Fluid limits for many-server systems with reneging under a priority policy[*]

Rami Atar[†] Haya Kaspi[‡] Nahum Shimkin[†]

March 27, 2012; last revised June 25, 2013

Abstract

A multi-class many-server system is considered, in which customers are served according to a non-preemptive priority policy and may renege while waiting to enter service. The service and reneging time distributions satisfy mild conditions. Building on an approach developed by Kaspi and Ramanan, the Law-of-Large-Numbers many-server asymptotics are characterized as the unique solution to a set of differential equations in a measure space, regarded as fluid model equations. In stationarity, convergence to the explicitly solved invariant state of the fluid model equations is established. An immediate consequence of the results in the case of exponential reneging is the asymptotic optimality of an index policy, called the $c\mu/\theta$ rule, for the problem of minimizing linear queue-length and reneging costs. A certain Skorohod map plays an important role in obtaining both uniqueness of solutions to the fluid model equations, and convergence.

AMS subject classifications: 60K25, 60F17, 68M20

Keywords: Many-server systems, reneging, fluid limits, measure-valued processes, Skorohod map, the $c\mu/\theta$ rule.

1 Introduction

A multi-class system with many servers is studied under a Law-of-Large-Numbers (LLN) scaling. In this system, customers of various classes are served according to a fixed non-preemptive priority policy; they may leave the system while waiting to enter service. The goal is to study the scaling limit of the queue-length and other processes of the model, via the approach of Kaspi and Ramanan [14] and Kang and Ramanan [12], [13]. In this approach, developed in [14] for the G/G/N queue, and extended in [12] to include customer reneging, the scaling limit is described in terms of a *fluid model*, comprising a system of differential equations in measure space. The relation to the fluid model equations (FME) is then used to show convergence of stationary laws of the queueing model

[*] Research supported in part by the ISF (Grants 1349/08, 81/09, 1319/11), the US-Israel BSF (Grants 2006379, 2008466), and the Technion fund for promotion of research

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

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

to the invariant state of the FME, and is applied to prove the asymptotic optimality of the so-called *$c\mu/\theta$ rule* for linear abandonment and queue-length costs, in the case of exponential reneging.

While multi-server queues are important as they arise in many applications, they are harder to analyze than single server queues. As was first observed by Halfin and Whitt [9], letting the number of servers increase to infinity may sometimes simplify the system description. In particular, in [9], a G/M/N queue was studied with scaled-up number of servers and arrival rate, and a central limit theorem (CLT) was established in which the limiting dynamics was identified as a one-dimensional diffusion process. It is well-understood that, whether in LLN or CLT scale, it is the exponential distribution assumption on the service time that enables to describe the limiting dynamics in terms of a (deterministic or stochastic) ordinary differential equation in one real variable. In Kaspi and Ramanan [14], the G/G/N queue was analyzed in a many-server LLN scaling, and the limit behavior was shown to be governed by a (deterministic) differential equation in measure space. In this approach, the Markovian state descriptor of the queueing model consists of the number-in-system process and a measure-valued process that records the age-in-service of each of the customers being served. The FME characterize the dynamics of the limits of a properly scaled version of these quantities. The extension by Kang and Ramanan [12] to a setting with reneging has an additional ingredient in the state descriptor, that accounts for the age-in-system of customers prior to reneging, and accordingly an extended set of FME. The limiting behavior, in LLN and CLT scales, was also identified by a different approach by Reed [17] and Puhalskii and Reed [16] (see [14], [12] for further references on many-server limit results).

This paper extends the results of [14] and [12] to the setting of multi-class systems with reneging, where the service allocation adheres to a fixed non-preemptive priority among the customer classes. Convergence of the scaled queueing model processes to a suitable set of FME is established, on a finite time interval, and in stationarity. The approach and much of the technique build on [14], [12] and [13], including the Markovian formulation, representation formulas for solutions to the FME, tightness of various processes, and the analysis of stationary measures and their convergence. In fact, this paper can be viewed as an attempt to demonstrate the applicability and versatility of the approach.

Yet, the techniques developed in the above papers alone fall short of covering the model under consideration; particularly, the uniqueness of solutions to the FME and the convergence of the queueing model processes to the FME solution do not follow directly from these treatments. As we show, a certain *Skorohod map* (SM) can be used to represent some of the model's processes (queue-length, idleness, arrival into service) as images of others (exogenous arrivals, departure, reneging). This representation turns out to capture a useful property of the priority policy. Indeed, continuity and other properties of the SM play a key role in the proofs of uniqueness and convergence alluded to above. While this is a simple example of a SM, to the best of our knowledge it has not been used before in a queueing setting.

A major motivation for this study arises from a natural dynamic control problem, in which scheduling is to be determined so as to (asymptotically) minimize a linear abandonment/queue-length cost in stationarity. While the problem is interesting under any reneging time distribution, we focus in this part of the paper on the relatively simple case of the (class-dependent) exponential distribution. Under this assumption, the cost can be expressed solely as a queue-length cost. This problem was considered in [1] and [2] in the case where also the *service times* are (class-dependent) exponential, in which the Markovian state descriptor is finite-dimensional. Denoting by c_i, μ_i and $\theta_i > 0$, respectively, the cost per customer per unit time, the rate of service, and the rate of reneging for a class-i customer, it was shown that a policy that prioritizes classes in the order of the index $c_i\mu_i/\theta_i$ (with highest priority to the largest index) achieves asymptotic optimality. In addition, a lower bound on the cost was established in [2] for *general* service time distributions. It was proposed in [1] to refer to this policy as the *$c\mu/\theta$ rule*, as it is reminiscent of the well-known $c\mu$ rule (which is, under suitable assumptions, optimal for multi-class scheduling in systems without reneging). It follows from the main results of the present paper that the aforementioned lower bound is achieved, in an asymptotic sense, by the $c\mu/\theta$ rule for a general service time distribution. Here, μ_i now stands for the reciprocal mean class-i service time. Although the priority rule is simple to state, the proof of the asymptotic optimality result is not so simple, and in fact uses the main results of this paper to their full strength.

In summary, the main contribution of this paper is the treatment of a multi-class many-server queueing system with non-preemptive priorities, with general service and reneging distribution, based on the approach of [12]–[14] and significantly extending it. This extension, that we believe may be of broader interest in the analysis of priority queues, includes the following:

- *The formulation of a set of FME for the multi-class many-server system with reneging, under a non-preemptive priority policy.* We establish uniqueness of solutions to this set of equations (Theorem 3.1), and identify their invariant state (Theorem 3.3). While the formulation of the FME follows the approach of [12]–[14], and several tools are borrowed from these works (Proposition 3.2), a crucial new tool is a certain two-dimensional Skorohod map, that effectively captures the nature of the priority discipline (Section 3.2).
- *Convergence analysis:* We establish convergence in law of the scaled queueing processes to the FME solution (Theorem 4.3), and consequently the convergence of any invariant state distribution of the scaled queueing processes to the invariant state of the FME (Theorem 4.4). Here, the methodology follows closely the framework of [12]–[14]. Continuity properties of the Skorohod map alluded to above play a role here.

As a corollary of the convergence results, we obtain

- Asymptotic optimality of the $c\mu/\theta$ priority rule for exponential reneging and general service time distribution (Theorem 5.1), significantly extending a known result for the case of exponential service.

We use the following notation. For $x \in \mathbb{R}$, $x^{\pm} = \max(\pm x, 0)$. For $x \in \mathbb{R}^k$, $\|x\| = \sum_{i=1}^k |x_i|$. For $y : \mathbb{R}_+ \to \mathbb{R}^k$ and $t > 0$, $\|y\|_t = \sup_{s\in[0,t]} \|y(s)\|$. The modulus of continuity of y is defined as

$$w(y, \theta, t) = \sup\{\|y(s) - y(u)\| : s, u \in [0, t], |s - u| \leq \theta\}, \qquad \theta, t > 0.$$

If $y : \mathbb{R}_+ \to \mathbb{R}$ is locally of bounded variation, we write $|y|_t$ for the variation of y over $[0, t]$. Note that we sometimes use $y(t)$ and y_t interchangeably as convenient.

Given a non-decreasing, right-continuous function $f : [0, \infty) \to [0, \infty)$, denote $f_* = \sup_{t\geq 0} f(t)$ and, in the case when $f_* = \infty$, define $f^{-1} : [0, \infty) \to [0, \infty)$ by $f^{-1}(t) = \inf\{s \geq 0 : f(s) \geq t\}$ for

$t \in [0,\infty)$. When $f_* < \infty$, let $f^{-1} : [0,\infty) \to [0,\infty]$ be defined as above for $t \in [0, f_*]$, and set $f^{-1}(t) = \infty$ for $t \in (f_*, \infty)$. This is the left-continuous inverse of f. For a measure m over $[0, H)$ (some $H \in (0, \infty]$), we will write $m[a, b)$ as shorthand for $m([a, b))$ and $m[a, b]$ for $m([a, b])$. We write $F^m(x)$ for $m[0, x]$ and denote

$$\langle f, m\rangle = \int_{[0,H)} f dm, \qquad f : [0, H) \to \mathbb{R}. \tag{1}$$

Note that

$$(F^m)^{-1}(y) = \inf\{x \geq 0 : m[0, x] \geq y\}. \tag{2}$$

For $a \in \mathbb{R}$, δ_a denotes the unit mass at a. For an event $A \in \mathcal{F}$, 1_A denotes the indicator of A.

Given a Polish space E, its Borel σ-field $\mathcal{E}$, and an E-valued random variable X on the probability space $(\varOmega, \mathcal{F}, \mathbb{P})$, the probability measure $\mathfrak{L}(X)$ on $(E, \mathcal{E})$, defined as $\mathbb{P} \circ X^{-1}$, is referred to as the *law* of X. Given random variables $X, X_1, X_2, \ldots$ taking values in E, we write $X_n \Rightarrow X$ for convergence *in law* defined as the weak convergence of the laws, $\mathfrak{L}(X_n) \to \mathfrak{L}(X)$, as probability measures over $(E, \mathcal{E})$. The sequence $\{X_n\}$ is said to be *tight* if the corresponding laws form a tight sequence in $\mathcal{P}(E, \mathcal{E})$. Denote by $\mathcal{D}_E[0, H)$ the space of RCLL paths from $[0, H)$ to E, equipped with the usual Skorohod topology. A sequence X_n of random variables taking values in this space is said to be *C-tight* if it is tight and every subsequential limit has continuous paths w.p.1.

We write $\mathcal{M}_F[0, H)$ for the space of finite measures on $[0, H)$, and endow it with the topology of weak convergence. All stochastic processes in this paper are assumed to have RCLL sample paths.

Finally, the dependence on $t \in [0,\infty)$ of a process, say X_i, will be denoted by $X_{i,t}$ and $X_i(t)$ interchangeably, whichever notation is more convenient.

The paper is organized as follows. The queueing model is introduced in Section 2. Section 3 describes the FME, establishes their uniqueness and identifies the invariant state. In Section 4 the convergence results are stated and proved. The results are then applied in Section 5 to prove the asymptotic optimality of the $c\mu/\theta$ rule under exponential reneging. Finally, certain properties of the SM are proved in the appendix.

2 The N-server system

In this section we give a precise description of the model. The system has N identical servers that serve customers of J classes. Each customer has a single service requirement, and leaves the system once his service is completed. Another possibility for a customer to leave the system is by reneging while waiting to be served. The system is considered under a work conserving, non-preemptive priority policy. Thus, customers that arrive into the system when one of the servers is idle are immediately assigned a server. Otherwise they are queued in a buffer (with infinite room), and are sent to the service as soon as a server becomes available. The order in which the customers are assigned to service follows a priority rule, where each class i has priority over all the classes $i + 1, \ldots, J$. Within the class, customers are sent to servers in a first-come-first-served manner.

The model is defined on a probability space $(\varOmega, \mathcal{F}, \mathbb{P})$. For the j-th customer of class i to enter the system we let

- $r_{i,j}$ be the patience time of the customer,
- $v_{i,j}$ be the service time requirement of the customer.

This means that the customer reneges if he waits in the queue $r_{i,j}$ units of time, and if the customer is assigned a server, he keeps the server busy for $v_{i,j}$ units of time. We assume that the patience times of class-i customers, $\{r_{i,j} : j = 1, 2, \dots\}$, are i.i.d. random variables with distribution G_i^r, density g_i^r, finite mean θ_i^{-1}, and hazard rate function $h_i^r(x) = (1 - G_i^r(x))^{-1} g_i^r(x)$, where, by convention, $0/0 = 0$. We denote $H_i^r = \inf\{x : G_i^r(x) = 1\}$. Similarly, the service times of class-i customers, $\{v_{i,j} : j = 1, 2, \dots\}$, are i.i.d. random variables with distribution G_i^s, density g_i^s, finite mean μ_i^{-1}, and hazard rate function $h_i^s(x) = (1 - G_i^s(x))^{-1} g_i^s(x)$. Also, $H_i^s = \inf\{x : G_i^s(x) = 1\}$.

We refer to the system containing N servers as the *N-server system*, or simply the *N-th system*. For each fixed N we consider the following arrival processes associated with the N-th system:

- $(e_{i,j}^N)$, where $e_{i,j}^N$ is the time of arrival into the system of the j-th customer of class i,
- E_i^N, the corresponding counting process of class-i arrivals into the system for $t \geq 0$, so that $E_i^N(t)$ is the number of class-i arrivals in $[0, t]$.

We further denote by $\mathcal{E}_{i,0}^N$ the number of class-i customers that have arrived before $t = 0$. We assign to these customers negative indices between $-\mathcal{E}_{i,0}^N + 1$ and 0, and to those that arrive at or after time 0 positive indices from 1 to ∞. Hence $e_{i,j}^N = (E_i^N)^{-1}(j)$ for $j \geq 1$.

We assume that the arrival processes $\{E_i^N : i = 1, \dots, J\}$ are mutually independent renewal processes with mean inter-arrival times $(\lambda_i^N)^{-1}$, respectively. It is further assumed that the collections $r_{i,\cdot}$, $v_{i,\cdot}$ and $e_{i,\cdot}^N$ (equivalently, E_i^N), $i = 1, \dots, J$, are mutually independent for each N. At this point of the article we are interested in the evolution of the systems for $t \geq 0$ starting from a given *initial state*, a term that refers to quantities associated with customers that are present in the system at time 0 (including their number, arrival time, and time already spent in service). Therefore, the distribution of the initial state is not specified Starting at Section 4, the initial state will be considered with a (generic) distribution. We will assume that, given the ages of the customers in service (as part of the initial state), their residual service time distribution is that of independent random variables with densities $\frac{g_i^s(x+y)}{1-G_i^s(x)}$ for a customer of class i with age x in service. A similar statement holds for ages in queue, with g_i^r and G_i^r replacing g_i^s and G_i^s.

We proceed to define some additional processes for the N-th system that depend on the above primitive variables, starting with

- $s_{i,j}^N$, the time of entrance into service of the j-th customer of class i. If the customer reneges before entering service we set $s_{i,j}^N = \infty$.
- K_i^N, the counting process of class-i customers that enter service for $t \geq 0$.

Define next the *age-in-service* measures, denoted by $\nu_{i,t}^N(dx)$. For class i and $t \in [0, \infty)$, $\nu_{i,t}^N$ puts a unit mass at every $x \in [0, \infty)$ for which a class-i customer, that is in service, has been there x

units of time at time t. More precisely,

$$\nu_{i,t}^N(dx) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \delta_{a_{i,j}^N(t)}(dx) 1_{\{0 \le t - s_{i,j}^N < v_{i,j}\}}, \tag{3}$$

where $\{0 \le t - s_{i,j}^N < v_{i,j}\}$ indicates that the customer entered service but has not completed it yet, and

$$a_{i,j}^N(t) = ((t - s_{i,j}^N) \vee 0) \wedge v_{i,j} \tag{4}$$

represents the age in service of the respective customer at time t. Next, consider the *potential queue* measures, $\eta_{i,t}^N(dx)$. These measures represent the age-in-queue, under a policy that never assigns servers to any customers (this policy is not actually implemented in our model, and is mentioned only as a means of describing the potential queue measures). Therefore, these measures encode information about arrival and reneging, but not service. Specifically,

$$\eta_{i,t}^N(dx) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \delta_{w_{i,j}^N(t)}(dx) 1_{\{0 \le t - e_{i,j}^N < r_{i,j}\}}, \tag{5}$$

where $\{0 \le t - e_{i,j}^N < r_{i,j}\}$ indicates a customer that has arrived prior to t but not reneged yet, and $w_{i,j}^N$ are the potential waiting times, defined by

$$w_{i,j}^N(t) = ((t - e_{i,j}^N) \vee 0) \wedge r_{i,j}. \tag{6}$$

Although $\eta_{i,t}^N$ encodes the age-in-queue under a fictitious policy, the information about the ages of customers in queue under the actual policy can be recovered from it, using additional ingredients of the system description, as we shall see below.

Let $B_i^N(t) = \langle 1, \nu_{i,t}^N \rangle$ denote the total mass of $\nu_{i,t}^N$, representing the number of class-i customers that are in service at time t, or equivalently, the number of servers busy with class-i customers. Let $Q_i^N(t)$ denote the number of class-i customers in the queue at time t. Let

$$X_i^N(t) = Q_i^N(t) + B_i^N(t) \tag{7}$$

denote the total number of class-i customers in the system at time t. Then we require that

$$N - \sum_{i=1}^J B_i^N(t) = \Big(N - \sum_{i=1}^J X_i^N(t)\Big)^+, \qquad t \ge 0. \tag{8}$$

This relation asserts that servers do not idle when there are customers waiting in the queue. It thus expresses the work conservation property.

Introduce the process $\chi_i^N(t)$ representing the waiting time of the "oldest" class-i customer in the queue, and set it equal zero when the class-i queue is empty. Namely,

$$\chi_i^N(t) = \inf\{x \ge 0 : \eta_{i,t}^N[0,x] \ge Q_i^N(t)\} = (F^{\eta_{i,t}^N})^{-1}(Q_i^N(t)), \tag{9}$$

where we recall the definition of the inverse in (2). Evidently,

$$Q_i^N(t) = \eta_{i,t}^N[0, \chi_i^N(t)].$$

The cumulative class-i departure-from-service process, denoted D_i^N, is given by

$$D_i^N(t) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \sum_{s\in[0,t]} 1_{\{\frac{da_{i,j}^N}{dt}(s-)>0,\, \frac{da_{i,j}^N}{dt}(s+)=0\}}, \tag{10}$$

where we denote by $(df/dt)(t+)$ and $(df/dt)(t-)$ the right- and, resp., left-derivative of f at t. The cumulative potential reneging of class-i customers in $[0,t]$, denoted $S_i^N(t)$, is equal to

$$S_i^N(t) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \sum_{s\in[0,t]} 1_{\{\frac{dw_{i,j}^N}{dt}(s-)>0, \frac{dw_{i,j}^N}{dt}(s+)=0\}}. \tag{11}$$

The cumulative reneging of class-i customers in $[0,t]$, denoted $R_i^N(t)$, is equal to

$$R_i^N(t) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \sum_{s\in[0,t]} 1_{\{w_{i,j}^N(s)<\chi_i^N(s-), \frac{dw_{i,j}^N}{dt}(s-)>0, \frac{dw_{i,j}^N}{dt}(s+)=0\}}. \tag{12}$$

Additional relations satisfied by these processes are the so-called balance equations, obtained by counting customers in the system (13), in the potential queue (14) and in service (15). Namely,

$$X_i^N = X_{i,0}^N + E_i^N - D_i^N - R_i^N, \tag{13}$$

$$\langle 1, \eta_i^N\rangle = \langle 1, \eta_{i,0}^N\rangle + E_i^N - S_i^N, \tag{14}$$

$$B_i^N = B_{i,0}^N + K_i^N - D_i^N. \tag{15}$$

The non-preemptive priority rule is expressed as

$$K_i^N(t) = \int_{[0,t]} 1_{\{\sum_{k=1}^{i-1} Q_k^N(s)=0\}} dK_i^N(s), \qquad i \ge 2,\, t \ge 0. \tag{16}$$

This relation imposes a necessary condition for a class-i customer to be sent to service at time s, namely that at time s no class-k customers are present in the queue, for $k < i$.

Remark 2.1. *While* (16) *captures precisely the nature of the priority rule, it may seem that the following variant is also a valid condition, namely*

$$K_i^N(t) = \int_{[0,t]} 1_{\{\sum_{k=1}^{i-1} Q_k^N(s-)=0\}} dK_i^N(s), \qquad i \ge 2,\, t \ge 0. \tag{17}$$

Here the respective queues are observed just before time s. However, as we explain below, (16) *and* (17) *are not equivalent, and* (17) *is not the right condition.*

(a) Condition (17) does not agree with the priority policy. *Fix $k < i$. Consider a scenario when two arrivals, of class k and class i, occur at the same time. Assume that just prior to this time the queues are all empty, and there is exactly one free server. The policy should assign the server to the new class-k customer. However, condition* (17) *allows for the class-i customer to be sent to service rather than k. Condition* (16) *prohibits this behavior.*

(b) Condition (17) contradicts work conservation. *Suppose two servers become idle at the same time, and just before that time there are one class-k customer and one class-i customer in the queue. Both should enter service. However,* (17) *prohibits this.*

We further consider the departure-from-service marked point processes, defined for bounded measurable φ on $[0, H_i^s) \times \mathbb{R}_+$ via

$$D_{i,\varphi}^N(t) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \sum_{s\in[0,t]} 1_{\{\frac{da_{i,j}^N}{ds}(s-)>0, \frac{da_{i,j}^N}{ds}(s)=0\}} \varphi(a_{i,j}^N(s), s), \tag{18}$$

and similarly the potential reneging marked point processes, defined for bounded measurable ψ on $[0, H_i^r) \times \mathbb{R}_+$ via

$$S_{i,\psi}^N(t) = \sum_{j=-\mathcal{E}_{i,0}^N+1}^{E_i^N(t)} \sum_{s\in[0,t]} 1_{\{\frac{dw_{i,j}^N}{ds}(s-)>0, \frac{dw_{i,j}^N}{ds}(s)=0\}} \psi(w_{i,j}^N(s), s). \tag{19}$$

Let

$$R_{i,\psi}^N(t) = S_{i,\theta_i^N\psi}^N(t), \tag{20}$$

where $(\theta_i^N\psi)(x, s) = \theta_i^N(x, s)\psi(x, s)$, and

$$\theta_i^N(x, s) = 1_{(x,\infty)}(\chi_i^N(s-)). \tag{21}$$

Then, the reneging process R_i^N is given by

$$R_i^N(t) = R_{i,1}^N = S_{i,\theta_i^N}^N(t). \tag{22}$$

For $h \in (0, \infty]$, we denote by $\mathcal{C}_c^{1,1}([0, h) \times \mathbb{R}_+)$ the space of compactly supported functions φ for which the directional derivative $\lim_{\Delta\to 0} \frac{\varphi(x+\Delta,t+\Delta)-\varphi(x,t)}{\Delta}$ exists for all $x \in [0, h), t \in \mathbb{R}_+$, and lies in $\mathcal{C}_c([0, h) \times \mathbb{R}_+)$. We shall abuse the notation slightly and denote this directional derivative by $\varphi_x + \varphi_t$ whether the partial derivatives φ_x and φ_t exist or not. For $\varphi \in \mathcal{C}_c^{1,1}([0, H_i^s) \times \mathbb{R}_+)$, the measure-valued processes satisfy the following relations:

$$\langle\varphi(\cdot, t), \nu_{i,t}^N\rangle = \langle\varphi(\cdot, 0), \nu_{i,0}^N\rangle + \int_0^t \langle\varphi_x(\cdot, s) + \varphi_t(\cdot, s), \nu_{i,s}^N\rangle ds - D_{i,\varphi}^N(t) + \int_0^t \varphi(0, s)dK_i^N(s), \tag{23}$$

where $\varphi_x + \varphi_t$ is the directional derivative alluded to above. Similarly, for $\psi \in \mathcal{C}_c^{1,1}([0, H_i^r) \times \mathbb{R}_+)$,

$$\langle\psi(\cdot, t), \eta_{i,t}^N\rangle = \langle\psi(\cdot, 0), \eta_{i,0}^N\rangle + \int_0^t \langle\psi_x(\cdot, s) + \psi_t(\cdot, s), \eta_{i,s}^N\rangle ds - S_{i,\psi}^N(t) + \int_0^t \psi(0, s)dE_i^N(s). \tag{24}$$

The proof that, given K_i^N and E_i^N, (23)–(24) are satisfied by the measure valued processes, is identical to that of Theorem 5.1 of [14]. The construction of collection of processes satisfying the N-server system equations (3)–(16), (18)–(22) is very similar to that in Appendix A of [12], with obvious adaptations to address the priority policy.

While the detailed proofs appear in [12] and [14], it is in order to give an explanation of the various terms in the above equations. First, $\theta_i^N(x,s)$ is the indicator of the event that the waiting time of the customer at the head of the queue, just before s, is larger than x. Hence $S^N_{i,\theta_i^N}(t)$ is the potential reneging applied to the function θ_i^N, which counts all reneging of customers while they are in queue, that is, the actual reneging in $[0,t]$. Equations (23)–(24) describe the evolution of the measures ν_i^N and η_i^N, where the second, third and fourth terms on the right correspond to three different causes of evolution. The second term is due to the fact that ages of customers in service (resp., waiting times of customers in queue) increase at rate 1. The variables (x,t) for the test functions φ (resp., ψ) correspond to age (resp., waiting time) and time. Since both these elements are affected by the flow of time, the directional derivative as defined above appears in these expressions. Clearly, in the special case when φ (resp., ψ) is a function of the space variable x alone, only the term $\langle\varphi_x(\cdot),\nu^N_{i,s}\rangle$ (resp., $\langle\psi_x,\eta^N_{i,s}\rangle$) will appear. Next, those customers that have left the system in $[0,t]$ due to end of service in (23) and because of reneging in (24), should be subtracted, resulting in the third term on the r.h.s. Finally, the last term represents entrance to the service (resp., the system) during $[0,t]$. The test functions appear here as $\varphi(0,s)$ (resp., $\psi(0,s)$) due to the fact that at the time customers enter, their age (resp., waiting time) is equal to 0.

3 The fluid model

In this section we analyze a deterministic *fluid model* that will be shown, in later sections, to govern the LLN behavior of the N-server system, as $N\to\infty$. It consists of a set of equations derived from the equations satisfied by the N-server model. The main issue addressed here is showing that the solution of the fluid-model equations is unique. We also provide here some additional properties of the fluid model and characterize its invariant state.

3.1 The fluid model equations

Write $\mathcal{D}^+_{\mathbb{R}^J}(\mathbb{R}_+)$ for the set of members of $\mathcal{D}_{\mathbb{R}^J}(\mathbb{R}_+)$ that are nonnegative and nondecreasing (componentwise). We are given data $E\in\mathcal{D}^+_{\mathbb{R}^J}(\mathbb{R}_+)$ and initial conditions $X_{i,0}\in[0,\infty)$, $\nu_{i,0}\in\mathcal{M}_F[0,H_i^s)$ and $\eta_{i,0}\in\mathcal{M}_F[0,H_i^r)$, for $i=1,\dots,J$. Set $B_{i,0}=\langle 1,\nu_{i,0}\rangle$. We consider equations satisfied by (B,X,Q,D,K,R,ν,η), where $B=(B_i)_{i=1,\dots,J}$, etc., and, for each i, B_i,X_i,Q_i,D_i,K_i,R_i are members of $\mathcal{D}_{\mathbb{R}}(\mathbb{R}_+)$, and ν_i and η_i are members of $\mathcal{D}_{\mathcal{M}_F[0,H_i^s)}(\mathbb{R}_+)$ and $\mathcal{D}_{\mathcal{M}_F[0,H_i^r)}(\mathbb{R}_+)$, respectively.

The measures ν_i and η_i are assumed to satisfy

$$\int_0^t\langle h_i^s,\nu_{i,\tau}\rangle d\tau<\infty,\qquad\int_0^t\langle h_i^r,\eta_{i,\tau}\rangle d\tau<\infty,\qquad t\ge 0.\tag{25}$$

Balance equations and basic relations (in analogy with (7), (13), (14), (15)) are expressed by

$$B_i = B_{i,0} - D_i + K_i, \tag{26}$$

$$X_i = X_{i,0} - D_i + E_i - R_i, \tag{27}$$

$$Q_i = X_i - B_i, \tag{28}$$

$$Q_i \text{ and } B_i \text{ are nonnegative} \tag{29}$$

Work conservation and non-preemptive priority (8), (16), correspond to

$$I := 1 - \sum_{i=1}^{J} B_i = \Big(1 - \sum_{i=1}^{J} X_i\Big)^+, \tag{30}$$

$$K_i \text{ are nonnegative, nondecreasing} \tag{31}$$

$$K_{i,t} = \int_{[0,t]} 1_{\{\sum_{j=1}^{i-1} Q_{j,s}=0\}} dK_{i,s}, \qquad i \geq 2,\, t \geq 0. \tag{32}$$

Note that one can deduce that X_i are nonnegative from the nonnegativity of Q_i and B_i, and that $\sum B_i \leq 1$ from (30). Also note that (30) imposes an assumption on the initial condition.

Further, in analogy with (23) and (24), we write the following integral equations. Namely, for $\varphi \in \mathcal{C}_c^1([0, H_i^s) \times \mathbb{R}_+)$ and $\psi \in \mathcal{C}_c^1([0, H_i^r) \times \mathbb{R}_+)$,

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_{i,t}\rangle &= \langle \varphi(\cdot,0), \nu_{i,0}\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_t(\cdot,s), \nu_{i,s}\rangle ds \\ &\quad - \int_0^t \langle h_i^s(\cdot)\varphi(\cdot,s), \nu_{i,s}\rangle ds + \int_0^t \varphi(0,s) dK_{i,s},\end{aligned} \tag{33}$$

$$\begin{aligned}\langle \psi(\cdot,t), \eta_{i,t}\rangle &= \langle \psi(\cdot,0), \eta_{i,0}\rangle + \int_0^t \langle \psi_x(\cdot,s) + \psi_t(\cdot,s), \eta_{i,s}\rangle ds \\ &\quad - \int_0^t \langle h_i^r(\cdot)\psi(\cdot,s), \eta_{i,s}\rangle ds + \int_0^t \psi(0,s) dE_{i,s}.\end{aligned} \tag{34}$$

Finally,

$$B_{i,t} = \langle 1, \nu_{i,t}\rangle, \tag{35}$$

$$D_{i,t} = \int_0^t \int_0^\infty h_i^s(x)\nu_{i,s}(dx)ds, \tag{36}$$

$$R_{i,t} = \int_0^t \int_0^\infty h_i^r(x) 1_{\{\eta_{i,s}[0,x] < Q_{i,s}\}} \eta_{i,s}(dx)ds. \tag{37}$$

Equations (25)–(37) are called the *fluid model equations* (FME). A tuple $(B, X, Q, D, K, R, \nu, \eta)$ satisfying these equations is said to be a *solution to the FME with initial conditions* (X_0, ν_0, η_0) *and data* E.

Remark 3.1.

(a) *Uniqueness of solutions to the FME is established in the next subsection. We do not address existence of solutions in this section. However, we will show (in Theorem 4.3 below), under suitable* *assumptions, that fluid-scaled versions of the processes associated with the N-server system do converge weakly to solutions to the FME, by which existence follows.*

(b) *As in the case of the N-server system equations, the evolution of the fluid measures $\nu_{i,t}$ and $\eta_{i,t}$ is due to three sources. First is the motion resulting from the age (resp., waiting times) increasing at unit rate. This accounts for the second term of* (33) *(resp.,* (34)*). The second is due to departures (resp., potential reneging) which correspond to the third term, and finally the last term is due to beginning of new service (resp., new arrivals into the system).*

(c) *Equation* (36) *describes the fluid departure process. Note that $\nu_{i,s}(dx)$ represents the fluid mass of customers with ages in $[x, x+dx)$ at time s and $h_i^s(x)$ represents the rate at which mass with age x departs from the system. Thus $\langle h_i^s, \nu_i \rangle$ gives the departure rate, which explains* (36)*. A similar explanation holds for the reneging process equation* (37)*, except that in this case the indicator of $\{\eta_{i,s}[0,x] < Q_{i,s}\}$ appears. This factor corrects for the fact that η corresponds to the potential, not the actual queue. Fluid mass of customers with waiting time within $[x, x+dx)$, where $\eta_i[0,x] > Q_i$, does not appear in the actual queue, and therefore its fictitious reneging must be deleted.*

We next recall Theorem 4.1 and Remark 4.3 of [14] which we state here as Proposition 3.1. This result establishes a representation of the solution to equations of the form (33) and (34).

Proposition 3.1. (Theorem 4.1 and Remark 4.3 of [14]) *Let G be a cumulative distribution function on $\mathbb{R}_+$ with density function g and hazard rate $h = \frac{g}{1-G}$. Let $H = \sup\{x : G(x) < 1\}$. suppose that $\{\bar{\nu}_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}_F[0,H)}[0,\infty)$ has the property that for every $m \in [0,H)$ and $T \in [0,\infty)$ there exists $C(m,T) < \infty$ such that*

$$\int_0^\infty \langle \varphi(\cdot, s) h(\cdot), \bar{\nu}_s \rangle ds \leq C(m,T) \|\varphi\|_\infty,$$

for every φ on $[0,H) \times \mathbb{R}_+$ continuous with support contained in $[0,m] \times [0,T]$, where $\|\varphi\|_\infty = \sup_{[0,H)\times\mathbb{R}_+} |\varphi|$. Then given any $\nu_0 \in \mathcal{M}_F[0,H)$, z that is locally of bounded variation on $[0,\infty)$ with $z(0) = 0$, one has that $\{\bar{\nu}_t\}_{t\geq 0}$ satisfies the integral equation

$$\langle \varphi(\cdot,t), \bar{\nu}_t \rangle = \langle \varphi(\cdot,0), \nu_0 \rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \bar{\nu}_s \rangle ds - \int_0^t \langle h(\cdot)\varphi(\cdot,s), \bar{\nu}_s \rangle ds + \int_0^t \varphi(0,s) dz(s),$$

for every $\varphi \in \mathcal{C}^{1,1}([0,H) \times \mathbb{R}_+)$ and $t \in [0,\infty)$, if and only if $\{\bar{\nu}_t\}_{t\geq 0}$ satisfies

$$\int_{[0,H)} f(x) \bar{\nu}_t(dx) = \int_{[0,H)} f(x+t) \frac{1 - G(x+t)}{1 - G(x)} \nu_0(dx) + \int_0^t f(t-s)(1 - G(t-s)) dz(s),$$

for every bounded, continuous function f on $[0,H)$ and $t \geq 0$. Moreover, for any bounded differentiable function f on $[0,H)$ and $t \geq 0$,

$$\int_0^t f(t-s)(1-G(t-s)) dz(s) = f(0)z(t) + \int_0^t f'(t-s)(1-G(t-s)) dz(s) - \int_0^t f(t-s) g(t-s) z(s) ds.$$

Applying the above results to $(G_i^s, h_i^s, \nu_{i,0}, \nu_i, K_i)$ and (33) and to $(G_i^r, h_i^r, \eta_{i,0}, \eta_i, E_i)$ and (34) and D_i in (37) we obtain the following.

Proposition 3.2. *Any solution to the FME satisfies the following for $\varphi \in \mathcal{C}_c^1([0,H_i^s)\times\mathbb{R}_+)$ and $\psi \in \mathcal{C}_c^1([0,H_i^r)\times\mathbb{R}_+)$,*

$$\langle\varphi(\cdot,t),\nu_{i,t}\rangle = \int_{[0,\infty)} \frac{1-G_i^s(x+t)}{1-G_i^s(x)}\,\varphi(x+t,t)\nu_{i,0}(dx) + \int_0^t (1-G_i^s(t-s))\varphi(t-s,t)dK_{i,s}, \tag{38}$$

$$\langle\psi(\cdot,t),\eta_{i,t}\rangle = \int_{[0,\infty)} \frac{1-G_i^r(x+t)}{1-G_i^r(x)}\,\psi(x+t,t)\eta_{i,0}(dx) + \int_0^t (1-G_i^r(t-s))\psi(t-s,t)dE_{i,s}, \tag{39}$$

$$D_{i,t} = \int_{[0,\infty)} \frac{G_i^s(x+t)-G_i^s(x)}{1-G_i^s(x)}\,\nu_{i,0}(dx) + \int_0^t g_i^s(t-s)K_{i,s}ds. \tag{40}$$

Proof. Identical to the proof for the case of one class, from [12], [14]. See Theorem 4.2 and Corollary 4.4 in [14] for their proofs. □

Equation (39) uniquely determines η. Indeed $\eta_{i,0}$ is part of the system initial conditions and E is the data. Clearly, a similar statement cannot be made about ν and (38), since K is a part of the solution, rather than the data.

3.2 Uniqueness of solutions

In this subsection we prove uniqueness of solutions to the FME. The proof is based on a representation of Q and K as images of (E,D,R) under a certain continuous mapping involving a two-dimensional Skorohod map. The crux of the argument shows up in the case of two classes ($J=2$) and no reneging, that is presented first. The continuity property is lifted to a general number of classes, by using essentially the same, two-dimensional argument. Uniqueness is addressed (for the full model, including reneging) by combining the continuity property with Proposition 3.2. An additional property (42) regarding the modulus of continuity is proved along the way; it is used in the next section.

The Skorohod problem (SP) of interest is concerned with constraining paths that reside in $\mathbb{R}^2$ to

$$G = \{x \in \mathbb{R}^2 : x_1 \ge 0 \text{ or } x_2 \le 0\},$$

via the fixed constraint direction $d = e_1 - e_2$. Here, $x=(x_1,x_2)\in\mathbb{R}^2$, $e_1=(1,0)$, $e_2=(0,1)$. Denote the interior of the set G by G^o.

Definition 3.1. (The SP (G,d)) *Let $\beta \in \mathcal{D}_{\mathbb{R}^2}(\mathbb{R}_+)$. Then (γ,η), $\gamma\in\mathcal{D}_{\mathbb{R}^2}(\mathbb{R}_+)$, $\eta\in\mathcal{D}_{\mathbb{R}_+}(\mathbb{R}_+)$ are said to solve the SP for β if*

- $\gamma = \beta + (e_1-e_2)\eta$,
- $\gamma_t \in G$ *for all* $t \ge 0$,
- η *is nondecreasing, and* $\int_{[0,\infty)} 1_{\{\gamma_s\in G^o\}}d\eta_s = 0$.

As shown in the appendix, this problem is uniquely solvable. The solution map $\beta\mapsto\gamma$ is denoted throughout by Γ. The solution map $\beta\mapsto(\gamma,\eta)$ is denoted by $\hat{\Gamma}$. The following two properties,

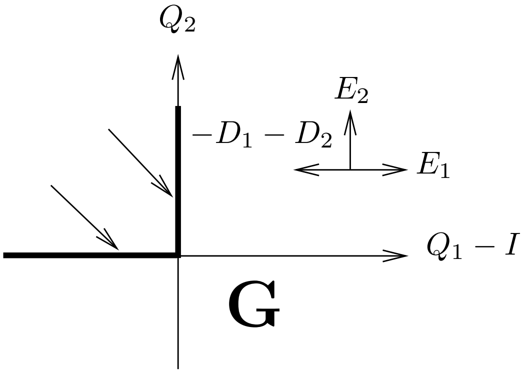

Figure 1: *The dynamics of $\tilde{Q} = (Q_1 - I, Q_2)$. The boundary of the set G is shown (thick line) along with the constraint direction $e_1 - e_2$. When $Q_1 > 0$ (equivalently, $Q_1 - I > 0$), an increase in E_1 (E_2) contributes to an increase in Q_1 (Q_2), while an increase in either D_1 or D_2 decreases Q_1.*

crucial to our treatment, are shown in Proposition A.1 in the appendix:

There exists a constant c such that

$$\|\gamma - \bar{\gamma}\|_t \le c\|\beta - \bar{\beta}\|_t, \quad t \ge 0, \quad \text{whenever } \gamma = \Gamma(\beta) \text{ and } \bar{\gamma} = \Gamma(\bar{\beta}), \tag{41}$$

$$w(\gamma, \theta, t) \le cw(\beta, \theta, t), \quad \theta, t > 0, \quad \text{whenever } \gamma = \Gamma(\beta). \tag{42}$$

The model without reneging is obtained by setting h_i^r and R_i to zero (of course, equation (34) then becomes redundant). Consider the model without reneging, with two classes ($J = 2$). Given a solution to the FME, denote

$$\tilde{Q} = (Q_1 - I, Q_2). \tag{43}$$

Note, as an immediate consequence of (28) and (30), that

$$Q = \tilde{Q}^+, \tag{44}$$

where for $x = (x_1, x_2) \in \mathbb{R}^2$ we write x^+ for (x_1^+, x_2^+). The main observation regarding the SP is the following fact, involving only equations (26)–(32). As a convention, for $H = B, X, Q, D$ or K and $\bar{H} = \bar{B}, \bar{X}, \bar{Q}, \bar{D}$ or $\bar{K}$, respectively, (possibly corresponding to solutions with different data, say E and $\bar{E}$), we write ΔH for $H - \bar{H}$ (as well as $\Delta E = E - \bar{E}$).

Lemma 3.1. *Consider the model without reneging, with $J = 2$. Given $E \in \mathcal{D}^+_{\mathbb{R}^2}(\mathbb{R}_+)$, let the tuple (B, X, Q, D, K) satisfy equations (26)–(32). Then $(\tilde{Q}, K_2)$ solve the SP for*

$$\tilde{E} := \tilde{Q}_0 + E - (D_1 + D_2, 0). \tag{45}$$

As a consequence, given E and $\bar{E}$, the corresponding solutions (with common initial conditions) satisfy

$$\|\Delta Q\|_t \le c(\|\Delta E\|_t + \|\Delta D\|_t), \qquad t > 0. \tag{46}$$

Moreover,

$$w(Q,\theta,t) \le c(w(E,\theta,t) + w(D,\theta,t)), \qquad t,\theta > 0. \tag{47}$$

Proof. Verifying the first bullet in the definition of the SP amounts to showing

$$Q_1 - I = Q_{1,0} - I_0 + E_1 - D_1 - D_2 + K_2, \qquad \text{and} \qquad Q_2 = Q_{2,0} + E_2 - K_2. \tag{48}$$

For the first equality, note by (26)–(28) (recalling we have set $R = 0$), that

$$Q_{1,0} - I_0 + E_1 - D_1 - D_2 + K_2 = Q_{1,0} - 1 + B_{1,0} + B_{2,0} + X_1 - X_{1,0} + B_2 - B_{2,0} = Q_1 - 1 + B_1 + B_2.$$

The second statement in (48) follows similarly.

That $\tilde{Q}$ resides in G will be shown by arguing that, for all $t \ge 0$,

$$\tilde{Q}_t \in \tilde{G} \subset G, \qquad t \ge 0,$$

where

$$\tilde{G} = \{(x_1, x_2) \in G : x_2 \ge 0\}.$$

Since $\tilde{Q}_2$ is nonnegative (see (43) and (29)), it suffices to show that $\tilde{Q}_2(t) = Q_2(t) > 0$ implies $\tilde{Q}_1(t) = Q_1(t) - I_t \ge 0$. By (28) and (30), if $Q_2(t) > 0$ then indeed $I_t = 1 - \sum_i B_{i,t} = 0$. This shows $\tilde{Q}_t \in \tilde{G}$ for all t.

Finally, K_2 is clearly nonnegative and nondecreasing by (31). Moreover, since $\tilde{Q}_t \in \tilde{G}$, the condition $\tilde{Q}_t \in G^o$ implies $\tilde{Q}_1(t) > 0$, and, in turn, $Q_1(t) > 0$. Hence

$$\int 1_{\{\tilde{Q}_s \in G^o\}} dK_{2,s} \le \int 1_{\{Q_{1,s} > 0\}} dK_{2,s} = 0,$$

by (32). This completes the proof of the first assertion.

The second and third assertions follow from the first on using (41), (42) and (44). □

Remark 3.2. *A review of the proof shows that the nonnegativity and the nondecreasing property of* E *are not used. Thus the result continues to hold when* $E, \bar{E} \in \mathcal{D}_{\mathbb{R}^2}(\mathbb{R}_+)$*. This observation will be used when the model with reneging is considered.*

We next argue that for general number of classes J, results similar to Lemma 3.1 continue to hold. To this end, fix $i_0 \in \{2, 3, \dots, J\}$ and write

$$B^{(1)} = \sum_{j=1}^{i_0 - 1} B_j, \qquad B^{(2)} = \sum_{j=i_0}^{J} B_j, \tag{49}$$

with a similar convention for X, Q, D, K and E. The key point that allows reducing the problem to a two-dimensional one is this. Given any solution $(B_i, X_i, Q_i, D_i, K_i, E_i)$, $i = 1, 2, \dots, J$ to (26)–(32), the quantities $(B^{(i)}, X^{(i)}, Q^{(i)}, D^{(i)}, K^{(i)}, E^{(i)})$, $i = 1, 2$ satisfy precisely the same relations. As a result, Lemma 3.1 is applicable. Since i_0 is arbitrary, we conclude that there exists a constant c_1, such that whenever (B, X, Q, D, K) and $(\bar{B}, \bar{X}, \bar{Q}, \bar{D}, \bar{K})$ are two solutions corresponding to some E and $\bar{E}$,

$$\|Q - \bar{Q}\|_t \le c_1(\|E - \bar{E}\|_t + \|D - \bar{D}\|_t), \qquad t > 0, \tag{50}$$

and

$$w(Q,\theta,t) \le c_1(w(E,\theta,t) + w(D,\theta,t)), \qquad t,\theta > 0. \tag{51}$$

Finally, for the full model ($J \ge 2$, with reneging) we have

Proposition 3.3. *Given E and $\bar{E}$ in $\mathcal{D}^+_{\mathbb{R}^J}(\mathbb{R}_+)$, let*

$$S = (B, X, Q, D, K, R) \quad \textit{and} \quad \bar{S} = (\bar{B}, \bar{X}, \bar{Q}, \bar{D}, \bar{K}, \bar{R})$$

be corresponding solutions (with common initial conditions) to equations (26)–(32)*. Then*

$$\|\Delta Q\|_t \le c_1(\|\Delta E\|_t + \|\Delta D\|_t + \|\Delta R\|_t), \qquad t > 0, \tag{52}$$

and

$$w(Q,\theta,t) \le c_1(w(E,\theta,t) + w(D,\theta,t) + w(R,\theta,t)), \qquad t,\theta > 0. \tag{53}$$

Proof. This follows from (50) and (51) upon replacing E by $E - R$, and recalling Remark 3.2 by which the data need not be nondecreasing. □

We can now prove

Theorem 3.1. *Assume h^r_i are bounded. Let S and $\bar{S}$ be two solutions to the FME* (25)–(37)*, corresponding to the same initial conditions and the same data E. Then $S = \bar{S}$.*

Proof. The structure of the FME is such that given any $T > 0$ and a solution $\{S_t\}_{t\ge0}$ corresponding to data $\{E_t\}_{t\ge0}$ (and some initial condition), $\{S_{T+t}\}_{t\ge0}$ is a solution corresponding to the data $\{E_{T+t} - E_T\}_{t\ge0}$ and initial condition (X_T, ν_T, η_T). Therefore, by the usual argument by contradiction, it suffices to prove that uniqueness holds over $[0,T]$ for however small $T > 0$.

By (26)–(28), $\Delta K = -\Delta Q - \Delta R$. Thus by Proposition 3.3, with $c_2 = c_1 + 1$,

$$\|\Delta K\|_t \le c_2(\|\Delta D\|_t + \|\Delta R\|_t), \qquad t \ge 0. \tag{54}$$

By (40),

$$\Delta D_{i,t} = \int_0^t g^s_i(t-s)\Delta K_{i,s}ds,$$

so

$$\|\Delta D\|_t \le \sum_i \int_0^t g^s_i(s)ds\, \|\Delta K_i\|_t \le \frac{1}{4c_2}\|\Delta K\|_t,$$

provided $t > 0$ is sufficiently small. Hence

$$\|\Delta D\|_t \le \frac{1}{4}(\|\Delta D\|_t + \|\Delta R\|_t).$$

Next, by (37), if c_3 is an upper bound on h^r_i,

$$|\Delta R_{i,t}| \le c_3 \int_0^t |\Delta Q_{i,s}|ds.$$

Thus, making $t > 0$ even smaller if necessary, we have

$$\|\Delta R\|_t \leq \frac{1}{4}(\|\Delta D\|_t + \|\Delta R\|_t).$$

As a result, for some $t > 0$,

$$\|\Delta D\|_t + \|\Delta R\|_t \leq \frac{1}{2}(\|\Delta D\|_t + \|\Delta R\|_t).$$

Thus $\Delta D = \Delta R = 0$ on $[0,t]$; by (52) and (54) a similar conclusion holds for ΔQ and ΔK. Finally, by (38), $\nu = \bar{\nu}$ on $[0,t]$. This completes the proof. □

We end this subsection with two properties of the FME not directly related to uniqueness (but used later in Section 4), regarding the two-dimensional versions of the form (49). Recall the map $\hat{\Gamma}$ defined in the paragraph following Definition 3.1.

Lemma 3.2. *Let data $E \in \mathcal{D}^+_{\mathbb{R}^J}(\mathbb{R}_+)$ be given.*

i. Let $S = (B, X, Q, D, K, R)$ be the corresponding solution to (26)–(32). Fix $i_0 \in \{2, 3, \dots, J\}$ and consider

$$(B^{(i)}, X^{(i)}, Q^{(i)}, D^{(i)}, K^{(i)}, R^{(i)}, E^{(i)}), \quad i = 1, 2,$$

defined as in (49) and in the discussion that follows (with the additional component $R^{(i)}$ defined similarly). Define

$$\hat{Q} = (Q^{(1)} - I, Q^{(2)})$$

(in analogy with (43)). Define

$$\hat{E} := \hat{Q}(0) + (E^{(1)}, E^{(2)}) - (R^{(1)}, R^{(2)}) - (D^{(1)} + D^{(2)}, 0)$$

(in analogy with (45), but taking into account R). Then $(\hat{Q}, K^{(2)}) = \hat{\Gamma}[\hat{E}]$, namely, $(\hat{Q}, K^{(2)})$ solve the SP for $\hat{E}$. (The transformation we have just defined from (B, X, Q, D, K, R) to $(\hat{Q}, K^{(2)}, \hat{E})$ will be denoted by Θ.)

ii. Let now $S = (B, X, Q, D, K, R)$ satisfy (26)–(31) (i.e., not including relation (32)). Assume, moreover, that for every $i_0 \in \{2, 3, \dots, J\}$,

$$\sum_{i=i_0}^{J} K_{i,t} = \int_{[0,t]} 1_{\{\sum_{j=1}^{i_0-1} Q_{j,s}=0\}} d\Big(\sum_{i=i_0}^{J} K_{i,s}\Big), \qquad t \geq 0. \tag{55}$$

Then S satisfies (32).

Proof. i. This follows from Lemma 3.1, the discussion following Remark 3.2 and considering $E - R$ in place of E.

ii. Owing to the nonnegativity of Q_j, the assumed condition (55) is equivalent to

$$\int_{[0,t]} 1_{\{\sum_{j=1}^{i_0-1} Q_{j,s}>0\}} d\Big(\sum_{i=i_0}^{J} K_{i,s}\Big) = 0, \qquad t \geq 0.$$

Thus

$$\int_{[0,t]} 1_{\{\sum_{j=1}^{i_0-1} Q_{j,s}>0\}} dK_{i_0,s} = 0, \qquad t \geq 0.$$

Since this holds for every $i_0 \in \{2, 3, \dots, J\}$, (32) follows. □

3.3 Some properties of the solution

We show that the entrance-into-service can be represented in terms of the entrance and departure (processes) in a way that reflects the priority discipline.

Theorem 3.2. *Assume that for every i, E_i is nondecreasing and absolutely continuous, and denote $\lambda_i(t) = \frac{d}{dt}E_i(t)$. Denote $\delta(t) = \frac{d}{dt}\sum_{i=1}^J D_{i,t} = \sum_{i=1}^J \langle h_i^s, \nu_{i,t}\rangle$ (see (36)). Then K_i are absolutely continuous, and the derivatives κ_i satisfy a.e., for $j = 1, 2, \dots, J$,*

$$\sum_{i=1}^j \kappa_i(t) = \begin{cases} \delta(t) & \textit{if} \quad \sum_{i=1}^j Q_{i,t} > 0, \\ \delta(t) \wedge \sum_{i=1}^j \lambda_i(t) & \textit{if} \quad \sum_{i=1}^j Q_{i,t} = 0,\ \sum_{i=1}^J B_{i,t} = 1, \\ \sum_{i=1}^j \lambda_i(t) & \textit{if} \quad \sum_{i=1}^J B_{i,t} < 1. \end{cases}$$

The second entry in the above formula corresponds to the case where the system is critically loaded, namely all servers are busy and all queues are empty. The rate at which mass is sent to service is then the minimum between the rate of arrival and the rate at which servers become available, as one intuitively might guess. However, as shown in the proof below, it is legitimate to replace the expression $\delta(t) \wedge \sum_{i=1}^j \lambda_i(t)$ by the simpler one $\sum_{i=1}^j \lambda_i(t)$.

Proof. Since E_i are absolutely continuous, so are X_i by (27), (36) and (37). As a result, so is $(1 - \sum_{i=1}^J X_i)^+$. In view of (30) and (26), one has that $\sum_{i=1}^J B_i$, and, in turn, $\sum_{i=1}^J K_i$ are absolutely continuous. But since K_i are nondecreasing (31), it follows that each K_i must be absolutely continuous. Denote by κ_i the corresponding densities.

If $\sum_{i=1}^J B_{i,t} < 1$ for some t, then by the work conservation condition (30), $\sum_{i=1}^J X_{i,t} < 1$, and by the continuity of the latter in t, this holds on a neighborhood of t. In such a neighborhood, it is seen, by combining (28) and (30), that $Q_i = 0$, and by (37), that R_i do not increase. Hence using (26), (27) and (28), for s in a neighborhood of t,

$$K_{i,s} - K_{i,t} = Q_{i,s} - Q_{i,t} + E_{i,s} - E_{i,t} = E_{i,s} - E_{i,t}.$$

This shows $\kappa_i(t) = \lambda_i(t)$, for all i.

On the other hand, if $\sum_{i=1}^j Q_{i,t} > 0$, then the same is true in a neighborhood, by continuity of Q_i (which follows from (26), (27) and (28), using the continuity of K_i, E_i and R_i). By (32), K_i remains constant on any such interval, for all $i \geq j+1$. Moreover, using (28) and (30), $\sum_{i=1}^J B_i$ is equal to one. Hence for s in a neighborhood of t,

$$\begin{aligned} \sum_{i=1}^j (K_{i,s} - K_{i,t}) = \sum_{i=1}^J (K_{i,s} - K_{i,t}) &= \sum_{i=1}^J (B_{i,s} - B_{i,t} + D_{i,s} - D_{i,t}) \\ &= \sum_{i=1}^J (D_{i,s} - D_{i,t}), \end{aligned}$$

where we used (26) for the second equality. This shows $\sum_{i=1}^j \kappa_i(t) = \delta(t)$ if $\sum_{i=1}^j Q_{i,t} > 0$.

Finally, since $\sum_{i=1}^{J} B_i$ and $\sum_{i=1}^{j} Q_i$ are absolutely continuous, it follows that $\frac{d}{dt}\sum_{i=1}^{J} B_i = 0$ a.e. on $A_1 := \{t : \sum_{i=1}^{J} B_{i,t} = 1\}$ and $\frac{d}{dt}\sum_{i=1}^{j} Q_{i,t} = 0$ a.e. on $A_2 := \{t : \sum_{i=1}^{j} Q_{i,t} = 0\}$ [8, Theorem A.6.3]. But

$$\frac{d}{dt}\sum_{i=1}^{J} B_i = \sum_{i=1}^{J} \kappa_i - \delta,$$

and

$$\frac{d}{dt}\sum_{i=1}^{j} Q_i = \sum_{i=1}^{j} (\lambda_i - \kappa_i - \frac{d}{dt}R_i).$$

Note by (37), that a.e. on $\{t : Q_i(t) = 0\}$, $\frac{d}{dt}R_i = 0$. Thus a.e. on $A = A_1 \cap A_2$, we have $\sum_{i=1}^{j} \kappa_i = \sum_{i=1}^{j} \lambda_i$ and $\sum_{i=1}^{J} \kappa_i = \delta$. Hence a.e. on A, $\sum_{i=1}^{j} \kappa_i = \sum_{i=1}^{j} \lambda_i = \delta \wedge \sum_{i=1}^{j} \lambda_i$. □

3.4 Characterization of the invariant state

We now consider the case where, for all i, $E_i(t) = \lambda_i t$ for $t \geq 0$, where $\lambda_i > 0$ are constants. Recall that $\mu_i \in (0, \infty)$ denote the reciprocal expected service times, that is,

$$\frac{1}{\mu_i} = \int_0^\infty (1 - G_i^s(x))dx, \qquad i = 1, \dots, J.$$

For each i, let $\rho_i = \lambda_i/\mu_i$.

A tuple $\Sigma_0 = (X_0, \nu_0, \eta_0)$ is said to be an *invariant state* if any solution

$$S = (B, X, Q, D, K, R, \nu, \eta)$$

to the FME with initial condition Σ_0, satisfies $(X(t), \nu(t), \eta(t)) = (X_0, \nu_0, \eta_0)$ for all $t \geq 0$. If Σ_0 is an invariant state and S is the corresponding solution then $B_{i,0} := B_i(0) = \langle 1, \nu_{i,0} \rangle$ and $Q_{i,0} := Q_i(0) = X_{i,0} - B_{i,0}$, as dictated by (35) and (28).

Denote

$$L = \inf\Big\{j : \sum_{i=1}^{j} \rho_i \geq 1\Big\}. \tag{56}$$

Theorem 3.3. *Let the hypotheses of Theorems 3.1 and 3.2 hold, and suppose that G_L^r is strictly increasing in $[0, H_L^r)$. Then there exists a unique invariant, given as follows.*

i. $\eta_{i,0}(dx) = \lambda_i(1 - G_i^r(x))dx =: \eta_{i,*}(dx)$.

ii. If $\sum_{i=1}^{J} \rho_i \leq 1$ *then* $\nu_{i,0}(dx) = \lambda_i(1 - G_i^s(x))dx$, $X_{i,0} = \langle 1, \nu_{i,0} \rangle$, $Q_{i,0} = 0$ *for all* i.

iii. If $\sum_{i=1}^{J} \rho_i > 1$, *let* $\hat{\rho} = \sum_{i=1}^{L-1} \rho_i < 1$ *(note that $L \leq J$ in this case). Then*

$$\nu_{i,0}(dx) = \lambda_i(1 - G_i^s(x))dx, \qquad i = 1, 2, \dots, L-1, \tag{57}$$

$$\nu_{L,0}(dx) = \mu_L(1 - \hat{\rho})(1 - G_L^s(x))dx, \tag{58}$$

$$\nu_{i,0}(dx) = 0, \qquad i > L. \tag{59}$$

$X_{i,0} = \langle 1, \nu_{i,0} \rangle = \rho_i$ for $i \leq L-1$. $X_{L,0} = \langle 1, \nu_{L,0} \rangle + Q_{L,0} = 1 - \hat{\rho} + b$, where $b > 0$ is uniquely determined (owing to the strict monotonicity of G_L^r) by

$$G_L^r(\chi_L(b)) = \frac{\sum_{i=1}^L \rho_i - 1}{\rho_L}, \tag{60}$$

$$\chi_L(y) = \inf\{x : \eta_{L,*}[0,x] \geq y\}.$$

Finally, for $i > L$, one has $X_{i,0} = Q_{i,0} \geq 0$, $R_i(t) = \lambda_i t$, $K_i(t) = 0$, and

$$Q_{i,0} = \lambda_i \int_0^\infty (1 - G_i^r(x))dx. \tag{61}$$

Proof. First we show that any invariant state satisfies assertions (i)–(iii) above. Suppose that (X_0, ν_0, η_0) is an invariant state. Since $X(t) = X_0$, $\nu(t) = \nu_0$, it follows that $B(t) = B_0$ and $Q(t) = Q_0$, $t \geq 0$. In addition, by Proposition 3.2, for $f \in C_b[0,\infty)$, $\langle f, \eta_{i,0} \rangle < \infty$,

$$\langle f, \eta_{i,0} \rangle = \int_{[0,\infty)} \frac{1 - G_i^r(x+t)}{1 - G_i^r(x)} f(x+t)\eta_{i,0}(dx) + \lambda_i \int_0^t (1 - G_i^r(t-s))f(t-s)ds, \qquad t \geq 0.$$

As $t \to \infty$, the first integral converges to zero by dominated convergence, and the second converges to $\lambda_i \int_0^\infty (1 - G_i^r(u))f(u)du$. Thus

$$\eta_{i,0}(dx) = \lambda_i(1 - G_i^r(x))dx = \eta_{i,*}(dx).$$

In addition,

$$\int_0^\infty h_i^r(x) 1_{\{\eta_{i,s}[0,x] < Q_{i,s}\}} \eta_{i,s}(dx) = \int_0^\infty h_i^r(x) 1_{\{\eta_{i,0}[0,x] < Q_{i,0}\}} \eta_{i,0}(dx) =: p_i, \tag{62}$$

so that, by (37),

$$R_i(t) = p_i t.$$

Owing to the strict monotonicity of G_L^r on $[0, H_L^r)$,

$$Q_{L,0} > 0 \quad \text{implies} \quad p_L > 0. \tag{63}$$

By (26), (27) and (28),

$$K_{i,t} = Q_{i,0} - Q_{i,t} + E_{i,t} - R_{i,t} = (\lambda_i - p_i)t.$$

It follows, again by Proposition 3.2, that

$$\langle f, \nu_{i,0} \rangle = \int_{[0,\infty)} \frac{1 - G_i^s(x+t)}{1 - G_i^s(x)} f(x+t)\nu_{i,0}(dx) + (\lambda_i - p_i) \int_0^t (1 - G_i^s(t-s))f(t-s)ds,$$

which converges, as $t \to \infty$, to $(\lambda_i - p_i) \int_0^\infty f(u)(1 - G_i^s(u))du$. Hence

$$\nu_{i,0}(dx) = (\lambda_i - p_i)(1 - G_i^s(x))dx = (\lambda_i - p_i)\nu_{i,*}(dx).$$

Let us show that for all $j < L$, $Q_{j,0} = 0$. Arguing by contradiction, assume $Q_{1,0}+\cdots+Q_{j,0} > 0$, for some $j < L$. Then by Theorem 3.2,

$$\kappa_1(t) + \cdots + \kappa_j(t) = \delta(t) = \sum_{i=1}^{J} \langle h_i^s, \nu_{i,*}\rangle = \sum_{i=1}^{J} (\lambda_i - p_i),$$

and $\kappa_{j+1}(t) = \cdots = \kappa_J(t) = 0$. This implies that for $i > j$, $\lambda_i = p_i$, so that $\nu_{i,0} = 0$. But

$$\sum_{i=1}^{J} B_{i,0} = \sum_{i=1}^{J} \langle 1, \nu_{i,0}\rangle = \sum_{i=1}^{j} \langle 1, \nu_{i,0}\rangle = \sum_{i=1}^{j} \frac{\lambda_i - p_i}{\mu_i} \le \sum_{i=1}^{j} \frac{\lambda_i}{\mu_i} < 1.$$

Due to the work conservation condition (30), this contradicts the assumption $\sum_{i=1}^{j} Q_{i,0} > 0$. This shows $Q_{j,0} = 0$ for all $j < L$. If $\sum_{j=1}^{L} \rho_j = 1$ and $Q_{L,0} > 0$, then by (63) $p_L > 0$, and this, together with $\kappa_{L+1}(t) = \cdots = \kappa_J(t) = 0$, implies that

$$\sum_{i=1}^{J} B_{i,0} = \sum_{i=1}^{J} \langle 1, \nu_{i,0}\rangle = \sum_{i=1}^{L} \frac{\lambda_i - p_i}{\mu_i} < \sum_{i=1}^{L} \frac{\lambda_i}{\mu_i} = 1,$$

which, again, contradicts the work conservation assumption.

As a result, for $j < L$, $R_j(t) = 0$, $p_j = 0$, and thus $\kappa_j(t) = \lambda_j$, and if $\sum_{j=1}^{L} \rho_j = 1$ then also $Q_{L,0} = 0, \quad R_L(t) = 0, \quad p_L = 0$ and $\kappa_L(t) = \lambda_L$.

So assume from now on that $\sum_{j=1}^{L} \rho_j > 1$. Suppose that $Q_{L,0} = 0$. Then $R_L(t) = 0$ and $p_L = 0$ so that $\nu_{L,0} = \lambda_L \nu_{L,*}$ and

$$\sum_{i=1}^{J} \langle 1, \nu_{i,0}\rangle \ge \sum_{i=1}^{L} \langle 1, \nu_{i,0}\rangle = \sum_{i=1}^{L} \lambda_i \int_0^\infty (1 - G_i^s(x))dx = \sum_{i=1}^{L} \frac{\lambda_i}{\mu_i} > 1,$$

which is impossible. Thus $Q_{L,0} > 0$ so that $\sum_{i=1}^{J} \langle 1, \nu_{i,0}\rangle = 1$. It follows by Theorem 3.2 that $K_i(t) = 0$ for $i \ge L+1$ and therefore $\nu_{i,0} = 0$ for $i \ge L+1$ and

$$\begin{aligned} 1 = \sum_{i=1}^{J} \langle 1, \nu_{i,0}\rangle = \sum_{i=1}^{L} \langle 1, \nu_{i,0}\rangle &= \sum_{i=1}^{L-1} \lambda_i \int_0^\infty (1 - G_i^s(x))dx + (\lambda_L - p_L) \int_0^\infty (1 - G_L^s(x))dx \\ &= \sum_{i=1}^{L-1} \frac{\lambda_i}{\mu_i} + \frac{\lambda_L - p_L}{\mu_L}. \end{aligned}$$

Hence

$$p_L = \Big(\sum_{i=1}^{L} \frac{\lambda_i}{\mu_i} - 1 \Big) \mu_L.$$

Since by its definition in (62) for all j, $p_j = \lambda_j G_j^r(\chi_j(Q_{j,0}))$, it follows that $Q_{L,0} = b$ where b is such that $G_L^r(\chi_L(b)) = \frac{\sum_{i=1}^{L} \rho_i - 1}{\rho_L}$.

For $j \ge L+1$, since $\nu_{j,0} = 0$, we have $X_{j,0} = Q_{j,0}$. But $p_j = \lambda_j G_j^r(\chi_j(Q_{j,0}))$ so that

$$\chi_j(Q_{j,0}) = \inf\{x : G_j^r(x) = 1\} = H_j^r.$$

That is

$$\inf \left\{ y : \lambda_j \int_0^y (1 - G_j^r(u))du \geq Q_{j,0} \right\} = H_j^r.$$

Hence $Q_{j,0} = \lambda_j \int_0^{H_j^r} (1 - G_j^r(u))du$.

We have thus shown that any invariant state satisfies (i)–(iii) above. It can be easily checked using similar calculations that the tuple (X_0, ν_0, η_0) specified by (i)–(iii) is an invariant state for the FME. □

4 Convergence of scaled processes

The goal of this section is to argue that the processes underlying the N-server model, normalized in fluid scale, converge to the corresponding quantities of the fluid model, both on a finite time interval and in stationarity. Given our treatment from Section 3, the main results presented here follow almost immediately from those of [14], [12], [13].

4.1 The N-server system as a Markov process

For $i = 1, \dots, J$ let

$$\alpha_i^N(t) = \inf\{s > t : E_i^N(s) > E_i^N(t)\} - t,$$

be the forward recurrence time at time t of the arrival process of class-i customers. Consider

$$Y^N = (\alpha^N, X^N, \nu^N, \eta^N) = \{\alpha_i^N(t), X_i^N(t), \nu_i^N(t), \eta_i^N(t) : i = 1, \dots, J,\, t \geq 0\}.$$

This process takes values in $\mathcal{Y} = \mathbb{R}_+^J \times \mathbb{N}^J \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H_i^s) \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H_i^r)$, where we recall that, for $0 \leq a < b \leq \infty$, $\mathcal{M}_F[a, b)$ is the space of finite measures on the measurable sets of $[a, b)$ and let $\mathcal{M}_D$ be the subset of $\mathcal{M}_F$ which consists of measures of the form $\sum_{1=1}^{l} \delta_{x_i}$ where δ_x is a point mass at x ($x \in [a, b)$). If we endow $\mathbb{R}_+$ with the Euclidean topology, $\mathbb{N}$ with the discrete topology, and $\mathcal{M}_D[a, b)$ with the weak topology, then $\mathcal{Y}$, endowed with the product topology, is a Polish space.

We hereafter consider Y^N as a stochastic process over $t \geq 0$, with initial conditions $Y^N(0)$. As implied by the system description in Section 2, it is assumed that $Y^N(0)$ is independent of $\{r_{i,j}, v_{i,j}, e_{i,j+1}^N - e_{i,j}^N,\, i = 1, \dots, J,\, j \geq 1\}$, namely the patience and service times of future arrivals, as well as their inter-arrival times.

Theorem 4.1. *Y^N is a strong Markov process on the state space $\mathcal{Y}$.*

Proof. The proof of this theorem follows along the lines of Appendices A and B of [12] for the one-class model, relying on the process being a *piecewise deterministic Markov process* (as defined, eg., in [10]). The construction of the process from the model primitives in our case is slightly more involved because of the priority classes, and differs due to the component α^N being the forward rather than backward recurrence time, but is quite straightforward, and we have therefore chosen to omit it here. Also, as in [12], the deterministic functions that govern the process between its jumps are continuous. Thus the strong Markov property then follows by Theorem 7.5.1 of [10]. □

From here to the end of the next subsection we fix N, consider the Markov process Y^N, and suppress N from our notation. We denote by $\mathbb{P}_y$ its law given $Y(0) = y$ and, given any probability measure μ over $\mathcal{Y}$, let $\mathbb{P}_\mu = \int \mathbb{P}_x(\cdot)\mu(dx)$. We denote by $\mathbb{E}_y$ and $\mathbb{E}_\mu$ the corresponding expectations.

For each bounded measurable function ψ on $\mathcal{Y}$ and $\lambda > 0$, let $U^\lambda\psi(y) = \mathbb{E}_y[\int_0^\infty e^{-\lambda t}\psi(Y_t)dt]$ denote the λ-potential of the process Y applied to the function ψ.

The state space contains points that we call 'special'. These are points $y = (\alpha, X, \nu, \eta) \in \mathcal{Y}$, having the property $\alpha_i = \alpha_j$ for some two distinct indices i and j. When starting from such a point, customers of two classes are scheduled to arrive at the same time. Write $\mathcal{Y}_s$ for the set of special points and $\mathcal{Y}_s^c = \mathcal{Y} \setminus \mathcal{Y}_s$.

Assumption 4.1. *For each i, the class-i inter-arrival distribution has a density.*

Lemma 4.1. *Let Assumption 4.1 hold. Then for each bounded, continuous function ψ on $\mathcal{Y}$, the function $y \mapsto U^\lambda\psi(y)$ is continuous at any $y \in \mathcal{Y}_s^c$.*

Proof. For e_λ an exponentially distributed random variable with parameter λ which is independent of Y, we can write

$$U^\lambda\psi(y) = \frac{1}{\lambda}\mathbb{E}_y\psi(Y(e_\lambda)).$$

Suppose that $y^m \to y^0 \in \mathcal{Y}_s^c$ as $m \to \infty$, that Y^m (respectively Y^0) is the process Y that starts at time 0 at y^m (respectively y^0).

For each $m \in \mathbb{Z}_+$, let Y^m be the state descriptor of a multi-class N-server queue with initial state

$$Y^m(0) = y^m = \Big(\alpha^m, x^{m,i}, \sum_{j=1}^{k^{m,i}} \delta_{u_j^{m,i}}, \sum_{j=1}^{l^{m,i}} \delta_{z_j^{m,i}}, i = 1, \ldots, J\Big) \in \mathcal{Y},$$

for some $k^{m,i} \in \{0, \ldots, N\}$, $l^{m,i} \in \mathbb{N}$, $i = 1, \ldots, J$. Suppose that all $\{Y^m : m \in \mathbb{Z}_+\}$ are defined on the same probability space constructed using all the inter-arrival, service and patience times as primitives. Suppose further that $y^m \to y^0$. This immediately implies that $x^{m,i} = x^{0,i}$, $k^{m,i} = k^{0,i}$, $l^{m,i} = l^{0,i}$ for m sufficiently large and that $\alpha^{m,i} \to \alpha^{0,i}$, $u_j^{m,i} \to u_j^{0,i}$, $z_{j'}^{m,i} \to z_{j'}^{0,i}$ for $0 \leq j \leq k^{0,i}$, $0 \leq j' \leq l^{0,i}$. We may assume without loss of generality that $k^{m,i} = k^{0,i}$, $l^{m,i} = l^{0,i}$, $x^{m,i} = x^{0,i}$, that the residual service the j-th customer of class i in service has the density $\frac{g^{s,i}(u_j^{m,i}+t)}{1-G^{s,i}(u_j^{m,i})}$ and the residual patience of the j'-th customer of the class i in the queue is $\frac{g^{r,i}(z_j^{m,i}+t)}{1-G^{r,i}(z_j^{m,i})}$. It is further assumed that

- all inter-arrivals after the first one, $\alpha^{m,i}$, are identical for each N-server process Y^m.
- all service times of customers that arrive after time zero are identical for each N-server process Y^m.

Since $z_j^{m,i} \to z_j^{0,i}$ and $u_j^{m,i} \to u_j^{0,i}$, it follows that the remaining patience of customers in the queue at time 0 and the remaining service times of customers in service at time 0 converge in distribution to those associated with $z_j^{0,i}$ and $u_j^{0,i}$. Since we are looking for the convergence of $U^\lambda\psi(y^m)$ to $U^\lambda\psi(y^0)$ which is the convergence of $\mathbb{E}_{y^m}(\psi(e_\lambda))$ to $\mathbb{E}_{y^0}(\psi(e_\lambda))$ we may as well assume that

- for each $j' = 1, \ldots, l^{0,i}$ the remaining patience times of the customer associated with $\delta_{z^{m,i}_{j'}}$ converges almost surely as $m \to \infty$ to the remaining patience time associated with the point mass $\delta_{z^{0,i}_{j'}}$ and for $j = 1, \ldots, k^{0,i}$ the remaining service time of the customer associated with $\delta_{u^{m,i}_j}$ converges almost surely to the remaining service time associated with the point mass $\delta_{u^{0,i}_j}$ (by using the Skorohod representation theorem).

Observe that for $y^0 \in \mathcal{Y}_s$, two customers of different classes arrive at the same time $\alpha^{0,i} = \alpha^{0,i'}$. If the number of busy servers at that point of time is equal to $N-1$, say, then it is possible that, for a subsequence m' along which $\alpha^{m',i} < \alpha^{m',i'}$ for m' large, customer from class i will go into service at time $\alpha^{m',i}$, and for another sequence m'', having $\alpha^{m'',i} > \alpha^{m'',i'}$ the opposite will occur. This may cause $\{Y^m(\alpha^{m,i})\}$ to have two limit points, which may in turn cause two different limits of $Y^m(\alpha^{m,i}+t)$. To avoid this possibility we have assumed that $y^0 \in \mathcal{Y}_s^c$. Since we have assumed that all inter-arrival time distributions have densities, the following event has probability zero, namely that the arrivals of two customers of different classes, at least one of which is not a first arrival after time 0, coincide. Further, Lemma 4.2 of [13] and the fact that the inter-arrival times and service times have densities, exclude the possibility that arrivals and departures will coincide. Those proofs carry over with no change to our situation of multi-class queues and via the same argument one can prove that arrivals and reneging do not coincide when the patience times have densities.

Let $\{\tau_n^m : n = 1, 2, \ldots\}$ be the jump times of Y^m and $\{\tau_n^0 : n = 1, 2, \ldots\}$ be the jump times of Y^0. Since ψ is bounded, by dominated convergence it suffices to show that $\psi(Y^m(e_\lambda)) \to \psi(Y^0(e_\lambda))$ a.s. Since e_λ is an exponential r.v., independent of the processes Y^m, $m = 1, 2, \ldots$, and Y^0, it suffices to show that

$$Y^m(t) \to Y^0(t) \text{ for every } t, \text{ a.s.} \tag{64}$$

Combining now the facts that

- the deterministic functions that govern the motion between the jumps are all continuous functions on $\mathcal{Y}$,
- for all t, $Y^0(t) \in \mathcal{Y}_s^c$ as we have explained above,
- no arrivals of two customers (for the process Y^m), beyond possibly those at $\alpha^{m,i}$, coincide.

One can now use the same argument as the one used in Lemma 4.1 of [13] to prove that for each $i \in \mathbb{N}$, $\tau_n^m \to \tau_n^0$, and $Y^m(\tau_n^m) \to Y^0(\tau_n^0)$. Finally, if t is not a jump time of Y^0 then there is an n so that $\tau_n^0 < t < \tau_{n+1}^0$ and therefore for sufficiently large m, $\tau_n^m < t < \tau_{n+1}^m$. By the continuity of the deterministic functions that govern the motion between jumps, it follows that $Y^m(t) \to Y^0(t)$ a.s. as $m \to \infty$, for such t, as we set out to prove. This completes the proof. □

4.2 Stationary distributions

In this subsection we show that the process Y has a stationary distribution. Since Y is a Markov process, this can be done by finding invariant distributions to its semigroup. For that we shall use the Krylov-Bogoliubov theorem (see Theorem 3.1.1 of [7]). The statement of this theorem requires the semigroup of the process to be Feller, a condition not met in our case, since we work with the forward recurrence time α^N. We therefore argue that the Krylov-Bogoliubov candidate for the invariant measure is invariant with respect to the 1-potential operator U^1 defined above, proven in Lemma 4.1 to map bounded continuous functions to bounded functions that are continuous on $\mathcal{Y}_s^c$. We then use Lemma 1 (p. 159) of Azema, Kaplan-Duflo, Revuz [4] to conclude that any measure that is invariant with respect to U^1 is invariant with respect to the Markovian semigroup $P_t\psi(y) = \mathbb{E}_y(\psi(Y_t))$, and therefore a stationary measure for Y.

For each measurable set $B \subset \mathcal{Y}$ and $t > 0$ define

$$L_t^\mu(B) = \frac{1}{t}\int_0^t \mathbb{P}_\mu(Y(s) \in B)ds, \tag{65}$$

where μ is any initial distribution for the process Y. Obviously, for each t, L_t is a probability measure on the measurable sets of $\mathcal{Y}$.

Theorem 4.2. *Let Assumption 4.1 hold. Assume, in addition, that for all $i = 1, \dots, J$, one has $\mathbb{E}_\mu\langle 1, \eta_{i,0}\rangle < \infty$. Then the family of measures $\{L_t^\mu\}_{t>0}$ is tight. Any subsequential limit of this family is an invariant measure for U^1, and thus for the semigroup of Y.*

Proof. The proof of the first assertion follows along lines similar to those of Lemma 4.4–4.8 of [13], proved for all classes in our case, and we shall not repeat it here. Since in the Krylov-Bogoliubov Theorem it is required that the semigroup be Feller, we shall show how to adjust its proof to our setting. Let $t_n \to \infty$ as $n \to \infty$ be a subsequence along which the sequence of probability measures $L_{t_n}^\mu$ converges weakly to a measure ξ.

Since by Assumption 4.1 the interarrival times have densities, with probability one no two arrivals occur at the same time, except possibly the first arrivals of some of the classes (i.e., in the case of starting at a special point of the state space). Hence

$$\lim_{t\to\infty} \mathbb{P}_\mu(Y_t \in \mathcal{Y}_s) = 0,$$

and therefore any limit of $L_{t_n}^\mu$ does not charge the set of special points, $\mathcal{Y}_s$.

Let ψ be a bounded continuous function on $\mathcal{Y}$. Then $U^1\psi$ is bounded, and by Lemma 4.1, it is continuous on $\mathcal{Y}_s^c$. Denote by $(P_r)_{r\geq 0}$ the semigroup of Y. Then

$$\begin{aligned}
\langle U^1\psi, \xi\rangle &= \langle U^1\psi, \lim_{t_n\to\infty} L^\mu_{t_n}\rangle \\
&= \lim_{t_n\to\infty} \Big\langle U^1\psi, \frac{1}{t_n}\int_0^{t_n} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle \\
&= \lim_{t_n\to\infty} \Big\langle \int_0^\infty e^{-r}P_r\psi dr, \frac{1}{t_n}\int_0^{t_n} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle \\
&= \lim_{t_n\to\infty} \int_0^\infty e^{-r}\Big\langle P_r\psi, \frac{1}{t_n}\int_0^{t_n} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle dr \\
&= \lim_{t_n\to\infty} \int_0^\infty e^{-r}\Big\langle \psi, \frac{1}{t_n}\int_r^{t_n+r} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle dr \\
&= \int_0^\infty e^{-r}\lim_{t_n\to\infty} \Big\langle \psi, \frac{1}{t_n}\int_r^{t_n+r} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle dr \\
&= \int_0^\infty e^{-r}\lim_{t_n\to\infty} \Big[\frac{1}{t_n}\Big\langle \psi, \int_0^{t_n} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle \\
&\qquad + \frac{1}{t_n}\Big\langle \psi, \int_{t_n}^{t_n+r} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle - \frac{1}{t_n}\Big\langle \psi, \int_0^{r} \mathbb{P}_\mu(Y_s\in\cdot)ds\Big\rangle\Big] dr \\
&= \int_0^\infty e^{-r}\langle\psi,\xi\rangle dr = \langle\psi,\xi\rangle,
\end{aligned}$$

where the second equality follows from weak convergence, the fact that $U^1\psi$ is bounded, and is continuous on a set of full ξ-measure, the fourth by Fubini's theorem, the fifth by the Markov property and the sixth by dominated convergence. □

The following result relating integration against L^μ_t to integration with respect to invariant measures will be used in Section 5.

Proposition 4.1. *Let Assumption 4.1 hold and assume that $\mathbb{E}_\mu(\langle 1,\eta_{i,0}\rangle) < \infty, i = 1,\dots,J$. Let $c_i, i = 1,\dots,J$ be non negative constants and recall that $Q_1,\dots,Q_J$ denote the queue lengths of the various classes. Let ξ be an invariant measure obtained as above, and t_n the corresponding subsequence. Then*

$$\langle c_1Q_1+\cdots+c_JQ_J,\xi\rangle = \lim_{t_n\to\infty}\frac{1}{t_n}\int_0^{t_n}\mathbb{E}_\mu(c_1Q_1(s)+\cdots+c_JQ_J(s))ds. \tag{66}$$

Consequently, there exists an invariant measure $\hat{\xi}$, such that

$$\langle c_1Q_1+\cdots+c_JQ_J,\hat{\xi}\rangle = \limsup_{T\to\infty}\frac{1}{T}\mathbb{E}_\mu\Big[\int_0^{T}(c_1Q_1(s)+\cdots+c_JQ_J(s))ds\Big]. \tag{67}$$

Proof. We first prove (66). Recall that for each $i = 1,\dots,J$, $Q_i(t) = X_i(t) - \langle 1,\nu_{i,t}\rangle$. Thus $Q_i(t)$ is obtained as a continuous function on the state space of the Markov process (α, X,η,ν). If it were bounded the result would follow from weak convergence. To obtain the result for the unbounded function at hand, we shall prove that $\{Q_i : i = 1,\dots,J\}$ are uniformly integrable with respect to the sequence of measures $L^\mu_{t_n}$. That is, that $\sup_n(\mathbb{E}_{L^\mu_{t_n}}(Q_i 1_{\{Q_i>K\}}))\to 0$ as $K\to\infty$, where $\mathbb{E}_{L^\mu_{t_n}}$

is the expected value w.r.t. the measure $L^{\mu}_{t_n}$. Note that for each i, $Q_i(t) \leq \langle 1, \eta_{i,t}\rangle$, and so the uniform integrability of Q_i will follow from that of $\langle 1, \eta_i\rangle$.

Using Theorem 4.9 of Chapter 3 of [6], and the fact that $\langle 1, \eta_{i,t}\rangle$ are non-negative, it suffices to show is that $\mathbb{E}_{L^{\mu}_{t_n}}(\langle 1, \eta_i\rangle) < \infty$, that $\langle\langle 1, \eta_i\rangle, \xi\rangle < \infty$ and that $\lim_{t_n\to\infty} \mathbb{E}_{L^{\mu}_{t_n}}(\langle 1, \eta_i\rangle) = \langle\langle 1, \eta_i\rangle, \xi\rangle$.

We first recall Lemma 4.4 of [13] that proves in the single-class case that $\sup_{t\geq 0} \mathbb{E}(\langle 1, \eta_t\rangle) < \infty$. Their proof carries over to our case with $\mathbb{E}_\mu$ replacing their $\mathbb{E}$, with no changes. This immediately implies that $\mathbb{E}_{L^{\mu}_{t_n}}(\langle 1, \eta_i\rangle) < \infty$. Next, recall that since ξ is a stationary distribution, and $\langle\langle 1, \eta_i\rangle, \xi\rangle$ is the expectation, under the stationary distribution, of the number of customers in a $G/G/\infty$ queueing system with the arrival process E^N_i and service distribution G^r_i, it follows from the Little's law [15] that it is equal to $\lambda^N_i \theta^{-1}_i$, where we recall that $(\lambda^N_i)^{-1}$ are the mean inter-arrival times and θ^{-1}_i are the mean patience times. It therefore remains to show that $\lim_{t_n\to\infty} \mathbb{E}_{L^{\mu}_{t_n}}(\langle 1, \eta_i\rangle) = \lambda_N \theta^{-1}_i$.

To lighten the notation we restrict ourselves to one class and suppress the symbol i. From [13, Proposition 2.2] we have

$$\mathbb{E}_\mu\langle 1, \eta_t\rangle = \mathbb{E}_\mu \int_0^\infty \frac{1 - G^r(x+t)}{1 - G^r(x)} \eta_0(dx) + \int_0^t (1 - G^r(t-s)) de(s), \tag{68}$$

where $e(s) = \mathbb{E}_\mu(E(s))$, and $(E(s))_{s\geq 0}$ is the arrival process (of class i customers in the N-server queue). We shall treat the two terms of (68) separately.

We first note that since $\mathbb{E}_\mu\langle 1, \eta_0\rangle < \infty$ and $\lim_{t\to\infty} \frac{1-G^r(x+t)}{1-G^r(x)} = 0$, it follow by dominated convergence that

$$\lim_{t\to\infty} \mathbb{E}_\mu \int_0^\infty \frac{1 - G^r(x+t)}{1 - G^r(x)} \eta_0(dx) = 0,$$

and therefore that

$$\lim_{t_n\to\infty} \frac{1}{t_n} \mathbb{E}_\mu \int_{s=0}^{t_n} \int_0^\infty \frac{1 - G^r(x+s)}{1 - G^r(x)} \eta_0(dx) ds = 0.$$

As for or the second term in (68)

$$\begin{aligned} \frac{1}{t_n} \int_{t=0}^{t_n} \int_{s=0}^{t} (1 - G^r(t-s)) de(s) dt &= \frac{1}{t_n} \int_{u=0}^{t_n} e(t_n - u)(1 - G^r(u)) du \\ = \quad \frac{1}{t_n} \int_{u=0}^{t_n/2} e(t_n - u)(1 - G^r(u)) du &+ \frac{1}{t_n} \int_{u=t_n/2}^{t_n} e(t_n - u)(1 - G^r(u)) du. \end{aligned}$$

We shall treat the two terms on the right hand side of the above equation separately. For the second term,

$$\frac{1}{t_n} \int_{u=t_n/2}^{t_n} e(t_n - u)(1 - G^r(u)) du \leq \frac{e(t_n/2)}{t_n} \int_{t_n/2}^{t_n} (1 - G^r(u)) du.$$

By the Elementary Renewal Theorem $t_n^{-1} e(t_n/2) \to \frac{1}{2}\lambda_N$ as $t_n \to \infty$ whereas $\int_{t_n/2}^{t_n} (1 - G^r(u)) du \to 0$, as $t_n \to \infty$, by our assumption that the patience time has a finite expectation. As to the first term, it is equal to

$$\int_{u=0}^{t_n/2} \frac{e(t_n - u)}{t_n - u} \frac{t_n - u}{t_n} (1 - G^r(u)) du.$$

One can choose t_n large enough so that $|\frac{e(t_n-u)}{t_n-u} - \lambda_N| < 1$, for all $u \in [0, t_n/2]$. Thus applying the dominated convergence theorem to the above term we have

$$\lim_{t_n\to\infty} \int_{u=0}^{t_n/2} \frac{e(t_n-u)}{t_n-u} \frac{t_n-u}{t_n}(1-G^r(u))du = \lambda_N \int_0^\infty (1-G^r(u)du = \frac{\lambda_N}{\theta^r}.$$

Summing all the above we have proved that

$$\lim_{t_n\to\infty} \mathbb{E}_{L^\mu_{t_n}}(\langle 1, \eta\rangle) = \frac{\lambda_N}{\theta^r},$$

as required. This being proved for each class, we have shown that $\langle 1, \eta_i\rangle$, $i = 1, \ldots, J$ are uniformly integrable under $\mathbb{E}_{L^\mu_{t_n}}$ and therefore so are Q_i, $i = 1, \ldots, J$. It follows that

$$\mathbb{E}_{L^\mu_{t_n}}\Big(\sum_{i=1}^J c_i Q_i\Big) \to \Big\langle \sum_{i=1}^J c_i Q_i, \xi\Big\rangle \quad \text{as} \quad t_n \to \infty$$

as claimed.

Next, to show that (67) follows, let $\{T_n\}$ be a sequence along which the r.h.s. of (67) is achieved. It follows from Theorem 4.2 that the sequence $\{L^\mu_{T_n}\}$ is tight, and that any subsequential limit is invariant. Select one such invariant measure and denote it by $\hat{\xi}$. Denote by $\{n'\}$ the corresponding subsequence. Then (67) follows from (66) by substituting $(\hat{\xi}, \{T_{n'}\})$ for $(\xi, \{t_n\})$. □

4.3 Convergence

We now relate a scaled version of the N-server system to the fluid model. The scaling is performed as follows. For the initial conditions we write $\bar{X}^N_{i,0} = N^{-1}X^N_{i,0}$ and $\bar{B}^N_{i,0} = N^{-1}B^N_{i,0}$. For the real-valued processes we let $\bar{X}^N_i = N^{-1}X^N_i$, and define $\bar{E}^N_i, \bar{B}^N_i, \bar{D}^N_i, \bar{K}^N_i, \bar{R}^N_i, \bar{Q}^N_i, \bar{I}^N_i$ analogously. For the measure-valued processes, $\bar{\nu}^N_{i,0} = N^{-1}\nu^N_{i,0}$, $\bar{\nu}^N_i = N^{-1}\nu^N_i$, and $\bar{\eta}^N_{i,0}$ and $\bar{\eta}^N_i$ are defined analogously.

The first two of the three items in the assumption below summarize the hypotheses considered in the main results of Section 3. Recall L defined in (56).

Assumption 4.2. *One has*

- *The hazard rates h^r_i are all bounded.*
- *G^r_L is strictly increasing in $[0, H^r_L)$.*
- *For each i, h^s_i is either bounded or lower semi-continuous on (L^s_i, H^s_i), for some $L^s_i < H^s_i$.*

Next are assumptions regarding convergence of the initial distributions and mean inter-arrival times (recall that for the N-th system, the class-i mean inter-arrival times are given by $(\lambda^N_i)^{-1}$). For simplicity we assume that the limiting initial conditions are deterministic.

Assumption 4.3. *As $N \to \infty$,*

- *$N^{-1}\lambda^N_i \to \lambda_i > 0$, for every i.*

- $\bar{X}^N_{i,0}$, $i = 1, \ldots, J$ *converges a.s.; its limit is denoted by* $X_{i,0}$.

- $\bar{\nu}^N_{i,0}$ *converges a.s., weakly in* $\mathcal{M}_F[0, H^s_i)$, *for every* i*; its limit is denoted by* $\nu_{i,0}$.

- $\bar{\eta}^N_{i,0}$ *converges a.s., weakly in* $\mathcal{M}_F[0, H^r_i)$ *for every* i*; its limit is denoted by* $\eta_{i,0}$. *Moreover, for each* i, $\eta_{i,0}[0, t)$ *are continuous in* t. *Finally, one has* $\mathbb{E}[\langle 1, \bar{\eta}^N_{i,0}\rangle] \to \langle 1, \eta_{i,0}\rangle$, *for every* i.

- $X_{i,0}$, $\nu_{i,0}$ *and* $\eta_{i,0}$ *are deterministic.*

Owing to the structure of the arrival processes (renewal with finite mean inter-arrival), $\bar{E}^N_i$ converge a.s., uniformly over finite time intervals, to E_i, where, here and in what follows,

$$E_i(t) = \lambda_i t, \qquad t \geq 0,\ i = 1, \ldots, J.$$

Recall our notation from Section 3 and denote by $\mathbf{S} = (B, X, Q, D, K, R, \nu, \eta)$ the solution to the FME with data E and initial condition (X_0, ν_0, η_0). Note that $\mathbf{S}$ is uniquely defined under the assumptions of this section and in view of the results of the previous section. We can now prove convergence of the scaled N-server system over a finite time interval. The process

$$(\bar{E}^N, \bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N, \bar{\nu}^N, \bar{\eta}^N)$$

takes values in $\hat{\mathcal{Y}}$, where $\hat{\mathcal{Y}} = \mathbb{R}^{7J}_+ \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H^s_i) \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H^r_i)$. We endow it with the product topology ($\mathbb{R}_+$ with Euclidean, $\mathcal{M}_D$ with weak topology). It is a Polish space. The process's sample paths belong to $\mathcal{D}_{\hat{\mathcal{Y}}}(\mathbb{R}_+)$, which we endow with the corresponding Skorohod topology.

Theorem 4.3. *Under Assumptions 4.1, 4.2 and 4.3,*

$$(\bar{E}^N, \bar{\mathbf{S}}^N) = (\bar{E}^N, \bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N, \bar{\nu}^N, \bar{\eta}^N) \Rightarrow (E, \mathbf{S}) = (E, B, X, Q, D, K, R, \nu, \eta).$$

Proof. First, as mentioned above, $\bar{E}^N_i$ converge to E_i, which have continuous sample paths, by which $\{\bar{E}^N_i\}$ are C-tight. Tightness of each of the sequences $\bar{D}^N_i$ and $\bar{R}^N_i$ follows precisely as in the case treated in Lemma 6.3 of [12]. We thus omit the details. Note that each of the jumps of these processes is of size N^{-1}. As a consequence, these processes are, in fact, C-tight (see [11], Proposition VI.3.26).

Note that, for each N, $\bar{E}^N$ and the components of $\bar{\mathbf{S}}^N$ satisfy equations (26)–(32), as follows from the equations listed in Section 2 for the unscaled processes. As a result, Proposition 3.3 applies for the scaled processes. Thus, for any $t, \theta > 0$,

$$w(\bar{Q}^N, \theta, t) \leq c_1(w(\bar{E}^N, \theta, t) + w(\bar{D}^N, \theta, t) + w(\bar{R}^N, \theta, t)). \tag{69}$$

Using (27) and (28) we also have

$$\|\bar{Q}^N\| \leq 1 + \|\bar{X}^N_0\| + \|\bar{E}^N\| + \|\bar{D}^N\| + \|\bar{R}^N\|. \tag{70}$$

The C-tightness of each of the sequences $\bar{E}^N_i$, $\bar{D}^N_i$ and $\bar{R}^N_i$ implies, in view of (69), that, for each $t > 0$, $\varepsilon > 0$, $\varepsilon' > 0$ there exists $\theta > 0$ such that $\mathbb{P}(w(\bar{Q}^N, \theta, t) > \varepsilon) < \varepsilon'$ for all large N. Using also the assumed convergence of $\bar{X}^N_0$ gives tightness of the r.h.s. of (70) (see, eg. Proposition VI.3.26 of [11]). As a result, C-tightness of each of the sequences $\bar{Q}^N_i$ follows (ibid.).

Next, recalling that the scaled processes satisfy (27), (28) and (26), it follows that each of the sequences $\bar{X}_i^N$, and in turn, $\bar{B}_i^N$ and $\bar{K}_i^N$ are C-tight as well.

Further, the measure-valued processes are tight. The argument follows closely that provided in Lemma 6.6 of [12], and we thus omit the details.

Since the scaled processes satisfy (26)–(31) and (35), any subsequential limit also satisfies these equations. The prelimit processes satisfy also (32). Let us argue via continuity of the Skorohod map that so do the limit processes. To this end, fix a subsequential limit (B, X, Q, D, K, R) of $(\bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N)$. Since the prelimit processes satisfy (26)–(32), Lemma 3.2(i) is applicable. Fix $i_0 \in \{2, 3, \dots, J\}$. Recalling the notation Θ from this lemma, as well as the solution map $\hat{\Gamma}$ from Definition 3.1, we have $(\hat{Q}^N, \hat{K}^{N,(2)}) = \hat{\Gamma}[\hat{E}^N]$ where $(\hat{Q}^N, \hat{K}^{N,(2)}, \hat{E}^N) = \Theta(\bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N)$. Recall that Γ is continuous in the uniform topology (41), and note, by Definition 3.1 and the definition of $\hat{\Gamma}$, that so is $\hat{\Gamma}$. As a result, if $(\hat{Q}, K^{(2)}, \hat{E}) = \Theta(B, X, Q, D, K, R)$, one has $(\hat{Q}, K^{(2)}) = \hat{\Gamma}[\hat{E}]$. By (44),

$$\Big(\sum_{j=1}^{i_0-1} Q_j, \sum_{j=i_0}^{J} Q_j\Big) = \hat{Q}^+. \tag{71}$$

Now, by the third bullet in Definition 3.1 we have, a.s.,

$$\int_{[0,\infty)} 1_{\{\hat{Q}(s) \in G^o\}} dK_s^{(2)} = 0.$$

By the structure of the set G and the nonnegativity of $\hat{Q}_2$, it follows that $\hat{Q}(s) \in G^o$ if and only if $\hat{Q}_1(s) > 0$, which, by (71) holds if and only if $\sum_{j=1}^{i_0-1} Q_j(s) > 0$. We thus obtain, a.s.,

$$\int_{[0,\infty)} 1_{\{\sum_{j=1}^{i_0-1} Q_j(s) > 0\}} dK_s^{(2)} = 0.$$

Recalling that i_0 is arbitrary and applying Lemma 3.2(ii) we obtain that the limit processes satisfy (32).

Now, any subsequential limit satisfies also equations (33), (34), (36) and (37). The argument here follows that of the proof of Theorem 7.1 of [12]. More precisely, the first inequality of (25) and identity (36) follow as that of (7.2) of [12]. This relation corresponds to (5.49) established in Proposition 5.17 of [14] that relies on Lemma 5.8(1) and Lemma 5.16 of that paper. Those continue to hold in the presence of abandonments and priorities. The second inequality of (25) and identity (37) are proved in Proposition 7.2 and Lemmas 7.3–7.6 that are a part of the proof of Theorem 7.1 of [12] and carry without any change for each class, to our model. The fact that (33) and (34) are satisfied follows as in the proof of Theorem 7.1 of [12] or of Theorem 5.15 of [14] applied to each class. We avoid repeating the details of these arguments here.

Having shown that any limit satisfies all of (25)–(37), we can apply Theorem 3.1, by which the limit must be equal to $(E, \mathbf{S})$ a.s. This shows the claimed convergence and completes the proof. □

Finally we present the result regarding convergence of invariant distributions. Let $\Sigma_0^* = (X_0^*, \nu_0^*, \eta_0^*)$ denote the unique invariant state of the fluid model, identified in Theorem 3.3. Let $\bar{L}_t^N$ be defined by

$$\bar{L}_t^N(B) = \frac{1}{t}\int_0^t \mathbb{P}((\bar{X}_s^N, \bar{\nu}_s^N, \bar{\eta}_s^N) \in B) ds,$$

for any measurable set $B \subset \bar{\mathcal{Y}} = \mathbb{R}_+^J \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H_i^s) \times \mathop{\times}\limits_{i=1}^{J} \mathcal{M}_D[0, H_i^r)$. We endow $\mathbb{R}_+$ with the Euclidean topology, and $\mathcal{M}_D$ with the weak topology and $\bar{\mathcal{Y}}$ with the corresponding product topology. It follows from Theorem 4.2 that these measures are tight in t, for each N.

Theorem 4.4. *Let Assumptions 4.1, 4.2 and 4.3 hold. For each N, fix a subsequential limit ξ^N of $\bar{L}_t^N$. Then $\xi^N \Rightarrow \delta_{\Sigma_0^*}$ as $N \to \infty$.*

Proof. Given the result of Theorem 4.3, this follows as in Theorem 3.3 of [13]. □

5 Application: the $c\mu/\theta$ rule

The main results of this paper are concerned with the behavior of the system under a particular policy, namely a policy of priority type. In this section we relate these results to a dynamic control problem in which a control policy is sought to minimize a given cost. In [1] and [2] such a control problem was studied for a multi-class many-server system with abandonment, under a LLN scaling, and a general lower bound on the asymptotic performance was obtained [2] for general service time distribution and exponential reneging time distribution (see Proposition 5.1 below). In addition, in the case of exponential service time, this bound was shown to be achieved by a simple fixed priority policy (the priority ordering is described below). The goal of this section is to show that this bound is achieved by the same policy for general service time distribution. The proof of this fact uses the results of this paper to their full strength.

To describe the control problem we consider a queueing system analogous to the one presented in Section 2, under a wide range of control policies. The fixed priority policy of Section 2 will be a special case. Thus, as before, N represents the number of (identical) servers, and E_i^N, B_i^N, X_i^N, Q_i^N, D_i^N, R_i^N are processes having the same meaning as in Section 2. The probabilistic and scaling limit assumptions that we shall impose on arrival, service and reneging will be consistent with the general framework of this paper, except that we will only be concerned with exponential reneging distributions.

A control policy is usually defined as a rule for scheduling jobs. For our purpose, however, specifying the set of rules is not necessary, and instead, a control will be associated with a collection of processes satisfying a minimal set of relations. More precisely, given N, let NJ mutually independent renewal processes $\tilde{D}_{i,k}$, $i = 1, 2, \ldots, J$, $k = 1, 2, \ldots, N$, be given, where $D_{i,k}$ specifies the service times of class i in server k. The inter-renewal times for each of these processes are distributed according to G_i^s (with mean μ_i^{-1}), and $D_{i,k}(0) = 0$ (i.e., no renewal counted at time 0). For each i, k, let $B_{i,k}^N$ be a process that takes values in $\{0, 1\}$, and indicates the business of server k with a class-i customer. The number of class-i service completions by server k, up to time t, is given by

$$D_{i,k}^N(t) = \tilde{D}_{i,k}\Big(\tilde{a}_{i,k}^N(0) + \int_0^t B_{i,k}^N(s)ds\Big), \tag{72}$$

where $\tilde{a}_{i,k}^N(0)$ denotes the time that a customer of class i that occupies server k at time 0 (if such a customer exists) has already spent there by then. The number of class-i customers in service and number of class-i departures, respectively, are given by

$$\sum_{k=1}^{N} B_{i,k}^N = B_i^N, \qquad \sum_{k=1}^{N} D_{i,k}^N = D_i^N. \tag{73}$$

It is assumed that interruption of service is not possible (i.e., a server that is assigned a new customer serves it until completion of the service requirement). The total number of customers reneging up to time t is given by

$$R_i^N(t) = \tilde{R}_i\Big(\theta_i \int_0^t Q_i^N(s)ds\Big), \tag{74}$$

where $\tilde{R}_i$ are mutually independent standard Poisson processes, and $\theta_i > 0$ are given parameters, representing the per-customer reneging rate. The arrival processes (E_i^N) are as defined in section 2.

For each N, the collections (E_i^N), $(\tilde{R}_i)$ and $(\tilde{D}_{i,k})$ are assumed to be mutually independent. Given are, in addition, initial conditions $(X_{i,0}^N, \nu_{0,i}^N)$. We refer to these stochastic processes and initial conditions as the *primitives.* Note that the initial age-in-queue measures $(\eta_{i,0}^N)$ are not relevant here due to the memoryless property of the exponential patience distribution. The initial age-in-service measure ν_0^N may be used to determine the parameters $\tilde{a}_{i,k}^N(0)$ in (72) in an obvious way (that is, each non-zero $\tilde{a}_{i,k}^N(0)$ corresponds to a unit point mass of $\nu_{0,i}^N$ at $x = \tilde{a}_{i,k}^N(0)$). In addition, the primitives are related to the assumptions made in the previous section. First, for each N, E_i^N are mutually independent renewal processes with inter-arrival distribution having mean $1/\lambda_i^N$, and satisfying Assumption 4.1 (regarding density) and the first item of Assumption 4.3 (convergence). Next, the first and second items of Assumption 4.2 are satisfied due to the exponential assumption on the patience. The last item of Assumption 4.2, regarding the service time distribution, is assumed, as well as all items of Assumption 4.3, regarding initial conditions. Thus all of Assumptions 4.1–4.3 are in force. Finally, it is assumed that the system is overloaded, in the sense that $\sum \rho_i = \sum \lambda_i/\mu_i > 1$. (The results below are still valid in the underloaded case but are trivial, as the fluid cost V in (79) is zero in this case.)

Clearly, we require

$$X_i^N = X_i^N(0) + E_i^N - R_i^N - D_i^N, \tag{75}$$

$$Q_i^N = X_i^N - B_i^N \geq 0, \tag{76}$$

and

$$B_i^N \geq 0, \quad \sum_{i=1}^{J} B_i^N \leq N. \tag{77}$$

Given the primitives, any collection of processes

$$\pi = ((B_{i,k}^N), B^N, (D_{i,k}^N), D^N, R^N, X^N, Q^N),$$

satisfying equations (72)–(77) is regarded a *policy* for the Nth system, and the set of all policies for the Nth system is denoted by $\mathit{\Pi}_N$. The priority policy analyzed in this paper (specialized to exponential reneging) is a valid policy according to this definition. As in the rest of this paper, normalized versions of X^N, Q^N and B^N are denoted by $\bar{X}^N = N^{-1}X^N$, $\bar{Q}^N = N^{-1}Q^N$ and $\bar{B}^N = N^{-1}B^N$. Fix $c = (c_1, \dots, c_J) \in (0, \infty)^J$. Given N and policy $\pi \in \Pi^N$, consider the long-run average, expected cost

$$\bar{C}^N(\pi) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\Big[\int_0^T \sum_i c_i\bar{Q}_i^N(t)dt\Big], \tag{78}$$

where $c_i\bar{Q}_i^N$ represents a linear holding cost for class i. See Remark 5.1 below for the incorporation of reneging penalties in this cost function. Let also

$$\underline{C}^N(\pi) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\Big[\int_0^T \sum_i c_i\bar{Q}_i^N(t)dt\Big].$$

The following is a result from [2]. Denote $\mathbb{S}^J = \{b \in \mathbb{R}_+^J : \sum b_i \le 1\}$.

Proposition 5.1. (Propositions 2.1 and A.1 of [2]) *Under any sequence of policies $\pi^N \in \Pi^N$, $N \in \mathbb{N}$,*

$$\liminf_{N\to\infty} \underline{C}^N(\pi^N) \ge V := \inf\big\{c \cdot q \,:\, (q, b) \in (\mathbb{R}_+^J, \mathbb{S}^J),\, \theta_i q_i + \mu_i b_i = \lambda_i,\, i = 1, 2, \dots, J\big\}. \tag{79}$$

It is easy to see what pair (q, b) achieves the infimum on the r.h.s. of (79). Namely, q_i are determined from b_i via the equations $\theta_i q_i + \mu_i b_i = \lambda_i$, while b_i are determined by the relations

$$\sum_{i=1}^j b_i = 1 \wedge \sum_{i=1}^j \frac{\lambda_j}{\mu_j}, \qquad j = 1, 2, \dots, J,$$

where the classes are labeled in such a way that, with $L_i = c_i\mu_i/\theta_i$,

$$L_1 \ge L_2 \ge \cdots \ge L_J. \tag{80}$$

In what follows, we assume that the labeling is as above. What is referred to in [1] and [2] as the *$c\mu/\theta$ rule* (in analogy with the well-known $c\mu$ rule) is the non-preemptive priority policy according to the ordering (80). The main point of this section is to show that prioritizing according to (80) is asymptotically optimal.

Theorem 5.1. *Let π^* denote the priority policy according to the class ordering* (80)*. Then*

$$\limsup_{N\to\infty} \bar{C}^N(\pi^*) = V\ .$$

Proof. This is a consequence of the main results of this paper. First, by Proposition 4.1, specifically (67), there exists, for each N, an invariant distribution $\hat{\xi} = \hat{\xi}^N$ such that

$$\bar{C}^{N,*} \doteq \bar{C}^N(\pi^*) = \frac{1}{N}\Big\langle \sum_{i=1}^J c_i Q_i^N, \hat{\xi}^N \Big\rangle = \Big\langle \sum_{i=1}^J c_i \bar{Q}_i^N, \hat{\xi}^N \Big\rangle.$$

Next, note that the hypotheses of Theorem 3.3 are satisfied and thus the invariant state of the fluid model is uniquely given by that result. Denote the invariant state and the corresponding quantities from Theorem 3.3 by $\mathbf{S}_0 = (B_0, X_0, Q_0, D_0, K_0, R_0, \nu_0, \eta_0)$. Arguing as in the proof of Theorem 6.2 of [13] one can show that any sequence of stationary measures for the processes $\bar{Y}^N$ are tight. Our convergence results of the previous section (particularly, Theorem 4.4) show that any subsequence of stationary distributions of the scaled N-server system converge to the unique stationary state of the fluid equations. Hence $\langle \sum_{i=1}^J c_i \bar{Q}_i^N, \hat{\xi}^N \rangle$ converges in distribution to $\sum_{i=1}^J c_i Q_{i,0}$. To deduce that

$$\lim_{N\to\infty} \bar{C}^{N,*} = \lim_{N\to\infty} \Big\langle \sum_{i=1}^J c_i \bar{Q}_i^N, \hat{\xi}^N \Big\rangle = \sum_{i=1}^J c_i Q_{i,0},$$

it suffices to show that $\bar{Q}_i^N$ are uniformly integrable with respect to the stationary measures $\hat{\xi}^N$. Again, to show that, it is enough to show that $\langle 1, \bar{\eta}_i^N \rangle$ are uniformly integrable with respect to $\hat{\xi}^N$. As in Section 3, one needs to show that $\sup_N \langle \langle 1, \bar{\eta}_i^N \rangle, \hat{\xi}^N \rangle < \infty$, $\langle 1, \eta_{i,0} \rangle < \infty$, where $\eta_{i,0}$ is the i-th component of η_0 of the unique invariant state of the fluid, and that $\lim_{N\to\infty} \langle \langle 1, \bar{\eta}_i^N \rangle, \hat{\xi}^N \rangle = \langle 1, \eta_{i,0} \rangle$. But as we have shown in Section 4, by Little's formula, $\langle \langle 1, \bar{\eta}_i^N \rangle, \hat{\xi}^N \rangle = \frac{\lambda_i^N}{N} \frac{1}{\theta_i}$ which converges as $N \to \infty$ to $\bar{\lambda}_i \frac{1}{\theta_i} = \bar{\lambda}_i \int_0^\infty (1 - G_i^r(u)) du = \langle 1, \eta_{i,0} \rangle$. We have thus shown that $\bar{C}^{N,*} \to c \cdot Q_0$ as $N \to \infty$.

It thus remains to show that $V = c \cdot Q_0$. In view of the discussion following Proposition 5.1, it suffices to show that the pair (Q_0, B_0) satisfies the equations

$$\lambda_i = \theta_i Q_{i,0} + \mu_i B_{i,0}, \qquad i = 1, 2, \ldots, J, \tag{81}$$

$$\sum_{i=1}^j B_{j,0} = 1 \wedge \sum_{i=1}^j \rho_j, \qquad j = 1, 2, \ldots, J. \tag{82}$$

To this end note that, for $i < L$, by (57),

$$B_{i,0} = \lambda_i \int_0^\infty (1 - G_i^s(x)) dx = \frac{\lambda_i}{\mu_i} = \rho_i,$$

whereas $Q_{i,0} = X_{i,0} - B_{i,0} = 0$. Equation (81) thus holds in this case. Since $B_{i,0} = X_{i,0} = \rho_i$, so does (82). For $i > L$, by (59), $B_{i,0} = 0$, while by (61), $Q_{i,0} = \frac{\lambda_i}{\theta_i}$, which again shows that (81) and (82) are valid. Finally, consider $i = L$. By (58), $B_{L,0} = 1 - \hat{\rho}$. Thus (82) holds. Moreover, $Q_{L,0} = b$, with $\chi_L(b) = x$, $\eta_{L,*}[0, x] = \lambda_L \int_0^x e^{-\theta_L a} da = b$. Along with equation (60) this gives

$$Q_{L,0} = \frac{\lambda_L}{\theta_L} \frac{\hat{\rho} + \rho_L - 1}{\rho_L}.$$

As a result, (81) holds for $i = L$ as well. This shows (81) and (82), and hence $\bar{C}^{N,*} \to c \cdot Q_0 = V$. □

Remark 5.1. *In addition to the holding cost treated above, it is reasonable to penalize abandonment of waiting customers. That is, replace the expected value in* (78) *with the augmented cost*

$$\mathbb{E}^\pi \Big[\sum_i c_i^a \bar{R}_i^N(T) + \sum_i \int_0^T c_i^b \bar{Q}_i^N(t) dt \Big],$$

where $\bar{R}^N = N^{-1} R^N$ *is the normalized cumulative reneging process. Recalling* (74)*, it may be verified that the equality*

$$\mathbb{E}^\pi[\bar{R}_i^N(T)] = E^\pi \Big[\theta_i \int_0^T \bar{Q}_i^N(t) dt \Big]$$

holds for all $T \geq 0$, provided that the policy π is non-anticipative, that is, $Q^N(t)$ is measurable on the history

$$\{R^N(s), E^N(s), D^N(s), s \leq t;\ Q^N(s), s < t\}\,.$$

(The non-anticipative property is needed to ensure that the integral $\tau_t = \theta_i \int_0^t Q_i^N(s)ds$ is independent of the future increments $\tilde{R}_i(\tau_t+s)-\tilde{R}_i(\tau_t)$ of the Poisson process $\tilde{R}_i(t)$.) See, e.g., Lemma 1 in [3] for a proof of the above equality via a martingale argument. Given this equality, the augmented cost is equivalent to the one in (78)*, with $c_i = \theta_i c_i^a + c_i^b$.*

A Appendix

Here we analyze the SP (G, d) (see Definition 3.1). Uniqueness of solutions to the SP (and Lipschitz continuity of the solution map) in convex polyhedral domains are well-understood, but on a non-convex polyhedron sufficient conditions are perhaps less standard. However, the particular setting under consideration is simple. Indeed, owing to the direction of constraint being fixed, questions of uniqueness, explicit representation, and Lipschitz property can be addressed via a one-dimensional SP on a time-varying domain.

Proposition A.1. *The SP (G, d) is uniquely solvable for any $\beta \in \mathcal{D}_{\mathbb{R}^2}(\mathbb{R}_+)$. Moreover, the Lipschitz property* (41) *and the statement* (42) *regarding the modulus of continuity hold.*

Proof. The proof is based on a result from [5] regarding a one-dimensional SP with moving boundary. Let $\ell \in \mathcal{D}_{\mathbb{R}}(\mathbb{R}_+)$ be fixed. Let a path $\hat{\beta} \in \mathcal{D}_{\mathbb{R}}(\mathbb{R}_+)$ be given. A pair $(\hat{\gamma}, \hat{\eta})$, $\hat{\gamma} \in \mathcal{D}_{\mathbb{R}}(\mathbb{R}_+)$, $\hat{\eta} \in \mathcal{D}_{\mathbb{R}_+}(\mathbb{R}_+)$, is said to *solve the SP on $[\ell(\cdot), \infty)$ for $\hat{\beta}$* if

- $\hat{\gamma} = \hat{\beta} + \hat{\eta}$,
- $\hat{\gamma}(t) \geq \ell(t)$ for all $t \geq 0$,
- $\hat{\eta}$ is nondecreasing, and $\int_{[0,\infty)} 1_{\{\hat{\gamma}(s) > \ell(s)\}} d\hat{\eta}(s) = 0$.

It follows from Theorem 2.6 and Remark 2.7 of [5] that for any path $\hat{\beta}$ there exists a unique pair $(\hat{\gamma}, \hat{\eta})$ that solves the SP on $[\ell(\cdot), \infty)$ for $\hat{\beta}$, and

$$\hat{\gamma}(t) = \hat{\beta}(t) + \sup_{s\in[0,t]} [\ell(s) - \hat{\beta}(s)]^+, \qquad t \geq 0.$$

Turning to the SP (G, d), let β be given, and consider a solution (γ, η). Denote $\tilde{e}_1 = d/\sqrt{2} = (e_1 - e_2)/\sqrt{2}$ and $\tilde{e}_2 = (e_1 + e_2)/\sqrt{2}$. Let (β_1, β_2) and (γ_1, γ_2) represent β and γ in the coordinate system $(\tilde{e}_1, \tilde{e}_2)$. By Definition 3.1, we have $\gamma_2 = \beta_2$. Moreover, letting

$$\ell(t) = -|\beta_2(t)|, \qquad t \geq 0,$$

it is straightforward to check that $(\gamma_1, \sqrt{2}\,\eta)$ solves the SP on $[\ell(\cdot), \infty)$ for β_1. Hence, by the result cited above, there exists a unique solution (γ, η) for the SP (G, d) for β, and γ is given by

$$\gamma_1(t) = \beta_1(t) + \sup_{s\in[0,t]} [-|\beta_2(s)| - \beta_1(s)]^+, \qquad t \geq 0,$$

$$\gamma_2 = \beta_2.$$

Both properties (41) and (42) follow from this explicit representation. This completes the proof of the proposition. □

Acknowledgement. We thank the referees and associate editor for their very useful comments.

References

[1] R. Atar, C. Giat and N. Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.*, Vol. 58 No. 5, 1427–1439 (2010)

[2] R. Atar, C. Giat and N. Shimkin. On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Que. Sys.*, Vol. 67 No. 2, 127–144 (2011)

[3] R. Atar, A. Mandelbaum and M. Reiman. Scheduling a multi class queue withmany exponenetial serversL asymprotic optimaly in heavy traffic. *Ann. Appl. Probab.*, Vol. 14 No. 3, 1084–1134 (2004)

[4] J. Azema, M. Kaplan-Duflo and D. Revuz. Mesure invariante sur les classes récurrentes des processus de Markov. *Z. Wahrscheinlichkeitstheorie verw. Geb.* 8, 157–181 (1967)

[5] K. Burdzy, W. Kang and K. Ramanan. The Skorokhod problem in a time-dependent interval. *Stoch. Proc. Appl.* 119 (2009) 428–452.

[6] E. Cinlar. *Probability and stochastics* Graduate texts in mathematics, Springer. (2010).

[7] G. Da Prato and J. Zabczyk. *Ergodicity for infinite dimensional systems* Cambridge University Press (1996).

[8] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations.* Wiley, New York, 1997

[9] S. Halfin and W. Whitt. Heavy-traffic limit theorems for queues with many servers. *Oper. Res.* 29 (1981), 567–588.

[10] M. Jacobsen. *Point process theory and applications: marked point and piecewise deterministic processes.* Birkhauser 2006.

[11] J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes.* Springer-Verlag, 1987.

[12] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* Volume 20, Number 6 (2010), 2204–2260.

[13] W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Ann. Appl. Probab.* Volume 22, Number 2 (2012), 477–521.

[14] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* Volume 21, Number 1 (2011), 33–114.

[15] J. D. C. Little. A Proof of the queueing formula $L = \lambda W$. *Oper. Res.*, 9 (1961) 383–387.

[16] A. A. Puhalskii and J. E. Reed. On many-server queues in heavy traffic. *Ann. Appl. Probab.* 20, No. 1 (2010), 129–195.

[17] J. E. Reed. The G/GI/N queue in the Halfin-Whitt regime I: infinite server queue system equations. *Ann. Appl. Probab.* 19 (2009), 2211–2269.