Operations Research Letters 50 (2022) 362–369

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Asymptotically optimal idling in the $GI/GI/N{+}GI$ queue

Yueyang Zhong [a,*], Amy R. Ward [a], Amber L. Puha [b,1]

[a] The University of Chicago Booth School of Business, United States of America
[b] California State University San Marcos, United States of America

ARTICLE INFO

Article history:
Received 18 June 2021
Received in revised form 16 February 2022
Accepted 7 April 2022
Available online 12 April 2022

Keywords:
$GI/GI/N{+}GI$
Fluid control problem
Asymptotically optimal idling

ABSTRACT

We formulate a control problem for a $GI/GI/N{+}GI$ queue, whose objective is to trade off the long-run average operational costs with server utilization costs. To solve the control problem, we consider an asymptotic regime in which the arrival rate and the number of servers grow large. The solution to an associated fluid control problem motivates that non-idling service disciplines are not in general optimal, unless some arrivals are turned away. We propose an admission control policy designed to ensure that servers have sufficient idle time, which we show is asymptotically optimal.

1. Introduction

One common assumption when studying the $GI/GI/N{+}GI$ queue is that the service discipline is non-idling; that is, that servers do not idle when customers are present in the queue ([14,7,8,16,6]). However, in the restricted $M/M/N{+}M$ setting, the paper [15] (see Theorem 1, Proposition 1, and Example 1 therein) shows that in the presence of server utilization costs, a non-idling service discipline may not be asymptotically optimal. Our purpose in this paper is to show that a similar phenomenon occurs in the $GI/GI/N{+}GI$ setting; that is, a non-idling service discipline might be suboptimal in the non-Markovian setting, when the system operates in a first-come, first-served (FCFS) manner.

The $GI/GI/N{+}GI$ queue is more difficult to analyze than the $M/M/N{+}M$ queue because the state descriptor is more complex. In particular, tracking the one-dimensional number-in-system process is sufficient when studying the $M/M/N{+}M$ queue, but more is needed when studying the $GI/GI/N{+}GI$ queue. This is because a Markovian state descriptor must also include knowledge regarding the time that has elapsed since the last arrival, the amount of time each job in service has been in service, and the amount of time each job in the queue has waited, resulting in a measure-valued state descriptor.

The control question is to determine when an available server should take the next customer into service, and when such a server should idle for some period of time. Too much idleness may lead to customer abandonment and excessive waiting, whereas too little rest may lead to server fatigue. To quantify these two competing interests, we consider an objective function that trades off the abandonment costs (and also, as an extension, holding costs) with server utilization costs. Exact analysis of the $GI/GI/N{+}GI$ queue is intractable, and, therefore, we study the queue in an overloaded asymptotic regime in which the arrival rate and the number of servers become large. In that regime, we formulate a fluid control problem, and find that the solution to the fluid control problem sometimes motivates idling servers when customers are waiting (when operational costs are small compared to utilization costs). The policy we propose, and show is asymptotically optimal (see our main results in Theorems 1 and 2, and their extension to incorporate holding costs in the online appendix), is one that "thins" the arrival process just enough to ensure the server utilization matches the solution to the fluid control problem.

Incorporating server utilization in the objective function is one way to ensure that the service discipline does not overwork servers. This can lead to increased employee retention, which can have performance benefits (discussed in [13]). Not overworking servers means ensuring sufficient idleness for all servers, an idea that arose earlier in papers that studied how to be fair to heterogeneous servers that can be grouped into statistically identical pools (see, e.g., [4], [12]), and how to exploit heterogeneous customers preferences so as to maximize revenue (see, e.g., [1], [9]).

Notation. We denote the set of integers endowed with the discrete topology by $\mathbb{Z}$, the set of non-negative integers by $\mathbb{Z}_+$, the set of positive integers by $\mathbb{N}$, the set of real numbers endowed with the Euclidean topology by $\mathbb{R}$, and the set of non-negative real numbers by $\mathbb{R}_+$. For F, a cumulative distribution function (abbreviated c.d.f. henceforth) on $\mathbb{R}_+$ with density f, we write $\bar{F} = 1 - F$ and recall that the right edge of the support is given by $x_r =$

* Corresponding author.
E-mail address: yzhong0@chicagobooth.edu (Y. Zhong).
[1] Research supported by NSF Grant DMS-1712974.

https://doi.org/10.1016/j.orl.2022.04.005

$\sup\{x \in \mathbb{R}_+ : \bar{F}(x) > 0\}$ and the hazard function is $x \mapsto f(x)/\bar{F}(x)$ for $x \in [0, x_r)$. For a measurable space $(S, \mathcal{F})$ and a measurable set $A \in \mathcal{F}$, 1_A is the indicator function of the set A, which is one when its argument is a member of the set A and is zero otherwise. In addition, when A is S, we use the shorthand notation 1 to mean 1_S. For $H \in (0, \infty]$, let $\mathbf{M}[0, H)$ denote the set of finite, non-negative Borel measures on $[0, H)$ endowed with the topology of weak convergence. For a given $\eta \in \mathbf{M}[0, H)$ and a Borel measurable function $f : [0, H) \to \mathbb{R}_+$ that is integrable with respect to η, we write $\langle f, \eta \rangle = \int_{[0,H)} f(x)\eta(dx)$. The set $\mathbf{M}[0, H)$ endowed with the topology of weak convergence is a Polish space ([10]). We let $\mathbf{0} \in \mathbf{M}[0, H)$ be the measure such that $\langle f, \mathbf{0} \rangle = 0$ for all Borel measurable functions $f : [0, H) \to \mathbb{R}_+$. Given $x \in [0, H)$, δ_x denotes the Dirac measure in $\mathbf{M}[0, H)$ such that for all Borel measurable functions $f : [0, H) \to \mathbb{R}_+$, $\langle f, \delta_x \rangle = f(x)$. Then let $\mathbf{M}_D[0, H)$ denote the subset of $\mathbf{M}[0, H)$ consisting of the measures $\eta \in \mathbf{M}[0, H)$ such that either $\eta = \mathbf{0}$ or η can be represented as a sum of finitely many Dirac measures, that is, $\eta = \sum_{i=1}^n a_i \delta_{x_i}$, for some finite $n \in \mathbb{N}$, $(a_1, \ldots, a_n) \in (0, \infty)^n$ and $(x_1, \ldots, x_n) \in [0, H)^n$. Given a Polish space $\mathbb{S}$, we use $\mathbf{D}(\mathbb{S})$ to denote the set of $\mathbb{S}$ valued functions of $\mathbb{R}_+$ that are right continuous with finite lefts, endowed with the usual Skorokhod J_1-topology. Finally, we use $\Rightarrow$ to denote weak convergence and $\stackrel{d}{=}$ to denote equivalence in distribution.

2. The model and admissible policy class

In this paper, we study a single-class many server queue with generally distributed inter-arrival, service, and patience times (i.e., a $GI/GI/N{+}GI$ queue) operating under a head-of-the-line (HL) control policy, that may or may not be non-idling. This is as specified in [11] specialized to a single customer class. In particular, we consider the model specified in [7], but with the non-idling condition [7, (2.30)] removed. Absent the non-idling condition, the system dynamics are not uniquely specified. Hence, one must specify a control policy to determine when each customer in system will commence service. Such control policies should satisfy natural conditions such as not using information about the future to make scheduling decisions. In what follows, we describe the model and admissible policy class in brief. We refer the interested reader to [11] for details.

The model. Customers arrive according to a delayed renewal process E with rate $\lambda \in \mathbb{R}_+$, each with a service time sampled from c.d.f. G^s having finite mean $1/\mu \in (0, \infty)$, and a patience time (also known as reneging time) sampled from a c.d.f. G^r having finite mean $1/\theta \in (0, \infty)$. We denote the c.d.f. for the inter-arrival distribution associated with the renewal arrival as G. We assume G, G^s and G^r are absolutely continuous with density functions g, g^s and g^r respectively that have right edges of support H, H^s and H^r respectively and hazard function h, h^s and h^r respectively. We assume that there exists $0 \le L^s < H^s$ such that h^s is either bounded or lower-semicontinuous on (L^s, H^s) and h^r is bounded and continuous. Boundedness of h^r implies that $H^r = \infty$. Finally, we assume G^r is strictly increasing with inverse function $(G^r)^{-1}$. The queue indexed by $N \in \mathbb{N}$ has N identical servers and is defined on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For the remainder of this paper, we superscript all quantities that depend on N by N, e.g., G^N, g^N, H^N, λ^N and E^N depend on N, but G^s and G^r do not vary with N.

Following the notation in Section 2.2 in [11], the state descriptor for the N-server queue is denoted by $y^N = (\alpha^N, x^N, \nu^N, \eta^N) \in \mathbb{Y}_D$, where $\mathbb{Y}_D = \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathbf{M}_D[0, H^s) \times \mathbf{M}_D[0, H^r)$. In particular, $\alpha^N \in [0, H^N)$ is the time that has elapsed since the last customer arrived to the system, $x^N \in \mathbb{Z}_+$ is the number of customers in system, $\nu^N \in \mathbf{M}_D[0, H^s)$ is a measure that has a unit mass at the age-in-service (amount of service received) of each customer currently in service, and $\eta^N \in \mathbf{M}_D[0, H^r)$ is a measure that has a unit mass at the potential waiting time of each customer "potentially" in system. (That is, each unit mass tracks the time passed since a customer's arrival, until that customer's patience time expires, at which point the unit atom is removed and tracking stops.) When $Y^N(0)$ denotes the initial state, the coordinate $\alpha^N(0)$ determines the distribution of the initial delay for E^N as the conditional distribution of G^N given $\alpha^N(0)$. That is, the initial delay distribution has density $g_0^N(x) = \frac{g^N(\alpha^N(0)+x)}{1-G^N(\alpha^N(0))}$ for $x \in [0, H^N - \alpha^N(0))$.

A state process for the N-server queue is a $\mathbb{Y}_D$ valued, right continuous process Y^N with finite left limits that satisfies a set of dynamic equations for the N-server queue consistent with HL service. These are given as equations (5)-(26) in [11], which we omit here due to space constraints. With these, customers can only enter service at or after their arrival time and prior to their patience time expiring. An available server may idle or may take the customer in queue with the largest waiting time, the HL customer, into service. Once a server commences serving a customer, it works at rate one on the work associated with that customer until completely fulfilling that customer's service requirement, at which point the customer departs.

The admissible policy class. The admissible policy class consists of all policies that only allow customers to enter service at moments of a customer departure or arrival, do not use information about the future, and are such that the state process Y^N is a Feller Markov process with respect to a natural filtration, and whose initial condition is policy compatible. The following leverages [11] to make this more precise.

As mentioned above, equations (5)-(26) in [11] do not uniquely specify the system dynamics. These are uniquely determined by the specification of an HL control policy $\pi^N = (\mathbb{S}^N, \{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N})$. Here, as in Definition 1 in [11], $\mathbb{S}^N$ is the Polish subspace of $\mathbb{Y}_D$ that corresponds to the set of states that are achievable under the control policy. Also, for each initial state $y \in \mathbb{S}^N$, $\mathbb{P}_y^N$ is a probability measure that uniquely determines the system dynamics when the system starts in state y. More formally, $\{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N}$ is a collection of probability measures indexed by $\mathbb{S}^N$ such that the mapping $y \mapsto \mathbb{P}_y^N(B)$ from $\mathbb{S}^N$ to $[0, 1]$ is Borel measurable for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$ and, for each $y \in \mathbb{S}^N$, $\mathbb{P}_y^N$ almost surely,

$$Y^N(0) = y,\ Y^N \in \mathbf{D}(\mathbb{S}^N) \text{ and satisfies (5) – (26) in [11].} \tag{1}$$

Given an HL control policy π^N, a state process Y^N satisfying (1) specifies an entry-into-service process K^N. Indeed, since a job has age-in-service equal to zero at the time of entering service, $\langle 1_{\{0\}}, \nu^N(t) \rangle$ is the number of jobs to enter service at time t, for each $t > 0$. Then K^N is a counting process such that $K^N(0) = 0$ and $K^N(t) - K^N(t-) = \langle 1_{\{0\}}, \nu^N(t) \rangle$ for each $t > 0$. In particular, $K^N(t)$ is the number of customers that enter service by time t for each $t \ge 0$. Then, for each $t \ge 0$, $D^N(t) = \langle 1, \nu^N(0) \rangle + K^N(t) - \langle 1, \nu^N(t) \rangle$ denotes the number of customers to depart the system due to service completion by time t. We restrict attention to HL policies that only allow customers to enter service at moments of a customer departure or arrival. We require that for each $y \in \mathbb{S}^N$, $\mathbb{P}_y^N$ almost surely, for all $t \ge 0$,[2]

$$K^N(t) - K^N(t-) \le E^N(t) - E^N(t-) + D^N(t) - D^N(t-). \tag{2}$$

We allow for random initial states that are compatible with a given HL control policy $\pi^N = (\mathbb{S}^N, \{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N})$. As in Definition 2 in [11], an initial distribution for π^N is a Borel probability measure ς^N on $\mathbb{S}^N$ that determines the distribution of the initial state $Y^N(0)$. In particular, for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$,

[2] This condition is sufficient for a tightness result to hold as shown in [11].

define $\mathbb{P}_\varsigma^N(B) = \int_{\mathbb{S}^N} \mathbb{P}_y^N(B)\varsigma^N(dy)$. Then $\mathbb{P}_\varsigma^N$ denotes the distribution of the state process Y^N under π^N for initial distribution ς^N. We say that an initial distribution ς^N for π^N is compatible if $\mathbb{E}_\varsigma^N\left[\langle 1, \eta^N(0)\rangle\right] < \infty$, where $\mathbb{E}_\varsigma^N$ denotes the expectation operator for $\mathbb{P}_\varsigma^N$. Given an HL control policy π^N and a compatible initial distribution ς^N, we refer to the process Y^N with law $\mathbb{P}_\varsigma^N$ as the state process for (π^N, ς^N).

In order to restrict attention to HL control policies that do not use information about the future, we require K^N to be non-anticipating. This amounts to requiring K^N to be adapted to a suitable filtration as in Definition 3 in [11]. Because we consider long-run average cost, we make a further restriction in the definition of admissible HL control policies, which is used in Section 6 to establish the existence of a stationary distribution.

Definition 1 *(Admissible policies).* An admissible HL control policy for E^N is an HL control policy π^N such that for any compatible initial distribution ς^N, the pair (π^N, ς^N) (i) satisfies Definition 3 in [11] and (2) and (ii) is such that the state process Y^N for (π^N, ς^N) is a Feller Markov process with respect to the filtration used in Definition 3 in [11].

Remark 1. Our admissible policies focus on HL (equivalently, FCFS) control policies due to their common use in practice. However, non-HL control policies can be optimal in some settings; see [5].

Let Π^N denote the set of admissible HL control policies for E^N in Definition 1. For $\pi^N \in \Pi^N$, we will sometimes write $Y^N(\pi^N, \cdot)$, $X^N(\pi^N, \cdot)$, $\nu^N(\pi^N, \cdot)$, $\eta^N(\pi^N, \cdot)$, $K^N(\pi^N, \cdot)$ or $D^N(\pi^N, \cdot)$ to make the dependence on π^N explicit.

Proposition 1. *For any $\pi^N \in \Pi^N$, there exists a compatible initial distribution ξ^N such that the state process Y_∞^N for (π^N, ξ^N) is a stationary process.*

Proposition 1 follows as a special case of Lemma 1 stated in Section 7.

Given $\pi^N \in \Pi^N$ and a compatible initial distribution ξ^N such that the state process Y_∞^N for (π^N, ξ^N) is a stationary process, we refer to ξ^N as a compatible stationary distribution for π^N and we let $\mathcal{S}(\pi^N)$ denote the set of all compatible stationary distributions for π^N.

3. The control problem

Each customer abandonment incurs a cost $a \in (0, \infty)$ and the strictly increasing, continuous and convex function $g_U : [0, 1] \to [0, \infty)$ captures the cost of server utilization. The trade-off is between working the servers as much as possible, which incurs high utilization cost but low abandonment cost, and giving the servers more rest, which incurs lower utilization cost but higher abandonment cost. In particular, given $\pi^N \in \Pi^N$ and a compatible initial distribution ς^N, we define the long-run average cost of (π^N, ς^N) as

$$C_\varsigma^N(\pi^N) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\varsigma^N\left[a\frac{R^N(\pi^N, T)}{N} + \int_0^T g_U\left(\frac{B^N(\pi^N, t)}{N}\right)dt\right],$$

where, for each $t > 0$, $R^N(\pi^N, t)$ is the cumulative number of abandonments by time t under π^N, and $B^N(\pi^N, t) \le N$ is the number of busy servers at time t under π^N.

Proposition 2. *For any $\pi^N \in \Pi^N$ and compatible initial distribution ς^N, there exists $\xi^N \in \mathcal{S}(\pi^N)$ such that $C_\varsigma^N(\pi^N) = C_\xi^N(\pi^N)$.*

Proposition 2 follows as a special case of Lemma 2 stated in Section 7.

Given $\pi^N \in \Pi^N$, let $C^N(\pi^N) := \sup_{\xi^N \in \mathcal{S}(\pi^N)} C_\xi^N(\pi^N)$ denote the worst case cost. By Proposition 2, $C^N(\pi^N)$ is the supremum of $C_\varsigma^N(\pi^N)$ over all compatible initial distributions ς^N. Our objective is to find an admissible control policy π_{opt}^N such that

$$C^N(\pi_{\text{opt}}^N) := \inf_{\pi^N \in \Pi^N} C^N(\pi^N). \tag{3}$$

The objective is such that a non-idling control policy is not in general optimal. Based on the discrete-event queuing model, it is not possible to solve for π_{opt}^N exactly. Thus, we leverage an analytically tractable approximating fluid control problem to postulate an HL control policy that one might expect to perform well for the objective (3). Then, we show that this policy is asymptotically optimal (see Theorems 1 and 2 in Section 6).

4. The fluid control problem

The fluid control problem is based on the fluid model and the fluid model solutions defined in [11]. Fluid model solutions arise as functional law of large numbers limits of sequences of state descriptors for the stochastic system under fluid scaling. For each $N \in \mathbb{N}$, we define the fluid scaling for the N-server system as follows. Recall the constant λ^N and the processes E^N, α^N, X^N, ν^N, η^N, K^N and D^N defined in Section 2, and the processes R^N and B^N defined in Section 3; also define the process $Q^N = X^N - B^N$ as the queue length, and the process $I^N = N - B^N$ as the number of idle servers. Then, let $\bar{\alpha}^N = \alpha^N$; also for $\triangle^N = \lambda^N, E^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N, I^N$, let $\bar{\triangle}^N = \triangle^N/N$. Then, the fluid-scaled state process for the N-server system is $\bar{Y}^N = (\bar{\alpha}^N, \bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)$. Under suitable asymptotic conditions, limit points exist and are fluid model solutions almost surely (see Lemma 4 in Section 7).

In particular, fluid model solutions are functions of time that take values in the set $\mathbb{X} = \mathbb{R}_+ \times \mathbf{M}[0, H^s) \times \mathbf{M}[0, H^r)$ endowed with the product topology. Then a state $(x, \nu, \eta) \in \mathbb{X}$ for the fluid model is a fluid analog of the state descriptor for the stochastic system with x, $\langle 1_{[0,z]}, \nu\rangle$ and $\langle 1_{[0,z]}, \eta\rangle$ corresponding to the total mass in system, the total mass in service with age-in-service less than or equal to z for each $z \in \mathbb{R}_+$, and the total mass potentially in system of age less than or equal to z for each $z \in \mathbb{R}_+$, respectively. They satisfy a set of conditions determined by a positive constant γ, which is the rate at which "fluid" or mass arrives to the system. These conditions are referred to as the fluid model for γ. We summarize the fluid model for γ and the definition of a fluid model solution for γ in Appendix A.

The invariant states for the fluid model for γ are fixed points of the fluid model for γ. From Proposition 1 in [11], an invariant state for γ is determined by the long-run average fraction of the collective server effort provided to the customers, denoted by b. It is clear that b must satisfy $b \in [0, \min\{1, \gamma/\mu\}]$, where we recall that μ is the reciprocal of the mean of G^s. Then, when the initial state for a fluid model solution for γ is an invariant state for γ, it turns out that the departure rate of the fluid from the system is $b\mu$ and so, by conservation of mass, $\gamma - b\mu$ must be the rate at which fluid abandons. This implies that the abandonment rate is insensitive to the patience time distribution, which has a similar flavor to the insensitivity result for a single server queue in the large deviations regime in [2].

Assumption 1. Let $\lambda \in (0, \infty)$. Suppose that $\lim_{N\to\infty} \bar{\lambda}^N = \lambda$.

Henceforth, λ satisfying the conditions in Assumption 1 is fixed. Our fluid control problem is based on the invariant states for λ. We expect to obtain the following fluid control problem for λ when letting $N \to \infty$ in problem (3).

Definition 2 *(The fluid control problem).* The fluid control problem for λ is given by

$$\min_{b\in[0,\min\{1,\lambda/\mu\}]} a(\lambda - b\mu) + g_U(b). \tag{4}$$

We denote the solution to (4) by b_* (which exists and is unique because (4) optimizes a convex function over a compact set).

Example 1. Suppose $a = 1$ and $g_U(b) = b^2$. Then, the solution to (4) is $b_* = \min\{1, \mu/2, \lambda/\mu\}$.

The solution to (4) motivates a control policy that we expect to have good performance with respect to the original objective (3) when the arrival rate λ^N and the number of servers N are large. When $b_* = \min\{1, \lambda/\mu\}$, we expect a non-idling control policy to be optimal for (3). Otherwise, when $b_* < \min\{1, \lambda/\mu\}$, the solution to the fluid control motivates defining a policy that uses customer abandonments to trim congestion, in order to reduce server workload, and provide (additional) server idle time. In this case, for each $N \in \mathbb{N}$, consider the HL control policy $\tilde{\pi}^N$ such that each server idles after each service completion for the difference between the desired expected time between service completions, $(b_*\mu)^{-1}$, and the expected time between service completions when the server is always busy, μ^{-1}; that is, for $(b_*\mu)^{-1} - \mu^{-1} = (1 - b_*)(b_*\mu)^{-1}$ time units. Such a policy seems quite reasonable, and should be asymptotically optimal. However, establishing that for any sequence of compatible initial distributions $\{\varsigma^N\}_{N\in\mathbb{N}}$,

$$\begin{aligned} &\lim_{N\to\infty}\lim_{t\to\infty}\frac{1}{t}\mathbb{E}^N_\varsigma\left[\bar{R}^N(\tilde{\pi}^N, t)\right] = \lambda - b_*\mu \text{ and} \\ &\lim_{N\to\infty}\lim_{t\to\infty}\mathbb{E}^N_\varsigma\left[g_U\left(\bar{B}^N(\tilde{\pi}^N, t)\right)\right] = g_U(b_*) \end{aligned} \tag{5}$$

is difficult. This difficulty is related to a lack of results providing sufficient conditions for fluid model solutions to converge to invariant states in the time infinity limit (see Section 7.1 in [8]). Instead, we propose to expand the admissible policy class to include thinned arrival processes and then rely on results in the literature for non-idling many server queues to show that (5) holds. If we can show a policy is asymptotically optimal for an enlarged policy class, then we know that no policy in the original smaller policy class can perform better.

5. The proposed policy π^N_*

The solution $0 \le b_* \le \min\{1, \lambda/\mu\}$ to (4) represents the optimal long-run average fraction of busy servers, which suggests that a control policy that thins the arrival process to rate $b_*\mu$ and forces the servers to work in a non-idling fashion, but builds in idleness due to admission control, should perform well for the original objective (3). This motivates us to enlarge the admissible policy class in Definition 1 to allow for admission control. Specifically, at the time of each arrival, let $p \in (0, 1]$ be the probability the arrival is admitted for service and $1 - p$ the probability the arrival is rejected, which incurs a cost a. Given $p \in (0, 1]$, we denote the admitted arrival process by E^N_p, and we refer to the N-server queue with arrival process E^N_p as the p-admitted queue. It is clear that the thinned arrival process E^N_p is a suitably delayed renewal process with arrival rate $p\lambda^N$, because the admitted arrivals remain i.i.d.

Definition 3 *(Enlarged admissible policies).* For any $p \in (0, 1]$, an admissible HL control policy for E^N_p satisfies Definition 1 with E^N replaced by E^N_p.

For $p \in (0, 1]$, let Π^N_p denote the set of admissible HL control polices for E^N_p. Note that $\Pi^N_1 = \Pi^N$. For $p \in (0, 1]$, $\pi^N_p \in \Pi^N_p$ and $\triangle^N = Y^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N$ or I^N, $\triangle^N(\pi^N_p, \cdot)$ refers to the process for the p-admitted queue under π^N_p.

Given $p \in (0, 1]$, $\pi^N_p \in \Pi^N_p$ and a compatible initial distribution ς^N, the long-run average cost of (π^N_p, ς^N) is

$$\begin{aligned} C^N_\varsigma(\pi^N_p) := \limsup_{T\to\infty}\frac{1}{T}\mathbb{E}^N_\varsigma\Bigg[a\left(\bar{E}^N(T) - \bar{E}^N_p(T) + \bar{R}^N(\pi^N_p, T)\right) \\ + \int_0^T g_U\left(\bar{B}^N(\pi^N_p, t)\right)dt\Bigg]. \end{aligned} \tag{6}$$

When the initial state for the fluid model for $p\lambda$ is an invariant state for $p\lambda$ associated with $b \in [0, p\lambda/\mu]$, $p\lambda - b\mu$ is the rate at which fluid abandons and $(1 - p)\lambda$ is the rate at which fluid is rejected. Since $p \in (0, 1]$ is a parameter that can be optimized over, the resulting fluid control problem is given by

$$\begin{aligned} &\min_{p\in(0,1],b\in[0,\min\{1,p\lambda/\mu\}]} a(1-p)\lambda + a(p\lambda - b\mu) + g_U(b) \\ &= \min_{b\in[0,\min\{1,\lambda/\mu\}]} a(\lambda - b\mu) + g_U(b). \end{aligned} \tag{7}$$

The solution to (7) does not depend on the admission control parameter $p \in (0, 1]$ and is identical to the solution to (4). This observation crucially relies on the abandonment cost being linear with the per unit cost equal to the per unit cost of rejection.

This gives us flexibility to propose a policy in Π^N_p for various choices of $p \in (0, 1]$. We first observe that an optimal admission control parameter must lie in $[b_*\mu/\lambda, 1]$, because otherwise the admitted arrivals would not be sufficient for servers to work at busyness level b_*. Let

$$p_* := b_*\mu/\lambda. \tag{8}$$

We next observe that if the p_*-admitted queue satisfies the non-idling condition (that is, the servers never idle when customers are waiting), the long-run average fraction of busy servers achieves b_*. The non-idling condition, together with (5)-(26) in [11] uniquely specifies $\mathbb{P}^N_y$ for each $y \in \mathbb{S}^N = \left\{y^N \in \mathbb{Y}^D : N - \langle 1, \nu^N\rangle = (N - x^N)^+ \text{ and } x^N \le \langle 1, \eta^N\rangle\right\}$ and satisfies (2). Moreover, for any compatible initial distribution, the state process that satisfies the non-idling condition is a Feller, strong Markov process (see Proposition 4.2 in [8]). Thus, for any $p \in (0, 1]$, the non-idling policy (the control policy that obeys the non-idling condition) is an admissible HL control policy for E^N_p, and thus is in Π^N_p.

Definition 4 *(The proposed policy).* For each $N \in \mathbb{N}$, let π^N_* be the non-idling policy in $\Pi^N_{p_*}$, where p_* is given by (8).

6. Asymptotic optimality of π^N_*

In this section, we state our main results concerning asymptotic optimality of $\{\pi^N_*\}_{N\in\mathbb{N}}$ under fluid scaling.

Theorem 1 *(Convergence under the proposed policy). Suppose that Assumption 1 holds and that h^s is non-increasing when $b_* = 1$. Then the sequence $\{\pi_*^N\}_{N\in\mathbb{N}}$ satisfies*

$$\lim_{N\to\infty} C^N\left(\pi_*^N\right) = a(\lambda - b_*\mu) + g_U(b_*).$$

Let $\hat{\Pi}^N := \cup_{p\in(0,1]}\Pi_p^N$ denote the enlarged policy class, and given $\hat{\pi}^N \in \hat{\Pi}^N$, let $\hat{p}^N \in (0,1]$ denote the associated admission control parameter.

Theorem 2 *(Asymptotic lower bound). Suppose that Assumption 1 holds, $\hat{\pi}^N \in \hat{\Pi}^N$ for each $N \in \mathbb{N}$ and the sequence $\{\hat{p}^N\}_{N\in\mathbb{N}}$ satisfies $\lim_{N\to\infty}\hat{p}^N = p$ for some $p \in (0,1]$. Then,*

$$\liminf_{N\to\infty} C^N(\hat{\pi}^N) \geq a(\lambda - b_*\mu) + g_U(b_*).$$

Remark 2. The condition that $\lim_{N\to\infty}\hat{p}^N = p$ for some $p \in (0,1]$ implies that $\{\hat{p}^N\bar{\lambda}^N\}_{N\in\mathbb{N}}$ satisfies $\lim_{N\to\infty}\hat{p}^N\bar{\lambda}^N = p\lambda$.

Theorem 1 establishes that the solution to the fluid control problem (4) is achieved in the limiting system, when, for each N, the N-server system operates under π_*^N in Definition 4, and in case $b_* = 1$, h^s is non-increasing. Theorem 2 establishes that the fluid control problem (4) is an asymptotic lower bound for the objective (6). As a consequence, we conclude that the proposed sequence of policies $\{\pi_*^N\}_{N\in\mathbb{N}}$ is asymptotically optimal.

The proof of Theorem 1 given in Section 8 is facilitated by the fact that, for each $N \in \mathbb{N}$, under π_*^N the p_*-admitted N-server queue is non-idling, and thus, we can appeal to results in [8,3] to establish the weak convergence of the sequence of fluid-scaled stationary distributions. The additional condition that h^s is non-increasing when $b_* = 1$, is needed for this in order to apply part (3) of Theorem 3.2 in [3] in that case. This implies that the limit is the unique invariant state with zero queue mass.

The proof of Theorem 2 in Section 8 requires first adapting one of the arguments in [8] (wherein the non-idling condition is assumed throughout) to show that a sequence of fluid-scaled stationary distributions is tight, and second arguing that the fluid control problem (7) provides an asymptotic lower bound on the cost along any convergent subsequence.

In the next section, we establish some preliminary results for stationary distributions (for both the stochastic N-server queue model and the fluid model) that help to prove Theorems 1 and 2, which may also be of independent interest. The proofs of Theorems 1 and 2 will be provided in Section 8.

7. Preliminary results

In order to prove our main results (Theorems 1 and 2), we begin by establishing two foundational results concerning stationary distributions for the N-server queue. Then, we provide a fluid limit theorem, which shows that the distributional limit points of stationary distributions are fluid model solutions almost surely under suitable asymptotic conditions. Finally, we show some properties of stationary fluid model solutions for γ. The proofs are delayed to the online appendix A.1.

Stationary distributions of the N-server queue. The following lemmas confirm the existence of a stationary distribution under any admissible HL control policy for E_p^N and $p \in (0,1]$, and derive an expression for the long-run average cost. We denote by $\triangle_\infty^N$ a stationary process associated with the process $\triangle^N$, for $\triangle^N = E^N, Y^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N, I^N$.

Lemma 1. *Let $p \in (0,1]$. For any $\pi_p^N \in \Pi_p^N$, there exists a compatible initial distribution ξ^N such that the state process Y_∞^N for (π_p^N, ξ^N) is stationary. Moreover, $\mathbb{E}_\xi^N\left[\langle 1, \eta_\infty^N(t)\rangle\right] = p\lambda^N\theta^{-1} < \infty$, for all $t \geq 0$.*

Remark 3. Proposition 1 in Section 2 follows by setting $p = 1$.

Given $p \in (0,1]$, $\pi_p^N \in \Pi_p^N$ and a compatible initial distribution ς^N, let

$$\chi^N(t) := \inf\{x \geq 0 : \left\langle 1_{[0,x]}, \eta^N(t)\right\rangle \geq Q^N(t)\} \tag{9}$$

represent the waiting time of the HL customer at time t for each $t \geq 0$. Then, for $t \geq 0$,

$$Q^N(t) = \left\langle 1_{[0,\chi^N(t)]}, \eta^N(t)\right\rangle. \tag{10}$$

The associated stationary process is denoted by χ_∞^N.

Lemma 2. *Let $p \in (0,1]$. For any $\pi_p^N \in \Pi_p^N$ and compatible initial distribution ς^N, there exists $\xi^N \in \mathcal{S}(\pi_p^N)$ such that*

$$\limsup_{T\to\infty}\mathbb{E}_\varsigma^N\left[\frac{\bar{R}^N(\pi_p^N, T)}{T}\right] = \mathbb{E}_\xi^N\left[\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle\right], \tag{11}$$

and

$$\limsup_{T\to\infty}\mathbb{E}_\varsigma^N\left[\frac{1}{T}\int_0^T g_U\left(\bar{B}^N(\pi_p^N, t)\right)dt\right] = \mathbb{E}_\xi^N\left[g_U\left(\bar{B}_\infty^N(0)\right)\right]. \tag{12}$$

If $\varsigma^N \in \mathcal{S}(\pi_p^N)$, then $\xi^N = \varsigma^N$.

In light of (10), one can interpret the right-hand side of (11) as an expected stationary reneging rate for the N-server queue.

Remark 4. For any $p \in (0,1]$, $\pi_p^N \in \Pi_p^N$ and compatible initial distribution ς^N, there exists $\xi^N \in \mathcal{S}(\pi_p^N)$ such that

$$C_\varsigma^N(\pi_p^N) = \mathbb{E}_\xi^N\left[a(1-p)\bar{\lambda}^N + a\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle + g_U\left(\bar{B}_\infty^N(0)\right)\right].$$

Proposition 2 in Section 2 follows by setting $p = 1$.

A fluid limit theorem. Here we provide asymptotic assumptions under which it is shown in [11] that fluid limit points are almost surely fluid model solutions. Such a result is crucial for the proof of Theorem 2, which will appear in Section 8.

Assumption 2. Suppose for each $N \in \mathbb{N}$, $p^N \in (0,1]$, $\pi_{p^N}^N \in \Pi_{p^N}^N$ for $E_{p^N}^N$ and ς^N is a compatible initial distribution for $\pi_{p^N}^N$. Assume that $\lim_{N\to\infty}p^N = p$ and $(\bar{X}^N(0), \bar{\nu}^N(0), \bar{\eta}^N(0)) \Rightarrow (X^0, \nu^0, \eta^0)$, as $N \to \infty$, for some random variable (X^0, ν^0, η^0) taking values in $\mathbb{X}$ such that $\sup_{N\in\mathbb{N}}\mathbb{E}_\varsigma^N\left[\langle 1, \bar{\eta}^N(0)\rangle\right] < \infty$.

Remark 5. Under Assumptions 1 and 2 and the conditions on $E_{p^N}^N$, K^N, G^s, g^s, h^s, G^r, g^r, and h^r specified in Sections 2 and 5, one can without loss of generality assume that the convergence of the initial condition in Assumption 2 is almost sure and then check that Assumptions 1, 2, 3(1), 3(3), 3(4), 4, 5(1) and 5(3) in [11] hold, i.e., Assumptions 3(2), 3(5) and 5(2) may not hold.

Lemma 3. *Suppose Assumptions 1 and 2 hold. Then, $\bar{\eta}^N \Rightarrow \eta$, as $N \to \infty$, where $\eta(0) \stackrel{d}{=} \eta^0$ and η satisfies (A.15) almost surely for $E(t) = p\lambda t$, $t \geq 0$.*

In fact, Assumptions 3(2) and 3(5) in [11] can be replaced by the condition $\sup_{N\in\mathbb{N}} \mathbb{E}_\xi^N\left[\left\langle 1, \bar{\eta}^N(0)\right\rangle\right] < \infty$ and Assumption 5(2) (η^0 has no atoms) is used to establish convergence of the scaled reneging processes to the expression in (A.8). Thus, the result in Theorem 1 in [11] continues to hold. We obtain the following slightly restated version of Theorem 1 in [11].

Lemma 4 *(Theorem 1 in [11]). Suppose that $\{(\pi^N, \varsigma^N)\}_{N\in\mathbb{N}}$ is such that Assumptions 1 and 2 hold, η^0 has no atoms, and (X, ν, η) is a distributional limit point of $\left\{\left(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N\right)\right\}_{N\in\mathbb{N}}$. Then $(X(0), \nu(0), \eta(0)) \stackrel{d}{=} (X^0, \nu^0, \eta^0)$ and (X, ν, η) is almost surely a fluid model solution for $p\lambda$.*

Properties of stationary fluid model solutions. Fix $\gamma > 0$. Here we consider the fluid model for γ with random initial states such that the resulting fluid model solution is a stationary process. Lemmas 5 and 6 below, provide properties of such solutions. The proof of Theorem 2 relies on Lemmas 5 and 6.

In what follows, we fix a fluid model solution $Z_\infty = (X_\infty, \nu_\infty, \eta_\infty)$ for γ such that Z_∞ is a stationary process. We denote the law of $Z_\infty(0)$ by ξ and the expectation operator by $\mathbb{E}_\xi$. In addition, we define a Borel probability measure η_e satisfying $d\eta_e(x) = \theta\bar{G}^r(x)dx$ for all $x \in \mathbb{R}_+$, where the subscript e is mnemonic for excess life distribution.

Lemma 5. *For all $t \geq 0$, $\eta_\infty(t) = \gamma\theta^{-1}\eta_e$. In particular, for all $t \geq 0$, $\eta_\infty(t)$ has no atoms, $x \mapsto \left\langle 1_{[0,x]}, \eta_\infty(t)\right\rangle$ is a continuous strictly increasing function on $\mathbb{R}_+$, and $\langle 1, \eta_\infty(t)\rangle = \gamma\theta^{-1}$.*

Lemma 6. *There exists $b \in [0, \min\{1, \gamma/\mu\}]$ such that for all $t \geq 0$, $\mathbb{E}_\xi[B_\infty(t)] = b$ and $\mathbb{E}_\xi\left[\left\langle 1_{[0,\chi_\infty(t)]}h^r, \eta_\infty(t)\right\rangle\right] = \gamma - b\mu$.*

8. Proofs of main results (Theorems 1 and 2)

Proof of Theorem 1. For each $N \in \mathbb{N}$, let $\xi^N \in \mathcal{S}(\pi_*^N)$ which exists by Lemma 1, and recall that $Y_\infty^N(0)$ has distribution ξ^N. Consider the sequence $\{(\bar{X}_\infty^N(0), \bar{\nu}_\infty^N(0), \bar{\eta}_\infty^N(0))\}_{N\in\mathbb{N}}$. We wish to show that $\lim_{N\to\infty} C(\pi_*^N) = a(\lambda - b_*\mu) + g_U(b_*)$. By Lemma 2, it suffices to show that,

$$\lim_{N\to\infty} \mathbb{E}_\xi^N\left[a(1-p_*)\bar{\lambda}^N + a\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle + g_U(\bar{B}_\infty^N(0))\right] = a(\lambda - b_*\mu) + g_U(b_*). \tag{13}$$

Note that $p_*\bar{\lambda}^N \to p_*\lambda$, as $N \to \infty$ (from Assumption 1). This, together with the assumptions on E^N (which $E_{p_*}^N$ inherits), G^s, g^s, h^s, G^r, g^r, and h^r given in Section 2, implies that Assumptions 3.1-3.5 in [8] hold for $\{(E_{p_*}^N, \pi_*^N, \xi^N)\}_{N\in\mathbb{N}}$. In addition, since it is assumed that h^s is non-increasing when $b_* = 1$, the result in Theorem 3.3 in [8] holds,[3] which establishes

$$(\bar{X}_\infty^N(0), \bar{\nu}_\infty^N(0), \bar{\eta}_\infty^N(0)) \Rightarrow (b_*, b_*\nu_e, p_*\lambda\theta^{-1}\eta_e), \tag{14}$$

as $N \to \infty$, where $d\nu_e(x) = \mu\bar{G}^s(x)dx$ and $d\eta_e(x) = \theta\bar{G}^r(x)dx$ for each $x \in \mathbb{R}_+$. This, together with (A.5), (A.6), and $p_* = b_*\mu/\lambda$, gives that as $N \to \infty$,

$$\bar{B}_\infty^N(0) = \left\langle 1, \bar{\nu}_\infty^N(0)\right\rangle \Rightarrow \langle 1, b_*\nu_e\rangle = b_*, \tag{15}$$

$$\bar{Q}_\infty^N(0) = \bar{X}_\infty^N(0) - \bar{B}_\infty^N(0) \Rightarrow b_* - b_* = 0. \tag{16}$$

The function g_U is continuous. Hence, by (15) and the continuous mapping theorem,

$$g_U(\bar{B}_\infty^N(0)) \Rightarrow g_U(b_*), \text{ as } N \to \infty. \tag{17}$$

Then, since g_U is bounded, (17) and the bounded convergence theorem yield that

$$\lim_{N\to\infty} \mathbb{E}_\xi^N\left[g_U\left(\bar{B}_\infty^N(0)\right)\right] = g_U(b_*). \tag{18}$$

From (10) and (16),

$$\left\langle 1_{[0,\chi_\infty^N(0)]}, \bar{\eta}_\infty^N(0)\right\rangle = \bar{Q}_\infty^N(0) \Rightarrow 0, \text{ as } N \to \infty. \tag{19}$$

Note that for each $N \in \mathbb{N}$,

$$0 \leq a\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle \leq a\|h^r\|_\infty \bar{Q}_\infty^N(0),$$

which, together with (19) and boundedness of h^r, implies

$$a\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle \Rightarrow 0, \text{ as } N \to \infty. \tag{20}$$

By Lemma 1, $\lim_{N\to\infty} p_*\bar{\lambda}^N = p_*\lambda$, and (14),

$$\lim_{N\to\infty} \mathbb{E}_\xi^N\left[\left\langle 1, \bar{\eta}_\infty^N(0)\right\rangle\right] = \lim_{N\to\infty} p_*\bar{\lambda}^N\theta^{-1} = p_*\lambda\theta^{-1} = \left\langle 1, p_*\lambda\theta^{-1}\eta_e\right\rangle.$$

This together with (14) implies that $\left\{\left\langle 1, \bar{\eta}_\infty^N(0)\right\rangle\right\}_{N\in\mathbb{N}}$ is uniformly integrable. Note that $\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle \leq \|h^r\|_\infty\left\langle 1, \bar{\eta}_\infty^N(0)\right\rangle$ for each $N \in \mathbb{N}$ and h^r is bounded. Thus, $\left\{\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N(0)\right\rangle\right\}_{N\in\mathbb{N}}$ is uniformly integrable. This together with (20) implies that

$$\lim_{N\to\infty} \mathbb{E}_\xi^N\left[a\left\langle 1_{[0,\chi_\infty^N(0)]}h^r, \bar{\eta}_\infty^N\right\rangle\right] = 0. \tag{21}$$

Finally, by Assumption 1, it follows that

$$\lim_{N\to\infty} a(1-p_*)\bar{\lambda}^N = a(1-p_*)\lambda = a(\lambda - b_*\mu). \tag{22}$$

Combining (18), (21) and (22) establishes (13), as desired. □

Proof of Theorem 2. Fix a sequence $\{\hat{\pi}^N\}_{N\in\mathbb{N}}$ satisfying the conditions of Theorem 2. For each $N \in \mathbb{N}$, let $\xi^N \in \mathcal{S}(\hat{\pi}^N)$ be such that $C^N(\hat{\pi}^N) = C_{\xi^N}^N(\hat{\pi}^N)$ which exists by Lemma 1 and the definition of $C^N(\hat{\pi}^N)$. For each $N \in \mathbb{N}$, let Y_∞^N be the state process for $(\hat{\pi}^N, \xi^N)$. It suffices to show that $\lim_{i\to\infty} C_{\xi^{N_i}}^{N_i}(\hat{\pi}^{N_i}) \geq a(\lambda - b_*\mu) + g_U(b_*)$, for any convergent subsequence of cost functions $\left\{C_{\xi^{N_i}}^{N_i}(\hat{\pi}^{N_i})\right\}_{i=1}^\infty$. Fix such a subsequence $\{N_i\}_{i=1}^\infty$. We consider the fluid scaled sequence $\{(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i})\}_{i=1}^\infty$. By Lemma 2, it suffices to show

$$\lim_{i\to\infty} \mathbb{E}_\xi^{N_i}\left[a\left(1-\hat{p}^{N_i}\right)\bar{\lambda}^{N_i} + a\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle + g_U\left(\bar{B}_\infty^{N_i}(0)\right)\right] \geq a(\lambda - b_*\mu) + g_U(b_*). \tag{23}$$

We begin by noting that the sequence $\{(\bar{X}_\infty^{N_i}(0), \bar{\nu}_\infty^{N_i}(0), \bar{\eta}_\infty^{N_i}(0))\}_{i=1}^\infty$ is tight. This follows by Theorem 6.2 in [8] and its proof since in the present setting, the result in Lemma 6.1

[3] There is a gap in the original proof of Theorem 3.3 in [8], where a stationary distribution for the fluid model is assumed to coincide with the invariant state, which is unique since G^r is strictly increasing. Under the conditions of Theorem 3.3 in [8], Theorem 3.2(1) in [3] implies that this is true when $b_* < 1$. With the added condition that h^s is non-increasing, Theorem 3.2(3) in [3] implies that this is true when $b_* = 1$. Hence, the result in Theorem 3.3 in [8] holds in the present setting. See the discussion in [3] that follows the statement of Theorem 3.2 for a detailed explanation.

in [8] holds, $\bar{K}_\infty^{N_i}(t) \le \bar{E}_\infty^{N_i}(t) + \langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle$ for all $i \in \mathbb{N}$ and $t \ge 0$, and $\bar{X}_\infty^{N_i}(0) \le 1 + \langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle$ for all $i \in \mathbb{N}$. Since $\{(\bar{X}_\infty^{N_i}(0), \bar{\nu}_\infty^{N_i}(0), \bar{\eta}_\infty^{N_i}(0))\}_{i=1}^\infty$ is tight, there exists a further subsequence $\{N_{i_k}\}_{k=1}^\infty$ such that

$$\left(\bar{X}_\infty^{N_{i_k}}(0), \bar{\nu}_\infty^{N_{i_k}}(0), \bar{\eta}_\infty^{N_{i_k}}(0)\right) \Rightarrow \left(X_\infty^0, \nu_\infty^0, \eta_\infty^0\right), \tag{24}$$

as $k \to \infty$. Without loss of generality, we can replace $\{N_{i_k}\}_{k=1}^\infty$ with $\{N_i\}_{i=1}^\infty$ by eliminating some members if necessary. In what follows, we verify that (23) holds along this subsequence. For this, we will first show that

$$\begin{aligned}\lim_{i\to\infty} \mathbb{E}_\xi^{N_i}\Big[a\left(1-\hat{p}^{N_i}\right)\bar{\lambda}^{N_i} + a\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle \\ + g_U\left(\bar{B}_\infty^{N_i}(0)\right)\Big] = a(1-p)\lambda + a\mathbb{E}_\xi\left[\left\langle 1_{[0,\chi_\infty^0]}h^r, \eta_\infty^0\right\rangle\right] \\ + \mathbb{E}_\xi\left[g_U(B_\infty^0)\right],\end{aligned} \tag{25}$$

where ξ denotes the distribution of $\left(X_\infty^0, \nu_\infty^0, \eta_\infty^0\right)$ and $\mathbb{E}_\xi$ is the expectation operator for ξ. Then we will establish process level convergence to a stationary fluid model solution for $p\lambda$ in order to apply Lemma 6 to the right-hand side of (25).

We begin by showing that η_∞^0 has no atoms and that $\left\{\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle\right\}_{i=1}^\infty$ is uniformly integrable. By Lemma 1, Assumption 1 and $\lim_{i\to\infty}\hat{p}^{N_i} = p$,

$$\lim_{i\to\infty} \mathbb{E}_\xi^{N_i}\left[\left\langle 1, \bar{\eta}_\infty^{N_i}(0)\right\rangle\right] = p\lambda\theta^{-1}, \tag{26}$$

so $\sup_{i\in\mathbb{N}} \mathbb{E}_\xi^{N_i}\left[\langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle\right] < \infty$. This together with (24) implies that Assumption 2 holds with $\{E_{\hat{p}^{N_i}}^{N_i}\}_{i=1}^\infty$, $\{\hat{\pi}^{N_i}\}_{i=1}^\infty$ and $\{\xi^{N_i}\}_{i=1}^\infty$ replacing $\{E_{p^N}^N\}_{N\in\mathbb{N}}$, $\{\pi_{p^N}^N\}_{N\in\mathbb{N}}$ and $\{\varsigma^N\}_{N\in\mathbb{N}}$ respectively. Thus, by Lemma 3, $\bar{\eta}_\infty^{N_i} \Rightarrow \eta_\infty$, as $i \to \infty$, where $\eta_\infty(0) \overset{d}{=} \eta_\infty^0$ and η_∞ satisfies (A.15) almost surely for $E_\infty(t) = p\lambda t$, $t \ge 0$. Moreover, since $\bar{\eta}_\infty^{N_i}$ is a stationary process for each $i \in \mathbb{N}$, η_∞ is a stationary process such that $\eta_\infty(t) \overset{d}{=} \eta_\infty^0$ for all $t \ge 0$. Hence, by Lemma 5, $\eta_\infty^0 = p\lambda\theta^{-1}\eta_e$, so that η_∞^0 has no atoms, $\langle 1, \eta_\infty^0\rangle = p\lambda\theta^{-1}$ and $x \mapsto \langle 1_{[0,x]}, \eta_\infty^0\rangle$ is a continuous, strictly increasing function on $\mathbb{R}_+$. Then recalling (26), $\lim_{i\to\infty}\mathbb{E}_\xi^{N_i}\left[\langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle\right] = \langle 1, \eta_\infty^0\rangle$. This together with (24) implies that $\left\{\langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle\right\}_{i=1}^\infty$ is uniformly integrable. Since h^r is bounded and $\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle \le \|h^r\|_\infty\langle 1, \bar{\eta}_\infty^{N_i}(0)\rangle$ for each $i \in \mathbb{N}$, uniform integrability of $\left\{\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle\right\}_{i=1}^\infty$ follows.

Next we show (25). For this without loss of generality, we assume that the convergence in (24) is almost sure which we abbreviate as a.s. By (24), we have

$$\begin{aligned}\lim_{i\to\infty}\bar{B}_\infty^{N_i}(0) &= \lim_{i\to\infty}\langle 1, \bar{\nu}_\infty^{N_i}(0)\rangle = \langle 1, \nu_\infty^0\rangle = B_\infty^0, \quad \text{a.s., and} \\ \lim_{i\to\infty}\bar{Q}_\infty^{N_i}(0) &= \lim_{i\to\infty}\bar{X}_\infty^{N_i}(0) - \lim_{i\to\infty}\bar{B}_\infty^{N_i}(0) = X_\infty^0 - B_\infty^0 \\ &= Q_\infty^0, \quad \text{a.s.}\end{aligned} \tag{27}$$

This implies that

$$\begin{aligned}\lim_{i\to\infty}\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}, \bar{\eta}_\infty^{N_i}(0)\right\rangle &= \lim_{i\to\infty}\bar{Q}_\infty^{N_i}(0) = Q_\infty^0 \\ &= \left\langle 1_{[0,\chi_\infty^0]}, \eta_\infty^0\right\rangle, \quad \text{a.s.}\end{aligned}$$

Thus, since $x \mapsto \langle 1_{[0,x]}, \eta_\infty^0\rangle$ is a continuous strictly increasing function on $\mathbb{R}_+$, $\lim_{i\to\infty}\chi_\infty^{N_i}(0) = \chi_\infty^0$ a.s. This together with the above display and that h^r is continuous and bounded implies

$$\lim_{i\to\infty}\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle = \left\langle 1_{[0,\chi_\infty^0]}h^r, \eta_\infty^0\right\rangle, \quad \text{a.s.} \tag{28}$$

Now, as in the proof of Theorem 1, (25) follows from $\lim_{i\to\infty}\hat{p}^{N_i}\bar{\lambda}^{N_i} = p\lambda$, (27) and g_U is bounded and continuous, and (28) and the uniform integrability of $\left\{\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle\right\}_{i=1}^\infty$.

Finally, we argue process level convergence to a stationary fluid model solution for $p\lambda$. Since Assumption 2 holds for $\{N_i\}_{i=1}^\infty$ (as noted above) and η_∞^0 has no atoms (also noted above), Lemma 4 implies that $\left(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i}\right) \Rightarrow (X_\infty, \nu_\infty, \eta_\infty)$, as $i \to \infty$, where $(X_\infty, \nu_\infty, \eta_\infty)$ is almost surely a fluid model solution for $p\lambda$ such that $(X_\infty(0), \nu_\infty(0), \eta_\infty(0)) \overset{d}{=} \left(X_\infty^0, \nu_\infty^0, \eta_\infty^0\right)$. Moreover, $(X_\infty, \nu_\infty, \eta_\infty)$ is a stationary fluid model solution for $p\lambda$ by the stationarity of $(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i})$ for each $i \in \mathbb{N}$. Then, from Lemma 6, there exists $b \in [0, \min\{1, p\lambda/\mu\}]$ such that $\mathbb{E}_\xi[B_\infty^0] = \mathbb{E}_\xi[B_\infty(0)] = b$. Since g_U is convex, Jensen's inequality further implies that

$$\mathbb{E}_\xi\left[g_U(B_\infty^0)\right] \ge g_U\left(\mathbb{E}_\xi\left[B_\infty^0\right]\right) = g_U(b).$$

This together with (25) and the second part of Lemma 6 gives

$$\begin{aligned}\lim_{i\to\infty}\mathbb{E}_\xi^{N_i}\Big[a(1-\hat{p}^{N_i})\bar{\lambda}^{N_i} + a\left\langle 1_{[0,\chi_\infty^{N_i}(0)]}h^r, \bar{\eta}_\infty^{N_i}(0)\right\rangle \\ + g_U\left(\bar{B}_\infty^{N_i}(0)\right)\Big] \ge a(1-p)\lambda + a(p\lambda - b\mu) + g_U(b) \\ = a(\lambda - b\mu) + g_U(b) \ge a(\lambda - b_*\mu) + g_U(b_*),\end{aligned} \tag{29}$$

which completes the proof that (23) holds, as desired. □

Appendix A. The fluid model for γ

We write the fluid model equations and write fluid model solutions for $\gamma > 0$ in this appendix. We refer the reader to Section 3.1 in [11] for details. Given a Polish space $\mathbb{S}$, we use $\mathbf{C}(\mathbb{S})$ to denote the set of functions having domain $\mathbb{R}_+$ and range $\mathbb{S}$ that are continuous in time.

The fluid model for γ has as an input a non-decreasing function $E(t) = \gamma t$, $t \ge 0$. We set $\mathbb{X} := \mathbb{R}_+ \times \mathbf{M}[0, H^s) \times \mathbf{M}[0, H^r)$, endowed with the product topology in a Polish space. To define the fluid model for γ, we consider $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ such that

$$\langle 1_{\{x\}}, \eta(0)\rangle = 0, \text{ for all } x \in [0, H^r), \tag{A.1}$$

and such that for each $t \ge 0$,

$$\langle 1, \nu(t)\rangle \le X(t) \le \langle 1, \nu(t)\rangle + \langle 1, \eta(t)\rangle, \tag{A.2}$$

$$\langle 1, \nu(t)\rangle \le 1, \tag{A.3}$$

$$\int_0^t \langle h^s, \nu(u)\rangle du < \infty \text{ and } \int_0^t \langle h^r, \eta(u)\rangle du < \infty. \tag{A.4}$$

Given $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ satisfying (A.1)-(A.4), we define auxiliary functions B, Q, χ, R, D, and K in $\mathbf{C}(\mathbb{R}_+)$ and I in $\mathbf{C}(\mathbb{R}_+)$ as follows: for each $t \ge 0$,

$$B(t) = \langle 1, \nu(t)\rangle, \tag{A.5}$$

$$Q(t) = X(t) - B(t), \tag{A.6}$$

$$\chi(t) = \inf\{x \ge 0 : \langle 1_{[0,x]}, \eta(t)\rangle \ge Q(t)\}, \tag{A.7}$$

$$R(t) = \int_0^t \left(\int_0^{\chi(u)} h^r(w)\eta(u)(dw)\right) du, \tag{A.8}$$

$$D(t) = \int_0^t \langle h^s, \nu(u) \rangle du, \tag{A.9}$$

$$K(t) = B(t) + D(t) - B(0), \tag{A.10}$$

$$I(t) = 1 - B(t). \tag{A.11}$$

Then B, Q, χ, R, D, K, and I are fluid analogs of the busy server, the queue length, the waiting time of the HL fluid in queue, the reneging, the departure, the entry-into-service, and the idleness processes, respectively.

Further some additional properties and equations that should be satisfied by $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ are as follows: for any continuous and bounded function f having domain $\mathbb{R}_+$, for each $t \geq 0$,

$$K \text{ is non-decreasing}, \tag{A.12}$$

$$X(t) = X(0) + E(t) - R(t) - D(t), \tag{A.13}$$

$$\langle f, \nu(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}^s(\cdot + t)}{\bar{G}^s(\cdot)}, \nu(0) \right\rangle + \int_0^t f(t-u) \bar{G}^s(t-u) dK(u), \tag{A.14}$$

$$\langle f, \eta(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}^r(\cdot + t)}{\bar{G}^r(\cdot)}, \eta(0) \right\rangle + \gamma \int_0^t f(t-u) \bar{G}^r(t-u) du. \tag{A.15}$$

Definition 5. A fluid model solution for $\gamma > 0$ is (X, ν, η) that satisfies (A.1)-(A.4), and (A.12)-(A.15).

Definition 6. A non-idling fluid model solution for $\gamma > 0$ is (X, ν, η) that satisfies Definition 5 and the following non-idling condition for each $t \geq 0$:

$$I(t) = (1 - X(t))^+. \tag{A.16}$$

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.orl.2022.04.005.

References

[1] Philipp Afeche, J. Michael Pavlin, Optimal price/lead-time menus for queues with customer choice: segmentation, pooling, and strategic delay, Manag. Sci. 62 (8) (2016) 2412–2436.
[2] Rami Atar, Amarjit Budhiraja, Paul Dupuis, Ruoyu Wu, Large deviations for the single server queue and the reneging paradox, arXiv preprint, arXiv:1903.06870, 2019.
[3] Rami Atar, Weining Kang, Haya Kaspi, Kavita Ramanan, Large-time limit of nonlinearly coupled measure-valued equations that model many-server queues with reneging, arXiv preprint, arXiv:2107.05226, 2021.
[4] Rami Atar, Yair Y. Shaki, Adam Shwartz, A blind policy for equalizing cumulative idleness, Queueing Syst. 67 (4) (2011) 275–293.
[5] Achal Bassamboo, Ramandeep Singh Randhawa, Scheduling homogeneous impatient customers, Manag. Sci. 62 (7) (2016) 2129–2147.
[6] Weining Kang, Guodong Pang, Equivalence of fluid models for Gt/GI/N+GI queues, in: Modeling, Stochastic Control, Optimization, and Applications, Springer, 2019, pp. 315–349.
[7] Weining Kang, Kavita Ramanan, Fluid limits of many-server queues with reneging, Ann. Appl. Probab. 20 (6) (2010) 2204–2260.
[8] Weining Kang, Kavita Ramanan, Asymptotic approximations for stationary distributions of many-server queues with abandonment, Ann. Appl. Probab. 22 (2) (2012) 477–521.
[9] Costis Maglaras, John Yao, Assaf Zeevi, Optimal price and delay differentiation in large-scale queueing systems, Manag. Sci. 64 (5) (2018) 2427–2444.
[10] Yu.V. Prokhorov, Convergence of random processes and limit theorems in probability theory, Theory Probab. Appl. 1 (2) (1956) 157–214.
[11] Amber L. Puha, Amy R. Ward, Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling, Math. Oper. Res. (2021).
[12] Amy R. Ward, Mor Armony, Blind fair routing in large-scale service systems with heterogeneous customers and servers, Oper. Res. 61 (1) (2013) 228–243.
[13] Ward Whitt, The impact of increased employee retention upon performance in a customer contact center, Manuf. Serv. Oper. Manag. (2006).
[14] Ward Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. 54 (1) (2006) 37–54.
[15] Dongyuan Zhan, Amy R. Ward, Staffing, routing, and payment to trade off speed and quality in large service systems, Oper. Res. 67 (2019) 1738–1751.
[16] Jiheng Zhang, Fluid models of many-server queues with abandonment, Queueing Syst. 73 (2) (2013) 147–193.