STOCHASTIC SYSTEMS
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1946-5238 | EISSN 1946-5238 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Fluid Limit for a Multi-Class $G/G/1+G$ Queue with Random Order of Service

Reza Aghajani
Google, raghajani01@gmail.com

Ruth J. Williams
Department of Mathematics, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, rjwilliams@ucsd.edu

Abstract here.

Key words: Fluid limit, random order of service, queueing system, measure-valued processes.

Contents

1. Introduction

1.1. Notation and Terminology

For $a, b \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the largest integer smaller than or equal to a, $\lceil a \rceil$ denote the smallest integer larger than or equal to a, and $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b, respectively. Also, $a^+ \doteq a \vee 0$ and $a^- \doteq -(a \wedge 0)$. For a set B, $\mathbb{1}(B)$ is the indicator function of the set B.

For $n \in \mathbb{N}$ and a closed set $V \subset \mathbb{R}^n$, $\mathbf{C}(V)$, $\mathbf{C}_b(V)$, $\mathbf{C}_c(V)$ and $\mathbf{C}_b^1(V)$ are respectively, the spaces of real-valued continuous functions on V, bounded continuous functions on V, continuous functions with compact support in V, once continuously differentiable functions that together with their first (partial) derivatives are bounded on V. When $V = [0, \infty)$, we will write $\mathbf{C}[0, \infty)$ for $\mathbf{C}([0, \infty))$ and analogously for the other spaces. For $f \in \mathbf{C}_b(V)$, we use $\|f\|_V$ to denote the supremum of $|f(s)|$, $s \in V$, and for $f \in \mathbf{C}_b[0, \infty)$, we let $\|f\|_\infty = \|f\|_{[0,\infty)}$.

For a metric space (E,d), let $\mathbb{D}_E[0,\infty)$ denote the set of functions defined from $[0,\infty)$ into E that are right continuous with finite left limits and let $\mathbb{C}_E[0,\infty)$ denote the subspace of continuous functions in $\mathbb{D}_E[0,\infty)$. We endow $\mathbb{D}_E[0,\infty)$ with the usual J_1 Skorokhod topology, see e.g., Ethier and Kurtz (1986).

Let $\mathcal{M}$ denote the space of finite non-negative Borel measures on $[0,\infty)$,endowed with the Prohorov metric d_p, defined by

$$d_p(\mu,\tilde{\mu}) := \inf\{\epsilon>0 \,|\, \mu(A)\leq \tilde{\mu}(A^\epsilon)+\epsilon \text{ and } \tilde{\mu}(A)\leq \mu(A^\epsilon)+\epsilon \text{ for all non-empty Borel sets } A\subset[0,\infty)\},$$

where A^ϵ is the open ϵ-neighborhood of A. The induced topology is that of weak convergence of measures. Moreover, $(\mathcal{M},d_p)$ is a separable metric space, and there is an equivalent metric under which it is complete. Hence it is a Polish space. For every bounded, real-valued, measurable function f on $[0,\infty)$ and every $\nu\in\mathcal{M}$, define

$$\langle f,\nu\rangle = \int_{[0,\infty)} f(x)\nu(dx).$$

Also, define $\mathcal{M}^L$ to be the L-fold product of spaces $(\mathcal{M},d_p)$, endowed with the metric d_p^L defined by

$$d_p^L(\mu,\tilde{\mu}) = \max(d_p(\mu_\ell,\tilde{\mu}_\ell);\ell=1,...,L),$$

for $\mu=(\mu_1,...,\mu_L),\tilde{\mu}=(\tilde{\mu}_1,...,\tilde{\mu}_L)\in\mathcal{M}^L$. This space is also a Polish space.

2. Model Description

We consider a single-server queue with infinite capacity waiting room that provides service to jobs of L different classes indexed by $\ell=1,...,L$. The server processes only one job at a time, working at unit rate; it is non-idling in the sense that it is never idle at a time when there is a job in the system, and it operates under a Discriminatory Random Order of Service policy described below. Jobs are considered to be impatient, that is, if a job has to wait more than a pre-determined patience time before entering service, it abandons the system without being served. Jobs have class-specific arrival rates, patience time distributions, and service time distributions. Fundamental stochastic processes associated with these are now defined.

2.1. Primitive stochastic processes

XXXShould we consider delayed renewal processes?XXX

Job arrivals. Jobs of different classes arrive according to independent processes E_ℓ, $\ell=1,...,L$, where for each $t\geq 0$, $E_\ell(t)$ is the number of job arrivals of class ℓ during $(0,t]$. The arrival processes are assumed to satisfy the following assumption.

ASSUMPTION 2.1. *(Arrival processes) For each $\ell = 1, 2, ..., L$, E_ℓ is a renewal process where the times between arrivals are independent and identically distributed (i.i.d.) with distribution function $G_{E,\ell}$ and reciprocal mean $\lambda_\ell \in (0, \infty)$. It is assumed that $G_{E,\ell}$ has a Borel measurable density $g_{E,\ell}$ with respect to Lebesgue measure on $[0, \infty)$.*

The letter E here indicates *exogenous* arrivals.

Patience times. The patience times of arriving jobs are given by L independent sequences of positive random variables $\{\varpi_{\ell,j}, j \geq 1\}, \ell = 1, ..., L$. For each $\ell \in \{1, \ldots, L\}$ and $j \geq 1$, $\varpi_{\ell,j}$ is the time that the j^{th} arrival to class ℓ is willing to wait in queue before reneging (i.e., abandoning the system). The common distribution function for class ℓ is denoted by $G_{R,\ell}$. We make the following assumption on the patience time distribution functions. (See Section 2.3 for a description of patience times for jobs that are initially in queue.)

ASSUMPTION 2.2. *(Patience times) For each $\ell = 1, ..., L$, the patience time distribution function $G_{R,\ell}$ has a Borel measurable density $g_{R,\ell}$ with respect to Lebesgue measure on $[0, \infty)$, reciprocal mean $\gamma_\ell \in (0, \infty)$, complementary cumulative distribution function $\bar{G}_{R,\ell} = 1 - G_{R,\ell}$ that is positive on $[0, \infty)$, and hazard rate function*

$$h_{R,\ell}(x) := \frac{g_{R,\ell}(x)}{\bar{G}_{R,\ell}(x)} \qquad x \in [0, \infty),$$

that is bounded[1]*.*

The subscript R is used here to indicate that the quantities are associated with reneging.

Service times. The service time requirements of jobs entering service after time zero are determined by L independent sequences of positive random variables $\{v_{\ell,i}, i \geq 1\}, \ell = 1, ..., L$. If the class of the i^{th} job to enter service after time zero is $\ell(i)$, then $v_{\ell(i),i}$ is the amount of service time required by that job. (The random variables $v_{\ell',i}$ for $\ell' \neq \ell(i)$ are not used.) For $\ell = 1, \ldots, L$, the common distribution function for the class ℓ service times is denoted by $G_{S,\ell}$ and is assumed to satisfy the following assumption. (See Section 2.3 for a description of the service time requirement of the job that is initially in service (if any).)

ASSUMPTION 2.3. *(Service times) For each $\ell = 1, ..., L$, the service time distribution $G_{S,\ell}$ has Borel measurable density $g_{S,\ell}$ on $[0, \infty)$, reciprocal mean $\mu_\ell \in (0, \infty)$, complementary cumulative distribution function $\bar{G}_{S,\ell} = 1 - G_{S,\ell}$ that is positive on $[0, \infty)$, and hazard rate function*

$$h_{S,\ell}(x) := \frac{g_{S,\ell}(x)}{\bar{G}_{S,\ell}(x)} \qquad x \in [0, \infty),$$

that is bounded.

[1] Since $\int_a^b h_{R,\ell}(x)dx = \ln(\bar{G}_{R,\ell}(b)) - \ln(\bar{G}_{R,\ell}(a))$ for $0 \leq a < b < \infty$, $h_{R,\ell}$ is not integrable over its support, and hence the bounded hazard rate assumption necessitates that $\bar{G}_{R,\ell}(x) > 0$ for all $x \geq 0$.

The subscript S is used here to indicate that the quantities are associated with service. Let $H < \infty$ be an upper bound on the hazard rate of all service and patience time distributions (which exists by Assumptions 2.2 and 2.3), i.e.,

$$h_{S,\ell}(x), h_{R,\ell}(x) \leq H, \qquad \text{for all } x \geq 0,\ \ell = 1, ..., L. \tag{2.1}$$

We assume that the processes/sequences of random variables, E_ℓ, $\{\varpi_{\ell,j}, j \geq 1\}$, $\{v_{\ell,i}, i \geq 1\}$, $\ell = 1, \ldots, L$ are all mutually independent.

2.2. Service policy: Discriminatory Random Order of Service

We consider the queue operating under the so-called Discriminatory Random Order of Service (DROS) policy. According to this policy, a job enters service immediately upon arrival if the server is idle, and waits in the queue if the server is busy. Once the service requirement is complete for the job in service, it departs the system and the server becomes available immediately to serve the next job. The next job to enter service is randomly chosen from the jobs waiting in queue (if any) according to the following probability law. Denoting the number of jobs of class ℓ waiting in queue at time t by $Q_\ell(t)$, if the server becomes available at time $t > 0$ and the queue is nonempty right before t, each given job of class ℓ waiting in queue just before t (if any) will be chosen for service entry at time t with probability

$$\frac{p_\ell}{\sum_{\ell'=1}^{L} p_{\ell'} Q_{\ell'}(t-)}, \tag{2.2}$$

where $(p_\ell, \ell = 1, ..., L)$ is a probability vector, satisfying $p_\ell > 0, \ell = 1, ..., L$, and $\sum_{\ell=1}^{L} p_\ell = 1$, that determines the preferences for different job classes. Here $Q_{\ell'}(t-)$ is the left limit value of $Q_{\ell'}$ at time t.

More precisely, the random selection of jobs for service entry is determined by a sequence of i.i.d. random variables $\{\kappa_i, i \geq 1\}$ uniformly distributed on $(0, 1]$. These random variables are independent of the primitive processes defined in the previous subsection. If the i^{th} service entry after time 0 happens at a time t and the queue is non-empty right before the service entry, the job chosen to enter service at time t is of class ℓ if $\kappa_i \in (a_\ell(t), b_\ell(t)]$, where

$$a_\ell(t) := \frac{\sum_{\ell'=1}^{\ell-1} p_{\ell'} Q_{\ell'}(t-)}{\sum_{\ell'=1}^{L} p_{\ell'} Q_{\ell'}(t-)}, \qquad b_\ell(t) := \frac{\sum_{\ell'=1}^{\ell} p_{\ell'} Q_{\ell'}(t-)}{\sum_{\ell'=1}^{L} p_{\ell'} Q_{\ell'}(t-)} = a_\ell(t) + \frac{p_\ell Q_\ell(t-)}{\sum_{\ell'=1}^{L} p_{\ell'} Q_{\ell'}(t-)}. \tag{2.3}$$

In addition, all jobs of class ℓ waiting in queue right before the entry time (if any) are equally likely to be selected, and hence, each such job is selected with a probability given by (2.2). On the other hand, if the i^{th} service entry results from a job arrival when the server is idle, no random selection for service entry is required and κ_i is discarded.

XXXIn addition to the random variables κ_i, it seems to me that you need a random variable, once the class is chosen, to choose which of the jobs of that class will enter service.XXXX

2.3. State Descriptor

For every $\ell = 1, ..., L$, recall that $Q_\ell(t)$ denotes the number of jobs of class ℓ waiting in queue at time t, and let $Q_\Sigma(t) = \sum_{\ell=1}^{L} Q_\ell(t)$ be the total number of jobs waiting in queue at time t. We index individual jobs of each class by integers in the order of their arrival as follows: jobs of class ℓ have their indices in the set

$$\mathcal{Z}_\ell := \{-Q_\ell(0) + 1, ..., -1, 0, 1, ...\},$$

where non-positive indices $-Q_\ell(0) + 1, ..., -1, 0$ are assigned to the $Q_\ell(0)$ jobs that are initially in queue (if any), and positive indices are assigned to the jobs that arrive after time $t = 0$. The jobs are ordered according to their arrival times with jobs having arrived earlier having smaller indices. Each job is then uniquely determined by its class-index pair $(\ell, j) : \ell \in \{1, ..., L\}, j \in \mathcal{Z}_\ell$.

Since the patience, service and interarrival times are not (necessarily) exponentially distributed, to have a Markovian state description of the system, we keep track of the age in queue for each job that is currently waiting in queue, the class and time in service of the job currently in service (if any), as well as the time since the last arrival to each class[2].

We first describe the ages of the jobs in queue. If a job with index (ℓ, j) is in queue at time t, let $w_{\ell,j}(t)$ denote the age in queue of the job at time t, which is the amount of time that the job has spent in queue since its arrival (see (4.1) for a full description). We record the age in queue of all waiting jobs at time t in a vector $\nu(t) = (\nu_\ell(t); \ell = 1, ..., L)$ of measure-valued variables defined as follows. At time t, let $\mathcal{Q}_\ell(t)$ be the set of indices of the jobs of class ℓ that are waiting in queue at time t, and define the measure $\nu_\ell(t)$ on $[0, \infty)$ by

$$\nu_\ell(t) := \sum_{j \in \mathcal{Q}_\ell(t)} \delta_{w_{\ell,j}(t)}, \tag{2.4}$$

where δ_a is the Dirac delta measure at a. Therefore, $\nu_\ell(t)$ is the sum of Dirac delta masses located at the age in queue of each job of class ℓ waiting in queue at time t, and takes values in the space $\mathcal{M}$ of finite non-negative Borel measures on $[0, \infty)$ (see (4.10) for a full description). Note that

$$Q_\ell(t) = \langle \mathbf{1}, \nu_\ell(t) \rangle, \tag{2.5}$$

where $\mathbf{1}$ is the constant function on $[0, \infty)$ with value 1.

To describe the state of the server, for each $t \geq 0$, we define the pair $(a(t), I(t))$ as follows: if the server is busy at time t, $a(t)$ and $I(t)$ are, respectively, the age in service (the time passed since entry to service) and the class of the job currently in service at time t; if the server is idle at time t (i.e., no job in service), we define $a(t) = 0$ and $I(t) = 0$.

[2] Alternatively, one could choose to keep track of residual patience, service and interarrival times.

For arrivals, for each $t \geq 0$ and $\ell \in \{1, \ldots, L\}$, we define $\{r_\ell(t), t \geq 0\}$ to be the backward recurrence time of the arrival process E_ℓ, so that $r_\ell(t)$ is the time since the last arrival in $[0, t]$. If there has been no arrival in $[0, t]$, then $r_\ell(t) = t$.

The process

$$\{(r_\ell(t), \nu_\ell(t); \ell = 1, ..., L, a(t), I(t)); t \geq 0\}$$

is a Markovian descriptor for our queueing system. However, we do not use the Markov property of this process directly, and hence do not prove this claim.

2.4. Initial Conditions

The initial condition for our queueing system is described by the collection of random variables

$$\mathfrak{Q}_0 := \{\nu(0) = (\nu_\ell(0); \ell = 1, ..., L), a(0), I(0)\}. \tag{2.6}$$

XXX Should we include a backward recurrence time for arrivals in this? XXX Note that a valid initial condition must be consistent with the non-idling assumption: when $I(0) = 0$ (which means the server is initially idle), one must have $a(0) = 0$ and $\nu_\ell(0)$ is the zero measure on $[0, \infty)$, for each $\ell = 1, ..., L$. The number of jobs initially in queue and their initial ages can be recovered from $\nu(0)$: for $\ell = 1, ..., L$, the number of jobs of class ℓ initially in queue is given by

$$Q_\ell(0) = \langle \mathbf{1}, \nu_\ell(0) \rangle, \tag{2.7}$$

and, given our indexing convention, the initial ages $w_{\ell,j}(0)$ of jobs with indices (ℓ, j) with $j = -Q_\ell(0) + 1, ..., -1, 0$ are given by the locations of the delta masses in $\nu_\ell(0)$ in a manner that the jobs with smaller indices j have larger initial ages in queue. For convenience, we let $w_{\ell,j}(0) = 0$ for $j \leq -Q_\ell(0)$.

Recall that the patience times of jobs of class ℓ arriving after time zero are given by the i.i.d. sequence $\{\varpi_{\ell,j}, j \geq 1\}$. However, for jobs initially in service, the patience times are defined slightly differently. Consider L independent sequences $\{\tilde{\varpi}_{\ell,j}, j \leq 0\}$, $\ell = 1, ..., L$, where for each ℓ, $\tilde{\varpi}_{\ell,j}, j \leq 0$, are conditionally independent of all other random elements given $\nu(0)$ and satisfy

$$\mathbb{P}\{\tilde{\varpi}_{\ell,j} > x | \nu(0)\} = \frac{1 - G_{R,\ell}(x + w_{\ell,j}(0))}{1 - G_{R,\ell}(w_{\ell,j}(0))}, \ x \geq 0.$$

Moreover, define

$$\varpi_{\ell,j} = w_{\ell,j}(0) + \tilde{\varpi}_{\ell,j}, \qquad \ell = 1, ..., L, j \leq 0. \tag{2.8}$$

For each $\ell = 1, ..., L$, the quantities $\{\varpi_{\ell,j}; j = -Q_\ell(0) + 1, ..., -1, 0\}$ are the patience times of jobs of class ℓ that are initially in service, while $\{\tilde{\varpi}_{\ell,j}; j = -Q_\ell(0) + 1, ..., -1, 0\}$ are the residual patience times of those jobs at time 0. The variables $\varpi_{\ell,j}(0)$ and $\tilde{\varpi}_{\ell,j}$ for $j \leq -Q_\ell(0)$ will not be used in

our model description, they are introduced for convenience so that we do not have sequences of random variables of random length.

Similarly, for the job initially in service, the service requirement time is defined as follows. Define random variables $\{\tilde{v}_{\ell,0}\}$, $\ell = 1, ..., L$, that are conditionally independent of all other random elements given $a(0)$, and satisfy

$$\mathbb{P}\left(\tilde{v}_{\ell,0} > x | a(0)\right) = \frac{1 - G_{S,\ell}(x + a(0))}{1 - G_{S,\ell}(a(0))}.$$

Moreover, define $v_{\ell,0} = a(0) + \tilde{v}_{\ell,0}$ for $\ell = 1, ..., 0$, and

$$v_0 := \sum_{\ell=1}^{L} \mathbb{1}\left(I(0) = \ell\right) v_{\ell,0}. \tag{2.9}$$

When the server is not initially idle (i.e. $I(0) \neq 0$), v_0 is the service requirement of the job initially in service, and $\tilde{v}_{\ell,0}$ is the residual service requirement of that job at time 0 if the job is of class ℓ.

REMARK 2.1. Given the initial conditions, the stochastic processes and random variables underlying the model are

- the arrival processes $\{E_\ell; \ell = 1, ..., L\}$,
- the patience times $\{\varpi_{\ell,j}; \ell = 1, ..., L; j \in \mathbb{Z}\}$,
- the job selection random variables $\{\kappa_i; i \geq 1\}$,
- the service time requirements $\{v_{\ell,i}; i \geq 0, \ell = 1, ..., L\}$.

Initial conditions and the random elements listed above are all defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All other random times and state variables can be constructed as functions of the above system primitives.

XXX DO WE NEED ANOTHER SEQUENCE OF RANDOM VARIABLES BESIDES THE κ_i. XXX

3. Main Results

3.1. Fluid scaling

To obtain a fluid limit result for the G/G/1+G queue, we accelerate the arrival processes and effective service rate by a positive integer scale factor of m, while keeping the patience time distribution unchanged. The initial conditions may change somewhat with m. We scale the queue length processes Q_ℓ, and the measure-valued processes ν_ℓ, by the factor $1/m$. The fluid limit is obtained by taking $m \to \infty$. We formally define the scaling below.

Consider a sequence of queueing systems as described above, indexed by a parameter m. The corresponding quantities in the m^{th} system are annotated with a superscript of m. In the m^{th} system, the arrival process E_ℓ^m for each class $\ell \in \{1, \ldots, L\}$ satisfies $E_\ell^m(t) = E_\ell(mt)$ for all $t \geq 0$, where E_ℓ is as in Section 2.1. The patience time distributions are kept unchanged, hence

$$\overline{G}_{R,\ell}^m(x) = \overline{G}_{R,\ell}(x), \qquad g_{R,\ell}^m(x) = g_{R,\ell}(x), \qquad h_{R,\ell}^m(x) = h_{R,\ell}(x), \quad x \in [0, \infty).$$

Indeed, we can take $\varpi^m_{\ell,j} = \varpi_{\ell,j}$ for $\ell = 1,\dots,L$ and $j \geq 1$, where the $\{\varpi_{\ell,j}, j \geq 1\}$, $\ell = 1,\dots,L$ are defined as in Section 2.1. The $\varpi^m_{\ell,j}$ for $j \leq 0$ can be defined as in Section 2.4, using $\nu^m(0)$ and the associated distributions $G^m_{R,\ell}$.

To produce an effective service rate of m in the m^{th} system, the server continues to work at unit rate, but the service time requirements in the m^{th} system are scaled by $1/m$, and hence, the complementary c.d.f. $\overline{G}^m_{S,\ell}$, the density $g^m_{S,\ell}$, and hazard rate function $h^m_{S,\ell}$ of the service times in the m^{th} system are given by

$$\overline{G}^m_{S,\ell}(x) = \overline{G}_{S,\ell}(mx), \qquad g^m_{S,\ell}(x) = m g_{S,\ell}(mx), \qquad h^m_{S,\ell}(x) = m h_{S,\ell}(mx), \quad x \in [0,\infty).$$

Indeed, the service times $\{v^m_{\ell,i}, i \geq 1\}$ can be defined by $v^m_{\ell,i} = \frac{1}{m} v_{\ell,i}$ where the $\{v_{\ell,i}, i \geq 1$ are defined as in Section 2.1 for $\ell = 1,\dots,L$. The $v^m_{\ell,0}$ can be defined as in Section 2.4 with $v^m_{\ell,0}, G^m_{S,\ell}, a^m(0)$ in place of $v_{\ell,0}, G_{S,\ell}, a(0)$, $\ell = 1,\dots,L$.

We define the fluid scaled processes $\bar{\nu}^m_\ell$ and $\bar{Q}^m_\ell$, $\ell = 1,...,L$, by

$$\bar{\nu}^m_\ell(t) = \frac{1}{m}\nu^m_\ell(t), \qquad \bar{Q}^m_\ell(t) = \frac{1}{m}\bar{Q}_\ell(t), \qquad t \geq 0.$$

REMARK 3.1. Alternatively, a sequence of queueing systems indexed by m can be defined by keeping the arrival processes and service times unchanged, scaling the patience times by the factor m, speeding up time, and scaling the "mass" of jobs by $1/m$. This scaling would lead to the same fluid limit as the procedure described above.

3.2. Fluid limit theorem

The main result of this paper is to establish a fluid limit result for the multiclass G/G/1+G queue with a Discriminatory Random Order of Service (DROS) policy as described above. The next assumption, on the convergence of initial conditions, is necessary for our results.

ASSUMPTION 3.1. *The sequence $\{\bar{\nu}^m(0)\}_{m\in\mathbb{N}}$ is tight in $\mathcal{M}^L$.*

Note that by (2.7) and the continuous mapping theorem, Assumption 3.1 implies that

$$\sup_m \mathbb{E}\left[\bar{Q}^m_\ell(0)\right] < \infty. \tag{3.1}$$

THEOREM 3.1. *The sequence $\{\bar{\nu}^m = (\bar{\nu}^m_1,...,\bar{\nu}^m_L)\}_{m\in\mathbb{N}}$ is $\mathbb{C}$-tight in $\mathbb{D}_{\mathcal{M}^L}[0,\infty)$. Moreover, for any weakly convergence subsequence $\{\bar{\nu}^{m_k}\}$ of $\{\bar{\nu}^m\}$, the weak limit $\bar{\nu} = (\bar{\nu}_1,...,\bar{\nu}_L)$ satisfies the following almost surely: for every $\ell = 1,...,L$ and every $f \in \mathbf{C}^1_b[0,\infty)$,*

$$\begin{aligned}\langle f, \bar{\nu}_\ell(t)\rangle =& \langle f, \bar{\nu}_\ell(0)\rangle + \int_0^t \langle f', \bar{\nu}_\ell(s)\rangle ds + \lambda_\ell f(0)\int_0^t \mathbb{1}\left(\bar{\nu}(s) \neq \mathbf{0}\right) ds - \int_0^t \langle f h_{R,\ell}, \bar{\nu}_\ell(s)\rangle ds \\ &- \int_0^t \mathbb{1}\left(\bar{\nu}(s) \neq \mathbf{0}\right) \frac{p_\ell \langle f, \bar{\nu}_\ell(s)\rangle}{\sum_{\ell'=1}^L \frac{p_{\ell'}}{\mu_{\ell'}} \langle \mathbf{1}, \bar{\nu}_{\ell'}(s)\rangle} ds, \qquad \text{for all } t \geq 0,\end{aligned} \tag{3.2}$$

where $\mathbf{0}$ is the vector of zero measures in $\mathcal{M}^L$.

REMARK 3.2. Note that we are not assuming uniqueness of the limit $\bar{\nu}$, this may depend on the subsequence chosen.

4. Dynamics of the State Variables

4.1. Some auxiliary processes

For each class $\ell = 1, ..., L$, recall that the arrival process E_ℓ of class ℓ is a renewal process satisfying Assumption 2.1, and that $\mathcal{Z}_\ell = \{-Q_\ell(0)+1, ..., -1, 0, 1, 2, ...\}$ is the index set of jobs of that class. The arrival time of each job (ℓ, j), $j \in \mathcal{Z}_\ell$, is denoted by $\alpha_{\ell,j}$. For jobs with indices $j \geq 1$ that arrive after time $t = 0$, the arrival times are given by the epoch times of the arrival process E_ℓ. However, for jobs that are initially in queue, the arrival times are non-positive and are determined by the initial conditions, namely, $\alpha_{\ell,j} = -w_{\ell,j}(0)$ for $j \in \{-Q_\ell(0)+1, ..., 0\}$, where $w_{\ell,j}(0)$ is the initial age in queue of the job (ℓ, j). Next, recall that for any $\ell = 1, ..., L$ and $j \in \mathcal{Z}_\ell$, the job with index (ℓ, j) has a patience time $\varpi_{\ell,j}$ that is a member of the i.i.d. sequence described in Assumption 2.2 for jobs that arrive after time zero, and by (2.8) for jobs initially in queue. The age in queue, $w_{\ell,j}(t)$, of the job with index (ℓ, j) at time $t \geq 0$ is given by

$$w_{\ell,j}(t) = \begin{cases} (t - \alpha_{\ell,j})^+ & \text{if } t < \alpha_{\ell,j} + \varpi_{\ell,j}, \\ \varpi_{\ell,j} & \text{if } t \geq \alpha_{\ell,j} + \varpi_{\ell,j}. \end{cases} \tag{4.1}$$

Namely, $w_{\ell,j}(t)$ is zero before the job arrives, then it grows linearly with rate one until it reaches the job's patience time at the time $\sigma_{\ell,j} := \alpha_{\ell,j} + \varpi_{\ell,j}$, which we call the *expiration time* of the job (ℓ, j). If the job (ℓ, j) is still in queue immediately prior to its expiration time (i.e. has not yet entered service), then it leaves the system at its expiration time without receiving service. We assume that jobs do not renege while receiving service.

Recall that a job may enter service either upon arrival (if the server is idle right before its arrival) or right after the service completion of another job (through being selected for service entry by the DROS policy after having waited in queue). The service entry time of a job (ℓ, j) is denoted by $\beta_{\ell,j}$. By convention, we define $\beta_{\ell,j} = +\infty$ if the job (ℓ, j) reneges before it could have entered service. Finally, recall that the service time requirements are determined by L sequences of i.i.d. random variables $\{v_{\ell,i}; i \geq 1\}$, $\ell = 1, ..., L$, whose distributions satisfy Assumption 2.3. Namely, to the i^{th} job to enter service, a service time requirement $v_{\ell,i}$ is assigned if that job is of class ℓ, and the other $v_{\ell',i}$ for $\ell' \neq \ell$ are discarded. Each job remains in service until its service requirement is completed, and then it departs the system and the server will be available immediately for the next job to enter service.

Using the notation introduced above, we can define auxiliary processes that help us described the dynamics of the state descriptors. First, note that the arrival processes satisfy

$$E_\ell(t) = \sum_{j \in \mathcal{Z}_\ell} \mathbb{1}\left(0 < \alpha_{\ell,j} \leq t\right) = \sum_{j=1}^{\infty} \mathbb{1}\left(\alpha_{\ell,j} \leq t\right). \tag{4.2}$$

Next, since the set $\mathcal{Q}_\ell(t)$ of jobs of class ℓ that are waiting in queue at time t is in fact the set of jobs of class ℓ that have arrived to the system, but have not yet reneged or entered service by time t, it can be expressed as

$$\mathcal{Q}_\ell(t) = \{j \geq -Q_\ell(0) + 1 : \alpha_{\ell,j} \leq t, \sigma_{\ell,j} > t, \beta_{\ell,j} > t\}.$$

The process $\{R_{\ell,j}(t); t \geq 0\}$ defined by[3]

$$R_{\ell,j}(t) := \mathbb{1}(\sigma_{\ell,j} \leq t)\, \mathbb{1}(j \in \mathcal{Q}_\ell(\sigma_{\ell,j}-)), \qquad t \geq 0, \tag{4.3}$$

indicates whether the job (ℓ, j) has reneged from the queue at any given time t, and the total number of renegings of jobs of class ℓ during $(0, t]$ is then given by

$$R_\ell(t) = \sum_{j \in \mathcal{Z}_\ell} R_{\ell,j}(t). \tag{4.4}$$

Moreover, the process $\{K_{\ell,j}(t); t \geq 0\}$ defined as

$$K_{\ell,j}(t) := \mathbb{1}(\beta_{\ell,j} \leq t),$$

indicates whether the job (ℓ, j) has entered service by time t, and therefore, the number of service entries $K_\ell(t)$ of jobs of class ℓ and the total number of service entries $K(t)$ during $(0, t]$ are given by

$$K_\ell(t) = \sum_{j \in \mathcal{Z}_\ell} K_{\ell,j}(t) \quad \text{and} \qquad K(t) = \sum_{\ell=1}^{L} K_\ell(t), \tag{4.5}$$

respectively. In addition, he process $\{K^Q_{\ell,j}(t); t \geq 0\}$ defined as

$$K^Q_{\ell,j}(t) := \mathbb{1}(j \in \mathcal{Q}_\ell(\beta_{\ell,j}-))\, \mathbb{1}(\beta_{\ell,j} \leq t) = \int_0^t \mathbb{1}(j \in \mathcal{Q}_\ell(u-))\, dK_{\ell,j}(u), \tag{4.6}$$

indicates whether the job (ℓ, j) has yet entered service by time t after waiting in queue and being selected by DROS policy, and

$$K^Q_\ell(t) := \sum_{j \in \mathcal{Z}_\ell} K^Q_{\ell,j}(t)$$

denotes the total number of such service entries for jobs of class ℓ during $(0, t]$.

For brevity of notation in the equations derived in the next section, we define the following functional-valued versions of the reneging and service entry processes. For every $f \in \mathbf{C}^1_b[0, \infty)$ and $\ell = 1, ..., L$, and $t \geq 0$, define

$$\mathcal{R}_\ell(f; t) := \sum_{j \in \mathcal{Z}_\ell} f(\varpi_{\ell,j}) \mathbb{1}(j \in \mathcal{Q}_\ell(\sigma_{\ell,j}-)) \mathbb{1}(\sigma_{\ell,j} \leq t) = \sum_{j \in \mathcal{Z}_\ell} \int_0^t f(w_{\ell,j}(u)) dR_{\ell,j}(u), \tag{4.7}$$

[3] By definition, the set valued function $u \mapsto \mathcal{Q}_\ell(u)$ is piecewise constant and hence the definition of $\mathcal{Q}_\ell(u-)$ is unambiguous.

and

$$\mathcal{K}_\ell(f;t) := \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(\beta_{\ell,j}))\mathbb{1}\left(j\in\mathcal{Q}_\ell(\beta_{\ell,j}-)\right)\mathbb{1}\left(\beta_{\ell,j}\le t\right) = \sum_{j\in\mathcal{Z}_\ell}\int_0^t f(w_{\ell,j}(s))dK^Q_{\ell,j}(u). \tag{4.8}$$

Roughly speaking, $\mathcal{R}_\ell(\cdot;t)$ keeps track of patience times (or equivalently, the age in queue at reneging time) of those jobs of class ℓ that reneged from the system during $[0,t]$, and $\mathcal{K}_\ell(\cdot;t)$ keeps track of the age in queue at the service entry time for those jobs of class ℓ that entered service during $[0,t]$ after waiting in queue. Note that jobs that enter service directly upon arrival are not counted in $\mathcal{K}_\ell$, and in fact,

$$K^Q_\ell(t) = \mathcal{K}_\ell(\mathbf{1};t). \tag{4.9}$$

Finally, using the notation introduced in this section we can express $\nu_\ell(t)$ as follows:

$$\nu_\ell(t) = \sum_{j=-Q_\ell(0)+1}^{E_\ell(t)} \delta_{w_{\ell,j}(t)}\mathbb{1}\left(\sigma_{\ell,j}>t\right)\mathbb{1}\left(\beta_{\ell,j}>t\right). \tag{4.10}$$

Remark 4.1. It follows from (4.10) and the identity $Q_\ell(t) = \langle \mathbf{1}, \nu_\ell(t)\rangle$ that

$$Q_\ell(t) \le Q_\ell(0) + E_\ell(t). \tag{4.11}$$

Moreover, since the expiration time and service entry time of a job are always after its arrival time, we obtain the elementary bound

$$R_\ell(t) \le Q_\ell(0) + E_\ell(t), \tag{4.12}$$

and,

$$K_\ell(t) \le Q_\ell(0) + E_\ell(t), \tag{4.13}$$

for all $t\ge 0$. These bounds are used later in Section 6.

Remark 4.2. The model has a simpler behavior on a subset of realizations $\tilde{\Omega}\subseteq\Omega$ on which all arrivals, renegings, and departures of all jobs of all classes occur on distinct times, the number of initial jobs in queue is finite, and for every finite time t, the total number of arrivals (and hence renegings and departures) occurred during $(0,t]$ is finite. It is easy to see that $\tilde{\Omega}$ has full measure. The proof is straightforward and hence is omitted; it follows from the fact that the distributions of interarrival times, patience times, and service times have densities, the assumption that the arrival processes have finite means, and assumption 3.1. The remainder of the results in this paper hold on $\tilde{\Omega}$, and hence, almost surely.

4.2. Dynamics of measure-valued state descriptors

To describe the dynamics of the vector of measure-valued state variables $\nu = (\nu_\ell; \ell = 1, ..., L)$, we study the actions of ν_ℓ, $\ell = 1, ..., L$, on the set of test functions $f \in \mathbf{C}_b^1[0, \infty)$.

PROPOSITION 4.1. *Almost surely, for every $\ell = 1, ..., L$, $t \geq 0$, and $f \in \mathbf{C}_b^1[0, \infty)$, ν_ℓ satisfies*

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) \int_0^t \mathbb{1}\left(I(s-) \neq 0\right) dE_\ell(s) - \mathcal{R}_\ell(f;t) - \mathcal{K}_\ell(f;t). \quad (4.14)$$

The terms on the right-hand side of the display above are due to, respectively, the initial condition, the linear growth of ages of jobs in queue, the arrival of new jobs (with age 0 at the time of their arrival), jobs leaving the queue due to abandonment, and jobs leaving the queue due to service entry. The rest of this section is devoted to the proof of Proposition 4.1.

Fix a test function $f \in \mathbf{C}_b^1[0, \infty)$ and $t \geq 0$. For every $\ell = 1, ..., L$, since ν_ℓ is right-continuous by construction, we have

$$\langle f, \nu_\ell(t)\rangle - \langle f, \nu_\ell(0)\rangle = \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \left\langle f, \nu_\ell\left(\frac{k+1}{n}\right) - \nu_\ell\left(\frac{k}{n}\right)\right\rangle = \mathcal{I}_1 + \mathcal{I}_2, \quad (4.15)$$

where

$$\mathcal{I}_1 := \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \left\langle f\left(\cdot + \frac{1}{n}\right), \nu_\ell\left(\frac{k}{n}\right)\right\rangle - \left\langle f, \nu_\ell\left(\frac{k}{n}\right)\right\rangle, \quad (4.16)$$

and

$$\mathcal{I}_2 := \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \left\langle f, \nu_\ell\left(\frac{k+1}{n}\right)\right\rangle - \left\langle f\left(\cdot + \frac{1}{n}\right), \nu_\ell\left(\frac{k}{n}\right)\right\rangle. \quad (4.17)$$

We first compute $\mathcal{I}_1$.

LEMMA 4.1. *Almost surely, for every $f \in \mathbf{C}_b^1[0, \infty)$ and $t \geq 0$,*

$$\mathcal{I}_1 = \int_0^t \langle f', \nu_\ell(s)\rangle ds. \quad (4.18)$$

Proof. For every $s, u \geq 0$, define

$$F(s, u) = \langle f(\cdot + s), \nu_\ell(u)\rangle = \int_{[0,\infty)} f(x + s)\nu_\ell(u, dx).$$

Using this notation, we can write

$$\sum_{k=0}^{\lfloor nt \rfloor} \left\langle f\left(\cdot + \frac{1}{n}\right), \nu_\ell\left(\frac{k}{n}\right)\right\rangle - \left\langle f, \nu_\ell\left(\frac{k}{n}\right)\right\rangle = \sum_{k=0}^{\lfloor nt \rfloor} \left\{F\left(\frac{1}{n}, \frac{k}{n}\right) - F\left(0, \frac{k}{n}\right)\right\}. \quad (4.19)$$

Since f is continuously differentiable with bounded derivative f' and $\nu_\ell(u)$ is a finite measure, for every $u \geq 0$, the mapping $s \mapsto F(s, u)$ is continuously differentiable with derivative $\partial_s F(s, u) =$

$\langle f'(\cdot+s),\nu_\ell(u)\rangle$. Therefore, by the mean value theorem, there exist $s_{n,k}\in(0,1/n)$, for $n\in\mathbb{N}, k=0,...,\lfloor nt\rfloor$, such that

$$\begin{aligned}\sum_{k=0}^{\lfloor nt\rfloor}\left\{F\left(\frac{1}{n},\frac{k}{n}\right)-F\left(0,\frac{k}{n}\right)\right\}&=\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\partial_s F\left(s_{n,k}\,,\frac{k}{n}\right)\\&=\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\partial_s F\left(0,\frac{k}{n}\right)+\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\left\{\partial_s F\left(s_{n,k},\frac{k}{n}\right)-\partial_s F\left(0,\frac{k}{n}\right)\right\}.\end{aligned}\tag{4.20}$$

First, since f' is bounded and continuous, the mapping $u\mapsto\partial_s F(0,u)=\langle f',\nu_\ell(u)\rangle$ is right-continuous and hence, is Riemann integrable, and thus,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\partial_s F\left(0,\frac{k}{n}\right)=\int_0^t\langle f',\nu_\ell(u)\rangle du.\tag{4.21}$$

To bound the other summation, fix any $\epsilon>0$. Define $T=\max\{w_{\ell,j}(0)\mid j=-Q_\ell(0)+1,...,0\}+t$, which is finite since $Q_\ell(0)$, and note that by definition, $w_{\ell,j}(u)\le T$ for every $j\in\mathcal{Z}_\ell$ and $u\in[0,t]$. This implies

$$\langle f',\nu_\ell(u)\rangle=\sum_{j\in\mathcal{Q}_\ell(u)}f'(w_{\ell,j}(u))=\int_{[0,T]}f'(x)\nu_\ell(u,dx),\qquad u\in[0,t].$$

Moreover, since f' is continuous, it is uniformly continuous on $[0,T+1]$ and thus for large enough n, $|f'(x+s_{n,k})-f'(x)|<\epsilon$ for every $x\in[0,T]$ and $k=1,...,\lfloor nt\rfloor$. Therefore, for all such n,

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\left|\partial_s F\left(s_{n,k},\frac{k}{n}\right)-\partial_s F\left(0,\frac{k}{n}\right)\right|&\le\frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\int_{[0,T]}|f'(x+s_{n,k})-f'(x)|\,\nu_\ell\left(\frac{k}{n},dx\right)\\&\le\frac{\epsilon}{n}\sum_{k=0}^{\lfloor nt\rfloor}Q_\ell\left(\frac{k}{n}\right)\\&\le\epsilon t\big(Q_\ell(0)+E_\ell(t)\big),\end{aligned}\tag{4.22}$$

where the last inequality uses (4.11). Recalling definition (4.16) and combining (4.19)-(4.22), we conclude that for every $\epsilon>0$,

$$\left|\mathcal{I}_1-\int_0^t\langle f',\nu_\ell(u)\rangle du\right|\le\epsilon t(Q_\ell(0)+E_\ell(t)).$$

Since $\epsilon>0$ was arbitrary, the desired result follows on letting $\epsilon\to0$. □

The computation of $\mathcal{I}_2$ is more involved. For every $t\ge0$ and $\delta>0$, define the set $\Omega_{t,\delta}$ to be the set of realizations for which there is at most one arrival, reneging, or departure happening during $(t,t+\delta]$.

LEMMA 4.2. *For every $\ell = 1, \ldots L$, $s \geq 0$, $\delta > 0$, and $f \in \mathbf{C}_b^1[0,\infty)$, on $\Omega_{s,\delta}$ we have*

$$\begin{aligned}\langle f, \nu_\ell(s+\delta)\rangle = \langle f(\cdot+\delta), \nu_\ell(s)\rangle &+ \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s+\delta))\mathbb{1}\left(I(s)\neq 0\right)\mathbb{1}\left(\alpha_{\ell,j}\in(s,s+\delta]\right) \\ &- \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s)+\delta)\mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right)\mathbb{1}\left(\sigma_{\ell,j}\in(s,s+\delta]\right) \\ &- \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s)+\delta)\mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right)\mathbb{1}\left(\beta_{\ell,j}\in(s,s+\delta]\right).\end{aligned} \tag{4.23}$$

The first term on the right-hand side above captures the effect of unit-rate growth of ages of jobs in queue, while the next three terms represent the adjustments due to a possible arrival, reneging, or service entry during $(s, s+\delta]$.

Proof. By (2.4),

$$\langle f, \nu_\ell(s+\delta)\rangle = \sum_{j\in\mathcal{Q}_\ell(s+\delta)} f(w_{\ell,j}(s+\delta)). \tag{4.24}$$

The set $\mathcal{Q}_\ell(s+\delta)$ of jobs of class ℓ waiting in queue at time $s+\delta$ admits the partition

$$\mathcal{Q}_\ell(s+\delta) = \mathcal{Q}_\ell^{\text{old}}(s+\delta)\,\dot\cup\,\mathcal{Q}_\ell^{\text{new}}(s+\delta), \tag{4.25}$$

where $\mathcal{Q}_\ell^{\text{old}}(s+\delta) := \mathcal{Q}_\ell(s+\delta)\cap\mathcal{Q}_\ell(s)$ is the set of jobs that have been in queue since time s, and $\mathcal{Q}_\ell^{\text{new}}(s+\delta) = \mathcal{Q}_\ell(s+\delta)\backslash\mathcal{Q}_\ell(s)$ is the set of jobs that have joined the queue during $(s, s+\delta]$. For a job $j\in\mathcal{Q}_\ell^{\text{old}}(s+\delta)$, the age in queue has increased by δ since time s, that is, $w_{\ell,j}(s+\delta) = w_{\ell,j}(s)+\delta$. Moreover, the jobs that are in $\mathcal{Q}_\ell(s)$ but not in $\mathcal{Q}_\ell^{\text{old}}(s+\delta)$ have left the queue, either due to reneging or service entry, during $(s, s+\delta]$. Therefore,

$$\begin{aligned}\sum_{j\in\mathcal{Q}_\ell^{\text{old}}(s+\delta)} f(w_{\ell,j}(s+\delta)) = &\sum_{j\in\mathcal{Q}_\ell(s)} f(w_{\ell,j}(s)+\delta) - \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s)+\delta)\mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right)\mathbb{1}\left(\sigma_{\ell,j}\in(s,s+\delta]\right) \\ &- \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s)+\delta)\mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right)\mathbb{1}\left(\beta_{\ell,j}\in(s,s+\delta]\right).\end{aligned} \tag{4.26}$$

On the other hand, on $\Omega_{s,\delta}$, if a job arrives during $(s, s+\delta]$, no other events (reneging, departure, or other arrivals) can occur during that interval. In that case, the newly arrived job either directly enters service if the system is empty at time s (i.e., if $I(s)=0$), or otherwise, joins the queue and remains in the queue until time $s+\delta$, and hence belongs to $\mathcal{Q}_\ell^{\text{new}}(s+\delta)$. Therefore, on $\Omega_{s,\delta}$,

$$\sum_{j\in\mathcal{Q}_\ell^{\text{new}}(s+\delta)} f(w_{\ell,j}(s+\delta)) = \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(s+\delta))\mathbb{1}\left(I(s)\neq 0\right)\mathbb{1}\left(\alpha_{\ell,j}\in(s,s+\delta]\right) \tag{4.27}$$

The equation (4.23) follows on substituting (4.25)-(4.27) in (4.24). □

For a fixed $\omega\in\tilde\Omega$ (see Remark 4.2) and $t\geq 0$, the time difference between any two events (arrival, reneging, or departure) is strictly positive and the total number of such events during $(0,t]$ is finite. Therefore, the minimum time distance δ^* between any two such events during $(0,t]$ is positive,

and hence, any interval of length less than δ^* contains at most one arrival, reneging or departure. Therefore, for every large n (namely, $n > 1/\delta^*$) and $k = 0, ..., \lfloor nt \rfloor$, we can invoke Lemma 4.2 and obtain (4.23) with $s = k/n$ and $\delta = 1/n$. Substituting (4.23) in (4.17) and changing the order of summation over k and j, we have

$$\mathcal{I}_2 = \lim_{n\to\infty} \Bigg\{ \sum_{j\in\mathcal{Z}_\ell} \sum_{k=0}^{\lfloor nt \rfloor} f\left(w_{\ell,j}\left(\frac{k+1}{n}\right)\right) \mathbb{1}\left(I\left(\frac{k}{n}\right) \neq 0\right) \mathbb{1}\left(\alpha_{\ell,j} \in \left(\frac{k}{n}, \frac{k+1}{n}\right]\right) \tag{4.28a}$$

$$- \sum_{j\in\mathcal{Z}_\ell} \sum_{k=0}^{\lfloor nt \rfloor} f\left(w_{\ell,j}\left(\frac{k}{n}\right) + \frac{1}{n}\right) \mathbb{1}\left(j \in \mathcal{Q}_\ell\left(\frac{k}{n}\right)\right) \mathbb{1}\left(\sigma_{\ell,j} \in \left(\frac{k}{n}, \frac{k+1}{n}\right]\right) \tag{4.28b}$$

$$- \sum_{j\in\mathcal{Z}_\ell} \sum_{k=0}^{\lfloor nt \rfloor} f\left(w_{\ell,j}\left(\frac{k}{n}\right) + \frac{1}{n}\right) \mathbb{1}\left(j \in \mathcal{Q}_\ell\left(\frac{k}{n}\right)\right) \mathbb{1}\left(\beta_{\ell,j} \in \left(\frac{k}{n}, \frac{k+1}{n}\right]\right) \Bigg\}. \tag{4.28c}$$

Note that the total number of arrivals, renegings, and service entries during $(0, t]$ is finite, which justifies changing the order of summation. We now compute the limit as $n \to \infty$ in each term on the right-hand side above separately.

In (4.28a), the summation over k is non-zero only for jobs j that arrive during $(0, (\lfloor nt \rfloor + 1)/n]$. For such a job j, the summand is non-zero only for a single interval $(k/n, k/n + 1/n]$ that contains $\alpha_{\ell,j}$, which corresponds to $k/n = \alpha_{\ell,j,n} := (\lceil n\alpha_{\ell,j} \rceil - 1)/n$. Therefore, the limit of the double sum in (4.28a) can be written as

$$\lim_{n\to\infty} \sum_{j\in\mathcal{Z}_\ell} f\left(w_{\ell,j}\left(\alpha_{\ell,j,n} + \frac{1}{n}\right)\right) \mathbb{1}\left(I(\alpha_{\ell,j,n}) \neq 0\right) \mathbb{1}\left(0 < \alpha_{\ell,j} \leq \frac{\lfloor nt \rfloor + 1}{n}\right).$$

Furthermore, we can interchange the order of summation and the limit because the summand is non-zero only for finitely many $j \in \mathcal{Z}_\ell$. Using the continuity of f and $w_{\ell,j}$, the fact that $\alpha_{\ell,j,n} \uparrow \alpha_{\ell,j}$ as $n \to \infty$, $w_{\ell,j}(\alpha_{\ell,j}) = 0$, and the fact that I is piecewise constant, the last display can be written as

$$\sum_{j\in\mathcal{Z}_\ell} f(0) \mathbb{1}(I(\alpha_{\ell,j}-) \neq 0) \mathbb{1}\left(0 < \alpha_{\ell,j} \leq t\right) = f(0) \int_0^t \mathbb{1}\left(I(s-) \neq 0\right) dE_\ell(s),$$

where the representation (4.2) of E_ℓ is used. Similarly, since $\mathcal{Q}_\ell(\cdot)$ is piecewise constant, the limit of the double sum in (4.28b) can be simplified as

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{j\in\mathcal{Z}_\ell} f\left(w_{\ell,j}\left(\sigma_{\ell,j,n}\right) + \frac{1}{n}\right) \mathbb{1}\left(j \in \mathcal{Q}_\ell(\sigma_{\ell,j,n})\right) \mathbb{1}\left(\sigma_{\ell,j} \leq \frac{\lfloor nt \rfloor + 1}{n}\right) \\
&\quad = \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(\sigma_{\ell,j}) \mathbb{1}(j \in \mathcal{Q}_\ell(\sigma_{\ell,j}-))) \mathbb{1}\left(\sigma_{\ell,j} \leq t\right) \\
&\quad = \mathcal{R}_\ell(f; t),
\end{aligned}$$

where $\sigma_{\ell,j,n} := (\lceil n\sigma_{\ell,j}\rceil - 1)/n \uparrow \sigma_{\ell,j}$ as $n \to \infty$, and the identity $w_{\ell,j}(\sigma_{\ell,j}) = \varpi_{\ell,j}$ and definition (4.7) of $\mathcal{R}_\ell$ are used. Finally, the limit of the double sum in (4.28c) can similarly be simplified as

$$\begin{aligned}\lim_{n\to\infty} \sum_{j\in\mathcal{Z}_\ell} f\left(w_{\ell,j}\left(\beta_{\ell,j,n}\right) + \frac{1}{n}\right) \mathbb{1}\left(j \in \mathcal{Q}_\ell(\beta_{\ell,j,n})\right) \mathbb{1}\left(\beta_{\ell,j} \le \frac{\lfloor nt\rfloor + 1}{n}\right) \\ = \sum_{j\in\mathcal{Z}_\ell} f(w_{\ell,j}(\beta_{\ell,j}))\mathbb{1}\left(j \in \mathcal{Q}_\ell(\beta_{\ell,j}-)\right) \mathbb{1}\left(\beta_{\ell,j} \le t\right) \\ = \mathcal{K}_\ell(f;t),\end{aligned}$$

where $\beta_{\ell,j,n} := (\lceil n\beta_{\ell,j}\rceil - 1)/n \uparrow \beta_{\ell,j}$ as $n \to \infty$, and definition (4.8) of $\mathcal{K}_\ell$ is used.

Substituting the last three displays in (4.28a)-(4.28c) for $\mathcal{I}_2$, we obtain

$$\mathcal{I}_2 = f(0)\int_0^t \mathbb{1}(I(s-) \neq 0) dE_\ell(s) - \mathcal{R}_\ell(f;t) - \mathcal{K}_\ell(f;t). \tag{4.29}$$

Equation (4.14) then follows on substituting equalities (4.18) and (4.29) in (4.15). This concludes the proof of Proposition 4.1.

5. Martingale Decompositions

The goal of this section is to construct martingale decompositions for auxiliary processes describing the dynamics of the state variables in equation (4.14). This is an important step in the proof of Theorem 3.1. The result is summarized Proposition 5.1 below.

We start by constructing a marked point process $\mathcal{T} = \{(\tau_k, z_k); k \in \mathbb{N}\}$ that records all activities in the system. The increasing sequence $(\tau_k; k \in \mathbb{N})$, which we call *event times*, contains all times of entry to the system, all times at which a job reneges, and all times at which a job departs the system (recall from Remark 4.2 that almost surely, all such times are distinct). The sequence $(z_k; k \in \mathbb{N})$ of *marks* appended to the event times is defined as follows.

- Each arrival time is marked by $(\mathfrak{E}, \ell, j)$, where $\mathfrak{E}$ indicates that the event is an arrival and ℓ and j are the class and the index of the job arrived, respectively.
- Each reneging time is marked by $(\mathfrak{R}, \ell, j)$ where $\mathfrak{R}$ indicates that the event is a reneging and ℓ and j are the class and the index of the job reneged.
- For departure times, recall that by the non-idling condition, when the queue is non-empty right before a departure, that departure is followed (immediately) by a service entry. Such a departure time is marked by $(\mathfrak{D}, \ell, j)$ where $\mathfrak{D}$ indicates that the event is a departure and ℓ and j are the class and index of the job that consequently entered service. For a departure that leaves the system empty (i.e. when the queue is empty right before the departure), the departure time is marked by $(\mathfrak{D}, \emptyset)$ where $\emptyset$ indicates that no service entry is followed by this departure.

The marked point process $\mathcal{T}$ is constructed inductively. Namely, the event time τ_{k+1} after τ_k is the minimum of the next arrival time, the next expiration time among the jobs waiting in queue, and the next departure time after τ_k. Therefore, defining $\tau_{k+1}^{\mathfrak{E},\ell}$ to be the next arrival time of a job of class ℓ after τ_k, and $\tau_{k+1}^{\mathfrak{D}}$ to be the next departure time after τ_k if the server is busy at τ_k and $\tau_{k+1}^{\mathfrak{D}} = +\infty$ otherwise, and recalling that $\sigma_{\ell,j}$ is the expiration time of the job (ℓ, j), one has

$$\tau_{k+1} = \min\{\tau_{k+1}^{\mathfrak{E},\ell}; \ell = 1, ..., L\} \wedge \min\{\sigma_{\ell,j}; \ell = 1, ..., L, j \in \mathcal{Q}_\ell(\tau_k)\} \wedge \tau_{k+1}^{\mathfrak{D}}. \tag{5.1}$$

Define $\{\mathcal{F}_t; t \geq 0\}$ to be the filtration generated by the marked point process $\mathcal{T}$ (also called the internal history of $\mathcal{T}$; see page 2 of Bremaud (1981)).

We can now state the main result of this section.

PROPOSITION 5.1. *For every $\ell = 1, ..., L$ and $f \in \mathbf{C}_b^1[0, \infty)$, there exist $\{\mathcal{F}_t\}$-martingales $\{M_{R,\ell}(f;t); t \geq 0\}$ and $\{M_{S,\ell}(f;t); t \geq 0\}$ such that, almost surely,*

$$\begin{aligned}\langle f, \nu_\ell(t)\rangle =& \langle f, \nu_\ell(0)\rangle + f(0) \int_0^t \mathbb{1}\left(I(s-) \neq 0\right) dE_\ell(s) - \int_0^t \langle f h_{R,\ell}, \nu_\ell(s)\rangle ds \\ &- \int_0^t h_{S,I(s-)}\left(a(s)\right) \frac{p_\ell \langle f, \nu_\ell(s)\rangle}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(s)} ds + M_{R,\ell}(f;t) + M_{S,\ell}(f;t), \qquad t \geq 0. \end{aligned} \tag{5.2}$$

The proposition is proved at the end of this section.

5.1. General theory of marked point processes

For a given mark z, define $\mathcal{T}(z;\cdot)$ to be the process that counts events with mark z that happened up to time t:

$$\mathcal{T}(z;t) = \sum_{k \in \mathbb{N}} \mathbb{1}\left(\tau_k \leq t\right) \mathbb{1}\left(z_k = z\right). \tag{5.3}$$

As detailed in the next two sections, the auxiliary processes describing the dynamics of the state variables in equation (4.14) can be written in terms of Stieltjes integrals with respect to counting processes of the from (5.3). We use the theory of marked point processes to characterize the stochastic intensities for the process $\mathcal{T}(z;t)$ (see (Bremaud 1981, Definition II.D7) for a definition of stochastic intensity), and then invoke the elementary Lemma 5.2 below to construct martingale decompositions for our auxiliary processes of interest.

PROPOSITION 5.2. *For a fixed mark z and at each event time τ_k, suppose the conditional distribution, given the past $\mathcal{F}_{\tau_k}$, of the next event time with mark z admit a density, that is, there exists a sequence of measurable functions $\{g_k^z; k \geq 1\}$ such that for every interval $A \subset [0, \infty)$,*

$$\mathbb{P}\left(\tau_{k+1} - \tau_k \in A, z_{k+1} = z | \mathcal{F}_{\tau_k}\right)(\omega) = \int_A g_{k+1}^z(\omega, x) dx, \tag{5.4}$$

almost surely. Then, the process

$$\sum_{k\geq 0} \frac{g^z_{k+1}(t-\tau_k)}{\mathbb{P}\left(\tau_{k+1}>t|\mathcal{F}_{\tau_k}\right)} \mathbb{1}\left(\tau_k \leq t < \tau_{k+1}\right). \tag{5.5}$$

is an $\{\mathcal{F}_t\}$*-stochastic intensity of* $\mathcal{T}(z;\cdot)$*.*

Proof. The result follows from Theorem T7, comment (β) below the theorem in (Bremaud 1981, page 61), and the fact that the sequence $\{\tau_k; k\geq 1\}$ of event times is non-explosive, almost surely (see Remark 4.2). □

To invoke Proposition 5.2 for finding stochastic intensities of counting processes of the form (5.3), we need to compute conditional distributions of the next arrival times, the next expiration times for jobs in queue, and the next departure time after τ_k, conditioned on $\mathcal{F}_{\tau_k}$. Recall the sequence of selection random variables $\{\kappa_i; i\geq 1\}$, and note that the value of $\kappa_{K(\tau_k)+1}$ determines the next job after τ_k to be drawn from the queue (if any) for service entry.

LEMMA 5.1. *For every* $k\geq 0$, $\tau^{\mathfrak{E},\ell}_{k+1}, \ell=1,...,L$, $\sigma_{\ell,j}, \ell=1,...,L$, $j\in\mathcal{Q}_\ell(\tau_k)$, $\tau^{\mathfrak{D}}_{k+1}$, *and* $\kappa_{K(\tau_k)+1}$ *are conditionally independent given* $\mathcal{F}_{\tau_k}$*. Moreover, almost surely, for* $\ell=1,...,L$ *and* $j\in\mathbb{Z}$*,*

$$\mathbb{1}\left(j\in\mathcal{Q}_\ell(\tau_k)\right)\mathbb{P}\left(\sigma_{\ell,j}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(j\in\mathcal{Q}_\ell(\tau_k)\right)\frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k)+x)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))}, \quad x>0, \tag{5.6}$$

$$\mathbb{1}\left(I(\tau_k)\neq 0\right)\mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(I(\tau_k)\neq 0\right)\frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+x)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}, \quad x>0, \tag{5.7}$$

and

$$\mathbb{P}\left(\kappa_{K(\tau_k)+1}>x|\mathcal{F}_{\tau_k}\right)=1-x, \quad x\in[0,1]. \tag{5.8}$$

The result in Lemma 5.1 is intuitive and follows from the independence of the interarrival, patience times, service times, and service entry selection random variable. However, a completely rigorous proof is rather technical, although involving fairly routine calculations. Hence, we omit the proof, and refer the reader to Appendix A for details.

LEMMA 5.2. *Let* N *be a point process adapted to* $\{\mathcal{F}_t\}$ *with an* $\{\mathcal{F}_t\}$*-intensity* u*, and let* θ *be a locally bounded,* $\{\mathcal{F}_t\}$*-predictable process. Define* $\zeta(t):=\int_0^t\theta(s)dN(s)$*. Then* $M(t):=\zeta(t)-\int_0^t\theta(s)u(s)ds, t\geq 0$*, is a local* $\{\mathcal{F}_t\}$*-martingale with quadratic variation*

$$[M](t)=\int_0^t\theta^2(s)dN(s), \qquad t\geq 0. \tag{5.9}$$

Proof. The proof is elementary and follows from (Bremaud 1981, Lemma II.L3) and (Rogers and Williams 2000, Equation (18.1) in Chapter IV)). □

5.2. Martingale Decomposition for Reneging

Recall that the event time at which a job (ℓ, j) reneges system is marked by $(\mathfrak{R}, \ell, j)$. Therefore, the process $R_{\ell,j}$ that indicates whether the job (ℓ, j) has yet reneged by time t has the representation

$$R_{\ell,j}(t) = \sum_{k\in\mathbb{N}} \mathbb{1}\left(\tau_k \le t\right) \mathbb{1}\left(z_k = (\mathfrak{R}, \ell, j)\right) = \mathcal{T}((\mathfrak{R}, \ell, j); t) \tag{5.10}$$

We can therefore use Proposition 5.2 to compute stochastic intensity of $R_{\ell,j}$.

LEMMA 5.3. *For every $\ell = 1, ..., L$ and $j \in \mathcal{Z}_\ell$, an $\{\mathcal{F}_t\}$-stochastic intensity of $R_{\ell,j}$ is given by*

$$\{\mathbb{1}\left(j \in \mathcal{Q}_\ell(t)\right) h_{R,\ell}\left(w_{\ell,j}(t)\right); t \ge 0\}. \tag{5.11}$$

Proof. To invoke Proposition 5.2, we first need to compute the conditional density $g_{k+1}^{(\mathfrak{R},\ell,j)}$ defined as

$$\mathbb{P}\left(\tau_{k+1} - \tau_k \in A, z_{k+1} = (\mathfrak{R}, \ell, j) \middle| \mathcal{F}_{\tau_k}\right) = \int_A g_{k+1}^{(\mathfrak{R},\ell,j)}(x)dx, \quad A \in \mathcal{B}[0, \infty), \tag{5.12}$$

The next event has a mark $z_{k+1} = (\mathfrak{R}, \ell, j)$ if the job (ℓ, j) is in the queue at time τ_k and its expiration time occurs prior to the other possible events after τ_k. Define $\tau_{k+1}^{-\mathfrak{R},\ell,j}$ to be the minimum of all possible event times after τ_k except for the expiration time of job (ℓ, j), that is,

$$\tau_{k+1}^{-\mathfrak{R},\ell,j} := \min\{\tau_{k+1}^{\mathfrak{E},\ell'}; \ell' = 1, ..., L\} \wedge \min\{\sigma_{\ell',j'}; \ell' = 1, ..., L, j' \in \mathcal{Q}_{\ell'}(\tau_k), (\ell', j') \ne (\ell, j)\} \wedge \tau_{k+1}^{\mathfrak{D}}.$$

Then, on $\{j \in \mathcal{Q}_\ell(\tau_k)\}$, $\tau_{k+1} = \sigma_{\ell,j} \wedge \tau_{k+1}^{-\mathfrak{R},\ell,j}$ and $\tau_{k+1}^{-\mathfrak{R},\ell,j}$ and $\sigma_{\ell,j}$ are conditionally independent given $\mathcal{F}_{\tau_k}$, according to Lemma 5.1. Therefore, by (5.6),

$$\begin{aligned}
&\mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{R}, \ell, j) \middle| \mathcal{F}_{\tau_k}\right) \\
&\quad = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\sigma_{\ell,j} > \tau_k + t, \tau_{k+1}^{-\mathfrak{R},\ell,j} > \sigma_{\ell,j} \middle| \mathcal{F}_{\tau_k}\right) \\
&\quad = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \int_t^\infty \frac{g_{R,\ell}(w(\tau_k) + s)}{\overline{G}_{R,\ell}(w(\tau_k))} \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{R},\ell,j} > \tau_k + s \middle| \mathcal{F}_{\tau_k}\right) ds
\end{aligned} \tag{5.13}$$

Moreover, on $\{j \notin \mathcal{Q}_\ell(\tau_k)\}$ where the job (j, ℓ) is not waiting in queue at time τ_k, the next event after τ_k cannot be the reneging time of (j, ℓ) and hence, the probability on the right-hand side of (5.12) is zero. Therefore, the conditional density $g_{k+1}^{(\mathfrak{R},\ell,j)}$ exists, and for $t \ge \tau_k$,

$$g_{k+1}^{(\mathfrak{R},\ell,j)}(t - \tau_k) = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \frac{g_{R,\ell}(w_{\ell,j}(\tau_k) + t - \tau_k)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))} \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{R},\ell,j} > t \middle| \mathcal{F}_{\tau_k}\right). \tag{5.14}$$

In addition, analogous to the calculations in (5.13), on $\{j \in \mathcal{Q}_\ell(\tau_k)\}$, for every $k \ge 0$ and $t \ge \tau_k$ we have

$$\begin{aligned}
\mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1} > t | \mathcal{F}_{\tau_k}\right) &= \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{R},\ell,j} > t | \mathcal{F}_{\tau_k}\right) \mathbb{P}\left(\sigma_{\ell,j} > t | \mathcal{F}_{\tau_k}\right) \\
&= \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{R},\ell,j} > t | \mathcal{F}_{\tau_k}\right) \frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k) + t - \tau_k)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))}.
\end{aligned} \tag{5.15}$$

Plugging (5.14) and (5.15) in (5.5), recalling that $h_{R,\ell} = g_{R,\ell}/\overline{G}_{R,\ell}$ is the hazard rate function associated to the patience times of jobs of class ℓ, and noting that $\mathcal{Q}_\ell$ is constant during each interval (τ_k, τ_{k+1}) and that $w_{\ell,j}$ is continuous and grows linearly during each interval (τ_k, τ_{k+1}) where $j \in \mathcal{Q}_\ell(\tau_k)$, we conclude that the process

$$\sum_{k=0}^{\infty} \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) h_{R,\ell}(w_{\ell,j}(\tau_k) + t - \tau_k) \mathbb{1}\left(\tau_k \le t < \tau_{k+1}\right) = \mathbb{1}\left(j \in \mathcal{Q}_\ell(t)\right) h_{R,\ell}(w_{\ell,j}(t))$$

is an $\{\mathcal{F}_t\}$-intensity of $R_{\ell,j}$. $\square$

LEMMA 5.4. *For every $f \in \mathbf{C}_b^1[0,\infty)$ and $\ell = 1, ..., L$, the process $\{\mathcal{R}_\ell(f;t); t \ge 0\}$ admits a decomposition*

$$\mathcal{R}_\ell(f;t) = \int_0^t \langle f h_{R,\ell}, \nu_\ell(s)\rangle ds + M_{R,\ell}(f;t) \tag{5.16}$$

where $\{M_{R,\ell}(f;t); t \ge 0\}$ is a local $\{\mathcal{F}_t\}$-martingale with quadratic variation

$$[M_{R,\ell}(f;\cdot)](t) = \mathcal{R}(f^2;t). \tag{5.17}$$

Proof. By definition (4.7), we can write $\mathcal{R}_\ell$ as

$$\mathcal{R}_\ell(f;t) = \sum_{j \in \mathcal{Z}_\ell} \mathcal{R}_{\ell,j}(f;t), \tag{5.18}$$

where

$$\mathcal{R}_{\ell,j}(f;t) = \int_0^t f(w_{\ell,j}(s-)) dR_{\ell,j}(s). \tag{5.19}$$

Consider the setup of Lemma 5.2 with N replaced by $R_{\ell,j}$, $\theta(t)$ replaced by $f(w_{\ell,j}(t-))$, and $\zeta(t)$ replaced by $\mathcal{R}_{\ell,j}(f;t)$. Using the from of stochastic intensity of $R_{\ell,j}$ given by Lemma 5.3, the process $\{M_{R,\ell,j}(f;t); t \ge 0\}$ defined as

$$M_{R,\ell,j}(f;t) := \mathcal{R}_{\ell,j}(f;t) - \int_0^t \mathbb{1}\left(j \in \mathcal{Q}_\ell(s)\right) f((w_{\ell,j}(s)) h_{R,\ell}(w_{\ell,j}(s)) ds \tag{5.20}$$

is a local martingale. Summing (5.20) over $j \in \mathcal{Z}_\ell$, by definition (2.4) of ν_ℓ, we obtain,

$$\begin{aligned}\mathcal{R}_\ell(f;t) &= \int_0^t \sum_{j \in \mathcal{Z}_\ell} \mathbb{1}\left(j \in \mathcal{Q}_\ell(s)\right) f((w_{\ell,j}(s)) h_{R,\ell}(w_{\ell,j}(s)) ds + \sum_{j \in \mathcal{Z}_\ell} M_{R,\ell,j}(f;t) \\ &= \int_0^t \langle f h_{R,\ell}, \nu_\ell(t)\rangle + M_{R,\ell}(f;t)\end{aligned}$$

where $M_{R,\ell}(f;\cdot) := \sum_{j \in \mathcal{Z}_\ell} M_{R,\ell,j}(f;\cdot)$ is a local martingale.

Moreover, almost surely, for $j, j' \in \mathcal{Z}_\ell$ and $j' \ne j$, $\mathcal{R}_{\ell,j}(f,\cdot)$ and $\mathcal{R}_{\ell,j'}(f,\cdot)$ are bounded pure jump processes with no common jump times. Therefore,

$$[\mathcal{R}_{\ell,j}(f;\cdot)](t) = \int_0^t f^2(w_{\ell,j}(s-)) dR_{\ell,j}(s) = \mathcal{R}_{\ell,j}(f^2;t),$$

and $[\mathcal{R}_{\ell,j}(f,\cdot),\mathcal{R}_{\ell,j'}(f,\cdot)] \equiv 0$. Together with the fact that $\mathcal{R}_\ell(f;\cdot) - M_{R,\ell}(f;\cdot)$ is a continuous process with finite variation, this implies

$$[M_{R,\ell}(f;\cdot)] = [\mathcal{R}_\ell(f;\cdot)] = [\sum_{j\in\mathcal{Z}_\ell} \mathcal{R}_{\ell,j}(f;\cdot)] = \sum_{j\in\mathcal{Z}_\ell} [\mathcal{R}_{\ell,j}(f;\cdot)] = \sum_{j\in\mathcal{Z}_\ell} \mathcal{R}_{\ell,j}(f^2;\cdot) = \mathcal{R}_\ell(f^2;\cdot).$$

□

5.3. Martingale Decomposition for Service Entry

Recall that a job may enter service either upon arrival if the server is idle right before its arrival, or right after departure of another job by being selected by the DROS policy while waiting in queue. The process $\{K^Q_{\ell,j}(t); t \geq 0\}$, defined in (4.6), indicates whether the job (ℓ, j) has yet entered service after waiting in queue through being selected by DROS policy by time t. Recalling that the time of this service entry is marked by $(\mathfrak{D}, \ell, j)$, $K^Q_{\ell,j}$ has the representation

$$K^Q_{\ell,j}(t) = \sum_{k\in\mathbb{N}} \mathbb{1}(\tau_k \leq t)\, \mathbb{1}(z_k = (\mathfrak{D},\ell,j)) = \mathcal{T}((\mathfrak{D},\ell,j);t). \tag{5.21}$$

LEMMA 5.5. *For every $\ell = 1, ..., L$ and $j \in \mathcal{Z}_\ell$, an $\mathcal{F}$-stochastic intensity of $K^Q_{\ell,j}$ is given by*

$$\{\mathbb{1}(j \in \mathcal{Q}_\ell(t))\, h_{S,I(t)}(a(t)) \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(t)}; t \geq 0\}. \tag{5.22}$$

Proof. To invoke Proposition 5.2, we first need to compute the conditional density $g^{(\mathfrak{D},\ell,j)}_{k+1}$ defined through

$$\mathbb{P}\left(\tau_{k+1} - \tau_k \in A, z_{k+1} = (\mathfrak{D},\ell,j) \middle| \mathcal{F}_{\tau_k}\right) = \int_A g^{(\mathfrak{D},\ell,j)}_{k+1}(x)dx, \quad A \in \mathcal{B}[0,\infty). \tag{5.23}$$

The next event after τ_k has a mark $z_{k+1} = (\mathfrak{D},\ell,j)$ if 1) the next event after τ_k is a departure, and 2) the job (ℓ, j) is waiting in queue at τ_k and is selected for service entry by DROS service policy. Recall that $\tau^{\mathfrak{D}}_{k+1}$ is the next departure time after τ_k, and define $\tau^{-\mathfrak{D}}_{k+1}$ to be the minimum of all other possible event times after τ_k except for the next departure, that is,

$$\tau^{-\mathfrak{D}}_{k+1} := \min\{\tau^{\mathfrak{E},\ell}_{k+1}; \ell = 1, ..., L\} \wedge \min\{\sigma_{\ell,j}; \ell = 1, ..., L,\, j \in \mathcal{Q}_\ell(\tau_k)\}.$$

Then $\tau_{k+1} = \tau^{\mathfrak{D}}_{k+1} \wedge \tau^{-\mathfrak{D}}_{k+1}$, and $\tau^{-\mathfrak{D}}_{k+1}$ and $\tau^{\mathfrak{D}}_{k+1}$ are conditionally independent given $\mathcal{F}_{\tau_k}$, according to Lemma 5.1. Moreover, on $j \in \mathcal{Q}_\ell(\tau_k)$ and when τ_{k+1} is a departure time, $K(\tau_{k+1}) = K(\tau_k) + 1$ indicates the total number of service entries up to time τ_{k+1}, and hence, the class and index of the next job selected by DROS policy to enter service (if any) at τ_{k+1} is determined by $\kappa_{K(\tau_k)+1}$, which is conditionally independent of $\tau^{-\mathfrak{D}}_{k+1}$ and $\tau^{\mathfrak{D}}_{k+1}$ given $\mathcal{F}_{\tau_k}$, by another application of Lemma 5.1. Therefore, by (5.7) and (5.8),

$$\mathbb{P}\left(\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{D},\ell,j) \middle| \mathcal{F}_{\tau_k}\right) = \mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1} > \tau_k + t, \tau^{-\mathfrak{D}}_{k+1} > \sigma_{\ell,j}, \kappa_{K(\tau_k)+1} \in I_{\ell,j}(\tau_k) \middle| \mathcal{F}_{\tau_k}\right),$$

where $I_{\ell,j}(\tau_k)$ is a subinterval of $(a_\ell(\tau_{k+1}), b_\ell(\tau_{k+1})]$, such that $\kappa_{K(\tau_k)+1} \in I_{\ell,j}(\tau_k)$ leads to the selection of (ℓ, j) for the next service entry. By (2.2), the length of the subinterval $I_{\ell,j}(\tau_k)$ equals

$$\frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(\tau_{k+1}-)} = \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(\tau_k)},$$

when $j \in \mathcal{Q}_\ell(\tau_k)$, and equals 0 otherwise. Therefore, the right-hand side of display above can be written as

$$\mathbb{P}\left(\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{D}, \ell, j) \middle| \mathcal{F}_{\tau_k}\right) = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(\tau_k)} \int_t^\infty \frac{g_{S,I(\tau_k)}(a(\tau_k)+s)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))} \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{D}} > \tau_k + s \middle| \mathcal{F}_{\tau_k}\right) ds. \tag{5.24}$$

Therefore, the conditional density $g_{k+1}^{(\mathfrak{D},\ell,j)}$ exists, and for $t \geq \tau_k$,

$$g_{k+1}^{(\mathfrak{D},\ell,j)}(t-\tau_k) = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \frac{g_{S,I(\tau_k)}(a(\tau_k)+t-\tau_k)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))} \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(\tau_k)} \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{D}} > t \middle| \mathcal{F}_{\tau_k}\right). \tag{5.25}$$

Similarly, using $\tau_{k+1} = \tau_{k+1}^{\mathfrak{D}} \wedge \tau_{k+1}^{-\mathfrak{D}}$ and independence of $\tau_{k+1}^{\mathfrak{D}}$ and $\tau_{k+1}^{-\mathfrak{D}}$, for every $k \geq 0$ and $t \geq \tau_k$ we have

$$\mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1} > t | \mathcal{F}_{\tau_k}\right) = \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) \mathbb{P}\left(\tau_{k+1}^{-\mathfrak{D}} > t | \mathcal{F}_{\tau_k}\right) \frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+t)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}. \tag{5.26}$$

Plugging (5.25) and (5.26) in (5.5) of Proposition 5.2, recalling that $h_{S,\ell} = g_{S,\ell}/\overline{G}_{S,\ell}$ is the hazard rate function associated to the service times of jobs of class ℓ, and noting that $\mathcal{Q}_\ell$ is constant during each interval (τ_k, τ_{k+1}) and that $w_{\ell,j}$ is continuous and grows linearly during each interval (τ_k, τ_{k+1}) where $j \in \mathcal{Q}_\ell(\tau_k)$, we conclude that the process

$$\begin{aligned}\sum_{k=0}^\infty \mathbb{1}\left(j \in \mathcal{Q}_\ell(\tau_k)\right) h_{S,I(\tau_k)}(a(\tau_k)+t-\tau_k) \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(\tau_k)} \mathbb{1}\left(\tau_k \leq t < \tau_{k+1}\right) \\ = \mathbb{1}\left(j \in \mathcal{Q}_\ell(t)\right) h_{S,I(t)}(a(t)) \frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(t)}.\end{aligned}$$

is an $\{\mathcal{F}_t\}$-intensity of $K_{\ell,j}^Q$. □

PROPOSITION 5.3. *For every $f \in \mathbf{C}_b^1[0,\infty)$ and $\ell = 1, ..., L$, the process $\{\mathcal{K}_\ell(f;t); t \geq 0\}$ admits a decomposition*

$$\mathcal{K}_\ell(f;t) = \int_0^t h_{S,I(s)}\left(a(s)\right) \frac{p_\ell \langle f, \nu_\ell(s)\rangle}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(s)} ds + M_{S,\ell}(f;t), \tag{5.27}$$

where $\{M_{S,\ell}(f;t); t \geq 0\}$ is a local $\{\mathcal{F}_t\}$-martingale with quadratic variation

$$[M_{S,\ell}](t) = \mathcal{K}_\ell(f^2;t).$$

Proof. We can write

$$\mathcal{K}_\ell(f;t) = \sum_{j\in\mathbb{Z}} \mathcal{K}_{\ell,j}(f;t), \tag{5.28}$$

where

$$\mathcal{K}_{\ell,j}(f;t) = \int_0^t f(w_{\ell,j}(s-))dK^Q_{\ell,j}(s). \tag{5.29}$$

Consider the setup of Lemma 5.2 with N replaced by $K^Q_{\ell,j}$, $\theta(t)$ replaced by $f(w_{\ell,j}(t-)$, and $\zeta(t)$ replaced by $\mathcal{K}_{\ell,j}(f;t)$. Using the from of stochastic intensity of $K_{\ell,j}$ given by Lemma 5.5, the process $\{M_{S,\ell,j}(f;t);t\geq 0\}$ defined as

$$M_{S,\ell,j}(f;t) := \mathcal{K}_{\ell,j}(f;t) - \int_0^t \mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right) h_{S,I(s)}\left(a(s)\right) f((w_{\ell,j}(s))\frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'}Q_{\ell'}(s)}ds \tag{5.30}$$

is a local martingale. Summing (5.30) over $j\in\mathcal{Z}_\ell$, by definition (2.4) of ν_ℓ we obtain

$$\begin{aligned}\mathcal{K}_\ell(f;t) &= \int_0^t \sum_{j\in\mathcal{Z}_\ell} \mathbb{1}\left(j\in\mathcal{Q}_\ell(s)\right) h_{S,I(s)}\left(a(s)\right) f((w_{\ell,j}(s))\frac{p_\ell}{\sum_{\ell'=1}^L p_{\ell'}Q_{\ell'}(t-)}ds + \sum_{j\in\mathcal{Z}_\ell} M_{S,\ell,j}(f;t)\\ &= \int_0^t h_{S,I(s)}\left(a(s)\right)\frac{p_\ell\langle f,\nu_\ell(s)\rangle}{\sum_{\ell'=1}^L p_{\ell'}Q_{\ell'}(s)}ds + M_{S,\ell}(f;t),\end{aligned}$$

where $M_{S,\ell}(f;\cdot) := \sum_{j\in\mathcal{Z}_\ell} M_{S,\ell,j}(f;\cdot)$ is a local martingale.

Moreover, almost surely, for $j,j'\in\mathcal{Z}_\ell$ and $j'\neq j$, $\mathcal{K}_{\ell,j}(f,\cdot)$ and $\mathcal{K}_{\ell,j'}(f,\cdot)$ are bounded pure jump processes with no common jump times. Therefore,

$$[\mathcal{K}_{\ell,j}(f;\cdot)](t) = \int_0^t f^2(w_{\ell,j}(s-))dK^Q_{\ell,j}(s) = \mathcal{K}_{\ell,j}(f^2;t).$$

and $[\mathcal{K}_{\ell,j}(f,\cdot),\mathcal{K}_{\ell,j'}(f,\cdot)]\equiv 0$. Together with the fact that $\mathcal{K}_\ell(f;\cdot) - M_{S,\ell}(f;\cdot)$ is a continuous process with finite variation, this implies

$$[M_{S,\ell}(f;\cdot)] = [\mathcal{K}_\ell(f;\cdot)] = [\sum_{j\in\mathcal{Z}_\ell}\mathcal{K}_{\ell,j}(f;\cdot)] = \sum_{j\in\mathcal{Z}_\ell}[\mathcal{K}_{\ell,j}(f;\cdot)] = \sum_{j\in\mathcal{Z}_\ell}\mathcal{K}_{\ell,j}(f^2;\cdot) = \mathcal{K}_\ell(f^2;\cdot).$$

□

6. Proof of Convergence

This section is devoted to the proof of Theorem 3.1. We Consider a sequence of scaled DROS queues with arrival processes, service times and patience time as described in Section 3.1, and study the asymptotic properties of of the sequence of state variables $\{\bar{\nu}^m\}_{m\in\mathbb{N}}$, by proving tightness and characterizing its subsequential limits. First, we study dynamic evolution of the scaled processes in Section 6.1. Section 6.2 contains a law of large numbers limit for primitive processes. In Section 6.3, we prove tightness for the fluid-scaled sequence of state variables and other auxiliary processes, and finally, subsequential limits are characterized in section 6.4.

Throughout this section, to simplify notation, for $f\in\mathbf{C}_b[0,\infty)$, we use $\|f\|$ to denote $\|f\|_\infty = \sup_{s\in[0,\infty)}|f(s)|$.

6.1. Dynamics of the Scaled Processes

Recall from section 3.1 that we consider a sequence of queueing systems indexed by m, where arrival process and service rates are accelerated. Writing equation (4.14) for the the queueing system indexed by m, dividing both sides of the equation by m, recalling that $\bar{E}_\ell^m = E_\ell^m/m$, and defining the fluid scaled quantities

$$\bar{R}_\ell^m(t) = \frac{1}{m}R_\ell^m(t), \quad \bar{\mathcal{R}}_\ell^m(f;t) = \frac{1}{m}\mathcal{R}_\ell^m(f;t), \quad \bar{K}_\ell^m(t) = \frac{1}{m}K_\ell^m(t), \quad \bar{\mathcal{K}}_\ell^m(f;t) = \frac{1}{m}\mathcal{K}_\ell^m(f;t), \tag{6.1}$$

we can write the equations governing the dynamics of the fluid-scaled processes $\bar{\nu}_\ell^m$ as

$$\langle f, \bar{\nu}_\ell^m(t)\rangle = \langle f, \bar{\nu}_\ell^m(0)\rangle + \int_0^t \langle f', \bar{\nu}_\ell^m(s)\rangle ds + f(0)\int_0^t \mathbb{1}\left(I^m(s-) \neq 0\right) d\bar{E}_\ell^m(s) - \bar{\mathcal{R}}_\ell^m(f;t) - \bar{\mathcal{K}}_\ell^m(f;t). \tag{6.2}$$

In addition, invoking Propositions 5.4 and 5.3 for the queueing system with index m, we can write the martingale decompositions for the fluid-scaled processes $\bar{\mathcal{R}}_\ell^m$ and $\bar{\mathcal{K}}_\ell^m$. Namely, recalling that $h_{R,\ell}^m = h_{R,\ell}$ and $h_{S,\ell}^m(x) = mh_{S,\ell}(mx)$, for every $f \in \mathbf{C}_b^1[0,\infty)$ we have

$$\bar{\mathcal{R}}_\ell^m(f;t) = \int_0^t \langle f h_{R,\ell}, \bar{\nu}_\ell^m(s)\rangle ds + \bar{M}_{R,\ell}^m(f;t), \tag{6.3}$$

where $\{\bar{M}_{R,\ell}^m(f;t); t \geq 0\}$ is a local $\mathcal{F}$-martingale with quadratic variation

$$[\bar{M}_{R,\ell}^m(f;\cdot)]_t = \frac{1}{m^2}\mathcal{R}_\ell^m(f^2;t), \tag{6.4}$$

and

$$\bar{\mathcal{K}}_\ell^m(f;t) = \int_0^t h_{S,I^m(s)}\left(ma^m(s)\right)\frac{p_\ell\langle f, \bar{\nu}_\ell^m(s)\rangle}{\sum_{\ell'=1}^L p_{\ell'}\bar{Q}_{\ell'}^m(s)} ds + \bar{M}_{S,\ell}^m(f;t), \tag{6.5}$$

where $\{\bar{M}_{S,\ell}^m(f;t); t \geq 0\}$ is a local $\mathcal{F}$-martingale with quadratic variation

$$[\bar{M}_{S,\ell}^m(f;\cdot)]_t = \frac{1}{m^2}\mathcal{K}_\ell^m(f^2;t). \tag{6.6}$$

Plugging (6.3) and (6.5) in (6.7), we obtain

$$\begin{aligned}\langle f, \bar{\nu}_\ell^m(t)\rangle =& \langle f, \bar{\nu}_\ell^m(0)\rangle + f(0)\int_0^t \mathbb{1}\left(\bar{X}(s-) > 0\right) d\bar{E}_\ell^m(s) + \int_0^t \langle f' - fh_{R,\ell}, \bar{\nu}_\ell^m(s)\rangle ds \\ &- \int_0^t h_{S,I^m(s)}\left(ma^m(s)\right)\frac{p_\ell\langle f, \bar{\nu}_\ell^m(s)\rangle}{\sum_{\ell'=1}^L p_{\ell'}\bar{Q}_{\ell'}^m(s)} ds + \bar{M}_{R,\ell}^m(t) + \bar{M}_{S,\ell}^m(t).\end{aligned} \tag{6.7}$$

6.2. Convergence of Primitives and Martingale Terms

LEMMA 6.1. *For every $\ell = 1, ..., L$, almost surely, $\bar{E}_\ell^m \to \lambda_\ell \boldsymbol{Id}$ in $\mathbb{D}[0,\infty)$ as $m \to \infty$, where $\boldsymbol{Id}(t) = t$ is the identity function. Also for every $t \geq 0$, $\mathbb{E}\left[\bar{E}_\ell^m(t)\right] \to \lambda_\ell t$ as $m \to \infty$.*

Proof. This first assertion follows from standard law of large numbers results for renewal processes, see e.g. (Chen and Yao 2001, Theorem 5.10). The second assertion follows from elementary renewal theorem; see (Asmussen 2003, Proposition V.1.4). □

COROLLARY 6.1. *It follows from Assumptions 3.1 and Lemma 6.1 that for every $T<\infty$, there exists a constant $C_T<\infty$ such that for $t\in[0,T]$,*

$$\sup_m \mathbb{E}\left[\bar{Q}_\ell^m(0)+\bar{E}_\ell^m(t)\right]<C_T. \tag{6.8}$$

For every $\ell=1,...,L$, define the sequences of (random) measures $\{\bar{\mathcal{L}}_\ell^m\}$ and $\{\bar{\mathcal{H}}_\ell^m\}$ as

$$\bar{\mathcal{L}}_\ell^m:=\frac{1}{m}\sum_{\ell=1}^m \delta_{\kappa_j} v_{\ell,j}, \qquad \bar{\mathcal{H}}_\ell^m:=\frac{1}{m}\sum_{\ell=1}^m \delta_{\kappa_j}\int_0^{v_{\ell,j}} h_{S,\ell}(u)du, \tag{6.9}$$

where recall that $\{\kappa_i; i=1,2,...\}$ is the i.i.d. sequence of random variables that determines the random job selection in the DROS algorithm. For every m, both measures above take values in $\mathcal{M}_{[0,1]}$, the set of finite non-negative measures on $[0,1]$. Note that both measures are constructed directly using primitives of the queueing system. The following elementary law of large number result will be used later in Section 6.4

LEMMA 6.2. *For every $\ell=1,...,L$,*

$$\bar{\mathcal{L}}_\ell^m \Rightarrow \frac{1}{\mu_\ell}\mathcal{L} \tag{6.10}$$

and

$$\bar{\mathcal{H}}_\ell^m \Rightarrow \mathcal{L} \tag{6.11}$$

as $m\to\infty$, where $\mathcal{L}$ is the Lebesgue measure on $[0,1]$.

Proof. The result basically follows from the law of large numbers, independence of κ_js and $v_{\ell,j}$s, and the facts that $\mathbb{E}[v_{\ell,j}]=1/\mu_\ell$ and

$$\mathbb{E}\left[\int_0^{v_{\ell,j}} h_{S,\ell}(u)du\right]=\int_0^\infty\left(\int_0^x h_{S,\ell}(u)du\right)g_{S,\ell}(x)dx=\int_0^\infty\frac{g_{S,\ell}(u)}{\bar{G}_{s,\ell}(u)}\int_u^\infty g_{S,\ell}(x)dxdu=\int_0^\infty g_{S,\ell}(u)du=1.$$

□

The next two lemmas prove that the martingale terms on the right hand of (6.7) vanish as $m\to\infty$.

LEMMA 6.3. *Fix $f\in\mathbf{C}_b^1[0,\infty)$ and $\ell=1,...,L$. Then,*

$$\bar{M}_{R,\ell}^m(f;\cdot)\Rightarrow 0$$

as $m\to\infty$, uniformly on compact sets.

Proof. Fix $T<\infty$. By (6.4), the definition (4.7) of $\mathcal{R}^m_\ell$, the elementary bound (4.12) (for the m^{th} queue), and the bound (6.8) obtained from Assumption 3.1, for $t\in[0,T]$ we have

$$\mathbb{E}\left[[\bar{M}^m_{R,\ell}(f;\cdot)]_t\right]\leq\frac{\|f^2\|}{m^2}\mathbb{E}\left[R^m_\ell(t)\right]\leq\frac{\|f^2\|}{m}\mathbb{E}\left[\bar{Q}^m_\ell(0)+\bar{E}^m_\ell(t)\right]\leq\frac{\|f\|^2}{m}C_T<\infty.$$

Therefore, by (Klebaner 2005, Theorem 7.35), $\{\bar{M}^m_{R,\ell}(f;t);t\geq 0\}$ is a square integrable $\mathcal{F}$-martingale. Moreover, by Doob's inequality, for every $\epsilon>0$,

$$\limsup_{m\to\infty}\mathbb{P}\left(\left|\sup_{0\leq t\leq T}\bar{M}^m_{R,\ell}(t)\right|>\epsilon\right)\leq\limsup_{m\to\infty}\frac{\|f\|^2}{m}C_T=0,$$

which in turn shows that, uniformly on compact sets, $\{\bar{M}^m_{R,\ell}(f;\cdot)\}_m$ converges to zero in probability, and hence in distribution. □

LEMMA 6.4. *Fix $f\in\mathbf{C}^1_b[0,\infty)$ and $\ell=1,...,L$. Then,*

$$\bar{M}^m_{S,\ell}(f;\cdot)\Rightarrow 0$$

as $m\to\infty$, uniformly on compact sets.

Proof. Fix $T<\infty$. By (6.6), the definition (4.8) of $\mathcal{K}^m_\ell$, the elementary bound (4.12) (for the m^{th} queue), and the bound (6.8) obtained from Assumption 3.1, for $t\in[0,T]$ we have

$$\mathbb{E}\left[[\bar{M}^m_{S,\ell}(f;\cdot)]_t\right]\leq\frac{\|f^2\|}{m^2}\mathbb{E}\left[K^m_\ell(t)\right]\leq\frac{\|f^2\|}{m}\mathbb{E}\left[\bar{Q}^m_\ell(0)+\bar{E}^m_\ell(t)\right]\leq\frac{\|f\|^2}{m}C_T<\infty.$$

Therefore, by (Klebaner 2005, Theorem 7.35), $\{\bar{M}^m_{S,\ell}(f;t);t\geq 0\}$ is a square integrable $\mathcal{F}$-martingale. Moreover, by Doob's inequality, for every $\epsilon>0$,

$$\limsup_{m\to\infty}\mathbb{P}\left(\left|\sup_{0\leq t\leq T}\bar{M}^m_{S,\ell}(t)\right|>\epsilon\right)\leq\limsup_{m\to\infty}\frac{\|f\|^2}{m}C_T=0,$$

which in turn shows that, uniformly on compact sets, $\{\bar{M}^m_{S,\ell}(f;\cdot)\}_m$ converges to zero in probability, and hence in distribution. □

6.3. Proof of Tightness

Recall that for a metric space (E,d), a tight sequence of $\mathbb{D}_E[0,\infty)$-valued random elements is called $\mathbb{C}$-tight if every subsequential limit takes values in $\mathbb{C}_E[0,\infty)$, almost surely.

PROPOSITION 6.1. *The random sequence $\{\bar{\nu}^m\}_{m\in\mathbb{N}}$ taking values in $\mathbb{D}_{\mathcal{M}^L}[0,\infty)$ is $\mathbb{C}$-tight.*

The proof is given at the end of this section.

6.3.1. Review of Tightness Criteria We first recall tightness criteria for sequences of random elements in different spaces. Recall from Section 1.1 that $w'(f,\cdot,\cdot)$ denotes the modulus of continuity of a function f in $\mathbb{D}[0,\infty)$.

PROPOSITION 6.2 **(Kurtz's criteria for tightness in $\mathbb{D}[0,\infty)$).** *A sequence of processes $\{Y^m\}_{m\in\mathbb{N}}$ with sample paths in $\mathbb{D}[0,\infty)$ is tight if and only if it satisfies the following:*
K1. For every rational $t\geq 0$,

$$\lim_{r\to\infty}\sup_m \mathbb{P}\left(|Y^m(t)|>r\right)=0; \tag{6.12}$$

K2a. For every $\eta>0$ and $T>0$, there exists $\delta>0$ such that

$$\sup_m \mathbb{P}\left(w'(Y^m,\delta,T)\geq\eta\right)\leq\eta. \tag{6.13}$$

Moreover, $\{Y^m\}_{m\in\mathbb{N}}$ is tight if it satisfies K1 and the following condition:
K2b. For each $T\geq 0$, there exists $\beta>0$ such that

$$\lim_{\delta\to 0}\sup_m \mathbb{E}\left[\sup_{0\leq t\leq T}|Y^m(t+\delta)-Y^m(t)|^\beta\right]=0. \tag{6.14}$$

Proof. The necessity and sufficiency of K1 and K2a follow from (Ethier and Kurtz 1986, Theorem 3.7.2) and the sufficiency of K1 and K2b follows from (Ethier and Kurtz 1986, Theorems 3.7.2 and 3.8.6 and Remark 3.8.7). □

The spaces $(\mathcal{M},d_p)$ is a separable metric space and thus, it is a completely regular topological space with metrizable compacts, and hence, so is $(\mathcal{M}^L,d_p^L)$. Therefore, we can invoke Jakubowski's criteria (Jakubowski 1986, Theorem 4.6) for tightness of $\mathcal{M}^L$-valued processes.

PROPOSITION 6.3 **(Jakubowski's criteria).** *A sequence $\{\pi^m=(\pi_1^m,...,\pi_L^m)\}_{m\in\mathbb{N}}$ of $\mathbb{D}_{\mathcal{M}^L}[0,\infty)$-valued random elements is tight if*
J1. (Compact containment condition) For each $T>0$ and $\eta>0$ there exists a compact set $C_{T,\eta}\subset\mathcal{M}$ such that for every $\ell=1,...,L$,

$$\liminf_m \mathbb{P}\left(\pi_\ell^m(t)\in C_{T,\eta} \text{ for all } t\in[0,T]\right)>1-\eta.$$

J2. For every $\ell=1,...,L$ and $f\in\mathbf{C}_c^1[0,\infty)$, the sequence $\{\langle f,\pi_\ell^m\rangle\}_{m\in\mathbb{N}}$ is tight in $\mathbb{D}[0,\infty)$.

Proof. Condition J1 clearly implies the compact containment condition for the sequence $\{\pi^m=(\pi_1^m,...,\pi_L^m)\}$, that is

$$\liminf_m \mathbb{P}\left(\pi^m(t)\in C_{T,\eta}^L \text{ for all } t\in[0,T]\right)>1-\eta,$$

where the L-fold product $C_{T,\eta}^L=C_{T,\eta}\times....\times C_{T,\eta}$ is a compact subset of $\mathcal{M}^L$. Moreover, define $\mathbb{F}$ to be the family of real-valued continuous functionals on $\mathcal{M}^L$ that have the form

$$F(\mu)=\sum_{n=1}^N\prod_{\ell=1}^L\langle f_\ell^n,\mu_\ell\rangle,\ \ \forall\mu=(\mu_1,...,\mu_L)\in\mathcal{M}^L$$

for some $N \in \mathbb{N}$ and $f_\ell^n \in \mathbf{C}_c^1[0,\infty), n = 1,...,N, \ell = 1,...,L$. Note that the set $\mathbb{F}$ is closed under addition, and separates points in $\mathcal{M}^L$, that is, for every distinct $\mu, \tilde{\mu} \in \mathcal{M}^L$, there exists a function $F \in \mathbb{F}$ such that $F(\mu) \neq F(\tilde{\mu})$. In addition, condition J2 implies that $\{\pi^m\}_{m\in\mathbb{N}}$ is $\mathbb{F}$-tight, that is, for every $F \in \mathbb{F}$, $\{F(\mu^m)\}_{m\in\mathbb{N}}$ is tight in $\mathbb{D}[0,\infty)$. The result then follows from (Jakubowski 1986, Theorem 4.6). □

For $\pi \in \mathbb{D}_{\mathcal{M}^L}[0,\infty)$, the jump size $J(\pi)$ of π is defined as

$$J(\pi) = \sup_{t>0} d_p^L(\pi(t), \pi(t-)).$$

The next result from Billingsley (1968) shows that when jump sizes of a sequences of random process in $\mathbb{D}_{\mathcal{M}^L}[0,\infty)$ vanish, the sequence is in fact $\mathbb{C}$-tight.

PROPOSITION 6.4. *A tight sequence $\{\pi^m = (\pi_1^m, ..., \pi_L^m)\}_{m\in\mathbb{N}}$ of $\mathbb{D}_{\mathcal{M}^L}[0,\infty)$-valued random elements is $\mathbb{C}$-tight if and only if $J(\pi^m) \Rightarrow 0$ as $m \to \infty$.*

Proof. The result follows from applying Theorem 13.4 in Billingsley (1968) to any converging subsequence of $\{\pi^m\}_{m\in\mathbb{N}}$. □

6.3.2. Proof of Compact Containment For any constants $M > 0$, define the subset $\tilde{\mathcal{C}}_M$ of $\mathcal{M}$ as

$$\tilde{\mathcal{C}}_M := \{\mu \in \mathcal{M} : \langle \mathbf{1}, \mu\rangle, \langle \chi, \mu\rangle < M\},$$

where recall that $\mathbf{1}(x) = 1$ and $\chi(x) = x$ for all $x \geq 0$. It follows from compactness conditions (Kallenberg 2017, Theorem 4.2) for subsets of $\mathcal{M}$ that the closure $\mathcal{C}_M$ of $\tilde{\mathcal{C}}_M$ is a compact subset of $\mathcal{M}$.

LEMMA 6.5. *For every $T < \infty$ and $\eta > 0$, there exists $M = M(T,\eta)$ such that for every $\ell = 1,...,L$,*

$$\liminf_m \mathbb{P}\left(\bar{\nu}_\ell^m(t) \in \mathcal{C}_M \text{ for all } t \in [0,T]\right) > 1 - \eta. \tag{6.15}$$

Proof. Recall that $\langle \mathbf{1}, \bar{\nu}_\ell^m\rangle = \bar{Q}_\ell^m(t)$ is the fluid-scaled number of jobs of class ℓ waiting in queue at time t. The mass balance inequality (4.11) and monotonicity of E_ℓ^m implies that for every $t \in [0,T]$, $\bar{Q}_\ell^m(t) \leq \bar{Q}_\ell^m(0) + \bar{E}_\ell^m(T)$, and hence, using the bound (6.8), we have

$$\limsup_m \mathbb{E}\left[\sup_{0\leq t\leq T} \langle \mathbf{1}, \bar{\nu}_\ell^m(t)\rangle\right] \leq \limsup_m \mathbb{E}\left[\bar{E}_\ell(T) + \bar{Q}_\ell(0)\right] \leq C_T.$$

Therefore, choosing $M > 2C_T/\eta$, using Markov's inequality one has

$$\limsup_m \mathbb{P}\left(\langle \mathbf{1}, \bar{\nu}_\ell^m(t)\rangle > M \text{ for some } t \in [0,T]\right) \leq \limsup_m \frac{1}{M}\mathbb{E}\left[\sup_{0\leq t\leq T} \langle \mathbf{1}, \bar{\nu}_\ell^m(t)\right] \leq \frac{\eta}{2} \tag{6.16}$$

Moreover, for every $t \in [0,T]$ and every job (ℓ, j), $w_{\ell,j}(t) \le \varpi_{\ell,j}$ and hence,

$$\langle \chi, \bar{\nu}_\ell^m(t) \rangle = \frac{1}{m} \sum_{j \in \mathcal{Q}_\ell(t)} w_{\ell,j}(t) \le \frac{1}{m} \sum_{j=-Q_\ell(0)+1}^{0} \varpi_{\ell,j} + \frac{1}{m} \sum_{j=1}^{E_\ell(T)} \varpi_{\ell,j}$$

Since the arrival process and the initial queue length are independent of patience times and the sequence of patience times $\{\varpi_{\ell,j}\}_j$ for jobs of class ℓ are iid with mean $1/\gamma_\ell$, by Wald's lemma (Asmussen 2003, Proposition A.10.2) we have

$$\limsup_m \mathbb{E}\left[\sup_{0\le t\le T} \langle \chi, \bar{\nu}_\ell^m(t) \rangle\right] \le \limsup_m \mathbb{E}\left[\bar{Q}_\ell^m(0) + \bar{E}_\ell^m(T)\right] \mathbb{E}\left[\varpi_{\ell,1}\right] \le \frac{C_T}{\gamma_\ell}.$$

Therefore, for $M > 2C_T/(\gamma_\ell \eta)$, by another application of Markov's inequality,

$$\limsup_m \mathbb{P}\left(\langle \chi, \bar{\nu}_\ell^m(t) \rangle > M \text{ for some } t \in [0,T]\right) \le \limsup_m \frac{1}{M} \mathbb{E}\left[\sup_{0\le t\le T} \langle \chi, \bar{\nu}_\ell^m(t) \right] \le \frac{\eta}{2}. \tag{6.17}$$

Equation (6.15) with $M > 2C_t(1+\gamma_\ell)/(\gamma_\ell \eta)$ then follows from (6.16) and (6.17), and the definitions of $\tilde{\mathcal{C}}_M$ and $\mathcal{C}_M$. □

6.3.3. Tightness of $\{\langle f, \bar{\nu}_\ell^m \rangle\}_{m=1}^\infty$ We now show that the sequence $\{\bar{\nu}_\ell^m\}_m$ satisfies condition J2 of Proposition 6.3 for any fixed $\ell \in \{1, 2, ..., L\}$. We first need to prove the tightness of an auxiliary process. For every $m \in \mathbb{N}$, $\ell = 1, ..., L$, and $f \in \mathbf{C}_b^1[0,\infty)$ define

$$\bar{A}_\ell^m(f;t) = \int_0^t h_{S,I^m(u)}\left(m a^m(u)\right) \frac{p_\ell \langle f, \bar{\nu}_\ell^m(u) \rangle}{\sum_{\ell'=1}^L p_{\ell'} \bar{Q}_{\ell'}^m(u)} du. \tag{6.18}$$

LEMMA 6.6. *For every $\ell = 1, ..., L$ and $f \in \mathbf{C}_b^1[0,\infty)$, the sequence $\{\bar{A}_\ell^m(f;\cdot)\}_{m\in\mathbb{N}}$ is tight in $\mathbb{D}[0,\infty)$.*

Proof. Recall by Assumption 2.3 that for every $\ell = 1, ..., L$, the hazard rate function $h_{S,\ell}$ is bounded by a constant H, and note that $|\langle f, \bar{\nu}_\ell^m(u) \rangle| \le \|f\| \bar{Q}_\ell^m(u)$ for all $u \ge 0$. Therefore, for every $0 \le s \le t$,

$$\left|\bar{A}_\ell^m(t) - \bar{A}_\ell^m(s)\right| \le \left| \int_s^t h_{S,I^m(u)}\left(m a^m(u)\right) \frac{p_\ell \langle f, \bar{\nu}_\ell^m(u) \rangle}{\sum_{\ell'=1}^L p_{\ell'} \bar{Q}_{\ell'}^m(u)} du \right| \le \|f\| H |t-s|. \tag{6.19}$$

We now verify Kurtz's criteria of Proposition 6.2 for $\{\bar{A}_\ell^m(f;\cdot)\}$ using the bound above. First, using (6.19) with $s = 0$ implies that for every $t \ge 0$, $\mathbb{P}(\sup_m |\bar{A}_\ell^m(t)| > r) = 0$ for $r \ge \|f\| H t$, and hence condition K1 holds. Moreover, by another application of (6.19) with $t = s + \delta$, we obtain

$$\lim_{\delta\to 0} \limsup_m \mathbb{E}\left[\sup_{0\le s\le T} |\bar{A}_\ell^m(s+\delta) - \bar{A}_\ell^m(s)|\right] \le \lim_{\delta\to 0} \limsup_m \|f\| H \delta = 0,$$

and hence condition K2a holds with $\beta = 1$. □

LEMMA 6.7. *For every $f \in \mathbf{C}_c^1[0,\infty)$, the sequence $\{\langle f, \bar{\nu}_\ell^m \rangle\}_m$ is tight in $\mathbb{D}[0,\infty)$.*

Proof. Recall that by equation (6.7) and definition (6.18) of $\bar{A}^m_\ell(f;\cdot)$, for $t\in[0,T]$,

$$\begin{aligned}\langle f,\bar{\nu}^m_\ell(t)\rangle =&\langle f,\bar{\nu}^m_\ell(0)\rangle + f(0)\int_0^t \mathbb{1}\left(\bar{X}(s-)>0\right)d\bar{E}^m_\ell(s)+\int_0^t\langle f'-fh_{R,\ell},\bar{\nu}^m_\ell(s)\rangle ds\\ &-\bar{A}^m_\ell(f;t)+\bar{M}^m_{R,\ell}(t)+\bar{M}^m_{S,\ell}(t).\end{aligned}\tag{6.20}$$

We show tightness for $\{\langle f,\bar{\nu}^m_\ell\rangle\}_m$ by verifying Kurtz's criteria of Proposition 6.2 for each term on the right-hand side of (6.20) separately.

For the sequence corresponding to the first term on the right-hand side of (6.20), note that by Markov's inequality, for every $m\in\mathbb{N}$ and $r>0$,

$$\sup_m \mathbb{P}\left(|\langle f,\bar{\nu}^m_\ell(0)|>r\right)\le \sup_m \mathbb{P}\left(\|f\|Q^m_\ell(0)>r\right)\le \frac{\|f\|}{r}\sup_m\mathbb{E}\left[Q^m_\ell(0)\right].$$

Since $\mathbb{E}\left[Q^m_\ell(0)\right]<0$ by Assumption 3.1, the sequence $\{\langle f,\bar{\nu}^m_\ell(0)\rangle\}_{m\in\mathbb{N}}$ satisfies condition K1. Condition K2a trivially holds for $\{\langle f,\bar{\nu}^m_\ell(0)\rangle\}_{m\in\mathbb{N}}$.

For the sequence corresponding to the first term on the right-hand side of (6.20), note that the bound

$$\int_s^t \mathbb{1}\left(I^m(u-)\neq 0\right)d\bar{E}^m_\ell(u)\le \bar{E}^m(t)-\bar{E}^m(s),\qquad \forall s,t\in[0,T],\tag{6.21}$$

implies

$$w'(\int_0^\cdot \mathbb{1}\left(I^m(s-)\neq 0\right)d\bar{E}^m_\ell(s),\delta,T)\le w'(\bar{E}^m_\ell,\delta,T).\tag{6.22}$$

Since the sequence $\{\bar{E}^m_\ell\}_m$ converges in $\mathbb{D}[0,\infty)$ to the deterministic limit $\lambda_\ell\mathbf{Id}$ almost surely (and hence in distribution) by Lemma 6.1, it is a tight sequence and thus satisfies conditions K1 and K2a of Proposition 6.2. The validity of conditions K1 and K2a for the sequence $\{\int_0^\cdot \mathbb{1}\left(I^m(s-)\neq 0\right)d\bar{E}^m_\ell(s)\}_m$ follow from the bounds (6.21) (with $s=0$) and (6.22), respectively, and the corresponding conditions for $\{\bar{E}^m_\ell\}_m$.

For the sequence corresponding to the first term on the right-hand side of (6.20), using the bound (4.11) and monotonicity of E^m_ℓ, we have $\bar{Q}^m_\ell(s)\le\bar{Q}^m_\ell(0)+\bar{E}^m_\ell(T)$ for all $s\in[0,T]$. Therefore, recalling that $\|h_{R,\ell}\|\le H$ by Assumption 2.2 and using the bound (6.8),

$$\begin{aligned}\sup_m\mathbb{E}\left[\sup_{t\in[0,T-\delta]}\left|\int_t^{t+\delta}\langle f'-fh_{R,\ell},\bar{\nu}^m_\ell(s)\rangle ds\right|\right]\le&(\|f'\|+\|f\|\|h_{R,\ell}\|)\sup_m\mathbb{E}\left[\sup_{t\in[0,T-\delta]}\int_t^{t+\delta}\bar{Q}^m_\ell(s)ds\right]\\ \le&\,\delta(\|f'\|+H\|f\|)\sup_m\mathbb{E}\left[\bar{Q}^m_\ell(0)+\bar{E}^m_\ell(T)\right]\\ \le&\,\delta(\|f'\|+H\|f\|)C_T.\end{aligned}\tag{6.23}$$

The sequence $\{\int_0^\cdot\langle f'-fh_{R,\ell},\bar{\nu}^m_\ell(s)\rangle ds\}_m$ therefore satisfies conditions K1 and K2b (with $\beta=1$).

Finally, tightness for sequence $\{\bar{A}^m_\ell(f;\cdot)\}_{m\in\mathbb{N}}$, $\{\bar{M}^m_{R,\ell}(f;\cdot)\}_{m\in\mathbb{N}}$, and $\{\bar{M}^m_{S,\ell}(f;\cdot)\}_m$ is proved in Lemmas 6.6, 6.3, and 6.4, respectively. This completes the proof. □

Proof of Proposition 6.1. Tightness of $\{\bar{\nu}^m = (\bar{\nu}^m_\ell; \ell = 1, ..., L)\}_{m\in\mathbb{N}}$ in $\mathcal{D}_{\mathcal{M}^L}[0,\infty)$ follows from Lemmas 6.5 and 6.7, and the Jakobovski's criteria of Proposition 6.3. To prove the $\mathcal{C}$-tightness, we need to show that the jump sizes vanish.

For every $m \in \mathbb{N}$, define $\tilde{\Omega}^m$ to be the set of realizations ω on which in queueing system with index m, all arrivals, renegings, and departures of all jobs of all classes occur on distinct times, and for every finite time t, the total number of arrivals, renegings, and departures occurred during $[0,t]$ is finite, which has full measure (see Remark 4.2). On $\tilde{\Omega}^m$, and hence almost surely, for every $t > 0$, the sets $\mathcal{Q}^m_\ell(t)$ and $\mathcal{Q}^m_\ell(t-)$ of job indices that are waiting in queue at time t and right before time t, respectively, differ at most by one job index. Therefore, since for every $j \in \mathbb{Z}$ and $\ell = 1, ..., L$, $t \mapsto w_{j,\ell}(t)$ is continuous by definition, $\nu^m_\ell(t)$ and $\nu^m_\ell(t-)$ at most differ by one delta mass, which implies $d_{TV}(\bar{\nu}^m_\ell(t), \bar{\nu}^m_\ell(t-)) \le 1/m$, where d_{TV} is the total variation metric. Finally, since the Prohorov metric is bounded by total variation metric (see (Huber 1981, Equation (2.24) of page 36)), we conclude

$$J(\bar{\nu}^m) = \sup_{t>0} d^L_p(\bar{\nu}^m(t), \bar{\nu}^m(t-)) = \sup_{t>0} \sup_{\ell=1,...,L} d_p(\bar{\nu}^m_\ell(t), \bar{\nu}^m_\ell(t-)) \le \frac{1}{m}, \tag{6.24}$$

and hence $J(\bar{\nu}^m) \Rightarrow 0$, as $m \to \infty$. $\mathcal{C}$-tightness of the sequence $\{\bar{\nu}^m\}_{m\in\mathbb{N}}$ then follows from Proposition 6.4. □

6.4. Characterization of Subsequential Limits

By Proposition 6.1, the sequence $\{\bar{\nu}^m = (\bar{\nu}^m_\ell; \ell = 1, ..., L)\}_{m\in\mathbb{N}}$ is tight in $\mathcal{D}_{\mathcal{M}^L}[0,\infty)$, and therefore every subsequence $\{\bar{\nu}^m\}_{m\in\tilde{\mathbf{m}}}$ has a further subsequence $\{\bar{\nu}^m\}_{m\in\hat{\mathbf{m}}}$ such that $\{\bar{\nu}^m\}$ converges to a limit $\bar{\nu} = (\bar{\nu}_\ell; \ell = 1, ..., L)$ along $\hat{\mathbf{m}}$. Our goal is to find a characterization for the subsequential limit $\bar{\nu}$.

Fix $f \in \mathbf{C}^1_b[0,\infty)$. Combining Assumption 3.1 and the results of Proposition 6.1 and Lemmas 6.1, 6.6, 6.3, 6.4, and 6.2, we conclude that the sequence $\{\bar{Y}^m\}_{m\in\mathbb{N}}$ defined as

$$\bar{Y}^m := (\bar{\nu}^m_\ell(0), \bar{\nu}^m_\ell, \bar{E}^m_\ell, \bar{A}^m_\ell(f,\cdot), \bar{M}^m_{R,\ell}(f;\cdot), \bar{M}^m_{S,\ell}(f;\cdot), \bar{K}^m, \bar{\mathcal{L}}^m_\ell, \bar{\mathcal{H}}^m_\ell; \ell = 1, ..., L),$$

is tight in the space $\mathcal{Y}$ defined as

$$\mathcal{Y} := \mathcal{M}^L \times \mathcal{D}_{\mathcal{M}^L}[0,\infty) \times \mathbb{D}[0,\infty)^{4L+1} \times \mathcal{M}^{2L},$$

and hence the subsequence $\{\bar{Y}^m\}_{m\in\hat{\mathbf{m}}}$ has a further subsequence $\{\bar{Y}^m\}_{m\in\mathring{\mathbf{m}}}$ that converges in distribution to

$$\bar{Y} = (\bar{\nu}_\ell(0), \bar{\nu}_\ell, \lambda_\ell \mathbf{Id}, \bar{A}_\ell(f,\cdot), \mathbf{0}, \mathbf{0}, \bar{K}, \frac{1}{\mu_\ell}\bar{\mathcal{L}}, \bar{\mathcal{L}}; \ell = 1, ..., L),$$

where $\mathbf{0}$ is the zero function in $\mathbb{D}[0,\infty)$.

As a first step to characterize $\bar{\nu}$, the next lemma shows the mapping $t \mapsto \bar{\nu}(t)$ is continuous.

LEMMA 6.8. *Almost surely, the mapping $t \mapsto \bar{\nu}(t)$ is continuous.*

Proof. Fix $f \in \mathbf{C}_b^1[0,\infty)$ and $T < \infty$. Since at most one event (arrival, reneging, or service entry) happens at each, the sets $\mathcal{Q}_\ell^m(t)$ of jobs of class ℓ waiting in queue only changes by at most one element from $t-$ to t, and hence,

$$|\langle f, \bar{\nu}_\ell^m(t)\rangle - \langle f, \bar{\nu}_\ell^m(t-)\rangle| \le \frac{1}{m}\|f\|,$$

for all $t \in [0,T]$, and hence $J_T(\langle f, \bar{\nu}_\ell^m\rangle) \le \|f\|/m$, where $J_T(x)$ is the maximum jump size of a function $x \in \mathbb{D}[0,\infty)$ on $[0,T]$. Therefore, by (Billingsley 1968, Theorem 13.4), the mapping $t \mapsto \langle f, \bar{\nu}_\ell(t)\rangle$ is continuous almost surely, which implies the result. □

By Skorokhod representation theorem, there exists a sequence (possibly on a different probability space) whose elements have the same distributions as $\bar{Y}^m, m \in \mathring{\mathbf{m}}$, that converges almost surely to a limit with the same distribution as $\bar{Y}$. Since we are only interested in the distribution of the limit process, we denote this new sequence and its limit by $\{\bar{Y}^m\}_{m\in\mathbb{N}}$ and $\bar{Y}$ again. Finally, since the limit process $\bar{Y}$ is continuous almost surely by Lemma 6.8, the convergence in the Skorokhod topology implies uniform convergence on compact sets.

In summary, for the rest of the article we can assume that there exists $\Omega_0 \subset \Omega$ with $P(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$,

$$\bar{Y}^m \to \bar{Y} \tag{6.25}$$

as $m \to \infty$, uniformly on compact sets.

6.4.1. Lipschitz Continuity of the Limit Process For the rest of this section we fix $f \in \mathbf{C}_b^1[0,\infty)$ and $\omega \in \Omega_0$.

LEMMA 6.9. *Fix $T < \infty$. There exist $\Lambda_f > 0$ and sequences $\{c_0^m\} \downarrow 0$ and $\{c_L^m\} \downarrow 0$ such that*

a. *for every $\ell = 1, ..., L$*

$$|\langle f, \bar{\nu}_\ell^m(0)\rangle - \langle f, \bar{\nu}_\ell(0)\rangle| \le c_0^m, \tag{6.26}$$

b. *for every $\ell = 1, ..., L$ and $s, t \in [0,T]$,*

$$|\langle f, \bar{\nu}_\ell^m(s)\rangle - \langle f, \bar{\nu}_\ell^m(t)\rangle| \le \Lambda_f |s-t| + c_L^m. \tag{6.27}$$

Proof. The bound (6.26) follows from convergence of the first component in (6.25). To see (6.27), recall that for every $0 \le s \le t$, by (6.7),

$$\begin{aligned}\langle f, \bar{\nu}_\ell^m(t)\rangle - \langle f, \bar{\nu}_\ell^m(s)\rangle = &\int_s^t \langle f' - f h_{R,\ell}, \bar{\nu}_\ell^m(u)\rangle du + f(0)\int_s^t \mathbb{1}\left(\bar{X}(u-) > 0\right) d\bar{E}_\ell^m(u) \\ &- \int_s^t h_{S,I^m(u)}\left(m a^m(u)\right) \frac{p_\ell \langle f, \bar{\nu}_\ell^m(u)\rangle}{\sum_{\ell'=1}^L p_{\ell'} \bar{Q}_{\ell'}^m(u)} du \\ &+ \bar{M}_{R,\ell}^m(f;t) - \bar{M}_{R,\ell}^m(f;s) + \bar{M}_{S,\ell}^m(f;t) - \bar{M}_{S,\ell}^m(f;s). \end{aligned} \tag{6.28}$$

For the first term on the right-hand side above, recalling from Assumption 2.2 that the hazard rate $h_{R,\ell}$ is bounded by constant H,

$$\langle f' - fh_{R,\ell}, \bar{\nu}_\ell^m(u)\rangle du \le (\|f'\| + \|f\|H) \int_s^t \bar{Q}_\ell^m(u)du,$$

By (6.25), $\bar{Q}_\ell^m = \langle \mathbf{1}, \bar{\nu}_\ell^m\rangle$ converges uniformly on $[s,t]$ to the continuous function $\bar{Q} = \langle \mathbf{1}, \bar{\nu}_\ell\rangle$, and therefore, there exists a constant $C_T = C_T(\omega)$ such that $\bar{Q}_\ell^m(u) \le C_T$ for every large enough m. Hence,

$$\langle f' - fh_{R,\ell}, \bar{\nu}_\ell^m(u)\rangle du \le C_T(\|f'\| + \|f\|H)|t-s|. \tag{6.29}$$

For the second term, since $\bar{E}_\ell^m$ converges to $\lambda_\ell \mathbf{Id}$ uniformly on $[s,t]$, there exists a sequence $\{c_1^m\} \downarrow 0$ such that $|\bar{E}_\ell^m(t) - \lambda_\ell t| \vee |\bar{E}_\ell^m(s) - \lambda_\ell s| < c_1^m$. Therefore,

$$\begin{aligned}\int_s^t \mathbb{1}\left(\bar{X}(u-) > 0\right) d\bar{E}_\ell^m(u) &\le \bar{E}_\ell^m(t) - \bar{E}_\ell^m(s) \\ &\le |\bar{E}_\ell^m(t) - \lambda_\ell t| + |\bar{E}_\ell^m(t) - \lambda_\ell t| + \lambda_\ell |t-s| \\ &\le 2c_1^m + \lambda_\ell |t-s| \end{aligned} \tag{6.30}$$

For the third term, recalling from Assumption 2.3 [4] that the hazard rate function $h_{S,\ell}$ is bounded by H and since the fraction in the integrand is bounded by $\|f\|$,

$$\int_s^t h_{S,I^m(u)}\left(ma^m(u)\right) \frac{p_\ell \langle f, \bar{\nu}_\ell^m(u)\rangle}{\sum_{\ell'=1}^L p_{\ell'} \bar{Q}_{\ell'}^m(u)} du \le H\|f\||t-s| \tag{6.31}$$

Finally, by (6.25), the sequences of martingales $\{\bar{M}_{R,\ell}(f;\cdot)\}$ and $\{\bar{M}_{S,\ell}(f;\cdot)\}$ converge to zero uniformly on compact sets, and hence there exists a sequence $\{c_2^m\} \downarrow 0$ such that

$$|\bar{M}_{R,\ell}^m(f;t) - \bar{M}_{R,\ell}^m(f;s) + \bar{M}_{S,\ell}^m(f;t) - \bar{M}_{S,\ell}^m(f;s)| \le c_2^m. \tag{6.32}$$

The bound (6.27) follows from (6.28)-(6.32) with

$$\Lambda_f = \max_{\ell=1,\dots,L} \{(1+C_T)(\|f'\| + 2H\|f\|) + \lambda_\ell\},$$

and $c_L^m = 2c_1^m + c_2^m$. □

Taking the limit as $m \to \infty$ in (6.27), we conclude that the limit points are Lipschitz functions.

COROLLARY 6.2. *For every $\omega \in \Omega_0$, $f \in \mathbf{C}_b^1[0,\infty)$ and $T < \infty$, $\langle f, \bar{\nu}_\ell\rangle$ is locally Lipschitz continuous on $[0,T]$ with a constant Λ_f, that is, for every $s,t \in [0,T]$,*

$$|\langle f, \bar{\nu}_\ell(s)\rangle - \langle f, \bar{\nu}_\ell(t)\rangle| \le \Lambda_f |s-t|. \tag{6.33}$$

[4] Can remove bounded hazard rate assumption on service time distribution.

6.4.2. Equations Governing the Fluid Limit Since Lipschitz continuity implies absolute continuity, we have thus proven that with probability one, trajectories of $\langle f, \bar{\nu}_\ell \rangle$ are absolutely continuous and hence have derivative almost everywhere. We show that its derivative is given by equation (3.2). For the rest of this section, we fix a realization $\omega \in \Omega_0$ and $f \in \mathbf{C}_b^1[0,\infty)$.

Define

$$\bar{Q}(t) = \sum_{\ell=1}^{L} \bar{Q}_\ell(t),$$

to be the fluid limit of the total number of jobs in queue (of all classes). The derivative of $\langle f, \bar{\nu}_\ell \rangle(t)$ depends on whether the queue is empty at time t.

LEMMA 6.10. *Let $t \geq 0$ be such that $\bar{Q}(t) = 0$ and $\langle f, \bar{\nu}_\ell \rangle$ is differentiable at t. Then,*

$$\frac{d}{dt} \langle f, \bar{\nu}_\ell(t) \rangle = 0.$$

Proof. Note that since $\bar{Q}(t) = 0$,

$$\langle f, \bar{\nu}_\ell(t) \rangle \leq \|f\| \langle \mathbf{1}, \bar{\nu}_\ell(t) \rangle = \|f\| \bar{Q}_\ell(t) \leq \|f\| \bar{Q}(t) = 0.$$

Since the function $u \mapsto \langle f, \bar{\nu}_\ell(u) \rangle$ is non-negative and differentiable at t, its derivative at t needs to be zero. □

LEMMA 6.11. *For every $\ell = 1, ..., L$, $t \in [0,T]$ and $\epsilon > 0$, there exists $M < \infty$ such that for all $m \geq M$,*

$$|\langle f, \bar{\nu}_\ell^m(u) \rangle - \langle f, \bar{\nu}_\ell(t) \rangle| \leq 2\Lambda_f \epsilon, \qquad \forall u \in [t, t+\epsilon]. \tag{6.34}$$

and in particular,

$$|\bar{Q}_\ell^m(u) - \bar{Q}_\ell(t)| \leq 2\Lambda_{\mathbf{1}} \epsilon, \qquad \forall u \in [t, t+\epsilon]. \tag{6.35}$$

In addition, if $\bar{Q}(t) > 0$ for some $t \geq 0$, then there exists $\epsilon > 0$ and $M < \infty$ such that $\bar{Q}^m(u) > 0$ for every $u \in [t, t+\epsilon]$ and $m > M$.

Proof. Fix $\epsilon > 0$. For every $u \in [t, t+\epsilon]$ and $m \in \mathbb{N}$,

$$|\langle f, \bar{\nu}_\ell^m(u) \rangle - \langle f, \bar{\nu}_\ell(t) \rangle| \leq |\langle f, \bar{\nu}_\ell^m(u) \rangle - \langle f, \bar{\nu}_\ell(u) \rangle| + |\langle f, \bar{\nu}_\ell(u) \rangle - \langle f, \bar{\nu}_\ell(t) \rangle|.$$

Since $\langle f, \bar{\nu}_\ell \rangle$ is Lipschitz continuous with constant Λ_f (Corollary 6.2), $|\langle f, \bar{\nu}_\ell(u) \rangle - \langle f, \bar{\nu}_\ell(t) \rangle| \leq \Lambda_f \epsilon$ for every $u \in [t, t+\epsilon]$. Moreover, by (6.25), $\langle f, \bar{\nu}_\ell^m \rangle$ converges to $\langle f, \bar{\nu}_\ell \rangle$ uniformly on $[t, t+\epsilon]$, and hence for large enough $m > M$, $|\langle f, \bar{\nu}_\ell(u) \rangle - \langle f, \bar{\nu}_\ell(t) \rangle| \leq \Lambda_f \epsilon$, and therefore (6.34) follows. Since $\bar{Q}_\ell = \langle \mathbf{1}, \bar{\nu}_\ell \rangle$, (6.35) follows on plugging $f = \mathbf{1}$ in (6.34). Finally, for $u \in [t, t+\epsilon]$,

$$|\bar{Q}^m(u) - \bar{Q}(t)| \leq \sum_{\ell-1}^{L} |\bar{Q}_\ell^m(u) - \bar{Q}_\ell(t)| \leq 2L\Lambda_{\mathbf{1}} \epsilon.$$

The last claim then follows from the display above with $\epsilon = \bar{Q}(t)/4L\Lambda_{\mathbf{1}} > 0$. □

LEMMA 6.12. *For every $t \geq 0$ with $\bar{Q}(t) > 0$, $\bar{K}$ is differentiable with derivative*

$$\frac{d}{dt}\bar{K}(t) = \frac{\sum_{\ell=1}^{L} p_\ell \bar{Q}_\ell(t)}{\sum_{\ell=1}^{L} \frac{p_\ell}{\mu_\ell} \bar{Q}_\ell(t)}. \tag{6.36}$$

Proof. By definition, the derivative is given by

$$\frac{d}{dt}\bar{K}(t) = \lim_{\epsilon \to 0} \lim_{m \to 0} \frac{\bar{K}^m(t+\epsilon) - \bar{K}^m(t)}{\epsilon}.$$

We compute the limit as $\epsilon \downarrow 0$; the computation of the left limit is exactly identical. Fix $\epsilon > 0$ as small as required by Lemma 6.11. Therefore, by that lemma, $\bar{Q}^m(u) > 0$ for all large enough m and $u \in [t, t+\epsilon]$, and hence the server is never idle during $[t, t+\epsilon]$. Therefore, service times of jobs that have been served during $[t, t+\epsilon]$ should approximately sum up to ϵ. To make this precise, recall that K^m is the total service entry process, and let $\{\eta_j^m;\}_{j\in\mathbb{N}}$ be the epoch times of K^m. Recall from Section 2.1 that the j^{th} job to enter service (at time η_j^m) will be of class ℓ if $\kappa_j \in [a_\ell^m(\eta_j^m), b_\ell^m(\eta_j^m))$ where a_ℓ^m and b_ℓ^m are defined in 2.3, we have

$$\sum_{j=K^m(t)+1}^{K^m(t+\epsilon)-1} \sum_{\ell=1}^{L} \mathbb{1}\left(\kappa_j \in [a_\ell^m(\eta_j^m), b_\ell^m(\eta_j^m))\right) \frac{v_{\ell,j}}{m} \leq \epsilon \leq \sum_{j=K^m(t)}^{K^m(t+\epsilon)} \sum_{\ell=1}^{L} \mathbb{1}\left(\kappa_j \in [a_\ell^m(\eta_j^m), b_\ell^m(\eta_j^m))\right) \frac{v_{\ell,j}}{m} \tag{6.37}$$

Since $\bar{K}^m$ converges to $\bar{K}$ uniformly on $[t, t+\epsilon]$ by (6.25), there exists an M such that for all $m \geq M$ and $u \in [t, t+\epsilon]$, $\bar{K}^m(u) \in \bar{K}(u) \pm \epsilon^2$, and hence

$$m\bar{K}(t) - m\epsilon^2 \leq K^m(u) \leq m\bar{K}(t+\epsilon) + m\epsilon^2. \tag{6.38}$$

Moreover, by (6.35), $\bar{Q}_\ell(t) - 2\Lambda_{\mathbf{1}}\epsilon \leq \bar{Q}_\ell^m(u) \leq \bar{Q}_\ell(t) + 2\Lambda_{\mathbf{1}}\epsilon$, and therefore,

$$[\, \underline{a}_\ell(\epsilon)\,,\, \underline{b}_\ell(\epsilon)\,) \subset [\, a_\ell^m(\eta_j^m)\,,\, b_\ell^m(\eta_j^m)\,) \subset [\, \bar{a}_\ell(\epsilon)\,,\, \bar{b}_\ell(\epsilon)\,), \tag{6.39}$$

with

$$\bar{a}_\ell(\epsilon) = \frac{\sum_{\ell'=1}^{\ell-1} p_{\ell'}\bar{Q}_{\ell'}(t) - 2L\Lambda_{\mathbf{1}}\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}\bar{Q}_{\ell'}(t) + 2L\Lambda_{\mathbf{1}}\epsilon}, \qquad \bar{b}_\ell(\epsilon) = \frac{\sum_{\ell'=1}^{\ell} p_{\ell'}\bar{Q}_{\ell'}(t) + 2L\Lambda_{\mathbf{1}}\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}\bar{Q}_{\ell'}(t) - 2L\Lambda_{\mathbf{1}}\epsilon}.$$
$$\underline{a}_\ell(\epsilon) = \frac{\sum_{\ell'=1}^{\ell-1} p_{\ell'}\bar{Q}_{\ell'}(t) + 2L\Lambda_{\mathbf{1}}\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}\bar{Q}_{\ell'}(t) - 2L\Lambda_{\mathbf{1}}\epsilon}, \qquad \underline{b}_\ell(\epsilon) = \frac{\sum_{\ell'=1}^{\ell} p_{\ell'}\bar{Q}_{\ell'}(t) - L2\Lambda_{\mathbf{1}}\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}\bar{Q}_{\ell'}(t) + 2L\Lambda_{\mathbf{1}}\epsilon},$$

when note that

$$\lim_{\epsilon\to 0} \bar{b}_\ell(\epsilon) - \bar{a}_\ell(\epsilon) = \lim_{\epsilon\to 0} \underline{b}_\ell(\epsilon) - \underline{a}_\ell(\epsilon) = \frac{p_\ell \bar{Q}_\ell(t)}{\sum_{\ell'=1}^{L} p_{\ell'}\bar{Q}_{\ell'}(t)}. \tag{6.40}$$

Therefore, plugging in (6.38) and (6.39) in (6.37) we have

$$\sum_{j=\lceil m\bar{K}(t)+m\epsilon^2\rceil+1}^{\lfloor m\bar{K}(t+\epsilon)+m\epsilon^2\rfloor-1} \sum_{\ell=1}^{L} \mathbb{1}\left(\kappa_j \in [\underline{a}_\ell(\epsilon), \underline{b}_\ell(\epsilon))\right) \frac{v_{\ell,j}}{m} \leq \epsilon \leq \sum_{j=\lfloor m\bar{K}(t)+m\epsilon^2\rfloor}^{\lceil m\bar{K}(t+\epsilon)+m\epsilon^2\rceil} \sum_{\ell=1}^{L} \mathbb{1}\left(\kappa_j \in [\bar{a}_\ell(\epsilon), \bar{b}_\ell(\epsilon))\right) \frac{v_{\ell,j}}{m}$$

Recall from the definition (6.9) of $\bar{\mathcal{L}}_\ell^m$ that for $\ell = 1, ..., L$ and $I = [\underline{a}_\ell(\epsilon), \underline{b}_\ell(\epsilon))$ or $I = [\bar{a}_\ell(\epsilon), \bar{b}_\ell(\epsilon))$,

$$\sum_{j=1}^{m} \mathbb{1}\left(\kappa_j \in I\right) \frac{v_{\ell,j}}{m} = \bar{\mathcal{L}}_\ell^m\left(I\right).$$

Therefore, taking $m \to \infty$ using (6.10) of Lemma 6.2, we obtain

$$\frac{\bar{K}(t+\epsilon) - \bar{K}(t)}{\epsilon} \sum_{\ell=1}^{L} \frac{\underline{b}_\ell(\epsilon) - \underline{a}_\ell(\epsilon)}{\mu_\ell} \le 1 \le \frac{\bar{K}(t+\epsilon) - \bar{K}(t)}{\epsilon} \sum_{\ell=1}^{L} \frac{\bar{b}_\ell(\epsilon) - \bar{a}_\ell(\epsilon)}{\mu_\ell} \tag{6.41}$$

Finally, taking $\epsilon \to 0$ and using (6.40), we have

$$\frac{d}{dt}\bar{K}(t) \sum_{\ell=1}^{L} \frac{1}{\mu_\ell} \frac{p_\ell \bar{Q}_\ell(t)}{\sum_{\ell=1}^{L} p_\ell \bar{Q}_\ell(t)} \le 1 \le \frac{d}{dt}\bar{K}(t) \sum_{\ell=1}^{L} \frac{1}{\mu_\ell} \frac{p_\ell \bar{Q}_\ell(t)}{\sum_{\ell=1}^{L} p_\ell \bar{Q}_\ell(t)} \tag{6.42}$$

which proves (6.36). □

LEMMA 6.13. *For every $\ell = 1, ..., L$ and $t > 0$ with $\bar{Q}(t) > 0$, $\bar{A}_\ell(f;\cdot)$ is differentiable at t with derivative*

$$\frac{d}{dt}\bar{A}_\ell(f;t) = \frac{p_\ell \langle f, \bar{\nu}_\ell(t)\rangle}{\sum_{\ell'=1}^{L} \frac{p_{\ell'}}{\mu_{\ell'}} \bar{Q}_{\ell'}(t)} \tag{6.43}$$

Proof. Fix $\epsilon > 0$ as small as required by Lemma 6.11. Plugging in the bounds (6.34) and (6.35) in the definition (6.18) of $A_\ell^m(f;\cdot)$, we obtain

$$\frac{p_\ell \langle f, \bar{\nu}_\ell(t)\rangle - p_\ell c\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}(\bar{Q}_{\ell'}(t) + c\epsilon)} \int_t^{t+\epsilon} h_{S,I^m(u)}\left(ma^m(u)\right) du \le \bar{A}^m(t+\epsilon) - \bar{A}^m(t) \tag{6.44}$$

$$\le \frac{p_\ell \langle f, \bar{\nu}_\ell(t)\rangle + p_\ell c\epsilon}{\sum_{\ell'=1}^{L} p_{\ell'}(\bar{Q}_{\ell'}(t) - c\epsilon)} \int_t^{t+\epsilon} h_{S,I^m(u)}\left(ma^m(u)\right) du \tag{6.45}$$

To find bounds for the integral in the upper and lower bounds above, note that by Lemma 6.11, $\bar{Q}^m(u) > 0$ for all large enough $m > 0$ and $u \in [t, t+\epsilon]$, and hence the server is never idle during $[t, t+\epsilon]$. Recalling that $\{\eta_j^m\}$ are the epoch times of the total service entry process K^m and using the bounds $\eta_{K^m(t)} \le t \le \eta_{K^m(t)+1}$, we have the following lower bound

$$\begin{aligned}
\int_t^{t+\epsilon} h_{S,I^m(u)}\left(ma^m(u)\right) du &\ge \int_{\eta_{K^m(t)+1}}^{\eta_{K^m(t+\epsilon)}} h_{S,I^m(u)}\left(ma^m(u)\right) du \\
&\ge \sum_{K^m(t)+1}^{K^m(t+\epsilon)-1} \int_{\eta_j}^{\eta_{j+1}} h_{S,I^m(u)}\left(ma^m(u)\right) du \\
&\ge \sum_{K^m(t)+1}^{K^m(t+\epsilon)-1} \sum_{\ell=1}^{L} \mathbb{1}\left(\kappa_j \in [a_\ell^m(\eta_j), b_\ell^m(\eta_j))\right) \int_{\eta_j}^{\eta_{j+1}} h_{S,\ell}\left(ma^m(u)\right) du
\end{aligned} \tag{6.46}$$

Since the server busy during $[t, t+\epsilon]$, the age $a^m(t)$ of the job in service is zero at each η_j, and grows linearly with unit speed during $[\eta_j, \eta_{j+1})$, and when the job in service during that interval

is of class ℓ, we have $\eta_{j+1}-\eta_j = v_{\ell,j}/m$. Therefore, on $\{\kappa_j \in [a_\ell^m(\eta_j), b_\ell^m(\eta_j))\}$, the last integral in the right-hand side of (6.46) can be written as

$$\int_{\eta_j}^{\eta_{j+1}} h_{S,\ell}\left(ma^m(u)\right) du = \int_0^{\eta_{j+1}-\eta_j} h_{S,\ell}\left(mu\right) du = \frac{1}{m}\int_0^{m(\eta_{j+1}-\eta_j)} h_{S,\ell}\left(u\right) du = \frac{1}{m}\int_0^{v_{\ell,j}} h_{S,\ell}\left(u\right) du \tag{6.47}$$

Plugging (6.47) in (6.46), using the bounds (6.38) and (6.39), and changing the order of summations, we obtain

$$\int_t^{t+\epsilon} h_{S,I^m(u)}\left(ma^m(u)\right) du \geq \frac{1}{m}\sum_{\ell=1}^{L}\sum_{j=\lceil m\bar{K}(t)+m\epsilon^2\rceil+1}^{\lfloor m\bar{K}(t+\epsilon)+m\epsilon^2\rfloor-1} \mathbb{1}\left(\kappa_j \in [\underline{\mathrm{a}}_\ell(\epsilon), \underline{\mathrm{b}}_\ell(\epsilon))\right) \int_0^{v_{\ell,j}} h_{S,\ell}\left(u\right) du, \tag{6.48}$$

We can obtain the upper bound bellow in a similar fashion:

$$\int_t^{t+\epsilon} h_{S,I^m(u)}\left(ma^m(u)\right) du \leq \frac{1}{m}\sum_{\ell=1}^{L}\sum_{j=\lfloor m\bar{K}(t)+m\epsilon^2\rfloor}^{\lceil m\bar{K}(t+\epsilon)+m\epsilon^2\rceil} \mathbb{1}\left(\kappa_j \in [\bar{a}_\ell(\epsilon), \bar{b}_\ell(\epsilon)\right) \int_0^{v_{\ell,j}} h_{S,\ell}\left(u\right) du \tag{6.49}$$

Taking the limit $m\to\infty$ from both sides of (6.48) and (6.4.2) and recalling from (6.25) that $\bar{\nu}_\ell^m$ converges to $\bar{\nu}_\ell$, $\bar{K}^m$ converges to $\bar{K}$, $\bar{A}_\ell^m(f;\cdot)$ converges to $\bar{A}_\ell(f;\cdot)$ and $\bar{\mathcal{H}}$ converges to the Lebesgue measure $\mathcal{L}$, we conclude

$$\begin{aligned}\frac{p_\ell\langle f,\bar{\nu}_\ell(t)\rangle - p_\ell c\epsilon}{\sum_{\ell'=1}^L p_{\ell'}(\bar{Q}_{\ell'}(t)+c\epsilon)}\left(\bar{K}(t+\epsilon)-\bar{K}(t)\right)\sum_{\ell=1}^{L}\underline{\mathrm{b}}_\ell(\epsilon)-\underline{\mathrm{a}}_\ell(\epsilon) &\leq \bar{A}(t+\epsilon)-\bar{A}(t)\\ &\leq \frac{p_\ell\langle f,\bar{\nu}_\ell(t)\rangle + p_\ell c\epsilon}{\sum_{\ell'=1}^L p_{\ell'}(\bar{Q}_{\ell'}(t)-c\epsilon)}\left(\bar{K}(t+\epsilon)-\bar{K}(t)\right)\sum_{\ell=1}^{L}\bar{b}_\ell(\epsilon)-\bar{a}_\ell(\epsilon)\end{aligned} \tag{6.50}$$

Dividing all sides by ϵ, taking the limit $\epsilon\to 0$,

$$\frac{p_\ell\langle f,\bar{\nu}_\ell(t)\rangle}{\sum_{\ell'=1}^L p_{\ell'}\bar{Q}_{\ell'}(t)}\frac{d\bar{K}(t)}{dt}\sum_{\ell'=1}^{L}\lim_{\epsilon\to 0}\underline{\mathrm{b}}_\ell(\epsilon)-\underline{\mathrm{a}}_\ell(\epsilon) \leq \frac{d}{dt}\bar{A}(t) \leq \frac{p_\ell\langle f,\bar{\nu}_\ell(t)\rangle}{\sum_{\ell'=1}^L p_{\ell'}\bar{Q}_{\ell'}(t)}\frac{d\bar{K}(t)}{dt}\sum_{\ell'=1}^{L}\lim_{\epsilon\to 0}\bar{b}_\ell(\epsilon)-\bar{a}_\ell(\epsilon)$$

Finally, substituting $d/dt\bar{K}(t)$ from (6.36) and using the fact that by (6.40)

$$\sum_{\ell'=1}^{L}\lim_{\epsilon\to 0}\underline{\mathrm{b}}_\ell(\epsilon)-\underline{\mathrm{a}}_\ell(\epsilon) = \sum_{\ell'=1}^{L}\lim_{\epsilon\to 0}\bar{b}_\ell(\epsilon)-\bar{a}_\ell(\epsilon) = 1,$$

(6.43) follows. □

Appendix A: On Marked Point Process T

In this section, we prove Lemma 5.1. We first recall the following lemma from Aghajani and Ramanan (2017) that is used in the proof of Lemma 5.1.

Let Y and $W_i, i\geq 1$, be $\mathbb{R}\cup\{\infty\}$-valued random variables defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, and let $\mathcal{G}\subset\mathcal{F}$ be a σ-algebra such that $W_i, i\geq 1$ and Y are conditionally independent given $\mathcal{G}$. For $i\geq 1$, define

$$\overline{F}_i(a) := \mathbb{P}\left(W_i > a|\mathcal{G}\right), \qquad a\geq 0.$$

Also, define $T \doteq \min(Y, W_i; i\geq 1)$ and let Z be a discrete-valued random variable such that $\{Y<\min(W_i; i\geq 1)\} = \{Z=z^*\}$ for some value z^*.

LEMMA A.1. *Suppose $Y, W_i; i \geq 1, Y, T$ and Z are as described above. Then, on $\{Z = z^*\}$, $W_i; i \geq 1$, are conditionally independent given $\mathcal{G}$, T and Z, that is, for every $n \geq 1$ and $b_1, ..., b_n \geq 0$,*

$$\mathbb{1}(Z = z^*)\,\mathbb{P}\left(W_i > T + b_i; i \geq 1 \middle| \mathcal{G}, T, Z\right) = \mathbb{1}(Z = z^*) \prod_{i=1}^{n} \mathbb{P}\left(W_i > T + b_i \middle| \mathcal{G}, T, Z\right), \tag{A.1}$$

and, for $b \geq 0$,

$$\mathbb{1}(Z = z^*)\,\mathbb{P}\left(W_i > T + b \middle| \mathcal{G}, T, Z\right) = \mathbb{1}(Z = z^*)\,\frac{\overline{F}_i(T+b)}{\overline{F}_i(T)}. \tag{A.2}$$

Lemma A.1 is proved in (Aghajani and Ramanan 2017, Lemma A.5).

Proof of Lemma 5.1. We prove the lemma by induction on k. Suppose by induction hypothesis that $\tau_k^{\mathfrak{E},\ell}, \ell = 1, ..., L$, $\sigma_{\ell,j}, \ell = 1, ..., L$, $j \in \mathcal{Q}_\ell(\tau_{k-1})$, and $\tau_k^{\mathfrak{D}}$ are conditionally independent given $\mathcal{F}_{\tau_{k-1}}$, and that equations (5.6)-(5.7) hold with k replaced by $k-1$, that is,

$$\mathbb{1}(j \in \mathcal{Q}_\ell(\tau_{k-1}))\,\mathbb{P}\left(\sigma_{\ell,j} - \tau_{k-1} > x \middle| \mathcal{F}_{\tau_{k-1}}\right) = \mathbb{1}(j \in \mathcal{Q}_\ell(\tau_{k-1}))\,\frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_{k-1}) + x)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_{k-1}))}, \quad x > 0, \tag{A.3}$$

$$\mathbb{1}(I(\tau_{k-1}) \neq 0)\,\mathbb{P}\left(\tau_k^{\mathfrak{D}} - \tau_{k-1} > x \middle| \mathcal{F}_{\tau_{k-1}}\right) = \mathbb{1}(I(\tau_{k-1}) \neq 0)\,\frac{\overline{G}_{S,I(\tau_{k-1})}(a(\tau_{k-1}) + x)}{\overline{G}_{S,I(\tau_{k-1})}(a(\tau_{k-1}))}, \quad x > 0, \tag{A.4}$$

We partition the sample space based on possible values of the mark z_k and the status of the server (idle or busy) at τ_{k-1}, and prove the result by constructing the underlying random variables on each of the sets in this partition. Recall from (5.1) that the next event after τ_{k-1} is given by

$$\tau_k = \min\{\tau_k^{\mathfrak{E},\ell}; \ell = 1, ..., L\} \wedge \min\{\sigma_{\ell,j}; \ell = 1, ..., L, j \in \mathcal{Q}_\ell(\tau_{k-1})\} \wedge \tau_k^{\mathfrak{D}}. \tag{A.5}$$

In the following, for every time $t \geq 0$, define $B_t := \{I(t) \neq 0\}$ to be subset of realizations on which the server is busy at time t. The following are all the possibilities for the event type at τ_k.

1. An arrival occurs at τ_k. For every pair (ℓ^*, j^*) with $l^* \in \{1, ..., L\}$ and $j^* \in \mathbb{Z}$, define the event $A_{\ell^*,j^*} := \{z_k = (\mathfrak{E}, \ell^*, j^*)\}$ on which the next event is the arrival of the job (ℓ^*, j^*). Note that on A_{ℓ^*,j^*}, $\tau_k = \tau_k^{\mathfrak{E},\ell^*}$, and the forward renewal time for the arrival process of jobs of class ℓ^* will reset after τ_k, while it remains unchanged for the arrival processes of jobs of other classes, that is,

$$\tau_{k+1}^{\mathfrak{E},\ell^*} = \tau_k + u_{\ell^*,E_\ell(\tau_k)+1}, \qquad \tau_{k+1}^{\mathfrak{E},\ell} = \tau_k^{\mathfrak{E},\ell} \text{ for } \ell \neq \ell^*,$$

where $u_{\ell^*,E_\ell(\tau_k)+1}$ is the next interarrival time of the renewal process E_ℓ^* after τ_k.

Moreover, when the server is busy right before this arrival, the job (ℓ^*, j^*) will be added to the set of job indices of class ℓ^*, and the next departure time after τ_k will remain unchanged. That is, on $A_{\ell^*,j^*} \cap B_{\tau_{k-1}}$,

$$Q_{\ell^*}(\tau_k) = Q_{\ell^*}(\tau_{k-1}) \cup \{j^*\}, \quad \sigma_{\ell^*,j^*} = \tau_k + \varpi_{\ell^*,j^*}, \quad Q_\ell(\tau_k) = Q_\ell(\tau_{k-1}) \text{ for } \ell \neq \ell^*, \quad \tau_{k+1}^{\mathfrak{D}} = \tau_k^{\mathfrak{D}}.$$

By an application of Lemma A.1, with $\mathcal{G}$ replaced by $\mathcal{F}_{\tau_{k-1}}$, T replaced by τ_k, Z replaced by z_k, z^* replaced by $(\mathfrak{E}, \ell^*, j^*)$, Y replaced by $t_k^{\mathfrak{E},\ell^*}$, and the sequence Z_i replaced by other random times on the right-hand side of (A.5), we conclude that on $A_{\ell^*,j^*} \cap B_{\tau_{k-1}}$, $\tau_{k+1}^{\mathfrak{E},\ell}, \ell \neq \ell^*$, $\sigma_{\ell,j}, \ell = 1, ..., L$, $j \in \mathcal{Q}_\ell(\tau_{k-1})$ and $\tau_k^{\mathfrak{D}}$ are conditionally independent given $\mathcal{F}_{\tau_k} = \mathcal{F}_{\tau_{k-1}} \wedge \sigma(\tau_k, z_k)$. Also, $\tau_{k+1}^{\mathfrak{E},\ell^*} - \tau_k = u_{\ell^*,E_\ell(\tau_k)+1}$ is the next interarrival time of the

renewal process E^*_ℓ after τ_k and hence is independent of other random variables, given $\mathcal{F}_{\tau_k}$. Moreover, the equation (A.2) of same lemma with F_i replaced by the conditional distribution given in (A.3), implies that for all $(\ell,j)\neq(\ell^*,j^*)$ and all $x\geq 0$

$$\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\mathbb{P}\left(\sigma_{\ell,j}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k)+x)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))},\tag{A.6}$$

and by another application of (A.2) with F_i replaced by the conditional distribution given in (A.4), for all $x>0$,

$$\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\right)\mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\right)\frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+x)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}.\tag{A.7}$$

Note that we have used the fact that on A_{ℓ^*,k^*}, $w(\tau_{k-1})+\tau_k-\tau_{k-1}=w(\tau_k)$ and $a(\tau_{k-1})+\tau_k-\tau_{k-1}=a(\tau_k)$. Moreover, since $\sigma_{\ell^*,j^*}-\tau_k=\varpi_{\ell^*,j^*}$ is the patience time of a newly arrived job, and hence is independent of the current observations $\mathcal{F}_{\tau_k}$ and satisfies

$$\mathbb{P}\left(\varpi_{\ell^*,j^*}>x\right)=\overline{G}_{R,\ell^*}(x)=\frac{\overline{G}_{R,\ell^*}(w_{\ell^*,j^*}(\tau_k)+x)}{\overline{G}_{R,\ell^*}(w_{\ell^*,j^*}(\tau_k))},$$

where the second equality holds because $w_{\ell^*,j^*}(\tau_k)=0$. Therefore, (A.6) hold for all pairs (ℓ,j), including (ℓ^*,j^*).

On the other hand, when the server is idle before the arrival, the job immediately enters service and hence, the queue remains empty and the time until the next departure at τ_k is set to be equal to the service time requirement of the job in service. That is, on $A_{\ell^*,j^*}\cap B^c_{\tau_{k-1}}$,

$$Q_\ell(\tau_k)=\emptyset \text{ for } \ell=1,...,L,\quad \tau^{\mathfrak{D}}_{k+1}=\tau_k+v_{\ell,K(\tau_k)}.$$

By an application of Lemma A.1, with $\mathcal{G}$ replaced by $\mathcal{F}_{\tau_{k-1}}$, T replaced by τ_k, Z replaced by z_k, z^* replaced by $(\mathfrak{E},\ell^*,j^*)$, Y replaced by $t^{\mathfrak{E},\ell^*}_k$, and the sequence Z_i replaced by other random times on the right-hand side of (A.5), we conclude that on $A_{\ell^*,j^*}\cap B^c_{\tau_{k-1}}$, $\tau^{\mathfrak{E},\ell}_{k+1},\ell=1,...,L$, and $\tau^{\mathfrak{D}}_k$ are conditionally independent given $\mathcal{F}_{\tau_k}=\mathcal{F}_{\tau_{k-1}}\wedge\sigma(\tau_k,z_k)$, and for all (ℓ,j) and all $x\geq 0$

$$A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}=\emptyset.\tag{A.8}$$

Moreover, since $\tau^{\mathfrak{D}}_{k+1}-\tau_k=v_{\ell,K(\tau_k)}$ is the service requirement time of the job just entered service, it is independent of the current observations $\mathcal{F}_{\tau_k}$ and satisfies

$$\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\right)\mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\overline{G}_{S,I(\tau_k)}(x)=\mathbb{1}\left(A_{\ell^*,j^*}\cap B_{\tau_{k-1}}\right)\frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+x)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}.\tag{A.9}$$

where the second equality holds because $a(\tau_k)=0$ on $A_{\ell^*,j^*}\cap B_{\tau_{k-1}}$.

2. A reneging occurs at τ_k. for every pair (ℓ^*,j^*) with $\ell^*\in\{1,...,L\}$ and $j^*\in\mathbb{Z}$, define the event $C_{\ell^*,j^*}:=\{z_k=(\mathfrak{R},\ell^*,j^*)\}$ on which the next event is the reneging of the job (ℓ^*,j^*) from the queue. Note that $C_{\ell^*,j^*}=\{\tau_k=\sigma_{\ell^*,j^*}\}$, and on C_{ℓ^*,j^*}

$$\tau^{\mathfrak{E},\ell}_{k+1}=\tau^{\mathfrak{E},\ell}_k \text{ for } \ell=1,...,L,\quad Q_{\ell^*}(\tau_k)=Q_{\ell^*}(\tau_{k-1})\backslash\{j^*\},\quad Q_\ell(\tau_k)=Q_\ell(\tau_{k-1}) \text{ for } \ell\neq\ell^*,\quad \tau^{\mathfrak{D}}_{k+1}=\tau^{\mathfrak{D}}_k.$$

By an application of Lemma A.1, with $\mathcal{G}$ replaced by $\mathcal{F}_{\tau_{k-1}}$, T replaced by τ_k, Z replaced by z_k, z^* replaced by $(\mathfrak{R},\ell^*,j^*)$, Y replaced by σ_{ℓ^*,j^*}, and the sequence Z_i replaced by other random times on the right-hand side of (A.5), we conclude that on C_{ℓ^*,j^*}, $\tau^{\mathfrak{E},\ell}_{k+1}, \ell=1,...,L$, $\sigma_{\ell,j}, \ell=1,...,L$, $j\in\mathcal{Q}_\ell(\tau_k)$ and $\tau^{\mathfrak{D}}_k$ are conditionally independent given $\mathcal{F}_{\tau_k}=\mathcal{F}_{\tau_{k-1}}\wedge\sigma(\tau_k,z_k)$. Moreover, as above, the equation (A.2) of same lemma with F_i replaced by the conditional distribution given in (A.3), implies that for all (ℓ,j) and all $x\geq 0$

$$\mathbb{1}\left(C_{\ell^*,j^*}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\mathbb{P}\left(\sigma_{\ell,j}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(C_{\ell^*,j^*}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k)+x)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))},\tag{A.10}$$

and by another application of (A.2) of same lemma with F_i replaced by the conditional distribution given in (A.4), for all $x>0$,

$$\mathbb{1}\left(C_{\ell^*,j^*}\right)\mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(C_{\ell^*,j^*}\right)\frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+x)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}.\tag{A.11}$$

3. A departure occurs at τ_k immediately followed by a service entry. For every pair (ℓ^*,j^*) with $l^*\in\{1,...,L\}$ and $j^*\in\mathbb{Z}$, define the event $D_{\ell^*,j^*}:=\{z_k=(\mathfrak{D},\ell^*,j^*)\}$ on which the next event is the departure of the job in service immediately followed by the entry of the job (ℓ^*,j^*) to the service. Note that on D_{ℓ^*,j^*}, $\tau_k=\tau^{\mathfrak{D}}_k$ and

$$t^{\mathfrak{E},\ell}_{k+1}=t^{\mathfrak{E},\ell}_k \text{ for } \ell=1,...,L,\quad Q_{\ell^*}(\tau_k)=Q_{\ell^*}(\tau_{k-1})\backslash\{j^*\},\quad Q_\ell(\tau_k)=Q_\ell(\tau_{k-1}) \text{ for } \ell\neq\ell^*,\quad \tau^{\mathfrak{D}}_{k+1}=\tau_k+v_{\ell,K(\tau_k)}.$$

By an application of Lemma A.1, with $\mathcal{G}$ replaced by $\mathcal{F}_{\tau_{k-1}}$, T replaced by τ_k, Z replaced by z_k, z^* replaced by $(\mathfrak{D},\ell^*,j^*)$, Y replaced by $\tau^{\mathfrak{D}}_k$, and the sequence Z_i replaced by other random times on the right-hand side of (A.5), we conclude that on D_{ℓ^*,j^*}, $\tau^{\mathfrak{E},\ell}_{k+1}, \ell=1,...,L$, and $\sigma_{\ell,j}, \ell=1,...,L$, $j\in\mathcal{Q}_\ell(\tau_k)$ are conditionally independent given $\mathcal{F}_{\tau_k}=\mathcal{F}_{\tau_{k-1}}\wedge\sigma(\tau_k,z_k)$. Also, $\tau^{\mathfrak{D}}_{k+1}-\tau_k=v_{\ell,K(\tau_k)}$ is the service requirement time of the job (ℓ^*,j^*) that enters service at τ_k, and hence is independent of other random variables, given $\mathcal{F}_{\tau_k}$. Again, the equation (A.2) of same lemma with F_i replaced by the conditional distribution given in (A.3), implies that for all (ℓ,j) and all $x\geq 0$

$$\mathbb{1}\left(D_{\ell^*,j^*}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\mathbb{P}\left(\sigma_{\ell,j}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\mathbb{1}\left(D_{\ell^*,j^*}\cap\{j\in\mathcal{Q}_\ell(\tau_k)\}\right)\frac{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k)+x)}{\overline{G}_{R,\ell}(w_{\ell,j}(\tau_k))},\tag{A.12}$$

Moreover, since $\tau^{\mathfrak{D}}_{k+1}-\tau_k=v_{\ell,K(\tau_k)}$ is the service requirement time of the job just entered service, it is independent of the current observations $\mathcal{F}_{\tau_k}$ and satisfies

$$\mathbb{1}\left(D_{\ell^*,j^*}\right)\mathbb{P}\left(\tau^{\mathfrak{D}}_{k+1}-\tau_k>x\big|\mathcal{F}_{\tau_k}\right)=\overline{G}_{S,I(\tau_k)}(x)=\mathbb{1}\left(D_{\ell^*,j^*}\right)\frac{\overline{G}_{S,I(\tau_k)}(a(\tau_k)+x)}{\overline{G}_{S,I(\tau_k)}(a(\tau_k))}.\tag{A.13}$$

where the second equality holds because $a(\tau_k)=0$ on D_{ℓ^*,j^*}.

4. A departure occurs at τ_k, that leaves the system empty. Finally, define the event $D_\emptyset:=\{z_k=(\mathfrak{D},\emptyset)\}$ on which the next event is the departure of the job in service and the service remains idle after the departure. Note that on $D_\emptyset$, $\tau_k=\tau^{\mathfrak{D}}_k$ and

$$t^{\mathfrak{E},\ell'}_{k+1}=t^{\mathfrak{E},\ell'}_k \text{ for } \ell'=1,...,L,\quad Q_{\ell'}(\tau_k)=\emptyset \text{ for } \ell'=1,...,L,,\quad \tau^{\mathfrak{D}}_{k+1}=\infty.$$

By an application of Lemma A.1, with $\mathcal{G}$ replaced by $\mathcal{F}_{\tau_{k-1}}$, T replaced by τ_k, Z replaced by z_k, z^* replaced by $(\mathfrak{D},\emptyset)$, Y replaced by $\tau^{\mathfrak{D}}_k$, and the sequence Z_i replaced by other random times on the right-hand side of

(A.5), we conclude that on $D_\emptyset$, $\tau^{\mathfrak{E},\ell}_{k+1}, \ell = 1, ..., L$, are conditionally independent given $\mathcal{F}_{\tau_k} = \mathcal{F}_{\tau_{k-1}} \wedge \sigma(\tau_k, z_k)$. Moreover, note that on $D_\emptyset$, there are no jobs in queue at time τ_k, that is, for all (ℓ, j),

$$D_{\ell^*,j^*} \cap \{j \in \mathcal{Q}_\ell(\tau_k)\} = \emptyset, \tag{A.14}$$

and the server is idle at τ_k, that is,

$$\{I(\tau_k) \neq 0\} \cap D_\emptyset = \emptyset. \tag{A.15}$$

Putting all the cases together, we conclude that $\tau^{\mathfrak{E},\ell}_{k+1}, \ell = 1, ..., L$, $\sigma_{\ell,j}, \ell = 1, ..., L$, $j \in \mathcal{Q}_\ell(\tau_k)$, and $\tau^{\mathfrak{D}}_{k+1}$ are conditionally independent given $\mathcal{F}_{\tau_k}$. Moreover, (5.6) follows on summing (A.6), (A.8), (A.10),(A.12), and (A.14) over all $\ell = 1, ...,$ and $j \in \mathbb{Z}$, and (5.7) follows on summing (A.7), (A.9), (A.11), (A.13), and (A.15) over all $\ell = 1, ...,$ and $j \in \mathbb{Z}$.

Finally, given $\mathcal{F}_{\tau_k}$, $\kappa_{K(\tau_k)+1}$ is drawn from an i.i.d. sequence of uniformly distributed random variables, that are independent of all other random variables, and hence, (5.8) is satisfied. □

Appendix B: List of Notation

B.1. Primitives

E_ℓ	Arrival process of class ℓ
$R_{\ell,j}$	indicator function of reneging of job j of class ℓ
$K_{\ell,j}$	Counting service entry process
$u_{\ell,j}$	Inter-arrival times of class ℓ
λ_ℓ	arrival rate of class ℓ
$\varpi_{\ell,j}$	Patience Time of class ℓ
γ_ℓ	reciprocal mean of patience time of class ℓ
$v_{\ell,j}$	Service Requirement of class ℓ
μ_ℓ	reciprocal mean of service requirement of class ℓ

B.2. State Spaces

$\mathcal{M}$	The space of non-negative Borel measures on $[0, \infty)$

B.3. State Variables

$\alpha_{\ell,_j}$	Arrival time of job j of class ℓ
$\beta_{\ell,_j}$	Service entry time of job j class ℓ
$\sigma_{\ell,_j}$	reneging time of job j class ℓ
Y	State Vector
r_ℓ	backward recurrence Time class ℓ
$\mathcal{Q}_\ell$	set of jobs in queue class ℓ
Q_ℓ	number of jobs of class ℓ waiting in queue
$w_{\ell,_j}$	current waiting time of job j class ℓ
a	age of the job in service
$I(t)$	class of the current job in service at time t
$J(t)$	index of the current job in service at time t

B.4. Constants

L	number of different classes
p_ℓ	weight of class ℓ in DROS algorithm

Acknowledgments

Research supported in part by NSF Grant DMS-1712974 and the Charles Lee Powell Foundation.

References

Aghajani R, Ramanan K (2017) The hydrodynamic limit of a randomized load balancing network, arXiv:1707.02005 [math.PR].

Asmussen S (2003) *Applied Probability and Queues* (Springer-Verlag), 2nd edition edition.

Billingsley P (1968) *Convergence of Probability Measures* (New York: John Wiley).

Bremaud P (1981) *Point Processes and Queues: Martingale Dynamics: Martingale Dynamics*. Advances in Physical Geochemistry (Springer), ISBN 9780387905365.

Chen H, Yao D (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Applications of mathematics : stochastic modelling and applied probability (Springer), ISBN 9780387951669.

Ethier S, Kurtz T (1986) *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics (Wiley).

Huber PJ (1981) *Robust Statistics* (New York: John Wiley & Sons).

Jakubowski A (1986) On the Skorokhod topology. *Ann. Inst. H. Poincaré Probab. Statist.* 22(3):263–285, ISSN 0246-0203.

Kallenberg O (2017) *Random Measures, Theory and Applications*. Probability Theory and Stochastic Modelling (Springer International Publishing), ISBN 9783319415987.

Klebaner FC (2005) *Introduction to Stochastic Calculus with Applications*. Introduction to Stochastic Calculus with Applications (Imperial College Press), ISBN 9781860945557.

Rogers LCG, Williams D (2000) *Diffusions, Markov processes, and Martingales. Vol. 2*. Cambridge Mathematical Library (Cambridge University Press, Cambridge), ISBN 0-521-77593-0.