Submitted to
manuscript (Please, provide the manuscript number!)

Staffing a Service System Where Capacity Is Random

Rouba Ibrahim
School of Management, University College London, London, U.K., WC1E 6BT
rouba.ibrahim@ucl.ac.uk

One prevalent assumption in queueing theory is that the number of servers in a queueing model is deterministic. However, randomness in the number of servers often arises in practice, particularly when the servers themselves may be viewed as strategic decision makers, e.g., in virtual call centers or ride-sharing services where agents are allowed to set their own schedules, often at very short time notice. In this paper, we study the problem of staffing many-server queues with generally-distributed customer abandonment and a random number of servers. We rely on fluid approximations to determine cost-minimizing staffing levels in that setting, and we demonstrate the asymptotic accuracy of fluid-based performance measures and staffing prescriptions. For an application, we characterize the optimal staffing policy with self-scheduling servers, and show that this policy is not straightforward: It may be optimal to either understaff or overstaff the system, depending on (i) self-scheduling agent behavior and (ii) the abandonment-time distribution.

Key words: staffing; many-server queues; fluid approximation; asymptotic accuracy; self-scheduling servers.

1. Introduction

There is a broad literature in queueing theory which studies the problem of staffing large-scale service systems. This literature has important practical implications for the design of those systems; e.g., for surveys of applications in call-center management, see Gans et al. (2003) and Akşin et al. (2007). Much of that body of research formulates staffing recommendations based on queueing models with several realistic features, such as time-varying parameters and non-standard network structures. However, one prevalent assumption in those models is that the (possibly time-varying)

number of servers is deterministic. As such, the *realized* staffing level in any given time period is assumed to be equal to the *planned* staffing level, set by the system manager, for that period.

The purpose of this paper is to study the staffing problem in many-server queueing systems with a *random* number of servers. (Throughout this paper, we use "servers" and "agents" interchangeably.) Uncertainty in the number of servers may arise due to novel work arrangements, such as in virtual call centers, where agents set their own schedules (Gurvich et al. 2015). It may also arise in work environments where there is significant non-adherence to planned schedules, e.g., when frequent absenteeism is common, as for nurses in healthcare systems (Green et al. 2013), or call center representatives in physical call centers (Whitt 2006b). In all of those examples, agents may be viewed as being *strategic*, i.e., they are decision makers who choose whether or not to be available for work in a given period so as to maximize their individual utilities.

1.1. Motivating Applications

Virtual call centers. In virtual call centers, call center representatives (agents) are home-based and have the flexibility of determining their own working shifts, often at very short time notice. We refer to such agents as *self-scheduling*. Managing systems with self-scheduling agents involves two different time scales: (i) weeks ahead of time, the system manager selects the total staffing level in the system e.g., to allow sufficient time for agent training and qualification; and, (ii) hours ahead of time, or less, agents select their own schedules. Since the agent population is both remote and large (up to hundreds of agents), system managers cannot simply solicit their agents' scheduling preferences ahead of time. Moreover, the promised scheduling flexibility constitutes the main appeal of this type of jobs, and cannot be simply restricted by the firm (Arise 2015, Cloud Source 2015, Great Virtual Works 2015, LiveOps 2015, West at Home 2015).

Ride-sharing services. Ride-sharing services, such as Uber, also allow their drivers to self-schedule. They use "surge pricing" (Uber 2015) to ensure the participation of a sufficient number of drivers in different time periods. Since flexibility is important for drivers and cannot be restricted (Hall and Krueger 2015), the firm is faced with the problem of deciding on the number of self-scheduling drivers to hire (Fast Company 2015, Gurvich et al. 2015).

Allowing agents to self-schedule raises important operational challenges: It could lead to low levels of service, i.e., high customer-related costs, in understaffed working periods. Conversely, it could also lead to overstaffed working periods where customer-related costs are low, while staffing costs are too high. In setting appropriate staffing levels, system managers need to account for the uncertainty in the numbers of agents who will be present in different shifts.

Agent absenteeism. Uncertainty in the number of available agents may also arise in work environments where there is significant non-adherence to planned schedules, e.g., due to prevalent agent absenteeism. For example, it is well known that nurse absenteeism is a considerable problem in healthcare settings (US Bureau of Labor Statistics 2008, BBC 2015). As a result, the actual number of nurses who show up for work, in a given period, is uncertain. The same phenomenon can also be observed in physical call centers which are plagued with low employee satisfaction and considerable absenteeism (Whitt 2006b). When setting staffing levels in such systems, it is important to account for subsequent agent absenteeism. Indeed, Green et al. (2013) show that failing to do so (with workload-dependent nurse absenteeism) typically results in understaffing the system. In contrast, we show here to it may lead to either overstaffing or understaffing the system, depending on strategic agent behavior and other model characteristics.

1.2. Framework, Contributions, and Organization

In this paper, we consider many-server queueing models with generally-distributed customer abandonment, generally-distributed service times, a stationary arrival process, and a random number of servers. We assume that there are k working periods, and that agents have inherent, heterogenous, *availabilities* or *preferences* for different periods; e.g., in virtual call centers, agents who are parents may value morning shifts more because they may be busy with family obligations in the afternoon. We assume that there is a fixed staffing cost per server, depending on the period. First, the system manager decides on the total staffing level, n. Then, each agent in the pool of size n has, independently of other agents, a fixed probability, r_j, of being available for work in period j, $1 \leq j \leq k$. The derivation of r_j depends on the specific application context in mind; we provide

details in §3. For example, depending on the specific application in mind, we may require that $\sum_{j=1}^{k} r_j = 1$. As a result of those system dynamics, the (marginal) distribution of the number of available agents in period j, N_j^n, is binomial with parameters n and r_j.

Our objective is to determine n optimally so as to minimize the expected total system cost, which is the sum of staffing and customer-related costs, across all working periods. In other words, we focus on the long-term planning problem faced by the managers of such systems. Since the staffing problem, even with a deterministic number of servers, is not amenable to exact analysis, we determine optimal staffing levels by solving its fluid approximation, drawing on Whitt (2006a,b) and Bassamboo and Randhawa (2010). Our modelling framework is closest to Harrison and Zeevi (2005), Bassamboo et al. (2005, 2006) and Bassamboo and Randhawa (2010) because they all exploit fluid models in setting staffing requirements. However, in all of those papers, unlike in ours, the number of servers is assumed to be deterministic. Here are the main contributions of this paper.

Methodology: Asymptotic accuracy of fluid prescriptions. Fluid approximations are well known to be useful in describing the behavior of large, especially overloaded, queueing systems (Dai and He 2012). However, there remains to show whether they could be reliably used to approximate expected performance measures in queueing systems where the number of servers is random. In such systems, those measures must be computed by conditioning (and unconditioning) on the possible realizations of the number of servers. Since those alternative realizations correspond to different workloads (e.g., the system could either be underloaded or overloaded depending on the realized number of servers), it is not clear, a priori, what the resulting "weighted average" system performance is, or what would be a good choice for an approximating fluid model.

In this paper, we define fluid approximations by substituting the *random* number of servers in the original queueing system by its *expected value*, and deriving fluid approximations for the resulting system. Our objectives are to investigate whether the ensuing approximations could be used to: (i) approximate expected performance measures in the original system; and (ii) determine staffing levels that minimize the expected cost in the original system.

With those objectives in mind, we extend the results of Bassamboo and Randhawa (2010) to systems with uncertainty in the number of servers. In particular, we characterize the asymptotic (i.e., when the arrival rate is large) accuracy of fluid approximations and prescriptions with a binomial number of servers in the overloaded, critically-loaded, and underloaded regimes. Since the actual number of servers is random, we define those operational regimes relative to the expected number of servers instead (§4). For analytical tractability, we assume exponentially-distributed service times and a Poisson arrival process (along with generally-distributed times to abandon).

It is insightful that the underloaded regime is of interest when the number of servers is random. Indeed, it may be asymptotically optimal to purposely overstaff the system in that case (§6). In contrast, the underloaded regime is never asymptotically optimal when the number of servers is deterministic. In the overloaded regime, we show that the accuracy gap of fluid approximations does not increase with the arrival rate, although the performance metrics themselves do. Thus, fluid approximations are "extremely accurate" in that case. In the underloaded regime, we show that the fluid accuracy gap decreases in the arrival rate; in the critically-loaded regime, we show that it is of the order of the square root of the arrival rate. In both the underloaded and critically-loaded regimes, performance measures are asymptotically negligible. For all three operational regimes, we demonstrate that fluid-based staffing prescriptions are *asymptotically accurate* (Theorem 2).

Our asymptotic results provide general theoretical support to the usefulness of simple fluid approximations in staffing queueing systems where there is uncertainty in the number of servers. Doing so is important because such queueing models are becoming increasingly relevant in practice, as explained in §1.1. Thus, studying ways of reliably approximating their dynamics is of interest.

Application: Staffing with self-scheduling servers. We apply fluid approximations to determine optimal staffing levels with self-scheduling servers, generally-distributed service and abandonment times, and a stationary arrival process. For simplicity, we assume that there are only two periods in any day, and that agents must select to work in exactly one of those two periods (similar insights can extend to multiple periods too). The two periods represent e.g., daily and nightly

shifts. In particular, we are thinking of virtual call centers who treat their agents as employees rather than independent contractors. Typically, virtual agents who are employees earn a base wage (and possibly other benefits), are subject to a requirement on the minimum number of working hours (usually 15-20 hours/week), and are otherwise free to select one of several available shifts; e.g., see Cloud Source (2015), Great Virtual Works (2015), and West at Home (2015). In contrast, virtual agents who are independent contractors are usually not guaranteed a minimum wage and have no requirement on the least number of working hours to be fulfilled; e.g., as for LiveOps (2015) or Arise (2015). We describe the optimal staffing policy, as follows.

In a system where the critically-loaded regime is asymptotically optimal for both periods (without self-scheduling), we show that the optimal staffing policy in response to self-scheduling is not straightforward. In particular, we demonstrate that it may be optimal to either *understaff* or *overstaff* the system, and characterize how the optimal staffing policy depends on both strategic agent behavior and the abandonment-time distribution. We also show that the optimal staffing policy depends solely on the *proportion* of agents who prefer one period over another, but is otherwise *independent* of further distributional assumptions on agents' per-period utilities.

Numerical study. We conduct a numerical study with alternative abandonment-time distributions to: (i) provide numerical confirmation for our asymptotic accuracy results and (ii) extend our analysis to more general settings. For one example, we verify numerically that our asymptotic results extend to general service times as well; see Tables 4 and 5 in the appendix. For a second example, we study staffing decisions in systems where the overloaded regime is asymptotically optimal without self-scheduling agents. To illustrate the practical usefulness of our asymptotic results, we consider, throughout, systems which are not too large in size (tens of servers), and show that our approximations remain useful in describing the expected performance of those systems.

The remainder of this paper is organized as follows. In §2, we review the relevant literature. In §3, we describe our modelling framework and the system manager's problem. In §4, we formulate our main theoretical results concerning the asymptotic accuracy of the fluid prescription. In §5, we

prove our main results for the overloaded regime. In §6, we formulate staffing recommendations with self-scheduling agents. In §7, we describe our numerical results. In §8, we draw conclusions. We relegate other technical proofs to the appendix.

2. Related Literature

Our paper is related to the extensive literature analyzing asymptotics of many-server queueing systems with impatient customers; see Garnett et al. (2002), Ward and Glynn (2003), Zeltyn and Mandelbaum (2005), Whitt (2004, 2006a), Bassamboo and Randhawa (2010), Bassamboo et al. (2010), and references therein. It is also related to the large literature on optimal staffing decisions in service systems, e.g., including Maglaras and Zeevi (2003), Borst et al. (2004), Harrison and Zeevi (2005), and Bassamboo et al. (2005, 2006); for other references, see Gans et al. (2003) and Akşin et al. (2007). However, none of those papers considers a random number of servers.

To the best of our knowledge, the only two papers that study queueing systems with a random number of servers are Whitt (2006b) and Atar (2008). Whitt (2006b) considers many-server queues with an uncertain arrival rate and an uncertain number of servers. However, the focus of that paper is primarily on a numerical investigation of fluid-based prescriptions. As such, our results provide further theoretical grounding to the insights of that paper. Atar (2008) derives a diffusion limit for the number of customers in a system with a random number of servers and random service rates, but no customer abandonment. However, the staffing question is not addressed in that paper.

Our work is also related to the area of queuing games, reviewed in Hassin and Haviv (2003), which mostly focuses on the impact of customers acting strategically. There is a body of work within this literature which considers strategic servers that may select their service rates; e.g., see Cachon and Harker (2002) and Cachon and Zhang (2007). However, such papers do not consider staffing decisions, and the maximum number of servers considered is two. Recent exceptions are Gopalakrishnan et al. (2015) and Zhan and Ward (2015) who study, respectively, optimal routing and staffing decisions, and optimal compensation schemes in many-server queueing models where servers strategically choose their service rates. Gopalakrishnan et al. (2015) find that an

asymptotically optimal staffing policy staffs strictly more than the common square-root staffing rule (corresponding to the critically-loaded regime, at fluid scale). The analysis in Zhan and Ward (2015) is in the same spirit as ours, but their problem setting is different. They use a fluid approximation to find staffing and employee compensation policies which minimize systems costs, and find that different operating regimes, including the underloaded, critically-loaded, or overloaded regimes, may emerge depending on specific modelling assumptions.

Our work is also related to Gurvich et al. (2015) who study optimal staffing and compensation schemes with self-scheduling agents and no customer abandonment. However, they do not investigate the asymptotic accuracy of the fluid model. Moreover, agents in their setting are *independent contractors* who need not make a choice between several available work shifts. In contrast, agents in our setting are *employees* of the firm who must make such a choice. In their setting, unlike in ours, Gurvich et al. find that self-scheduling always leads to reducing the system's staffing level.

3. Modelling Framework

In this section, we describe our modelling framework. In §3.1, we present our agent valuation model. In §3.2, we describe our queueing framework. Finally, in §3.3, we formulate the optimization problem faced by the system manager.

3.1. Agent Valuation Model

We assume that agents are risk-neutral, independent, and heterogeneous. The agent pool size is n, and there are k working periods. We assume that there is a fixed staffing cost, c_j, per server in period j, $1 \leq j \leq k$. Let X_j quantify the (random) personal utility that an agent obtains from working in period j, e.g., due to personal preferences or availability for that specific period. An agent i, $1 \leq i \leq n$, draws a random variate, $x_{i,j}$ from the distribution of X_j; the distribution itself is assumed to be common across all agents. Different agents make such draws independently, and the value of $x_{i,j}$ is private information to agent i. We assume that $\{X_j, 1 \leq j \leq k\}$ are independent. However, the distribution of X_j, as well as all other model parameters, are assumed to be common knowledge to the agents and the system manager. Based on the values of his/her personal utility $x_{i,j}$ and the compensation c_j, agent i decides whether or not to work in period j.

Let r_j denote the probability that a randomly selected agent, in the pool of size n, will work in period j. Letting V_j denote the total valuation of an agent for period j, we have that:

$$V_j = c_j + X_j. \tag{1}$$

With self-scheduling agents, an agent must select to work in exactly one of the k periods. In this case, an agent selects to work in period j_0, if, and only if, $X_{j_0} + c_{j_0} \geq X_j + c_j$ for all j, i.e.,

$$r_{j_0} = \mathbb{P}\left(X_{j_0} + c_{j_0} = \max_{1 \leq j \leq k} \{X_j + c_j\} \text{ and } c_{j_0} + X_{j_0} \geq 0 \right), \tag{2}$$

where we require the nonnegativity of $V_{j_0} = c_{j_0} + X_{j_0}$ because agents would otherwise not be interested in working for the system manager. Therefore, for each period j, the number of available servers, N_j^n, has a binomial distribution with parameters n and r_j. Also, the joint distribution of $\{N_j^n, 1 \leq j \leq k\}$ is multinomial with parameter n and success probabilities $\{r_j, 1 \leq j \leq k\}$.

We may also consider a different application context. For example, with agent absenteeism: An agent may select to show up for work in a period if, and only if, his valuation for that period is nonnegative, i.e., $r_{j_0} = \mathbb{P}(c_{j_0} + X_{j_0} \geq 0)$, and s/he may elect to work in multiple periods. Then, $\{N_j^n, 1 \leq j \leq k\}$ are independent but the marginal distribution of N_j^n remains binomial.

3.2. Queueing Model

We consider single-class $M/M/N+GI$ queueing models in steady state (we consider the $G/G/N+GI$ model in §6). We assume that customers arrive to the system according to independent Poisson processes with rates λ_j, $1 \leq j \leq k$. We assume that there is no service overlap between the different periods, i.e., customers who arrive during a period must be served by agents who are assigned to that period. The number of servers in period j, N_j^n, has a (marginal) binomial distribution $N_j^n \sim Bin(n, r_j)$. Conditional on the number of servers in a given period, queueing dynamics in different periods are independent.

We let service times be independent and identically distributed (i.i.d.) exponential random variables with rate μ. We assume that $\mu = 1$; we do so without loss of generality, because we are free to choose the time units in our system, and this assumption amounts to measuring time in units

of mean service time. Each arriving customer will abandon if he is unable to start service before a random amount of time, which we refer to as his patience time. Patience times are i.i.d. across customers, and have a cumulative distribution function (cdf) F, complementary cdf (ccdf) F^c, density function f, hazard-rate function f_a, and mean $1/\theta$ for some $\theta > 0$. Irrespective of the value of N_j^n, customer abandonment makes the system stable; see Baccelli et al. (1984). The arrival, service, and abandonment processes are all mutually independent, also independent of N_j^n. There is unlimited waiting space, and we use the first-come-first-served service discipline.

3.3. System Manager's Problem

As in Bassamboo and Randhawa (2010), we consider two quality-of-service costs, indexed by the period j: (i) A delay cost, h_j, which is incurred per customer for each unit of time that this customer spends waiting to be served, and (ii) an abandonment penalty cost, p_j, incurred per customer who abandons before being served. For period j, let $Q_{N_j^n}$ denote the steady-state queue length and $\alpha_{N_j^n}$ denote the net customer abandonment rate. The system manager's staffing problem is:

$$\min_{n\,\in\mathbb{N}} \Pi(n) \equiv \sum_{1\le j\le k} \Big(c_j \cdot \mathbb{E}[N_j^n] + p_j \cdot \mathbb{E}[\alpha_{N_j^n}] + h_j \cdot \mathbb{E}[Q_{N_j^n}] \Big), \tag{3}$$

where $\mathbb{N}$ denotes the set of natural integers. Since the staffing problem in (3) is not amenable to exact analysis, we consider a steady-state fluid approximation of the system instead (Whitt 2006a). We define $\rho_j \equiv \lambda_j/\mathbb{E}[N_j^n] = \lambda_j/nr_j$. We let $\bar{q}_{\rho_j}$ and $\bar{\alpha}_{\rho_j}$ denote the fluid-scaled limits for the queue-length and net abandonment rates in the corresponding $M/M/\mathbb{E}[N_j^n]+GI$ model with traffic intensity ρ_j and a deterministic number of servers $\mathbb{E}[N_j^n]$; this is a slight abuse of notation since $\mathbb{E}[N_j^n] = nr_j$ may not be integer valued. In other words, $\bar{\alpha}_{\rho_j} = (\rho_j - 1)^+$ and $\bar{q}_{\rho_j} = \rho_j \int_0^{w_j} F^c(x)dx$, where the fluid waiting time w_j is such that $\rho_j F(w_j) = \bar{\alpha}_{\rho_j}$ (Whitt 2006a). The fluid approximation to problem (3) is therefore given by:

$$\min_{n\,\in\mathbb{N}} \Pi_f(n) \equiv \sum_{1\le j\le k} \big(c_j \cdot \mathbb{E}[N_j^n] + p_j \cdot \mathbb{E}[N_j^n] \cdot \bar{\alpha}_{\rho_j} + h_j \cdot \mathbb{E}[N_j^n] \cdot \bar{q}_{\rho_j} \big). \tag{4}$$

Next, we establish the asymptotic accuracy of fluid approximations and staffing prescriptions.

4. Asymptotic Accuracy of Fluid Approximations and Prescriptions

Conditional on the number of servers in the period, queueing dynamics in different periods are assumed to be independent. Moreover, the number of servers in each period has a marginal binomial distribution. Thus, to establish the desired asymptotic accuracy, it suffices to focus on a single period and a system with a binomial number of servers. The proof for multiple periods can then be obtained by a simple argument, exploiting a similar conditioning argument as the one that we use in what follows, along with the conditional independence across periods. In other words, due to this conditional independence, the joint distribution of the numbers of servers across periods does not matter, so long as we can show fluid asymptotic accuracy in each period. In this section, for clarity of exposition, we consider a single-period setting (thus, we drop dependence on j).

4.1. Asymptotic Framework

We consider a sequence of queueing models indexed by the arrival rate λ, and study system performance as λ increases without bound. The number of servers in the λ^{th} system is $N_\lambda \sim Bin(n_\lambda, r)$. We assume that $\rho \equiv \lambda/\mathbb{E}[N_\lambda] = \lambda/rn_\lambda$ remains fixed as λ increases. Let Q_{N_λ} denote the steady-state queue length and α_{N_λ} the net customer abandonment rate in the $M/M/N_\lambda + GI$ queue. We refer to the cases with $\rho > 1$, $\rho < 1$, and $\rho = 1$ as the overloaded, underloaded, and critically loaded regimes, respectively. Note that since N_λ is random, an $M/M/N_\lambda + GI$ system with e.g., $\rho > 1$ may or may not be overloaded, i.e., having $\lambda > N_\lambda$. We summarize our main results in Theorem 1. In §5, we prove these results for the overloaded regime; the remaining proofs are in the appendix.

THEOREM 1. *Consider an $M/M/N_\lambda + GI$ queueing model with $N_\lambda \sim Bin(n_\lambda, r)$ and let $\rho \equiv \lambda/rn_\lambda$.*

(a) *If $\rho > 1$ (overloaded regime), then there exists a finite constant $K > 0$ such that*

$$\limsup_{\lambda\to\infty} \left|\mathbb{E}[Q_{N_\lambda}] - rn_\lambda \bar{q}_\rho\right| \le K \text{ and } \lim_{\lambda\to\infty} \left|\mathbb{E}[\alpha_{N_\lambda}] - rn_\lambda \bar{\alpha}_\rho\right| \to 0.$$

(b) *If $\rho = 1$ (critically-loaded regime), then there exist finite constants $K_1', K_2' > 0$ such that*

$$\limsup_{\lambda\to\infty} \mathbb{E}[Q_{N_\lambda}] \le K_1'\sqrt{\lambda} \text{ and } \limsup_{\lambda\to\infty} \mathbb{E}[\alpha_{N_\lambda}] \le K_2'\sqrt{\lambda}.$$

(c) *If* $\rho < 1$ *(underloaded regime), then*

$$\lim_{\lambda\to\infty} \mathbb{E}[Q_{N_\lambda}] \to 0 \text{ and } \lim_{\lambda\to\infty} \mathbb{E}[\alpha_{N_\lambda}] \to 0.$$

4.2. Interpretation of Theorem 1

Intuitively, the results of Theorem 1 hold because the binomial distribution concentrates "tightly" around its expected value, in a sense that will be made formal in §5. This is why fluid approximations which are based on approximating the random number of servers by its expected value are useful. To interpret the results of Theorem 1, we need the following definitions.

DEFINITION 1. Let f and g be two functions defined on some subset of $\mathbb{R}$. Then, as $n \to \infty$,

(a) $f(n) = \mathcal{O}(g(n))$ if there exists $M > 0$ and $C > 0$ such that $|f(n)| \leq M|g(n)|$ for $n \geq C$;

(b) $f(n) = o(g(n))$ if for all $\epsilon > 0$, there exists N such that $|f(n)| \leq \epsilon|g(n)|$ for all $n \geq N$.

Theorem 1 shows that, in the overloaded system, the fluid approximation for the expected queue length is asymptotically accurate up to $\mathcal{O}(1)$, and the fluid approximation for the net abandonment rate is asymptotically accurate up to $o(1)$, i.e., the corresponding error is asymptotically bounded in the former case, and it decreases with the arrival rate in the latter case. In other words, fluid approximations are "extremely accurate" in the overloaded regime. In the critically-loaded system, those fluid-approximation errors are $\mathcal{O}(\sqrt{\lambda})$, i.e., they grow in the square-root of the size of the system. In the underloaded regime, fluid approximations are $o(1)$-accurate since errors for both performance measures decrease with the arrival rate. We will show in §4.3 that fluid-based staffing prescriptions are asymptotically accurate in all those cases.

4.3. Staffing Prescriptions

We are now ready to establish the asymptotic accuracy of fluid-based staffing prescriptions, by exploiting the results of Theorem 1. We do so in Theorem 2, whose proof proceeds along similar lines as Theorem 3 in Bassamboo and Randhawa (2010).

THEOREM 2. *The fluid-based prescription,* n^*_λ*, i.e., the optimal solution to problem (4), is asymptotically optimal in the overloaded, critically-loaded and underloaded regimes in the sense that*

$$\lim_{\lambda\to\infty} \frac{\Pi^*_\lambda}{\Pi_\lambda(n^*_\lambda)} = 1,$$

where Π_λ^* *is the optimal objective value for (3) and* $\Pi_\lambda(n_\lambda^*)$ *is the value of its objective evaluated at* n_λ^**. If, in addition,* n_λ^* *is such that the system is overloaded, then there exists* $K'' > 0$ *such that*

$$\limsup_{\lambda\to\infty} |\Pi_\lambda^* - \Pi_\lambda(n_\lambda^*)| \leq K'',$$

i.e., the fluid staffing prescription is asymptotically $\mathcal{O}(1)$*-accurate in the overloaded regime.*

Next, we present simulation results which substantiate Theorem 1. In §7, we present additional results which substantiate Theorem 2 in the case of self-scheduling servers.

4.4. A Numerical Study

In Table 1, we present simulation estimates quantifying the accuracy of fluid performance measures in the overloaded regime, in the $M/M/N+GI$ model with $N \sim Bin(n, 0.4)$ and alternative values of n. Corresponding results for the critically-loaded and underloaded regimes are given in Table 3 of the appendix. We consider three different abandonment-time distributions: (i) exponential with mean 1, (ii) Pareto with shape parameter 2 and mean 1, i.e., $F^c(x) = 1/(1+x)^2$, and (3) uniform over [0.5, 1.5]. Our simulation estimates are based each on 400 independent replications of length 50,000 arrival events per replication, with an initial transient of length 2,000 arrival events removed from each replication (to ensure steady-state conditions).

We let n increase while holding $\rho = \lambda/nr = 1.4$. In addition to point estimates, we report half-widths of 95% confidence intervals. The results of Table 1 clearly illustrate the $\mathcal{O}(1)$-accuracy of fluid queue length, since the errors in the approximations for the expected steady-state queue length do not increase in the magnitude of the arrival rate. For the net abandonment rate, the reported errors converge to 0, as desired, so that we have $o(1)$-accuracy in that case.

5. Proof of Theorem 1: The Overloaded Regime

5.1. Sketch of the Proof

In order to derive our desired results, we must condition (and uncondition) on the possible realizations of the random variable N_λ. We classify those realizations as being either "concentrated" around the mean, $n_\lambda r$, or "far away" from this mean. Conditional on each set of realizations, we quantify the resulting error in the fluid approximation; see Lemmas 1 and 2.

Exponential Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	5.12 ±0.21	5.14 ±0.21	4.8	4.8	0.32	0.34
50	8.13 ± 0.31	8.15 ± 0.31	8	8	0.13	0.15
70	11.2 ± 0.38	11.2 ± 0.38	11.2	11.2	0.038	0.026
100	16.0 ± 0.46	16.0 ± 0.46	16	16	0.014	0.00028

Pareto Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	8.48 ±0.20	5.00 ±0.23	9.70	4.8	1.2	0.20
50	15.0 ± 0.25	8.12 ±0.33	16.2	8	1.2	0.12
70	21.7 ±0.25	11.3 ±0.36	22.6	11.2	0.90	0.097
100	31.7 ±0.27	16.0 ±0.47	32.3	16	0.68	0.013

Uniform Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	11.0 ±0.43	4.92 ±0.37	12.5	4.8	1.5	0.12
50	19.4 ±0.57	8.05 ±0.54	20.9	8	1.5	0.050
70	27.9 ± 0.60	11.1 ±0.63	29.2	11.2	1.3	0.063
100	40.7 ±0.64	16.0 ±0.76	41.7	16	1.0	0.025

Table 1 **Asymptotic accuracy of fluid performance measures in the overloaded regime in the $M/M/N+GI$ model with $N \sim Bin(n, 0.4)$ and $\rho = 1.4$.**

We exploit a concentration inequality for the binomial distribution, which shows that the number of servers asymptotically concentrates tightly around its expected value. Recall that we consider that $\rho > 1$, i.e., $\lambda/n_\lambda r > 1$ and the system is overloaded in an expected sense. Thus, the expected system's performance is asymptotically well approximated by measures which are based on the overloaded fluid model with $n_\lambda r$ servers. We then exploit existing results quantifying the "strong" asymptotic accuracy of fluid approximations with a deterministic number of servers in the over-

loaded regime (Bassamboo and Randhawa 2010). Combining the above steps establishes the strong asymptotic accuracy of fluid approximations in the overloaded regime with a binomial number of servers. Our proofs for the critically-loaded and underloaded cases proceed similarly.

5.2. Proof Details

5.2.1. $\mathcal{O}(1)$-Accuracy for the Fluid Queue Length. We begin by establishing the asymptotic $\mathcal{O}(1)$-accuracy for the expected queue length. Let $0<\epsilon<r$ and define $k_1 \equiv r-\epsilon$ and $k_2 \equiv r+\epsilon$. Assume that ϵ is small enough so that $\rho r/(r+\epsilon)>1$. Denote $\mathbb{E}[Q_{N_\lambda}|N_\lambda=s] \equiv \mathbb{E}[Q_s]$ where Q_s is the steady-state queue length in the corresponding $M/M/s+GI$ queue with the same arrival rate.

Conditioning and unconditioning on N_λ. Conditioning on N_λ, we can write:

$$\begin{aligned}
|\mathbb{E}[Q_{N_\lambda}]-rn_\lambda\bar{q}_\rho| &= \left|\sum_{s\geq 0}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda=s)-rn_\lambda\bar{q}_\rho\right| \\
&= \left|\sum_{s\geq 0}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right| \qquad \text{since } \mathbb{E}[N_\lambda]=rn_\lambda=\sum_{s\geq 0}s\mathbb{P}(N_\lambda=s), \\
&= \left|\sum_{s<k_1n_\lambda \text{ or } s>k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)+\sum_{k_1n_\lambda\leq s\leq k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right| \\
&\leq \left|\sum_{s<k_1n_\lambda \text{ or } s>k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right|+\left|\sum_{k_1n_\lambda\leq s\leq k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right|.
\end{aligned}$$

We now turn to establishing asymptotic bounds for A_λ and B_λ, defined as follows:

$$A_\lambda \equiv \left|\sum_{s<k_1n_\lambda \text{ or } s>k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right| \text{ and } B_\lambda \equiv \left|\sum_{k_1n_\lambda\leq s\leq k_2n_\lambda}(\mathbb{E}[Q_s]-s\bar{q}_\rho)\mathbb{P}(N_\lambda=s)\right|.$$

Asymptotic bound for N_λ far from $n_\lambda r$. We begin by showing that A_λ is asymptotically negligible.

LEMMA 1. $\lim_{\lambda\to\infty}A_\lambda=0$.

PROOF. We can write,

$$\begin{aligned}
A_\lambda &= \left|\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda=s)-\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}s\bar{q}_\rho\mathbb{P}(N_\lambda=s)\right|, \\
&\leq \mathbb{E}[Q_0]\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{P}(N_\lambda=s)+\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}s\bar{q}_\rho\mathbb{P}(N_\lambda=s).
\end{aligned}$$

Also, define $A_\lambda^{(1)} \equiv \mathbb{E}[Q_0]\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{P}(N_\lambda=s)$ and $A_\lambda^{(2)} \equiv \sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}s\bar{q}_\rho\mathbb{P}(N_\lambda=s)$. Note that Q_0 has the same distribution as the steady-state number in the system in an $M/GI/\infty$

model with Poisson arrivals at rate $\lambda = rn_\lambda\rho$ and i.i.d. generally distributed service times having the same distribution, F, as the abandonment times in our original model. Therefore, exploiting standard results for the infinite-server queue, Q_0 has a Poisson distribution with mean $\lambda/\theta = rn_\lambda\rho/\theta$, i.e., $\mathbb{E}[Q_0] = \mathcal{O}(\lambda)$. Applying Hoeffding's inequality to the binomial distribution: $\mathbb{P}\left(k_1 n_\lambda \le N_\lambda \le k_2 n_\lambda\right) \ge 1 - 2e^{-2\epsilon^2 n_\lambda}$; equivalently, $\mathbb{P}\left(k_1 n_\lambda > N_\lambda \text{ or } N_\lambda > k_2 n_\lambda\right) \le 2e^{-2\epsilon^2 n_\lambda}$. Thus,

$$A_\lambda^{(1)} = \mathbb{E}[Q_0] \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} \mathbb{P}(N_\lambda = s) = \mathbb{E}[Q_0] \cdot \mathbb{P}\left(k_1 n_\lambda > N_\lambda \text{ or } N_\lambda > k_2 n_\lambda\right) \to 0 \text{ as } \lambda \to \infty.$$

We now turn to showing that $A_\lambda^{(2)}$ is asymptotically negligible as well. Note that:

$$A_\lambda^{(2)} = \bar{q}_\rho \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} s\mathbb{P}(N_\lambda = s) = \bar{q}_\rho \mathbb{E}[N_\lambda \mathbb{1}\{N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda\}],$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator random variable. By the Cauchy-Schwarz inequality:

$$\begin{aligned}\mathbb{E}[N_\lambda \mathbb{1}\{N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda\}] &\le \sqrt{\mathbb{E}[N_\lambda^2]\mathbb{P}(N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda)} \\ &= \sqrt{(n_\lambda r(1-r) + n_\lambda^2 r^2)\mathbb{P}(N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda)} \to 0 \text{ as } \lambda \to \infty.\end{aligned}$$

Therefore, $A_\lambda^{(2)} \to 0$ as $\lambda \to \infty$. Combining the above, we obtain that $A_\lambda \to 0$ as well. ■

Asymptotic bound for N_λ close to $n_\lambda r$. We now characterize B_λ for large λ.

LEMMA 2. *There exists a finite constant $C > 0$ such that $\limsup_{\lambda\to\infty} B_\lambda \le C$.*

PROOF. We begin by writing B_λ as follows,

$$B_\lambda \le \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \, \mathbb{P}(N_\lambda = s) + \left| \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} s(\bar{q}_{\rho_s} - \bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|, \tag{5}$$

where $\rho_s \equiv n_\lambda r\rho/s$ and $\bar{q}_{\rho_s}$ is the fluid limit for the queue length in the $M/M/s+GI$ queue with traffic intensity ρ_s (the arrival rate is $\lambda = rn_\lambda\rho$ and the number of servers is s). Let,

$$B_\lambda^{(1)} \equiv \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \mathbb{P}(N_\lambda = s) \text{ and } B_\lambda^{(2)} \equiv \left| \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} s(\bar{q}_{\rho_s} - \bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|.$$

First, we consider $B_\lambda^{(1)}$ and show that it is asymptotically bounded. Fix n_λ and note that to each $k_1 n_\lambda \le s \le k_2 n_\lambda$ corresponds a traffic intensity ρ_s in the $M/M/s+GI$ system, where $\rho_s = n_\lambda r\rho/s$

and $1 < \rho r/(r+\epsilon) \leq \rho_s \leq \rho r/(r-\epsilon)$. By Theorem 5 of Bassamboo and Randhawa (2010), assuming that f is strictly positive and continuously differentiable,

$$\limsup_{\lambda\to\infty} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \leq \sqrt{f(\bar{w}_{\rho_s})}\left(\frac{3|f'(\bar{w}_{\rho_s})|}{\rho_s f^2(\bar{w}_{\rho_s})} + 1/2\right), \tag{6}$$

where $\bar{w}_{\rho_s}$ is the fluid limit for the steady-state waiting time in the overloaded $M/M/s+GI$ queue with traffic intensity ρ_s. Note that for $\rho r/(r+\epsilon) \leq \rho_s \leq \rho r/(r-\epsilon)$, we have that $\bar{w}_{\rho r/(r+\epsilon)} \leq \bar{w}_{\rho_s} \leq \bar{w}_{\rho r/(r-\epsilon)}$. By the continuity of the bounding function in (6) and the boundedness theorem, we conclude that there exists a finite constant $C_1 > 0$ such that

$$\sup_{k_1 n \leq s \leq k_2 n} \sqrt{f(\bar{w}_{\rho_s})}\frac{3|f'(\bar{w}_{\rho_s})|}{\rho' f^2(\bar{w}_{\rho_s})} + 1/2 \leq C_1. \tag{7}$$

Thus, $\limsup_{\lambda\to\infty} B_\lambda^{(1)} \leq C_1$. There remains to study the asymptotic behaviour of $B_\lambda^{(2)}$. Note that $\bar{q}_{\rho_s} = \rho_s \int_0^{(F^c)^{-1}(1/\rho_s)} F^c(u)du$, e.g., by equations (3.6) and (3.7) in Whitt (2006a). Consider,

$$\begin{aligned}
&\left|\sum_{s\geq 0} s\left(\rho_s \int_0^{(F^c)^{-1}(1/\rho_s)} F^c(x)dx - \rho\int_0^{(F^c)^{-1}(1/\rho)} F^c(u)du\right)\mathbb{P}(N_\lambda = s)\right| \\
&= \left|\sum_{s\geq 0}\left(n_\lambda r\rho \int_0^{(F^c)^{-1}(s/n_\lambda r\rho)} F^c(u)du - s\rho\int_0^{(F^c)^{-1}(1/\rho)} F^c(u)du\right)\mathbb{P}(N_\lambda = s)\right|, \\
&= \left|\mathbb{E}\left[\left(n_\lambda r\rho \int_0^{(F^c)^{-1}(N_\lambda/n_\lambda r\rho)} F^c(u)du - N_\lambda\rho\int_0^{(F^c)^{-1}(1/\rho)} F^c(u)du\right)\right]\right|, \\
&= \left|n_\lambda\rho r\mathbb{E}\left[\left(\int_0^{(F^c)^{-1}(N_\lambda/n_\lambda r\rho)} F^c(u)du - \int_0^{(F^c)^{-1}(1/\rho)} F^c(u)du\right)\right]\right|, \\
&= \left|n_\lambda\rho r\mathbb{E}\left[\left(\int_{(F^c)^{-1}(1/\rho)}^{(F^c)^{-1}(N_\lambda/n_\lambda r\rho)} F^c(u)du\right)\right]\right|.
\end{aligned}$$

We now show that there must exist a finite constant $C_2 > 0$ such that $\left|n_\lambda\rho r\mathbb{E}\left[\left(\int_{(F^c)^{-1}(1/\rho)}^{(F^c)^{-1}(N_\lambda/n_\lambda r\rho)} F^c(u)du\right)\right]\right| \leq C_2$ for λ large enough. To this aim, define the function

$$g_\lambda(x) = n_\lambda\rho r\int_{(F^c)^{-1}(1/\rho)}^{(F^c)^{-1}(x/n_\lambda r\rho)} F^c(u)du \text{ for } x \geq 0.$$

For a given λ, we use a Taylor-series expansion of $\mathbb{E}[g_\lambda(N_\lambda)]$ around $\mathbb{E}[N_\lambda] = n_\lambda r$ (we can do this since g_λ is sufficiently differentiable and the moments of N_λ are finite):

$$|\mathbb{E}[g_\lambda(N_\lambda)]| \approx \left|\mathbb{E}\left[g_\lambda(n_\lambda r) + g'_\lambda(n_\lambda r)\left(N_\lambda - n_\lambda r\right) + \frac{1}{2}g''_\lambda(n_\lambda r)(N_\lambda - rn_\lambda)^2\right]\right|,$$

where "$\approx$" denotes equality up to an $\mathcal{O}(1/\lambda)$ term. Indeed, by computing the centralized moments of N_λ and higher-order derivatives of g_λ, it can be shown that the remainder term in the Taylor series is $\mathcal{O}(1/\lambda)$. Also, $g_\lambda(n_\lambda r) = 0$ and

$$g'_\lambda(n_\lambda r) = -\frac{1/\rho}{f(F^c)^{-1}(1/\rho)} \text{ and } g''_\lambda(n_\lambda r) = -\frac{1}{rn_\lambda\rho}\frac{h_1(\rho)+(1/\rho)h_2(\rho)/h_1(\rho)}{h_1^2(\rho)},$$

where $h_1(\rho) = f(F^{c^{-1}}(1/\rho))$ and $h_2(\rho) = f'(F^{c^{-1}}(1/\rho))$. Thus, there exists $C_2 > 0$ such that:

$$|\mathbb{E}[g_\lambda(N_\lambda)]| \approx |\frac{1}{2}g''_\lambda(n_\lambda r)n_\lambda r(1-r)| \leq C_2 \text{ for } \lambda \text{ large enough.}$$

We now turn to the asymptotic behaviour of $B_\lambda^{(2)}$. Note that:

$$B_\lambda^{(2)} = |\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}]|, \text{ and}$$

$$|\mathbb{E}[g_\lambda(N_\lambda)]| = |\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}] + \mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}]|.$$

Bounding the second term in the last equality,

$$\begin{aligned}\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}] &\leq |\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}]| \\ &\leq \sqrt{\mathbb{E}[g_\lambda^2(N_\lambda)]}\mathbb{P}(N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]) \quad \text{(Cauchy Schwarz inequality)} \\ &\to 0,\end{aligned}$$

since $\mathbb{P}(N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda])$ vanishes exponentially fast as $\lambda \to \infty$, and $\mathbb{E}[g_\lambda^2(N_\lambda)] = \mathcal{O}(\lambda^2)$ since $\int_{(F^c)^{-1}(1/\rho)}^{(F^c)^{-1}(N_\lambda/n_\lambda r\rho)} F^c(u)du \leq 1/\theta$. Thus, $\limsup_{\lambda\to\infty} B_\lambda^{(2)} = \limsup_{\lambda\to\infty}|\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}]| \leq C_2$. Combining the above, there exists $C > 0$ such that $\limsup_{\lambda\to\infty} B_\lambda \leq C$. ■

$\mathcal{O}(1)$-*accuracy.* Since both A_λ and B_λ are asymptotically bounded, there must exist $K > 0$ such that, as desired:

$$\limsup_{\lambda\to\infty} \left|\mathbb{E}[Q_{N_\lambda}] - rn_\lambda \bar{q}_\rho\right| \leq K.$$

5.2.2. $o(1)$-Accuracy for the Fluid Net Abandonment Rate. The proof for the net abandonment rate proceeds along similar lines, so we will be brief. Paralleling (6), and denoting $\mathbb{E}[\alpha_{N_\lambda}|N_\lambda = s] \equiv \mathbb{E}[\alpha_s]$, we can exploit Theorem 5 in Bassamboo and Randhawa (2010) to show that $\sum_{k_1 n_\lambda \le s \le k_2 n_\lambda}(\mathbb{E}[\alpha_s] - s\bar{\alpha}_{\rho_s})\mathbb{P}(N_\lambda = s) \to 0$ as $\lambda \to \infty$. Moreover, by equation (3.3) in Whitt (2006a): $\bar{\alpha}_{\rho_s} = \rho_s - 1$; thus, $s(\bar{\alpha}_{\rho_s} - \bar{\alpha_\rho}) = \rho(n_\lambda r - s)$. We can then write:

$$\sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} s(\bar{\alpha}_{\rho_s} - \bar{\alpha}_\rho)\mathbb{P}(N_\lambda = s) = \rho\mathbb{E}[(nr - N_\lambda)\mathbb{1}(k_1 n_\lambda \le N_\lambda \le k_2 n_\lambda)],$$

and deduce that $\mathbb{E}[(nr - N_\lambda)\mathbb{1}(k_1 n_\lambda \le N_\lambda \le k_2 n_\lambda)] \to 0$ since $\mathbb{E}[N_\lambda] = rn_\lambda$.

6. Application: Staffing with Self-Scheduling Servers

We now use fluid approximations to solve the staffing problem with self-scheduling agents. We extend our modelling framework and consider the $G/GI/N + GI$ queueing model instead, i.e., we consider generally-distributed i.i.d. service times and a general stationary arrival process. In the appendix (Tables 4-5), we verify numerically that our asymptotic results continue to hold in this more general setting. We assume that agents select to work in one of two periods (assuming that they derive positive utility for that period). Since we have virtual call centers in mind, we assume that there is a fixed base salary $c = c_1 = c_2$ per agent, e.g., this is consistent with the widely used per-hour or per-minute fixed compensation schemes. Since we consider that there is a single customer class, irrespective of the period, we let $p = p_1 = p_2$ and $h = h_1 = h_2$; allowing unequal costs is a simple extension.

We begin by solving the problem without self-scheduling, and then describe how self-scheduling affects the optimal staffing policy. We are particularly interested in characterizing the roles of the abandonment-time distribution and strategic agent behavior. Without self-scheduling, the system manager can independently select the optimal staffing levels, n_j^*, in periods $j = 1, 2$. However, with self-scheduling, the system manager can only choose the total staffing level, n^*, and allow agents in the pool of size n^* to self schedule.

6.1. No Self-Scheduling

The density of the fluid that has been waiting for exactly u time units, in period j, is equal to $\lambda_j F^c(u)$. Therefore, the corresponding (unscaled) queue length is given by $q_j = \int_0^{w_j} \lambda_j F^c(u)du$, where w_j denotes the waiting time. The net abandonment rate (unscaled) in period j is equal to $\lambda_j F(w_j)$. In the absence of self-scheduling, we must have that $n_j^* = \lambda_j F^c(w_j^*) \leq \lambda_j$ where w_j^* is the optimal waiting time in period j; indeed, it is then suboptimal to staff more than λ_j agents in period j. The fluid approximation to the system manager's problem for period j is:

$$\min_{w_j \geq 0} \lambda_j \left((c-p)F^c(w_j) + h \int_0^{w_j} F^c(u)du \right). \tag{8}$$

Hereafter, we make the following simplifying assumption:

ASSUMPTION 1. *The density function, f, is differentiable and strictly positive on $[0,\infty)$. Additionally, the hazard-rate function, f_a, is monotonic.*

Bassamboo and Randhawa (2010) derive optimal fluid prescriptions when servers do not self-schedule; the following is a restatement of Propositions 4 and 5 of that paper.

PROPOSITION 1. *Let $j \in \{1,2\}$. Under Assumption 1, for the benchmark problem in (8):*

(i) if f_a is monotonically decreasing and there exists $w_j^* > 0$ such that $c = p + h/f_a(w_j^*)$, then it is asymptotically optimal to operate period j overloaded with $n_t^* = \lambda_t F^c(w_j^*)$;

(ii) if f_a is monotonically increasing and $c < p + h/\theta$, or if $c < p$, then it is asymptotically optimal to operate period j critically loaded, i.e., $n_j^* = \lambda_j$.

Proposition 1 shows that optimal staffing decisions generally depend on both cost parameters and the abandonment-time distribution. We also assume the following:

ASSUMPTION 2. *The staffing cost is sufficiently inexpensive, i.e., $c < \min\{h/f_a(0) + p, h/\theta + p\}$.*

Under Assumptions 1 and 2, it is easy to establish the following result for the asymptotically optimal solution to problem (8). Unless otherwise specified, Assumptions 1 and 2 are assumed to hold hereafter; in §7, we consider systems where Assumption 2 fails to hold.

PROPOSITION 2. *Under Assumptions 1 and 2, in a system with no self-scheduling servers, it is asymptotically optimal to operate each period in the* ***critically-loaded*** *regime.*

6.2. Self-Scheduling Servers

We now investigate the effect of self-scheduling on cost-minimizing staffing decisions. Let r be the probability of selecting period 1, e.g., based on (2). With a total staffing level n, let $n_1 = nr$ and $n_2 = n(1-r)$ denote the resulting staffing levels in periods 1 and 2. The net abandonment rate is given by $\lambda_j F(w_j)$. We define $m_j \equiv \lambda_j F^c(w_j)$. Note that m_j may be different from the number of servers in period j, n_j, since it may now be optimal to overstaff the system, unlike in §6.1. It is convenient to express the system manager's staffing problem in terms of m_j and n, as follows:

$$\min cn + \sum_{j=1,2} \left(p(\lambda_j - m_j) + h \int_0^{F^{c-1}(m_j/\lambda_j)} \lambda_j F^c(u) du \right) \tag{9}$$

subject to

$$m_1 = \min\{\lambda_1, nr\}; \quad m_2 = \min\{\lambda_2, n(1-r)\}; \quad m_1, m_2, n \geq 0.$$

Proposition 2 shows that, if servers do not self-schedule, then $n_j^* = \lambda_j^*$. In this case, the proportion of agents who are assigned to period 1 is, at optimum, defined to be $n_1^*/(n_1^* + n_2^*) \equiv r^*$. Thus, self-scheduling has no effect on the operational management of the system if, and only if, $r = r^*$. In this case, both periods are **critically loaded** at optimum. We are interested in investigating optimal staffing decisions for each period when $r \neq r^*$. Since fully characterising the solution of problem (9) is algebraically complicated, we focus on deriving sufficient conditions for the asymptotic optimality of different regimes instead. Motivated by the importance of the shape of the abandonment-time distribution in the results of Bassamboo and Randhawa (2010), we begin by considering abandonment-time distributions with a monotonically increasing hazard rate.

6.3. Monotonically Increasing Hazard Rate

Let $r < r^*$, i.e., agents disproportionately select period 2 over period 1. In the proof of the following proposition, we show that $c/(p+h/f_a(0)) < c/(p+h/\theta)$, i.e., the intervals on r are non-overlapping.

PROPOSITION 3. *Under Assumptions 1 and 2, if f_a is monotonically increasing and $r < r^*$,*

(i) if $r < \min\{c/(p+h/f_a(0)), r^*\}$, then it is asymptotically optimal to operate period 1 ***overloaded****, and period* 2 ***critically-loaded****;*

(ii) if $c/(p+h/\theta) < r < r^*$*, then it is asymptotically optimal to operate period* 1 ***critically-loaded****, and period* 2 ***underloaded****.*

It is optimal for the system manager to *overstaff* his system, i.e., let $n^* > \lambda_1 + \lambda_2$, if r is smaller than but relatively close to r^*, as in item (ii) of Proposition 3. In this case, the system manager ensures that period 1, which is the least preferred period, is asymptotically critically loaded. Otherwise, if r is much smaller than r^*, then it is optimal for the system manager to *understaff* his system, i.e., let $n^* < \lambda_1 + \lambda_2$; this corresponds to item (i) in Proposition 3. In this case, period 2, which is the most preferred period, is asymptotically critically loaded. The optimal solution to problem (9) for $r > r^*$ is summarized in the following proposition. Since those results are in the same vein as above (reversing the roles of the two periods), we omit discussing them.

PROPOSITION 4. *Under Assumptions 1 and 2, if* f_a *is monotonically increasing and* $r > r^*$*,*

(i) if $r^* < r < 1 - c/(p+h/\theta)$*, then it is asymptotically optimal to operate period* 2 ***critically-loaded*** *and period* 1 ***underloaded****;*

(ii) if $\max\{1 - c/(p+h/f_a(0)), r^*\} < r$*, then it is asymptotically optimal to operate period* 2 ***overloaded****, and period* 1 ***critically loaded****.*

6.4. Monotonically Decreasing Hazard Rate

For a monotonically decreasing hazard rate, we make the additional simplifying assumption.

ASSUMPTION 3. *The hazard-rate function,* f_a*, is such that* $\lim_{x\to\infty} f_a(x) = 0$.

For example, Assumption 3 is satisfied for heavy-tailed distributions, such as the Pareto or log-normal distributions. We summarize our results in Proposition 5. Interestingly, with an decreasing hazard rate, it may be optimal to operate one period overloaded, while the other period is underloaded (we show this in the proofs of Proposition 5). That is never optimal with a monotonically increasing hazard rate function, or with the exponential distribution (constant hazard rate). In the proof of the following proposition, we show that it is possible to choose r_0 sufficiently small so that $r_0 < c/(h/f_a(0)+p)$, i.e., the intervals in items (i) and (ii) are non-overlapping.

PROPOSITION 5. *Under Assumptions 1, 2, and 3, a monotonically decreasing f_a, and $r < r^*$,*

(i) there exists $r_0 < r^*$ such that if $r < r_0$, then it is asymptotically optimal to operate period 1 ***overloaded*** *and period* 2 ***critically-loaded****;*

(ii) if $c/(p+h/f_a(0)) < r < r^*$, then it is asymptotically optimal to operate period 1 ***critically-loaded*** *and period* 2 ***underloaded****,*

(iii) if $r^* < r < 1 - c/(h/f_a(0)+p)$, it is asymptotically optimal to operate period 2 ***critically loaded*** *and period* 1 ***underloaded****.*

(iv) there exists $r_1 > r^*$ such that if $r > r_1$, then it is asymptotically optimal to operate period 2 ***overloaded*** *and period* 1 ***critically-loaded****.*

6.5. Main Takeaways

6.5.1. Independence of the distribution of X_j. So far, we have expressed the optimal staffing policy as a function of the value of r. However, it is also of interest to characterise how the distributions of per-period utilities, X_j, affect the optimal staffing policy in the system. As a consequence of (2), $r = \mathbb{P}(X_1 - X_2 \geq 0 \text{ and } X_1 \geq -c)$. Assuming that $X_j \geq -c$ (i.e., each agent receives some nonnegative utility from working in either period), we can write $r = \mathbb{P}(X \geq 0)$ where $X \equiv X_1 - X_2$ is defined as the difference between the per-period utilities. This shows that *the optimal staffing policy depends on distribution of X only through its value at 0, and is independent of the individual distributions of agent utilities for either period.* In practical terms, this implies that the optimal staffing policy for the system manager can be fully specified if the *proportion* of agents who prefer one period over another is known, irrespective of other distributional assumptions. Such a proportion can be estimated in practice, e.g., by surveying agents upon hire in the system. Next, we describe the optimal staffing policy as a function of r.

6.5.2. Dependence on r. Our results for the optimal staffing policy with self-scheduling servers show that, unlike Bassamboo and Randhawa (2010), the abandonment-time distribution is *no longer the sole determinant* of the asymptotically optimal operational regime in the system. Indeed, Propositions 3-5 show that, e.g., for small or large r, it is asymptotically optimal to overload one of the periods and critically load the other, irrespective of the abandonment distribution.

At first thought, we anticipate that the system manager must always compensate for self-scheduling by *increasing* the total staffing level, leading to $n^* > \lambda_1 + \lambda_2$. Our analysis shows that this is true for values of r which are relatively close to r^*, i.e., where agent decisions are not too "extreme". In these cases, the system manager increases the staffing level to ensure that the least preferred period is asymptotically critically loaded. Consequently, the more preferred period is over-staffed, i.e., it is underloaded at optimum. Thus, self-scheduling leads to an overall increase in the level of service provided by the system manager, and customers benefit from agent self-scheduling.

However, when agents exhibit disproportionately strong preferences for one of the two periods (relative to r^*), ensuring that the least preferred period is asymptotically critically loaded would entail staffing too many agents, which would cause the other period to be significantly overstaffed. The system manager is then forced to *understaff* the least preferred period and operate it in the overloaded regime instead. The strongly preferred period is operated in the critically-loaded regime, and we have that $n^* < \lambda_1 + \lambda_2$. Thus, self-scheduling leads to an overall decrease in the level of service provided by the system manager, and customers are disadvantaged by agent self-scheduling.

7. Numerical Study: Self-Scheduling Servers

In this section, we describe results from a numerical study substantiating and extending our results from previous sections.

7.1. Asymptotic Accuracy (Theorem 2)

We begin by numerically exploring the asymptotic accuracy of fluid-based staffing prescriptions. To illustrate, we consider exponentially-distributed customer abandonment (and exponential service times). We let $c = 0.3$, $p = 0.5$, $h = 0.5$, and $\theta = 1$. In Figure 1, we let $\lambda_1 = 50$ and $\lambda_2 = 35$, and in Figure 2, we let $\lambda_1 = 200$ and $\lambda_2 = 140$. In each case, we plot the percent relative errors in the system manager's cost, i.e., $100 \cdot |\Pi^*_\lambda - \Pi_\lambda(n^*_\lambda)|/\Pi^*_\lambda$ as a function of r. In the lower subplots of each figure, we plot the corresponding fluid-based prescriptions. We present additional results, for the Pareto and uniform distributions, in the appendix (Figures 7 and 8).

Because the two periods are operating in different regimes, it is not clear what would be the resulting magnitude of error in the fluid approximation. Our numerical results show that fluid-based

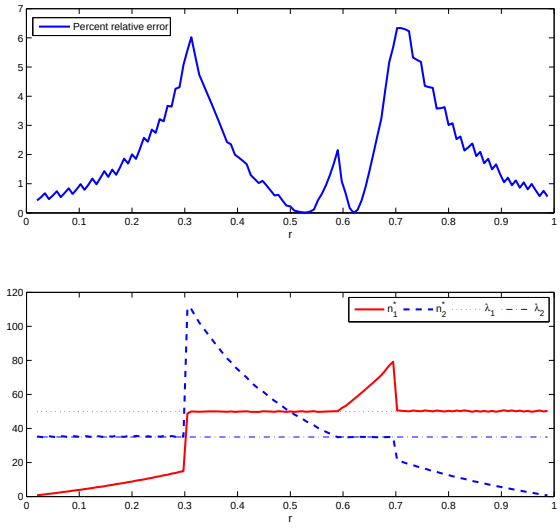

Figure 1 **Exponential abandonment in small systems ($\lambda_1 = 50, \lambda_2 = 35$).**

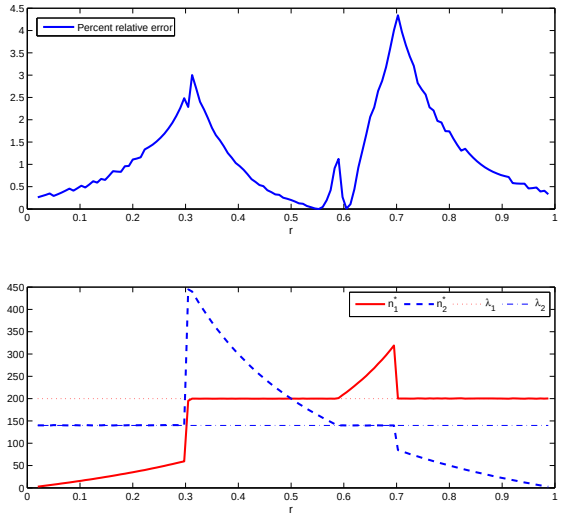

Figure 2 **Exponential abandonment in large systems ($\lambda_1 = 200, \lambda_2 = 140$).**

staffing prescriptions are remarkably accurate, as expected, with percent relative errors consistently falling below 6% in both figures. This is particularly important for the case in Figure 1, because it substantiates the usefulness of fluid approximations even in systems which are not too large. The "peaks" in the upper subplots in Figures 1 and 2 correspond to values of r at which the fluid-based optimal operational regime switches, e.g., at $r = 0.3$, it switches from overloading period 1 and critically loading period 2, to critically loading period 1 and underloading period 2. At such boundary points, the objective function in the fluid model is "flat", so that the resulting errors in fluid approximations are greater. Nevertheless, our results show that fluid-based prescriptions remain reliable, even in those extreme cases.

7.2. Fluid-Based Prescriptions (Propositions 3-5)

Increasing hazard rate. We now illustrate the results of Propositions 3 and 4 in Figure 3. There, we consider an abandonment-time distribution which is uniformly distributed over $(0, 1)$. The uniform distribution has an increasing hazard rate function. Throughout, we assume that service times are exponentially distributed. We let $c = 0.5$, $p = 0.7$, $h = 1$, $\lambda_1 = 125$, and $\lambda_2 = 75$, and present optimal staffing decisions in the system. With those parameter values, $r^* = 0.625$. Also,

$c/(p+h/f_a(0)) = 0.29$ and $c/(p+h/\theta) = 0.41$. As per Proposition 3, Figure 3 shows that: if $r < 0.29$, then period 1 is overloaded and period 2 is critically loaded. Also, if $0.41 < r < 0.625$, then period 1 is critically loaded and period 2 is underloaded (this can be seen in the figure by comparing the arrival rate value to the optimal staffing level prescribed).

Figure 3 also illustrates the optimal staffing policy for r values in the interval $c/(p + h/f_a(0), c/(p+h/\theta)$, which are not included in Proposition 3: In this case, period 1 must be overloaded as well, and period 2 remains critically loaded. We note in passing that $1 - c/(p+h/\theta) = 0.58 < r^*$; thus, the interval in case (i) of Proposition 4 is empty. Finally, consistent with case (ii) in Proposition 4, if $r > 1 - c/(p+h/f_a(0)) = 0.71$, then it is asymptotically optimal to let period 2 be overloaded and period 1 be critically loaded.

Decreasing hazard rate. We illustrate Proposition 5 in Figure 4 by numerically solving problem (9) with a Pareto abandonment-time distribution with shape parameter equal to 2 and mean equal to 1, i.e., with $F^c(x) = 1/(1+x)^2$ and $f_a(x) = 2/(1+x)$, for $x \geq 0$. The Pareto distribution has a decreasing hazard rate. We use the same remaining parameter values as in Figure 3. Figure 4 shows that, e.g., in Proposition 5, r_0 is roughly equal to 0.3 and r_1 is roughly equal to 0.6. Additionally, Figure 4 shows that for values of r ranging between 0.3 and 0.375, it is asymptotically optimal to operate period 1 overloaded concurrently with operating period 2 underloaded. Noting that $c/(p+h/f_a(0)) = 0.41$, it is easy to check that the remaining results in Proposition 5 hold. Next, we numerically investigate the optimal solution for problem (8) when Assumption 2 fails.

7.3. Optimality of the Overloaded Regime

When servers do not self-schedule, Proposition 2 shows that, under Assumptions 1 and 2, it is asymptotically optimal to operate both periods in the critically-loaded regime. However, if Assumption 2 does not hold, then this may no longer be the case. In particular, it may be asymptotically optimal to operate both periods in the overloaded regime instead. In what follows, we consider an example of such a system. Our objective is to investigate the optimal staffing policy in this case.

By Proposition 2, it is asymptotically optimal to operate a period overloaded if f_a is monotonically decreasing and there exists $w > 0$ such that $c = p + h/f_a(w)$. In Figures 5 and 6, we present

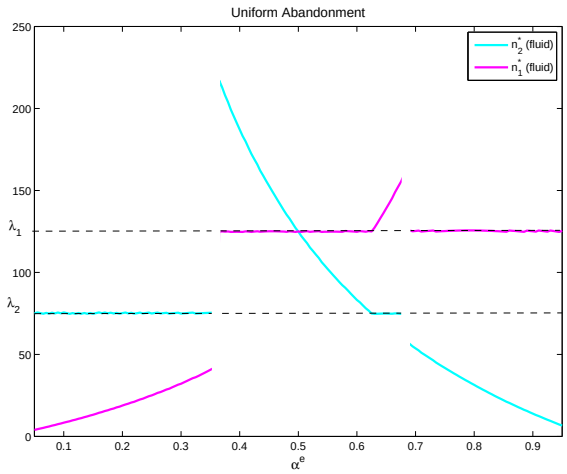

Figure 3 **Numerical solution for problem (9) for uniform (0,1) abandonment.**

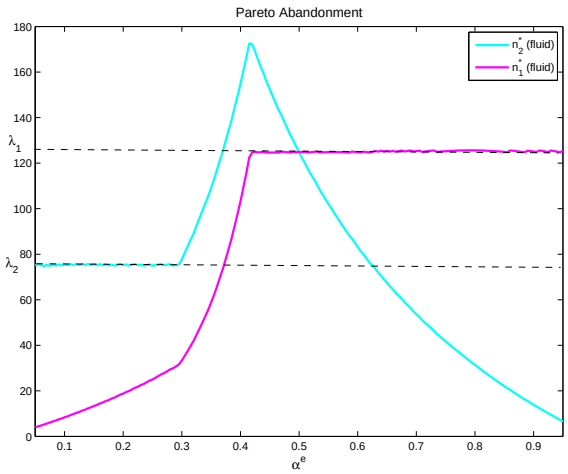

Figure 4 **Numerical solution for problem (9) for Pareto abandonment with shape 2.**

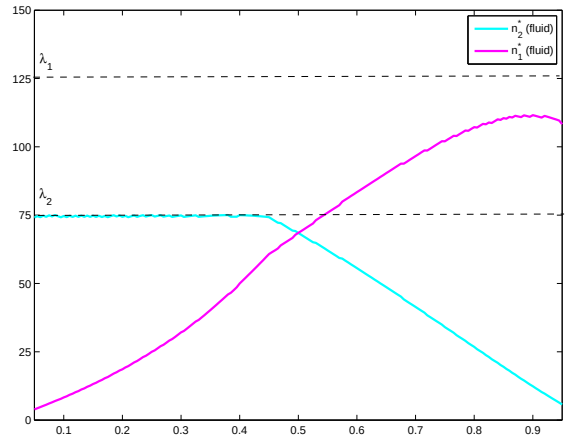

Figure 5 **Solution for problem (9) for Pareto abandonment with shape 2 and $c = 1.3$.**

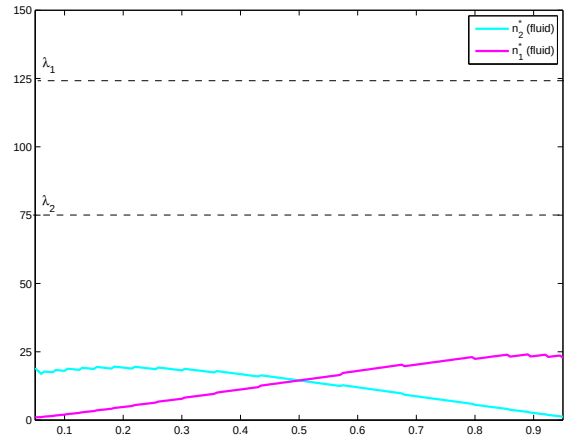

Figure 6 **Solution for problem (9) for Pareto abandonment with shape 2 and $c = 2$.**

the optimal solution to problem (9) with a Pareto abandonment-time distribution with shape parameter equal to 2 and mean equal to 1. We vary the value of c so as to violate the condition $c < p + h/f_a(0)$, i.e., to violate Assumption 2; otherwise, we consider the same parameter values as in Figure 4. If Assumption 2 does not hold, then it is easy to check that there must exist $w > 0$ such that $c = p + h/f_a(w)$. In this case, if servers do not self schedule, then it is asymptotically optimal to operate both periods in the overloaded regime. In Figure 5, we let $c = 1.3$ and in Figure 6, we

let $c = 2$. It can be easily verified that $c > p + h/f_a(0)$ for those two values of c. Figures 5 and 6 illustrate the optimal staffing policy. In particular, in Figure 5: irrespective of r, it is asymptotically optimal to operate period 1 overloaded. Similarly, Figure 6 shows that, for c large enough, both periods 1 and 2 are asymptotically overloaded at optimum, irrespective of r.

8. Conclusions

In this paper, we studied the staffing problem in large-scale service systems with a binomial number of servers. The randomness in the number of servers follows from strategic agent behavior, which has recently been gaining increased attention from practitioners and academics alike.

Our theoretical results support the usefulness of fluid approximations in analyzing systems where servers are strategic. As an application, we studied a system manager's problem of managing a service system with self-scheduling agents. We showed that optimal staffing decisions lead to different asymptotically optimal operational regimes in the system, depending on *both* self-scheduling behavior and the abandonment-time distribution. In particular, we demonstrated that the optimal staffing policy is not straightforward, and that it may be optimal to either understaff or overstaff the system. This, interestingly, shows that customers may either be disadvantaged (former case) or benefit (latter case) from self-scheduling.

Queueing systems with a random number of servers have not been sufficiently studied so far, and this paper constitutes one of recent efforts taken towards narrowing that gap in the literature. Further exploration of the dynamics of such systems is of interest for future research. In this paper, we focused solely on describing expected performance measures in the system. However, we did not investigate the variability of the system's performance measures about their expected values. We also did not establish supporting many-server heavy-traffic limits for different stochastic processes in the system, such as the queue length (corresponding to a functional law of large numbers). Such an investigation would enable a deeper understanding of the system's dynamics.

We also focused exclusively on the long-term decision of determining the total pool of agents to hire. However, we did not consider medium or short-term controls that the system manager could

use in order to incite the right number of agents, of the pool of size n (assumed fixed), to work in different time periods of a day. For example, it would be interesting to investigate how different compensation schemes may affect the numbers of agents available, and what would be the correct incentive to use so as to incite enough agents to participate, at both diffusion and fluid scales. Of particular interest are compensation schemes which are based on the workload, such as in surge pricing, since these are already used in practice, e.g., at Uber or at Zappos (Fortune 2015).

References

Akşin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Arise. 2015. Flexibility. URL `http://www.arise.com/Arise-Value/Flexibility`.

Atar, R. 2008. Central limit theorem for a many-server queue with random service rates. *The Annals of Applied Probability* **18**(4) 1548–1568.

Baccelli, F., P. Boyer, G. Hebuterne. 1984. Single-server queues with impatient customers. *Advances in Applied Probability* 887–905.

Bassamboo, A., M. J. Harrison, A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* **51**(3-4) 249–285.

Bassamboo, A., M. J. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research* **54**(3) 419–435.

Bassamboo, A., R. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research* **58**(5) 1398–1413.

Bassamboo, A., R. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

BBC. 2015. Hospital staff absences for mental health reasons double. URL `http://www.bbc.co.uk/news/uk-england-32022114`.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations research* **52**(1) 17–34.

Cachon, G., P. Harker. 2002. Competition and outsourcing with scale economies. *Management Science* **48**(10) 1314–1333.

Cachon, G., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science* **53**(3) 408–420.

Cloud Source. 2015. Cloud source employment details. URL `http://www.sutherlandcloudsource.com/SGS-Work-At-Home-Employee-Benefits.aspx`.

Dai, JG, Shuangchi He. 2012. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering* **21**(1) 1–36.

Fast Company. 2015. How uber changed the way they hire. URL `http://www.fastcompany.com/3028390/bottom-line/how-uber-changed-the-way-they-hire`.

Fortune. 2015. Zappos is bringing uber-like surge pay to the workplace. URL `http://fortune.com/2015/01/28/zappos-employee-pay/`.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5** 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208–227.

Gopalakrishnan, R., S. Doroudi, A. Ward, A. Wierman. 2015. Routing and staffing when servers are strategic. *arXiv preprint arXiv:1402.3606* .

Great Virtual Works. 2015. Flexible home based phone job. URL `http://www.dreamhomebasedwork.com/2014/04/great-virtual-works.html/`.

Green, L., S. Savin, N. Savva. 2013. nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237–2256.

Gurvich, I., M. Lariviere, T. Moreno-Garcia. 2015. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Northwestern University, working paper.

Hall, J., A. Krueger. 2015. An analysis of the labor market for ubers driver-partners in the united states.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management* **7**(1) 20–36.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer Science & Business Media.

LiveOps. 2015. Work from home with liveops. URL `http://join.liveops.com/`.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Uber. 2015. What is surge pricing? URL `https://help.uber.com/h/6c8065cf-5535-4a8b-9940-d292ffdce119`.

US Bureau of Labor Statistics. 2008. Industry injury and illness data. URL `http://www.bls.gov/iif/oshwc/osh/os/osnr0032.pdf`.

Ward, A., P. Glynn. 2003. A diffusion approximation for a markovian queue with reneging. *Queueing Systems* **43**(1-2) 103–128.

West at Home. 2015. West at home: Become an agent. URL `http://www.apply.westathome.com/index.html`.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Operations Research* **54** 37–54.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15** 88–102.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems: Theory and Applications* **51**(3-4) 361–402.

Zhan, D., A. Ward. 2015. Compensation and staffing to trade off speed and quality in large service systems University of Southern California, working paper.

Appendix

A. Proof of Theorem 1

A.1. The underloaded regime

PROOF. Let $0<\epsilon<r$ be small enough so that $\rho r/(r-\epsilon)<1$, and recall that $k_1 \equiv r-\epsilon$ and $k_2 \equiv r+\epsilon$. Then, conditioning on N_λ:

$$\begin{aligned}\mathbb{E}[Q_{N_\lambda}] &= \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \sum_{k_1 n_\lambda > s \text{ or } s > k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s),\\ &\le \sum_{k_1 n \le s \le k_2 n} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \mathbb{E}[Q_0] \sum_{k_1 n_\lambda > s \text{ or } s > k_2 n_\lambda} \mathbb{P}(N_\lambda = s).\end{aligned}$$

As in the proof of Theorem 1, we can show that: $\mathbb{E}[Q_0]\sum_{k_1 n > s \text{ or } s > k_2 n}\mathbb{P}(N=s) \to 0$ as $\lambda \to \infty$. Also, $\sum_{k_1 n \le s \le k_2 n}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \le \mathbb{E}[Q(k_1 n_\lambda)]\sum_{k_1 n \le s \le k_2 n}\mathbb{P}(N=s)$. Since $\mathbb{E}[Q(k_1 n_\lambda)]$ is the expected steady-state queue length in an underloaded queue, it converges to 0 as $\lambda \to \infty$, e.g, see Theorem 5.1 in Zeltyn and Mandelbaum (2005). The limit for the net abandonment follows similarly. ■

A.2. The Critically-Loaded Regime

PROOF. We condition on N_λ:

$$\begin{aligned}\mathbb{E}[Q_{N_\lambda}] &= \sum_{k_1 n_\lambda \le s < n_\lambda r} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \mathbb{E}[Q_{n_\lambda r}]\mathbb{P}(N_\lambda = n_\lambda r),\\ &\le \sum_{k_1 n_\lambda \le s < n_\lambda r} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) + \sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s)\\ &\quad + \mathbb{E}[Q(n_\lambda r)]\mathbb{P}(N_\lambda = n_\lambda r), \qquad (10)\end{aligned}$$

where $\rho_s = \lambda\rho n_\lambda/s$. Paralleling (6) and (7), we can show that there exists a finite constant C_1' such that for large λ: $\sum_{k_1 n_\lambda \le s < nr}|\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) \le C_1'$ since $\rho_s > 1$ for all $k_1 n_\lambda \le s < n_\lambda r$. Also,

$$\begin{aligned}\sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) &= \sum_{k_1 n_\lambda \le s < n_\lambda r} n_\lambda r \left(\int_0^{(F^c)^{-1}(s/n_\lambda r)} F^c(x)dx\right)\mathbb{P}(N_\lambda = s) \qquad (11)\\ &= \mathbb{E}\left[\left(n_\lambda r \int_0^{(F^c)^{-1}(N/n_\lambda r)} F^c(x)dx\right)\mathbb{1}(N_\lambda \in [k_1 n_\lambda, n_\lambda r))\right].\end{aligned}$$

Using arguments as in Theorem 1, we can show that there exists a finite $C_2' > 0$ such that

$$\limsup_{\lambda\to\infty}\mathbb{E}\left[\left(n_\lambda r \int_0^{(F^c)^{-1}(N/n_\lambda r)} F^c(x)dx\right)\mathbb{1}(N_\lambda \in [k_1 n_\lambda, n_\lambda r))\right] \le C_2'.$$

By Theorem 4.1 of Zeltyn and Mandelbaum (2005), there exists $K' > 0$ such that $\mathbb{E}[Q_{n_\lambda r}] \le K'\sqrt{\lambda}$ for large enough λ. Given that $\sum_{n_\lambda r < s \le k_2 n_\lambda}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \to 0$ as $\lambda \to \infty$ (underloaded regime), we obtain that the entire expression in (10) is $\mathcal{O}(\sqrt{\lambda})$. The proof for the abandonment rate follows along similar lines, so we omit the relevant details. ■

B. Proof of Proposition 2

Based on item (ii) in Proposition 1: if f_a is monotonically increasing and $c < p + h/\theta$, then it is asymptotically optimal to operate period j in the critically-loaded regime. There remains to establish that a similar conclusion holds if f_a is monotonically decreasing, provided that Assumption 2 holds. The first-order derivative of the objective in (8) is given by $\lambda_j F^c(w_j)(h-(c-p)f_a(w_j))$. If $c < h/f_a(0)+p$, and f_a is monotonically decreasing, then $c < h/f_a(w_j)+p$ for any $w_j > 0$. In other words, the objective is increasing in w_j. Thus, the optimal solution is to set $w_j^* = 0$, i.e., let period j be critically loaded.

C. Proof of Propositions 3 and 4

In any optimal solution to problem (9), we must have that $n = \max\{m_1/r, m_2/(1-r)\}$. This is so because the objective function is increasing in n (assuming that $c > 0$) and so it is never optimal to set n strictly larger than $\max\{m_1/r, m_2/(1-r)\}$, which satisfies the remaining constraints. Our objective is to study the solution space to problem (9). To this aim, we consider two cases: 1. $n = m_1/r$ at optimum, and 2. $n = m_2/(1-r)$ at optimum. Recall that $r^* = \lambda_1/(\lambda_1+\lambda_2)$.

1. $\underline{n = m_1/r}$. Note that we must then have that $n_1 = nr = m_1 \le \lambda_1$, which means that **period** 1 **is either critically loaded or overloaded**. Also note that we must have that $n(1-r) \ge m_2$. We consider two subcases: (a) $m_2 < n(1-r)$ at optimum and (b) $m_2 = n(1-r)$ at optimum.

   (a) $\underline{n = m_1/r \text{ and } m_2 < n(1-r)}$. If $m_2 < n(1-r)$, then $m_2 = \lambda_2 < n(1-r)$. That is, **period** 2 **must be underloaded** in this optimal solution. We then obtain that:

   $$\frac{\lambda_2}{1-r} = \frac{m_2}{1-r} < n = \frac{m_1}{r} \le \frac{\lambda_1}{r}.$$

   This implies that we must have that $r < r^*$ in an optimal solution where $n = m_1/r$ and $m_2 < n(1-r)$.

   (b) $\underline{n = m_1/r \text{ and } m_2 = n(1-r)}$. In this case, we must also consider two subcases: (i) $m_2 = n(1-r) = \lambda_2$ and (ii) $m_2 = n(1-r) < \lambda_2$.

      i. $\underline{n = m_1/r \text{ and } m_2 = n(1-r) = \lambda_2}$. In this case, we must have that 2 **is critically loaded**. Then,

      $$\frac{\lambda_1}{r} \ge \frac{m_1}{r} = n = \frac{\lambda_2}{1-r}.$$

      This implies that $r \le r^*$, which must hold if $n = m_1/r$ and $m_2 = n(1-r) = \lambda_2$ holds at optimum.

ii. $\underline{n = m_1/r \text{ and } m_2 = n(1-r) < \lambda_2}$. Then, we must consider two new subcases: 1. $nr = m_1 = \lambda_1$ and $m_2 = n(1-r) < \lambda_2$ and 2. $nr = m_1 < \lambda_1$ and $m_2 = n(1-r) < \lambda_2$.

A. $\underline{nr = m_1 = \lambda_1 \text{ and } m_2 = n(1-r) < \lambda_2}$. This means that 1 **is critically loaded and** 2 **is overloaded**. In this case, $n = \lambda_1/r$. This implies that

$$n(1-r) = \frac{\lambda_1(1-r)}{r} < \lambda_2.$$

This implies that $r > r^*$.

B. $\underline{nr = m_1 < \lambda_1 \text{ and } m_2 = n(1-r) < \lambda_2}$. This means that **both** 1 **and** 2 **are overloaded**. We now show that under the assumption that $c/(h/f_a(0)+p) < 1$ (Assumption 2), this case cannot happen.

To prove that this case cannot happen, we consider solutions where both periods are overloaded, and find the optimal waiting times $w_1 \geq 0$ and $w_2 \geq 0$ in this case. That is, we assume that we have all servers busy in 1, i.e., $m_1 = nr = n_1$, and all servers busy in 2, i.e., $m_2 = n(1-r) = n_2$, and show that, under our initial assumption that $c/(h/f_a(0)+p) < 1$, we cannot have that at optimum both $w_1 > 0$ and $w_2 > 0$. We are interested in non-trivial solutions where w_1 and w_2 are not both infinite, which corresponds to not staffing any servers. Since we seek solutions where all servers in both periods are busy, then we must have that $nr = n_1 = m_1 = \lambda_1 F^c(w_1)$ and $n(1-r) = n_2 = m_2 = \lambda_2 F^c(w_2)$. In other words, we need to solve the following problem:

$$\begin{aligned} \min_{w_1 \geq 0, w_2 \geq 0} \; & c(\lambda_1 F^c(w_1) + \lambda_2 F^c(w_2)) \\ & + p(\lambda_1 F(w_1) + \lambda_2 F(w_2)) \\ & + h\left(\int_0^{w_1} \lambda_1 F^c(u)du + \int_0^{w_2} \lambda_2 F^c(u)du\right) \end{aligned} \tag{12}$$

subject to

$$(1-r)\lambda_1 F^c(w_1) = r\lambda_2 F^c(w_2).$$

The constraint in problem (12) is due to the fact that $n = n_1/r = \lambda_1 F^c(w_1)/r = n_2/(1-r) = \lambda_2 F^c(w_2)/(1-r)$. In order to solve this problem, we use the method of direct substitution. That is, we substitute w_2 in problem (12) by the following expression:

$$w_2 = (F^c)^{-1}\left(\frac{1-r}{r}\frac{\lambda_1}{\lambda_2}F^c(w_1)\right).$$

The problem that we obtain (ignoring the constant $p\lambda_1 + p\lambda_2$ in the objective) is the following:

$$\min_{w_1 \geq 0} (c-p)\frac{\lambda_1 F^c(w_1)}{r} + h\left(\int_0^{w_1} \lambda_1 F^c(u)du + \int_0^{(F^c)^{-1}\left(\frac{1-r}{r}\frac{\lambda_1}{\lambda_2}F^c(w_1)\right)} \lambda_2 F^c(u)du\right). \tag{13}$$

The derivative of the objective function is:

$$\text{derivative } = \lambda_1 h F^c(w_1)\left(1 - \frac{c-p}{rh}f_a(w_1) + \left(\frac{1-r}{r}\right)^2 \frac{\lambda_1}{\lambda_2}\frac{f(w_1)}{f\left((F^c)^{-1}\left(\frac{(1-r)\lambda_1}{r\lambda_2}F^c(w_1)\right)\right)}\right). \tag{14}$$

If $f_a(\cdot)$ is increasing, then the possible optimal solutions are at the boundary, i.e., either $w_1 = 0$ or $w_1 = \infty$ (which corresponds to no staffing at all). Recall that we are investigating whether it is possible to have a non-trivial solution where $w_1 > 0$ at optimum. So, that is not possible if $f_a(\cdot)$ is increasing.

Now, let us assume that $f_a(\cdot)$ is decreasing. Then, a minimizing w_1 must satisfy the first-order condition:

$$0 = 1 - \frac{c-p}{rh}f_a(w_1) + \left(\frac{1-r}{r}\right)^2 \frac{\lambda_1}{\lambda_2}\frac{f(w_1)}{f\left((F^c)^{-1}\left(\frac{(1-r)\lambda_1}{r\lambda_2}F^c(w_1)\right)\right)}.$$

By some algebra, and replacing $(F^c)^{-1}\left(\frac{1-r}{r}\frac{\lambda_1}{\lambda_2}F^c(w_1)\right)$ by w_2, we obtain that we must have at optimum:

$$\frac{1}{f_a(w_1)} + \frac{1-r}{r}\frac{1}{f_a(w_2)} = \frac{c-p}{rh}.$$

Recall that f_a is assumed to be decreasing. Thus, $\frac{1}{f_a(w_1)} + \frac{1-r}{r}\frac{1}{f_a(w_2)}$ is increasing in both w_1 and w_2. So, if we impose that its value at $w_1 = w_2 = 0$ to be larger than or equal to $\frac{c-p}{rh}$, then there cannot exist $w_1 > 0$ and $w_2 > 0$ which satisfy this equation, implying that we cannot have $w_1 > 0$ and $w_2 > 0$ concurrently at optimum; i.e., we cannot have both periods 1 and 2 overloaded at optimum. We need to impose that:

$$\frac{1}{f_a(0)} + \frac{1-r}{r}\frac{1}{f_a(0)} > \frac{c-p}{rh}.$$

It is not hard to see that this is equivalent to:

$$\frac{c}{\frac{h}{f_a(0)}+p}<1,$$

which is the condition that we impose in Assumption 2. Thus, we cannot have a non-trivial solution where both 1 and 2 are overloaded at optimum.

2. $\underline{n=m_2/(1-r)}$. In this case, we must have that 2 **is critically-loaded or overloaded**. We must also have that $n\geq m_1/r$ and $m_1\leq\lambda_1$. We consider two subcases: (a) $nr=m_1$ and (b) $nr>m_1=\lambda_1$.

   (a) $\underline{n=m_2/(1-r) \text{ and } nr>m_1=\lambda_1}$. Since $nr>\lambda_1$, we must have that 1 **is underloaded**. In this case, $n(1-r)=m_2$ implies that $\lambda_2/(1-r)\geq n=m_2/(1-r)>m_1/r=\lambda_1/r$, which implies that $r>r^*$.

   (b) $\underline{n=m_2/(1-r) \text{ and } nr=m_1}$. We must then consider two subcases: (i) $n=m_2/(1-r)$ and $nr=m_1=\lambda_1$ and (ii) $n=m_2/(1-r)$ and $nr=m_1<\lambda_1$.

      i. $\underline{n=m_2/(1-r) \text{ and } nr=m_1=\lambda_1}$. In this case, 1 **must be critically loaded**. Then, $nr=\frac{m_2r}{1-r}=\lambda_1\leq\frac{\lambda_2r}{1-r}$. This implies that $r\geq r^*$.

      ii. $\underline{n=m_2/(1-r) \text{ and } nr=m_1<\lambda_1}$. In this case, we consider two new subcases: 1. $n(1-r)=m_2=\lambda_2$ and $nr=m_1<\lambda_1$ and 2. $n(1-r)=m_2<\lambda_2$ and $nr=m_1<\lambda_1$.

         A. $\underline{n(1-r)=m_2=\lambda_2 \text{ and } nr=m_1<\lambda_1}$. In this case, we must have that 2 **is critically loaded and** 1 **overloaded**. But, since $nr<\lambda_1$, $nr=\frac{m_2r}{1-r}=\frac{\lambda_2r}{1-r}<\lambda_1$, which implies that $r<r^*$.

         B. $\underline{n(1-r)=m_2<\lambda_2 \text{ and } nr=m_1<\lambda_1}$. In this case, **both periods** 1 **and** 2 **are overloaded**, which we demonstrated above cannot happen under our initial assumption on $f_a(0)$.

We are now ready to summarize the entire solution space to problem (9), and corresponding necessary values on r. Based on the analysis above, we obtain the following partition of the solution space:

From Table 2, we see that if $r<r^*$, then it is not possible to have period 2 overloaded or that period 1 is underloaded. Also, if $r>r^*$, then it is not possible to have period 1 overloaded or that period 2 is underloaded. And, it is not hard to see that if both periods are critically loaded at optimum then we must have that $n=\lambda_1/r=\lambda_2/(1-r)$ which implies that $r=r^*$. We now go further and express our problem in ways that will be useful to proving the remaining propositions of the section.

Period 1	**Period** 2	**Necessary conditions on** r
critically loaded, overloaded	underloaded	$r < r^*$
critically loaded, overloaded	critically loaded	$r \leq r^*$
critically loaded	overloaded	$r > r^*$
underloaded	critically loaded, overloaded	$r > r^*$
critically loaded	critically loaded, overloaded	$r \geq r^*$
overloaded	critically loaded	$r < r^*$

Table 2 Solution space of problem (9).

If $r < r^*$, then we must have that $m_2 = \lambda_2$ in the solution to problem (9). Additionally, all servers in 1 are busy so that $n_1 = m_1 = nr$ and $n = m_1/r$. If $r > r^*$, then we must have that $m_1 = \lambda_1$ and that $n_2 = n(1-r) = m_2$ so that $n = m_2/(1-r)$. So, if $r < r^*$, then finding the optimal solution to problem (9) amounts to solving the following problem:

$$\min_{(F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right) \geq w_1 \geq 0} p\lambda_1 + \lambda_1 \left(\frac{c}{r} - p\right) F^c(w_1) + \int_0^{w_1} \lambda_1 h F^c(u) du. \tag{15}$$

The upper bound on w_1 is because n is at least $\lambda_2/(1-r)$ since 2 cannot be overloaded.

If $r > r^*$, then finding the optimal solution to problem (9) amounts to solving the following problem:

$$\min_{(F^c)^{-1}\left(\frac{(1-r)\lambda_1}{r\lambda_2}\right) \geq w_2 \geq 0} p\lambda_2 + \lambda_2 \left(\frac{c}{1-r} - p\right) F^c(w_2) + \int_0^{w_2} \lambda_2 h F^c(u) du. \tag{16}$$

The upper bound on w_2 is because n is at least λ_1/r since 1 cannot be overloaded.

We assume that $f_a(\cdot)$ has increasing hazard rate. For Proposition 3, we restrict attention to $r < r^*$ and solve problem (15). For Proposition 4, we restrict attention to $r > r^*$ and solve problem (16).

Proof of Proposition 3. We begin with the case $r < r^*$. We focus on case (i) in Proposition 3. Then, differentiating the objective of problem (15), we obtain:

$$\text{derivative } = \lambda_1 h F^c(w_1) \left(1 - \frac{c/r - p}{h} f_a(w_1)\right).$$

With an increasing hazard rate, we see that the only possible solutions for problem (15) are at the boundary, i.e., we must have that either $w_1 = 0$ or $w_1 = (F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right)$. Now, if we assume that the derivative of the objective is negative at 0, then the objective function must be decreasing on its entire domain. Thus, the optimal solution to problem (15) must be at $w_1 = (F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right)$. At this boundary point, we must have

that 2 is critically loaded and 1 is overloaded (since $w_1 > 0$ and $n(1-r) = (\lambda_1 F^c(w_1)/r)(1-r) = \lambda_2$). There remains to specify conditions on r for which the derivative of the objective is negative at 0. It is not hard to see that this amounts to $1-(c/r-p)f_a(0)/h < 0$ which is equivalent to $\underline{r < c/(h/f_a(0)+p)}$, which under Assumption 2 is assumed to be strictly less than 1.

For case (ii), we need to impose initial conditions such that the optimal solution to problem (15) is at $w_1 = 0$. For this, given the shape of the objective function under the assumption that $f_a(\cdot)$ is increasing, it suffices to show that the objective value at 0 is smaller than the objective value at $w_1 = \infty$. We obtain:

$$\text{value of objective at } 0 = p\lambda_1 + \lambda_1(c/r - p),$$

$$\text{value of objective at } \infty = p\lambda_1 + \lambda_1 h/\theta,$$

using the fact that $\int_0^\infty F^c(u)du = 1/\theta$. Thus, imposing that the objective is minimized at 0 is equivalent to imposing that $\underline{r > c/(h/\theta+p)}$ (while also imposing that $r < r^*$). In this case, 1 is critically loaded at optimum, and 2 is underloaded.

Non-overlapping intervals. There remains to check that the intervals for r specified in parts (i) and (ii) are non-overlapping. For that, note that for distributions with increasing hazard rates, the mean residual lifetime (MRL) must be decreasing. Let m_t denote the mean residual lifetime. Then, it is well-known that:

$$f_a(t) = \frac{m_t' + 1}{m_t}.$$

Now, plug in 0: $f_a(0) = \frac{m_0'+1}{m_0}$; but $m_0' < 0$ since the MRL is decreasing. Thus, we must have that $f_a(0) < \theta$ since $m_0 = 1/\theta$. This implies that $c/(h/\theta+p) > c/(h/f_a(0)+p)$.

Proof of Proof of Proposition 4. We now restrict attention to $\underline{r > r^*}$, and solve problem (16). We focus on case (i). We begin by differentiating the objective of the problem:

$$\text{derivative } = \lambda_2 h F^c(w_2)\left(1 - \frac{c/(1-r)-p}{h} f_a(w_2)\right).$$

With an increasing hazard rate, the optimum value of problem (16) must occur at the boundary, i.e., we must have that either $w_2 = 0$ or $w_2 = (F^c)^{-1}\left(\frac{(1-r)\lambda_1}{r\lambda_2}\right)$. If we impose that the value of the objective is smaller at $w_2 = 0$ than at $w_2 = \infty$, then it must be that this function is minimized at 0 . We have that:

$$\text{value of objective at } 0 = p\lambda_2 + \lambda_2(c/(1-r) - p),$$

$$\text{value of objective at } \infty = p\lambda_2 + \lambda_2 h/\theta,$$

Thus, the objective is smaller at 0 if, and only if: $\underline{r < 1 - c/(h/\theta + p)}$. For these values of r, we have that 2 is critically loaded and 1 is underloaded.

For case (ii), it suffices to impose that the derivative of the objective is negative at 0. This amounts to imposing that $\underline{r > 1 - c/(p + h/f_a(0))}$. Finally, we need to ensure that the intervals in cases (i) and (ii) are non-overlapping. This follows directly from the MRL argument used above.

D. Proof of Proposition 5

We now assume that f_a is decreasing. We begin by considering $\underline{r < r^*}$, and solving problem (15). Recall that the derivative of the objective is given by:

$$\text{derivative } = \lambda_1 h F^c(w_1) \left(1 - \frac{c/r - p}{h} f_a(w_1)\right).$$

Then, it is readily seen that with a decreasing hazard rate, we may have an interior optimal solution $w^* > 0$ if there exists such w^* such that $1 - \frac{c/r-p}{h} f_a(w^*) = 0$. Under our assumption, $c < p + h/f_a(0)$. Now, pick $w' > 0$. Then, we must have that $c < p + h/f_a(w')$ since $f_a(w') < f_a(0)$. Thus, there must exist $r' \in (0,1)$ such that $c/r' = p + h/f_a(w')$, in particular: $0 < r' = c/(p + h/f_a(w')) < 1$. Since f_a is continuous and heavy-tailed, it must be that $\lim_{x\to\infty} f_a(x) = 0$. Then, there exists $w'' > 0$ such that $f_a(w'') = y$ for every $y < h/(c/r' - p)$. In particular, for every $r'' \le r'$, let $y = h/(c/r'' - p) \le h/(c/r' - p)$. Then, there must exist $w'' > 0$ such that $f_a(w'') = h/(c/r'' - p)$ i.e., that $(\frac{c}{r''} - p) f_a(w'') = h$. We must also have that $w'' > w'$ since f_a is decreasing. In this case, for all $r \le r'$, we must have that 1 is overloaded at optimum. Denote by w_1^* the optimal wait time. If $w_1^* < (F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right)$, then 2 is underloaded at optimum. Otherwise, we have that 2 is critically loaded at optimum. By choosing r small enough, we can ensure that $w > (F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right)$. This can be seen by considering an r such that $r' > r \to 0$. Then, there exists w such that $c/r = h/f_a(w) + p$, and this $w \to \infty$. So, eventually, w_1^* will exceed $(F^c)^{-1}\left(\frac{r\lambda_2}{(1-r)\lambda_1}\right)$. Thus, there exist an r_0 such as in case (i).

We now prove case (ii). If we impose that the derivative of the objective is positive at 0, then we guarantee that the objective is minimized at $w_1 = 0$. In this case, 1 is critically loaded and 2 is underloaded. Imposing that the derivative of the objective is positive at 0 is equivalent to imposing that: $\underline{r > c/(h/f_a(0) + p)}$ (recall that we assume $r < r^*$ as well).

We now consider $\underline{r > r^*}$, and solving problem (16). Recall that the derivative of the objective is given by:

$$\text{derivative } = \lambda_2 h F^c(w_2) \left(1 - \frac{c/(1-r) - p}{h} f_a(w_2)\right).$$

If we impose that this derivative is positive at the origin, then we guarantee that the objective is minimized at $w_2 = 0$, for which 2 is critically loaded and 1 is underloaded. Imposing that the derivative of the objective is positive at 0 is equivalent to imposing that: $r < 1 - c/(h/f_a(0) + p)$ (recall that we assume that $r > r^*$). This proves case (iii). We now turn to proving case (iv). It is readily seen that with a decreasing hazard rate, we may have an interior optimal solution $w^* > 0$ if there exists such w^* such that $1 - \frac{c/(1-r)-p}{h} f_a(w^*) = 0$. In this case, we must have that 2 is overloaded since $w_2 = w^* > 0$. If $w^* < (F^c)^{-1}\left(\frac{(1-r)\lambda_1}{r\lambda_2}\right)$, then 1 is underloaded at optimum. Otherwise, we have that 1 is critically loaded at optimum. The proof follows from here by arguments very similar to case (i) in Proposition 5.

E. Additional Numerical Results

E.1. Critically-loaded and Underloaded Regimes: Theorem 1

The results for the asymptotic accuracy of fluid performance measures in the critically-loaded and underloaded regimes with an exponential distribution are summarized in Table 3. These results substantiate the results of Theorem 1.

E.2. Generally-distributed Service Times: Theorem 1

In Tables 4 and 5, we illustrate that our asymptotic results of Theorem 1 continue to hold with non-exponential service times as well: There, we consider lognormal service times with mean 1 and variance $e-1$. We consider a lognormal distribution because service times in practice have been shown to be well-modelled by lognormal distribution.

E.3. Alternative Abandonment-Time distributions: Theorem 2

The results for the asymptotic accuracy of fluid-based staffing prescriptions for the uniform on (0,2) and Pareto distribution with shape parameter 2 and mean equal to 1 are given in Figures 7 and 8. The remaining parameters are as for the exponential distribution in Figures 1 and 2. As can be seen from the figures, the numerical results obtained substantiate Theorem 2.

Critically loaded ($\rho = 1$)

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N] - rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N] - rn\bar{\alpha}\rvert$
30	1.74 ±0.14	1.78 ±0.14	0	0	1.7	1.8
50	2.30 ±0.18	2.34 ±0.18	0	0	2.3	2.3
70	2.61 ±0.19	2.66 ±0.19	0	0	2.6	2.7
100	3.22 ±0.25	3.27 ±0.25	0	0	3.2	3.3

Underloaded regime ($\rho = 0.85$)

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N] - rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N] - rn\bar{\alpha}\rvert$
300	0.513 ±0.084	0.578 ±0.087	0	0	0.51	0.58
500	0.230 ±0.047	0.276 ±0.050	0	0	0.23	0.28
700	0.164 ±0.042	0.190 ±0.044	0	0	0.16	0.19
1000	0.0667 ±0.027	0.0754 ± 0.029	0	0	0.067	0.075

Table 3 **Asymptotic accuracy of fluid approximations with exponential abandonment and $N \sim Bin(n, 0.4)$.**

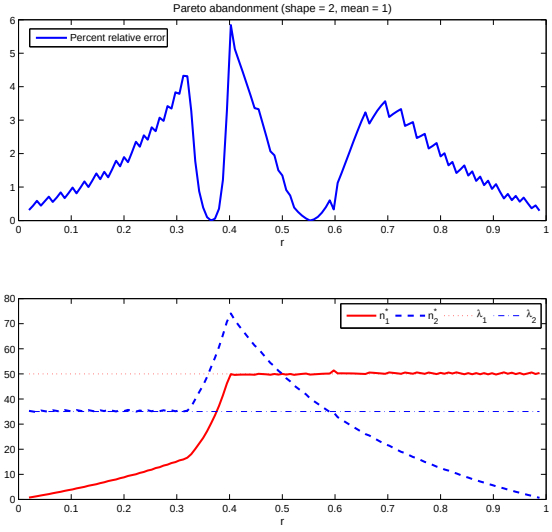

Figure 7 **Relative percent error of fluid prescriptions with Pareto abandonment and $\lambda_1 = 50, \lambda_2 = 35$.**

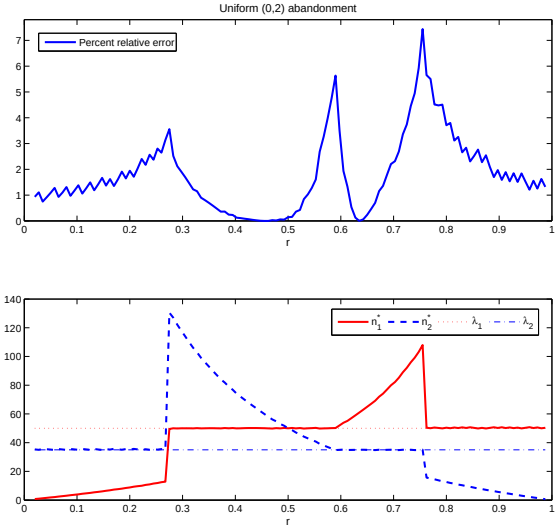

Figure 8 **Relative percent error of fluid prescriptions with uniform abandonment and $\lambda_1 = 50, \lambda_2 = 35$.**

Exponential Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	5.48 ±0.23	5.49 ±0.22	4.8	4.8	0.68	0.69
50	8.50 ±0.30	8.52 ±0.30	8	8	0.50	0.52
70	11.4 ±0.37	11.4 ±0.37	11.2	11.2	0.19	0.20
100	16.1 ±0.47	16.1 ±0.47	16	16	0.084	0.094

Pareto Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	8.59 ±0.20	5.18 ±0.22	9.70	4.8	1.1	0.38
50	15.1 ±0.22	8.30 ±0.30	16.2	8	1.0	0.30
70	21.5 ±0.27	11.1 ±0.41	22.6	11.2	1.14	0.101
100	31.7 ±0.25	16.0 ±0.46	32.3	16	0.68	0.027

Uniform Abandonment

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N]-rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N]-rn\bar{\alpha}\rvert$
30	10.8 ±0.46	4.82 ±0.37	12.5	4.8	1.7	0.019
50	19.3 ±0.54	7.95 ±0.51	20.9	8	1.6	0.051
70	27.8 ±0.64	11.2 ±0.63	29.2	11.2	1.4	0.0064
100	40.7 ±0.65	16.1 ±0.79	41.7	16	1.1	0.057

Table 4 **Asymptotic accuracy of fluid performance measures in the overloaded regime ($\rho = 1.4$) with lognormal service times with mean 1 and variance $e-1$ and $N \sim Bin(n, 0.4)$.**

Critically loaded ($\rho = 1$)

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N] - rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N] - rn\bar{\alpha}\rvert$
30	1.80 ±0.14	1.85 ±0.14	0	0	1.8	1.9
50	2.24 ±0.17	2.29 ±0.17	0	0	2.2	2.3
70	2.62 ±0.22	2.68 ±0.22	0	0	2.6	2.7
100	3.19 ±0.26	3.26 ±0.26	0	0	3.19	3.26

Underloaded regime ($\rho = 0.85$)

n	$\mathbb{E}[Q_N]$	$\mathbb{E}[\alpha_N]$	$rn\bar{q}$	$rn\bar{\alpha}$	$\lvert\mathbb{E}[Q_N] - rn\bar{q}\rvert$	$\lvert\mathbb{E}[\alpha_N] - rn\bar{\alpha}\rvert$
300	0.488 ±0.074	0.552 ±0.076	0	0	0.49	0.55
500	0.275 ±0.066	0.318 ±0.069	0	0	0.28	0.32
700	0.148 ±0.043	0.178 ±0.048	0	0	0.15	0.18
1000	0.0645 ±0.024	0.0772 ±0.026	0	0	0.065	0.077

Table 5 **Asymptotic accuracy of fluid approximations with exponential abandonment, $N \sim Bin(n, 0.4)$, and lognormal service times with mean 1 and variance $e - 1$.**