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
$^b$California State University San Marcos

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 (i.e., abandonment costs and holding 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 servers have sufficient idle time that we show is asymptotically optimal.

Keywords: $GI/GI/N+GI$, fluid control problem, asymptotically optimal idling

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 ($[1, 8, 11, 13, 2]$). However, in the restricted $M/M/N+M$ setting, the paper $[16]$ (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 generalize that observation to the $GI/GI/N+GI$ setting.

The $GI/GI/N+GI$ queue is much more difficult to analyze than the $M/M/N+M$ queue, because the state descriptor is much 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. That 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.

$^1$Research supported by NSF Grant DMS-1712974
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 operational costs (i.e., abandonment costs and 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 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), 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 the service discipline does not overwork servers, which could lead to increased employee retention, which can have performance benefits (discussed in [14]). 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., [3], [13]). Interestingly, the motivation to idle servers can also arise from the customer side, in order to exploit heterogeneous customers preferences so as to maximize revenue (as in [3], [2], [10]).

Notation

Some notation used in this paper are summarized as follows. We denote the set of integers endowed with the discrete topology by \(\mathbb{Z}\), the set of non-negative integers by \(\mathbb{N}\), the set of positive integers by \(\mathbb{N}^+\), the set of real numbers by \(\mathbb{R}\), and the set of non-negative real numbers by \(\mathbb{R}^+\). For \(H \in (0, \infty]\), let \(M(0, H)\) denote the set of finite, non-negative Borel measures on \([0, H)\) endowed with the topology of weak convergence. For a given \(\eta \in M(0, H)\) and a Borel measurable function \(f : [0, H) \to \mathbb{R}^+\) that is integrable with respect to \(\eta\), we write \(\langle f, \eta \rangle = \int_{[0, H)} f(x) \eta(dx)\). In particular, we let \(0 \in M(0, H)\) be the measure such that \(\langle f, 0 \rangle = 0\) for all Borel measurable functions \(f : [0, H) \to \mathbb{R}^+\). Given \(x \in [0, H)\), \(\delta_x\) denotes the Dirac measure in \(M(0, H)\) such that for all Borel measurable functions \(f : [0, H) \to \mathbb{R}^+\), \(\langle f, \delta_x \rangle = f(x)\). Then let \(M_D(0, H)\) denote the subset of \(M(0, H)\) consisting of the measures \(\eta \in M(0, H)\) such that either \(\eta = 0\) or \(\eta\) 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 \((x_1, \ldots, x_n) \in [0, H)^n\).
2. The Model

We consider 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. Customers arrive according to a renewal process $E$ with rate $\lambda \in \mathbb{R}_+$, each with a service time sampled from a cumulative distribution function (c.d.f.) $G^s$ having finite mean $1/\mu \in (0, \infty)$, and a patience time sampled from a c.d.f. $G^r$ having finite mean $1/\theta \in (0, \infty)$. We assume the inter-arrival distribution associated with the renewal arrival process is absolutely continuous with respect to Lebesgue measure. We assume $G^s$ and $G^r$ are absolutely continuous with density functions $g^s$ and $g^r$ that have right edges of support $H^s := \sup\{x \in [0, \infty) : 1 - G^s(x) > 0\}$ and $H^r := \sup\{x \in [0, \infty) : 1 - G^r(x) > 0\}$, respectively. Let $h^s(t) := g^s(t)/(1 - G^s(t))$ for $t \in [0, H^s)$ and $h^r(t) := g^r(t)/(1 - G^r(t))$ for $t \in [0, H^r)$ be the associated hazard rate functions. We assume that there exists $0 \leq L^s < H^s$ such that $h^s$ is either bounded or lower-semicontinuous on $(L^s, H^s)$. Moreover, we assume $G^r$ is strictly increasing with inverse function $(G^r)^{-1}$, where by convention $(G^r)^{-1}(1) = H^r$, and $h^r$ is bounded. Boundness of $h^r$ implies that $H^r = \infty$. 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., $\lambda^N$ 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 [12], the state descriptor for the $N$-server queue is denoted by $y^N = (\alpha^N, x^N, \nu^N, \eta^N) \in \mathcal{Y}_D$, where $\mathcal{Y}_D = \mathbb{R}_+ \times \mathbb{Z}_+ \times M_D[0, H^s) \times M_D[0, H^r)$. In particular, $\alpha^N \in \mathbb{R}_+$ 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, $\nu^N \in M_D[0, H^s)$ is a measure that has a unit mass at the age-in-service (amount of service received) of each customer currently in service, and $\eta^N \in M_D[0, H^r)$ 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.)

As in [12], a state process for the $N$-server queue is a $\mathcal{Y}_D$ 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. See equations (5)-(26) in [12]. In HL service, 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 state process $Y^N$ is uniquely determined by an HL control policy which determines the number $K^N(t)$ of customers to enter service by time $t$, for each $t > 0$; see Section 2.4.1 and Definition 1 in [12] for further details. The initial condition $Y^N(0)$ for a state process $Y^N$ is assumed to be independent of the stochastic primitive inputs (see Section 2.1 in [12] for details). We denote the
distribution of $Y^N(0)$ by $c^N$. Then $\mathbb{P}^N_\varsigma$ denotes the distribution of the process $Y^N$ and $\mathbb{E}^N_\varsigma$ denotes the expectation operator under $\mathbb{P}^N_\varsigma$. It is assumed that $\mathbb{E}^N_\varsigma[X^N(0) + \langle 1, \eta^N(0) \rangle] < \infty$. See Section 2.5 and Definition 2 in [12] for further details.

As in [12], we restrict attention to HL control policies that do not use information about the future to determine when a customer will enter service. This amounts to requiring $K^N$ to be adapted to a suitable filtration as in Definition 3 in [12]. 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.** An admissible HL control policy (i) satisfies Definition 3 in [12], and (ii) is such that $Y^N$ is a Feller Markov process with respect to the filtration used in Definition 3 in [12]. Let $\Pi^N$ denote the set of admissible HL control policies.

In summary, the conditions of Definition 1 guarantee that the system dynamics take place on a subspace of $\mathbb{Y}_D$ such that the dynamics of $Y^N$ are consistent with that of a HL $N$-server queue, the entry-into-service process $K^N$ is adapted and non-anticipating, and, as we will show in Lemma 2 in Section 6, a stationary distribution exists.

### 3. The Scheduling Problem

Each customer abandonment incurs 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. An admissible HL control policy determines how many customers enter service by time $t > 0$. 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. Our objective is to find an admissible control policy $\pi^N \in \Pi^N$ that minimizes the long-run average cost,

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

where the process $R^N(\pi^N, t)$ tracks the cumulative number of abandonments by time $t$ under $\pi^N$, and the process $B^N(\pi^N, t) \leq N$ tracks the number of busy servers at time $t$ under $\pi^N$. More specifically, we want to determine $\pi_{opt}^N$ and $C^N(\pi_{opt}^N)$ such that

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

(1)

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 $\pi_{opt}^N$ exactly. Thus, we leverage a fluid model because of its analytical tractability and justified approximation properties.
4. The Fluid Control Problem

In this section, we formulate a fluid control problem based on the fluid model and fluid model solutions as defined in [12]. Fluid model solutions are functions of time that take values in the set $\mathbb{X} = \mathbb{R}_+ \times \mathbb{M}(0, H^s) \times \mathbb{M}(0, H^r)$ endowed with the product topology. For this, $(x, \nu, \eta) \in \mathbb{X}$ is a fluid analog of the state descriptor for the stochastic system with $x$ corresponding to the total mass in system, $\langle 1_{[0, x]}, \nu \rangle$ corresponding to the total mass in service with age-in-service less or equal to $x$ for each $x \in \mathbb{R}_+$, and $\langle 1_{[0, x]}, \eta \rangle$ corresponding to the total mass potentially in system of age less or equal to $x$ for each $x \in \mathbb{R}_+$. They satisfy a set of conditions determined by an arrival function $E$, which is a non-decreasing, continuous function of time such that $E(0) = 0$. These conditions are referred to as the fluid model for arrival function $E$. We summarize the fluid model and definition of fluid model solutions for an arrival function $E$ in Appendix A for the reader’s convenience.

Recall the processes $X^N, \nu^N, \eta^N, E^N$ and $K^N$ defined in Section 2, and the processes $B^N$ and $R^N$ defined in Section 3 also define the process $Q^N = X^N - B^N$ as the queue length, the process $D^N$ as the cumulative number of departures, and the process $I^N = N - B^N$ as the number of idle servers. For each $N \in \mathbb{N}$, we define the fluid scaling for the $N$-server system as follows. Let $\tilde{\alpha}^N = \alpha^N$; also for $H^N = X^N, \nu^N, \eta^N, E^N, K^N, B^N, Q^N, R^N, D^N, I^N$, let $\tilde{H}^N = H^N/N$. Then, the fluid-scaled state processes for the $N$-server system is $\tilde{Y}^N = (\tilde{\alpha}^N, \tilde{X}^N, \tilde{\nu}^N, \tilde{\eta}^N)$.

Assumption 1. Let $\lambda \in (0, \infty)$, $\pi^N \in \Pi^N$, for each $N \in \mathbb{N}$. Assume that $\lim_{N \to \infty} \lambda^N/N = \lambda$, and $\{K^N(\pi^N, \cdot)\}_{N \in \mathbb{N}}$ is C-tight. Moreover, assume that $(X^N(0), \tilde{\nu}^N(0), \tilde{\eta}^N(0)) \to (X^0, \tilde{\nu}^0, \tilde{\eta}^0)$ almost surely, as $N \to \infty$, for some random variable $(X^0, \tilde{\nu}^0, \tilde{\eta}^0)$, taking values in $\mathbb{X}$ such that $\tilde{\eta}^0$ has no atoms. Finally, assume that $\lim_{N \to \infty} \mathbb{E}[\langle 1, \tilde{\eta}^N(0) \rangle] = \mathbb{E}[\langle 1, \tilde{\eta}^0 \rangle] < \infty$ and $\lim_{N \to \infty} \mathbb{E}[X^N(0)] = \mathbb{E}[X(0)] < \infty$.

Remark 1. Under Assumption 1 and the conditions on $G^s$, $g^s$, $h^s$, $G^r$, $g^r$, and $h^r$ specified in Section 4, one can check that Assumptions 1-5 in [12] are satisfied.

Henceforth, $\lambda$ satisfying the conditions in Assumption 1 is fixed, and we refer to it as the limiting fluid-scaled arrival rate, or simply the fluid arrival rate for short. Define $\Lambda(t) = \lambda t$, for $t \geq 0$. We remark that $\{\pi^N\}_{N \in \mathbb{N}}$ is not necessarily fixed.

Lemma 1 (Theorem 1 in [12]). Suppose that $\{\pi^N\}_{N \in \mathbb{N}}$ is such that Assumption 1 holds and $(X, \nu, \eta)$ is a distributional limit point of $\{(X^N, \tilde{\nu}^N, \tilde{\eta}^N)\}_{N \in \mathbb{N}}$. Then $(X(0), \nu(0), \eta(0)) \overset{d}{=} (X^0, \tilde{\nu}^0, \tilde{\eta}^0)$ and $(X, \nu, \eta)$ is almost surely a fluid model solution for $\Lambda$.

Since the arrival function $\Lambda$ is uniquely determined by the arrival rate $\lambda$, we adopt the shorthand terminology “fluid model for $\lambda$” in place of “fluid
model for $\Lambda$. A similar convention applies to fluid model solutions. The invariant states for the fluid model for $\lambda$ are fixed points of the fluid model for $\lambda$ defined in Appendix A. More specifically, an invariant state for $\lambda$ is an element $(\mathcal{X}_\infty, \nu_\infty, \eta_\infty)$ of $\mathcal{X}$ such that a function of time identically equal to $(\mathcal{X}_\infty, \nu_\infty, \eta_\infty)$ is a fluid model solution for $\lambda$. From Proposition 1 in [12], the invariant states for $\lambda$ are 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 $b \in [0, \min\{1, \lambda/\mu\}]$, where we recall that $\mu$ is the reciprocal of the mean of $G^\ast$. Then, when the initial condition for a fluid model solution for $\lambda$ is an invariant state for $\lambda$, it turns out that the departure rate of the fluid from the system is $b\mu$ and so, by conservation of mass, $\lambda - b\mu$ must be the rate at which fluid abandons. Thus, we expect to obtain the following fluid control problem for $\lambda$ when letting $N \to \infty$ in problem (1).

**Definition 2 (The Fluid Control Problem).** The fluid control problem for $\lambda$ is given by

$$
\min_{b \in [0, \min\{1, \lambda/\mu\}]} a(\lambda - b\mu) + g_U(b). \tag{2}
$$

We denote the solution to (2) by $b_\ast$ (which is unique and guaranteed to exist because (2) optimizes a convex function over a compact set).

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

The optimization problem (2) is not sensitive to the patience time distribution, because the fluid-scaled number of reneging customers does not depend on the patience time distribution. A related insensitivity result for a single server queue in the large deviations regime is found in [3].

The solution to (2) motivates a scheduling policy that we expect to have good performance with respect to the original objective (1) when the number of servers $N$ is large. When $b_\ast = \min\{1, \lambda/\mu\}$, we expect a non-idling control policy to be optimal for (1). Otherwise, when $b_\ast < \min\{1, \lambda/\mu\}$, we expect that an optimal control policy will have the servers idle while customers are waiting, such that the long-run average amount of time that the servers should be busy is close to $b_\ast$. In this case, suppose that for each $N \in \mathbb{N}$, consider the admissible 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, $(b_\ast\mu)^{-1}$, and the expected time between service completions when the server is always busy, $\mu^{-1}$; that is, for $(b_\ast\mu)^{-1} - \mu^{-1} = (1 - b_\ast)(b_\ast\mu)^{-1}$ time units. Such a policy seems quite reasonable, and should be asymptotically optimal. However, establishing that

$$
\lim_{N \to \infty} \lim_{t \to \infty} E\left[\frac{R^N(\tilde{\pi}^N, t)}{N}\right] = \lambda - b_\ast\mu \quad \text{and} \quad \lim_{N \to \infty} \lim_{t \to \infty} E\left[g_U\left(\frac{B^N(\tilde{\pi}^N, t)}{N}\right)\right] = g_U(b_\ast) \tag{3}
$$

6
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 [9]). 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 (3) holds. This is described more fully in the next section.

5. The Proposed Policy $\pi^*_N$

The solution $b_*$ < min{1, $\lambda/\mu$} to (2) 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 (1). 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 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.

Definition 3. For any admission control parameter $p \in (0, 1]$, an admissible HL control policy for $E^N_p$ satisfies Definition 1 with $E^N$ replaced by $E^N_p$. For each $p \in (0, 1]$, let $\Pi^N_p$ denote the set of admissible HL control policies for $E^N_p$.

When $p = 1$, Definition 3 reduces to Definition 1 i.e., $\Pi^N = \Pi^N_1$.

Given $p \in (0, 1]$ and $\pi^N_p \in \Pi^N_p$, the long-run average cost under $\pi^N_p$ is

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

(4)

Here, $R^N(\pi^N_p, t)$ denotes the cumulative number of customer abandonments from the $p$-admitted queue by time $t$, and $B^N(\pi^N_p, t) \leq N$ denotes the number of busy servers in the $p$-admitted queue at time $t$. When the initial condition for the fluid model for $p\lambda$ is an invariant state 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. Provided that $p \in (0, 1]$ is a parameter that can be optimized over, the resulting fluid control problem is given by

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

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

(5)
Note that the solution to (5) does not rely on the admission control parameter $p \in (0, 1]$ and is identical to the solution to (2).

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

$$p_* := b_\mu/\lambda. \quad (6)$$

We next observe that if $p = p_*$, then when the $p_*$-admitted queue is non-idling, the long-run average fraction of busy servers achieves $b_\mu$ (and non-idling queues are easier to analyze than queues that do not obey this condition).

**Definition 4 (The Proposed Policy).** For each $N \in \mathbb{N}$, let $\pi_*^N$ be the non-idling policy in $\Pi_{p_*}^N$, where $p_*$ is given by (6).

Our main results stated in Section 6 establish asymptotic optimality of $\{\pi_*^N\}_{N \in \mathbb{N}}$ under fluid scaling.

6. Asymptotic Optimality of $\pi_*^N$

In this section, we state our main results concerning asymptotic optimality of $\{\pi_*^N\}_{N \in \mathbb{N}}$. For this, let $\hat{\Pi}^N := \cup_{p \in (0, 1]} \Pi_p^N$ be 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 1 (Asymptotic Lower Bound).** Suppose that $\{\hat{\pi}^N\}_{N \in \mathbb{N}}$ is a sequence of admissible control policies such that $\hat{\pi}^N \in \hat{\Pi}^N$, for each $N \in \mathbb{N}$ and the sequence of admission control parameters $\{\hat{p}^N\}_{N \in \mathbb{N}}$ satisfies $\lim_{N \to \infty} \hat{p}^N = p$ for some $p \in (0, 1]$, and $\{\hat{K}^N(\hat{\pi}^N, \cdot)\}_{N \in \mathbb{N}}$ is $C$-tight. Then, Assumption 1 holds and

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

**Remark 2.** The thinned arrival process $E_{\hat{p}^N}$ is a renewal process for each $N \in \mathbb{N}$, because the admitted arrivals remain i.i.d., and the condition that $\lim_{N \to \infty} \hat{p}^N = p$ for some $p \in (0, 1]$ implies that $\{E_{\hat{p}^N}\}_{N \in \mathbb{N}}$ satisfies Assumption 7 with $p\lambda$ in place of $\lambda$; that is, $\lim_{N \to \infty} \hat{p}^N \lambda^N / N = p\lambda$. This together with $C$-tightness of $\{\hat{K}^N(\hat{\pi}^N, \cdot)\}_{N \in \mathbb{N}}$ implies that Assumption 7 holds for $\{\hat{\pi}^N\}_{N \in \mathbb{N}}$.

**Theorem 2 (Convergence under the Proposed Policy).** The sequence $\{\pi_*^N\}_{N \in \mathbb{N}}$ satisfies

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

Theorem 1 establishes that the fluid control problem (2) is an asymptotic lower bound on the objective (4), and Theorem 2 establishes that the solution to the fluid control problem (2) is achieved in the limiting system, when, for each
N, the N-server system operates under \( \pi^N \) in Definition 4. As a consequence, we can conclude that the proposed policy \( \pi^N \) is asymptotically optimal.

The proof of Theorem 1 requires first adapting one of the arguments in [9] to show that a sequence of fluid-scaled stationary distributions is tight, and second arguing that the fluid control problem (5) provides an asymptotic lower bound on the cost incurred \( C^N(\hat{\pi}^N) \) on any convergent subsequence. The proof of Theorem 2 is facilitated by the fact that the admitted queue is non-idling, and, therefore, we can appeal to a result in [9] to establish the weak convergence of a sequence of fluid-scaled stationary distributions. This implies that limit points are non-idling fluid model solutions for \( p\lambda \). These are fluid model solutions for \( p\lambda \) that also satisfy the non-idling equation; see Definition 7 in Appendix A. We will start in Section 6.1 with the proof of Theorem 2 which invokes an existing fluid limit theorem and thus is relatively easier, then the proof of Theorem 1, given in Section 6.2.

Let \( p \in (0, 1] \) and \( \pi^N_p \in \Pi^N_p \), when a stationary measure exists, the long-run average cost under \( \pi^N_p \) in (4) can be equivalently written as an expectation with respect to that stationary measure. The following lemmas, whose proofs are delayed to Section 7, confirm the existence of a stationary measure under any admissible HL control policy, and derive an expression for the long-run average cost.

**Lemma 2.** Let \( p \in (0, 1] \). For any admissible HL control policy \( \pi^N_p \in \Pi^N_p \), there exists a stationary measure \( \xi^N \).

Let

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

represent the waiting time of the HL customer at time \( t \) for \( t \geq 0 \). We denote by \( H^\infty \) the random variable that has the stationary distribution \( \xi^N \) associated with the process \( H^N \), for \( H^N = X^N, \nu^N, \eta^N, E^N, K^N, B^N, Q^N, R^N, D^N, I^N, \chi^N \).

**Lemma 3.** Let \( p \in (0, 1] \). For any admissible HL control policy \( \pi^N_p \in \Pi^N_p \), when the system is initialized according to a stationary measure \( \xi^N \),

\[
C^N(\pi^N_p) = \mathbb{E}_\xi^N \left[ a \cdot \frac{1}{N}(1-p)\lambda^N + a \cdot \frac{1}{N} \langle 1_{[0,\chi^N)}, h^*, \eta^N \rangle + g_U \left( \frac{B^N}{N} \right) \right].
\]

6.1. Proof of Theorem 2

Under policy \( \pi^N \), the arrivals join the non-idling \( p_* \)-admitted queue with probability \( p_* = b_*\mu/\lambda \) (from Definition 4). Thus, the non-idling equation, \( I^N(t) = (N - X^N(t))^+, t \geq 0 \) (the stochastic version of (A.16) in Appendix A), is satisfied under \( \pi^N \). From Lemma 3.1 in [9], since \( G^* \) is strictly increasing, the non-idling fluid model for \( p_* \lambda \) has a unique invariant state, denoted as \( (X^\infty, \nu^\infty, \eta^\infty) \) as specified in Proposition 1 in [12]. Let \( \{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N \in \mathbb{N}} \) be a sequence of scaled stationary distributions for \( \bar{E}^N_{\bar{p} N} \), where \( (\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N) \) is
a stationary distribution for the N-server system. Provided that we can show that all of the conditions of Theorem 3.3 in [9] are satisfied, it follows that (\(\bar{X}_\infty^N, \bar{\nu}_\infty^N, \bar{\eta}_\infty^N\)) weakly converges to the unique invariant state (\(X_\infty, \nu_\infty, \eta_\infty\)) under \(\pi_\infty^N\), i.e.,

\[
(\bar{X}_\infty^N, \bar{\nu}_\infty^N, \bar{\eta}_\infty^N) \Rightarrow (X_\infty, \nu_\infty, \eta_\infty), \quad \text{as } N \to \infty,
\]

and thus, from (A.5) in Appendix A, the fluid scaled busy server number satisfies

\[
\bar{B}_\infty^N = \langle 1, \bar{\nu}_\infty^N \rangle \Rightarrow \langle 1, \nu_\infty \rangle = b_*, \quad \text{as } N \to \infty. \tag{8}
\]

As noted in Remark 2, Assumption 1 holds for \(\{E^N_{\hat{p}_N}\}^N\in\mathbb{N}\). Then, under Assumption 1 and the conditions on \(h^s\) and \(h^r\) specified in Section 2, one can easily check that the conditions of Theorem 3.3 in [9] are satisfied.

Recall Definition 4, under \(\pi_\infty^*, p_* = b_* \mu / \lambda\), and so

\[
a \cdot \frac{1}{N} (1 - p_*) \lambda^N \to a (1 - p_*) \lambda = a (\lambda - b_* \mu), \quad \text{as } N \to \infty.
\]

From (8) and the continuous mapping theorem (\(g_U\) is a continuous function),

\[
g_U(\bar{B}_\infty^N) \Rightarrow g_U(b_*), \quad \text{as } N \to \infty.
\]

Since \(\bar{B}_\infty^N \in [0, 1]\), and \(g_U\) is continuous, \(g_U(\bar{B}_\infty^N)\) is bounded. Thus, the bounded convergence theorem establishes

\[
E^N_{\xi} \left[ g_U \left( \bar{B}_\infty^N \right) \right] \to g_U(b_*), \quad \text{as } N \to \infty.
\]

Hence, if we can show

\[
E^N_{\xi} \left[ a \cdot \langle 1, \chi_N \rangle h^r, \bar{\eta}_\infty^N \right] \to 0, \quad \text{as } N \to \infty, \tag{9}
\]

where \(\xi^N\) is a stationary distribution, then from Lemma 3, we can conclude

\[
C^N(\pi_\infty^N) \to a (\lambda - b_* \mu) - g_U(b_*), \quad \text{as } N \to \infty,
\]

which completes the proof.

To see (9), first note that \(\bar{\eta}_\infty^N \Rightarrow \eta_\infty\) with continuous \(\eta_\infty\) (by Theorem 3.3, Lemma 3.1, and the definition in (3.4) in [9]), and \(\chi_N \Rightarrow \chi_\infty\) with continuous \(\chi_\infty\) (by (I), (A.7), \(Q_N^r \Rightarrow Q_\infty, \bar{\eta}_\infty^N \Rightarrow \eta_\infty\), and continuity of \(Q_\infty\) and \(\eta_\infty\)), then the continuous mapping theorem implies

\[
a \cdot \langle 1, \chi_\infty \rangle h^r, \eta_\infty^N \Rightarrow a \cdot \langle 1, \chi_\infty \rangle h^r, \eta_\infty, \quad \text{as } N \to \infty.
\]
Since \( \{\langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle \}_{N \in \mathbb{N}} \) is a bounded sequence (because \( h^r \) is bounded), the bounded convergence theorem implies
\[
\mathbb{E}_\xi^N [a \cdot \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle] \to \mathbb{E}_\xi [a \cdot \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle], \text{ as } N \to \infty. \tag{10}
\]

From Proposition 1 in [12],
\[
Q_{\infty} = p_\ast \lambda \int_0^{(G_{\ast})^{-1} (1-b_\ast \mu/p_\ast \lambda)} (1-G_{\ast}(x)) dx,
\]
where \( p_\ast \lambda \) is the limiting fluid-scaled arrival rate for the admitted arrival process under the proposed policy. Since \( p_\ast \lambda = (b_\ast \mu/\lambda) \cdot \lambda = b_\ast \mu \), the above equation implies \( Q_{\infty} = 0 \). Since also the definition of \( \chi_{\infty} \) in (7) implies \( \bar{Q}_{\infty} = \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle \), so that taking the limit as \( N \to \infty \) yields \( Q_{\infty} = \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle \).

Thus, substituting for \( Q_{\infty} = 0 \) into the above display yields
\[
\langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle = 0.
\]

Hence, from (10),
\[
\mathbb{E}_\xi^N [a \cdot \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle] \to \mathbb{E}_\xi [a \cdot \langle 1_{[0,\chi_{\infty}]} h^r, \eta_{\infty} \rangle] = 0, \text{ as } N \to \infty,
\]
which establishes (9).

6.2. Proof of Theorem 1

Fix a sequence \( \{\hat{\pi}_N\}_{N \in \mathbb{N}} \) such that \( \hat{\pi}_N \in \hat{\Pi}^N \) and \( \lim_{N \to \infty} \bar{p}_N = p \) for some \( p \in (0, 1) \), and for each \( N \), let \( (\hat{X}_{\infty}^N, \bar{\nu}_{\infty}^N, \bar{\eta}_{\infty}^N) \) be an associated stationary process (that is, initialized under a stationary measure \( \xi_{\infty} \), which exists by Lemma 2).

Following the arguments in Section 6 of [9], it suffices to show the following two statements,

(i) \( \{(\hat{X}_{\infty}^N, \bar{\nu}_{\infty}^N, \bar{\eta}_{\infty}^N)\}_{N \in \mathbb{N}} \) is tight;

(ii) on every convergent subsequence, \( a(\lambda - b_\ast \mu) + g_U(b_\ast) \) is a lower bound on the limiting cost.

We begin by observing that the proof of Theorem 6.2 in [9] does not require the non-idling condition \( I_N(t) = (N - X_N(t))^+ \), \( t \geq 0 \) (the stochastic version of (A.16) in Appendix A), and so the same arguments establish (i) and that \( \sup_{N \in \mathbb{N}} \mathbb{E}[(1, \eta_{\infty}^N)] < \infty \). Hence, Assumption 1 holds and so, from Lemma 1 for any convergent subsequence \( (\hat{X}_{\infty}^N, \bar{\nu}_{\infty}^N, \bar{\eta}_{\infty}^N) \),
\[
(\hat{X}_{\infty}^N, \bar{\nu}_{\infty}^N, \bar{\eta}_{\infty}^N) \Rightarrow (X_{\infty}, \nu_{\infty}, \eta_{\infty}), \text{ as } N_i \to \infty,
\]
with \((X_\infty, \nu_\infty, \eta_\infty)\) being almost surely a fluid model solution for \(p\lambda\). Moreover, 
\((X_\infty, \nu_\infty, \eta_\infty)\) is stationary by the stationarity for each \(N \in \mathbb{N}\). Thus, as in the proof of Theorem 2

\[
\mathbb{E}^N_{\xi}\left[ a \cdot \frac{1}{N_i} \left(1 - \hat{p}^N_i\right) \lambda^N_i + a \cdot \frac{1}{N_i} \left(\mathbb{1}_{[0,\infty)} h^r, \eta^N_i\right) + g_U \left(\frac{B^N_i}{N_i}\right)\right] 
\to \mathbb{E}_\xi \left[ a \cdot (1 - p) \lambda + a \cdot \left(\mathbb{1}_{[0,\infty)} h^r, \eta_\infty\right) + g_U(B_\infty)\right], \text{ as } N_i \to \infty,
\]

by noting that \(\lim_{N \to \infty} \frac{N_i}{N} = \lambda\) and \(\lim_{N \to \infty} \frac{B^N_i}{N} = p\lambda\).

Suppose we can establish the following claim.

**Claim.** \(\mathbb{E}_\xi \left[ a \cdot (1 - p) \lambda + a \cdot \left(\mathbb{1}_{[0,\infty)} h^r, \eta_\infty\right)\right] = a(\lambda - b\mu)\) and \(\mathbb{E}_\xi [B_\infty] = b\),

for some \(b \in [0, \min\{1, p\lambda/\mu\}]\).

Then, since \(g_U\) is convex, Jensen’s inequality implies that

\[
\mathbb{E}_\xi [g_U(B_\infty)] \geq g_U \left(\mathbb{E}_\xi [B_\infty]\right) = g_U(b),
\]

and so

\[
\lim_{N_i \to \infty} \mathbb{E}^N_{\xi}\left[ a \cdot \frac{1}{N_i} (1 - \hat{p}^N_i) \lambda^N_i + a \cdot \frac{1}{N_i} \left(\mathbb{1}_{[0,\infty)} h^r, \eta^N_i\right) + g_U \left(\frac{B^N_i}{N_i}\right)\right] 
\geq a(\lambda - b\mu) + g_U(b) \geq a(\lambda - b_\ast\mu) + g_U(b_\ast),
\]

which completes the proof.

**Proof of Claim.** We have

\[
\mathbb{E}_\xi[K_\infty(t)] = (1) \mathbb{E}_\xi[D_\infty(t)] = (2) \mathbb{E}_\xi\left[ \int_0^t \langle h^r, \nu_\infty(u) \rangle \, du \right] = (3) \mathbb{E}_\xi[\langle h^r, \nu_\infty(0) \rangle] \cdot t,
\]

where (1) follows from (A.10) in Appendix A and the fact that \(\mathbb{E}_\xi[B_\infty(t)] = \mathbb{E}_\xi[B_\infty(0)]\) because \(B_\infty\) is a stationary process, (2) follows from (A.9) in Appendix A and (3) follows from stationarity.

From (A.14) in Appendix A

\[
(1, \nu_\infty(t)) = \int_0^\infty \frac{1 - G^\ast(x + t)}{1 - G^\ast(x)} \nu_\infty(0)(dx) + \int_0^t (1 - G^\ast(t - u))\,dK_\infty(u).
\]

Taking expectation on both sides of the above display and noting that \((1, \nu_\infty(t)) = B_\infty(t)\) from (A.5) in Appendix A yield

\[
\mathbb{E}_\xi[B_\infty(t)] = \mathbb{E}_\xi\left[ \int_0^\infty \frac{1 - G^\ast(x + t)}{1 - G^\ast(x)} \nu_\infty(0)(dx) \right] + \mathbb{E}_\xi\left[ \int_0^t (1 - G^\ast(t - u))\,dK_\infty(u) \right] 
\to 0 + \mathbb{E}_\xi[\langle h^r, \nu_\infty(0) \rangle] \int_0^t (1 - G^\ast(t - u))\,du, \text{ as } t \to \infty,
\]
where the first integral converges to zero by dominated convergence theorem. Hence, for all \( t \geq 0 \),
\[
E\xi [B_\infty (t)] = E\xi [(h^*, \nu_\infty (0))] \int_0^\infty (1 - G^*(u))du = E\xi [(h^*, \nu_\infty (0))] / \mu = E\xi [D_\infty (t)] / (t\mu).
\]
Furthermore, since \( X(t) = X(0) + E^p(t) - R(t) - D(t) \) (from \( \text{Appendix A} \)) and \( E\xi [X_\infty (t)] = E\xi [X_\infty (0)] \) due to stationarity, \( E\xi [D_\infty (t)] = E\xi [E^p(t)] - E\xi [R_\infty (t)] \leq E\xi [E^p(t)] = p\lambda t \) for all \( t \geq 0 \), the above display implies that \( E\xi [B_\infty (t)] \leq p\lambda / \mu \). Hence, there exists \( b \in [0, \min\{1, p\lambda / \mu\}] \) such that \( E\xi [D_\infty (t)] = b \) for all \( t \geq 0 \), and hence from the above equation,
\[
E\xi [B_\infty (t)] = b \mu t,
\]
which implies
\[
E\xi [R_\infty (t)] = E\xi [E^p(t)] - E\xi [D_\infty (t)] = (p\lambda - b\mu) \cdot t.
\]
From Lemma 3 in \([12]\) and stationarity,
\[
E\xi [R_\infty (t)] = E\xi \left[ \int_0^t \langle 1_{[0,\chi_\infty (u)]} h^r, \eta_\infty \rangle du \right] = E\xi \left[ \langle 1_{[0,\chi_\infty ]} h^r, \eta_\infty \rangle \right] \cdot t.
\]
The above two displays imply that \( E\xi \left[ \langle 1_{[0,\chi_\infty ]} h^r, \eta_\infty \rangle \right] = (p\lambda - b\mu) \). Hence, it follows that
\[
E\xi \left[ a \cdot (1 - p)\lambda + a \cdot \langle 1_{[0,\chi_\infty ]} h^r, \eta_\infty \rangle \right] = a \cdot [(1 - p)\lambda + (p\lambda - b\mu)] = a(\lambda - b\mu).
\]

7. Proofs from Section 5

Given \( p \in (0, 1] \) and \( \pi^N_p \in \Pi^N_p \), let \( Y^N_p \) be the state process for \( \pi^N_p \). Let \( A \) be a Borel set in the subspace of state space \( Y^N_D \). Define \( L^N_0 (A) := \mathbb{P}^N (Y^N_p (0) \in A) \) and for \( t > 0 \), define
\[
L^N_t (A) := \frac{1}{t} \int_0^t \mathbb{P}^N (Y^N_p (s) \in A) ds.
\]

7.1. Proof of Lemma 3

From Lemma 4.8 in \([9]\), the family of probability measures \( \{L^N_t\}_{t \geq 0} \) is tight. Furthermore, provided that \( \{Y^N_p (t)\}_{t \geq 0} \) is a Feller Markov process (from Definition \([3]\), the Krylov-Bogoliubov theorem (see Corollary 3.1.2 in \([6]\)) implies that any limit point \( \xi^N \) of \( \{L^N_t\}_{t \geq 0} \) is a stationary distribution, as shown in Theorem 4.9 in \([9]\). Thus, for each \( N \in \mathbb{N} \), there exists a stationary distribution under any admissible HL control policy.
7.2. Proof of Lemma 3

Define a stochastic process \( \{ U(t) : t \geq 0 \} \), where \( U(t) := g^N_b \left( B^N \left( \pi^N_p, t/N \right) / N \right) \), for \( t \geq 0 \). From the proof of Lemma 2, for each \( N \in \mathbb{N} \), any limit point \( \xi^N \) of \( \{ L^N_{\tau(n)} \}_{t \geq 0} \) is a stationary distribution. Let \( \{ \tau(n) \}_{n \in \mathbb{N} \subset \mathbb{R}^+} \) be a strictly increasing subsequence along which \( L^N_{\tau(n)} \) converges to \( \xi^N \). On that subsequence, since \( g^N_b \) is continuous and bounded,

\[
\lim_{n \to \infty} \frac{1}{\tau(n)} \mathbb{E}^N \left[ \int_0^{\tau(n)} g^N_b \left( B^N \left( \pi^N_p, t/N \right) / N \right) dt \right] = \lim_{n \to \infty} \frac{1}{\tau(n)} \mathbb{E}^N \left[ \int_0^{\tau(n)} U(t) dt \right] = \lim_{n \to \infty} \frac{1}{\tau(n)} \int_0^{\tau(n)} \left( \int_0^\infty \mathbb{P}^N_x \left( U(t) > x \right) dx \right) dt = \lim_{n \to \infty} \mathbb{E}^N_{L^N_{\tau(n)}} [U] = \mathbb{E}^N \left[ g^N_b \left( B^N / N \right) \right],
\]

where the last equality follows from the bounded convergence theorem, by noting that \( B^N \left( \pi^N_p, t \right) \leq N \) is a bounded continuous function on the state space of the Feller Markov process \( Y^N_p \). From Lemma 4 in [12],

\[
M^N \left( \pi^N_p, t \right) := R^N \left( \pi^N_p, t \right) - \int_0^t \left( \mathbb{1}_{[0, \lambda^N \left( u- \right)]} h^r, \eta^N \left( u \right) \right) du
\]

is a martingale (with respect to the filtration \( \mathcal{F}^N_t \) defined in [12]), because \( \mathbb{1}_{[0, \lambda^N \left( u- \right)]} \) is an almost surely bounded, measurable, real-valued function on \([0, h^r] \times \mathbb{R}_+ \). Hence,

\[
\mathbb{E}^N \left[ \frac{R^N \left( \pi^N_p, \tau(n) \right)}{\tau(n)} \right] = \mathbb{E}^N \left[ \frac{1}{\tau(n)} \int_0^{\tau(n)} \left( \mathbb{1}_{[0, \lambda^N \left( u- \right)]} h^r, \eta^N \left( u \right) \right) du \right]
\]

\[
= \mathbb{E}^N \left[ \frac{1}{\tau(n)} \int_0^{\tau(n)} \left( \mathbb{1}_{[0, \lambda^N \left( u \right)]} h^r, \eta^N \left( u \right) \right) du \right]
\]

\[
= \frac{1}{\tau(n)} \int_0^{\tau(n)} \left( \int_0^\infty \mathbb{P}^N_x \left( \mathbb{1}_{[0, \lambda^N \left( u \right)]} h^r, \eta^N \left( u \right) > x \right) dx \right) du
\]

\[
= \mathbb{E}^N_{L^N_{\tau(n)}} \left[ \left( \mathbb{1}_{[0, \lambda^N]} h^r, \eta^N \right) \right],
\]

where the second equality follows by noting that \( \{ t \geq 0 : \lambda^N \left( t- \right) \neq \lambda^N(t) \} \) has Lebesgue measure zero. Since \( h^r \) is bounded,

\[
\mathbb{1}_{[0, \lambda^N(t)]} h^r, \eta^N(t) \leq \mathbb{1}_{h^r, \eta^N(t)} \leq \| h^r \|_{\infty} \langle 1, \eta^N(t) \rangle,
\]

which implies \( \left\{ \langle \mathbb{1}_{[0, \lambda^N(t)]} h^r, \eta^N(t) \rangle \right\}_{t \geq 0} \) is uniformly integrable, and so

\[
\lim_{n \to \infty} \mathbb{E}^N \left[ \frac{R^N \left( \pi^N_p, \tau(n) \right)}{\tau(n)} \right] = \lim_{n \to \infty} \mathbb{E}^N_{L^N_{\tau(n)}} \left[ \left( \mathbb{1}_{[0, \lambda^N]} h^r, \eta^N \right) \right] = \mathbb{E}^N \left[ \left( \mathbb{1}_{[0, \lambda^N]} h^r, \eta^N \right) \right].
\]
From (11) and (12), it follows that

\[
C_N(\pi_N) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\xi \left[ \frac{a}{N} \left( E_N^N(T) - E_{p,N}^N(T) \right) + \int_0^T gU \left( \frac{B_N(\pi_N^N, t)}{N} \right) dt \right]
\]

\[
= \limsup_{T \to \infty} a \cdot \frac{1}{N} \mathbb{E}_\xi \left[ \frac{E_N^N(T) - E_{p,N}^N(T)}{T} \right] + \limsup_{T \to \infty} \frac{1}{N} \mathbb{E}_\xi \left[ R_N(\pi_N^N, T) \right]
\]

\[
+ \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\xi \left[ \int_0^T gU \left( \frac{B_N(\pi_N^N, t)}{N} \right) dt \right]
\]

\[
= \mathbb{E}_\xi \left[ a \cdot \frac{1}{N} (1 - p) \lambda^N + a \cdot \frac{1}{N} \mathbb{E}_\xi \left[ \mathbb{1}_{[0, \infty]} h^r \eta^N_\infty \right] + gU \left( \frac{B_N}{N} \right) \right],
\]

which establishes the statement.

8. Extension: Holding Cost

In this section, we include holding costs in the objective function to penalize congestion, and we show similar results as in the case with abandonment cost only, using the enlarged admissible policy class with admission control.

Let \( c \) be the holding cost incurred per customer per unit time. Then, the long-run average cost is modified as

\[
C_N^{hc}(\pi_N) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\xi \left[ \left( a \cdot \frac{1}{N} \left( E_N^N(T) - E_{p,N}^N(T) \right) + \int_0^T gU \left( \frac{B_N(\pi_N^N, t)}{N} \right) dt \right] + a R_N(\pi_N^N, T) \right]
\]

\[
= \mathbb{E}_\xi \left[ a \cdot \frac{1}{N} (1 - p) \lambda^N + a \cdot \frac{1}{N} \mathbb{E}_\xi \left[ \mathbb{1}_{[0, \infty]} h^r \eta^N_\infty \right] + gU \left( \frac{B_N}{N} \right) \right]
\]

and the objective is to determine \( \pi_N^{opt, hc} \) and \( C_N^{hc}(\pi_N^{opt, hc}) \) such that

\[
C_N^{hc}(\pi_N^{opt, hc}, hc) := \inf_{\pi_N^N \in \Pi_N} C_N^{hc}(\pi_N^N).
\] (13)

We begin by noting that the fluid model equations are not changed, and hence the fluid control problem can be obtained based on the unchanged fluid model invariant states. Also, the weak convergence result (Lemma 1) continues to hold, because the proof of Lemma 1 does not rely on the objective function.

Assumption 2. The function \( h^r \) is non-increasing.

Assumption 2 is crucial to prove the main asymptotic optimality result, Theorem 1, when holding costs are considered. To explain this point, for \( p \in (0, 1) \) and \( b \in [0, \min\{1, p\lambda/\mu\}] \), define

\[
q(b, p) := p\lambda \int_0^{(G^*)^{-1}(1-b\mu/p\lambda)} (1 - G^r(x)) dx.
\] (14)
From Equation (54) in [12], \( q(b, p) \) represents the invariant fluid queue length for \( p \in (0, 1) \) and \( b \in [0, \min\{1, p\lambda/\mu\}] \) when the arrival rate is thinned to \( p\lambda \). As a consequence of Assumption 2, \( q(\cdot, p) \) is a convex function on \([0, \min\{1, p\lambda/\mu\}]\) for each \( p \in (0, 1) \) (see Remark 10 in [11]). This convexity plays an integral role in our analysis. The modified fluid control problem is

\[
\min_{b \in [0, \min\{1, \lambda/\mu\}]} c \cdot q(b, 1) + a(\lambda - b\mu) + g_U(b). \tag{15}
\]

Different from (2) when the holding costs were not included, the optimization problem (15) becomes sensitive to the patience distribution, because the fluid queue length depends on the patience distribution; see the right-hand side of (14). We denote the solution to (15) by \( q^*_b \). As in Section 4, if \( b^{\text{ad}} < \min\{1, \lambda/\mu\} \), then we expect an idling control policy to be optimal for (15). In particular, servers can be allowed to take a rest for \( 1 - b^{\text{ad}} \) time units after each service completion.

As in Sections 5 and 6, in order to show the asymptotic optimality property, we work with the enlarged admissible policy class \( \Pi^N \), which incorporates the potential of admission control. The unit abandonment and holding cost for the admitted arrivals remain \( a \) and \( c \). For every rejected arrival, in addition to a cost of \( a \) (as in 5), we need to further account for the holding cost that would have been incurred if under a control policy in \( \Pi^N \) without admission control that may idle. Suppose that \( \tilde{C}(b, p) \) is the overall holding costs for the rejected arrivals for \( p \in (0, 1] \) and \( b \in [0, \min\{1, p\lambda/\mu\}] \), and we call it the fluid-scaled holding cost compensator.

**Definition 5 (Fluid-Scaled Holding Cost Compensator).** Given \( \lambda \in (0, \infty) \), \( p \in (0, 1] \) and \( b \in [0, \min\{1, p\lambda/\mu\}] \), \( \tilde{C}(b, p) \) is given by

\[
\tilde{C}(b, p) = c \cdot (q(b, 1) - q(b, p)).
\]

Then, the modified fluid control problem under \( \Pi^N \) is

\[
\min_{p \in (0, 1], b \in [0, \min\{1, p\lambda/\mu\}]} c \cdot q(b, p) + \tilde{C}(b, p) + a(p\lambda - b\mu) + a(1 - p)\lambda + g_U(b)
= \min_{p \in (0, 1], b \in [0, \min\{1, p\lambda/\mu\}]} c \cdot q(b, p) + \tilde{C}(b, p) + a(\lambda - b\mu) + g_U(b). \tag{16}
\]

Under Definition 5, it is clear that the solution to (16) is identical to the solution to (15). Then, from (16), the modified objective function under \( \pi_p^N \in \Pi_p^N \) (Definition 3) is given by

\[
\tilde{C}_b^N(\pi_p^N) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_N \left[ c \int_0^T \frac{q_N(\pi_p^N, t)}{N} dt + \int_0^T \tilde{C} \left( \frac{B_N(\pi_p^N, t)}{N}, \pi_p^N \right) dt + a \frac{(E_N^N(T) - E_p^N(T)) + R_N(\pi_p^N, T)}{N} + \int_0^T g_U \left( \frac{B_N(\pi_p^N, t)}{N} \right) dt \right], \tag{17}
\]
where the process $Q^N(\pi^N_p, t)$ tracks the $p$-admitted queue length at time $t$.

Noting that the fluid solution to (15), $b_{*,hc}$, is independent of $p$, we can consider the same proposed policy as defined in Definition 4 with $b_*$ replaced by $b_{*,hc}$; denote it by $\pi^N_{*,hc}$. In the remainder of the section, we outline the proof of asymptotic optimality for $\pi^N_{*,hc}$.

Lemma 2 continues to hold, because its proof does not rely on the objective function. For the objective (15), Lemma 5 can be easily modified such that for any admissible HL control policy $\pi^N_p \in \Pi^N_p$, when the system is initialized according to a stationary measure $\xi^N$,

$$C^N_{hc}(\pi^N_p) = \mathbb{E}_\xi^N \left[ c \cdot \frac{Q^N_p}{N} + \tilde{C} \left( \frac{B^N_p}{N}, p \right) + a \cdot \frac{1}{N} (1-p) \lambda^N + a \cdot \frac{1}{N} \left( 1_{[0,\chi^N]} h^r, \eta^N \right) + \left( \frac{B^N_p}{N} \right) \right].$$

This is because for each $N \in \mathbb{N}$, if $\{\tau(n)\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ is a strictly increasing subsequence along which $L^N_{\tau(n)}$ converges to $\xi^N$, then

$$\lim_{n \to \infty} \frac{1}{\tau(n)} \mathbb{E}_\xi^N \left[ \int_0^{\tau(n)} Q^N(\pi^N_p, t) dt \right] = \mathbb{E}_\xi^N \left[ Q^N \right],$$

because $Q^N(t) \leq \langle 1, \eta^N(t) \rangle$ for all $t \geq 0$, and $\langle 1, \eta^N(t) \rangle$ is uniformly integrable by similar argument as that in the proof of Proposition 4.1 in [4].

Thus, following a similar proof to that of Theorem 2 the sequence $\{\pi^N_{*,hc}\}_{N \in \mathbb{N}}$ satisfies

$$\lim_{N \to \infty} C^N_{hc}(\pi^N_{*,hc}) = c \cdot q(b_{*,hc}, p) + \tilde{C}(b_{*,hc}, p) + a \left( \lambda - b_{*,hc} \mu \right) + g_U(b_{*,hc})$$

$$= c \cdot q(b_{*,hc}, 1) + a \left( \lambda - b_{*,hc} \mu \right) + g_U(b_{*,hc}).$$

We wish to show that for $\{\hat{\pi}^N\}_{N \in \mathbb{N}}$ as in the statement of Theorem 1

$$\lim_{N \to \infty} C^N_{hc}(\hat{\pi}^N) \geq c \cdot q(b_{*,hc}, 1) + a \left( \lambda - b_{*,hc} \mu \right) + g_U(b_{*,hc}).$$

As in the proof of Theorem 1 on any convergent subsequence $(\tilde{X}^N, \tilde{\nu}^N, \tilde{\eta}^N)$,

$$(\tilde{X}^N, \tilde{\nu}^N, \tilde{\eta}^N) \Rightarrow (X^\infty, \nu^\infty, \eta^\infty), \text{ as } N_i \to \infty,$$

with $(X^\infty, \nu^\infty, \eta^\infty)$ being almost surely a stationary fluid model solution, and

$$\lim_{N_i \to \infty} \mathbb{E}_{\xi^N_i} \left[ c \cdot \frac{Q^N_{\tilde{X}^N_i}}{N_i} + \tilde{C} \left( \frac{B^N_{\tilde{X}^N_i}}{N_i}, p \right) \right. $$

$$\left. + a \cdot \frac{1}{N_i} (1-p) \lambda^N_i + a \cdot \frac{1}{N_i} \left( 1_{[0,\tilde{X}^N_i]} h^r, \eta^N_i \right) + \left( \frac{B^N_{\tilde{X}^N_i}}{N_i} \right) \right]$$

$$= \mathbb{E}_{\xi} \left[ c \cdot Q + \tilde{C}(B^\infty, p) + a \cdot (1-p) \lambda + a \cdot \left( 1_{[0,\xi^\infty]} h^r, \eta^\infty \right) + g_U(B^\infty) \right]$$

$$\geq c \cdot q(b, 1) + a \left( \lambda - b \mu \right) + g_U(b)$$

$$\geq c \cdot q(b_{*,hc}, 1) + a \left( \lambda - b_{*,hc} \mu \right) + g_U(b_{*,hc}),$$
where the first inequality follows because
\[
E[\xi a \cdot (1 - p)\lambda + a \cdot \langle 1_{[0,\infty]}, h^r, \eta \rangle + g_U(B_\infty)] \geq a(\lambda - b\mu) + g_U(b) \quad \text{(from the proof of Theorem 4), and}
\]
\[
E[\xi c \cdot Q_\infty + \tilde{C}(B_\infty, p)] \geq c \cdot q(b, 1). \quad (18)
\]
To see (18), note that, by definition,
\[
c \cdot Q_\infty + \tilde{C}(p, B_\infty) = c \cdot q(B_\infty, p) + c \cdot (q(B_\infty, 1) - q(B_\infty, p)) = c \cdot q(B_\infty, 1).
\]
Since \(q(\cdot, 1)\) is convex on \([0, \min\{1, \lambda/\mu\}]\), Jensen’s inequality implies that
\[
E[c \cdot Q_\infty + \tilde{C}(p, B_\infty)] = E[c \cdot q(B_\infty, 1)] \geq c \cdot q(E[B_\infty], 1) = c \cdot q(b, 1).
\]
Hence, the proposed policy \(\pi_{N^*, hc}\) with \(p^*, hc = b^*, hc \mu/\lambda\) is asymptotically optimal, with the generalized objective function involving holding costs.

Remark 3. All the results in this section continue to stand if we consider a non-decreasing convex holding cost rather than a linear holding cost. In that case, the corrected fluid-scaled holding cost compensator defined in Definition 5 is the difference of the holding cost function evaluated at \(q(b, 1)\) and at \(q(b, p)\).

Appendix A. The Fluid Model

For the reader’s convenience, we write the fluid model equations and fluid model solutions in this appendix. We refer the reader to Section 3.1 in [12] for details. More notation is needed. Given a Polish space \(S\), we use \(C(S)\) to denote the set of functions having domain \(\mathbb{R}_+\) and range \(S\) that are continuous in time.

The fluid model has as an input a non-decreasing function \(E \in C(\mathbb{R}_+)\) such that \(E(0) = 0\), which we refer to as an arrival function. We set \(X := \mathbb{R}_+ \times [0, H^r] \times M[0, H^r]\), endowed with the product topology in a Polish space. To define the fluid model for an arrival function \(E\), we consider \((X, \nu, \eta) \in C(X)\) such that
\[
\langle 1_{\{x\}}, \eta(0) \rangle = 0, \quad \text{for all } x \in [0, H^r), \quad (A.1)
\]
and such that for each \(t \geq 0\),
\[
\langle 1, \nu(t) \rangle \leq X(t) \leq \langle 1, \nu(t) \rangle + \langle 1, \eta(t) \rangle, \quad (A.2)
\]
\[
\langle 1, \nu(t) \rangle \leq 1, \quad (A.3)
\]
\[
\int_0^t \langle h^s, \nu(u) \rangle \, du < \infty \quad \text{and} \quad \int_0^t \langle h^r, \eta(u) \rangle \, du < \infty. \quad (A.4)
\]
Given \((X, \nu, \eta) \in \mathbf{C}(\mathbb{X})\) satisfying \([A.1]-[A.4]\), we define auxiliary functions \(B, Q, \chi, R, D, K\) in \(\mathbf{C}(\mathbb{R}_+)\) and \(I\) in \(\mathbf{C}(\mathbb{R}_+)\) as follows: for each \(t \geq 0\),

\[
B(t) = \langle 1, \nu(t) \rangle, \tag{A.5}
\]
\[
Q(t) = X(t) - B(t), \tag{A.6}
\]
\[
\chi(t) := \inf \{ x \geq 0 : \langle 1_{[0,x]}, \eta(t) \rangle \geq Q(t) \}, \tag{A.7}
\]
\[
R(t) = \int_0^t \left( \int_0^{\chi(u)} h^r(w)\eta(u)(dw) \right) du, \tag{A.8}
\]
\[
D(t) = \int_0^t \langle h^s, \nu(u) \rangle du, \tag{A.9}
\]
\[
K(t) = B(t) + D(t) - B(0), \tag{A.10}
\]
\[
I(t) = 1 - B(t). \tag{A.11}
\]

Then \(B, Q, \chi, R, D, K,\) 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 \((X, \nu, \eta) \in \mathbf{C}(\mathbb{X})\) for an arrival function \(E\) are as follows: for any continuous and bounded function \(f\) having domain \(\mathbb{R}_+\), for each \(t \geq 0\),

\[
K\text{ is non-decreasing}, \tag{A.12}
\]
\[
X(t) = X(0) + E(t) - R(t) - D(t), \tag{A.13}
\]
\[
\langle f, \nu(t) \rangle = \left\langle f(\cdot + t) \frac{1 - G^s(\cdot + t)}{1 - G^s(\cdot)}, \nu(0) \right\rangle + \int_0^t f(t - u)(1 - G^r(t - u))dK(u), \tag{A.14}
\]
\[
\langle f, \eta(t) \rangle = \left\langle f(\cdot + t) \frac{1 - G^r(\cdot + t)}{1 - G^r(\cdot)}, \eta(0) \right\rangle + \int_0^t f(t - u)(1 - G^r(t - u))dE(u). \tag{A.15}
\]

**Definition 6.** Let \(E\) be an arrival function. A fluid model solution for \(E\) is \((X, \nu, \eta)\) that satisfies \([A.1]-[A.4]\), and \([A.12]-[A.17]\).

**Definition 7.** Let \(E\) be an arrival function. A non-idling fluid model solution for \(E\) is \((X, \nu, \eta)\) that satisfies Definition 6 and the following non-idling condition for each \(t \geq 0\):

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

**References**

[1] Afche, Philipp. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* 15(3) 423–443.
[2] Afeche, Philipp, J Michael Pavlin. 2016. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science* **62**(8) 2412–2436.

[3] Atar, Rami, Amarjit Budhiraja, Paul Dupuis, Ruoyu Wu. 2019. Large deviations for the single server queue and the reneging paradox. *arXiv preprint arXiv:1903.06870* .

[4] Atar, Rami, Haya Kaspi, Nahum Shimkin. 2014. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research* **39**(3) 672–696.

[5] Atar, Rami, Yair Y. Shaki, Adam Shwartz. 2011. A blind policy for equalizing cumulative idleness. *Queueing Systems* **67**(4) 275–293.

[6] Da Prato, Giuseppe, Jerzy Zabczyk, J Zabczyk. 1996. *Ergodicity for infinite dimensional systems*, vol. 229. Cambridge University Press.

[7] Kang, Weining, Guodong Pang. 2019. Equivalence of fluid models for Gt/GI/N+ GI queues. *Modeling, Stochastic Control, Optimization, and Applications*. Springer, 315–349.

[8] Kang, Weining, Kavita Ramanan, et al. 2010. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* **20**(6) 2204–2260.

[9] Kang, Weining, Kavita Ramanan, et al. 2012. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability* **22**(2) 477–521.

[10] Maglaras, Costis, John Yao, Assaf Zeevi. 2018. Optimal price and delay differentiation in large-scale queueing systems. *Management Science* **64**(5) 2427–2444.

[11] Puha, Amber L, Amy R Ward. 2019. Scheduling an overloaded multiclass many-server queue with impatient customers. *Operations Research & Management Science in the Age of Analytics*. INFORMS, 189–217.

[12] Puha, Amber L, Amy R Ward. 2021. Fluid limits for multiclass many server queues with general reneging distributions and head-of-line scheduling. *Mathematics of Operations Research* .

[13] Ward, Amy R, Mor Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.

[14] Ward, Whitt. 2006. The impact of increased employee retention upon performance in a customer contact center. *Manufacturing & Service Operations Management* .

[15] Whitt, Ward. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54**(1) 37–54.
[16] Zhan, D., A. R. Ward. 2019. Staffing, routing, and payment to trade off speed and quality in large service systems. *Operations Research* 67 1738–1751.

[17] Zhang, Jiheng. 2013. Fluid models of many-server queues with abandonment. *Queueing Systems* 73(2) 147–193.