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

Managing Supply in the On-Demand Economy: Flexible Workers or Full-Time Employees?

Jing Dong
Columbia University, 3022 Broadway, New York, NY 10027 jing.dong@gsb.columbia.edu

Rouba Ibrahim
University College London, 1 Canada Square, London E14 5AB rouba.ibrahim@ucl.ac.uk

There are different workforce models in the "gig" economy. While some on-demand service providers rely strictly on either traditional employees or independent contractors, others rely on a blended workforce which melds a layer of contingent workers with a core of permanent employees. In deciding on the "right number of right people to staff at the right time", managers must appropriately weigh the pertinent tradeoffs. In this paper, we study cost-minimizing staffing decisions in service systems where the manager must decide on how many flexible (contractors) and/or fixed (employees) agents to staff in order to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers. We consider a queueing-theoretic framework where the number of servers is random because part of the workforce is flexible. Since the staffing problem with a random number of servers is analytically intractable, we formulate two problem relaxations, based on fluid and stochastic-fluid formulations, and establish their accuracies in large systems by relying on an asymptotic, many-server, mode of analysis. We derive the optimal staffing policy, and glean insights into the appropriateness of alternative workforce models in on-demand services. We also shed light on the distinction between demand-side (customer arrival rates) and supply-side (number of servers) uncertainties in queueing systems. Finally, we explore, through a numerical study, the impact of variability in the number of servers on the quality of service offered to customers and illustrate that, contrary to intuition, *more* variability in the agent pool may lead to an *improved* quality of service, i.e., customers may be advantaged by increased supply-side variability.

Key words: on-demand workforce; sharing economy; random capacity; many-server queues.

1. Introduction

The gig or on-demand economy has gradually become an integral part of the global economy, and it is projected to continue to grow in the coming years (PWC 2017). Naturally, not all on-demand services are delivered in the same manner. For example, ride-sharing applications, such as

Uber (*uber.com*) and Lyft (*lyft.com*), rely solely on independent contractors to fulfill ride requests from customers. In such settings, the supply of workers available in each time period is uncertain because those contractors are self-scheduling, i.e., they are free to set their own work schedules. In contrast, several on-demand startups, such as Instacart (*instacart.com*) and Sprig (*sprig.com*), have recently shifted away from staffing a workforce of independent contractors and rely on full-time employees instead. There are also multiple companies, such as Walmart (*walmart.com*) and Netflix (*netflix.com*), which rely on a blended workforce i.e., they meld, as a deliberate business strategy, a layer of contingent workers with a core of permanent employees (Forbes 2015).

Given that diverse landscape of alternative workforce models, a service provider must decide, as a long-term business strategy in an initial planning stage, on the numbers of flexible (contractors) and/or fixed (employees) agents to staff in order to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers. This is the problem that we address in this paper.

1.1. Modelling Framework

We study a cost-minimizing service provider's staffing problem in the context of a stylized queueing model. We assume that both types of workers, fixed or flexible, have the same processing speeds, i.e., service rates. However, the two types of workers differ in their unit operating costs, required working periods, and show-up rates. Specifically, while a fixed worker must adhere to a given schedule, a flexible worker is free to choose whether or not to show up to work in a given shift. Customers are both impatient and delay sensitive. There are multiple working periods, customer demand rates are deterministic yet time-varying, and the agent pool may include either fixed or flexible agents (or both). Hereafter, we will use "agent" and "server" interchangeably. A fixed server is compensated c_{fix} per unit time. If a flexible server is available, then she earns c_{flex} per unit time. That is, c_{fix} and c_{flex} are staffing costs. When there are flexible servers in the agent pool, the total number of available servers is random.

Why is our problem challenging? Since the number of servers in our queueing system is random, we are facing a decision-making problem under parameter uncertainty. Because the optimization problem faced by the system manager is analytically intractable, we rely instead on an asymptotic mode of analysis. In particular, we consider a sequence of queueing systems indexed by the arrival rate, λ, and we allow λ to increase without bound.

At a high level, systems with parameter uncertainty involve two "layers" of variability: (i) stochastic variability, for any given realized value of the underlying uncertain parameter, because interarrival, service, and patience times are random; and (ii) parameter uncertainty, because the parameter itself, here the number of servers, is random. We address our capacity-planning question

by considering two alternative problem formulations, which correspond to two regimes, respectively. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. The second formulation assumes that both uncertainty effects and stochastic fluctuations are negligible. In this regime, we derive the optimal staffing levels by solving a fluid optimization problem instead.

Our modelling approach is close to Bassamboo et al. (2010) who derive optimal staffing policies with uncertain arrival rates. However, it is important to emphasize that *the distinction between uncertainty in input (arrival rates) and uncertainty in supply (number of servers) is not a minor technical point.* Indeed, in capacity planning, the appropriate staffing level typically consists of a nominal capacity requirement and an additional capacity hedge against **exogenous** uncertainty. With self-scheduling servers, variability is **endogenous** because the distribution of the random number of available servers depends, itself, on the selected pool size. For example, with endogenous uncertainty, staffing a larger pool could also lead to increased variability and, potentially, a worse service level in the system. Thus, it is unclear, a priori, what the optimal staffing policy should be, and whether it would have a similar structure as with exogenous uncertainty. Indeed, we will demonstrate that the optimal staffing policy in systems with endogenous uncertainty, i.e., in supply, gives rise to **different hedging regimes** than with exogenous uncertainty, i.e., in demand.

1.2. Main Contributions

Here is a summary of the main theoretical and managerial contributions of this paper.

- We derive optimal staffing policies based on fluid and stochastic-fluid approximations with time-varying demand, and rigorously justify their accuracies by quantifying their corresponding errors in large systems. In particular, we demonstrate that stochastic-fluid approximations are "extremely" accurate, especially when the magnitude of uncertainty in supply is large.
- For the optimal staffing policy, we distinguish between **four regimes**, depending on the magnitude of variability of the random number of servers. Letting n denote the expected number of servers, and $\sigma_n = an^q$, for $a > 0$ and $0 \leq q \leq 1$, its standard deviation, the four regimes that we identify are: (i) variability-dominated, for $0 \leq q \leq 1/2$, where there is no concrete benefit from an uncertainty hedge over the regular square-root staffing hedge; (ii) "moderately" uncertainty-dominated, for $1/2 < q \leq 3/4$, where the uncertainty hedge embodied in a simple newsvendor-problem-based solution, which *ignores* the dependence between variability in supply and staffing prescription, i.e., approximates σ_n by $\sigma_{\lambda/\mu}$ where λ is the arrival rate and μ is the service rate, is extremely accurate; (iii) "strongly" uncertainty-dominated, for $3/4 < q < 1$, where there is additional benefit from using an uncertainty hedge which *accounts* for the dependence between variability in supply and staffing prescription, i.e., where using the simple newsvendor-problem-based solution above leads to a considerable loss in accuracy;

and (iv) "extremely" uncertainty-dominated, for $q = 1$, where the uncertainty hedge accounting for that dependence is on the order of the mean number of servers (in the first three regimes, it is of a smaller order than the mean), so that the system can be either underloaded or overloaded. In this case, using the simple newsvendor-problem-based solution above also leads to a considerable loss in accuracy.

We illustrate those theoretical contributions in Figures 1 and 2. Our objective for now is to convey key insights, so we keep our exposition here at a high level. In Figure 1, we consider a queueing system with a random number of servers and a single period. We plot three curves: The optimal staffing policy, i.e., the cost-minimizing prescription for the staffing problem specified in §3, the simple newsvendor-based solution which ignores the dependence between variability in supply and staffing prescription, and the refined solution which accounts for that dependence. Figure 1 illustrates that, while the newsvendor-based and refined solutions are almost indistinguishable in the variability-dominated and moderately-uncertainty-dominated regimes, they are considerably different when the level of uncertainty in supply is sufficiently large. For example, for the choice of parameters in the figure, the percent error for the newsvendor solution, relative to the optimal solution, is over 20% and, for the refined solution, it is less than 1%. Naturally, the staffing levels in the figures depend on our specific choice of parameters. Nevertheless, we deliberately include those numbers in the figures to illustrate the considerable differences between the alternative staffing prescriptions. In other words, we see that there can be significant loss in accuracy when ignoring the dependence between variability in supply and capacity prescription. However, this is not the case when considering staffing decisions in queueing systems with other forms of parameter uncertainty, e.g., with random arrival rates, as we illustrate in Figure 2. In Figure 2, we keep the same parameters as in Figure 1, and consider the same distribution for the random arrival rate as for the random number of servers in Figure 1. It is shown in Bassamboo et al. (2010) that the staffing policy with a random arrival rate gives rise to **two regimes** (instead of the four regimes above), variability-dominated, for $0 \leq q \leq 1/2$, and uncertainty-dominated, for $1/2 < q \leq 1$. In the uncertainty-dominated regime, a simple newsvendor-problem-based capacity prescription is extremely accurate for all values of q; moreover, it is increasingly accurate for more variable demand, i.e., as q increases. Those two regimes are illustrated in Figure 2.

- Finally, through a numerical study, we examine the impact of randomness in supply on the quality of service experienced by customers. A well-known maxim in service science is that "variability hurts in queueing". Thus, we would expect that customers will be disadvantaged by supply-side uncertainty. That is, we would expect that customers will be worse off with a more variable pool of servers. We illustrate that the impact of variability in supply is actually

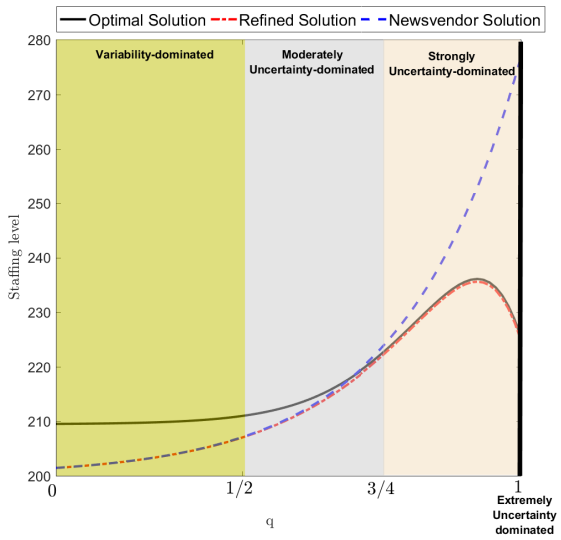

Figure 1 **Optimal staffing policy with a random number of servers $N(n)$ with $n = \mathbb{E}[N(n)]$ and standard deviation $\sigma_n = n^q$.**

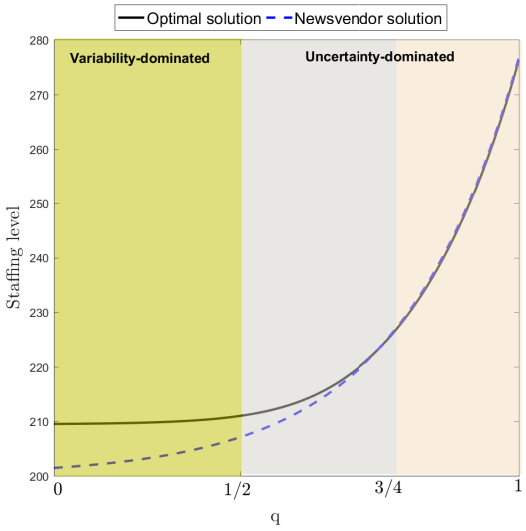

Figure 2 **Optimal staffing policy with a random arrival rate $\Lambda(\lambda)$ with $\lambda = \mathbb{E}[\Lambda(\lambda)]$ and standard deviation $\sigma_\lambda = \lambda^q$.**

more intricate, and that it need not hurt customer-service levels. In fact, those levels may actually *improve* with increased variability in supply. This is the case when all else is fixed in the system, including the pool size, and only variability in the number of servers is increased. We also investigate, numerically, the impact of blending on the quality of service offered to customers. We do so by comparing a blended system to systems where the manager relies on strictly one of the two resources, and where optimal staffing decisions are made in each system. We find that blending the workforce usually leads to a more "balanced" level of service.

The rest of this paper is organized as follows. In §2, we review the relevant literature. In §3, we formulate our capacity-sizing problem, as well as its stochastic-fluid and fluid relaxations. In §4, we consider the fluid approximation. In §5, we consider the stochastic-fluid approximation. In §6, we quantify the impact of supply-side uncertainty on the quality of service offered, i.e., we take the perspective of customers. In §7, we consider a general (not exponential) distribution for patience time. In §8, we draw conclusions. We relegate all proofs to the Appendices.

2. Related Literature

Our paper is part of the literature on controlling queueing systems under parameter uncertainty; e.g., see Harrison and Zeevi (2005) and Bassamboo et al. (2010). Our paper is also broadly related to the extensive literature analyzing asymptotics of many-server queueing systems with impatient customers (e.g., see Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Whitt (2006a), Bassamboo and Randhawa (2010)), and to the large literature on optimal staffing decisions in service

systems (e.g., see Maglaras and Zeevi (2003), Borst et al. (2004), Bassamboo et al. (2005)). However, none of those papers considers a random number of servers. Whitt (2006b) studies staffing decisions in many-server queues with an uncertain arrival rate, an uncertain number of servers, and a single period. Here, we go beyond the fluid approximation of that paper, and study optimal staffing policies in multiple periods where we may have both fixed and flexible servers. Atar (2008) derives a diffusion limit for the number of customers with a random number of servers and random service rates. However, the staffing question is not addressed in that paper.

Our work is also related to papers on nurse staffing with absenteeism, such as Green et al. (2013) and Wang and Gupta (2014), however our asymptotic mode of analysis is different, as well as our consideration of a blended workforce. Our work is also related to the literature on volunteer operations, e.g., harvest gleaning operations, where the managers of such systems are also faced with the uncertain availability of volunteers; see Ata et al. (2018) and references therein. However, this stream of papers does not consider the problem of blending both certain and uncertain supplies, and typically focuses on a binomially-distributed number of servers.

This paper is most closely related to recent papers on queues with a self-scheduling capacity. Gurvich et al. (2018) were the first to study the operational management of systems with self-scheduling agents. They consider a profit-maximizing firm which has three different levers of agent control at its disposal: the pool size, a cap on the number of allowed agents, and the compensation paid to agents. Ibrahim (2018) studies the capacity-sizing problem with a binomially-distributed number of servers and impatient customers, and proposes using delay announcements as an effective control in systems with self-scheduling agents. However, both Gurvich et al. (2018) and Ibrahim (2018) rely solely on fluid approximations of the system, and do not consider a blended workforce.

More generally, there is a growing stream of literature on the management of on-demand service platforms, e.g., see Ozkan and Ward (2018), Hu and Zhou (2018), Taylor (2018), and Cachon et al. (2017). Our work compliments that line of literature.

3. Capacity Sizing in an On-Demand Service Platform

In this section, we describe our modelling framework and formulate the capacity-sizing problem faced by the system manager.

3.1. Queueing Model

We consider a single-class $M/M/N+M$ queueing model, with a random number of servers N. For now, we do not distinguish between fixed and flexible capacity; we will do so later when investigating optimal staffing decisions in the system. Service times are independent and identically distributed (i.i.d.) exponential random variables with rate μ. Customers are impatient, and their patience times are i.i.d. exponentially distributed with rate θ. Customers are processed in the order

in which they arrive, i.e., we use the first-come-first-served discipline. The total number of servers, N, including both fixed and flexible servers, is a nonnegative integer-valued random variable. The arrival, service, and abandonment processes are all mutually independent, also independent of N.

Abandonment makes the system stable, even when N is random (Whitt 2006b). Specifically, conditional on a particular realization of N, a proper steady-state distribution always exists. Stability with a random N follows by conditioning and unconditioning on N. In this paper, we focus on steady-state performances throughout. We assume that there are k periods, and that period i has length T_i. The different periods may correspond to different work shifts in a single day, e.g., morning, afternoon, and evening shifts, or to successive days, weeks, months, etc., depending on the time scale at which the manager decides on her staffing requirements. The arrival rate of the Poisson arrival process in period i is given by λ_i. We fix $\lambda > 0$ and let $\lambda_i \equiv \lambda\xi_i$, where $\xi_i \geq 0$ for each i. We index all relevant quantities by λ, to indicate dependence on the arrival rates. In our asymptotic analysis, we let λ grow without bound while keeping each ξ_i constant.

3.2. A Random Number of Servers

We assume that the pool sizes of flexible agents may vary across periods, i.e., they can be scaled to meet seasonal demand fluctuations. In contrast, we assume that the number of fixed servers is fixed throughout the horizon[1]. Let m_λ denote the number of fixed servers, and n_λ^i denote the total pool size of flexible servers in period i. Let $N_{flex}(n_\lambda^i)$ denote the random number of flexible servers who show up in period i, which depends on n_λ^i. Without loss of generality, we assume that

$$N_{flex}(n_\lambda^i) = \eta_{n_\lambda^i} + \epsilon_{n_\lambda^i}, \tag{1}$$

where $\mathbb{E}[N_{flex}(n_\lambda^i)] = \eta_{n_\lambda^i}$ and $\epsilon_{n_\lambda^i}$ is a random variable with $\mathbb{E}[\epsilon_{n_\lambda^i}] = 0$ and $\mathrm{Var}[\epsilon_{n_\lambda^i}] = \sigma^2_{n_\lambda^i}$. The total number of servers in period i is given by:

$$N(m_\lambda, n_\lambda^i) = m_\lambda + N_{flex}(n_\lambda^i) = m_\lambda + \eta_{n_\lambda^i} + \epsilon_{n_\lambda^i}. \tag{2}$$

In (2), we ignore the integrality assumptions on m_λ, n_λ, and $N(m_\lambda, n_\lambda)$: This is reasonable when the system is large, which is the case of primary interest to us. We also note that the expected queue length expression for the Markovian Erlang-A queueing model can be extended to real values of the number of servers (Mandelbaum and Zeltyn 2007), i.e. the staffing problem faced by the manager is defined for both integer and non-integer values of m_λ and n_λ.

[1] We also considered an alternative setup where fixed workers are not required to show up in every period and are, instead, subject to a requirement on the minimal number of periods during which they must be available. The main insights that we obtain under either modelling framework are similar.

3.3. Binomial Model and Extensions

In this paper, our aim is to characterize the manager's optimal staffing decisions, i.e., what m_λ and n_λ^i in (2) should be. To be able to do so, we must relate the show-up decisions of flexible workers, in the pool of size n_λ^i, to the distribution of $N_{flex}(n_\lambda^i)$ in (1). Since different agent show-up models may be appropriate, depending on the specific application context in mind, we do not attempt here to propose a single model to "fit all" settings. Instead, we explore how different agent-participation models, which may emerge in practice, impact the distribution of $N_{flex}(n_\lambda^i)$. We are especially interested in quantifying the resulting order of magnitude for $\sigma^2_{n_\lambda^i}$ because, as we will demonstrate later, this variance term strongly affects the structure of the optimal staffing policy.

For ease of exposition, we focus here on a single period; thus, we drop dependence on i. For $1 \le j \le n_\lambda$, we define the Bernoulli random variable $I_j = 1$ if agent j is available for work, and $I_j = 0$ otherwise. Then, $N_{flex}(n_\lambda)$ in (1) can be written as follows:

$$N_{flex}(n_\lambda) = \sum_{j=1}^{n_\lambda} I_j. \tag{3}$$

The classical Binomial model. We begin by assuming that I_j in (3) are i.i.d. Bernoulli random variables with a constant and deterministic success probability p; e.g., as in Ibrahim (2018). In this case, it is readily seen that $N_{flex}(n_\lambda)$ has a binomial distribution with $\eta_{n_\lambda} = n_\lambda \cdot p$ and $\sigma^2_{n_\lambda} = n_\lambda \cdot p(1-p)$. In particular, σ_{n_λ} is on the order of magnitude of $\sqrt{n_\lambda}$. That is, the magnitude of uncertainty in the number of servers is of the same order as stochastic variability in the system.

Despite its analytical tractability, the classical Binomial model has several shortcomings; thus, there is a need to consider alternative models as well. For example, it assumes that each agent makes her participation decision independently of other agents. In practice, agent decisions typically exhibit correlations, e.g., because of coordinated joining and leaving decisions facilitated by social-media platforms[2]. Such correlations lead to over-dispersion, i.e., additional variability, compared with the classical Binomial model. For another example, joining probabilities may be neither homogeneous across agents, nor deterministic. Indeed, Chen et al. (2017) provide empirical evidence that each Uber driver faces a hierarchy of random, unforseen, shocks, e.g., linked to weather conditions or promotional events from ride-sharing competitors. A labor supply decision, for each driver, depends on specific (independent across drivers) heterogeneous realizations of those shocks. Thus, ex-ante, the show-up probability of an agent is, itself, a random variable, which also leads to over-dispersion. We now describe extensions to the Binomial model to capture such properties.

[2] `https://warwick.ac.uk/newsandevents/pressreleases/uber_drivers_are/`

Correlated Bernoulli sequences. We begin by describing a model for capturing correlations between agent joining decisions, i.e., I_j in (3). While modelling correlated Bernoulli sequences is not new, our intention here is not to provide an exhaustive review of the relevant literature. Rather, we show how one intuitively appealing model, the Generalized Binomial model proposed by Drezner and Farnum (1993), could be used to explain different orders of magnitude for $\sigma^2_{n_\lambda}$.

We begin by assuming, without loss of generality, that agent j is the j^{th} agent in the pool of size n_λ to make a joining decision. We define $\bar{I}_j \equiv (1/j)\sum_{k=1}^{j} I_k$ and let $\mathcal{F}_j$ be the σ-field generated by the history $\{I_1, ..., I_j\}$. As in Drezner and Farnum (1993), we assume that

$$\mathbb{P}(I_{j+1} = 1|\mathcal{F}_j) = (1-\alpha)\cdot p + \alpha \cdot \bar{I}_j, \tag{4}$$

for some probability p and $\alpha \in [0, 1)$. In other words, (4) assumes that an agent's joining decision is a convex combination of p and the relative frequency of agents who have already joined. In particular, if $\bar{I}_j > p$ $(< p)$, then $\mathbb{P}(I_{j+1} = 1|\mathcal{F}_j) > p$ $(< p)$. That is, the more agents join, the more likely it is that additional agents will join as well. We also note that letting $\alpha = 0$ in (4) allows us to retrieve the classical Binomial model. Heyde (2004) derives asymptotic properties for the variance, $\sigma^2_{n_\lambda}$, implied by (4). In particular, as $n_\lambda \to \infty$, the following is shown to hold[3]:

$$\sigma^2_{n_\lambda} \sim \begin{cases} \frac{p(1-p)n_\lambda}{1-2\alpha} & \text{for } \alpha < 1/2, \\ p(1-p)n_\lambda \log(n_\lambda) & \text{for } \alpha = 1/2, \\ \frac{p(1-p)n_\lambda^{2\alpha}}{(2\alpha-1)\Gamma(2\alpha)} & \text{for } \alpha > 1/2. \end{cases} \tag{5}$$

Based on (5), we note that for $\alpha > 1/2$, σ_{n_λ} is of a larger order of magnitude than $\sqrt{n_\lambda}$, i.e., this model allows for over-dispersion relative to the classical Binomial model. We also note that this case corresponds to Bernoulli sequences with long-range dependence (Heyde 2004). Asymptotic properties of Bernoulli sequences with more general long-range dependence structures can also be found in Romano and Wolf (2000).

A random joining probability. For an alternative extension of the classical Binomial framework, we could also assume that for each period, the joining probability is a random variate drawn from a distribution $\mathcal{P}_{n_\lambda}$ with $\mathbb{E}[\mathcal{P}_{n_\lambda}] = p$. Given $\mathcal{P}_{n_\lambda} = p'$, each agent makes a joining decision with probability p', independently of other agent. In this case, the expected number of agents who are available is $\eta_{n_\lambda} = \mathbb{E}[\mathbb{E}[N_{flex}(n_\lambda)|\mathcal{P}_{n_\lambda}]] = n_\lambda \cdot \mathbb{E}[\mathcal{P}_{n_\lambda}] = n_\lambda p$, and its variance is given by the conditional variance formula:

$$\begin{aligned} \sigma^2_{n_\lambda} = \mathrm{Var}[N_{flex}(n_\lambda)] &= \mathbb{E}[\mathrm{Var}[N_{flex}(n_\lambda)|\mathcal{P}_{n_\lambda}]] + \mathrm{Var}[\mathbb{E}[N_{flex}(n_\lambda)|\mathcal{P}_{n_\lambda}]] \\ &= n_\lambda \cdot \mathbb{E}[\mathcal{P}_{n_\lambda}(1-\mathcal{P}_{n_\lambda})] + n_\lambda^2 \cdot \mathrm{Var}[\mathcal{P}_{n_\lambda}]. \end{aligned}$$

It is readily seen that the order of magnitude of σ_{n_λ} depends on the distribution of $\mathcal{P}_{n_\lambda}$. For example, if $\mathrm{Var}[\mathcal{P}_{n_\lambda}] > c$ for some $c > 0$, then σ_{n_λ} is on the order of magnitude of n_λ.

[3] We write $A_n \sim B_n$, if $\frac{A_n}{B_n} \to 1$ as $n \to \infty$.

3.4. This Paper's Model

In §3.3, we considered alternative agent show-up models which may arise in practice, and illustrated how different modelling assumptions lead to different orders of magnitude for σ_{n_λ}. This is important because, as we will demonstrate in subsequent sections, σ_{n_λ} plays a central role in the optimal staffing policy. However, we emphasize that the fine detail of those specific agent show-up models are of secondary importance for our purposes. Thus, we directly make assumptions on the distribution of $N(m_\lambda, n_\lambda^i)$ instead. In particular, we consider, hereafter, the simplified model:

$$N(m_\lambda, n_\lambda^i) = m_\lambda + \eta_{n_\lambda^i} + \sigma_{\eta_{n_\lambda^i}} \epsilon_i, \tag{6}$$

for i.i.d. random variables $-1 \leq \epsilon_i \leq 1$ with $\mathbb{E}[\epsilon_i] = 0$, that do not depend on n_λ^i. We assume that ϵ_i has a strictly positive probability density function (pdf), f_ϵ, on $(-1, 1)$. Thus, its cumulative distribution function (cdf), F_ϵ, is invertible on that domain. For simplicity of exposition, we also assume the specific form $\sigma_n = an^q$, for some $a > 0$ and $0 < q \leq 1$. For $q = 1$, we also impose that $a < 1$ to ensure that $N_{flex}(n_\lambda) \geq 0$. Based on (6), we can express the manager's staffing decisions directly in terms of m_λ and the expected number of flexible servers available in period i, $\eta_{n_\lambda^i}$. We assume that η_n is strictly increasing in n. Thus there is a one-to-one correspondence between the pool size of flexible servers, n_λ^i, and $\eta_{n_\lambda^i}$, so that those staffing decisions could also be equivalently formulated in terms of n_λ^i instead. For example, in the three agent models proposed above, it holds that $\eta_{n_\lambda^i} = n_\lambda^i \cdot p$. By a slight abuse of notation, and for ease of exposition, we denote the expected pool size by n_λ^i for the remainder of this paper. In other words, we will consider the following model, which is equivalent to (6):

$$N(m_\lambda, n_\lambda^i) = m_\lambda + n_\lambda^i + \sigma_{n_\lambda^i} \epsilon_i, \tag{7}$$

where the manager's objective is to determine cost-effective m_λ and n_λ^i, as we explain next.

3.5. A Long-Term Staffing Problem

In this paper, we consider the long-term staffing question that the manager faces, in the initial planning stage. In practice, managers must make staffing decisions ahead of time to allow for agent training. The timeline of decision-making is as follows: At time zero, i.e., the initial planning stage, the manager makes a staffing decision on the flexible pool size n^i for each period i or, equivalently, on the average numbers of flexible agents desired, and the fixed pool size m. Then, at the beginning of each period i, the staffing level realizes, i.e., we observe a specific realization, $N(m, n^i) = s^i$, which is drawn from the distribution of the random variable $N(m, n^i)$. For the remainder of period i, the system operates like a Markovian Erlang-A queueing system with s^i servers. We assume that the arrival rate, λ_i, is deterministic and constant for each period. We also assume that λ_i

is known, a priori, to the manager, e.g., it is calculated based on demand forecasts which are made at time zero. Note that, with a deterministic λ_i, the optimal staffing level would remain the same so long as the manager must decide on her staffing level, n^i, before the start of period i [4]. Notice that stochastic variations that impact system behavior on the short-time scale, for given model parameters, are less important when uncertainty is introduced on a longer time scale, i.e., in the model parameters themselves. This motivates us to look at the stochastic-fluid optimization problem which ignores stochastic variability.

Consistently with Bassamboo and Randhawa (2010) and Bassamboo et al. (2010), we consider two customer-related costs: (i) A delay cost, h, per customer for each unit of time that this customer spends waiting to be served, and (ii) an abandonment penalty cost, r, incurred per customer who abandons before being served. The per unit of time staffing costs are given by c_{fix} for a fixed server, and c_{flex} for a flexible server. Throughout, we assume that $c_{flex}, c_{fix} < (h/\theta + r)\mu$. This ensures that the fixed and flexible resources are cheap enough to avoid pathological cases where the system manager would not staff any of the two resources. Without loss of generality, we assume that the alternative periods are numbered in order of increasing λ_i values, i.e., $\lambda_i \leq \lambda_j$ for $i \leq j$. In other words, we re-index the different shifts so that the λ_i values are ordered.

We let $Q^i(m_\lambda, n^i_\lambda)$ and $\Xi^i(m_\lambda, n^i_\lambda)$ denote the steady-state queue length and steady-state rate of customer abandonment in period i. We let $X^i(m_\lambda, n^i_\lambda)$ denote the steady-state number of customers in the system, in period i, so that:

$$Q^i(m_\lambda, n^i_\lambda) = (X^i(m_\lambda, n^i_\lambda) - N(m_\lambda, n^i_\lambda))^+,$$

where $x^+ \equiv \max\{x, 0\}$. With exponentially-distributed patience times, it is also well known that:

$$\Xi^i(m_\lambda, n^i_\lambda) = \theta \cdot \mathbb{E}[Q^i(m_\lambda, n^i_\lambda)],$$

where θ is the rate of the patience-time distribution (Mandelbaum and Zeltyn 2007). Letting $\mathbf{n}_\lambda \equiv (n^1_\lambda, \ldots, n^k_\lambda)$, the system manager's long-run staffing problem is given by:

$$\begin{aligned} \min_{m_\lambda, \mathbf{n}_\lambda} \quad & \Pi_\lambda(m_\lambda, n^i_\lambda) \\ & \equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n^i_\lambda + h \cdot \mathbb{E}[Q^i(m_\lambda, n^i_\lambda)] + r \cdot \Xi(m_\lambda, n^i_\lambda) \right), \\ & = \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n^i_\lambda + (h/\theta + r)\mathbb{E}[\Xi(m_\lambda, n^i_\lambda)] \right), \end{aligned} \tag{8}$$

[4] In some applications, we could also consider the case where additional information could be gathered to yield improved demand forecasts as we get closer to the start of the period. In that case, we may want to update our staffing decision over time. As flexible servers may be more flexible to call upon in the last minute, these servers will bring an extra layer of benefit. This case is beyond the scope of the current paper.

$$= \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta)\mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] \right),$$

The problem formulation in (8) is prohibitively difficult to solve in closed form, because our choices of m_λ and n_λ^i affect the distribution of the number of servers which, in turn, affects the distributions of the queue-length, Q, and the abandonment rate, Ξ. We next formulate a stochastic-fluid (ignoring stochastic variability) relaxation and a fluid (ignoring both stochastic variability and parameter uncertainty) relaxation of (8).

4. Fluid Approximation

We are now ready to formulate the fluid relaxation of our problem. For this, we ignore both uncertainty effects and stochastic fluctuations in the system. In particular, the fluid abandonment rate in our problem is given by $(\lambda_i - m_\lambda \mu - n_\lambda^i \mu)^+$, which is obtained by substituting the random number of servers, $N(m_\lambda, n_\lambda^i)$, by its expected value, $m_\lambda + n_\lambda^i$. This leads to the following:

$$\min_{m_\lambda, \mathbf{n}_\lambda} \tilde{\Pi}_\lambda(m_\lambda, n_\lambda^i) \equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \left(\frac{h}{\theta} + r \right) \mu \left(\lambda_i/\mu - m_\lambda - n_\lambda^i \right)^+ \right). \tag{9}$$

Given its simple form, the fluid approximation in (9) is appealing, provided that it does not entail a significant loss in accuracy. Next, we characterize when that is indeed the case.

4.1. Optimality Gap

We now study the accuracy of the fluid staffing prescription in (9), in a regime where the arrival rate, λ, is large. For ease of exposition, we focus here on a single-period setting or, equivalently, the stationary-demand case, and relegate the generalization to the multi-period case to the Appendix (§A.3). Thus, we drop for now the dependence on the period's index, i. Let $(m_\lambda^*, n_\lambda^*)$ denote the optimal staffing levels to the original staffing problem in (8). Let $(\tilde{m}_\lambda, \tilde{n}_\lambda)$ denote the optimal solution to the fluid relaxation in (9). In what follows, we will 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 $|f(n)| \le M|g(n)|$ for $n \ge C$;
(b) $f(n) = o(g(n))$ if for any $\xi > 0$, there exists $N(\xi)$ such that $|f(n)| \le \xi|g(n)|$ for all $n \ge N(\xi)$;
(c) $f(n) = \Theta(g(n))$ if there exist $M > 0$, $L > 0$ and $C > 0$ such that $L|g(n)| \le |f(n)| \le M|g(n)|$ for $n \ge C$.

We are now ready to state the main theorem of this section.

Theorem 1. *For large λ,*

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}\left(\max\left\{ \sigma_{\tilde{n}_\lambda}, \sqrt{\lambda} \right\} \right);$$

i.e., if $\tilde{n}_\lambda = \Theta(\lambda)$, then $\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}\left(\max\left\{ \sigma_\lambda, \sqrt{\lambda} \right\} \right)$.

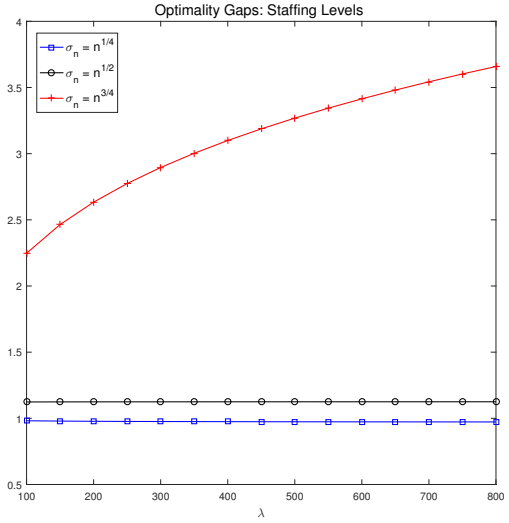

Figure 3 **Scaled errors for optimal fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ.**

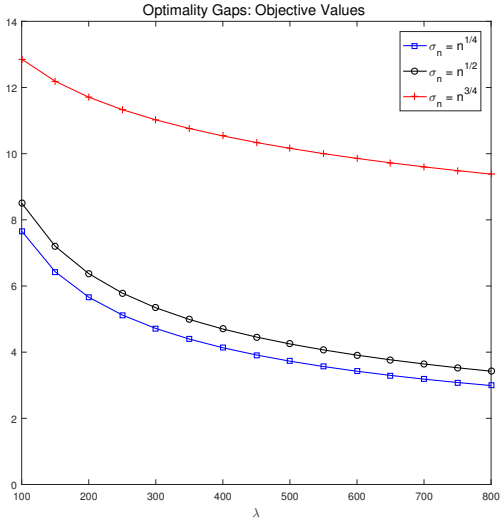

Figure 4 **Optimality gap in objective values, $100 \cdot |\Pi(\tilde{n}_\lambda) - \Pi(n^*_\lambda)|/\Pi(n^*_\lambda)$, as a function of λ.**

Theorem 1 shows that the accuracy of a first-order fluid approximation degrades as the uncertainty in the number of servers increases. In particular, when σ_λ is "small", i.e., of an order which is smaller than the square-root order of stochastic fluctuations in the system, the optimality gap for the fluid solution is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. However, when σ_λ is "large", i.e., of an order which is larger than the square-root order of stochastic fluctuations in the system, fluid approximations may lead to a considerable loss in accuracy.

Numerical example. We illustrate the results of Theorem 1 in Figures 3 and 4. In these figures, we let $N_{flex}(n_\lambda) = n_\lambda + \sigma_{n_\lambda}\epsilon$ and assume that ϵ has a uniform distribution, $\epsilon \sim U(-1,1)$. We consider $\sigma_n = n^q$ for $q = 1/4, 1/2$, and $3/4$. We let $c_{flex} = 1/3$, $c_{fix} = 1/2$, and $p = h = \mu = \theta = 1$. Solving (9), we obtain $\tilde{m}_\lambda = m^*_\lambda = 0$, i.e., there are no fixed servers at optimum. In Figure 3, we plot the scaled staffing-level errors, between the fluid and original solutions, $|n^*_\lambda - \tilde{n}_\lambda|/\sqrt{\lambda}$, as λ increases. In Figure 4, we plot the corresponding relative percent errors in the objectives, i.e., we plot $100 \cdot |\Pi(\tilde{n}_\lambda) - \Pi(n^*_\lambda)|/\Pi(n^*_\lambda)$. Figure 3 illustrates the orders of magnitude for the asymptotic accuracy of fluid prescriptions, as given by Theorem 1. In particular, the fluid approximation's accuracy degrades as the level of uncertainty in the number of servers increases. Figure 4 illustrates how the differences in staffing levels impact the respective objective values: While all solutions improve in accuracy as λ increases, the fluid approximation is considerably worse for larger values of q, which correspond to the uppermost curve in the figures.

4.2. Optimal Solution

For the derivation of the optimal solution, it is useful to define the "modified" cost:

$$c_{fix}^h \equiv c_{fix} \cdot \frac{\sum_{i=1}^k T_i}{\sum_{i=h}^k T_i} \text{ for } h \geq 1. \tag{10}$$

We are now ready to describe the optimal solution to the fluid staffing problem. This solution depends on both the staffing costs and the lengths of the individual periods, as follows.

LEMMA 1. *The solution to the fluid problem in (9), with time-varying demand, is given by*

$$\begin{cases} \tilde{m}_\lambda = 0, & \text{for } k_0 = 0, \\ \tilde{m}_\lambda = \frac{\lambda_{k_0}}{\mu} & \text{for } k_0 > 0, \\ \tilde{n}_\lambda^i = 0, & \text{for } 1 \leq i \leq k_0, \\ \tilde{n}_\lambda^i = \frac{\lambda_i - \lambda_{k_0}}{\mu}, & \text{for } k_0 < i \leq k, \end{cases}$$

where k_0 is defined as follows:

$$k_0 = \begin{cases} 0, & \text{if } c_{flex} < c_{fix}, \\ \max\{1 \leq h \leq k : c_{flex} \geq c_{fix}^h\}, & \text{otherwise.} \end{cases}$$

Lemma 1 coins the advantage of staffing a pool of flexible agents: Such a pool can be dynamically adjusted to meet seasonality in customer demand. The only case where the manager would staff fixed servers is if $c_{flex} \geq c_{fix}$, i.e., they are cheaper. However, even in this case, she may still staff the more expensive flexible servers, i.e., she would blend her workforce, unless fixed servers are "very" cheap, e.g., if $c_{fix} < \frac{T_k}{\sum_{i=1}^k T_i} c_{flex}$. The general form of the optimal staffing policy in Lemma 1 shows that a manager who blends should rely solely on the fixed resource in the low-demand periods, and blend in the high-demand periods. The extent to which the manager should rely on the fixed resource only depends on the cost of that resource: The cheaper the fixed resource, the more periods the manager relies strictly on that resource, i.e., does not resort to blending. While the optimal solution to the fluid staffing problem is readily obtainable and easy to interpret, Theorem 1 shows that it may not be reliable when uncertainty in the number of available servers is large. Thus, there is a need to consider a more refined approximation, which we do next.

5. Stochastic-Fluid Approximation

In this section, we define the stochastic-fluid approximation to the original staffing problem in (8). We then quantify the optimality gap entailed, and derive the optimal stochastic-fluid solution.

For the stochastic-fluid relaxation, we ignore stochastic fluctuations in the system. In particular, customers arrive in period i at the rate of λ_i per unit of time. The processing capacity is $N(m_\lambda, n_\lambda)\mu$

and, by conservation of flow, the resulting stochastic-fluid abandonment rate is given by $(\lambda_i - N(m_\lambda, n_\lambda)\mu)^+$. Thus, the resulting stochastic-fluid approximation to (8) is:

$$\begin{aligned}\min_{m_\lambda, n_\lambda^i} \bar{\Pi}_\lambda(m_\lambda, n_\lambda^i) &\equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta) \frac{1}{\theta} \mathbb{E}\left[\left(\lambda_i - N(m_\lambda, n_\lambda^i) \cdot \mu \right)^+ \right] \right), \qquad (11) \\ &= \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \left(\frac{h}{\theta} + r \right) \mu \mathbb{E}\left[\left(\lambda_i/\mu - m_\lambda - n_\lambda^i - \sigma_{n_\lambda^i} \epsilon_i \right)^+ \right] \right)\end{aligned}$$

5.1. Optimality Gaps

Paralleling Theorem 1, we now study the accuracy of the stochastic-fluid staffing prescription in (11), in a regime where the arrival rate, λ, is large. We focus here on a single-period setting as well, and relegate the generalization to the multi-period case to the Appendix (§A.3). Recall that $(m_\lambda^*, n_\lambda^*)$ are the optimal staffing levels to the original staffing problem in (8). Let $(\bar{m}_\lambda, \bar{n}_\lambda)$ denote the optimal solution to the stochastic-fluid relaxation in (11).

THEOREM 2. *For large λ,*

$$\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}\left(\min\left\{ \lambda/\sigma_{\bar{n}_\lambda}, \sqrt{\lambda} \right\} \right);$$

i.e., if $\bar{n}_\lambda = \Theta(\lambda)$, then $\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m_\lambda^*, n_\lambda^*) + \mathcal{O}\left(\min\left\{ \lambda/\sigma_\lambda, \sqrt{\lambda} \right\} \right)$.

When σ_λ is "large", i.e., of an order which is larger than the square-root order of stochastic fluctuations in the system, stochastic-fluid approximations are remarkably accurate. Indeed, the optimality gap for the stochastic-fluid solution is on the order $\mathcal{O}(\lambda/\sigma_\lambda)$. In other words, the stochastic-fluid approximation becomes *increasingly* accurate as the variability in the number of servers increases. On the other hand, when σ_λ is "small", i.e., of an order which is smaller than the square-root order of stochastic fluctuations in the system, the optimality gap for the stochastic-fluid solution is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. In other words, when the variability in the number of servers is small, there is no distinct advantage from using stochastic-fluid approximations over fluid approximations to the system (cf. Theorem 1). This implies that a first-order fluid approximation is sufficiently accurate in that case.

Numerical example. We illustrate the asymptotic results of Theorem 2 in Figures 5 and 6. In these figures, we compare the optimal stochastic-fluid solution, $\bar{n}_\lambda$, to the optimal solution of the original problem, n_λ^*. We consider the same system parameters as in Figures 3 and 4. In contrast with the fluid solution, Figures 5 and 6 illustrate the improvement in accuracy for $\bar{n}_\lambda$ as the uncertainty in the number of servers increases. Indeed, for $\sigma_n = n^{3/4}$ (bottom curve in the plots), n_λ^* and $\bar{n}_\lambda$ are practically indistinguishable. For $\sigma_n = n^q$ and $q \leq 1/2$, comparing Figures 3 and 5 reveals that the improvement in accuracy entailed in refining the fluid solution, and relying on the stochastic-fluid solution instead, is asymptotically negligible.

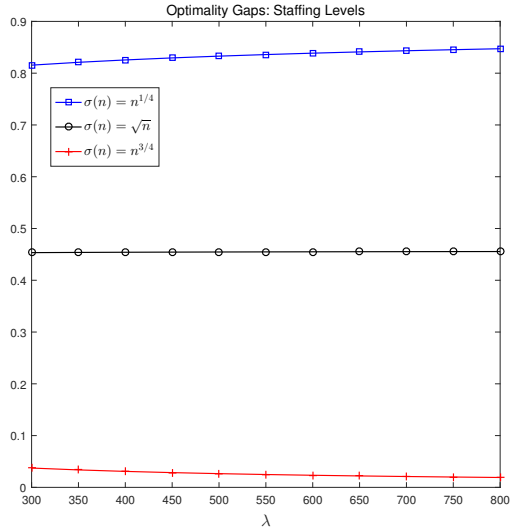

Figure 5 **Scaled errors for optimal stochastic-fluid levels, $|\bar{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ.**

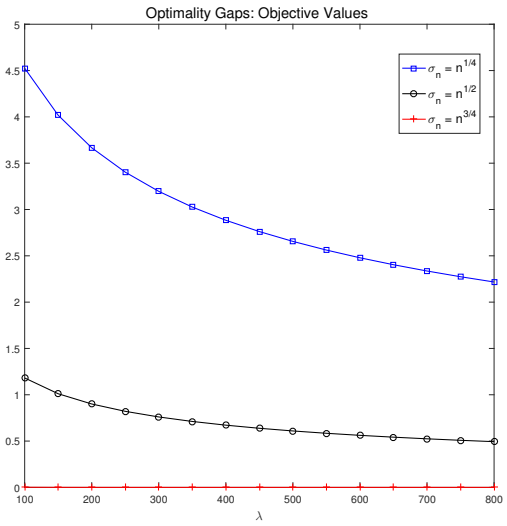

Figure 6 **Optimality gaps in objective values, $100 \cdot |\Pi(\bar{n}_\lambda) - \Pi(n^*_\lambda)|/\Pi(n^*_\lambda)$, as a function of λ.**

5.2. Optimal Solution

Contrasting Theorems 1 and 2 quantifies the improvement in accuracy which is entailed in refining the fluid solution, and relying on the stochastic-fluid solution instead. There remains to investigate the structure of the stochastic-fluid staffing policy, to which we devote this section. In Theorem 3, we derive the optimal solution to a stochastic-fluid problem with a single period. Then, we demonstrate how the solution to the multi-period stochastic-fluid problem in (11) can be derived by solving appropriate single-period stochastic-fluid problems instead. Building on this fact, we derive the optimal solution to the multi-period stochastic-fluid problem in Theorem 4.

5.2.1. Solution to the Single-Period Problem. In this section, we present the solution to the single-period stochastic-fluid problem. Our main theorem in this section is Theorem 3, where we show that uncertainty in supply gives rise to **four** regimes, depending on the magnitude of variability in the number of servers. Indeed, when supply is uncertain, the variability of the random number of servers depends, itself, on the staffing prescription. The question is, then, whether ignoring that dependence, i.e., substituting σ_n in (7) by $\sigma_{\lambda/\mu}$, would lead to an asymptotically non-negligible loss in accuracy, i.e., to errors that are of a higher order of magnitude.

If we substitute σ_n in (7) by $\sigma_{\lambda/\mu}$, then the staffing problem in (11) becomes a standard newsvendor problem. We show that if variability in supply is "small" or "moderate", then there is no loss of accuracy entailed in doing that substitution. This corresponds to cases I and II in Theorem 3, where we show that the solutions to simpler problems, such as a fluid problem in case I, or a

standard newsvendor problem in case II, yield the same optimality gap as the exact solution to (11). In contrast, if variability in supply is "strong" or "extreme", then there is loss in accuracy when ignoring that dependence. This corresponds to cases III and IV in Theorem 3, where the newsvendor prescription described above performs poorly. For expositional ease, let:

$$\beta \equiv (h/\theta + r)\mu. \tag{12}$$

For now, we will assume that there is only one type of resource: flexible. Later, we will include the fixed resource in our problem as well. Thus, our stochastic-fluid problem can be written as:

$$\min_{n_\lambda \geq 0} \bar{\Pi}_\lambda(n_\lambda) \equiv c_{flex} n_\lambda + \beta \cdot \mathbb{E}[(\lambda/\mu - n_\lambda - \sigma_{n_\lambda}\epsilon)^+]. \tag{13}$$

Here is our main theorem.

THEOREM 3. *For the solution of problem (13), we let $\sigma_n = an^q$, for $a > 0$ and $0 \leq q \leq 1$, and distinguish among four cases:*

(I) ***[Variability-dominated.]*** *If $0 \leq q \leq 1/2$, we set $\hat{n}_\lambda = \lambda/\mu$. In this case, we have*

$$\Pi_\lambda(\hat{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\sqrt{\lambda}).$$

(II) ***[Moderately uncertainty-dominated.]*** *If $1/2 < q \leq 3/4$, we set $\hat{n}_\lambda = \lambda/\mu + y^*_1 \sigma_{\lambda/\mu}$, where $y^*_1 = -F^{-1}_\epsilon(c_{flex}/\beta)$ is the solution to:*

$$\min_y c_{flex} y + \beta \mathbb{E}[(-y - \epsilon)^+], \tag{14}$$

for β in (12). In this case, we have

$$\Pi_\lambda(\hat{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda).$$

(III) ***[Strongly uncertainty-dominated.]*** *If $3/4 < q < 1$, then $\bar{n}_\lambda = \lambda/\mu + y^*_2 \sigma_{\bar{n}_\lambda}$, where y^*_2 solves:*

$$\min_y (c_{flex} y + \beta \mathbb{E}[(-y - \epsilon)^+]) s_\lambda(y), \tag{15}$$

for $s_\lambda(y) = \sigma_{n_\lambda(y)}$ and $n_\lambda(y) = \lambda/\mu + y\sigma_{n_\lambda(y)}$. In this case,

$$\Pi_\lambda(\bar{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda).$$

(IV) ***[Extremely uncertainty-dominated.]*** *If $q = 1$ and $0 < a < 1$, then $\bar{n}_\lambda = \frac{\lambda}{\mu}\gamma^*$, where γ^* denotes the solution to*

$$c_{flex} + \beta a \int_{-1}^{1/(a\gamma^*) - 1/a} F_\epsilon(u) du - \frac{\beta}{\gamma^*} F_\epsilon\left(\frac{1}{a\gamma^*} - \frac{1}{a}\right) = 0. \tag{16}$$

In this case,

$$\Pi_\lambda(\bar{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(1).$$

We emphasize that in cases I and II, $\hat{n}_\lambda$ is *not* the optimal solution to (13), but rather a simplified solution which we prove yields, asymptotically, the same order of accuracy as the actual optimal solution (cf. Theorem 2). In contrast, $\bar{n}_\lambda$ in cases III and IV is the actual solution to that problem. Indeed, simplifying as we did in the former cases would then lead to substantial errors. We quantify the asymptotic order of magnitude for that loss in accuracy in the following corollary.

COROLLARY 1. *Let $\sigma_n = an^q$, for $0 \le q \le 1$ and $a > 0$, $\bar{n}_\lambda$ be the optimal solution to (13), and $\hat{n}_\lambda$ be the optimal solution when σ_{n_λ} in (13) is replaced by $\sigma_{\lambda/\mu}$. Then,*

$$\Pi_\lambda(\bar{n}_\lambda) - \Pi_\lambda(\hat{n}_\lambda) = \mathcal{O}(\lambda^{3q-2}).$$

It is also interesting to note that in case IV of Theorem 3, i.e., with $q = 1$, it may be beneficial to underload or overload the system, i.e., to use an uncertainty hedge which is of the same order as the expected number of servers. For example, we obtain from (16) that $\gamma^* \ge 1$, i.e., we underload, if $c_{flex} \le \beta F_\epsilon(0) - \beta a \int_{-1}^{0} F_\epsilon(u)du$, i.e., capacity is cheap. This lies in contrast with all cases where $q < 1$, since the uncertainty hedge is of a smaller order than the expected number of servers there.

Numerical evidence. We present numerical evidence substantiating the results of Theorem 3 in Figures 7 and 8. The dashed curves in those figures correspond to the percent error in using the solution from case (II) of Theorem 3 relative to the optimal solution of the original problem in (8); the solid curves correspond to using the solution from case (III) instead. We let $\sigma_n = n^q$ for $0 < q \le 0.95$, i.e., we exclude case (IV) in the theorem, and assume that ϵ in (6) is uniformly distributed over $(-1, 1)$. In both figures, we assume that $\lambda = 200$, $c_{fix} = 1/2$, $c_{flex} = 1/4$, and $r = h = \mu = 1$. In Figure 7 we let $\theta = 1$, and in Figure 8 we let $\theta = 2$, which gives a smaller β as defined in (12).

Clearly, there is considerable loss in accuracy in ignoring the dependence between the variability in supply and the staffing prescription, particularly when there is considerable variability, i.e., q is large. This loss is also exacerbated for more impatient customers (Figure 8). However, when σ_n is not too large ($q < 3/4$ in the figures, as in Theorem 3), there is no asymptotically discernable gain from taking that dependence into account.

5.2.2. Solution to the Multi-Period Problem. We now derive the optimal staffing policy for problem (11), by making use of the single-period results of the previous section. We consider two cases, depending on σ_n: Case 1, where $\sigma_n = n^q$ for $q < 1$, and case 2, where $\sigma_n = an$ for $0 < a < 1$. In Theorem 4, we show that, when variability in supply is not too large (case 1), the optimal staffing level consists of a base capacity which is the fluid-based prescription of Lemma 1, along with an uncertainty hedge which is derived based on the single-period solution. However, when variability is large (case 2), then we can no longer rely on the fluid-based prescription of Lemma 1. Indeed, in

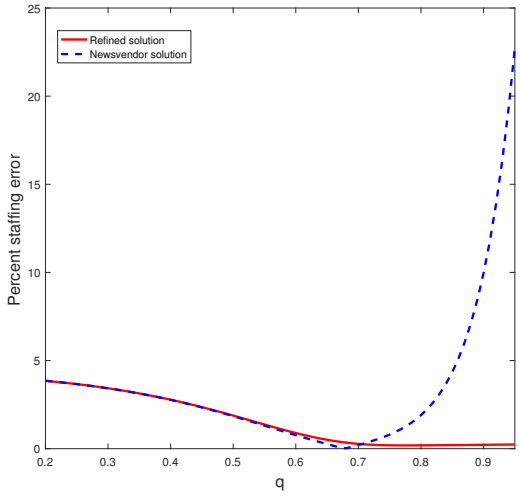

Figure 7 **Stochastic-fluid solution: Percent relative errors in staffing for $\theta = 1$.**

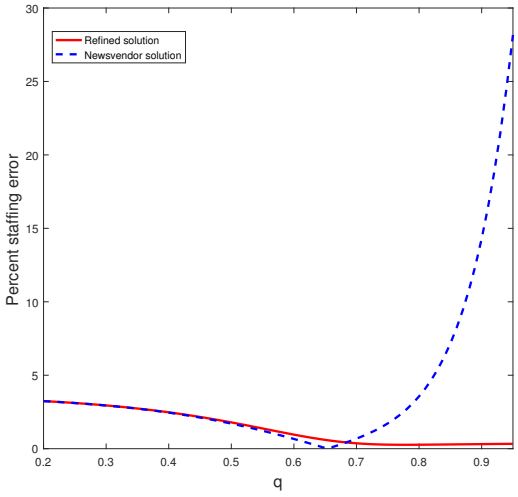

Figure 8 **Stochastic-fluid solution: Percent relative errors in staffing for $\theta = 2$.**

case 2, supply-side variability is on the same order as the mean number of flexible servers. As the uncertainty hedge is now of the same order as λ, it can be optimal to either overload or underload the system, i.e., the ratio between the arrival rate and the mean service capacity is either strictly smaller or strictly larger than 1; this is unlike all cases where $q < 1$. Moreover, when deciding which resource to staff from, we have to add a "risk premium" on the cost of flexible servers, i.e., even if flexible servers are cheaper, we may still use the fixed resource instead.

Recall that $(\tilde{m}_\lambda, \tilde{n}_\lambda)$ denotes the optimal solution to the fluid relaxation in (9) and $(\bar{m}_\lambda, \bar{n}_\lambda)$ denotes the optimal solution to the stochastic-fluid problem in (11). We also define the function:

$$g(c) \equiv c + caF_\epsilon^{-1}(c/\beta) - \beta a \int_{-1}^{F_\epsilon^{-1}(c/\beta)} F_\epsilon(u)du. \tag{17}$$

THEOREM 4. *For the solution of the multi-period problem in (11), there are two cases:*

- ***Case 1. Variability-dominated, moderately and strongly uncertainty-dominated regimes.*** *If $\sigma_n = an^q$ for $0 < q < 1$ and $a > 0$, then for $\tilde{m}_\lambda$ and k_0 as given by Lemma 1, the optimal solution is:*
  - *$\bar{m}_\lambda = \tilde{m}_\lambda$,*
  - *For $i \le k_0$, $\bar{n}_\lambda^i = 0$,*
  - *For $i > k_0$, $\bar{n}_\lambda^i$ is the minimizer of the following single-period problem:*

$$\min_{n_\lambda \ge 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon)^+], \tag{18}$$

  as given by cases I, II, and III of Theorem 3, depending on the value of q.

- ***Case 2. Extremely uncertainty-dominated regime.*** *If $\sigma_n = an$, for $0 < a < 1$, then for c^h_{fix} in (10) and $g(\cdot)$ in (17), let:*

$$k_1 = \begin{cases} 0, & \text{if } c_{flex} < g(c_{fix}) \\ \max\{1 \le h \le k : c_{flex} \ge g(c^h_{fix})\}, & \text{otherwise.} \end{cases}$$

The optimal solution is:

— $\bar{m}_\lambda = \lambda_{k_1}/\mu$,

— *For $i \le k_1$, $\bar{n}^i_\lambda = 0$,*

— *For $i > k_1$, $\bar{n}^i_\lambda$ is the minimizer of the following single-period problem:*

$$\min_{n_\lambda \ge 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - a n_\lambda \epsilon)^+], \tag{19}$$

as given by case IV in Theorem 3.

Recall from Lemma 1 that the optimal solution to the fluid problem is to rely strictly on the fixed resource in lower-demand periods, up to some period index k_0, and to blend resources in higher-demand periods. Theorem 4 shows that the optimal staffing policy for the stochastic-fluid problem has a similar structure when $\sigma_n = an^q$ for $q < 1$. Indeed, in lower-demand periods, up to period k_0, the manager should also rely strictly on the fixed resource. Moreover, the optimal staffing level for the fixed resource in the stochastic-fluid problem remains the same as for the fluid problem, i.e., $\tilde{m}_\lambda = \bar{m}_\lambda$ and $\bar{n}^i_\lambda = 0$ for $i \le k_0$. In higher-demand periods, i.e., periods whose index exceeds k_0, the manager should blend her workforce. However, the staffing levels for the flexible resource in the stochastic-fluid problem are potentially different than those given by the fluid solution in Lemma 1. In particular, for each period, we must solve (18), as in Theorem 3. This adds an uncertainty hedging to the flexible server pool.

Similar insights hold when $\sigma_n = an$, for $0 < a < 1$, where it remains optimal to rely on the fixed resource in low-demand periods and to blend in high demand periods. However, the fluid-based prescription of Lemma 1 could lead to substantial errors in this case. In particular, the period indices k_0 and k_1 in cases 1 and 2 of Theorem 4 may be *different*, and we may have $k_1 > k_0$. For example, with all other parameters held equal, it may be optimal to rely on the flexible resource in case 1 but not in case 2; we illustrate this in the following numerical study (see Figure 15 and 16 and the discussion in §6.3). We also note that $g(c^h_{fix}) < c^h_{fix}$, so that there is a "risk premium" that is incurred on the cost of the flexible resource.

6. How is the Quality of Service Impacted by Uncertainty in Supply?

In this section, we describe results from a supporting numerical study. In particular, we take the customers' perspective and study the impact of supply-side uncertainty on the quality of service, experienced by customers in the system.

To measure the quality of service, we consider both the probability of delay and the expected queue length, in steady state. It is customary to consider the probability of delay as a measure of the quality of service, e.g., see Halfin and Whitt (1981) and Garnett et al. (2002). Indeed, in the asymptotic regimes that arise at optimum for our staffing problem, the queue length is generally small (of a smaller order of magnitude than the average number of servers), and the waiting time negligible when the system is large enough. Thus, to study the impact of variability in supply on the quality of service, it would be more appropriate to focus on the non-trivial probability of delay instead. Nevertheless, we also consider the expected queue length for completeness.

A well-known maxim in service science is that "variability hurts performance" (Daskin 2011). In other words, we anticipate that, for a fixed agent pool size, a more variable supply should lead to worse performance in the system, i.e., a larger probability of delay and longer delays on average. In this section, we illustrate that the impact of variability in supply is more intricate, and may lead to either an improvement or deterioration in the quality of service, depending on system specifics as well as the quality-of-service measure considered.

In §6.1, we consider the case where the agent pool size is fixed, i.e., held equal to some $n > 0$. We consider alternative values for the variance of the number of available agents. We illustrate that, as that variance increases, the probability of delay may either increase or decrease, depending on the specific value of n that is selected, whereas the average queue length generally increases. In §6.2, we assume that the manager adjusts her pool size, in an optimal manner, as that variance increases. In particular, we let n be equal to the cost-minimizing staffing level. We make similar observations in this case as well. In §6.1 and §6.2, we focus on systems where the agent pool consists only of flexible agents. Then, in §6.3, we consider a blended system.

6.1. Fixed Pool Size

We begin by fixing the flexible agent pool size, n. For the random number of flexible agents, we assume that $N(n) = n + \sigma\epsilon$, where ϵ is uniformly distributed over $(-1, 1)$, i.e., $N(n) \sim U(n - \sigma, n + \sigma)$. In other words, each fixed value of n corresponds to a different distribution for $N(n)$. Also, for each n, $N(n)$ is more variable for a larger σ.

To illustrate the effect of σ on the quality of service, we consider two values: $\sigma_1 = 10$ and $\sigma_2 = 40$. We consider the $M/M/N(n) + M$ system where we let the arrival rate $\lambda = 100$ and the service rate $\mu = 1$. We let the abandonment rate $\theta = 1$. For each value of n, and other parameters held fixed, we condition (and then, uncondition) on the random number of servers, $N(n)$. In particular, conditional on a specific realization of the random variable $N(n)$, we numerically calculate the corresponding value for the probability of delay, in steady state. Then, we take the expected value (i.e., we average over realizations) by numerically integrating over the uniform distribution of $N(n)$.

Thus, we obtain values for the probability of delay, $\mathbb{P}(W(n)>0)$, as a function of n. Similarly, we obtain values for the expected queue length, $\mathbb{E}[Q(n)]$. We plot curves for $\mathbb{E}[Q(n)]$ and $\mathbb{P}(W(n)>0)$, as a function of n, in Figure 9. Each point on those curves corresponds to a fixed value of the flexible pool size, n.

Figure 9 illustrates that for small values of n (e.g., $n<100$ in the figure), $\mathbb{P}(W(n)>0)$ is *smaller* when the variability in supply is *larger*. Indeed, the dashed curve, corresponding to $\sigma_2=40$, falls below the solid curve, corresponding to $\sigma_1=10$, for such values of n. However, we observe the opposite effect when n is large (e.g., $n\geq 100$ in the figure). This seems to suggest that when the system is overloaded on average, variability in supply may help customers, on average. In this case, customers are advantaged by large realizations of $N(n)$. In the lower subplot of Figure 9, we plot $\mathbb{E}[Q(n)]$: We observe that $\mathbb{E}[Q(n)]$ is larger for a more variable supply (dashed curve is above solid curve). We also note that the queue length is small for $n\geq 100$.

To further investigate the improvement resulting from increased variability in supply in a system which is overloaded on average, we focus in Figure 10 on the case where $N(n)=n+n^q\cdot\epsilon$, for alternative values of $0<q<1$. In Figure 10, we calculate $\mathbb{P}(W(n)>0)$ and $\mathbb{E}[Q(n)]$ for each value of q, but hold $n=90$ fixed, with other parameters kept as in Figure 9. Figure 10 illustrates that as q increases, $\mathbb{P}(W(n)>0)$ decreases considerably, from over 0.85 for $q=0.2$ to below 0.6 for $q=0.95$. We ran similar experiments for systems that are underloaded on average, e.g., letting $n=110$ instead. As expected, $\mathbb{P}(W(n)>0)$ then increases instead from about 0.2 for $q=0.2$ to about 0.45 for $q=0.95$. Figure 10 illustrates that $\mathbb{E}[Q(n)]$ increases as q increases.

We close this section by presenting some explanations, grounded in theoretic results, for the numerical observations above. First, we note some properties of the expected steady-state queue length and probability of delay, as a function of the number of servers, n, in an $M/M/n+M$ queueing system where n is deterministic. We conjecture, but do not provide a proof, that the expected queue length is a convex function of n, whereas the convexity of the probability of delay depends on the value of n. Particularly, the probability of delay is concave for $n\ll\lambda/\mu$, and convex for $n\gg\lambda/\mu$ (see supporting Figures 21 and 22 in Appendix E.) We would expect that for a performance measure that is a convex function of n, more variability hurts performance. However, for a performance measure that is a concave function of n, more variability improves performance [5]. This is consistent with our numerical observations above: While the expected queue length

[5] Let $0\leq q_1<q_2\leq 1$, and $N_1=n+n^{q_1}\epsilon$ and $N_2=n+n^{q_2}\epsilon$ where $\mathbb{E}[\epsilon]=0$. Then one can show that $\mathbb{E}[\phi(N_1)]\leq\mathbb{E}[\phi(N_2)]$, for any convex function $\phi(\cdot)$. Indeed, we only need to prove that $\mathbb{E}[\phi(n^{q_1-q_2}\epsilon)]\leq\mathbb{E}[\phi(\epsilon)]$ for any convex $\phi(\cdot)$, and then the result follows as the composition of a convex function with a linear function is again convex. Note that $\mathbb{E}[\phi(n^{q_1-q_2}\epsilon)]=\int\phi(n^{q_1-q_2}x)f_\epsilon(x)dx\leq\int((1-n^{q_1-q_2})\phi(0)+n^{q_1-q_2}\phi(x))f_\epsilon(x)dx\leq(1-n^{q_1-q_2})\mathbb{E}[\phi(\epsilon)]+n^{q_1-q_2}\mathbb{E}[\phi(\epsilon)]=\mathbb{E}[\phi(\epsilon)]$, where we used Jensen's inequality for $\phi(0)\leq\mathbb{E}[\phi(\epsilon)]$.

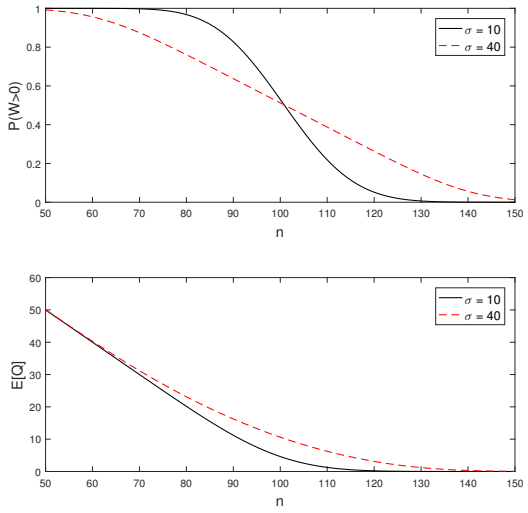

Figure 9 $\mathbb{P}(W(n) > 0)$ **and** $\mathbb{E}[Q(n)]$ **for** $N(n) \sim \text{Unif}(n-\sigma, n+\sigma)$.

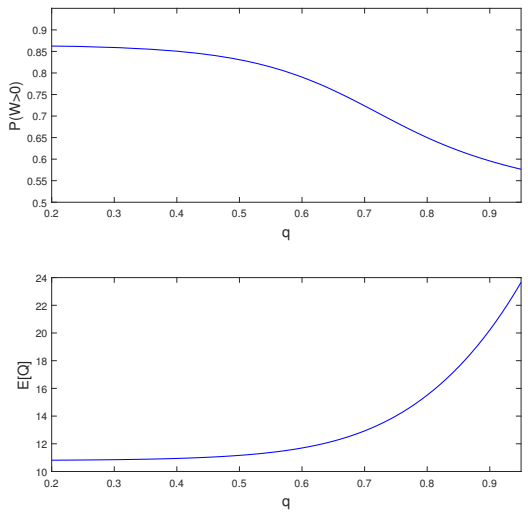

Figure 10 $\mathbb{P}(W(n) > 0)$ **and** $\mathbb{E}[Q(n)]$ **for** $N(n) \sim \text{Unif}(n-n^q, n+n^q)$.

increases as variability in the number of servers increases, the monotonicity of the probability of delay depends on whether the system is underloaded or overloaded, on average.

The numerical evidence above illustrates that the impact of variability in supply depends on the size of the agent pool. Thus, it is natural to investigate that impact in the case where the manager adjusts her pool size, optimally, in response to that variability in supply. We do so next.

6.2. Optimal Staffing Level

We now consider systems where the server pool size is according to the optimal staffing policy (8). That is, compared with §6.1, the agent pool size itself now also changes with the level of uncertainty in the number of available servers. We denote the optimal staffing level by $n^*_\lambda(q)$ to indicate, explicitly, its dependence on q. Building on the observations of §6.1, we highlight the following two cases where the dependence on the variability in supply is not straightforward:

(i) $n^*_\lambda(q) < \lambda/\mu$ **and is decreasing in** q**.** In this case, we note that if $n_\lambda(q)$ is held fixed, then the probability of delay will decrease as q increases. Now, as $n^*_\lambda(q)$ also decreases in q, the probability of delay may either decrease or increase in q. In particular, if $n^*_\lambda(q)$ decreases very slowly, then the probability of delay may still decrease in q. However, if it decreases very fast, then we expect the probability of delay to increase in q, because the system's congestion increases, on average, counteracting the improvement in performance observed with fixed $n_\lambda(q)$.

(ii) $n^*_\lambda(q) > \lambda/\mu$ **and is increasing in** q**.** In this case, if $n_\lambda(q)$ is held fixed, then the probability of delay will increase in q, but now as $n^*_\lambda(q)$ is also increasing in q, the probability of delay

may either increase or decrease in q depending on how fast $n^*_\lambda(q)$ increases. Specifically, if $n^*_\lambda(q)$ increases very slowly, then the probability of delay may still increase in q. However, if it increases fast enough, then the probability of delay will decrease in q.

We continue to assume that $\sigma_n = n^q$ for $0 < q < 1$. For the optimal staffing level $n^*_\lambda(q)$, we numerically solve the original problem in (8), where we let $N(n_\lambda) = n_\lambda + \sigma_{n_\lambda} \cdot \epsilon$, and assume that ϵ is uniformly distributed on $(-1, 1)$. We compute estimates for $\mathbb{P}(W(n^*_\lambda) > 0)$ based on the corresponding birth-and-death system equations, conditional on specific realizations for the random number of servers $N(n^*_\lambda)$; we then numerically integrate over the uniform distribution of $N(n^*_\lambda)$. We fix $\lambda = 100$ and $\mu = 1$, and consider two values of θ to study the impact of customer impatience. We also consider different values for the cost of the flexible resource, c_{flex}. Our numerical study illustrates that the impact of variability in supply on the quality of service is intricate, and that it strongly depends on the specific setting considered.

Customer impatience. In Figures 11 and 12, we plot $\mathbb{P}(W(n^*_\lambda) > 0)$ in the upper subplots and the optimal staffing level, n^*_λ, in the lower subplots. We keep all parameters fixed across the two figures, but vary the abandonment rate: In Figure 11, we assume that $\theta = 0.2$, and in Figure 12, we assume that $\theta = 4$. Contrasting Figures 11 and 12 illustrates the impact of customer impatience. In both cases, the manager responds to the increase in uncertainty, i.e., in q, by staffing a larger pool of flexible servers. However, the staffing level is smaller with more impatient customers since congestion costs are smaller in this case; i.e. $\beta = p + h/\theta$ is decreasing in θ. In Figure 11, the increase in the staffing level for larger q is large enough to cause a decrease in $\mathbb{P}(W(n^*_\lambda) > 0)$ as q increases. In contrast, Figure 12 illustrates the *non-monotonicity* of $\mathbb{P}(W(n^*_\lambda) > 0)$ as a function of $n^*_\lambda(q)$. In this case, for large values of q, $n^*_\lambda(q)$ does not increase fast enough, so that we observe an increase in $\mathbb{P}(W(n^*_\lambda) > 0)$ as q increases.

Staffing costs. In Figures 13 and 14, we keep all parameters as in Figure 11, but vary the cost of the flexible capacity: In Figure 13, we consider a cheap flexible capacity with $c_{flex} = 0.5$, and in Figure 14, we consider an expensive flexible capacity with $c_{flex} = 3.5$. Contrasting Figures 13 and 14 illustrates the effect of the cost of the flexible capacity on the quality of service experienced by customers. In particular, Figure 13 illustrates that when c_{flex} is small, the staffing level increases in q, and it increases fast enough so that $\mathbb{P}(W(n^*_\lambda) > 0)$ decreases in q. In contrast, Figure 14 shows that when c_{flex} is large, the staffing level decreases in q. For small values of q, $n^*_\lambda(q)$ decreases rather slowly, and $\mathbb{P}(W(n^*_\lambda) > 0)$ continues to decrease in q. But, for large values of q (e.g. $q > 0.75$), $n^*_\lambda(q)$ decreases too fast, leading to an increase in $\mathbb{P}(W(n^*_\lambda) > 0)$ as q increases.

6.3. Time-Varying Demand: Effect of Blending on the Quality of Service

So far, we focused on quantifying the quality of service when the manager staffs only flexible servers with stationary demand. We now turn to the case where she uses a blended workforce in a

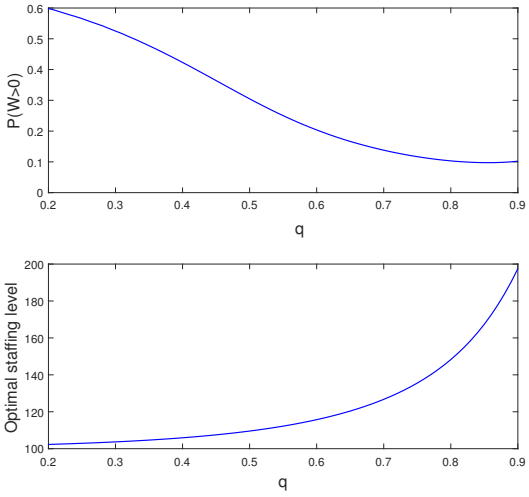

Figure 11 $\mathbb{P}(W(n_\lambda^*) > 0)$ **in the** $M/M/N(n_\lambda^*) + M$ **model with** $\lambda = 100$, $p = h = \mu = 1$, $c_{flex} = 1/4$, **and** $\theta = 0.2$.

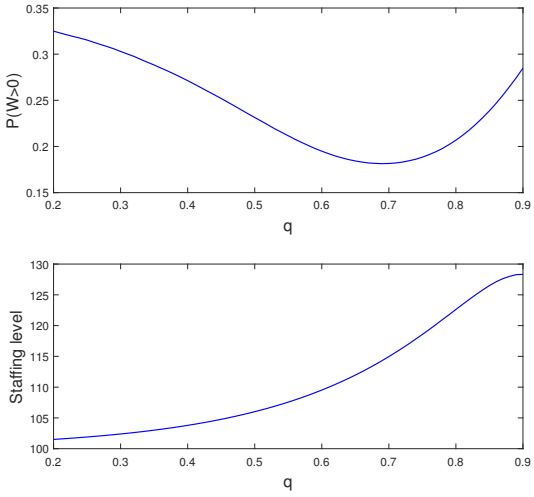

Figure 12 $\mathbb{P}(W(n_\lambda^*) > 0)$ **in the** $M/M/N(n_\lambda^*) + M$ **model with** $\lambda = 100$, $p = h = \mu = 1$, $c_{flex} = 1/4$, **and** $\theta = 4$.

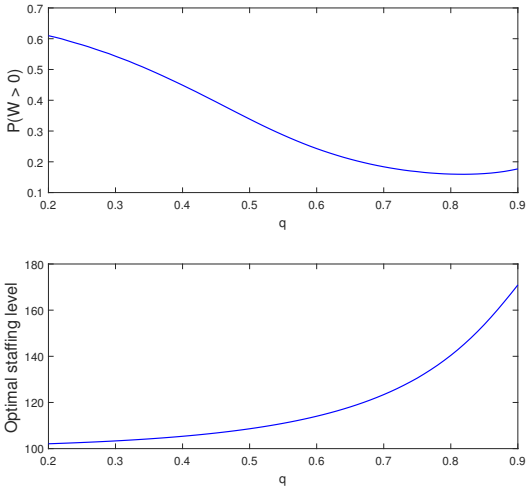

Figure 13 $\mathbb{P}(W(n_\lambda^*) > 0)$ **in the** $M/M/N(n_\lambda^*) + M$ **model with** $\lambda = 100$, $p = h = \mu = 1$, $c_{flex} = 1/2$, **and** $\theta = 0.2$.

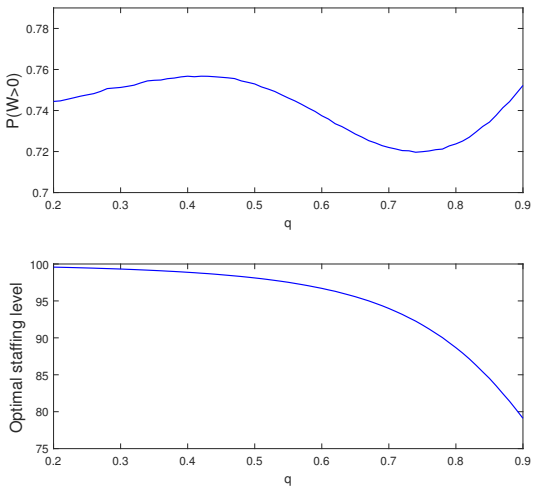

Figure 14 $\mathbb{P}(W(n_\lambda^*) > 0)$ **in the** $M/M/N(n_\lambda^*) + M$ **model with** $\lambda = 100$, $p = h = \mu = 1$, $c_{flex} = 3.5$, **and** $\theta = 0.2$.

time-varying setting. In Figures 15 and 16, we consider three scenarios: (i) the manager may use both flexible and fixed servers; (ii) the manager is restricted to using only the flexible resource, and that resource has the same per-unit cost as in (i); and (iii) the manager is restricted to using only the fixed resource, and that resource has the same per-unit cost as in (i). We plot both the optimal staffing levels for variants of our staffing problem under each alternative, as well as delay

probabilities in the high and low-demand periods. For the optimal staffing levels, we numerically solve the original staffing problem in (8). Figures 15 and 16 illustrate a case where the index k_0 in Lemma 1 is different than k_1 in Theorem 4. Indeed, for $q < 1$, the optimal solution in the stochastic-fluid problem is to blend the resources in the high period i.e, we have that $k_0 = 1$. On the other hand, for $q = 1$, we have that $k_1 = 2$, so that it is optimal not to use a blended workforce because of the extreme variability in the flexible resource when $q = 1$ (as can be seen in the figure, the staffing level for the flexible resource drops to 0 in the high period as q approaches 1).

Our objective is to illustrate the impact of blending the workforce on the quality of service offered in the system. For our choice of system parameters, the optimal staffing solution in problem (8) is to use a blended workforce, i.e., the solution in scenario (i) is the most cost-effective from the point of view of the manager; however, it is unclear whether customers will experience a higher quality of service under that scenario. Figures 15 and 16 illustrate that the impact of blending on customers depends on the period. In the low-demand period (Figure 15), the smallest delay probability corresponds to staffing solely from the fixed resource. This is because the manager staffs a high-enough level to match demand in the high period (Figure 16), which leaves the low-demand period overstaffed (we considered an alternative solution for this problem where the manager matches demand in the low-period instead, and reached the same conclusion). However, customers in the high-demand period are worse off when the manager staffs strictly from the fixed resource. This is because when flexible servers are used, the manager hedges against uncertainty by staffing a larger pool.

Figures 15 and 16 also show that blending may either help or hurt customers, compared with staffing from only the flexible resource. In the low-demand period, the uncertainty hedge is not large enough, so that customers are benefitted from blending (in the low demand period, only fixed servers are used, even when blending is allowed). In the high-demand period, the uncertainty hedge is large, and customers benefit from this. For the blended case, the hedge is not as large because of less uncertainty in the pool due to the presence of fixed servers.

7. General Abandonment

Our results so far are all under the assumption of exponentially-distributed patience times. Since there is statistical evidence indicating that patience times are typically not exponentially-distributed (Brown et al. 2005), it is important to go beyond that assumption. We do so in this section by describing results from a numerical study quantifying the optimality gaps for problems (9) and (11) with a non-exponential abandonment distribution.

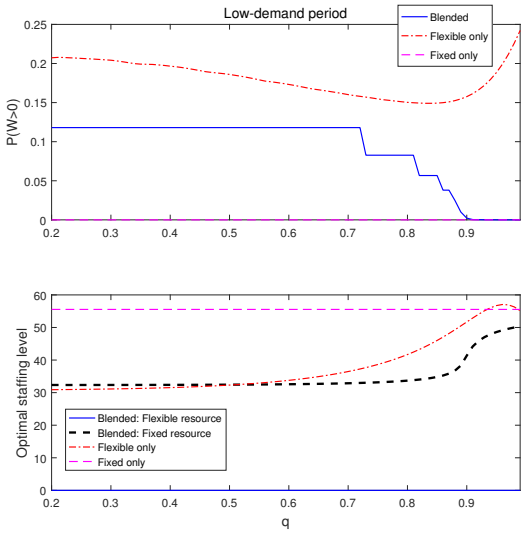

Figure 15 **Probability of delay and optimal staffing levels in the low-demand period.**

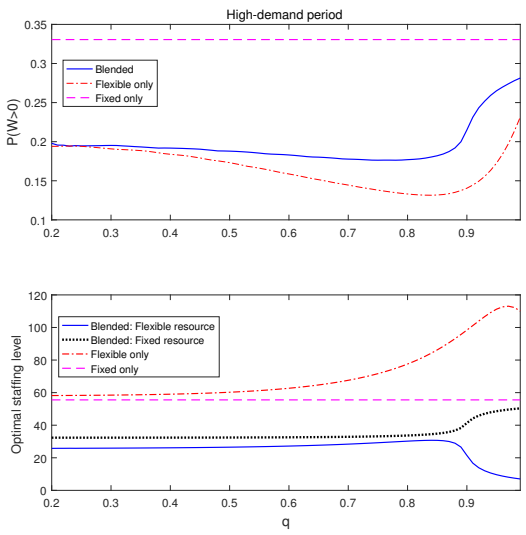

Figure 16 **Probability of delay and optimal staffing levels in the high-demand period.**

Staffing problem and relaxations. We begin by formulating the firm's optimization problem when times to abandon have a general distribution. As in (8), the firm's original problem is given by:

$$\min_{m_\lambda, n_\lambda^i} \Pi_\lambda(m_\lambda, n_\lambda^i) \equiv \min_{m_\lambda, n_\lambda^i} \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + h \cdot \mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] + r \cdot \Xi(m_\lambda, n_\lambda^i) \right).$$

Because of the difficulty in solving this staffing problem, we now describe both the stochastic-fluid and fluid relaxations of the problem. To do so, we let G denote the cdf of the abandonment-time distribution, $\bar{G}$ its ccdf, and g its pdf. We assume that g is strictly positive so that $\bar{G}$ is invertible. The stochastic-fluid formulation of the problem is given by:

$$\begin{aligned} &\min_{m_\lambda, n_\lambda^i} \bar{\Pi}_\lambda(m_\lambda, n_\lambda^i) \\ &\equiv \min_{m_\lambda, n_\lambda^i} \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + r\mathbb{E}[(\lambda_i - N(m_\lambda, n_\lambda^i)\mu)^+] + h\mathbb{E}\left[\int_0^{\bar{w}^i(m_\lambda, n_\lambda^i)} \lambda_i \bar{G}(u) du \right] \right), \end{aligned}$$

where $\bar{w}^i(m_\lambda, n_\lambda^i) = \bar{G}^{-1}(N(m_\lambda, n_\lambda^i)\mu/\lambda_i)$ denotes the stochastic-fluid approximation of the waiting time in the i^{th} period. Accordingly, the fluid approximation to the problem is given by:

$$\begin{aligned} &\min_{m_\lambda, n_\lambda^i} \tilde{\Pi}_\lambda(m_\lambda, n_\lambda^i) \\ &\equiv \min_{m_\lambda, n_\lambda^i} \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + r(\lambda_i - m_\lambda \mu - n_\lambda^i \mu)^+ + h \left(\int_0^{\tilde{w}^i(m_\lambda, n_\lambda^i)} \lambda_i \bar{G}(u) du \right) \right), \end{aligned}$$

where $\tilde{w}^i(m_\lambda, n_\lambda^i) = \bar{G}^{-1}((m_\lambda \mu + n_\lambda^i \mu)/\lambda_i)$ denotes the fluid approximation of the waiting time in the i^{th} period.

Numerical results. In Figures 17-20, we consider a problem with a single period. Moreover, we restrict attention to a setting with only flexible capacity, since our objective here is to quantify the accuracies of the alternative staffing-problem relaxations, and those accuracies depend on the variance of the pool of flexible servers. For the abandonment distribution, we consider Pareto (mean 1, shape 2) and Weibull (mean 1, shape 2) abandonment. We choose these two distributions because they exhibit, for those selected parameter values, different properties for their failure-rate functions: While the Pareto distribution had a decreasing failure rate, the Weibull distribution has an increasing failure rate. We consider the following cost parameters: $c_{flex} = 1,\ h = 1,\ p = 0.45$, and $\mu = 1$. For each distribution, we compare the fluid, $\tilde{n}_\lambda$, stochastic-fluid, $\bar{n}_\lambda$, and original, n^*_λ, optimal solutions, by plotting their respective differences for varying arrival rates.

We first discuss our numerical results with Pareto abandonment. In this case, the overloaded regime is asymptotically optimal at fluid scale, i.e., the optimal prescription is not to match mean demand and mean supply. Thus, we expect that fluid prescriptions should be extremely accurate, i.e., with absolute errors on the order of magnitude of $\mathcal{O}(1)$ (Bassamboo and Randhawa 2010). In other words, we expect that stochastic-fluid prescriptions would not lead to a substantial improvement over their fluid counterparts; this is confirmed by Figures 17 and 19, where we plot **unscaled** absolute differences $|\tilde{n}_\lambda - n^*_\lambda|$ and $|\bar{n}_\lambda - n^*_\lambda|$. It is unclear, a priori, how the fluid and stochastic-fluid solutions would perform when σ_n is large, i.e., is of an order of magnitude equal to $\mathcal{O}(n)$, because the uncertainty in the number of servers is on the same order as the offered load in this case. We find that, while stochastic-fluid approximations are more accurate in this case, the difference in performance in not too great either.

With Weibull abandonment, the fluid solution prescribes a critically-loaded regime, i.e., to match the mean demand and the mean supply. Because this is the same asymptotic regime prescribed with exponential abandonment, we expect the optimality gaps of our respective solutions to be close to those with exponential abandonment. Figures 18 and 20 confirm that this is indeed the case. In particular, the stochastic-fluid formulation is remarkably accurate, yielding an order of magnitude improvement over the fluid prescription (in most cases, n^*_λ and $\bar{n}_\lambda$ are indistinguishable). Moreover, Figures 18 and 20 show that the optimality gaps obtained are consistent with those reported in earlier sections of the paper, with exponentially-distributed abandonment.

8. Concluding Remarks

In this paper, we studied the problem of staffing a service system where the manager must decide on cost-minimizing levels of fixed and/ or flexible agents. Our analysis suggests that it may be cost-effective to staff either strictly one of the two resources, or to use a blended workforce instead, depending on the interaction between three competing factors: (i) operational costs in the system;

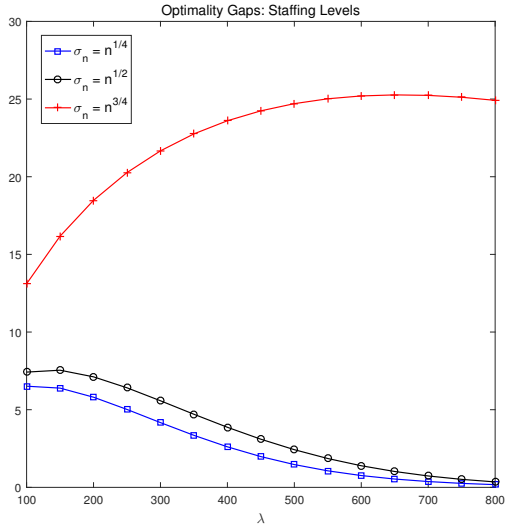

Figure 17 **Unscaled errors for optimal fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|$, as a function of λ, with Pareto abandonment.**

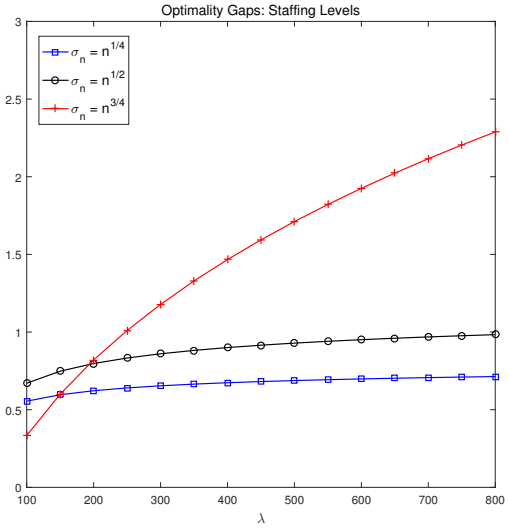

Figure 18 **Scaled errors for optimal fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ, with Weibull abandonment.**

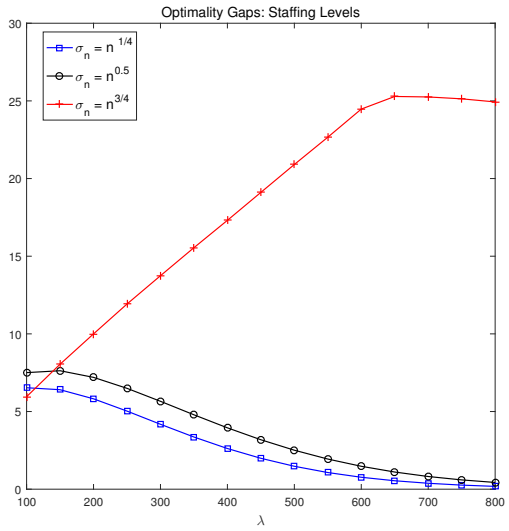

Figure 19 **Unscaled errors for optimal stochastic-fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|$, as a function of λ, with Pareto abandonment.**

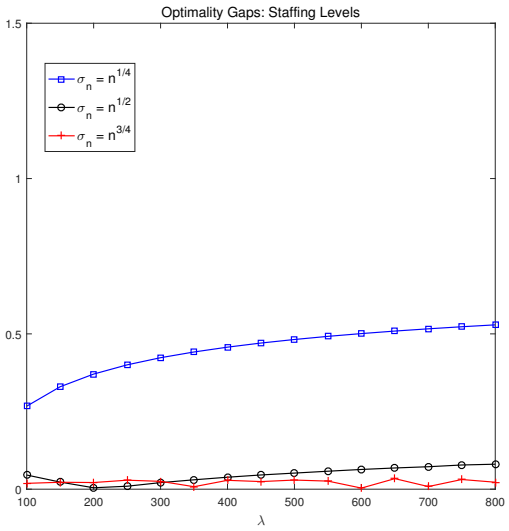

Figure 20 **Scaled errors for optimal stochastic-fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ, with Weibull abandonment.**

(ii) the time-variation in customer demand; and (iii) the supply-side uncertainty which is associated with staffing flexible agents. In broad terms, we showed that the optimal staffing levels involve both a base capacity, which is used to match mean demand, and an additional safety capacity which

hedges against both stochastic fluctuations in the system and the variability due to the randomness in supply. This additional safety capacity hedge may or may not be consistent with the square-root staffing hedge (Garnett et al. 2002), and generally depends on the mix of fixed and flexible resources in the staffed pool. We also investigated the impact of variability in the server pool on the quality of service experienced by customers and found that, perhaps contrary to intuition, that customers need not be disadvantaged by supply-side uncertainty, and may instead benefit from it.

In this work, we focused solely on the long-run staffing decision in the system. Our focus on that long-run strategic planning decision was motivated by: (1) the longer time scale which is associated with the staffing decision in practice, e.g., to allow for the training of agents, and (2) the fact that even though real-time pricing is used by some on-demand service platforms, such as ride-sharing services, most such platforms have to commit to the prices that they offer their agents well in advance (Taylor 2018). Nevertheless, it would be an interesting future research to investigate the dynamic compensation decision in a setting with a blended workforce.

Appendix A: Theorems 1 and 2: Optimality Gaps

We first state and prove some auxiliary lemmas (Lemmas 2-4), which will be useful for the proof of our theorems. The proofs follow similar lines of the arguments in Bassamboo et al. (2010). We consider the case with a single period first, and then extend to multiple periods in §A.3.

A.1. Additional Lemmas

LEMMA 2. *When* $\mu=\theta$,

$$\mathbb{E}[(\lambda/\mu - N(m_\lambda,n_\lambda))^+] \le \mathbb{E}[(X(m_\lambda,n_\lambda) - N(m_\lambda,n_\lambda))^+] \le \mathbb{E}[(\lambda/\mu - N(m_\lambda,n_\lambda))^+] + \mathcal{O}(\sqrt{\lambda}),$$

where $N(m_\lambda,n_\lambda) = m_\lambda + n_\lambda + \sigma_{n_\lambda}\epsilon$. *Moreover, if* $n_\lambda = \Theta(\lambda)$ *and* $\sigma_\lambda = \Theta(\lambda^q)$ *for* $q > 1/2$ *then:*

$$\mathbb{E}[(X(m_\lambda,n_\lambda) - N(m_\lambda,n_\lambda))^+] \le \mathbb{E}[(\lambda/\mu - N(m_\lambda,n_\lambda))^+] + \mathcal{O}\left(\lambda/\sigma_\lambda\right).$$

PROOF. When $\mu=\theta$, $X(m_\lambda,n_\lambda) \sim$Poisson$(\lambda/\mu)$. By Lemma 3 of Bassamboo et al. (2010):

$$\begin{aligned}\left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^+ &\le \mathbb{E}[(X(m_\lambda,n_\lambda) - N(m_\lambda,n_\lambda))^+ | N(m_\lambda,n_\lambda)] \\ &\le \left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^+ + \sqrt{\frac{4\pi\lambda}{\mu}} \exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^2\right) + \frac{1}{\log 2}\end{aligned}$$

Then as $\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^2\right) \le 1$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^+\right] \le\ & \mathbb{E}[X(m_\lambda,n_\lambda) - N(m_\lambda,n_\lambda))^+] \\ & \le \mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda,n_\lambda)\right)^+\right] + \mathcal{O}(\sqrt{\lambda}).\end{aligned}$$

For the second part of the lemma, we let $f_N^\lambda(s)$ denote the pdf of $N(m_\lambda,n_\lambda)$ and define, for $y \ge 0$:

$$M_\lambda(y) \equiv \sup_{y-\sqrt{\lambda}\log\lambda < s < y+\sqrt{\lambda}\log\lambda} \lambda f_N^\lambda(s).$$

We can then write:

$$\begin{aligned}
&E\left[\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-N(m_\lambda,n_\lambda)\right)^2\right)\right]\\
&=\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\quad+\int_{\lambda/\mu+\sqrt{\lambda}\log\lambda}^{\infty}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\quad+\int_{-\infty}^{\lambda/\mu-\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds.
\end{aligned}$$

Letting F_N^λ denote the cdf of $N(m_\lambda,n_\lambda)$, we see that: $F_N^\lambda(s)=\mathbb{P}\left(\epsilon\le\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)=F_\epsilon\left(\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)$. Thus, $f_N^\lambda(s)=\frac{1}{\sigma_{n_\lambda}}f_\epsilon\left(\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)$. That is, when $n_\lambda=\Theta(\lambda)$, it must be that $M_\lambda(y)=\mathcal{O}(\lambda/\sigma_\lambda)$. Now,

$$\begin{aligned}
&\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\le M_\lambda(\lambda/\mu)\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi}{\mu}}\frac{\sqrt{\lambda}}{\lambda}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)ds\\
&\le M_\lambda(\lambda/\mu)\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\frac{K_1}{\sqrt{\lambda}}\exp\left(-\frac{K_2}{\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)ds \text{ for some } K_1,K_2>0,\\
&=\mathcal{O}(\lambda/\sigma_\lambda).
\end{aligned}$$

In addition,

$$\begin{aligned}
&\int_{\lambda/\mu+\sqrt{\lambda}\log\lambda}^{\infty}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\le\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\lambda(\log\lambda)^2\right)\\
&=o(1).
\end{aligned}$$

Similarly,

$$\int_{-\infty}^{\lambda/\mu-\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds=o(1).$$

Thus, if $n_\lambda=\Theta(\lambda)$ then:

$$\mathbb{E}\left[(X(m_\lambda,n_\lambda)-N(m_\lambda,n_\lambda))^+\right]\le\mathbb{E}\left[(\lambda/\mu-N(m_\lambda,n_\lambda))^+\right]+\mathcal{O}(\lambda/\sigma_\lambda).$$

■

Lemma 3.

$$\bar{\Pi}_\lambda(m_\lambda,n_\lambda)\le\Pi_\lambda(m_\lambda,n_\lambda)\le\bar{\Pi}_\lambda(m_\lambda,n_\lambda)+\mathcal{O}(\sqrt{\lambda})$$

Moreover, when $n_\lambda=\Theta(\lambda)$ *and* $\sigma_\lambda=\Theta(\lambda^q)$ *for* $q>1/2$*:*

$$\bar{\Pi}_\lambda(m_\lambda,n_\lambda)\le\Pi_\lambda(m_\lambda,n_\lambda)\le\bar{\Pi}_\lambda(m_\lambda,n_\lambda)+O\left(\lambda/\sigma_\lambda\right).$$

PROOF. We prove the first statement in detail. The second statement, where $n_\lambda = \Theta(\lambda)$ and $\sigma_{n_\lambda} \geq \sqrt{n_\lambda}$, follows along the same line of arguments.

When $\mu = \theta$, the result follows directly from Lemma 2.

When $\mu > \theta$**,** we first consider an auxiliary "upper bound" system with abandonment rate μ. On each sample path, we assume that the two systems have the same (randomly drawn) number of servers. Let $A_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \mu, \theta) - N(m_\lambda, n_\lambda))^+]$ and $A^I_\lambda(N(m_\lambda, n_\lambda)) \equiv \mu\mathbb{E}[(X(m_\lambda, n_\lambda; \mu, \mu) - N(m_\lambda, n_\lambda))^+]$ where $X(m_\lambda, n_\lambda; x, y)$ is the steady-state number-in-system with service rate x and abandonment rate y. As $A_\lambda(N(m_\lambda, n_\lambda)) \leq A^I_\lambda(N(m_\lambda, n_\lambda))$:

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\leq c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^I_\lambda(N(m_\lambda, n_\lambda)) \\
&\leq c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] + \mathcal{O}(\sqrt{\lambda}) \quad \text{by Lemma 2} \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

We then consider an auxiliary "lower bound" system with service rate θ. On each sample path, we assume that the two systems have the same (randomly drawn) number of servers. Let $A^{II}_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \geq A^{II}_\lambda(N(m_\lambda, n_\lambda)\mu/\theta)$ (Bassamboo et al. 2010):

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda\left(N(m_\lambda, n_\lambda)\right) \\
&\geq c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\geq c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] \quad \text{by Lemma 2} \\
&= c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}\left[(\lambda/\mu - N(m_\lambda, n_\lambda))^+\right] \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}$$

When $\mu < \theta$**,** the proof is similar to the case of $\mu > \theta$. We first consider an auxiliary "upper bound" system with service rate θ. Let $A^{II}_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \leq A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)$:

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\leq c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\leq c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} m_\lambda - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] + \mathcal{O}(\sqrt{\lambda}) \text{ by Lemma 2} \\
&= c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

We then consider an auxiliary "lower upper" bound system with abandonment rate μ. Let $A_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(N(m_\lambda, n_\lambda); \mu, \theta) - N(m_\lambda, n_\lambda))^+]$ and $A^I_\lambda(N(m_\lambda, n_\lambda)) \equiv \mu\mathbb{E}[(X(N(m_\lambda, n_\lambda); \mu, \mu) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \geq A^I_\lambda(N(m_\lambda, n_\lambda))$:

$$\Pi_\lambda(m_\lambda, n_\lambda) = c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda))$$

$$\begin{aligned}
&\geq c_{fix}m_\lambda + c_{flex}n_\lambda + (h/\theta + r)A^I_\lambda(N(m_\lambda, n_\lambda)) \\
&\geq c_{fix}m_\lambda + c_{flex}n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \quad \text{by Lemma 2} \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}$$

■

Lemma 4.

$$\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ \leq \mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] \leq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ + \mathcal{O}(\sigma_{n_\lambda}).$$

Proof. We notice that by Jensen's inequality,

$$\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] \geq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+.$$

For the upper bound, as $-1 < \epsilon < 1$,

$$\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] = \mathbb{E}\left[\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon\right)^+\right] =$$

$$\begin{cases}
0 & \text{for } \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} < -1, \\
\sigma_{n_\lambda}\int_{-1}^{\frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}}} F_\epsilon(x)dx & \text{for } -1 \leq \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} \leq 1, \\
\lambda/\mu - m_\lambda - n_\lambda & \text{for } \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} > 1.
\end{cases}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] &= \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right) \cdot \mathbf{1}\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda > \sigma_{n_\lambda}\right) + \mathcal{O}(\sigma_{n_\lambda}) \\
&\leq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ + \mathcal{O}(\sigma_{n_\lambda}).
\end{aligned}$$

■

A.2. Proof of Theorems 1 and 2

From Lemma 3, we have

$$\begin{aligned}
\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) &\leq \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&\leq \bar{\Pi}_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&\leq \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\sqrt{\lambda});
\end{aligned}$$

Moreover, if $\bar{n}_\lambda = \Theta(\lambda)$ and $\sigma_\lambda = \Theta(\lambda^q)$ for $q > 1/2$:

$$\begin{aligned}
\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) &\leq \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda) \\
&\leq \bar{\Pi}_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda) \\
&\leq \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda).
\end{aligned}$$

From Lemmas 3 & 4, we have:

$$\begin{aligned}
\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) &\le \bar{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\tilde{n}_\lambda}\}) \\
&\le \tilde{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\tilde{n}_\lambda}\}) + \mathcal{O}(\sigma_{\tilde{n}_\lambda}) \\
&\le \tilde{\Pi}_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\tilde{n}_\lambda}\}) + \mathcal{O}(\sigma_{\tilde{n}_\lambda}) \\
&\le \bar{\Pi}_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\tilde{n}_\lambda}\}) + \mathcal{O}(\sigma_{\tilde{n}_\lambda}) \\
&\le \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\tilde{n}_\lambda}\}) + \mathcal{O}(\sigma_{\tilde{n}_\lambda}).
\end{aligned}$$

In particular, if $\tilde{n}_\lambda = \Theta(\lambda)$ then

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) \le \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\max\{\sqrt{\lambda}, \sigma_\lambda\}).$$

A.3. Asymptotic Accuracy with Time-Varying Demand

Theorems 1 and 2 are generalizable to the multi-period case. In particular, following Lemmas 3 and 4, we have:

THEOREM 5. *For large λ and time-varying demand,*

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\max\{\sqrt{\lambda}, \sigma_\lambda\}\right); \tag{20}$$

and

$$\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\sqrt{\lambda}\right). \tag{21}$$

Note that the orders of magnitude of errors in the theorem are only upper bounds. In particular, if $\tilde{n}_\lambda = 0$, then the order of accuracy in (20) is $\mathcal{O}(\sqrt{\lambda})$ instead. Moreover, if $\bar{m}_\lambda = 0$, then the order of accuracy in (21) is $\mathcal{O}(\lambda/\sigma_\lambda)$ instead.

Appendix B: Lemma 1: Fluid Solution with Multiple Periods

In this section, we shall omit the index λ when there is no confusion. If $c_{flex} < c_{fix}$, then $n_i = \lambda_i$ for all i, and $m = 0$. Consider that $c_{flex} > c_{fix}$.

Recall that $c_{flex}, c_{fix} < \beta$ for β in (12). λ_i's are arranged in an increasing order.

Fix m and solve for n_i. The problem for period i where we drop T_i is:

$$\min_{n_i} c_{flex} n_i + \beta(\lambda_i - m - n_i)^+.$$

The solution is:

- If $\lambda_i < m$ then $n_i = 0$;
- If $\lambda_i \ge m$ then $n_i = \lambda_i - m$.

Solve for m. Plugging in the above solution, the problem becomes:

$$\begin{aligned}
&\min_m \quad \left(\sum_{\{i:\lambda_i<m\}} T_i c_{fix} m + \sum_{\{i:\lambda_i\geq m\}} T_i[c_{fix}m + c_{flex}(\lambda_i - m)] \right) \\
\equiv &\min_m \quad \left(\sum_{i=1}^{k} T_i c_{fix} m - \sum_{\{i:\lambda_i\geq m\}} T_i c_{flex} m + \sum_{\{i:\lambda_i\geq m\}} T_i c_{flex}\lambda_i \right) \\
\equiv &\min_m \quad \left(m \cdot \left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{\{i:\lambda_i\geq m\}} T_i c_{flex} \right) + \sum_{\{i:\lambda_i\geq m\}} T_i c_{flex}\lambda_i \right).
\end{aligned}$$

It is easy to see that there exist $1 \leq k_0 \leq k$ such that:

$$\left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{i=k_0}^{k} T_i c_{flex} \right) \leq 0 \quad \text{and} \quad \left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{i=k_0+1}^{k} T_i c_{flex} \right) > 0.$$

The optimal solution is given by:

$$\begin{cases} \tilde{m} = \lambda_{k_0} & \\ \tilde{n}_i = 0 & \text{for } i \leq k_0 \\ \tilde{n}_i = \lambda_i - \lambda_{k_0} & \text{for } i > k_0. \end{cases}$$

That is, we use only the fixed capacity in low-demand periods, and we blend in the higher-demand periods. In the special case when $c_{fix} \leq \frac{T_k}{\sum_{i=1}^{k} T_i} c_{flex}$, we set $\tilde{m} = \lambda_k$ and $\tilde{n}_i = 0$ for all i.

Appendix C: Theorems 3 and Corollary 1: Stochastic-Fluid Problem with a Single Period

Recall that we denote by $\bar{n}_\lambda$ and $\bar{m}_\lambda$ the optimal solutions of $\bar{\Pi}_\lambda$; we shall drop the index λ when doing so does not cause any confusion. The statement of Theorem 3 focuses on the flexible resource only. Here, we present the proof for a system with both flexible and fixed resources, and a single period. The proof of the statement in Theorem 3 follows immediately by assuming that the price of the fixed resource is larger than that of the flexible resource; this corresponds to sections C.2., C.3., and part of C.4. below.

C.1. $c_{fix} \leq c_{flex}$.

In any optimal solution, we must have that $\bar{n} = 0$. Aiming at a contradiction, assume that $\bar{n} > 0$. Then, for β in (12):

$$\begin{aligned}
\bar{\Pi}(\bar{m}+\bar{n}, 0) &= c_{fix}(\bar{m}+\bar{n}) + \beta(\lambda - \bar{m} - \bar{n})^+ \\
&< c_{fix}\bar{m} + c_{flex}\bar{n} + \beta\mathbb{E}[(\lambda - \bar{m} - \bar{n} - \sigma_{\bar{n}} \cdot \epsilon)^+] = \bar{\Pi}(\bar{m}, \bar{n}).
\end{aligned}$$

The inequality follows from the fact that $c_{fix} < c_{flex}$ and Jensen's inequality. Fixing $\bar{n} = 0$, we have $\bar{\Pi}(m, 0) = \tilde{\Pi}(m, 0)$. Thus, a unique optimal solution exists, and is the solution to the fluid problem.

C.2. $c_{fix} > c_{flex}$ and $0 \leq q \leq 1/2$.

Plugging in the fluid optimal solution, i.e., setting $m = 0$ and $n = \lambda/\mu$, we have

$$\Pi(m, n) = \Pi(m^*, n^*) + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_\lambda) = \Pi(m^*, n^*) + \mathcal{O}(\sqrt{\lambda}),$$

where the first equality follows from Theorem 1. Thus, $(\bar{m}, \bar{n}) = (0, \lambda/\mu)$ is optimal. **This proves Case I in Theorem 3**.

C.3. $c_{fix} > c_{flex}$ **and** $1/2 < q < 1$**.**

We consider relatively large systems with

$$(c_{fix} - c_{flex})\lambda/\mu > \sigma_{\lambda/\mu}\mathbb{E}[(-\epsilon)^+]. \tag{22}$$

We first show that $\bar{n}$ is $\Theta(\lambda/\mu)$ and $\bar{m}$ is $\mathcal{O}(\sigma_{\lambda/\mu})$. The proof of this part is divided into four steps.

Step 1. Show that $\bar{n} > 0$**.** For $n = 0$, under assumption (22), we have

$$\bar{\Pi}(m, 0) \geq \bar{\Pi}(\lambda/\mu, 0) > \bar{\Pi}(0, \lambda/\mu).$$

Thus, we must have $\bar{n} > 0$.

Step 2. Show that $\bar{n} \leq \lambda/\mu + \frac{\beta\mathbb{E}[(-\epsilon)^+]}{c_{flex}}\sigma_{\lambda/\mu}$**.** To see why, assume that $\bar{n} > \lambda/\mu + \frac{\beta\mathbb{E}[(-\epsilon)^+]}{c_{flex}}\sigma_{\lambda/\mu}$. Then,

$$\bar{\Pi}(m, \bar{n}) > c_{flex}\bar{n} > c_{flex}\frac{\lambda}{\mu} + \beta\mathbb{E}[(-\epsilon)^+]\sigma_{\lambda/\mu} = \bar{\Pi}(0, \lambda/\mu).$$

We thus get a contradiction.

Step 3. Show that $\bar{n} \geq \lambda/\mu - \frac{\beta\mathbb{E}[(-\epsilon)^+] + c_{fix}\max\{F_\epsilon^{-1}(c_{fix}/\beta), 0\}}{c_{fix} - c_{flex}}\sigma_{\lambda/\mu}$**.**

For fixed n, let $m^*(n) = \arg\min_{m \geq 0} \bar{\Pi}(m, n)$. In particular,

- if $\frac{\lambda/\mu - n}{\sigma_n} > F_\epsilon^{-1}(c_{fix}/\beta)$, then $m^*(n) = \lambda/\mu - n - F_\epsilon^{-1}(c_{fix}/\beta)\sigma_n$;
- else, $m^*(n) = 0$.

In what follows, we shall suppress the dependence of m^* on n where there is no confusion. Assume $n^* < \lambda/\mu - \frac{\beta\mathbb{E}[(-\epsilon)^+] + c_{fix}\max\{F_\epsilon^{-1}(c_{fix}/\beta), 0\}}{c_{fix} - c_{flex}}\sigma_{\lambda/\mu}$. If $m^* = 0$,

$$\begin{aligned}
\bar{\Pi}(m^*, n^*) &= \bar{\Pi}(0, n^*) \\
&\geq c_{flex}n^* + \beta(\lambda/\mu - n^*) \\
&> c_{flex}\lambda/\mu + (\beta - c_{flex})\frac{\beta\mathbb{E}[(-\epsilon)^+] + c_{fix}\max\{F_\epsilon^{-1}(c_{fix}/\beta), 0\}}{c_{fix} - c_{flex}}\sigma_{\lambda/\mu} \\
&> c_{flex}\lambda/\mu + \beta\mathbb{E}[(-\epsilon)^+]\sigma_{\lambda/\mu} = \bar{\Pi}(0, \lambda/\mu).
\end{aligned}$$

We thus get a contradiction.

If $m^* = \lambda/\mu - n^* - F_\epsilon^{-1}(c_{fix}/\beta)\sigma_{n^*}$:

- When $F_\epsilon^{-1}(c_{fix}/\beta) > 0$,

$$\begin{aligned}
\bar{\Pi}(m^*, n^*) &> c_{fix}(\lambda/\mu - n^*) - c_{fix}F_\epsilon^{-1}(c_{fix}/\beta)\sigma_{\lambda/\mu} + c_{flex}n^* \\
&> c_{flex}(\lambda/\mu - n^*) + \beta\mathbb{E}[(-\epsilon)^+]\sigma_{\lambda/\mu} + c_{flex}n^* = \bar{\Pi}(0, \lambda/\mu).
\end{aligned}$$

- When $F_\epsilon^{-1}(c_{fix}/\beta) \leq 0$,

$$\begin{aligned}
\bar{\Pi}(m^*, n^*) &> c_{fix}(\lambda/\mu - n^*) + c_{flex}n^* \\
&> c_{flex}\lambda/\mu + \beta\mathbb{E}[(-\epsilon)^+]\sigma_{\lambda/\mu} = \bar{\Pi}(0, \lambda/\mu).
\end{aligned}$$

We get again a contradiction.

Step 4. Show that $\bar{m} \leq \left(\frac{\beta\mathbb{E}[(-\epsilon)^+]}{c_{fix} - c_{flex}} + 2|F_\epsilon^{-1}(c_{fix}/\beta)|\right)\sigma_{\lambda/\mu}$. This follows from the analysis in Step 3.

Based on the analysis above, we write

$$m = x\sigma_n \text{ and } n = \lambda/\mu + y\sigma_n.$$

We notice that there is a one-to-one correspondence between (x, y) and (m, n). Then, minimizing $\bar{\Pi}(m, n)$ is equivalent to minimizing

$$\Gamma(x, y) = \left(c_{fix}x + c_{flex}y + \beta\mathbb{E}[(-x - y - \epsilon)^+]\right) s_\lambda(y)$$

where $s_\lambda(y) = \sigma_n$. We denote the optimal solution to $\min_{x,y} \Gamma(x, y)$ as $(\bar{x}_\lambda, \bar{y}_\lambda)$. Steps 1-4 indicate that we can find $0 < M < \infty$, such that for all λ,

$$0 \le \bar{x}_\lambda < M \text{ and } -M < \bar{y}_\lambda < M.$$

This implies that we will use the cheaper flexible capacity to meet the mean demand. The more expensive fixed capacity, if used, is only used to hedge the uncertainty in the flexible capacity.

We now focus on $\min_{x,y} \Gamma(x, y)$. We first notice that the optimal solution to

$$\min_{x,y} c_{fix}x + c_{flex}y + \beta\mathbb{E}[(-x - y - \epsilon)^+]$$

is $(0, -F_\epsilon^{-1}(c_{flex}/\beta))$. We also observe that

$$\frac{\partial\Gamma(x, y)}{\partial x} = (c_{fix} - \beta F_\epsilon(-x - y))s_\lambda(y), \quad \frac{\partial^2\Gamma(x, y)}{\partial x^2} = \beta f_\epsilon(-x - y)s_\lambda(y) \ge 0$$

and

$$\frac{\partial\Gamma(x, y)}{\partial y} = (c_{flex} - \beta F_\epsilon(-x - y))s_\lambda(y) + (c_{fix}x + c_{flex}y + \beta\mathbb{E}[(-x - y - \epsilon)^+])s'_\lambda(y),$$

$$\frac{\partial^2\Gamma(x, y)}{\partial y^2} = \beta f_\epsilon(-x - y)s_\lambda(y) + 2(c_{flex} - \beta F_\epsilon(-x - y))s'_\lambda(y) + (c_{fix}x + c_{flex}y + \beta\mathbb{E}[(-x - y - \epsilon)^+])s''_\lambda(y).$$

In particular, $\Gamma(x, y)$ is convex in x, but may not be convex in y. We next analyze the sign of $\frac{\partial^2\Gamma(x,y)}{\partial y^2}$ for large systems. Recall that $\sigma_n = an^q$. Without loss of generality, we set $a = 1$ for the ease of exposition. Then, we can write $n(y) = \lambda/\mu + y\sigma_{n(y)} = \lambda/\mu + yn(y)^q$. We first notice that $n(y) = \Theta(\lambda)$. As

$$n'(y) = n(y)^q + qy \cdot n'(y) \cdot n(y)^{q-1},$$

$n'(y) = \frac{n(y)^q}{1 - qy \cdot n(y)^{q-1}} = \Theta(\lambda^q)$ and $s'(y) = q \cdot n(y)^{q-1} \cdot n'(y) = \Theta(\lambda^{2q-1})$. Similarly, as

$$n''(y) = \frac{2qn'(y) \cdot n(y)^{q-1} + q(q-1)y(n'(y))^2 \cdot n(y)^{q-2}}{1 - qyn(y)^{q-1}},$$

$n''(y) = \Theta(\lambda^{2q-1})$. In addition, as $s''_\lambda(y) = qn''(y)n(y)^{q-1} + q(q-1)(n'(y))^2 n(y)^{q-2}$, we have $|s''_\lambda(y)| = \Theta(\lambda^{3q-2})$. This leads us to make the following observations for λ large enough.

i) The first term in $\frac{\partial^2\Gamma(x,y)}{\partial y^2}$ will dominate. Thus, $\frac{\partial^2\Gamma(x,y)}{\partial y^2} \ge 0$ for $y \in [-M, M]$. Then there is a unique optimal solution, which satisfies $\frac{\partial\Gamma(x,y)}{\partial y} = 0$.

ii) As $c_{flex} < c_{fix}$, when $\frac{\partial\Gamma(x,y)}{\partial y} = 0$, we have that $\frac{\partial\Gamma(x,y)}{\partial x} > 0$. Thus, $\bar{x}_\lambda = 0$, and $\bar{y}$ is solution of

$$\beta F_\epsilon(-y) = c_{flex} + \left(c_{flex}y + \beta\mathbb{E}[(-y - \epsilon)^+]\right) \frac{s'_\lambda(y)}{s_\lambda(y)}. \tag{23}$$

Notice that $-M \le y \le M$ and that $\frac{s'_\lambda(y)}{s_\lambda(y)} = \Theta(\lambda^{q-1})$. Applying Taylor expansion to $F_\epsilon^{-1}(\cdot)$, we obtain that

$$\bar{y}_\lambda = -F_\epsilon^{-1}(c_{flex}/\beta) + \mathcal{O}(\lambda^{q-1}). \tag{24}$$

Based on these observations, and applying Taylor expansion to $\Gamma(0,\bar{y}_\lambda)$ around $y_1^* := -F_\epsilon^{-1}(c_{flex}/\beta)$, we have

$$\begin{aligned}\Gamma(0,\bar{y}_\lambda) &= \Gamma(0,y_1^*) + \frac{\partial\Gamma(0,y_1^*)}{\partial y}(\bar{y}_\lambda - y_1^*) + \mathcal{O}(\lambda^{3q-2}) \\ &= \Gamma(0,-F_\epsilon^{-1}(c_{flex}/\beta)) + \mathcal{O}(\lambda^{3q-2}). \end{aligned} \tag{25}$$

To get the second equality, we notice that $(c_{flex} - \beta F_\epsilon(-y_1^*))s_\lambda(y_1^*) = 0$.

In particular, we can consider two cases:

- Moderately uncertainty-dominated regime: $1/2 < q \le 3/4$. In this case, $3q-2 < 1-q$. Thus, we do not need to worry about the $\mathcal{O}(\lambda^{3q-2})$ term in (25). Using $n_1 = \lambda/\mu + y_1^*\sigma_{n_1}$ is $\mathcal{O}(\lambda/\sigma_\lambda)$ optimal. In addition, if instead of $n_1 = \lambda/\mu + y_1^*\sigma_{n_1}$, we use $\hat{n} = \lambda/\mu + y_1^*\sigma_{\lambda/\mu}$. We write $\hat{n} = \lambda/\mu + \hat{y}\sigma_{\hat{n}}$, then

$$y_1^* - \hat{y} = \frac{y_1^*(\sigma_{\hat{n}} - \sigma_{\lambda/\mu})}{\sigma_{\hat{n}}} = \mathcal{O}\left(\lambda^{q-1}\right).$$

  Applying again Taylor expansion to $\Gamma(0,\hat{y})$ around y_1^*, we have

$$\Gamma(0,\hat{y}) = \Gamma(0,y_1^*) + \mathcal{O}(\lambda^{3q-2}).$$

  Thus, using $\hat{n} = \lambda/\mu + y_1^*\sigma_{\lambda/\mu}$ is also $\mathcal{O}(\lambda/\sigma_\lambda)$ optimal. **This proves both Case II in Theorem 3 and Corollary 1.**

- Strongly uncertainty-dominated regime: $3/4 < q < 1$. When $3/4 < q < 1$, we should solve (14) exactly to achieve $\mathcal{O}(\lambda/\sigma_\lambda)$ optimality. **This proves Case III in Theorem 3.**

C.4. Case IV in Theorem 3: $\sigma_n = an$, i.e., $q = 1$, for $a < 1$.

Let $m_\lambda = x\lambda/\mu$, $n_\lambda = y\lambda/\mu$ for $x, y \in \mathbb{R}^+$. Then, optimizing $\bar{\Pi}_\lambda(m_\lambda, n_\lambda)$ is equivalent to optimizing:

$$V(x,y) = \frac{\bar{\Pi}_\lambda(m_\lambda,n_\lambda)}{\lambda/\mu} = c_{fix}x + c_{flex}y + \beta\mathbb{E}\left[(1-x-y-ay\epsilon)^+\right].$$

We denote the optimal solution to $V(x,y)$ as x^*, y^*. It suffices to consider x, y such that $-ay \le 1-x-y \le ay$.

That is, we need to solve:

$$\min_{x,y} V(x,y) \tag{26}$$

where

$$V(x,y) = c_{fix}x + c_{flex}y + \beta ay \int_{-1}^{\frac{1-x-y}{ay}} F_\epsilon(u)du$$

We first notice that for fixed y, we have

$$\begin{aligned}\frac{\partial V(x,y)}{\partial x} &= c_{fix} - \beta F_\epsilon\left(\frac{1-x-y}{ay}\right), \\ \frac{\partial^2 V(x,y)}{\partial x^2} &= \beta\frac{1}{ay} f_\epsilon\left(\frac{1-x-y}{ay}\right) \ge 0 \text{ implying that } V(x,y) \text{ is convex in } x.\end{aligned}$$

Let $\beta_1 \equiv c_{fix}/\beta$. Then, we divide the analysis into two cases:

- **Case a:** If $1 - y - ayF_\epsilon^{-1}(\beta_1) \ge 0$, $x^*(y) = 1 - y - ayF_\epsilon^{-1}(\beta_1)$.
- **Case b:** If $1 - y - ayF_\epsilon^{-1}(\beta_1) \le 0$, $x^*(y) = 0$.

In **Case a**, we have $\frac{1-x^*(y)-y}{ay} = F_\epsilon^{-1}(\beta_1)$. Then, we find y that minimizes:

$$\Xi_1(y) \equiv V(x^*(y), y) = c_{fix} + \left(c_{flex} - c_{fix} - c_{fix} a F_\epsilon^{-1}(\beta_1) + a\beta \int_{-1}^{F_\epsilon^{-1}(\beta_1)} F_\epsilon(u)du \right) y,$$

which is readily seen to be linear in y. Thus, if

$$c_{flex} > c_{fix} + c_{fix} a F_\epsilon^{-1}(\beta_1) - a\beta \int_{-1}^{F_\epsilon^{-1}(\beta_1)} F_\epsilon(u)du,$$

then $x^* = 1$, $y^* = 0$; otherwise, $x^* = 0$, $y^* = (1 + aF_\epsilon^{-1}(\beta_1))^{-1}$ (to be in Case a).

In **Case b**, we find y that minimizes:

$$\Xi_2(y) \equiv V(0, y) = c_{flex} y + \beta a y \int_{-1}^{1/(ay)-1/a} F_\epsilon(u)du.$$

We notice that

$$\Xi_2'(y) = c_{flex} + \beta a \int_{-1}^{1/(ay)-1/a} F_\epsilon(x)dx - \frac{\beta}{y} F_\epsilon\left(\frac{1}{ay} - \frac{1}{a}\right)$$

and

$$\Xi_2''(y) = \frac{\beta}{ay^3} f_\epsilon\left(\frac{1}{ay} - \frac{1}{a}\right) > 0 \text{ implying that } V(0, y) \text{ is convex in } y.$$

To be in Case b, we must have that $y \geq (1 + aF_\epsilon^{-1}(\beta_1))^{-1}$. If $\Xi_2'\left(\frac{1}{1+aF_\epsilon^{-1}(\beta_1)}\right) \geq 0$ then $y^* = \frac{1}{1+aF_\epsilon^{-1}(\beta_1)}$. Note that $\Xi_2'\left(\frac{1}{1+aF_\epsilon^{-1}(\beta_1)}\right) \geq 0$ is equivalent to

$$c_{flex} > c_{fix} + c_{fix} a F_\epsilon^{-1}(\beta_1) - a\beta \int_{-1}^{F_\epsilon^{-1}(\beta_1)} F_\epsilon(u)du, \tag{27}$$

in which case we need to compare $\Xi_1(0)$ to $\Xi_2\left(\frac{1}{1+aF_\epsilon^{-1}(\beta_1)}\right)$ to find the overall optimal value of $V(x, y)$. It is readily seen that if (27) holds, then $\Xi_1(0) < \Xi_2\left(\frac{1}{1+aF_\epsilon^{-1}(\beta_1)}\right)$, so that $y^* = 0$ in this case. Otherwise, y^* is the solution of

$$\Xi_2'(y) = c_{flex} + \beta a \int_{-1}^{1/(ay)-1/a} F_\epsilon(u)du - \frac{\beta}{y} F_\epsilon\left(\frac{1}{ay} - \frac{1}{a}\right) = 0. \tag{28}$$

For the solution to (28), we observe that

$$\Xi_2(y^*) \leq \Xi_2\left(\frac{1}{1 + aF_\epsilon^{-1}(\beta_1)}\right) = \Xi_1\left(\frac{1}{aF_\epsilon^{-1}(\beta_1) + 1}\right).$$

Therefore, the solution to (26) is given by:

- If $c_{flex} > c_{fix} + c_{fix} a F_\epsilon^{-1}(\beta_1) - a\beta \int_{-1}^{F_\epsilon^{-1}(\beta_1)} F_\epsilon(x)dx$, then: $x^* = 1$ and $y^* = 0$;
- Otherwise, $x^* = 0$ and y^* solves (28).

This proves case IV in Theorem 3.

Appendix D: Theorem 4: Stochastic-Fluid Problem with Multiple Periods

D.1. Case 1 in Theorem 4: $\sigma_n = an^q$ for $0 < q < 1$

Letting $\mathbf{n}_\lambda \equiv (n_\lambda^1, n_\lambda^2, ..., n_\lambda^k)$, we can write:

$$\begin{aligned}
\bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) &= \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta)\frac{\mu}{\theta} E\left[\left(\frac{\lambda_i}{\mu} - N(m_\lambda, n_\lambda^i) \right)^+ \right] \right), \\
&= \sum_{i=1}^k T_i \bar{\Pi}_\lambda^i(m_\lambda, n_\lambda^i), \\
&= \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta)\frac{\mu}{\theta} 1\{n_\lambda^i > 0\} \sigma_{n_\lambda^i} \int_{-1}^{\frac{\lambda_i/\mu - m_\lambda - n_\lambda^i}{\sigma_{n_\lambda^i}}} F_\epsilon(u)du \right).
\end{aligned}$$

When plugging the fluid solution $(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda)$, and recalling that $\beta \equiv (h/\theta + r)\mu$, we have

$$\bar{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) = \tilde{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) + \beta \sum_{i=k_0+1}^{k} \sigma_{\tilde{n}^i_\lambda} \int_{-1}^{0} F_\epsilon(u)du, \tag{29}$$

where k_0 is defined in Lemma 1. The expression in (29) suggests that the optimal policy for the stochastic-fluid problem will have the same number of fixed servers as its counterpart fluid solution, but will add some flexible resource to its counterpart fluid solution for an additional "hedge" against variability; this hedge should be on the order of σ_λ. We will prove that this is indeed the case in what follows, breaking down our proof into 3 steps.

Step 1. In an optimal solution $(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda)$ to the stochastic-fluid problem, for all $1 \le i \le k$:

(a) If $\bar{m}_\lambda < \lambda_i/\mu$, then $\lambda_i/\mu - \sigma_{\bar{n}^i_\lambda} \le \bar{m}_\lambda + \bar{n}^i_\lambda \le \lambda_i/\mu + \sigma_{\bar{n}^i_\lambda}$;

(b) If $\bar{m}_\lambda \ge \lambda_i/\mu$, then $\bar{n}^i_\lambda = 0$.

The second part of the statement is straightforward. We prove the first part of the statement by contradiction.

i) Aiming at a contradiction, suppose that, for period i, we have $\bar{m}_\lambda + \bar{n}^i_\lambda > \lambda_i/\mu + \sigma_{\bar{n}^i_\lambda}$. Then,

$$\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) = \tilde{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) > \tilde{\Pi}^i_\lambda(\bar{m}_\lambda, 0) = \bar{\Pi}^i_\lambda(\bar{m}_\lambda, 0).$$

The first equality holds because $\epsilon \in (-1, 1)$. In this case, by sending $\bar{n}^i_\lambda$ to zero, we reduce the cost of period i without changing the cost of any other period. We thus get a contradiction.

ii) Now suppose that, for period i, $\bar{m}_\lambda + \bar{n}^i_\lambda < \lambda_i/\mu - \sigma_{\bar{n}^i_\lambda}$. Choose n' such that $\bar{m}_\lambda + n' = \lambda_i/\mu - \sigma_{n'}$. Notice that $n' > \bar{n}^i_\lambda$ as σ_n is increasing in n. Then,

$$\begin{aligned}\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) &= c_{fix}\bar{m}_\lambda + c_{flex}\bar{n}^i_\lambda + \beta\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - \bar{n}^i_\lambda\right) \\ &> c_{fix}\bar{m}_\lambda + c_{flex}n' + \beta\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - n'\right) = \bar{\Pi}^i_\lambda(\bar{m}_\lambda, n').\end{aligned}$$

The inequality follows from the fact that $c_{flex} < \beta$ and $n' > \bar{n}^i_\lambda$. In this case, by increasing $\bar{n}^i_\lambda$ to n', we reduce the cost in period i without changing the cost of any other period. We thus get a contradiction. This concludes the proof of Step 1.

Step 2. For k_0 as defined in Lemma 1, i.e., such that $\tilde{m}_{k_0} = \lambda_{k_0}/\mu$:

(a) $\bar{m}_\lambda \ge \lambda_{k_0}/\mu$;

(b) For all $i \le k_0$, we have that $\bar{n}^i_\lambda = 0$.

We, again, prove this statement by contradiction. We begin by establishing part (a).

i) Aiming at a contradiction, suppose that $\bar{m}_\lambda < \lambda_{k_0}/\mu$. Recall that $\tilde{m}_\lambda = \lambda_{k_0}/\mu$. For $i \le k_0$, let $n^{i\prime}_\lambda = 0$; and for $k_0 < i \le k$, let $n^{i\prime}_\lambda = \max\{\bar{n}^i_\lambda - (\tilde{m}_\lambda - \bar{m}_\lambda), 0\}$. For $i < k_0$, we have

$$\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) \ge c_{fix}\bar{m}_\lambda.$$

For $i = k_0$, let $x = \lambda_{k_0}/\mu - \bar{m}_\lambda - \bar{n}^i_\lambda = \tilde{m}_\lambda - \bar{m}_\lambda - \bar{n}^i_\lambda$. Then we have

$$\begin{aligned}
&\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) - \bar{\Pi}^i_\lambda(\tilde{m}_\lambda, 0) \\
&= -c_{fix}(\tilde{m}_\lambda - \bar{m}_\lambda) + c_{flex}\bar{n}^i_\lambda + \beta\sigma_{\bar{n}^i_\lambda} \int_{-1}^{\frac{\lambda_i/\mu - \bar{m}_\lambda - \bar{n}^i_\lambda}{\sigma_{\bar{n}^i_\lambda}}} F(u)du \\
&= (-c_{fix} + c_{flex})(\tilde{m}_\lambda - \bar{m}_\lambda) - c_{flex}x + \beta\sigma_{\bar{n}^i_\lambda} \int_{-1}^{\frac{x}{\sigma_{\bar{n}^i_\lambda}}} F(u)du \\
&\geq (-c_{fix} + c_{flex})(\tilde{m}_\lambda - \bar{m}_\lambda) - c_{flex}\sigma_{\bar{n}^i_\lambda} F^{-1}(c_{flex}/\beta) + \beta\sigma_{\bar{n}^i_\lambda} \int_{-1}^{F^{-1}(c_{flex}/\beta)} F(u)du \\
&= (-c_{fix} + c_{flex})(\tilde{m}_\lambda - \bar{m}_\lambda) - \beta\sigma_{\bar{n}^i_\lambda} \int_{-1}^{F^{-1}(c_{flex}/\beta)} uf(u)du \\
&> (-c_{fix} + c_{flex})(\tilde{m}_\lambda - \bar{m}_\lambda).
\end{aligned}$$

Thus,

$$\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) \geq c_{fix}\bar{m}_\lambda + c_{flex}(\tilde{m}_\lambda - \bar{m}_\lambda).$$

For $i > k_0$, from Step 1, we have $\bar{m}_\lambda + \bar{n}^i_\lambda > \lambda_i/\mu - \sigma_{\bar{n}^i_\lambda}$ and $\bar{n}^i_\lambda < \lambda_i/\mu + \sigma_{\bar{n}^i_\lambda}$. Then for λ large enough, we have: $\bar{n}^i_\lambda - (\tilde{m}_\lambda - \bar{m}_\lambda) > \lambda_i/\mu - \lambda_{k_0}/\mu - \sigma_{\bar{n}^i_\lambda} > 0$. Thus, $n^{i'}_\lambda = \bar{n}^i_\lambda - (\tilde{m}_\lambda - \bar{m}_\lambda)$ for λ large enough, and

$$\begin{aligned}
\bar{\Pi}^i_\lambda(\bar{m}_\lambda, \bar{n}^i_\lambda) &= c_{fix}\bar{m}_\lambda + c_{flex}(\tilde{m}_\lambda - \bar{m}_\lambda) + c_{flex}n^{i'}_\lambda + \beta\left(\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - \bar{n}^i_\lambda - \sigma_{\bar{n}^i_\lambda}\epsilon\right)^+\right]\right) \\
&> c_{fix}\bar{m}_\lambda + c_{flex}(\tilde{m}_\lambda - \bar{m}_\lambda) + c_{flex}n^{i'}_\lambda + \beta\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - \bar{n}^i_\lambda - \sigma_{n^{i'}_\lambda}\epsilon\right)^+\right] \\
&= c_{fix}\bar{m}_\lambda + c_{flex}(\tilde{m}_\lambda - \bar{m}_\lambda) + c_{flex}n^{i'}_\lambda + \beta\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - n^{i'}_\lambda - \sigma_{n^{i'}_\lambda}\epsilon\right)^+\right]
\end{aligned}$$

where the last inequality follows from the fact that for fixed $s > 0$, $\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - s - \sigma_n\epsilon\right)^+\right]$ is increasing in n, and the last equality follows from the fact that $\bar{m}_\lambda + \bar{n}^i_\lambda = \tilde{m}_\lambda + n^{i'}_\lambda$

From the proof of Lemma 1, we have $\sum_{i=1}^k T_i c_{fix} \leq \sum_{i=k_0}^k T_i c_{flex}$. Combining the bound for different values of i, we have

$$\begin{aligned}
&\bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) \\
&= \sum_{i=1}^k T_i \left(c_{fix}\bar{m}_\lambda + c_{flex}\bar{n}^i_\lambda + \beta \cdot \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - \bar{n}^i_\lambda - \sigma_{\bar{n}^i_\lambda}\epsilon\right)^+\right]\right) \\
&> \sum_{i=1}^k T_i c_{fix}\tilde{m}_\lambda - \sum_{i=1}^k T_i c_{fix}(\tilde{m}_\lambda - \bar{m}_\lambda) + \sum_{i=k_0}^k T_i c_{flex}(\tilde{m}_\lambda - \bar{m}_\lambda) + \sum_{i=k_0+1}^k T_i c_{flex}n^{i'}_\lambda \\
&\quad + \beta \sum_{i=k_0+1}^k \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - n^{i'}_\lambda - \sigma_{n^{i'}_\lambda}\epsilon\right)^+\right] \\
&\geq \sum_{i=1}^k T_i c_{fix}\tilde{m}_\lambda + \sum_{i=k_0+1}^k T_i c_{flex}n^{i'}_\lambda + \beta \sum_{i=k_0+1}^k \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - n^{i'}_\lambda - \sigma_{n^{i'}}\epsilon\right)^+\right] \\
&= \bar{\Pi}_\lambda(\tilde{m}_\lambda, \mathbf{n}'_\lambda).
\end{aligned}$$

We therefore get a contradiction, which proves part (a). Next, we show part (b).

i) Suppose that $\bar{m}_\lambda \geq \lambda_{k_0}/\mu$ and $\bar{n}^i_\lambda > 0$ for some $i \leq k_0$. Then, it is easy to see that by decreasing such $\bar{n}^i_\lambda$

to zero, we reduce the value of the objective $\bar{\Pi}_\lambda$. We thus get a contradiction. This concludes the proof of Step 2.

Step 3. In an optimal solution to the stochastic-fluid problem, we must have that $\bar{m}_\lambda = \tilde{m}_\lambda$. To show this, we denote $\hat{m}_\lambda = \bar{m}_\lambda - \tilde{m}_\lambda$ and $\hat{\mathbf{n}}_\lambda = (n_\lambda^{k_0+1}, \ldots, n_\lambda^k)$. Based on Step 2, we can write

$$\begin{aligned}
&\min_{m_\lambda \geq 0, \mathbf{n}_\lambda \geq 0} \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \\
&= c_{fix}\tilde{m}_\lambda \sum_{i=1}^{k} T_i \\
&\quad + \min_{\hat{m}_\lambda \geq 0, \hat{\mathbf{n}}_\lambda \geq 0} \sum_{i=k_0+1}^{k} T_i \left(\left(c_{fix} \frac{\sum_{j=1}^{k} T_j}{\sum_{i=k_0+1}^{k} T_i} \right) \hat{m}_\lambda + c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \tilde{m}_\lambda - N(\hat{m}_\lambda, n_\lambda^i))^+ \right] \right)
\end{aligned}$$

As $c_{fix}\sum_{j=1}^{k} T_j > c_{flex}\sum_{i=k_0+1}^{k} T_i$, we must have at optimum that $\hat{m}_\lambda = 0$. To see why, we notice that:

$$\begin{aligned}
&\min_{\hat{m}_\lambda \geq 0, \hat{\mathbf{n}}_\lambda \geq 0} \sum_{i=k_0+1}^{k} T_i \left(\left(c_{fix} \frac{\sum_{j=1}^{k} T_j}{\sum_{i=k_0+1}^{k} T_i} \right) \hat{m}_\lambda + c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \tilde{m}_\lambda - N(\hat{m}_\lambda, n_\lambda^i))^+ \right] \right) \\
&\geq \sum_{i=k_0+1}^{k} T_i \min_{m_\lambda^i \geq 0, n_\lambda^i \geq 0} \left\{ \tilde{c}_{fix} m_\lambda^i + c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \tilde{m}_\lambda - N(m_\lambda^i, n_\lambda^i))^+ \right] \right\} \\
&= \sum_{i=k_0+1}^{k} T_i \min_{n_\lambda^i \geq 0} \left\{ c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \tilde{m}_\lambda - n_\lambda^i - \sigma_{n_\lambda^i}\epsilon)^+ \right] \right\}
\end{aligned}$$

where $\tilde{c}_{fix} = c_{fix}\sum_{j=1}^{k} T_j / \sum_{i=k_0+1}^{k} T_i > c_{flex}$, and the last equality follows from the analysis in Appendix C for the single-period case. Most importantly, the above analysis suggests that we can fully decompose the stochastic-fluid optimization problem into k single period problems with the number of fixed servers equal to $\tilde{m}_\lambda$. In particular,

$$\begin{aligned}
\min_{m_\lambda \geq 0, \mathbf{n}_\lambda \geq 0} \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) = c_{fix}\tilde{m}_\lambda \sum_{i=1}^{k} T_i \\
+ \sum_{i=k_0+1}^{k} T_i \min_{n_\lambda^i \geq 0} \left\{ c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \tilde{m}_\lambda - n_\lambda^i - \sigma_{n_\lambda^i}\epsilon)^+ \right] \right\}.
\end{aligned}$$

This includes the special case where $k_0 = 0$ and $\tilde{m}_\lambda = 0$.

D.2. Case 2 in Theorem 4: $\sigma_n = an$ for $a < 1$:

Let

$$g(\gamma) = c_{flex} + \beta a \int_{-1}^{(1-\gamma)/(a\gamma)} F_\epsilon(u)du - \frac{\beta}{\gamma} F_\epsilon\left(\frac{1-\gamma}{a\gamma} \right).$$

Recall that γ^* is the solution of $g(\gamma) = 0$. For a fixed value of m, we can solve for the corresponding optimal $\bar{n}(m)$ by optimizing each period individually. Particularly, for period i, we choose $n_i(m)$ that maximizes

$$\bar{\Pi}_\lambda^i(m, n_i(m)) = c_{fix}m + c_{flex}n_i(m) + \beta \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - N(m, n_i(m)) \right)^+ \right].$$

From our analysis in the single period case in Appendix C, we have $\bar{n}_i(m) = \gamma^*(\lambda_i/\mu - m)^+$, for γ^* in (16). Then if $\lambda_i/\mu > m$,

$$\begin{aligned}\bar{\Pi}^i_\lambda(m, \bar{n}_i(m)) &= c_{fix}m + c_{flex}\gamma^*(\lambda_i/\mu - m) + a\gamma^*(\lambda_i/\mu - m)\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du \\ &= \left(c_{fix} - c_{flex}\gamma^* - a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right) m \\ &\quad + \frac{\lambda_i}{\mu}\left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right).\end{aligned}$$

Now define $\kappa(m)$ as the first period we start blending when the fixed pool is set at m. Particularly, if $\lambda_k/\mu \le m$, we write $\kappa(m) \equiv k+1$ (recall that we assumed, without loss of generality, that the periods are ordered in increasing λ_i values). Otherwise, set $\kappa(m) \equiv \min\{i \ge 1 : \lambda_i/\mu > m\}$. Define $\sum_{i=k+1}^k z_i \equiv 0$. Then our goal is to optimize

$$\begin{aligned}\min_m \sum_{i=1}^k T_i \bar{\Pi}^i_\lambda(m, \bar{n}_i(m)) &= \left(\sum_{i=1}^k T_i c_{fix} - \sum_{i=\kappa(m)}^k T_i \left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right)\right) \cdot m \\ &\quad + \sum_{i=\kappa(m)}^k T_i \frac{\lambda_i}{\mu}\left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right). \qquad (30)\end{aligned}$$

Following the same line of argument as the proof of Lemma 1, one can show that the solution of (30) is $m = \lambda_{k_1}/\mu$, where $k_1 = 0$, if $c_{fix} > c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du$; and

$$k_1 = \max\left\{1 \le h \le k : \sum_{i=1}^k T_i c_{fix} \le \sum_{i=h}^k T_i \left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right)\right\},$$

otherwise. Particularly, if $k_1 = 0$, the we use flexible servers only. If $k_1 = k$, then we use fixed servers only. If $1 \le k_1 < k$, then $k_1 + 1$ is the first period where we start blending.

We next take a closer look at the event that determines k_1.

$$\begin{aligned}&\sum_{i=1}^k T_i c_{fix} \le \sum_{i=h}^k T_i \left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du\right) \\ \iff &\frac{\sum_{i=1}^k T_i}{\sum_{i=h}^k T_i} c_{fix} \le c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u)du \\ \iff &\frac{\sum_{i=1}^k T_i}{\sum_{i=h}^k T_i} c_{fix} \le \beta F_\epsilon\left(\frac{1-\gamma^*}{a\gamma^*}\right) \quad \text{as } g(\gamma^*) = 0. \\ \iff &\gamma^* \le \frac{1}{1 + aF_\epsilon^{-1}\left(c^h_{fix}/\beta\right)} \qquad (31)\end{aligned}$$

where $\iff$ means equivalent to and $c^h_{fix} = c_{fix}\frac{\sum_{i=1}^k T_i}{\sum_{i=h}^k T_i}$.

As $g(\gamma)$ is an increasing function of γ and γ^* is the solution of $g(\gamma) = 0$, we can check whether

$$g\left(1/\left(1 + aF_\epsilon^{-1}\left(c^h_{fix}/\beta\right)\right)\right) \ge 0 \qquad (32)$$

to check if the inequality (31) holds. As

$$g\left(\frac{1}{1 + aF_\epsilon^{-1}\left(c^h_{fix}/\beta\right)}\right) = c_{flex} + \beta a \int_{-1}^{F_\epsilon^{-1}\left(c^h_{fix}/\beta\right)} F_\epsilon(u)du - c^h_{fix}\left(1 + aF_\epsilon^{-1}\left(c^h_{fix}/\beta\right)\right),$$

then the inequality (32) is equivalent to

$$c_{flex} \ge c^h_{fix} + c^h_{fix} aF_\epsilon^{-1}\left(c^h_{fix}/\beta\right) - \beta a \int_{-1}^{F_\epsilon^{-1}\left(c^h_{fix}/\beta\right)} F_\epsilon(u)du.$$

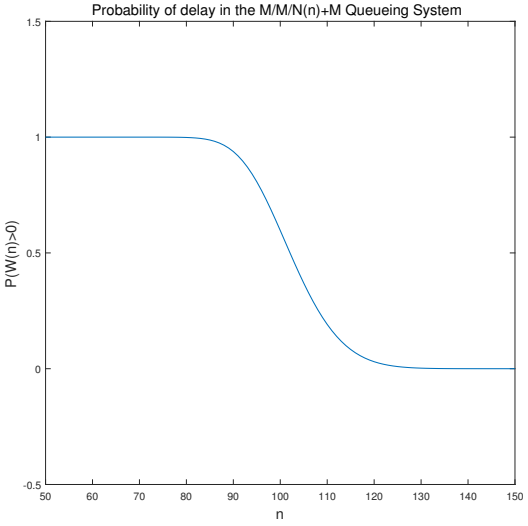

Figure 21 **Probability of delay $P(W(n) > 0)$ in the $M/M/N(n) + M$ queue as a function of n.**

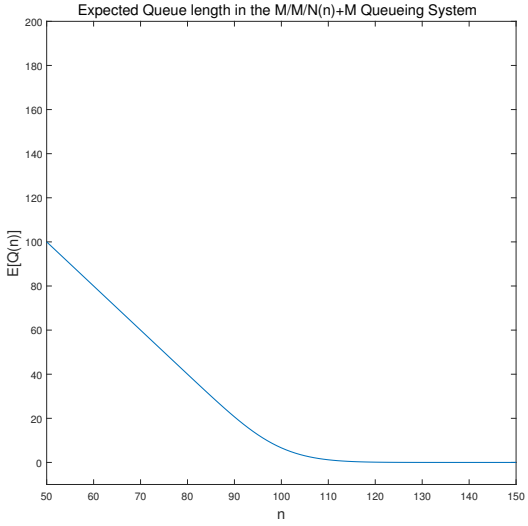

Figure 22 **Probability of delay $\mathbb{E}[Q(n)]$ in the $M/M/N(n) + M$ queue as a function of n.**

Appendix E: Supporting Figures: Convexity of Measures

In Figures 21 and 22, we consider an $M/M/N(n) + M$ queueing system with $\lambda = 100$ and $\theta = \mu = 1$. Figure 21 illustrate that the probability of delay, $P(W(n) > 0)$, is concave for $n \leq \lambda/\mu$, and convex otherwise. On the other hand, Figure 22 illustrates that $\mathbb{E}[Q(n)]$ is convex in n.

References

Ata, B., D. Lee, E. Sonmez. 2018. Dynamic staffing of volunteer gleaning operations. University of Chicago, working paper.

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

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., 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.

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

Brown, Lawrence, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, Linda Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* **100**(469) 36–50.

Cachon, Gerard P, Kaitlin M Daniels, Ruben Lobel. 2017. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* .

Chen, M Keith, Judith A Chevalier, Peter E Rossi, Emily Oehlsen. 2017. The value of flexible work: Evidence from uber drivers. Tech. rep., National Bureau of Economic Research.

Daskin, Mark S. 2011. *Service science*. John Wiley & Sons.

Drezner, Zvi, Nicholas Farnum. 1993. A generalized binomial distribution. *Communications in Statistics-Theory and Methods* **22**(11) 3051–3063.

Forbes. 2015. 3 secrets to leading a multi-everything blended workforce. `http://www.forbes.com/sites/meghanbiro/2015/11/07/3-secrets-to-leading-a-multi-everything-blended-workforce/#7cfdd938311a`. Accessed: 2017-02-08.

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

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. 2018. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Northwestern University, working paper.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.

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.

Heyde, CC. 2004. Asymptotics and criticality for a correlated bernoulli process. *Australian & New Zealand Journal of Statistics* **46**(1) 53–57.

Hu, Ming, Yun Zhou. 2018. Price, wage and fixed commission in on-demand matching. University of Toronto, working paper.

Ibrahim, R. 2018. Staffing a service system with a random service capacity and impatient customers. *Production and Operations Management* .

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.

Mandelbaum, A., A.S. Zeltyn. 2007. Service engineering in action: the Palm/Erlang-A queue, with appications to call centers. *Advances in Services Innovations*. Springer-Verlag.

Ozkan, E., A. Ward. 2018. Dynamic matching for real-time ridesharing. University of Southern California, working paper.

PWC. 2017. The sharing economy grows up. `https://www.pwc.co.uk/issues/megatrends/collisions/sharingeconomy/outlook-for-the-sharing-economy-in-the-uk-2016.html`. Accessed: 2017-12-4.

Romano, Joseph P, Michael Wolf. 2000. A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & probability letters* **47**(2) 115–124.

Taylor, T. 2018. On-demand service platforms. University of California Berkeley, working paper.

Wang, W., D. Gupta. 2014. Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing and Service Operations Management* **16**(3) 439–454.

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.