The Annals of Applied Probability
2012, Vol. 22, No. 2, 477–521
DOI: 10.1214/10-AAP738

ASYMPTOTIC APPROXIMATIONS FOR STATIONARY DISTRIBUTIONS OF MANY-SERVER QUEUES WITH ABANDONMENT

BY WEINING KANG AND KAVITA RAMANAN[1]

University of Maryland, Baltimore County and Brown University

A many-server queueing system is considered in which customers arrive according to a renewal process and have service and patience times that are drawn from two independent sequences of independent, identically distributed random variables. Customers enter service in the order of arrival and are assumed to abandon the queue if the waiting time in queue exceeds the patience time. The state of the system with N servers is represented by a four-component process that consists of the forward recurrence time of the arrival process, a pair of measure-valued processes, one that keeps track of the waiting times of customers in queue and the other that keeps track of the amounts of time customers present in the system have been in service and a real-valued process that represents the total number of customers in the system. Under general assumptions, it is shown that the state process is a Feller process, admits a stationary distribution and is ergodic. It is also shown that the associated sequence of scaled stationary distributions is tight, and that any subsequence converges to an invariant state for the fluid limit. In particular, this implies that when the associated fluid limit has a unique invariant state, then the sequence of stationary distributions converges, as $N \to \infty$, to the invariant state. In addition, a simple example is given to illustrate that, both in the presence and absence of abandonments, the $N \to \infty$ and $t \to \infty$ limits cannot always be interchanged.

1. Introduction.

1.1. *Description.* An N-server queueing system is considered in which customers arrive according to a renewal process, have independent and identically distributed (i.i.d.) service requirements that are drawn from a general distribution with finite mean and also carry i.i.d. patience times that are drawn from another general distribution. Customers enter service in the order of arrival as soon as an idle server is available, service is nonpreemptive, and customers abandon the queue if the time spent waiting in queue reaches the patience time. This system is also sometimes referred to as the $GI/GI/N+G$ model. In this work, it is assumed that the sequences of service requirements and patience times are mutually

Received March 2010; revised August 2010.

[1]Supported in part by NSF Grants CMMI-0728064, CMMI-0928154.

MSC2010 subject classifications. Primary 60K25, 68M20, 90B22; secondary 60F99.

Key words and phrases. Multi-server queues, stationary distribution, ergodicity, measure-valued processes, abandonment, reneging, interchange of limits, mean-field limits, call centers.

independent, and that the interarrival, service and patience time distributions have densities.

The state of the N-server system is represented by a four component process $Y^{(N)}$, consisting of the forward recurrence time process associated with the renewal arrival process, a measure-valued process that keeps track of the amounts of time customers currently in service have been in service, another measure-valued process that encodes the times elapsed since customers have entered the system (for all customers for which this time has not yet exceeded their patience times) and a real-valued process that keeps track of the total number of customers in the system. This infinite-dimensional state representation was shown in Lemma B.1 of Kang and Ramanan [15] to lead to a Markovian description of the dynamics (with respect to a suitable filtration). In addition, a fluid limit for this model was also established in [15], that is, under suitable assumptions, it was shown that almost surely, $\overline{Y}^{(N)} = Y^{(N)}/N$ converges, as $N \to \infty$, to a limit process $\overline{Y}$ which is characterized as the unique solution to a set of coupled integral equations (see Definition 5.1). The process $\overline{Y}$ will be referred to as the fluid limit.

The present work focuses on obtaining first-order approximations to the stationary distribution of $Y^{(N)}$ which is of fundamental interest for the performance analysis of many-server queues. It is first shown that for each N, $Y^{(N)}$ is a Feller, strong Markov process and has a stationary distribution. Under an additional assumption (Assumption 7.1), uniqueness of the stationary distribution and ergodicity of each $Y^{(N)}$ is also established. The main result, Theorem 3.3, shows that under fairly general assumptions the sequence of stationary distributions is tight and that any subsequential limit is an invariant state for the fluid limit. In particular, if the fluid limit has a unique invariant state, this implies that the sequence of scaled stationary distributions (indexed by the number of servers N) converges, as $N \to \infty$, to this unique invariant state. More generally, this work seeks to illustrate how an infinite-dimensional Markovian representation of a stochastic network can facilitate the (first-order) characterization of the associated stationary distributions. Furthermore, examples are presented to illustrate several subtleties in the dynamics. Specifically, it is shown that the presence of a unique invariant state is not a necessary condition for the sequence of scaled stationary distributions to have a limit and that even when such a limit exists, the $t \to \infty$ and $N \to \infty$ limits cannot in general be interchanged.

1.2. *Motivation and context.* The study of many-server queueing systems with abandonment is motivated by applications to telephone call centers and (more generally) customer contact centers. The incorporation of customer abandonment captures the effect of customers' impatience, which has a substantial impact on the performance of the system. For example, customer abandonment can stabilize a system even when it is overloaded. A considerable body of work has been devoted to the study of various steady-state or stationary performance measures of many-server queues, both with and without abandonment. In the absence of abandonment, when the interarrival times and service times are exponential, an explicit expression for the steady state queue length can be found in Bocharov et al. [4]. In the discrete-time setting, when the i.i.d. interarrival and service times are generally distributed, the classical work of Kiefer and Wolfowitz [18] (see also Foss [7]) establishes the convergence in distribution, as time goes to infinity, of the waiting time vectors to a stationary limit. The generalization to continuous time is dealt with in Asmussen and Foss [2]. For a many-server queue with stationary renewal arrivals, deterministic service times and no abandonments, Jelenkovic, Mandelbaum and Momčilović [13] showed that on the diffusive scale, the scaled stationary waiting times converge in distribution to the supremum of a Gaussian random walk with negative drift. For a many-server queue with stationary renewal arrivals, a finitely supported, lattice-valued service time distribution and no abandonments, in the so-called Halfin–Whitt asymptotic regime where the number of servers N goes to infinity and the corresponding arrival rate grows as $N - \beta\sqrt{N}$ for some $\beta > 0$, Gamarnik and Momčilović [8] characterized the limit of the scaled stationary queue length distribution in terms of the stationary distribution of an explicitly constructed Markov chain and obtained an explicit expression for the exponential decay rate of the moment generating function of this limiting stationary distribution.

For many-server queues with abandonment whose interarrival, service and abandonment distributions are exponential, Garnett, Mandelbaum and Reiman [10] provide exact calculations of various steady state performance measures and their approximations in the Halfin–Whitt asymptotic regime, both in the case of finite waiting rooms ($M/M/N/B+M$) and infinite waiting rooms ($M/M/N+M$). In the case of Poisson arrivals, exponential service distribution and general abandonment distribution ($M/M/N+G$), explicit formulae for the steady state distributions of the queue length and virtual waiting time were obtained by Baccelli and Hebuterne [3] (see Sections IV and V.2 therein), whereas several other steady state performance measures and their approximations in the Halfin–Whitt asymptotic regime were derived by Mandelbaum and Zeltyn [23].

In the previously mentioned works on characterization of stationary distributions of many-server queues, either the interarrival times and service times are assumed to be exponential or it is assumed that the service time distribution is discrete and has a finite support, and that there is no abandonment. However, statistical analysis of real call centers has shown that both service times and patience times are typically not exponentially distributed (see Brown et al. [5] and Mandelbaum and Zeltyn [23]). In general, it is difficult to derive explicit expressions for the stationary distributions of many-server queues, especially in the more realistic case when service times are not exponential and there is abandonment. This is also the case for many other classes of stochastic networks. To circumvent this problem, a common approach that is taken is to identify the long-time limits of the fluid or diffusion approximations, which are often more tractable, and then use these limits as approximations of the stationary distribution of the original system. Such an approach relies on the premise that the long-time behavior of the fluid limit can be characterized and also requires an argument that justifies the interchange of (the $N \to \infty$ and $t \to \infty$) limits (see, e.g., Gamarnik and Zeevi [9] for an interchange of limits result in the context of generalized Jackson networks). However, we show that this approach may not always be appropriate for stochastic network models. Indeed, for the case of many-server queues whose service distributions are not exponential, the long-time behavior of the fluid is subtle and difficult to characterize in large part due to the complexity in the dynamics introduced by the coupling of the measure-valued component of the fluid limit with the positive real-valued component by the nonidling condition. Furthermore, as the example we construct in Section 7 demonstrates, in general, the order of the $N \to \infty$ and $t \to \infty$ limits cannot be interchanged.

Instead we take a different approach to showing convergence that is more appropriate for mean-field limits, which involves establishing tightness of the stationary distributions and showing that any subsequence converges to an invariant state. A more detailed description of the approach is provided in Section 3.2 and additional discussion is provided in Section 7. The present work is also related to the work of Whitt [22] who analyzed a discrete time version of the model, proposed a fluid limit model and made several conjectures on the associated steady-state quantities. A comparison of our results with those of Whitt [22] is also given in Section 3.2 after the statement of our main results.

1.3. *Outline.* The outline of the paper is as follows. A precise mathematical description of the model is provided in Section 2. Section 3 introduces the basic assumptions and states the main result. The Feller property and the existence of stationary distributions of the state descriptor are proved in Section 4. The fluid equations and the invariant manifold are described in Section 5 and the asymptotics of the stationary distributions is established in Section 6. Finally, Section 7 contains a discussion of the positive Harris recurrence and ergodicity of the state descriptor, the long time behavior of the fluid limit and an example that shows that the "interchange of limits" property does not always hold. In the remainder of this section, we introduce some common notation used in the paper.

1.4. *Notation and terminology.* The following notation will be used throughout the paper. $\mathbb{Z}$ is the set of integers, $\mathbb{N}$ is the set of positive integers, $\mathbb{R}$ is the set of real numbers, $\mathbb{Z}_+$ is the set of nonnegative integers and $\mathbb{R}_+$ the set of nonnegative real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum of a and b, $a \wedge b$ the minimum of a and b and the short-hand a^+ is used for $a \vee 0$. $\mathbb{1}_B$ denotes the indicator function of the set B [i.e., $\mathbb{1}_B(x) = 1$ if $x \in B$ and $\mathbb{1}_B(x) = 0$ otherwise].

1.4.1. *Function and measure spaces.* Given any metric space E, $\mathcal{C}_b(E)$ and $\mathcal{C}_c(E)$ are, respectively, the space of bounded, continuous functions and the space of continuous real-valued functions with compact support defined on E. The support of a function φ is denoted by $\operatorname{supp}(\varphi)$. We denote by $\mathcal{D}_E[0,T]$ (resp., $\mathcal{D}_E[0,\infty)$) the space of E-valued, càdlàg functions on $[0,T]$ (resp., $[0,\infty)$) and we endow this space with the usual Skorokhod J_1-topology [21]. When E is Polish then $\mathcal{D}_E[0,T]$ and $\mathcal{D}_E[0,\infty)$ are also Polish spaces (see [21]). Let $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ be the subset of nondecreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0)=0$. Given $f \in \mathcal{I}_{\mathbb{R}_+}[0,\infty)$, f^{-1} denotes the inverse function of f defined by

$$f^{-1}(y) = \inf\{x \geq 0 : f(x) \geq y\}. \tag{1.1}$$

The space of Radon measures on a complete separable metric space E, endowed with the Borel σ-algebra, is denoted by $\mathcal{M}(E)$, while $\mathcal{M}_F(E)$ is the subspace of finite measures in $\mathcal{M}(E)$. Recall that a Radon measure on E is one that assigns finite measure to every relatively compact subset of E. The space $\mathcal{M}_F(E)$ is equipped with the weak topology, that is, a sequence of measures $\{\mu_n\}$ in $\mathcal{M}_F(E)$ is said to converge to μ in the weak topology (denoted $\mu_n \stackrel{w}{\to} \mu$) if and only if for every $\varphi \in \mathcal{C}_b(E)$,

$$\int_E \varphi(x)\mu_n(dx) \to \int_E \varphi(x)\mu(dx) \qquad \text{as } n \to \infty. \tag{1.2}$$

As is well known, $\mathcal{M}_F(E)$, endowed with the weak topology is a Polish space. The symbol δ_x will be used to denote the measure with unit mass at the point x and, by some abuse of notation, we will use $\mathbf{0}$ to denote the identically zero Radon measure on E. When E is an interval, say $[0,H)$ for some $H \in (0,\infty]$, we will often write $\mathcal{M}[0,H)$ and $\mathcal{M}_F[0,H)$ instead of $\mathcal{M}([0,H))$ and $\mathcal{M}_F([0,H))$, respectively. For any $\mu \in \mathcal{M}_F[0,H)$, we define

$$F^{\mu}(x) \doteq \mu[0,x], \qquad x \in [0,H). \tag{1.3}$$

For any Borel measurable function $f:[0,H) \to \mathbb{R}$ that is integrable with respect to $\xi \in \mathcal{M}[0,H)$, we often use the short-hand notation

$$\langle f, \xi\rangle \doteq \int_{[0,H)} f(x)\xi(dx).$$

Also, for ease of notation, given $\xi \in \mathcal{M}[0,H)$ and an interval $(a,b) \subset [0,M)$, we will use $\xi(a,b)$ to denote $\xi((a,b))$.

1.4.2. *Measure-valued stochastic processes.* In this work, we will be interested in càdlàg $\mathcal{H}$-valued stochastic processes where $\mathcal{H} = \mathcal{M}_F[0,H)$ for some $H \leq \infty$. These are random elements that are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in $\mathcal{D}_{\mathcal{H}}[0,\infty)$, equipped with the Borel σ-algebra (generated by open sets under the Skorokhod J_1-topology). A sequence $\{X_n\}_{n\in\mathbb{N}}$ of càdlàg, $\mathcal{H}$-valued processes, with X_n defined on the probability space

$(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, is said to converge in distribution to a càdlàg $\mathcal{H}$-valued process X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ if, for every bounded, continuous functional $\Xi : \mathcal{D}_{\mathcal{H}}[0,\infty) \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{E}_n[\Xi(X_n)] = \mathbb{E}[\Xi(X)],$$

where $\mathbb{E}_n$ and $\mathbb{E}$ are the expectation operators with respect to the probability measures $\mathbb{P}_n$ and $\mathbb{P}$, respectively. Convergence in distribution of X_n to X will be denoted by $X_n \Rightarrow X$.

2. Description of model and state dynamics. In Section 2.1 we describe the basic model, which is sometimes referred to as the $GI/GI/N+G$ model. In Section 2.2 we introduce the state descriptor and some auxiliary processes and also describe the state dynamics. In Section 2.3 we obtain a convenient representation formula for expectations of linear functionals of the measure-valued components of the state process. In Section 2.4 we introduce a filtration with respect to which the state descriptor is an adapted, strong Markov process. This model was also considered in [15], where a functional strong law of large numbers limit for the state descriptor was established as the number of servers and the mean arrival rate both tend to infinity.

2.1. *Model description and primitive data.* Consider a queueing system with N identical servers in which arriving customers are served in a nonidling, first-come-first-serve (FCFS) manner, that is, a newly arriving customer immediately enters service if there are any idle servers or, if all servers are busy, then the customer joins the back of the queue, and the customer at the head of the queue (if one is present) enters service as soon as a server becomes free.

It is assumed that customers are impatient and that a customer reneges from the queue as soon as the amount of time he or she has waited in the queue reaches his or her patience time. Service is nonpreemptive and customers do not renege once they have entered service. The patience times of customers are given by an i.i.d. sequence, $\{r_i, i \in \mathbb{Z}\}$, with common cumulative distribution function G^r on $[0,\infty]$, while the service requirements of customers are given by another i.i.d. sequence, $\{v_i, i \in \mathbb{Z}\}$, with common cumulative distribution function G^s on $[0,\infty)$. For $i \in \mathbb{N}$, r_i and v_i, respectively, represent the patience time and the service requirement of the ith customer to enter the system after time zero, whereas $\{r_i, i \in -\mathbb{N} \cup \{0\}\}$ and $\{v_i, i \in -\mathbb{N} \cup \{0\}\}$, respectively, represent the patience times and the service requirements of customers that arrived prior to time zero (if such customers exist), ordered according to their arrival times (prior to time zero). We assume that G^s has density g^s and G^r, restricted to $[0,\infty)$, has density g^r, with G^r possibly having some mass at infinity. This implies, in particular, that $G^r(0+) = G^s(0+) = 0$. We define $h^r = g^r/(1-G^r)$ and $h^s = g^s/(1-G^s)$ to be the corresponding hazard rate functions associated with G^r and G^s. Let

$$H^r \doteq \sup\{x \in [0,\infty): G^r(x) < 1\},$$
$$H^s \doteq \sup\{x \in [0,\infty): G^s(x) < 1\}.$$

The superscript (N) will be used to refer to quantities associated with the system with N servers.

Let $E^{(N)}$ denote the cumulative arrival process associated with the system that has N servers, with $E^{(N)}(t)$ representing the total number of customers that arrive into the system in the time interval $[0,t]$. We assume that $E^{(N)}$ is a renewal process with a common interarrival distribution function $F^{(N)}$, which has finite mean and satisfies $F^{(N)}(0) = 0$. Let $\lambda^{(N)}$ be the inverse of the mean of $F^{(N)}$, that is,

$$\lambda^{(N)} \int_0^\infty x F^{(N)}(dx) = 1.$$

The number $\lambda^{(N)}$ represents the long-run average arrival rate of customers to the system with N servers. We assume $E^{(N)}$, the sequence of service requirements $\{v_j, j \in \mathbb{Z}\}$ and the sequence of patience times $\{r_j, j \in \mathbb{Z}\}$ are mutually independent. Let $\alpha_E^{(N)}$ be a càdlàg, real-valued process defined by $\alpha_E^{(N)}(s) \doteq \alpha_E^{(N)}(0) + s$ if $E^{(N)}(s) = 0$ and, if $E^{(N)}(s) > 0$, then

$$\alpha_E^{(N)}(s) \doteq s - \sup\{u < s : E^{(N)}(u) < E^{(N)}(s)\}.$$

Observe that the quantity $\alpha_E^{(N)}(s)$ denotes the time to s since the last arrival, and coincides with the backward recurrence time process. Moreover, the process $\alpha_E^{(N)}$ determines the process $E^{(N)}$. Let $\mathcal{E}_0^{(N)}$ be an a.s. finite $\mathbb{Z}_+$-valued random variable that represents the number of customers that entered the system prior to time zero. This random variable does not play an important role in the analysis. It is used merely for bookkeeping purposes, to keep track of the indices of customers.

2.2. *State descriptor.* A Markovian description of the state of the system with N servers would require one to keep track of the residual or elapsed patience times and the residual or elapsed service times of each customer present in the queue or in service. In order to do this in a succinct manner, with a common state space for all N-server systems, we use the representation introduced in [15]. In this representation, the state of the N-server system consists of the backward recurrence time $\alpha_E^{(N)}$ of the renewal arrival process, a nonnegative real-valued process $X^{(N)}$, which represents the total number of customers in system with N servers (including those in service and those in queue) and a pair of measure-valued processes, the "age measure" process, $\nu^{(N)}$, which encodes the amounts of time that customers currently receiving service have been in service and the "potential queue measure" process, $\eta^{(N)}$, which keeps track not only of the waiting times of customers in queue but also of the potential waiting times (defined to be the times since entry into system) of every customer (irrespective of whether the customer has already entered service and possibly departed the system) for whom the potential waiting time has not yet exceeded the patience time. Thus, the state of the system, denoted by $Y^{(N)}$, takes the form

$$Y^{(N)} = \left(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)}\right). \tag{2.1}$$

Note that $X^{(N)}$ and $\eta^{(N)}$, together, yield the number and waiting times of customers currently in queue. Indeed, for $t \in [0, \infty)$, let $Q^{(N)}(t)$ be the number of customers waiting in queue at time t. Because the head-of-the-line customer is the customer in queue with the longest waiting time, the quantity

$$\chi^{(N)}(t) \doteq \inf\{x > 0 : \eta_t^{(N)}[0, x] \geq Q^{(N)}(t)\} = \left(F^{\eta_t^{(N)}}\right)^{-1}\left(Q^{(N)}(t)\right) \tag{2.2}$$

represents the waiting time of the head-of-the-line customer in the queue at time t. Here, the function $F^{\eta_t^{(N)}}$ and its inverse are as defined in (1.3) and (1.1), respectively. Since this is an FCFS system, any mass in $\eta_t^{(N)}$ that lies to the right of $\chi^{(N)}(t)$ represents a customer that has already entered service by time t, and all masses in $[0, \chi^{(N)}(t)]$ are still in queue. Therefore, the queue length process $Q^{(N)}$ can be expressed in terms of $\chi^{(N)}$ and $\eta^{(N)}$:

$$Q^{(N)}(t) = \eta_t^{(N)}\left[0, \chi^{(N)}(t)\right], \qquad t \in [0, \infty), \tag{2.3}$$

and the restriction of $\eta_t^{(N)}$ to $[0, \chi^{(N)}(t)]$ encapsulates the waiting times of all customers in queue at time t. As explained in Section 2.2 of [15], we include in the state the measure-valued process $\eta^{(N)}$ rather than a measure-valued process that only keeps track of the waiting times of customers in queue because the dynamics of the former is easier to analyze.

We note that, due to the nonidling condition, the queue length process also admits the following alternative representation in terms of $X^{(N)}$:

$$Q^{(N)}(t) = \left[X^{(N)}(t) - N\right]^+.$$

Moreover, because

$$X^{(N)} = \langle \mathbf{1}, \nu^{(N)} \rangle + Q^{(N)}, \tag{2.4}$$

the nonidling condition is equivalent to

$$N - \langle \mathbf{1}, \nu^{(N)} \rangle = \left[N - X^{(N)}\right]^+. \tag{2.5}$$

The following auxiliary processes are useful for the evolution of the system and can be recovered from the state of the system $Y^{(N)}$ by using equations (2.9)–(2.11) and (2.14) in [15]:

- the cumulative reneging process $R^{(N)}$, where $R^{(N)}(t)$ is the cumulative number of customers that have reneged from the system in the time interval $[0, t]$;

- the cumulative potential reneging process $S^{(N)}$, where $S^{(N)}(t)$ represents the cumulative number of customers whose potential waiting times reached their patience times in the interval $[0, t]$;
- the cumulative departure process $D^{(N)}$, where $D^{(N)}(t)$ is the cumulative number of customers that departed the system after completion of service in the interval $[0, t]$;
- the process $K^{(N)}$, where $K^{(N)}(t)$ represents the cumulative number of customers that entered service in the interval $[0, t]$.

It is easy to see from (2.16) in [15] that the following mass balance for the number of customers in queue holds:

$$Q^{(N)}(0) + E^{(N)} = Q^{(N)} + R^{(N)} + K^{(N)}. \tag{2.6}$$

2.3. *A useful representation formula.* We now establish representation formulae (in Proposition 2.2) for expectations of linear functionals of the age and potential queue measure-valued processes. These are used to establish tightness of the sequence of stationary distributions in Section 4.2. This representation formula is similar to that obtained for the fluid in Theorem 4.1 of [17]. The representation can be deduced from a result given in Proposition 4.1 of [15] which, for convenience, we first reproduce below as Proposition 2.1.

PROPOSITION 2.1. *Let G be the cumulative distribution function of a probability distribution with density g and hazard rate function $h = g/(1 - G)$, let $H \doteq \sup\{x \in [0, \infty) : G(x) < 1\}$. Suppose $\overline{\pi} \in \mathcal{D}_{\mathcal{M}_F[0,H)}[0, \infty)$ has the property that for every $L \in [0, H)$ and $T \in [0, \infty)$, there exists $C(L, T) < \infty$ such that*

$$\int_0^\infty < \varphi(\cdot, s)h(\cdot), \qquad \overline{\pi}_s > ds < C(L, T)\|\varphi\|_\infty, \tag{2.7}$$

for every $\varphi \in \mathcal{C}_c((-\infty, H) \times \mathbb{R})$ with $\operatorname{supp}(\varphi) \subset [0, L] \times [0, T]$. Then, given any $\overline{\pi}_0 \in \mathcal{M}_F[0, H)$ and $\overline{Z} \in \mathcal{I}_{\mathbb{R}_+}[0, \infty)$, $\overline{\pi}$ satisfies the integral equation

$$\begin{aligned} \langle \varphi(\cdot, t), \overline{\pi}_t \rangle = \langle \varphi(\cdot, 0), \overline{\pi}_0 \rangle &+ \int_0^t \langle \varphi_s(\cdot, s), \overline{\pi}_s \rangle\, ds + \int_0^t \langle \varphi_x(\cdot, s), \overline{\pi}_s \rangle\, ds \\ &- \int_0^t \langle \varphi(\cdot, s)h(\cdot), \overline{\pi}_s \rangle\, ds + \int_{[0,t]} \varphi(0, s)\, d\overline{Z}(s) \end{aligned} \tag{2.8}$$

for every $\varphi \in \mathcal{C}_c((-\infty, H) \times \mathbb{R})$ and $t \in [0, \infty)$, if and only if $\overline{\pi}$ satisfies

$$\begin{aligned} \int_{[0,M)} f(x)\overline{\pi}_t(dx) = \int_{[0,M)} & f(x + t)\frac{1 - G(x + t)}{1 - G(x)}\overline{\pi}_0(dx) \\ &+ \int_{[0,t]} f(t - s)\big(1 - G(t - s)\big)\, d\overline{Z}(s) \end{aligned} \tag{2.9}$$

for every $f \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \in (0, \infty)$.

We now use this general result to obtain a useful representation formula, which can also alternatively be deduced by taking expectations in the representation formula provided in Proposition 6.4 of [16].

PROPOSITION 2.2. *Suppose that* $\mathbb{E}[\langle \mathbf{1}, \eta_0^{(N)}\rangle] < \infty$ *and* $\mathbb{E}[\langle \mathbf{1}, \nu_0^{(N)}\rangle] < \infty$. *Then for each bounded measurable function* f *on* $\mathbb{R}_+$ *and* $t \geq 0$,

$$\begin{aligned}
\mathbb{E}[\langle f, \eta_t^{(N)}\rangle] &= \mathbb{E}\left[\int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\eta_0^{(N)}(dx)\right] \\
&\quad + \mathbb{E}\left[\int_{[0,t]} f(t-s)\bigl(1-G^r(t-s)\bigr)\,dE^{(N)}(s)\right]
\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}
\mathbb{E}[\langle f, \nu_t^{(N)}\rangle] &= \mathbb{E}\left[\int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\nu_0^{(N)}(dx)\right] \\
&\quad + \mathbb{E}\left[\int_{[0,t]} f(t-s)\bigl(1-G^s(t-s)\bigr)\,dK^{(N)}(s)\right].
\end{aligned} \tag{2.11}$$

PROOF. We provide the details of the proof of (2.10) only, because the proof of (2.11) is exactly analogous. Fix $N \in \mathbb{N}$ and define $\overline{\pi} \doteq \mathbb{E}[\eta^{(N)}]$ and $\overline{Z} \doteq \mathbb{E}[E^{(N)}]$, $G \doteq G^r$ and $h \doteq h^r$. By Proposition 2.1, in order to establish (2.10) it suffices to show that (2.7) and (2.8) are satisfied with $\overline{\pi}$ and $\overline{Z}$ defined as above. However, these are easily deduced from properties established in [15]. Indeed, by the analog of (5.4) of Proposition 5.1(2) in [15], we know that

$$\mathbb{E}\left[\int_0^T \langle \varphi(\cdot,s)h^r(\cdot), \eta_s^{(N)}\rangle\,ds\right] \leq C(L,T)\|\varphi\|_\infty, \tag{2.12}$$

where $C(L,T) \doteq (\int_0^L h^r(x)\,dx)\mathbb{E}[X^{(N)}(0) + E^{(N)}(T)]$ is finite because of the supposition of the proposition, the relation $\overline{X}(0) \leq \langle \mathbf{1}, \eta_0^{(N)}\rangle + \langle \mathbf{1}, \nu_0^{(N)}\rangle$ and the fact that $E^{(N)}$ is a renewal process with finite mean. Thus, (2.12) implies (2.7). On the other hand, for every $\varphi \in \mathcal{C}_c^1([0,H^r) \times \mathbb{R}_+)$, (2.28) of Theorem 2.1 of [15] implies that for every $t \in (0,\infty)$,

$$\begin{aligned}
\langle \varphi(\cdot,t), \eta_t^{(N)}\rangle &= \langle \varphi(\cdot,0), \eta_0^{(N)}\rangle + \int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s), \eta_s^{(N)}\rangle\,ds \\
&\quad - S_\varphi^{(N)}(t) + \int_{[0,t]} \varphi(0,s)\,dE^{(N)}(s),
\end{aligned} \tag{2.13}$$

and Proposition 5.1(2) of [15] shows that

$$M_{\varphi,\eta}^{(N)} \doteq S_\varphi^N - \int_0^t \langle \varphi(\cdot,s)h^r(\cdot), \eta_s^{(N)}\rangle\,ds$$

is a local $\{\mathcal{F}_t^{(N)}\}$ martingale. In fact, $M_{\varphi,\eta}^{(N)}$ is an $\{\mathcal{F}_t^{(N)}\}$-martingale because

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in[0,t]}\big|M_{\varphi,\eta}^{(N)}(s)\big|\Big] &\le \mathbb{E}[S_\varphi^N(t)] + \mathbb{E}\bigg[\int_0^T \langle|\varphi(\cdot,s)|h^r(\cdot), \eta_s^{(N)}\rangle\,ds\bigg]\\ &\le \|\varphi\|_\infty \mathbb{E}\big[E^{(N)}(t)\big] + C(L,T)\|\varphi\|_\infty < \infty,\end{aligned}$$

where the finiteness follows from the assumption that $E^{(N)}$ is a renewal process with finite mean. The relation (2.8) then follows on taking expectations of both sides of (2.13) and interchanging the expectation with integration. Hence, the representation (2.10) follows. □

2.4. *State space and filtration.* The total number of customers in service at time t is given by

$$\langle\mathbf{1}, \nu_t^{(N)}\rangle = \nu_t^{(N)}[0, H^s)$$

and is bounded above by the number of servers N. On the other hand, it is clear (see, e.g., (2.13) of [15]) that a.s., for every $t \in [0,\infty)$,

$$\langle\mathbf{1}, \eta_t^{(N)}\rangle = \eta_t^{(N)}[0, H^r) \le E^{(N)}(t) + \langle\mathbf{1}, \eta_0^{(N)}\rangle \le E^{(N)}(t) + \mathcal{E}_0^{(N)} < \infty.$$

Therefore, a.s., for every $t \in [0,\infty)$, $\nu_t^{(N)} \in \mathcal{M}_F[0, H^s)$ and $\eta_t^{(N)} \in \mathcal{M}_F[0, H^r)$.

Let $\mathcal{M}_D[0, H^s)$ be the subset of measures in $\mathcal{M}_F[0, H^s)$ that can be represented as the sum of a finite number of unit Dirac measures in $[0, H^s)$, that is, measures that take the form $\sum_{i=1}^k \delta_{x_i}$ for some $k \in \mathbb{Z}_+$ and $x_i \in [0, H^s)$, $i = 1, \dots, k$. Analogously, let $\mathcal{M}_D[0, H^r)$ be the subset of $\mathcal{M}_F[0, H^r)$ that can be expressed as the sum of a finite number of unit Dirac measures in $[0, H^r)$. Also, define

$$\begin{aligned}\mathcal{Y}^{(N)} \doteq \{&(\alpha, x, \mu, \pi) \in \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_D[0, H^s) \times \mathcal{M}_D[0, H^r) :\\ &x \le \langle\mathbf{1}, \mu\rangle + \langle\mathbf{1}, \pi\rangle, \langle\mathbf{1}, \mu\rangle \le N\},\end{aligned} \tag{2.14}$$

where $\mathbb{R}_+$ is endowed with the Euclidean topology d, $\mathbb{Z}_+$ is endowed with the discrete topology ρ and $\mathcal{M}_D[0, H^s)$ and $\mathcal{M}_D[0, H^r)$ are both endowed with the topology of weak convergence. The space $\mathcal{Y}^{(N)}$ is a closed subset of $\mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ and is endowed with the usual product topology. Since $\mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ is a Polish space, the closed subset $\mathcal{Y}^{(N)}$ is also a Polish space. It follows from the representations for $\nu_t^{(N)}$ and $\eta_t^{(N)}$ given in (2.3) and (2.8) of [15] that a.s., the state descriptor $Y^{(N)}(t)$ takes values in $\mathcal{Y}^{(N)}$ for every $t \in [0,\infty)$.

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

$$\begin{aligned}\big\{&\mathcal{E}_0^{(N)}, X^{(N)}(0), \alpha_E^{(N)}(s), w_j^{(N)}(s), a_j^{(N)}(s), s_j^{(N)},\\ &j \in \{-\mathcal{E}_0^{(N)} + 1, \dots, 0\} \cup \mathbb{N}, s \in [0,t]\big\},\end{aligned}$$

where $s^{(N)} \doteq (s_j^{(N)}, j \in \mathbb{Z})$ is the "station process," defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $t \in [0, \infty)$, if customer j has already entered service by time t, then $s_j^{(N)}(t)$ is equal to the index $i \in \{1, \dots, N\}$ of the station at which customer j receives service and $s_j^{(N)}(t) \doteq 0$ otherwise. Let $\{\mathcal{F}_t^{(N)}\}$ denote the associated right-continuous filtration, completed with respect to $\mathbb{P}$. It is proved in Appendix A of [15] that the state descriptor $Y^{(N)}$ and the auxiliary processes $E^{(N)}$, $Q^{(N)}$, $S^{(N)}$, $R^{(N)}$, $D^{(N)}$ and $K^{(N)}$ are càdlàg and adapted to the filtration $\{\mathcal{F}_t^{(N)}\}$. Moreover, from Lemma B.1 of [15] it follows that $Y^{(N)}$ is a strong Markov process with respect to the filtration $\{\mathcal{F}_t^{(N)}\}$.

3. Assumptions and main results.

The main focus of this paper is to obtain a "first-order" approximation for the stationary distribution of the N-server queue, which is accurate in the limit as the number of servers goes to infinity.

3.1. *Basic assumptions*. We impose the following mild first moment assumption on the patience and service time distribution functions G^r and G^s. Without loss of generality, we can normalize the service time distribution so that its mean equals 1.

ASSUMPTION 3.1. The mean patience and service times are finite:

$$\theta^r \doteq \int_{[0,\infty)} x g^r(x)\, dx = \int_{[0,\infty)} \big(1 - G^r(x)\big)\, dx < \infty \tag{3.1}$$

and

$$\int_{[0,\infty)} x g^s(x)\, dx = \int_{[0,\infty)} \big(1 - G^s(x)\big)\, dx = 1. \tag{3.2}$$

Let ν_* and η_* be the probability measures defined as follows:

$$\nu_*[0, x) \doteq \int_0^x \big(1 - G^s(y)\big)\, dy, \qquad x \in [0, H^s), \tag{3.3}$$

$$\eta_*[0, x) \doteq \int_0^x \big(1 - G^r(y)\big)\, dy, \qquad x \in [0, H^r). \tag{3.4}$$

Note that ν_* and η_* are well defined due to Assumption 3.1. For $\lambda \geq 1$, define the set B_λ as follows:

$$B_\lambda \doteq \left\{ x \in [1, \infty) : G^r\big((F^{\lambda\eta_*})^{-1}\big((x-1)^+\big)\big) = \frac{\lambda - 1}{\lambda} \right\}. \tag{3.5}$$

Let

$$b_l^\lambda \doteq \inf\{x \in [1, \infty) : x \in B_\lambda\} \quad \text{and} \quad b_r^\lambda \doteq \sup\{x \in [1, \infty) : x \in B_\lambda\}.$$

Since the functions G^r and $F^{\lambda\eta_*}$ are continuous and nondecreasing, we have $B_\lambda = [b_l^\lambda, b_r^\lambda]$. Let $\mathcal{I}_\lambda$ be the set of states defined by

$$\mathcal{I}_\lambda \doteq \begin{cases} \{(\lambda, \lambda\nu_*, \lambda\eta_*)\}, & \text{if } \lambda < 1, \\ \{(x_*, \nu_*, \lambda\eta_*) : x_* \in B_\lambda\}, & \text{if } \lambda \geq 1. \end{cases} \tag{3.6}$$

We show in Theorem 5.5 that $\mathcal{I}_\lambda$ describes the so-called invariant manifold for the fluid limit. Suppose that $\mathcal{I}_\lambda$ satisfies the following assumption.

ASSUMPTION 3.2. The set $\mathcal{I}_\lambda$ has a single element.

Note that this is a nontrivial restriction only when $\lambda \geq 1$. A deterministic fluid limit of the $GI/GI/N + G$ queue was conjectured to exist in Conjecture 2.2 of [22], and Theorem 3.1 of [22] states that this fluid limit has a unique steady state. However, as shown in the example in Section 7.2, in general there need not be a unique invariant state (or, equivalently, a unique steady state in the sense of [22]) due to the possibility of the existence of multiple solutions to the equation (3.7) below. Thus, we explicitly assume uniqueness of the steady state to obtain the full convergence result. We now provide a general sufficient condition for Assumption 3.2 to hold.

LEMMA 3.1. *If either $\lambda < 1$ or $\lambda \in [1, \infty)$ and the equation*

$$G^r(x) = \frac{\lambda - 1}{\lambda} \tag{3.7}$$

has a unique solution, then Assumption 3.2 holds. In particular, this is true if G^r is strictly increasing.

PROOF. Fix $\lambda \in [1, \infty)$. It suffices to show that the set B_λ in (3.5) consists of a single point. Since the equation in (3.7) has a unique solution and the function $(F^{\lambda\eta_*})^{-1}(\cdot)$ is strictly increasing on $[0, \lambda\theta^r)$, the equation

$$G^r\big((F^{\lambda\eta_*})^{-1}\big((x-1)^+\big)\big) = \frac{\lambda - 1}{\lambda}$$

has a unique solution. Thus, B_λ has a single element and the lemma follows. □

For each $N \in \mathbb{N}$, let $\overline{Y}^{(N)} = (\overline{\alpha}_E^{(N)}, \overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ be the fluid scaled state descriptor defined as follows: for $t \in [0, \infty)$ and any Borel subset B of $\mathbb{R}_+$,

$$\begin{aligned} \overline{\alpha}_E^{(N)}(t) &\doteq \alpha_E^{(N)}(t), \qquad & \overline{X}^{(N)}(t) &\doteq \frac{X^{(N)}(t)}{N}, \\ \overline{\nu}_t^{(N)}(B) &\doteq \frac{\nu_t^{(N)}(B)}{N}, \qquad & \overline{\eta}_t^{(N)}(B) &\doteq \frac{\eta_t^{(N)}(B)}{N}. \end{aligned} \tag{3.8}$$

Analogously, for $I = E, D, K, Q, R, S$, define

$$\overline{I}^{(N)} \doteq \frac{I^{(N)}}{N}. \tag{3.9}$$

The following standard assumption is imposed on the sequences of fluid scaled external arrival processes $\{\overline{E}^{(N)}\}$ and initial conditions $(\eta_0^{(N)}, \nu_0^{(N)})$, $N \in \mathbb{N}$.

ASSUMPTION 3.3. The following conditions are satisfied:

(1) There exists $\lambda \in [0, \infty)$ such that $\overline{\lambda}^{(N)} = \lambda^{(N)}/N \to \lambda$ as $N \to \infty$;
(2) As $N \to \infty$, $\overline{E}^{(N)} \to \overline{E}$ in $\mathcal{D}_{\mathbb{R}_+}[0, \infty)$ $\mathbb{P}$-a.s., where $\overline{E}(t) = \lambda t$;
(3) $\mathbb{E}[\langle \mathbf{1}, \eta_0^{(N)} \rangle] < \infty$ and $\mathbb{E}[\langle \mathbf{1}, \nu_0^{(N)} \rangle] < \infty$ for each $N \in \mathbb{N}$.

The following technical assumption was imposed on the hazard rate functions in [15] to establish the fluid limit theorem.

ASSUMPTION 3.4. There exists $L^s < H^s$ such that h^s is either bounded or lower-semicontinuous on (L^s, H^s), and likewise, there exists $L^r < H^r$ such that h^r is either bounded or lower-semicontinuous on (L^r, H^r).

We conclude with a mild assumption on the interarrival distribution function $F^{(N)}$.

ASSUMPTION 3.5. The interarrival distribution $F^{(N)}$ has a density.

3.2. *Main results*. The first result focuses on the existence of a stationary distribution for the state process.

THEOREM 3.2. *For each N, under Assumption 3.5, $\{Y_t^{(N)}, \mathcal{F}_t^{(N)}\}$ is a Feller process that has a stationary distribution.*

The Feller property is proved in Proposition 4.2 and the existence of a stationary distribution is established in Theorem 4.9. In Theorem 7.1, the state process is also shown to be ergodic under an additional condition (Assumption 7.1) which holds, for example, when the interarrival, reneging and service densities are strictly positive and the latter two have support on $(0, \infty)$.

We now state the main result, which provides a first-order approximation for stationary distributions of N-server queues.

THEOREM 3.3. *Suppose Assumptions 3.1, 3.3 and 3.5 hold and for $N \in \mathbb{N}$, let $\overline{Y}_*^{(N)} = (\overline{\alpha}_{E,*}^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$ be a scaled stationary distribution for the N-*

server queue with abandonment. Then the sequence $\overline{Y}_*^{(N)}$, $N \in \mathbb{N}$, *is tight. If, in addition, Assumption* 3.4 *holds, then the limit of any convergent subsequence of the sequence* $(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$, $N \in \mathbb{N}$, *almost surely takes values on the invariant manifold* $\mathcal{I}_\lambda$. *Furthermore, if Assumption* 3.2 *also holds, then the sequence* $(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$, $N \in \mathbb{N}$, *converges to the unique element of* $\mathcal{I}_\lambda$.

A related discrete-time result was conjectured in Theorem 7.2 of Whitt [22]. In particular, Theorem 7.2 of [22] states that if the discrete model introduced in [22] satisfies the assumptions that (i) each N-server queueing system converges (for large times) to a unique stationary distribution; (ii) the sequence of fluid-scaled stationary distributions is tight; and (iii) the sequence of fluid-scaled stationary distributions has a weak limit as $N \to \infty$, this limit must be equal to the unique steady-state associated with the fluid model described in [22]. The validity of properties (ii) and (iii) was not established in [23]. In contrast, we consider the continuous model, and for this model establish tightness and (under the additional assumption that there is a unique invariant state) existence of a weak limit. The proof of Theorem 3.3 is given in Section 6 and consists of the following main steps. In Theorem 3.2, the Markovian nature of the state representation is used to establish the existence of a stationary distribution for each N-server system. In Theorem 6.2 a convenient representation for the state dynamics in the N-server system (see Proposition 2.2) is used to establish tightness of any sequence of fluid-scaled stationary distributions. It is shown in Section 7.2 that, in general, the steady state (equivalently an element of the invariant manifold) need not in fact be unique. Nevertheless, it is shown that any subsequential limit must be an invariant state, and that when there is a unique invariant state, the desired convergence follows. Sufficient conditions for uniqueness of the invariant state are given in Lemma 3.1.

The characterization of the stationary distribution and a better understanding of the possible metastable behavior of the N-server queue in the presence of multiple invariant states for the fluid remains a subject for future investigation.

4. Stationary distribution of the N-server queue. We now establish the existence of a stationary distribution for the Markovian state descriptor $\{Y_t^{(N)}, \mathcal{F}_t^{(N)}\}$ for the system with N servers, under Assumption 3.5. First, in Section 4.1, $\{Y_t^{(N)}, \mathcal{F}_t^{(N)}\}_{t\geq 0}$ is shown to be a Feller process (see Proposition 4.2). Then, in Section 4.2, the Krylov–Bogoliubov existence theorem (cf. Corollary 3.1.2 of [6]) is used to show that $\{Y_t^{(N)}, \mathcal{F}_t^{(N)}\}_{t\geq 0}$ has a stationary distribution. Finally, in the Appendix, ergodicity and positive Harris recurrence of the process $\{Y_t^{(N)}, \mathcal{F}_t^{(N)}\}_{t\geq 0}$ is established under an additional condition (Assumption 7.1). For conciseness, in the rest of this section, N is fixed and the dependence on N is omitted from the notation.

4.1. *Feller property.* It follows from the definition of Y in (2.1) and Lemma B.1 of [15] that Y is a so-called piecewise deterministic Markov process with jump times $\{\tau_1, \tau_2, \dots\}$ (see [12] for a precise definition of piecewise deterministic Markov processes), where each jump time is either the arrival time of a new customer, the time of a service completion or the end of a patience time. Note that the set of jump times also includes the time of entry into service of each customer because, due to the nonidling condition, each such entry time coincides with either the arrival time of that customer or the time of service completion of another customer. Let $\tau_0 = 0$. For each $i \in \mathbb{Z}_+$, Y evolves in a deterministic fashion on $[\tau_i, \tau_{i+1})$,

$$Y(\tau_i + t) = \phi_{Y(\tau_i)}(t), \qquad t \in [0, \tau_{i+1} - \tau_i),$$

where, for each $y \in \mathcal{Y}$ of the form $y = (\alpha, x, \sum_{i=1}^k \delta_{u_i}, \sum_{j=1}^l \delta_{z_j})$, $k, l \in \mathbb{N}$, $k \le N$, we define

$$\phi_y(t) \doteq \left(\alpha + t, x, \sum_{i=1}^{k} \delta_{u_i+t}, \sum_{j=1}^{l} \delta_{z_j+t}\right), \qquad t \ge 0. \tag{4.1}$$

The Markovian semigroup of Y is defined in the usual way: for each $t \ge 0$, $y \in \mathcal{Y}$ and $A \in \mathcal{B}(\mathcal{Y})$, the set of Borel subsets of $\mathcal{Y}$, let

$$P_t(y, A) \doteq \mathbb{P}\big(Y(t) \in A | Y(0) = y\big). \tag{4.2}$$

Moreover, for any measurable function ψ defined on $\mathcal{Y}$ and $t \ge 0$, let $P_t\psi$ be the function on $\mathcal{Y}$ given by

$$P_t\psi(y) \doteq \mathbb{E}[\psi(Y(t)) | Y(0) = y], \qquad y \in \mathcal{Y}. \tag{4.3}$$

We now show that the semigroup $\{P_t, t \ge 0\}$ is Feller in the sense of [6] (see the beginning of Section 3.1 therein), that is, we show that for any $\psi \in C_b(\mathcal{Y})$ and $t \ge 0$, $P_t\psi \in C_b(\mathcal{Y})$.

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

$$Y^m(0) = y^m = \left(\alpha^m, x^m, \sum_{i=1}^{k^m} \delta_{u_i^m}, \sum_{j=1}^{l^m} \delta_{z_j^m}\right) \in \mathcal{Y}$$

for some $k^m \in \{0, \dots, N\}$ and $l^m \in \mathbb{N}$. Suppose that $\{Y^m, m \in \mathbb{Z}_+\}$ are defined on the same probability space and y^m converges to y^0 as $m \to \infty$. Due to the nature of the topology on $\mathcal{Y}$, the convergence of y^m to y^0 implies that $x^m = x^0, k^m = k^0, l^m = l^0$ for all sufficiently large m and, as $m \to \infty$, $\alpha^m \to \alpha^0$, $u_i^m \to u_i^0$ and $z_j^m \to z_j^0$ for each $1 \le i \le k^0, 1 \le j \le l^0$. Without loss of generality, we may assume that $x^m = x^0, k^m = k^0, l^m = l^0$ for every $m \in \mathbb{Z}_+$. For the mth N-server system, $m \in \mathbb{Z}_+$, the time since the arrival of the last customer before time 0 is α^m

and hence, the random time to the arrival of the first customer after time 0 has distribution function $F(\alpha^m + \cdot)/(1 - F(\alpha^m))$, which has a density by Assumption 3.5. Likewise, the distribution of the residual patience time of the initial customer associated with the point mass $\delta_{z_j^m}$ has density $g^r(z_j^m + \cdot)/(1 - G^r(z_j^m))$ and the distribution of the residual service time of the initial customer associated with the point mass $\delta_{u_i^m}$ has density $g^s(u_i^m + \cdot)/(1 - G^s(u_i^m))$. For simplicity, henceforth we will denote k^0, l^0, x^0 simply by k, l, x. We assume that the elements of the sequence $\{Y^m, m \in \mathbb{Z}_+\}$ are coupled so that:

- the interarrival times after the first arrival and the sequences of service times and patience times of customers that arrive after time 0 are identical for each N-server queue $Y^m, m \in \mathbb{Z}_+$;
- the first arrival time of a new customer in the mth N-server queue converges to the first arrival time in the 0th N-server queue (note that this is equivalent to the convergence of the residual interarrival times at time zero in the corresponding systems);
- for each $j = 1, \dots, l$, the residual patience time of the customer associated with the point mass $\delta_{z_j^m}$ converges, as $m \to \infty$, to the residual patience time of the customer associated with the point mass $\delta_{z_j^0}$;
- for each $i = 1, \dots, k$, the residual service time of the customer associated with the point mass $\delta_{u_i^m}$ converges, as $m \to \infty$, to the residual service time of the customer associated with the point mass $\delta_{u_i^0}$.

LEMMA 4.1. *Suppose Assumption 3.5 holds. For each $m \in \mathbb{Z}_+$ and $n \in \mathbb{N}$, let τ_n^m be the nth jump time of Y^m. Then for each $n \in \mathbb{N}$, τ_n^m converges to τ_n^0 and $Y^m(\tau_n^m)$ converges in $\mathcal{Y}$ to $Y^0(\tau_n^0)$ a.s., as $m \to \infty$.*

PROOF. We prove the lemma by an induction argument. First, consider $n = 1$. For each $m \in \mathbb{Z}_+$, the first jump time τ_1^m is the minimum of the first arrival time of a new customer, the residual patience times of initial customers with potential waiting times in the set $\{z_j^m, 1 \le j \le l\}$ and the residual service times of initial customers associated with ages in the set $\{u_i^m, 1 \le i \le k\}$. It follows directly from the assumptions on $\{Y^m, m \in \mathbb{Z}_+\}$ that for every realization,

$$\tau_1^m \to \tau_1^0, \qquad \text{as } m \to \infty. \tag{4.4}$$

Since the interarrival distribution F, the service time distribution function G^s and the patience time distribution function G^r are independent and have densities, with probability 1, τ_1^0 coincides with exactly one of the following in the 0th system: the first arrival time of a new customer, the residual patience time of an initial customer with initial waiting time $z_j^0, 1 \le j \le l$, or the residual service time of an initial customer with age $u_i^0, 1 \le i \le k$. Let us fix a realization such that τ_1^0 is equal to the first arrival time of a new customer in the 0th system. The remaining two cases can be handled similarly. In this case, by the convergence of τ_1^m to τ_1^0, the convergence of the other quantities stated above and the coupling construction, for all sufficiently large m, τ_1^m is equal to the first arrival time of a new customer in the mth system. Hence, for all sufficiently large m, the first jump of Y^m is due to the first arrival of a new customer in the mth system. For such m, since Y^m evolves in a deterministic fashion on $[0, \tau_1^m)$ described by the continuous function ϕ introduced in (4.1), we have

$$Y^m(\tau_1^m-) = \left(\alpha^m + \tau_1^m, x, \sum_{i=1}^k \delta_{u_i^m+\tau_1^m}, \sum_{j=1}^l \delta_{z_j^m+\tau_1^m}\right).$$

If $k = N$ and $x \geq k = N$, then all the servers are busy and the customer that arrives at τ_1^m will have to wait in queue. Thus, by the coupling construction,

$$Y^m(\tau_1^m) = \left(0, x+1, \sum_{i=1}^k \delta_{u_i^m+\tau_1^m}, \sum_{j=1}^l \delta_{z_j^m+\tau_1^m} + \delta_0\right).$$

On the other hand, if $k < N$, then $x = k$ and there is at least one idle server present. Hence, the customer will join service immediately upon arrival at time τ_1^m. Thus, in this case,

$$Y^m(\tau_1^m) = \left(0, x+1, \sum_{i=1}^k \delta_{u_i^m+\tau_1^m} + \delta_0, \sum_{j=1}^l \delta_{z_j^m+\tau_1^m} + \delta_0\right).$$

In both cases, for the chosen realization, we have $Y^m(\tau_1^m) \to Y^0(\tau_1^0)$ as $m \to \infty$.

Now, suppose that τ_i^m converges to τ_i^0 and $Y^m(\tau_i^m)$ converges to $Y^0(\tau_i^0)$ a.s., as $m \to \infty$, for $1 \leq i \leq n$, and consider $i = n+1$. Fix a realization such that τ_n^m converges to τ_n^0 and $Y^m(\tau_n^m)$ converges to $Y^0(\tau_n^0)$ as $m \to \infty$. By the same argument as in the case $n = 1$, we may assume, without loss of generality, that for the chosen realization and $m \in \mathbb{Z}_+$, the jump at τ_n^m for Y^m is due to the arrival of a new customer. Then, for each $m \in \mathbb{Z}_+$, $Y^m(\tau_n^m)$ has the following representation:

$$Y^m(\tau_n^m) = \left(0, x_n^m, \sum_{i=1}^{k_n^m} \delta_{u_{i,n}^m}, \sum_{j=1}^{l_n^m} \delta_{z_{j,n}^m}\right)$$

for some $k_n^m, l_n^m, x_n^m \in \mathbb{Z}_+$, $u_{i,n}^m, z_{j,n}^m \in \mathbb{R}_+$ with $x_n^m \leq k_n^m + l_n^m$, $k_n^m \leq N$. Due to the induction hypothesis and the topology of $\mathcal{Y}$, for all sufficiently large m, $x_n^m = x_n^0$, $k_n^m = k_n^0$, $l_n^m = l_n^0$, $u_{i,n}^m \to u_{i,n}^0$ and $z_{j,n}^m \to z_{j,n}^0$ as $m \to \infty$ for each $1 \leq i \leq k_n^0$ and $1 \leq j \leq l_n^0$. The argument that was used for the case $n = 1$ can be used again to show that τ_{n+1}^m converges to τ_{n+1}^0 and $Y^m(\tau_{n+1}^m)$ converges to $Y^0(\tau_{n+1}^0)$ a.s., as $m \to \infty$. This completes the induction argument and hence, proves the lemma. □

PROPOSITION 4.2. *Suppose that the interarrival distribution F has a density. Then the semigroup $\{P_t, t \geq 0\}$ is Feller.*

PROOF. It is easy to see from the definition of the function $P_t\psi$ in (4.3) that when ψ is bounded, $P_t\psi$ is also bounded. To prove the proposition, it suffices to show that $P_t\psi$ is a continuous function with respect to the topology on $\mathcal{Y}$. Fix $t \geq 0$. Let $y^0 = (\alpha^0, x^0, \mu^0, \pi^0) \in \mathcal{Y}$ and $y^m = (\alpha^m, x^m, \mu^m, \pi^m)$, $m \in \mathbb{Z}_+$, be points in $\mathcal{Y}$ such that, as $m \to \infty$, y^m converges in $\mathcal{Y}$ to y^0. Since $\mathbb{Z}_+$ is a discrete space and $x^m \to x^0$ as $m \to \infty$, it must be that for all sufficiently large m, $x^m = x^0$. Without loss of generality, we assume that $x^m = x^0$ for each $m \in \mathbb{N}$. Consider a sequence of coupled N-server queues $\{Y^m, m \in \mathbb{Z}_+\}$ carried out earlier such that $Y^m(0) = y^m$ for each $m \in \mathbb{Z}_+$. Then $P_t\psi(y^m) = \mathbb{E}[\psi(Y^m(t))]$. To prove the continuity of $P_t\psi$, it suffices to show that $Y^m(t) \to Y^0(t)$ a.s., as $m \to \infty$. Indeed, since $\psi \in C_b(\mathcal{Y})$, the latter convergence would imply that $\psi(Y^m(t)) \to \psi(Y^0(t))$ and hence, by the bounded convergence theorem, that $P_t\psi(y^m) \to P_t\psi(y^0)$ as $m \to \infty$, which would show that $\{P_t, t \geq 0\}$ is Feller.

It only remains to prove that almost surely, $Y^m(t) \to Y^0(t)$ as $m \to \infty$. Since the interarrival distribution F, service distribution G^s and patience distribution G^r all have densities, with probability one t does not belong to the set $\{\tau_n^0, n \in \mathbb{N}\}$ of jump times of Y^0. Fix a realization such that t does not belong to the set $\{\tau_n^0, n \in \mathbb{N}\}$ and such that for each $n \in \mathbb{N}$, τ_n^m converges to τ_n^0 and $Y^m(\tau_n^m)$ converges in $\mathcal{Y}$ to $Y^0(\tau_n^0)$, as $m \to \infty$. By Lemma 4.1, this can be done on a set of probability one. Let $r \doteq \sup\{n : \tau_n^0 < t\}$. Then $\tau_r^0 < t < \tau_{r+1}^0$ and hence, for all sufficiently large m, $\tau_r^m < t < \tau_{r+1}^m$. By the convergence of τ_r^m to τ_r^0 and $Y^m(\tau_r^m)$ to $Y^0(\tau_r^0)$, as $m \to \infty$, as well as the definition of ϕ in (4.1), we conclude that $Y^m(t) \to Y^0(t)$, as $m \to \infty$. Thus, we have shown that $Y^m(t) \to Y^0(t)$ a.s., as $m \to \infty$. □

4.2. *Existence of stationary distributions.* In this section, it is shown that the Feller process $\{Y_t, \mathcal{F}_t\}_{t\geq 0}$ admits a stationary distribution. To achieve this, we apply the Krylov–Bogoliubov theorem (cf. Corollary 3.1.2 of [6]) which requires showing that the following family $\{L_t, t \geq 0\}$ of probability measures associated with $\{Y_t, \mathcal{F}_t\}_{t\geq 0}$ is tight. For each measurable set $B \subset \mathcal{Y}$ and $t > 0$, define

$$L_t(B) \doteq \frac{1}{t}\int_0^t \mathbb{P}(Y(s) \in B)\, ds.$$

Obviously, for each $t \geq 0$, L_t is a probability measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. We now recall some useful criteria for tightness of a family of random measures, which can be derived from A7.5 of [14] (see also Exercise 4.11 of [14]).

PROPOSITION 4.3. *A family $\{\pi_t\}_{t\geq 0}$ of $\mathcal{M}_F[0, H)$-valued random variables is tight if the following two conditions hold*:

(1) $\sup_{t\geq 0} \mathbb{E}[\langle \mathbf{1}, \pi_t\rangle] < \infty$;
(2) $\lim_{c\to H} \sup_{t\geq 0} \mathbb{E}[\pi_t[c, H)] \to 0$.

LEMMA 4.4. *Suppose Assumption* 3.1 *holds and* $\mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] < \infty$. *Then* $\sup_{t\geq 0} \mathbb{E}[\langle \mathbf{1}, \eta_t\rangle] < \infty$ *and* $\sup_{t\geq 0} \mathbb{E}[\langle \mathbf{1}, \nu_t\rangle] < \infty$.

PROOF. Let $f = \mathbf{1}$ in (2.10) and (recalling that the superscript N is being suppressed from the notation) let $e(t) \doteq \mathbb{E}[E(t)]$, $t \geq 0$. Using integration-by-parts, it follows that

$$\begin{aligned}\mathbb{E}[\langle \mathbf{1}, \eta_t\rangle] &\leq \mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] + \int_0^t \big(1 - G^r(t-s)\big)\, de(s) \\ &= \mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] + e(t) - \int_0^t e(s) g^r(t-s)\, ds \\ &= \mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] + e(t)\big(1 - G^r(t)\big) - \int_0^t \big(e(t) - e(t-s)\big) g^r(s)\, ds.\end{aligned}$$

Since E is a renewal process with rate λ, $e(t)/t \to \lambda$ as $t \to \infty$ by the key renewal theorem. Moreover, the finite mean condition (3.1) implies $t(1 - G^r(t)) \to 0$ as $t \to \infty$. Therefore, we have $\sup_{t\geq 0} e(t)(1 - G^r(t)) < \infty$. The Blackwell renewal theorem (cf. Theorem 4.3 of [1]) implies that $e(t) - e(t-s) \to s\lambda$ as $t \to \infty$ and hence, that $\sup_{t\geq 0} \int_0^t (e(t) - e(t-s)) g^r(s)\, ds < \infty$. Combining these relations with (3) of Assumption 3.3 and the last display, we conclude that $\sup_{t\geq 0} \mathbb{E}[\langle \mathbf{1}, \eta_t\rangle] < \infty$.

On the other hand, since each ν_t is the sum of at most N unit Dirac masses, it trivially follows that $\sup_{t\geq 0} \mathbb{E}[\langle \mathbf{1}, \nu_t\rangle] \leq N < \infty$. □

To show that $\{\eta_t\}_{t\geq 0}$ and $\{\nu_t\}_{t\geq 0}$ satisfy the second property in Proposition 4.3, note that by choosing $f = \mathbb{1}_{[c,H^r)}$, $c > 0$, in (2.10), we obtain for $t \geq 0$,

$$\begin{aligned}\mathbb{E}[\eta_t[c, H^r)] &\leq \mathbb{E}\bigg[\int_{[0,H^r)} \mathbb{1}_{[c,H^r)}(x+t) \frac{1 - G^r(x+t)}{1 - G^r(x)} \eta_0(dx)\bigg] \\ &\quad + \int_0^t \mathbb{1}_{[c,H^r)}(t-s)\big(1 - G^r(t-s)\big)\, de(s)\end{aligned} \tag{4.5}$$

and, likewise, by choosing $f = \mathbb{1}_{[c,H^s)}$ in (2.11) it follows that for $t \geq 0$,

$$\begin{aligned}\mathbb{E}[\nu_t[c, H^s)] &= \mathbb{E}\bigg[\int_{[0,H^s)} \mathbb{1}_{[c,H^s)}(x+t) \frac{1 - G^s(x+t)}{1 - G^s(x)} \nu_0(dx)\bigg] \\ &\quad + \mathbb{E}\bigg[\int_0^t \mathbb{1}_{[c,H^s)}(t-s)\big(1 - G^s(t-s)\big)\, dK(s)\bigg].\end{aligned} \tag{4.6}$$

We now establish two supporting lemmas.

LEMMA 4.5. *Suppose Assumption* 3.1 *holds and* $\mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] < \infty$. *We have*

$$\lim_{c\to H^r} \sup_{t\geq 0} \mathbb{E}\bigg[\int_{[0,H^r)} \mathbb{1}_{[c,H^r)}(x+t) \frac{1 - G^r(x+t)}{1 - G^r(x)} \eta_0(dx)\bigg] = 0 \tag{4.7}$$

and

$$\lim_{c\to H^s}\sup_{t\ge 0}\mathbb{E}\left[\int_{[0,H^s)}\mathbb{1}_{[c,H^s)}(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\nu_0(dx)\right]=0. \tag{4.8}$$

PROOF. When $H^r<\infty$, we have

$$\begin{aligned}
&\sup_{t\ge 0}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\eta_0(dx)\right]\\
&\quad\le \sup_{t\ge 0}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\eta_0(dx)\right]\\
&\quad= \sup_{t\in[0,c)}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\eta_0(dx)\right]\\
&\qquad\vee \sup_{t\in[c,H^r)}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\eta_0(dx)\right].
\end{aligned}$$

Using $\mathbb{E}[\langle \mathbf{1},\eta_0\rangle]<\infty$ to justify the application of the dominated convergence theorem, we obtain

$$\lim_{c\to H^r}\sup_{t\in[c,H^r)}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\eta_0(dx)\right]\le \lim_{c\to H^r}\mathbb{E}\big[\eta_0[0,H^r-c)\big]=0.$$

On the other hand, we know that

$$\begin{aligned}
&\sup_{t\in[0,c)}\mathbb{E}\left[\int_{[0,H^r)}\mathbb{1}_{[c,H^r)}(x+t)\eta_0(dx)\right]\\
&\qquad\le \sup_{t\in[0,c)}\mathbb{E}[\eta_0(c-t,H^r-t)].
\end{aligned}$$

We show by contradiction that $\sup_{t\in[0,c)}\mathbb{E}[\eta_0(c-t,H^r-t)]\to 0$ as $c\to H^r$. Suppose this is not true. Then there exist $\delta>0$ and sequences $\{c_n\}_{n\in\mathbb{N}}$ and $\{t_n\}_{n\in\mathbb{N}}$ such that $c_n\to H^r$ as $n\to\infty$, $t_n\in[0,c_n)$ for each $n\in\mathbb{N}$ and $\mathbb{E}[\eta_0(c_n-t_n,H^r-t_n)]>\delta$ for each $n\in\mathbb{N}$. Because we are considering the case $H^r<\infty$, $\{t_n\}_{n\in\mathbb{N}}$ is bounded and so we can take a subsequence, which we call again $\{t_n\}_{n\in\mathbb{N}}$, such that $\lim_{n\to\infty}t_n=t_*\in[0,H^r]$. In turn, this implies

$$\lim_{n\to\infty}\mathbb{E}[\eta_0(c_n-t_n,H^r-t_n)]=0,$$

which contradicts the initial hypothesis. Thus, $\sup_{t\in[0,c)}\mathbb{E}[\eta_0(c-t,H^r-t)]\to 0$. Together with the last three displays, this implies that (4.7) holds when $H^r<\infty$.

On the other hand, when $H^r = \infty$ we have

$$\begin{aligned}
&\sup_{t\geq 0} \mathbb{E}\biggl[\int_{[0,H^r)} \mathbb{1}_{[c,H^r)}(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\eta_0(dx)\biggr] \\
&\quad \leq \max\biggl(\sup_{t\in[0,c/2)} \mathbb{E}\biggl[\int_{[0,\infty)} \mathbb{1}_{[c,\infty)}(x+t)\eta_0(dx)\biggr], \\
&\qquad \sup_{t\in[c/2,\infty)} \mathbb{E}\biggl[\int_{[0,\infty)} \frac{1-G^r(x+t)}{1-G^r(x)}\eta_0(dx)\biggr]\biggr) \\
&\quad \leq \mathbb{E}[\eta_0(c/2,\infty)] \vee \mathbb{E}\biggl[\int_{[0,\infty)} \frac{1-G^r(x+c/2)}{1-G^r(x)}\eta_0(dx)\biggr].
\end{aligned}$$

Sending $c \to \infty$ on both sides, and using the fact that $\mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] < \infty$, an application of the dominated convergence theorem shows that the right-hand side vanishes and thus (4.7) holds in this case too. The proof of (4.8) is exactly analogous and is thus omitted. $\square$

LEMMA 4.6. *Suppose Assumption 3.1 holds and let $e(t) \doteq \mathbb{E}[E(t)], t \geq 0$. For $(H, G) = (H^r, G^r)$ and $(H, G) = (H^s, G^s)$, we have*

$$\lim_{c\to H} \sup_{t\geq 0} \int_0^t \mathbb{1}_{[c,H)}(t-s)\bigl(1-G(t-s)\bigr)\,de(s) = 0. \tag{4.9}$$

PROOF. E is a (delayed) renewal process with rate λ and due to Assumption 3.1 and Proposition 4.1 in Chapter V of [1], the function $x \mapsto \mathbb{1}_{[c,H)}(x)(1 - G(x))$ is directly Riemann integrable. Thus, by the key renewal theorem (cf. Theorem 4.7 of [1]) we obtain

$$\lim_{t\to\infty} \int_0^t \mathbb{1}_{[c,H)}(t-s)\bigl(1-G(t-s)\bigr)\,de(s) = \frac{1}{\lambda}\int_{[0,\infty)} \mathbb{1}_{[c,H)}(x)\bigl(1-G(x)\bigr)\,dx.$$

Since the integrability condition imposed in Assumption 3.1 implies that $\int_{[0,\infty)} \mathbb{1}_{[c,H)}(x)(1-G(x))\,dx \to 0$ as $c \to H$, we have the desired result. $\square$

LEMMA 4.7. *Suppose Assumption 3.1 holds and the initial condition satisfies $\mathbb{E}[\langle \mathbf{1}, \eta_0] < \infty$. Then the family $\{\eta_t\}_{t\geq 0}$ of $\mathcal{M}_F[0, H^r)$-valued random variables and the family $\{\nu_t\}_{t\geq 0}$ of $\mathcal{M}_F[0, H^s)$-valued random variables are tight.*

PROOF. Both families satisfy the first condition of Proposition 4.3 due to Lemma 4.4. Combining (4.5) with (4.7) and Lemma 4.6 for the case $(H, G) = (H^r, G^r)$, it follows that $\{\eta_t\}_{t\geq 0}$ also satisfies the second condition of Proposition 4.3 and is thus tight.

It only remains to show that $\{\nu_t\}_{t\geq 0}$ also satisfies the second condition of Proposition 4.3. For this, it suffices to show that as $c \to H^s$, the supremum (over t) of the right-hand side of (4.6) goes to zero. Now, let $k(t) \doteq \mathbb{E}[K(t)]$ for $t \geq 0$. Applying the integration-by-parts and change of variable formulae to the second term on the right-hand side of (4.6), we see that

$$
\begin{aligned}
\sup_{t\geq 0} \mathbb{E}\Big[\int_0^t & \mathbb{1}_{[c,H^s)}(t-s)\big(1-G^s(t-s)\big)\,dK(s)\Big] \\
&= \sup_{t>c} \int_0^t \mathbb{1}_{[c,H^s)}(t-s)\big(1-G^s(t-s)\big)\,dk(s) \\
&= \sup_{t>c}\Big(k(t-c)\big(1-G^s(c)\big) - k\big((t-H^s)^+\big)\big(1-G^s(t\wedge H^s)\big) \\
&\qquad\qquad - \int_c^{t\wedge H^s} k(t-s) g^s(s)\,ds\Big) \qquad (4.10)\\
&\leq \sup_{t>c}\Big(k(t-c)\big(1-G^s(t)\big) + \int_c^{t\wedge H^s} \big(k(t-c)-k(t-s)\big) g^s(s)\,ds\Big).
\end{aligned}
$$

Taking expectations of both sides of (2.6), we obtain for each $t \geq 0$,

$$
\mathbb{E}[Q(0)] + e(t) = \mathbb{E}[Q(t)] + \mathbb{E}[R(t)] + k(t).
$$

Since Q and R are nonnegative and R is increasing, it follows that

$$
k(t-c) \leq e(t-c) + \mathbb{E}[Q(0)]
$$

and

$$
k(t-c) - k(t-s) \leq e(t-c) - e(t-s) + \big(\mathbb{E}[Q(t-s)] - \mathbb{E}[Q(t-c)]\big).
$$

Substituting these inequalities into (4.10) and carrying out another integration-by-parts, we obtain

$$
\begin{aligned}
\sup_{t>c} \int_0^t & \mathbb{1}_{[c,H^s)}(t-s)\big(1-G^s(t-s)\big)\,dk(s) \\
&\leq \sup_{t>0} \int_0^t \mathbb{1}_{[c,H^s)}(t-s)\big(1-G^s(t-s)\big)\,de(s) \\
&\quad + \sup_{t>c} \mathbb{E}[Q(0)]\big(1-G^s(t)\big) \qquad (4.11)\\
&\quad + \sup_{t>c} \int_c^{t\wedge H^s} \big(\mathbb{E}[Q(t-s)] - \mathbb{E}[Q(t-c)]\big) g^s(s)\,ds.
\end{aligned}
$$

Applying Lemma 4.6, with $(H,G) = (H^s, G^s)$, we have

$$
\lim_{c\to H^s} \sup_{t\geq 0} \int_0^t \mathbb{1}_{[c,H^s)}(t-s)\big(1-G^s(t-s)\big)\,de(s) = 0.
$$

Moreover,

$$\lim_{c\to H^s}\sup_{t>c}\mathbb{E}[Q(0)]\big(1-G^s(t)\big)=\mathbb{E}[Q(0)]\lim_{c\to H^s}\big(1-G^s(c)\big)=0.$$

Also, since $Q(t)\le\langle\mathbf{1},\eta_t\rangle$ by (2.3), we have

$$\sup_{t>c}\int_c^{t\wedge H^s}\big(\mathbb{E}[Q(t-s)]-\mathbb{E}[Q(t-c)]\big)g^s(s)\,ds\le 2\sup_{t\ge 0}\mathbb{E}[\langle\mathbf{1},\eta_t\rangle]\big(1-G^s(c)\big).$$

Since Lemma 4.4 implies $\sup_{t\ge0}\mathbb{E}[\langle\mathbf{1},\eta_t\rangle]<\infty$, the right-hand side of the above inequality tends to zero as $c\to H^s$. Combining the last five assertions with (4.10) and (4.11), it follows that as $c\to H^s$, the supremum over $t\ge0$ of the second term on the right-hand side of (4.6) vanishes to zero. On the other hand, as $c\to H^s$, the supremum over $t\ge0$ of the first term on the right-hand side of (4.6) also vanishes to zero by (4.8). Thus, we have shown that $\sup_{t\ge0}\mathbb{E}[\nu_t[c,H^s)]\to0$ as $c\to H^s$, and the proof of the lemma is complete. □

LEMMA 4.8. *Suppose Assumption 3.1 holds and $\mathbb{E}[\langle\mathbf{1},\eta_0\rangle]<\infty$. The family of probability measures $\{L_t\}_{t\ge0}$ is tight.*

PROOF. By Lemma 4.7, we know that for each $\delta>0$, there exist two compact subsets $\tilde{C}_\delta\subset\mathcal{M}_F[0,H^s)$ and $\tilde{D}_\delta\subset\mathcal{M}_F[0,H^r)$ such that

$$\begin{aligned}&\inf_{t\ge0}\mathbb{P}(\nu_t\in\tilde{C}_\delta)\ge1-\delta/2,\\ &\inf_{t\ge0}\mathbb{P}(\eta_t\in\tilde{D}_\delta)\ge1-\delta/2.\end{aligned}\tag{4.12}$$

It follows from (2.3) and (2.4) that $X(t)\le\langle\mathbf{1},\nu_t\rangle+\langle\mathbf{1},\eta_t\rangle$ for each $t\ge0$. Together with (4.12) and the fact that the map $\mu\to\langle\mathbf{1},\mu\rangle$ is continuous, this implies that there exists $b>0$ such that

$$\inf_{t\ge0}\mathbb{P}\big(X(t)\le b\big)\ge1-\delta.\tag{4.13}$$

On the other hand, by Theorem 4.5 in Chapter V of [1], it follows that as $t\to\infty$, $\alpha_E(t)$ converges weakly to the distribution

$$F_0(t)\doteq\lambda\int_0^t\big(1-F(y)\big)\,dy.\tag{4.14}$$

Thus, there exist $T_0>0$ and $c>0$ such that for all $t\ge T_0$,

$$\mathbb{P}\big(\alpha_E(t)\le a\big)\ge F_0(a)-\delta/2\ge1-\delta.$$

By choosing a large enough, we may assume without loss of generality, that

$$\inf_{t\in[0,T_0]}\mathbb{P}\big(\alpha_E(t)\le a\big)\ge1-\delta.$$

Define $C_\delta \doteq [0,a] \times [0,b] \times \tilde{C}_\delta \times \tilde{D}_\delta$. Then the set C_δ is compact and $L_t(C_\delta) \geq 1-\delta$ for each $t \geq 0$, which proves the lemma. □

Since $\{Y_t, \mathcal{F}_t\}_{t\geq 0}$ is a Feller process by Proposition 4.2, and Lemma 4.8 is applicable when the initial condition satisfies $\mathbb{E}[\langle \mathbf{1}, \eta_0\rangle] < \infty$, the Krylov–Bogoliubov theorem immediately yields the following result.

THEOREM 4.9. *Suppose that Assumptions 3.1 and 3.5 hold. Then the state descriptor (α_E, X, ν, η) has a stationary distribution $(\alpha_{E,*}, X_*, \nu_*, \eta_*)$ that satisfies $\mathbb{E}[\langle \mathbf{1}, \eta_*\rangle] < \infty$.*

5. Fluid limit. In Section 5.1, we describe a deterministic dynamical system that was shown in Theorems 3.5 and 3.6 of [15] to arise as the so-called fluid limit of a many-server queue with abandonment that has service time and patience time distribution functions G^s and G^r, respectively. In Section 5.2, we identify the invariant manifold associated with the fluid limit, which is then used in Section 6 to obtain a first-order asymptotic approximation to the stationary distribution of the fluid scaled state descriptor $\overline{Y}^{(N)}$.

5.1. *Fluid equations.* The state of the fluid system at time t is represented by the triplet

$$(\overline{X}(t), \overline{\nu}_t, \overline{\eta}_t) \in \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r).$$

Here, $\overline{X}(t)$ represents the mass (or, equivalently, limiting scaled number of customers) in the system at time t, $\overline{\nu}_t[0,x)$ represents the mass of customers in service at time t who have been in service for less than x units of time, whereas $\overline{\eta}_t[0,x)$ represents the mass of customers in the system who, at time t, have been in the system no more than x units of time and whose patience time exceeds their time in system (which implies, in particular, that they have not yet abandoned the system). The inputs to the system are the (limiting) cumulative arrival process $\overline{E}$ and the initial conditions $\overline{X}(0)$, $\overline{\nu}_0$ and $\overline{\eta}_0$. Thus, $\langle \mathbf{1}, \overline{\nu}_0\rangle$ represents the total mass of customers in service at time 0 and the fluid analog of the nonidling condition (2.5) is

$$1 - \langle \mathbf{1}, \overline{\nu}_0\rangle = [1 - \overline{X}(0)]^+. \tag{5.1}$$

The quantity $\langle \mathbf{1}, \overline{\eta}_0\rangle$ represents the total mass of customers at time 0 whose residual patience times are positive. Hence, we have

$$[\overline{X}(0) - 1]^+ \leq \langle \mathbf{1}, \overline{\eta}_0\rangle.$$

Thus, the space of possible input data for the fluid equations is given by

$$\begin{aligned} \mathcal{S}_0 \doteq \{(e,x,\nu,\eta) \in \mathcal{I}_{\mathbb{R}_+}[0,\infty) \times \mathbb{R}_+ \times \mathcal{M}_F[0,H^s) \times \mathcal{M}_F[0,H^r) :& \\ 1 - \langle \mathbf{1}, \nu\rangle = [1-x]^+, [x-1]^+ \leq \langle \mathbf{1}, \eta\rangle\},& \end{aligned} \tag{5.2}$$

where recall that $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ is the subset of nondecreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0)=0$. Let $F^{\overline{\eta}_t}(x)$ denote $\overline{\eta}_t[0,x]$ for each $x \in [0, H^r)$.

DEFINITION 5.1 (Fluid equations). Given any $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, we say that the càdlàg function $(\overline{X}, \overline{\nu}, \overline{\eta})$ taking values in $\mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ satisfies the associated fluid equations if for every $t \in [0,\infty)$,

$$\int_0^t \langle h^r, \overline{\eta}_s\rangle\, ds < \infty, \qquad \int_0^t \langle h^s, \overline{\nu}_s\rangle\, ds < \infty \tag{5.3}$$

for every bounded Borel measurable function f defined on $\mathbb{R}_+$,

$$\begin{aligned}\int_{[0,H^s)} f(x)\overline{\nu}_t(dx) = {}& \int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\overline{\nu}_0(dx)\\ &+ \int_0^t f(t-s)\bigl(1-G^s(t-s)\bigr)\, d\overline{K}(s)\end{aligned} \tag{5.4}$$

and

$$\begin{aligned}\int_{[0,H^r)} f(x)\overline{\eta}_t(dx) = {}& \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\overline{\eta}_0(dx)\\ &+ \int_0^t f(t-s)\bigl(1-G^r(t-s)\bigr)\, d\overline{E}(s),\end{aligned} \tag{5.5}$$

where

$$\overline{K}(t) = [\overline{X}(0)-1]^+ - [\overline{X}(t)-1]^+ + \overline{E}(t) - \overline{R}(t); \tag{5.6}$$

$$\overline{X}(t) = \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h^s, \overline{\nu}_s\rangle\, ds - \overline{R}(t); \tag{5.7}$$

$$\overline{R}(t) = \int_0^t \left(\int_0^{[\overline{X}(s)-1]^+} h^r((F^{\overline{\eta}_s})^{-1}(y))\, dy\right) ds; \tag{5.8}$$

$$1 - \langle \mathbf{1}, \overline{\nu}_t\rangle = [1-\overline{X}(t)]^+; \tag{5.9}$$

$$[\overline{X}(t)-1]^+ \le \langle \mathbf{1}, \overline{\eta}_t\rangle. \tag{5.10}$$

Note that these fluid equations are not of the same form as those given in Definition 3.3 of [15] because the analogs of (5.4) and (5.5) are presented in dynamical form in [15] and are only required to be satisfied for continuous functions with compact support (in particular, see equations (3.9) and (3.11) of [15]). However, these two pairs of equations are equivalent due to Theorem 4.1 of [17] or, equivalently, Proposition 4.1 of [15], and can be shown to hold for the larger class bounded measurable functions using standard monotone convergence arguments. Theorems 3.5 and 3.6 of [15] show that under some mild assumptions on the input data $\overline{E}$, $\overline{\nu}_0$ and $\overline{\eta}_0$ and the hazard rate functions h^r and h^s (which are stated as Assumptions 3.3 and 3.4 here), there exists a unique solution to the fluid equations.

For future purposes, note that if $(\overline{X}, \overline{\nu}, \overline{\eta})$ satisfy the fluid equations for some $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, then $\overline{K}$ also satisfies

$$\overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_0^t \langle h^s, \overline{\nu}_s \rangle \, ds. \tag{5.11}$$

Indeed, this is simply the mass balance equation for the fluid in service and can be derived from (5.6), (5.7) and (5.9). Moreover, combining (5.11) and (5.4), with $f = \mathbf{1}$, and using an integration-by-parts argument (see Corollary 4.2 of [15]), it is easy to see that $\overline{K}$ satisfies the renewal equation

$$\begin{aligned} \overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_{[0,H^s)} \frac{G^s(x+t) - G^s(x)}{1 - G^s(x)} \overline{\nu}_0(dx) \\ + \int_0^t g^s(t-s) \overline{K}(s) \, ds. \end{aligned} \tag{5.12}$$

Since the first two terms on the right-hand side are bounded, by the key renewal theorem (see, e.g., Theorem 4.3 in Chapter V of [1]), $\overline{K}$ admits the representation

$$\begin{aligned} \overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_{[0,H^s)} \frac{G^s(x+t) - G^s(x)}{1 - G^s(x)} \overline{\nu}_0(dx) \\ + \int_0^t \biggl(\langle \mathbf{1}, \overline{\nu}_{t-s} \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle \\ + \int_{[0,H^s)} \frac{G^s(x+t-s) - G^s(x)}{1 - G^s(x)} \overline{\nu}_0(dx) \biggr) u^s(s) \, ds, \end{aligned} \tag{5.13}$$

where u^s is the density of the renewal function U^s associated with G^s (u^s exists because G^s is assumed to have a density). Also, it will prove convenient to introduce the fluid queue length process $\overline{Q}$ defined by

$$\overline{Q}(t) \doteq [\overline{X}(t) - 1]^+, \qquad t \in [0, \infty). \tag{5.14}$$

For every $t \in [0, \infty)$, the inequality in (5.10) implies

$$\overline{Q}(t) \le \langle \mathbf{1}, \overline{\eta}_t \rangle, \tag{5.15}$$

and (5.6) and (5.14), when combined, show that

$$\overline{Q}(0) + \overline{E}(t) = \overline{Q}(t) + \overline{K}(t) + \overline{R}(t). \tag{5.16}$$

The fluid equations without abandonment can be defined in a similar fashion. Let

$$\begin{aligned} \tilde{\mathcal{S}}_0 \doteq \{(e, x, \nu) \in \mathcal{I}_{\mathbb{R}_+}[0, \infty) \times \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) : \\ 1 - \langle \mathbf{1}, \nu \rangle = [1 - x]^+\}. \end{aligned} \tag{5.17}$$

DEFINITION 5.2. Given any $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \tilde{\mathcal{S}}_0$, we say $(\overline{X}, \overline{\nu}) \in \mathbb{R}_+ \times \mathcal{M}_F[0, H^s)$ is a solution to the associated fluid equations in the absence of abandonment if for every $t \in [0, \infty)$, the second inequality in (5.3) holds, and equations (5.4), (5.6), (5.7) and (5.9) hold with $\overline{R} \equiv 0$.

REMARK 5.3. The case when customers do not renege corresponds to the case when the patience time distribution G^r has unit mass at ∞. Formally setting $dG^r = \delta_\infty$ in Definition 5.1, we obtain the fluid limit equations in the absence of abandonment specified in Definition 5.2 (also refer to Definition 3.3 in [17]). In fact, in this case $G^r(x) = 0$ and hence, $h^r(x) = 0$ for all $x \in [0, \infty)$. From this and (5.8) we see that $\overline{R}(t) = 0$ for all $t \geq 0$. Also, note that (5.3), (5.4), (5.7), (5.9) and (5.11) are equivalent to (3.4)–(3.8) of Definition 3.3 in [17]. At last, by letting $f = \mathbf{1}$ in (5.5), since G^r is zero on $[0, \infty)$, we have $\langle \mathbf{1}, \overline{\eta}_t \rangle = \langle \mathbf{1}, \overline{\eta}_0 \rangle + \overline{E}(t)$. On the other hand, by (5.7) and (5.2), we have

$$[\overline{X}(t) - 1]^+ \leq \big[[\overline{X}(0) - 1]^+ + \overline{E}(t)\big]^+ \leq [\langle \mathbf{1}, \overline{\eta}_0 \rangle + \overline{E}(t)]^+ = \langle \mathbf{1}, \overline{\eta}_t \rangle.$$

This shows that (5.10) holds automatically when there is no abandonment.

5.2. *Invariant manifold.* We now introduce a set of states associated with the fluid equations described in Definition 5.1, which we call the *invariant manifold*. As shown in Section 6, when the invariant manifold consists of a single point, it is the limit of the scaled sequence of convergent stationary distributions $(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)}) = \frac{1}{N}(X_*^{(N)}, \nu_*^{(N)}, \eta_*^{(N)})$.

DEFINITION 5.4 (Invariant manifold). Given $\lambda \in (0, \infty)$, a state $(x_0, \nu_0, \eta_0) \in \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ such that $(\lambda \mathbf{1}, x_0, \nu_0, \eta_0) \in \mathcal{S}_0$ is said to be invariant for the fluid equations described in Definition 5.1 with arrival rate λ if the solution $(\overline{X}, \overline{\nu}, \overline{\eta})$ to the fluid equations associated with the input data $(\lambda \mathbf{1}, x_0, \nu_0, \eta_0)$ satisfies $(\overline{X}(t), \overline{\nu}_t, \overline{\eta}_t) = (x_0, \nu_0, \eta_0)$ for all $t \geq 0$. The set of all invariant states for the fluid equation with rate λ will be referred to as the *invariant manifold* (associated with the fluid equations with rate λ).

THEOREM 5.5 (Characterization of the invariant manifold). *Given $\lambda \in (0, \infty)$, the set $\mathcal{I}_\lambda$ defined in* (3.6) *is the invariant manifold associated with the fluid equations with arrival rate λ.*

Theorem 5.5 is a consequence of the next two lemmas. Let $\lambda \in (0, \infty)$ and (x_0, ν_0, η_0) be an invariant state according to Definition 5.4. Then the unique solution $(\overline{X}, \overline{\nu}, \overline{\eta})$ to the fluid equations associated with the input data $(\lambda \mathbf{1}, x_0, \nu_0, \eta_0) \in \mathcal{S}_0$ satisfies $(\overline{X}(t), \overline{\nu}_t, \overline{\eta}_t) = (x_0, \nu_0, \eta_0)$ for all $t \geq 0$. Let $\overline{Q}, \overline{R}, \overline{K}$ be the associated auxiliary processes satisfying (5.14), (5.8), (5.6), and recall the definition of the measures ν_* and η_* given in (3.3) and (3.4), respectively.

LEMMA 5.6. *If* (x_0, ν_0, η_0) *is an invariant state, then* $\eta_0(dx) = \lambda(1 - G^r(x))\,dx = \lambda\eta_*(dx)$.

PROOF. On substituting the relation $\eta_t = \eta_0$, $t \ge 0$, into (5.5), we see that for every $f \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \in [0, \infty)$,

$$\begin{aligned}\int_{[0,H^r)} f(x)\eta_0(dx) \\ = \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\eta_0(dx) \\ + \lambda\int_0^t f(s)\bigl(1-G^r(s)\bigr)\,ds.\end{aligned} \tag{5.18}$$

Sending $t \to \infty$ and applying the dominated convergence theorem, the first term vanishes and we obtain

$$\int_{[0,H^r)} f(x)\eta_0(dx) = \lambda\int_0^\infty f(s)\bigl(1-G^r(s)\bigr)\,ds = \int_{[0,H^r)} f(s)\lambda\bigl(1-G^r(s)\bigr)\,ds.$$

It then follows that $\eta_0(dx) = \lambda\eta_*(dx)$. $\square$

LEMMA 5.7. *If* (x_0, ν_0, η_0) *is an invariant state, then* $\nu_0(dx) = (\lambda \wedge 1)\nu_*(dx)$, $x_0 = \lambda$ *if* $\lambda < 1$ *and* $x_0 \in B_\lambda$ *if* $\lambda \ge 1$. *Moreover, if either* $x_0 = \lambda < 1$, *or* $\lambda > 1$ *and* $x_0 \in B_\lambda$, *then* $(x_0, (\lambda \wedge 1)\nu_*, \lambda\eta_*)$ *is an invariant state.*

PROOF. Suppose (x_0, ν_0, η_0) is an invariant state. Since $\overline{X}(t) = x_0$, we have $\overline{Q}(t) = \overline{Q}(0)$ by (5.14). Since, in addition, $\overline{\eta}_t = \eta_0 = \lambda\eta_*$ by Lemma 5.6, we have

$$\int_0^{[\overline{X}(t)-1]^+} h^r((F^{\overline{\eta}_t})^{-1}(y))\,dy = \int_0^{[x_0-1]^+} h^r((F^{\lambda\eta_*})^{-1}(y))\,dy.$$

Let p denote the term on the right-hand side of the above display. Then for each $t \ge 0$, by (5.8) we have $\overline{R}(t) = pt$ and by (5.16) we have $\overline{K}(t) = (\lambda - p)t$. Substituting $\overline{\nu}_t = \nu_0$ in (5.4), we obtain for every $f \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \in [0, \infty)$,

$$\begin{aligned}\int_{[0,H^s)} f(x)\nu_0(dx) \\ = \int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\nu_0(dx) \\ + \int_0^t f(s)\bigl(1-G^s(s)\bigr)(\lambda-p)\,ds.\end{aligned} \tag{5.19}$$

Sending $t \to \infty$ and applying the dominated convergence theorem, we obtain

$$\begin{aligned}\int_{[0,H^s)} f(x)\nu_0(dx) &= (\lambda-p)\int_0^\infty f(s)\bigl(1-G^s(s)\bigr)\,ds \\ &= (\lambda-p)\int_{[0,H^r)} f(s)\bigl(1-G^s(s)\bigr)\,ds.\end{aligned}$$

Thus, $\nu_0(dx) = (\lambda - p)\nu_*(dx)$ and hence, $\langle \mathbf{1}, \nu_0\rangle = \lambda - p$.

To show that $\nu_0(dx) = (\lambda \wedge 1)\nu_*(dx)$, it suffices to show that $\lambda - p = \langle \mathbf{1}, \nu_0\rangle = \lambda \wedge 1$. If $x_0 \le 1$, then $p = 0$ by its definition. Hence, $\nu_0(dx) = \lambda\nu_*(dx)$ and $\lambda = \langle \mathbf{1}, \nu_0\rangle \le 1$. Thus, in this case, $\lambda - p = \lambda \wedge 1$. On the other hand, if $x_0 > 1$, it follows from (5.9) that $\langle \mathbf{1}, \nu_0\rangle = 1$. Since we also have $\langle \mathbf{1}, \nu_0\rangle = \lambda - p$, it follows that $\lambda = p + 1 \ge 1$. Thus, in this case too, we have $\lambda - p = \lambda \wedge 1$. This proves the first assertion of the lemma.

For the second assertion of the lemma, we observe that when $\lambda < 1$, the equality $\lambda - p = \lambda \wedge 1$ implies $p = 0$ and $\langle \mathbf{1}, \nu_0\rangle = \lambda < 1$. Hence, (5.1) implies $x_0 = \langle \mathbf{1}, \nu_0\rangle = \lambda$. If $\lambda \ge 1$, we have $\nu_0(dx) = \nu_*(dx)$ and the equality $\lambda - p = \lambda \wedge 1$ implies $p = \lambda - 1$. Then $x_0 \ge \langle \mathbf{1}, \nu_0\rangle = 1$ and

$$\lambda G^r\big((F^{\lambda\eta_*})^{-1}\big((x_0 - 1)^+\big)\big) = \int_0^{(x_0-1)^+} h^r\big((F^{\lambda\eta_*})^{-1}(y)\big)\,dy = p = \lambda - 1.$$

Hence, x_0 belongs to the set B_λ defined in (3.5). The last assertion can be verified directly by substituting the initial condition into the fluid equations. This completes the proof of the lemma. □

6. The limit of scaled stationary distributions. This section is devoted to the proof of Theorem 3.3. Suppose Assumptions 3.1 and 3.5 hold and let $\overline{Y}_*^{(N)} = (\overline{\alpha}_{E,*}^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$, $N \in \mathbb{N}$, be a sequence of scaled stationary distributions for the N-server queue, which exists by Theorem 4.9. When Assumption 3.2 also holds, let $(x_*, (\lambda \wedge 1)\nu_*, \lambda\eta_*)$ be the unique element of the invariant manifold $\mathcal{I}_\lambda$. The main result of this section is to show that, as $N \to \infty$,

$$\big(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)}\big) \Rightarrow \big(x_*, (\lambda \wedge 1)\nu_*, \lambda\eta_*\big). \tag{6.1}$$

We first show in Section 6.1 that the sequence $\{(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)}), N \in \mathbb{N}\}$ is tight. Then, in Section 6.2, we show that (without imposing Assumption 3.2) the weak limit of every convergent subsequence must almost surely be an invariant state. When there is a unique invariant state, this proves (6.1). Note that the method of proof does not explicitly require that the stationary distribution for each N-server queue be unique. For each $N \in \mathbb{N}$, recall the definition given in (3.8) of the fluid-scaled state process

$$\overline{Y}^{(N)} = \big(\overline{\alpha}_E^{(N)}, \overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)}\big) \tag{6.2}$$

for the N-server queue with abandonment associated with the initial condition $\overline{Y}^{(N)}(0) = (\overline{\alpha}_{E,*}^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$. Let $\overline{Q}^{(N)}, \overline{R}^{(N)}, \overline{K}^{(N)}$ be the fluid-scaled auxiliary processes associated with $\overline{Y}^{(N)}$ that were introduced in Section 2.2.

6.1. *Tightness*. To establish tightness of the sequence $\{\overline{Y}_*^{(N)}\}_{N\in\mathbb{N}}$, we will make use of the criteria for tightness of measure-valued random variables given in Proposition 4.3.

LEMMA 6.1. *Let $c \in [0, H^r)$. Then, for each integer $n \geq 2$, $\overline{\eta}_*^{(N)}$ and $\overline{\nu}_*^{(N)}$ satisfy the following relations*:

$$\begin{aligned}
&\mathbb{E}[\overline{\eta}_*^{(N)}[c, H^r)] \\
&\quad = \mathbb{E}\biggl[\int_{[0,H^r)} \frac{1 - G^r(x + nc)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\biggr] \\
&\qquad + \mathbb{E}\biggl[\int_{[0,c]} \sum_{j=2}^{n} \bigl(1 - G^r(jc - s)\bigr)\, d\overline{E}^{(N)}(s)\biggr],
\end{aligned} \tag{6.3}$$

$$\begin{aligned}
&\mathbb{E}[\overline{\nu}_*^{(N)}[c, H^s)] \\
&\quad = \mathbb{E}\biggl[\int_{[0,H^s)} \frac{1 - G^s(x + nc)}{1 - G^s(x)} \overline{\nu}_*^{(N)}(dx)\biggr] \\
&\qquad + \mathbb{E}\biggl[\int_{[0,c]} \sum_{j=2}^{n} \bigl(1 - G^s(jc - s)\bigr)\, d\overline{K}^{(N)}(s)\biggr].
\end{aligned} \tag{6.4}$$

PROOF. We only prove (6.3) because (6.4) can be proved in the same way. Fix $c \in [0, H^r)$. Dividing both sides of (2.10) by N and setting $\overline{\eta}_0^{(N)} = \overline{\eta}_*^{(N)}$, we obtain for each bounded measurable function f on $\mathbb{R}_+$ and $t > 0$,

$$\begin{aligned}
\mathbb{E}[\langle f, \overline{\eta}_t^{(N)}\rangle] &= \mathbb{E}\biggl[\int_{[0,H^r)} f(x+t) \frac{1 - G^r(x + t)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\biggr] \\
&\quad + \mathbb{E}\biggl[\int_{[0,t]} f(t-s)\bigl(1 - G^r(t - s)\bigr)\, d\overline{E}^{(N)}(s)\biggr].
\end{aligned} \tag{6.5}$$

Since the initial conditions are stationary, $\overline{\eta}_t^{(N)}$ has the same distribution as $\overline{\eta}_*^{(N)}$ for every $t \geq 0$. Therefore, by substituting $f = \mathbb{1}_{[c,H^r)}$ and $t = c$ in (6.5), and noting that $F^{(N)}(0) = 0$, $\mathbb{1}_{[c,H^r)}(x + c) = 1$ for every $x \geq 0$ and $\mathbb{1}_{[c,H^r)}(c - s) = 0$ for every $s \in (0, c]$, we obtain

$$\begin{aligned}
\mathbb{E}[\overline{\eta}_*^{(N)}[c, H^r)] &= \mathbb{E}[\overline{\eta}_c^{(N)}[c, H^r)] \\
&= \mathbb{E}\biggl[\int_{[0,H^r)} \frac{1 - G^r(x + c)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\biggr] \\
&= \mathbb{E}\biggl[\int_{[0,H^r)} \frac{1 - G^r(x + c)}{1 - G^r(x)} \overline{\eta}_c^{(N)}(dx)\biggr].
\end{aligned}$$

Next, choosing $f = (1 - G^r(\cdot + c))/(1 - G^r(\cdot))$ and $t = c$ in (6.5), we obtain

$$\mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + c)}{1 - G^r(x)} \overline{\eta}_c^{(N)}(dx)\right] = \mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + 2c)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\right] + \mathbb{E}\left[\int_{[0,c]} \big(1 - G^r(2c - s)\big)\, d\overline{E}^{(N)}(s)\right].$$

Combining the last two displays, we see that

$$\mathbb{E}[\overline{\eta}_*^{(N)}[c, H^r)] = \mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + 2c)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\right] + \mathbb{E}\left[\int_{[0,c]} \big(1 - G^r(2c - s)\big)\, d\overline{E}^{(N)}(s)\right].$$

Thus, we have shown that (6.3) holds for $n = 2$. Suppose that for some integer $m \geq 2$, (6.3) holds for $n = m$, that is,

$$\mathbb{E}[\overline{\eta}_*^{(N)}[c, H^r)] = \mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + mc)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\right] + \mathbb{E}\left[\int_{[0,c]} \sum_{j=2}^{m} \big(1 - G^r(jc - s)\big)\, d\overline{E}^{(N)}(s)\right]. \tag{6.6}$$

Substituting $f = (1 - G^r(\cdot + mc))/(1 - G^r(\cdot))$ and $t = c$ in (6.5) and using the fact that $\overline{\eta}_c^{(N)}$ has the same distribution as $\overline{\eta}_*^{(N)}$, we obtain

$$\begin{aligned} &\mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + mc)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\right] \\ &\quad = \mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + mc)}{1 - G^r(x)} \overline{\eta}_c^{(N)}(dx)\right] \\ &\quad = \mathbb{E}\left[\int_{[0,H^r)} \frac{1 - G^r(x + (m+1)c)}{1 - G^r(x)} \overline{\eta}_*^{(N)}(dx)\right] \\ &\qquad + \mathbb{E}\left[\int_{[0,c]} \big(1 - G^r\big((m+1)c - s\big)\big)\, d\overline{E}^{(N)}(s)\right]. \end{aligned} \tag{6.7}$$

This, together with (6.6), yields (6.3) with $n = m + 1$. This completes the induction argument and we have the desired result. $\square$

THEOREM 6.2. *If Assumptions 3.1 and 3.5 are satisfied and $\overline{\lambda}^{(N)} \to \lambda \in (0, \infty)$, then the sequence $\{(\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})\}_{N \in \mathbb{N}}$ is tight. Moreover,*

$$\sup_{N \in \mathbb{N}} \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_*^{(N)} \rangle] < \infty. \tag{6.8}$$

PROOF. We first show that $\{\overline{\eta}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight. Note that $\langle \mathbf{1}, \overline{\eta}_*^{(N)}\rangle$ can be viewed as the fluid scaled queue length process associated with an infinite-server queue with arrival process $\overline{E}^{(N)}$ and service distribution function G^r. By Little's law (cf. Theorem 2 of [19]), we know that $\mathbb{E}[\langle \mathbf{1}, \overline{\eta}_*^{(N)}\rangle] = \overline{\lambda}^{(N)}\theta^r$, where θ^r, the mean of G^r, is finite by Assumption 3.1. Due to the assumed convergence of $\overline{\lambda}^{(N)}$ to λ, this implies (6.8).

Next, note that for each n, the function $(1-G^r(\cdot+nc))/(1-G^r(\cdot))$ is bounded by 1 and converges to 0 as $n\to\infty$. Therefore, it follows from the dominated convergence theorem that

$$\lim_{n\to\infty}\mathbb{E}\left[\int_{[0,H^r)}\frac{1-G^r(x+nc)}{1-G^r(x)}\overline{\eta}_*^{(N)}(dx)\right]=0. \tag{6.9}$$

Sending $n\to\infty$ on the right-hand side of (6.3), and using (6.9) and the monotone convergence theorem, we have

$$\mathbb{E}[\overline{\eta}_*^{(N)}[c,H^r)] = \mathbb{E}\left[\int_{[0,c]}\sum_{j=2}^{\infty}\bigl(1-G^r(jc-s)\bigr)\,d\overline{E}^{(N)}(s)\right]. \tag{6.10}$$

On the other hand, we also have the simple estimate

$$\begin{aligned}\mathbb{E}\left[\int_{[0,c]}\bigl(1-G^r(2c-s)\bigr)\,d\overline{E}^{(N)}(s)\right] &\le (1-G^r(c))\mathbb{E}[\overline{E}^{(N)}(c)]\\ &= c\bigl(1-G^r(c)\bigr)\frac{\mathbb{E}[\overline{E}^{(N)}(c)]}{c}.\end{aligned} \tag{6.11}$$

Carrying out integration-by-parts on $\int_0^\infty(1-G^r(x))\,dx$, it follows that

$$\int_{[0,H^r)}\bigl(1-G^r(x)\bigr)\,dx = \lim_{x\to H^r}x\bigl(1-G^r(x)\bigr)+\int_{[0,H^r)}xg^r(x)\,dx.$$

However, since the mean θ^r is finite by (3.1), it follows that $c(1-G^r(c))\to 0$ as $c\to H^r$. In addition, because the elementary renewal theorem implies that $\mathbb{E}[\overline{E}^{(N)}(c)]/c\to\overline{\lambda}^{(N)}$ as $c\to\infty$ and $\overline{\lambda}^{(N)}\to\lambda$ as $N\to\infty$, it follows that

$$\limsup_{c\to H^r}\sup_N\frac{\mathbb{E}[\overline{E}^{(N)}(c)]}{c}<\infty. \tag{6.12}$$

Thus, taking first the supremum over N and then the limit as $c\to H^r$ in (6.11), we obtain

$$\lim_{c\to H^r}\sup_N\mathbb{E}\left[\int_{[0,c]}\bigl(1-G^r(2c-s)\bigr)\,d\overline{E}^{(N)}(s)\right]=0. \tag{6.13}$$

Since $1-G^r(\cdot)$ is a decreasing function, for $s\in[0,c]$,

$$\sum_{j=3}^{\infty}c\bigl(1-G^r(jc-s)\bigr)\le\int_{[2c-s,H^r)}\bigl(1-G^r(x)\bigr)\,dx\le\int_{[c,H^r)}\bigl(1-G^r(x)\bigr)\,dx.$$

Therefore, we have

$$\sup_N \mathbb{E}\left[\int_{[0,c]} \sum_{j=3}^{\infty} \big(1 - G^r(jc - s)\big)\, d\overline{E}^{(N)}(s)\right]$$
$$\leq \sup_N \frac{\mathbb{E}[\overline{E}^{(N)}(c)]}{c} \int_{[c,H^r)} \big(1 - G^r(x)\big)\, dx,$$

which tends to zero as $c \to H^r$ because of (6.12) and Assumption 3.1. Combining the last assertion with (6.10) and (6.13), we see that

$$\lim_{c\to H^r} \sup_N \mathbb{E}\big[\overline{\eta}_*^{(N)}[c, H^r)\big] = 0, \tag{6.14}$$

which establishes the second criterion for tightness. Thus, the sequence $\{\overline{\eta}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight.

We next show that $\{\overline{\nu}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight. The analog of (6.8) holds for $\{\overline{\nu}_*^{(N)}\}_{N\in\mathbb{N}}$ automatically because $\langle \mathbf{1}, \overline{\nu}_*^{(N)}\rangle \leq 1$ for each N. On the other hand, the analog of (6.14) can be shown to hold for $\{\overline{\nu}_*^{(N)}\}_{N\in\mathbb{N}}$ by using (6.4) and an argument similar to that used above to establish (6.14), along with the additional observation that $\mathbb{E}[\overline{K}^{(N)}(c)] \leq \mathbb{E}[\overline{E}^{(N)}(c)] + \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_*^{(N)}\rangle]$ implies $\limsup_{c\to H^s} \sup_N \mathbb{E}[\overline{K}^{(N)}(c)]/c < \infty$. Thus, the sequence $\{\overline{\nu}_*^{(N)}\}_{N\in\mathbb{N}}$ is also tight.

Finally, we show that the sequence of $\mathbb{R}_+$-valued random variables $\{\overline{X}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight. Since $\overline{X}_*^{(N)} \leq 1 + \langle \mathbf{1}, \overline{\eta}_*^{(N)}\rangle$ for each N, $\sup_N \mathbb{E}[\overline{X}_*^{(N)}] \leq 1 + \sup_N \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_*^{(N)}\rangle]$, which is finite due to (6.8). The tightness of $\{\overline{X}_*^{(N)}\}_{N\in\mathbb{N}}$ is a direct consequence of Markov's inequality. □

6.2. *The limit of the stationary distributions*. We now present the proof of our main result.

PROOF OF THEOREM 3.3. For each $N \in \mathbb{N}$, let $\overline{Y}_*^{(N)} = (\overline{\alpha}_{E,*}^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$ be a fluid scaled stationary distribution for the N-server system. We will invoke the fluid limit theorem established in Theorem 3.6 of [15] to establish the result. For each $N \in \mathbb{N}$, let $\overline{Z}^{(N)} = (\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ be the (fluid scaled) state process for the N-server queue with initial data $(\overline{E}_*^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$. Since the hazard rate functions h^s and h^r satisfy Assumption 3.4 (which corresponds to Assumption 3.3 of [15]), it follows from Remark 3.2 and Theorem 3.6 of [15] that if (a) the sequence (or subsequence) of initial data $(\overline{E}_*^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$ converges in distribution to some random element $(\overline{E}_*, \tilde{X}_*, \tilde{\nu}_*, \tilde{\eta}_*)$ in the sense of Assumption 3.1 of [15], (b) $\overline{E}_*$ is continuous and (c) $\tilde{\eta}_*$ is a continuous distribution, then the sequence (subsequence) $\overline{Z}^{(N)}$ converges to a process $\overline{Z} = (\overline{X}, \overline{\nu}, \overline{\eta})$ that is the unique solution to the fluid equations with initial data $(\overline{E}_*, \tilde{X}_*, \tilde{\nu}_*, \tilde{\eta}_*)$.

However, by stationarity for each $N \in \mathbb{N}$ and $t > 0$, $\overline{Z}^{(N)}(t)$ has the same distribution as $\overline{Z}^{(N)}(0)$. This implies that $\overline{Z}$ is the constant process that is identically equal to the initial data $(\tilde{X}_*, \tilde{\nu}_*, \tilde{\eta}_*)$, which in turn implies that $(\tilde{X}_*, \tilde{\nu}_*, \tilde{\eta}_*)$ is an invariant state for the fluid limit.

Thus, to establish the theorem, it only remains to verify properties (a)–(c) above. Since Assumptions 3.1, 3.3(1) and 3.5 hold, by the tightness result established in Theorem 6.2, it follows that the sequence of stationary "initial conditions" $\{\overline{Y}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight. On the other hand, by basic properties of renewal processes and the assumption that $\overline{\lambda}^{(N)} \to \lambda$, the sequence of scaled stationary arrival processes $\{\overline{E}_*^{(N)}\}_{N\in\mathbb{N}}$ satisfies $\overline{E}_*^{(N)} \Rightarrow \overline{E}_*$ as $N \to \infty$, where $\overline{E}_*(t) = \lambda t$ for $t \in [0, \infty)$. Therefore, there exists a convergent subsequence, which by some abuse of notation we denote again by $\{\overline{Y}_*^{(N)}\}_{N\in\mathbb{N}}$, that converges weakly to some limit $\overline{Y}_*$ of the form $\overline{Y}_* = (\lambda\mathbf{1}, \tilde{X}_*, \tilde{\nu}_*, \tilde{\eta}_*)$. This immediately shows that properties (a) and (b) above are satisfied. It only remains to show that $\tilde{\eta}_*$ has a continuous distribution. Now, by the proof of Theorem 7.1 of [15] (note that Assumption 3.2 of [15] is not used for this part of the proof) it follows that the inequality (3.39) of [15] holds and that $\overline{\eta}$ satisfies the dynamical equation (3.42) of [15], with $\overline{\eta}_0 = \tilde{\eta}_*$ and $\overline{E}$ replaced by $\overline{E}_*$. By Theorem 4.1 of [17] (equivalently, Proposition 4.1 of [15]), it follows that $\overline{\eta}$ satisfies the fluid equation (5.5) with $\overline{\eta}_0 = \tilde{\eta}_*$. In particular, also using the fact that $\overline{E}_*(t) = \lambda t$ and $\overline{\eta}_t$ has the same distribution as $\tilde{\eta}_*$, this implies that for every bounded measurable f on $[0, H_r)$ and any $t > 0$,

$$
\begin{aligned}
\langle f, \tilde{\eta}_*\rangle \overset{(d)}{=} & \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\tilde{\eta}_*(dx) \\
& + \int_0^t f(t-s)\big(1-G^r(t-s)\big)\lambda\, ds.
\end{aligned} \tag{6.15}
$$

Now, sending $t \to \infty$ on the right-hand side, using the dominated convergence theorem (which is justified by the bound $\langle 1, \tilde{\eta}_*\rangle < \infty$ a.s. established in Theorem 6.2), we see that the right-hand side equals $\lambda\langle f, \eta_*\rangle$. This shows that $\tilde{\eta}_*$ has the same distribution as $\lambda\eta_*$, which in particular proves that $\tilde{\eta}_*$ is a continuous distribution. This completes the proof of property (c). Thus, we have shown that any convergence subsequence of the stationary distribution converges to an invariant state of the fluid limit. When the manifold consists of a single element, the usual argument by contradiction then shows that the original sequence of stationary distributions converges to this point. □

7. Concluding remarks. We can establish ergodicity of the state processes under an additional condition. Let

$$\varrho^r \doteq \sup\{u \in [0, H^r) : g^r = 0 \text{ a.e. on } [a, a+u] \text{ for some } a \in [0, \infty)\}$$

and

$$\varrho^s \doteq \sup\{u \in [0, H^s) : g^s = 0 \text{ a.e. on } [a, a+u] \text{ for some } a \in [0, \infty)\}.$$

$$\begin{tikzcd}
(\overline{X}^{(N)}(t), \overline{\nu}_t^{(N)}, \overline{\eta}_t^{(N)}) \arrow[r] \arrow[d] & (\overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)}) \arrow[d] \\
(\overline{X}(t), \overline{\nu}_t, \overline{\eta}_t) \arrow[r, "?"] & (\overline{X}_*, \overline{\nu}_*, \overline{\eta}_*)
\end{tikzcd}$$

FIG. 1. *Interchange of limits diagram.*

ASSUMPTION 7.1. The following three conditions hold:

(1) $H^r = H^s = \infty$;
(2) $\varrho \doteq \varrho^r \vee \varrho^s < \infty$;
(3) For every interval $[a, b] \subset [0, \infty)$ with $b - a > 0$, $F^{(N)}(b) - F^{(N)}(a) > 0$.

THEOREM 7.1. *Suppose Assumptions 3.1–3.5 and 7.1 hold. Then the Markov process $\{Y_t, \mathcal{F}_t, t \geq 0\}$ is ergodic in the sense that it has a unique stationary distribution and the distribution of $Y(t)$ converges in total variation, as $t \to \infty$, to this unique stationary distribution.*

Theorem 7.1, whose proof is deferred to the Appendix, validates the rightward arrow at the top of the "interchange of limits" diagram presented in Figure 1. On the other hand, the fluid limit theorem (Theorem 3.6 of [15]) justifies the downward arrow on the left-hand side of Figure 1. The focus of this work has been on understanding the convergence represented by the downward arrow on the right-hand side of Figure 1. When there is a unique invariant state, this convergence is established in Theorem 3.3. Although this question is not directly relevant to the characterization of the stationary distributions, it is natural in this setting to ask whether the diagram in Figure 1 commutes, namely, whether the fluid limit from any initial condition converges as $t \to \infty$ to the unique invariant state. In Section 7.1 we briefly discuss why the study of the long-time behavior of the fluid limit is a nontrivial task. Furthermore, in Section 7.2 we provide a very simple counterexample that shows that the diagram in Figure 1 need not commute and thus, the limits $N \to \infty$ and $t \to \infty$ cannot always be interchanged.

7.1. *Long-time behavior of the fluid limit.* The long-time behavior of the fluid limit is nontrivial even in the absence of abandonment. For example, in the absence of abandonment, it was proved in Theorem 3.9 of [17] that $\overline{\nu}_t \to \lambda\nu_*$ as $t \to \infty$ when $\lambda \in [0, 1]$, the service time distribution G^s has a second moment and its hazard rate function h^s is either bounded or lower-semicontinuous on (m_0, H^s) for some $m_0 < H^s$. The question of whether the second moment condition on the distribution is necessary for this convergence is still unresolved. Even under the second moment assumption, the long-time behavior of the component $\overline{X}$ of the fluid limit is not easy to describe except in the cases when

(i) the system is subcritical ($\lambda < 1$) or (ii) when the system is critical or supercritical ($\lambda \geq 1$) and the service distribution is exponential. In case (i), it follows from Theorem 3.9 of [17] that $\overline{X}(t) \to \lambda\langle \mathbf{1}, \nu_* \rangle$ as $t \to \infty$, whereas in case (ii), if the initial condition satisfies $\overline{X}(0) \geq 1$ and $\overline{\nu}_0 \in \mathcal{M}_F[0,\infty)$, then it is easy to see that the fluid limit is given explicitly by $\overline{X}(t) = \overline{X}(0) + (\lambda - 1)t$ and $\overline{\nu}_t(dx) = \mathbb{1}_{[0,t]} e^{-x}\,dx + \mathbb{1}_{(t,\infty)}(x)e^{-t}\overline{\nu}_0(d(x-t))$. Therefore, at criticality ($\lambda = 1$), if $\overline{X}(0) = 1$ then $\overline{X}(t) = \overline{X}(0)$ for every $t > 0$. In particular, $\overline{X}(t) \to 1$ as $t \to \infty$. However, as the following example demonstrates, the critical fluid limit need not converge to 1 [even if critically loaded and with initial condition $\overline{X}(0) = 1$] when the service distribution is not exponential.

EXAMPLE 7.2. Let the fluid arrival rate be $\overline{E}(t) = t$, $t > 0$, and let the service time distribution G^s be the Erlang distribution with density

$$g^s(x) = 4xe^{-2x}, \qquad x \geq 0.$$

A simple calculation shows that $\int_0^\infty (1 - G^s(x))\,dx = 1$. Let $(\overline{X}, \overline{\nu})$ be the solution to the fluid equations without abandonment (see Definition 5.2) associated with the initial condition $(\mathbf{1}, 1, \delta_0)$. We show below that in this case, $\lim_{t\to\infty} \overline{X}(t) = 5/4$, which is bigger than $1 = \overline{X}(0)$. In fact, since $\overline{\nu}_0 = \delta_0$, a straightforward calculation shows that

$$\langle h^s, \overline{\nu}_0 \rangle = \int_0^\infty \frac{g^s(x)}{1 - G^s(x)} \overline{\nu}_0(dx) = \frac{g^s(0)}{1 - G^s(0)} = g^s(0) = 0.$$

Define

$$\kappa \doteq \inf\{t \geq 0 : \langle h^s, \overline{\nu}_t \rangle \geq 1\}.$$

The hazard rate function h^s is bounded and continuous and $\langle h^s, \overline{\nu}_0 \rangle < \lambda = 1$. Therefore, substituting h^s in (5.4), it is clear that $\kappa > 0$ and $\langle h^s, \overline{\nu}_t \rangle < \lambda = 1$ for $t \in [0, \kappa)$. In turn, by the nonidling condition, this implies $\langle \mathbf{1}, \overline{\nu}_t \rangle = 1$ and $d\overline{K}/dt = \langle h^s, \overline{\nu}_t \rangle$ and, by (5.4), for $t \in [0, \kappa)$,

$$\langle h^s, \overline{\nu}_t \rangle = g^s(t) + \int_0^t g^s(t-s) \frac{d\overline{K}}{dt}(s)\,ds = g^s(t) + \int_0^t g^s(t-s)\langle h^s, \overline{\nu}_s \rangle\,ds.$$

Applying the key renewal theorem to the above equation, it follows that

$$\langle h^s, \overline{\nu}_t \rangle = u^s(t) = 1 - e^{-4t}.$$

Since $u^s(t) < 1$ for all $t \geq 0$, we must have that $\kappa = \infty$, $\langle \mathbf{1}, \overline{\nu}_t \rangle = 1$ for all $t \geq 0$, and

$$\lim_{t\to\infty} \overline{Q}(t) = \int_0^\infty \big(1 - u^s(t)\big)\,dt = \int_0^\infty e^{-4t}\,dt = 1/4,$$

which yields the convergence of $\overline{X}(t)$ to 5/4 as $t \to \infty$.

To emphasize that this phenomenon is not the consequence of the fact that the initial condition was chosen to be singular with respect to Lebesgue measure, we show that we can modify the above example by choosing $\overline{\nu}_0$ to be absolutely continuous with respect to the Lebesgue measure. For example, for some $\alpha \in (0,\infty)$, define

$$q(x) \doteq \begin{cases} \dfrac{1+2x}{\alpha+\alpha^2}, & \text{if } x \in [0,\alpha], \\ 0, & \text{otherwise,} \end{cases}$$

and let $\overline{\nu}_0(dx) = q(x)\,dx$. Then $\langle \mathbf{1}, \overline{\nu}_0\rangle = \int_0^\alpha q(x)\,dx = 1$, $\langle h^s, \overline{\nu}_t\rangle = 1 - ((1-\alpha)/(\alpha+1))e^{-4t}$ for each $t \geq 0$. Hence, when $\alpha < 1$ we have $\langle h^s, \overline{\nu}_t\rangle < 1$ and $\langle \mathbf{1}, \overline{\nu}_t\rangle = 1$ for all $t \geq 0$. This implies that, when $\alpha < 1$,

$$\lim_{t\to\infty} \overline{Q}(t) = \int_0^\infty \frac{1-\alpha}{\alpha+1} e^{-4t}\,dt = \frac{1-\alpha}{4(\alpha+1)} > 0,$$

showing that $\lim_{t\to\infty} \overline{X}(t) > 1$.

7.2. *A counterexample* (*invalidity of the interchange of limits*). In this section we show that even for an $M/M/N$ queue (both with and without abandonments), the "interchange of limits" need not hold, that is, the diagram presented in Figure 1 may not commute.

Consider the sequence of state processes $(X^{(N)}, \nu^{(N)})$, $N \in \mathbb{N}$, of N-server queues without abandonment, where the service time distribution G^s is exponential with rate 1. For the Nth queue, let the arrival process $E^{(N)}$ be a Poisson process with parameter $\lambda^{(N)} = N-1$ and suppose that there exists $\overline{\nu}_0 \in \mathcal{M}_F[0,\infty)$ with $\langle \mathbf{1}, \overline{\nu}_0\rangle \leq 1$ such that a.s., as $N \to \infty$,

$$(\overline{X}^{(N)}(0), \overline{\nu}_0^{(N)}) \to (2, \overline{\nu}_0). \tag{7.1}$$

Given the exponentiality of the service time distribution, it immediately follows that Assumption 2 of [17] is satisfied. Moreover, because (7.1) holds and $\overline{\lambda}^{(N)} = (N-1)/N \to 1$ as $N \to \infty$, it follows that Assumption 1 of [17] also holds with $\lambda = 1$. On the other hand, since $G^r(x) = 0$ for all $x \in [0,\infty)$, Assumption 2 fails to hold because in this case $B_1 = [1,\infty)$ and so the invariant manifold has uncountably many points.

Now, because Assumptions 1 and 2 of [17] are satisfied, we can apply the fluid limit result in Theorem 3.7 of [17] to conclude that, almost surely, as $N \to \infty$, $(\overline{X}^{(N)}, \overline{\nu}^{(N)})$ converges weakly to the unique solution $(\overline{X}, \overline{\nu})$ of the fluid equations associated with initial data $(\mathbf{1}, 2, \overline{\nu}_0)$, and using the exponentiality of the service time distribution, it is easily verified that the fluid limit is given explicitly by $\overline{X}(t) = \overline{X}(0) = 2$ and $\overline{\nu}_t(dx) = \mathbb{1}_{[0,t]} e^{-x}\,dx + \mathbb{1}_{(t,\infty)}(x) e^{-t} \overline{\nu}_0(d(x-t))$.

For each $N \in \mathbb{N}$, because the arrival rate, which equals $N-1$, is less than the total service rate N, by (3.2.4) and (3.2.5) of [4] it follows that $X^{(N)}$ is ergodic and has the following stationary distribution:

$$\mathbb{P}(X_*^{(N)} = k) = \begin{cases} \dfrac{(N-1)^k}{k!} p_0, & \text{if } k = 0, 1, \ldots, N-1, \\ \dfrac{(N-1)^k}{N! N^{k-N}} p_0, & \text{if } k = N, N+1, \ldots, \end{cases}$$

where

$$p_0 \doteq \left\{ \sum_{i=0}^{N-1} \frac{(N-1)^i}{i!} + \frac{(N-1)^N}{(N-1)!} \right\}^{-1}.$$

It follows from Stirling's formula that

$$\begin{aligned} \lim_{N\to\infty} \frac{\sum_{i=0}^{N-1} (N-1)^i/i!}{(N-1)^N/(N-1)!} &= \lim_{N\to\infty} \frac{\sum_{i=0}^{N-1} (N-1)^i/i!}{(1/\sqrt{2\pi})\sqrt{N-1}e^{N-1}} \\ &\leq \lim_{N\to\infty} \frac{\sum_{i=0}^{\infty} (N-1)^i/i!}{(1/\sqrt{2\pi})\sqrt{N-1}e^{N-1}} = 0. \end{aligned}$$

For each $\varepsilon > 0$, elementary calculations show that

$$\begin{aligned} \mathbb{P}(X_*^{(N)} \geq N + \varepsilon N) &= \sum_{k=N+\varepsilon N}^{\infty} \frac{(N-1)^k}{N! N^{k-N}} p_0 \\ &= \frac{N^N}{N!} p_0 \sum_{k=N+\varepsilon N}^{\infty} \left(\frac{N-1}{N}\right)^k \\ &= \frac{N^N}{N!} p_0 \left(\frac{N-1}{N}\right)^{N+\varepsilon N} N \\ &= \frac{(N-1)^N}{(N-1)!} p_0 \left(\frac{N-1}{N}\right)^{\varepsilon N} \end{aligned}$$

and

$$\mathbb{P}(X_*^{(N)} \leq N - \varepsilon N) = \sum_{k=0}^{N-\varepsilon N} \frac{(N-1)^k}{k!} p_0.$$

Combining the above three displays, we then have for each $\varepsilon > 0$

$$\lim_{N\to\infty} \mathbb{P}(\overline{X}_*^{(N)} \geq 1 + \varepsilon) = \lim_{N\to\infty} \mathbb{P}(X_*^{(N)} \leq 1 - \varepsilon) = 0. \tag{7.2}$$

Using the distribution of $X_*^{(N)}$ it can also be shown that

$$\sup_{N\in\mathbb{N}} \mathbb{E}[\overline{X}_*^{(N)}] = \sup_{N\in\mathbb{N}} \frac{\mathbb{E}[X_*^{(N)}]}{N} \leq 3.$$

An application of Markov's inequality then shows that the sequence $\{\overline{X}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight. Let $\overline{x}_*$ denote a subsequential weak limit of $\{\overline{X}_*^{(N)}\}_{N\in\mathbb{N}}$. Then (7.2) clearly shows that almost surely, $\overline{x}_* = 1$. Thus, as $N \to \infty$ $\overline{X}_*^{(N)}$ converges weakly to 1 (see Theorem 1 of [11] for a more refined calculation that also identifies the limit of the sequence of stationary distributions centered around N and divided by $\sqrt{N}$). We have shown that the sequence of stationary distributions does not converge (even along a subsequence) to the value 2, thus demonstrating that the interchange of limits does not hold even in this simple setting.

In addition, this example also demonstrates that even in the presence of multiple invariant states, the sequence of scaled stationary distributions $(\overline{X}_*^{(N)}, \overline{\eta}_*^{(N)}, \overline{\nu}_*^{(N)})$, $N \in \mathbb{N}$, could still converge to a limit. In the above example, the explicit formula of the stationary distribution of the $M/M/N$ queue was used to compute this limit, which defeats the whole purpose of the approximation. It is unclear whether, in the presence of multiple invariant states, there is a general methodology that does not rely on a priori knowledge of the stationary distributions of the N-server queues, but that would nevertheless allow one to identify when a limit exists and, in that case, identify which invariant state corresponds to the limit.

A minor modification of the above example shows that the interchange of limits can also fail to hold in the presence of abandonment. For the same sequence of queues described above, suppose that customers abandon the queue according to a nontrivial patience time distribution G^r satisfying Assumption 3.4 and having support in $(3, \infty)$. For each $N \in \mathbb{N}$, consider the marginal state process $(X^{(N)}, \nu^{(N)}, \eta^{(N)})$. Suppose that there exists $(2, \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$ such that almost surely, as $N \to \infty$,

$$(7.3) \qquad \big(\overline{X}^{(N)}(0), \overline{\nu}_0^{(N)}, \overline{\eta}_0^{(N)}\big) \to (2, \overline{\nu}_0, \overline{\eta}_0).$$

Given the assumption imposed on the patience time distribution, Assumption 2 fails to hold because in this case $B_1 = [1, 3]$. By the previous argument, Assumptions 3.1, 3.3 and 3.4 are satisfied. Therefore, by the fluid limit result stated as Theorem 3.6 of [15] (see also the proof of Theorem 3.3 of the current paper) it follows that almost surely, as $N \to \infty$, $(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ converges weakly to the unique solution $(\overline{X}, \overline{\nu}, \overline{\eta})$ of the fluid equations associated with $(\mathbf{1}, 2, \overline{\nu}_0, \overline{\eta}_0)$. By the exponentiality of the service time distribution, we have $\overline{X}(t) = \overline{X}(0) = 2$ and $\overline{R}(t) = 0$ for each $t \geq 0$. On the other hand, let $\overline{Y}_*^{(N)} = (\overline{\alpha}_{E,*}^{(N)}, \overline{X}_*^{(N)}, \overline{\nu}_*^{(N)}, \overline{\eta}_*^{(N)})$ be the stationary distribution associated with the fluid-scaled state process, which exists by Theorem 4.9. By a simple coupling argument, it can be shown that $X^{(N)}$ is stochastically dominated by the corresponding state $\tilde{X}^{(N)}$ of an $M/M/N$ queue without abandonment that has the same arrival process $E^{(N)}$ and the same initial condition [i.e., $\mathbb{P}(\tilde{X}^{(N)} \geq c) \geq \mathbb{P}(X^{(N)} \geq c)$ for every $c > 0$]. Together with the previous discussion of the case without abandonment, this can be used to show that $\{\overline{X}_*^{(N)}\}_{N\in\mathbb{N}}$ is tight and, for any $\varepsilon > 0$, $\limsup_{N\to\infty} \mathbb{P}(\overline{X}_*^{(N)} \geq (1 + \varepsilon)) = 0$, from which one can conclude that $\overline{X}_*^{(N)} \to 1$. Thus, in this case too,

$$\lim_{N\to\infty}\lim_{t\to\infty}\overline{X}^{(N)}(t) = \lim_{N\to\infty}\overline{X}_*^{(N)} = 1 \neq 2 = \lim_{t\to\infty}\overline{X}(t) = \lim_{t\to\infty}\lim_{N\to\infty}\overline{X}^{(N)}(t),$$

where the limits are all in the sense of weak convergence.

APPENDIX: PROOF OF THEOREM 7.1

By Theorem 6.1 of [20], to show that the Feller process $\{Y_t, \mathcal{F}_t, t \geq 0\}$ is ergodic, it suffices to show that the skeleton chain $\{Y_n\}_{n\in\mathbb{N}}$ is ψ-irreducible and that $\{Y_t, \mathcal{F}_t, t \geq 0\}$ is positive Harris recurrent. This is done in Lemma A.3 and Theorem A.5 below. Let ϱ be the quantity defined in condition (2) of Assumption 7.1, and define

$$\mathcal{Z} \doteq \{(\alpha, 0, \mathbf{0}, \mathbf{0}) \in \mathcal{Y} : \alpha \in [\varrho + 1, \infty)\}.$$

For each Borel subset A of $\mathcal{Z}$, let $\Gamma_A \subset [1 + \varrho, \infty)$ be the Borel subset obtained by projecting $\mathcal{Z}$ to its first coordinate:

$$\Gamma_A \doteq \{\alpha \in \mathbb{R} : (\alpha, 0, \mathbf{0}, \mathbf{0}) \in A\}. \tag{A.1}$$

LEMMA A.1. *There exists a strictly positive continuous function C on $\mathcal{Y}$ such that for every $y = (\alpha, x, \sum_{i=1}^k \delta_{u_i}, \sum_{j=1}^l \delta_{z_j}) \in \mathcal{Y}$, every Borel subset $A \subset \mathcal{Z}$ and every $t > 2\varrho + 1$,*

$$\begin{aligned} &\mathbb{P}_y\big(Y(t) \in A\big) \\ &\qquad \geq C(y) \int_{\alpha+2\varrho+1}^{\alpha+t} \mathbb{1}_{\Gamma_A}(\alpha + t - s)\big(1 - F(\alpha + t - s)\big)\, dF(s). \end{aligned} \tag{A.2}$$

PROOF. At time t, if the state $Y(t)$ is in the set $A \subset \mathcal{Z}$, this means that, by time t, all customers in service at time 0 with residual service times $\{u_i, 1 \leq i \leq k\}$, all customers in queue at time 0 with residual patience times $\{z_j, 1 \leq j \leq l\}$ and those new customers that arrived in the interval $[0, t)$ have completed service (if they entered service before time t) and have run out of their patience (irrespective of whether or not they entered service). Now, we consider a subset of $\{\omega : Y(t, \omega) \in A\}$, in which (a) by time $2\varrho + 1 < t$, all the initial customers with residual patience times $\{z_j, 1 \leq j \leq l\}$ and residual service times $\{u_i, 1 \leq i \leq k\}$ have finished service (if they entered service) and run out of their patience (irrespective of whether or nor they entered service), (b) the first new customer arrived after $2\varrho + 1$, finished service before t and ran out of his/her patience time before t, (c) the difference between t and the arrival time of that customer lies in Γ_A, and (d) the second new customer arrived after time t. Let $\mathcal{Q}_a$, $\mathcal{Q}_{ad}$ and $\mathcal{Q}_{bd}$, respectively, be the events that property (a) holds, properties (a)–(d) hold and properties (b)–(d) hold. Then, for $y \in \mathcal{Y}$,

$$\mathbb{P}_y\big(Y(t) \in A\big) \geq \mathbb{P}_y(\mathcal{Q}_{ad}) = \mathbb{P}_y(\mathcal{Q}_a)\mathbb{P}_y(\mathcal{Q}_{bd}|\mathcal{Q}_a),$$

and, due to the independence assumptions on the service, patience and interarrival distributions, $\mathbb{P}_y(\mathcal{Q}_{bd}|\mathcal{Q}_a)$ is greater than or equal to

$$\begin{aligned}
&\int_{\alpha+2\varrho+1}^{\alpha+t} G^r(\alpha+t-s)G^s(\alpha+t-s) \\
&\qquad \times \mathbb{1}_{\Gamma_A}(\alpha+t-s)\big(1-F(\alpha+t-s)\big)\frac{dF(s)}{1-F(\alpha)} \\
&\qquad \geq \frac{G^r(\varrho+1)G^s(\varrho+1)}{1-F(\alpha)}\int_{\alpha+2\varrho+1}^{\alpha+t} \mathbb{1}_{\Gamma_A}(\alpha+t-s)\big(1-F(\alpha+t-s)\big)\,dF(s),
\end{aligned}$$

where the last inequality holds because $\alpha+t-s \geq \varrho+1$ when $\alpha+t-s \in \Gamma_A$. Let $C(y) \doteq (\mathbb{P}_y(\mathcal{Q}_a)G^r(\varrho+1)G^s(\varrho+1))/(1-F(\alpha))$. Since, due to Assumption 7.1(2), $G^r(A) > 0$ and $G^s(A) > 0$ for any interval A with length bigger than ϱ, $\mathbb{P}_y(\mathcal{Q}_a)$, as a function of $y \in \mathcal{Y}$, is strictly positive and continuous. Thus C is a strictly positive and continuous function on $\mathcal{Y}$, and the lemma is proved. $\square$

DEFINITION A.2. Any Markov process $\{X_t\}$ with topological state space $\mathcal{X}$ is said to be ψ-irreducible if and only if there exists a σ-finite measure ψ on $\mathcal{B}(\mathcal{X})$, the Borel σ-algebra on $\mathcal{X}$ such that for every $x \in \mathcal{X}$ and $B \in \mathcal{B}(\mathcal{X})$,

$$\int_0^\infty \mathbb{P}_x\big(X(t) \in B\big)\,dt > 0 \qquad \text{if } \psi(B) > 0.$$

Let $\psi = m \times \delta_0 \times \delta_0 \times \delta_0$, where $m(A) = \overline{m}(A \cap [\varrho+1,\infty))$, where $\overline{m}$ is Lebesgue measure. Clearly, ψ is a σ-finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$.

LEMMA A.3. *The Markov process $\{Y_t, \mathcal{F}_t\}$ is ψ-irreducible and the discrete-time Markov chain $\{Y(n)\}_{n\in\mathbb{N}}$ is ψ-irreducible.*

PROOF. Let $B \in \mathcal{B}(\mathcal{Y})$ be such that $\psi(B) > 0$. Then $\psi(B \cap \mathcal{Z}) > 0$ by the definition of ψ. Let $\Gamma_{B\cap\mathcal{Z}}$ be the set defined in (A.1) with $A = B \cap \mathcal{Z}$ and suppose $m(\Gamma_{B\cap\mathcal{Z}}) > 0$. Fix $y \in \mathcal{Y}$. It follows from Lemma A.1 that there exists a strictly positive function C on $\mathcal{Y}$ such that

$$\begin{aligned}
&\int_0^\infty \mathbb{P}_y\big(Y(t) \in B \cap \mathcal{Z}\big)\,dt \\
&\qquad \geq \int_{2\varrho+1}^\infty \mathbb{P}_y\big(Y(t) \in B \cap \mathcal{Z}\big)\,dt \\
&\qquad \geq \int_{2\varrho+1}^\infty C(y)\bigg(\int_{\alpha+2\varrho+1}^{\alpha+t} \mathbb{1}_{\Gamma_{B\cap\mathcal{Z}}}(\alpha+t-s)\big(1-F(\alpha+t-s)\big)\,dF(s)\bigg)\,dt \\
&\qquad = C(y)\big(1-F(\alpha+2\varrho+1)\big)\int_{\Gamma_{B\cap\mathcal{Z}}}\big(1-F(t)\big)\,dt \\
&\qquad > 0,
\end{aligned}$$

where the equality follows from Fubini's theorem and the last inequality holds because $C(y) > 0$, $m(\Gamma_{B\cap\mathcal{Z}}) > 0$ and $1 - F(x) > 0$ for every $x \in [0, \infty)$ by Assumption 7.1(3). This establishes the first assertion. On the other hand, for $n > 2\varrho + 1$,

$$\mathbb{P}_y\big(Y(n) \in B\big) \geq C(y) \int_{\alpha+2\varrho+1}^{\alpha+n} \mathbb{1}_{\Gamma_{B\cap\mathcal{Z}}}(\alpha + n - s)\big(1 - F(\alpha + n - s)\big)\, dF(s).$$

By Assumption 7.1(3) and the fact that $m(\Gamma_{B\cap\mathcal{Z}}) > 0$, it follows that $\mathbb{P}_y(Y(n) \in B) > 0$ for all n sufficiently large. Hence, $\{Y(n)\}_{n\in\mathbb{N}}$ is ψ-irreducible. $\square$

For each $y \in \mathcal{Y}$, $B \in \mathcal{B}(\mathcal{Y})$ and each probability measure Π on $[0, \infty)$, let

$$\mathcal{K}_\Pi(y, B) = \int_0^\infty \mathbb{P}_y\big(Y(t) \in B\big)\Pi(dt).$$

LEMMA A.4. *There exists a probability measure Π on $[0, \infty)$ and a function $T : \mathcal{Y} \times \mathcal{B}(\mathcal{Y}) \to \mathbb{R}_+$ such that*:

(1) $\mathcal{K}_\Pi(y, B) \geq T(y, B)$ *for all $y \in \mathcal{Y}$ and every Borel measurable set $B \in \mathcal{B}(\mathcal{Y})$*;

(2) $T(y, \mathcal{Y}) > 0$ *for all $y \in \mathcal{Y}$*;

(3) $T(\cdot, B)$ *is lower-semicontinuous for every $B \in \mathcal{B}(\mathcal{Y})$*.

PROOF. Let C be the strictly positive, continuous function C of Lemma A.1. Let Π be a probability measure with density function $e^{-(t-2\varrho-1)}$ on $[2\varrho + 1, \infty)$. For each $y \in \mathcal{Y}$ and $B \subset \mathcal{Z}$, define

$$T(y, B) \doteq C(y)e^{\alpha+2\varrho+1} \int_{\alpha+2\varrho+1}^\infty e^{-s}\, dF(s) \int_0^\infty \big(1 - F(t)\big)\mathbb{1}_{\Gamma_B}(t)e^{-t}\, dt,$$

and $T(y, \mathcal{Y} \setminus \mathcal{Z}) = 0$. It is easy to see that for any Borel measurable set $B \in \mathcal{B}(\mathcal{Y})$, $T(y, B) = T(y, B \cap \mathcal{Z})$ and $T(\cdot, B)$ is continuous. Moreover, $T(y, \mathcal{Y}) = T(y, \mathcal{Z}) > 0$. Now, fix $y \in \mathcal{Y}$ and $B \in \mathcal{B}(\mathcal{Y})$. By Lemma A.1, we have

$$\begin{aligned}
&\mathcal{K}_\Pi(y, B)\\
&\quad= \int_0^\infty \mathbb{P}_y\big(Y(t) \in B\big)e^{-(t-2\varrho-1)}\, dt\\
&\quad\geq \int_{2\varrho+1}^\infty \mathbb{P}_y\big(Y(t) \in B \cap \mathcal{Z}\big)e^{-(t-2\varrho-1)}\, dt\\
&\quad\geq \int_{2\varrho+1}^\infty C(y)\bigg(\int_{\alpha+2\varrho+1}^{\alpha+t} \mathbb{1}_{\Gamma_{B\cap\mathcal{Z}}}(\alpha + t - s)\big(1 - F(\alpha + t - s)\big)\, dF(s)\bigg)\\
&\qquad\qquad \times e^{-(t-2\varrho-1)}\, dt\\
&\quad= C(y)e^{\alpha+2\varrho+1} \int_{\alpha+2\varrho+1}^\infty e^{-s}\, dF(s) \int_0^\infty \big(1 - F(t)\big)\mathbb{1}_{\Gamma_{B\cap\mathcal{Z}}}(t)e^{-t}\, dt\\
&\quad= T(y, B \cap \mathcal{Z}) = T(y, B).
\end{aligned}$$

Thus we have proved the lemma. □

THEOREM A.5. *The Markov process Y is positive Harris recurrent.*

PROOF. Lemma A.4 shows that Y is a so-called T process (cf. Section 3.2 of [20]) and Lemma A.3 shows that Y is ψ-irreducible. Now, Theorem 3.2 of [20] states that any ψ-irreducible T process Y is positive Harris recurrent if Y is bounded in probability on average, that is, for each $y \in \mathcal{Y}$ and $\varepsilon > 0$, there exists a compact set $B \in \mathcal{B}(\mathcal{Y})$ such that

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{P}_y\big(Y(s) \in B\big)\, ds \geq 1 - \varepsilon.$$

However, this is satisfied by the state process Y due to Lemma 4.8. So we have the desired result. □

Acknowledgment. We would like to thank Haya Kaspi for observing that the state process would not be Feller if α_E were chosen to be the forward, rather than the backward, recurrence time.

REFERENCES

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues*, 2nd ed. *Applications of Mathematics* (*New York*) **51**. Springer, New York. MR1978607

[2] ASMUSSEN, S. and FOSS, S. (1993). Renovation, regeneration and coupling in multiple-server queues in continuous time. In *Frontiers in Pure and Applied Probability* (H. Niemi, G. Högnas, A. N. Shiryaev and A. V. Melnikov, eds.) 1–6. BRILL, The Netherlands.

[3] BACCELLI, F. and HEBUTERNE, G. (1981). On queues with impatient customers. In *Performance* '81 (*Amsterdam*, 1981) 159–179. North-Holland, Amsterdam. MR0727303

[4] BOCHAROV, P. P., D'APICE, C., PECHINKIN, A. V. and SALERNO, S. (2004). *Queueing Theory*. de Gruyter, Berlin.

[5] BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. and ZHAO, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50. MR2166068

[6] DA PRATO, G. and ZABCZYK, J. (1996). *Ergodicity for Infinite-Dimensional Systems*. *London Mathematical Society Lecture Note Series* **229**. Cambridge Univ. Press, Cambridge. MR1417491

[7] FOSS, S. G. (1983). Ergodicity conditions in multiserver queueing systems with waiting time. *Sib. Math. J.* **24** 168–175. MR0731053

[8] GAMARNIK, D. and MOMČILOVIĆ, P. (2008). Steady-state analysis of a multiserver queue in the Halfin–Whitt regime. *Adv. in Appl. Probab.* **40** 548–577. MR2433709

[9] GAMARNIK, D. and ZEEVI, A. (2006). Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.* **16** 56–90. MR2209336

[10] GARNETT, O., MANDELBAUM, A. and REIMAN, M. I. (2002). Designing a call center with impatient customers. *Manufac. Serv. Oper. Management* **4** 208–227.

[11] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588. MR0629195

[12] JACOBSEN, M. (2006). *Point Process Theory and Applications*: *Marked Point and Piecewise Deterministic Processes*. Birkhäuser, Boston, MA. MR2189574

[13] JELENKOVIĆ, P., MANDELBAUM, A. and MOMČILOVIĆ, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Syst*. **47** 53–69. MR2074672

[14] KALLENBERG, O. (1976). *Random Measures*. Akademie-Verlag, Berlin. MR0431373

[15] KANG, W. and RAMANAN, K. (2010). Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* **20** 2204–2260. MR2759733

[16] KASPI, H. and RAMANAN, K. (2010). SPDE limits for many-server queues. Preprint.

[17] KASPI, H. and RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* **21** 33–114. MR2759196

[18] KIEFER, J. and WOLFOWITZ, J. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc*. **78** 1–18. MR0066587

[19] LITTLE, J. D. C. (1961). A proof for the queuing formula: $L = \lambda W$. *Oper. Res*. **9** 383–387. MR0125644

[20] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes. II. Continuous-time processes and sampled chains. *Adv. in Appl. Probab.* **25** 487–517. MR1234294

[21] PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. *Probability and Mathematical Statistics* **3**. Academic Press, New York. MR0226684

[22] WHITT, W. (2006). Fluid models for multiserver queues with abandonments. *Oper. Res*. **54** 37–54. MR2201245

[23] ZELTYN, S. and MANDELBAUM, A. (2005). Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst*. **51** 361–402. MR2189598

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MARYLAND, BALTIMORE COUNTY
1000 HILLTOP CIRCLE
BALTIMORE, MARYLAND 21250
USA
E-MAIL: wkang@umbc.edu

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: kavita@dam.brown.edu