x+t)}{1-G^s_k(x)}\,\overline{\nu}^k_0(dx)\\
&\qquad\qquad + \int_{[0,t]} f(t-s)(1-G^s_k(t-s))\, d\overline{L}_k(s). &(7.5)
\end{aligned}
$$

Proof of Lemma 6.5. First, (6.26) follows directly from (6.25) and (6.18). Now, we focus on showing (6.25). Fix $k \in \mathcal{K}$ and $T > 0$. It follows from the definition of $A^{(N),j,k}_{\theta^{(N)}_k,\eta}$ in (6.13) 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

$$
\overline{C}^{(N),j}_1(t,\kappa) \doteq \left| \int_0^t \left(\int_{[0,H^r_{j,k})} \left(\mathbb{1}_{[0,\chi^{(N)}_k(s-)\wedge\kappa]}(u) - \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) \right) h^r_{j,k}(u)\, \overline{\eta}^{(N),j,k}_s(du) \right) ds \right|, \quad (7.6)
$$

$$
\overline{C}^{(N),j}_2(t,\kappa) \doteq \Big| \int_0^t \Big(\int_{[0,H^r_{j,k})} \Big(\mathbb{1}_{(\chi^{(N)}_k(s-)\wedge\kappa,\chi^{(N)}_k(s-)]}(u) - \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u) \Big) h^r_{j,k}(u)\, \overline{\eta}^{(N),j,k}_s(du) \Big) ds \Big|, \quad (7.7)
$$

$$
\begin{aligned}
\overline{C}^{(N),j}_3(t,\kappa) \doteq \Bigg| &\int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) h^r_{j,k}(u) \overline{\eta}^{(N),j,k}_s(du) \right) ds\\
&- \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds \Bigg|
\end{aligned} \quad (7.8)
$$

and

$$
\begin{aligned}
\overline{C}^{(N),j}_4(t,\kappa) \doteq \Bigg| &\int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{(N),j,k}_s(du) \right) ds\\
&- \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds \Bigg|.
\end{aligned} \quad (7.9)
$$

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

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

From this, to prove (6.25), 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 [14] by taking $h = 1$ therein that for $s \ge 0$,

$$\eta_s^{(N),k}[0, \chi_k^{(N)}(s-)] = Q_k^{(N)}(s) + \iota_k^{(N)}(s),$$

where

$$\iota_k^{(N)}(s) \doteq \begin{cases} 0 & \text{if } (\chi_k^{(N)}(s-) - \chi_k^{(N)}(s))(L_k^{(N)}(s) - L_k^{(N)}(s-)) = 0, \\ 1 & \text{if } (\chi_k^{(N)}(s-) - \chi_k^{(N)}(s))(L_k^{(N)}(s) - L_k^{(N)}(s-)) > 0. \end{cases}$$

Firstly, note that $\chi_k^{(N)}(s-) \ge \chi_k^{(N)}(s)$ for all $s \ge 0$. It follows from (7.6) and (2.8) that

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

where $\overline{\iota}_k^{(N)}(s) = \iota_k^{(N)}(s)/N$. Since $\overline{E}_k$ and $\overline{\eta}_0^k = \overline{\eta}_0^{1,k} + \overline{\eta}_0^{2,k}$ are continuous, then by (7.3) and (7.4), $\overline{\eta}^k_s$ is also continuous. Thus, by the convergence of $\overline{Q}_k^{(N)}$, $\overline{\iota}_k^{(N)}$ and $\overline{\eta}^{(N),k}$ to $\overline{Q}_k$, $\mathbf{0}$, $\overline{\eta}^k$, respectively, we have for each $s \ge 0$,

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

Note that by (2.18),

$$\begin{aligned}
&\mathbb{E}\left[\int_0^T \left| (\overline{Q}_k^{(N)}(s) + \overline{\iota}_k^{(N)}(s)) \wedge \overline{\eta}_s^{(N),k}[0,\kappa] - \overline{\eta}_s^{(N),k}[0, \overline{\chi}_k(s) \wedge \kappa] \right| ds \right] \\
&\quad \le \mathbb{E}\left[\int_0^T \left(\overline{\eta}_s^{(N),k}[0,\kappa] + \overline{\eta}_s^{(N),k}[0,\kappa] \right) ds \right] \le T\mathbb{E}\left[\langle \mathbf{1}, \eta_0^{(N),k} \rangle + E_k^{(N)}(T) + I_k^{(N)}(T) \right] < \infty.
\end{aligned}$$

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

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

Secondly, by (7.7) and an application of triangle inequality, we have that

$$\overline{C}_2^{(N),j}(t,\kappa) \le 2 \int_0^t \int_{[\kappa, H^r_{j,k})} h^r_{j,k}(u)\, \overline{\eta}_s^{(N),j,k}(du) ds.$$

Moreover, by (7.9), we also have

$$\overline{C}_4^{(N),j}(t,\kappa) \le \int_0^t \int_{[\kappa, H^r_{j,k})} h^r_{j,k}(u)\overline{\eta}_s^{(N),j,k}(du)ds + \int_0^t \int_{[\kappa, H^r_{j,k})} h^r_{j,k}(u)\overline{\eta}_s^{j,k}(du)ds.$$

Thus, by a similar argument in showing (7.30) of [14], we have

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

Lastly, we show that

$$\lim_{\kappa\to H^r_{j,k}} \lim_{N\to\infty} \mathbb{E}\left[\sup_{0\le t\le T} \overline{C}_3^{(N),j}(t,\kappa)\right] = 0. \tag{7.10}$$

Since $1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u)h^r_{j,k}(u)$ lies in $\mathcal{L}^1_{loc}[0,H^r_{j,k})$ and is nonnegative, there exists a sequence of nonnegative continuous functions $\{h_n\}_{n\ge1}$ on $[0,H^r_{j,k})$ such that

$$\lim_{n\to\infty}\int_0^\kappa |1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u)h^r_{j,k}(x) - h_n(x)|dx = 0$$

and h_n has common compact support in $[0,\overline{\kappa}]$, where $\kappa < \overline{\kappa} < H^r$. For each $n\in\mathbb{N}$, by the convergence of $\overline{\eta}^{(N),j,k}$ to $\overline{\eta}^{j,k}$, we have for each $s\in[0,T]$,

$$\lim_{N\to\infty}\int_{[0,H^r_{j,k})} h_n(u)\overline{\eta}_s^{(N),j,k}(du) = \int_{[0,H^r_{j,k})} h_n(u)\overline{\eta}_s^{j,k}(du),$$

and hence another application of the dominated convergence theorem yields that (7.10) holds with h_n in place of $1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]}h^r_{j,k}$. Let $l_n = |h_n - 1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]}h^r_{j,k}|$ for each $n\ge1$. Then, in order to prove (7.10), it clearly suffices to show that the following two limits hold: almost everywhere,

$$\lim_{n\to\infty}\sup_N \int_0^T\left(\int_{[0,\overline{\kappa}]} l_n(u)\overline{\eta}_s^{(N),j,k}(du)\right)ds = 0, \tag{7.11}$$

and

$$\lim_{n\to\infty}\int_0^T\left(\int_{[0,\overline{\kappa}]} l_n(u)\overline{\eta}_s^{j,k}(du)\right)ds = 0. \tag{7.12}$$

By the representations of $\eta^{(N),1,k}$ and $\eta^{(N),2,k}$ in (2.3) and (2.4), respectively, we have

$$\textstyle\int_0^T\left(\int_{[0,\overline{\kappa}]} l_n(u)\overline{\eta}_s^{(N),1,k}(du)\right)ds \le \frac{1}{N}\sum\limits_{j=-\mathcal{E}_k^{(N)}+1}^{0}\int_0^T l_n(w_j^{(N),1,k}(0)+s)1\!\!1_{\{w_j^{(N),1,k}(0)+s\le\overline{\kappa}\wedge r_j^{1,k}\}}\,ds$$
$$+\frac{1}{N}\sum_{j=1}^{E_k^{(N)}(T)}\int_{\zeta_j^{(N),1,k}}^T l_n(s-\zeta_j^{(N),1,k})1\!\!1_{\{s-\zeta_j^{(N),1,k}\le\overline{\kappa}\}}\,ds \le \sup_N\left(\left\langle 1,\overline{\eta}_0^{(N),1,k}\right\rangle + \overline{E}_k^{(N)}(T)\right)\int_0^{\overline{\kappa}} l_n(x)\,dx$$

and

$$\textstyle\int_0^T\left(\int_{[0,\overline{\kappa}]} l_n(u)\overline{\eta}_s^{(N),2,k}(du)\right)ds \le \frac{1}{N}\sum\limits_{j=-\mathcal{C}_k+1}^{0}\int_0^T l_n(w_j^{(N),2,k}(0)+s)1\!\!1_{\{w_j^{(N),2,k}(0)+s\le\overline{\kappa}\wedge r_j^{2,k}\}}\,ds$$
$$+\frac{1}{N}\sum_{j=1}^{I_k^{(N)}(T)}\int_{\zeta_j^{(N),2,k}}^T l_n(s-\zeta_j^{(N),2,k})1\!\!1_{\{s-\zeta_j^{(N),2,k}\le\overline{\kappa}\}}\,ds \le \sup_N\left(\left\langle 1,\overline{\eta}_0^{(N),2,k}\right\rangle + \overline{I}_k^{(N)}(T)\right)\int_0^{\overline{\kappa}} l_n(x)\,dx.$$

Since $\sup_N \left(\left\langle 1, \overline{\eta}_0^{(N),1,k} \right\rangle + \overline{E}_k^{(N)}(T) \right) < \infty$ almost surely, due to Assumption 6.1, $\sup_N \left(\left\langle 1, \overline{\eta}_0^{(N),2,k} \right\rangle + \overline{I}_k^{(N)}(T) \right) < \infty$ almost surely, due to (6.19) and Assumption 6.1, and h_n converges in $\mathcal{L}^1_{loc}[0, H^r_{j,k})$ to $1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]} h^r_{j,k}$, we obtain (7.11). On the other hand, observe that, by (7.14) of Lemma 7.4 of [14] applied to $l = l_n$, there exists a constant $\tilde{\kappa}(\overline{\kappa}, T) < \infty$ such that

$$\int_0^T \left(\int_{[0,\overline{\kappa}]} l_n(u) \overline{\eta}_s^{j,k}(du) \right) ds \leq \tilde{\kappa}(\overline{\kappa}, T) \int_0^{\overline{\kappa}} l_n(x) dx.$$

By the convergence of h_n to $1\!\!1_{[0,\overline{\chi}_k(s)\wedge\kappa]} h^r_{j,k}$ in $\mathcal{L}^1_{loc}[0, H^r_{j,k})$, the term on the right-hand side of the above display converges to 0, as $n \to \infty$, and (7.12) follows. ■