Flexible Workers or Full-Time Employees?
On Staffing Systems with a Blended Workforce

Jing Dong
Northwestern University, 2145 Sheridan Road, M239, Evanston, IL 60208 jing.dong@northwestern.edu

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

The rise of the blended workforce, which is identified as one of the top workplace trends in 2017, is prompting firms to re-evaluate their staffing strategies. A blended workforce melds, as a deliberate business strategy, contingent workers, e.g., independent contractors or freelancers, with permanent employees. In this paper, we study optimal staffing decisions in service systems with a blended workforce, in the context of a queueing-theoretic framework. Because part of the workforce is flexible, the number of servers in our queueing model is random. Since the staffing problem with a random number of servers is analytically intractable, we formulate two problem relaxations and demonstrate their accuracies in large systems by relying on an asymptotic, many-server, mode of analysis. Based on those relaxations, we make staffing recommendations for systems with a blended workforce. We demonstrate that staffing decisions in such systems are not straightforward. Indeed, we show these decisions depend on three main factors: (i) the supply-side uncertainty of the flexible agent pool, (ii) operating costs in the system, and (iii) fluctuations in incoming customer demand, and it may or may not be cost-effective to staff a blended workforce depending on the interplay between those factors.

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

1. Introduction

Major multinational companies (McKinsey&Company 2015, Deloitte 2016, Accenture 2016) and leading business periodicals (The Economist 2013, Harvard Business Review 2016, Forbes 2016) identify the rise of the blended workforce as one of the top global workplace trends which are reshaping the modern business landscape. A blended workforce melds, as a deliberate business strategy, a layer of contingent workers, e.g., independent contractors or freelancers, with a core of permanent, full-time, employees. In order to effectively manage a blended workforce, companies must begin by “reevaluating their staffing models” (Forbes 2015). Indeed, a major problem facing those companies is how to decide on the “right number of right people at the right time”\(^1\), by appropriately weighing the pertinent tradeoffs. This is the problem that we address in this paper.

\(^1\) https://www.beeline.com/blog/10-tips-for-the-right-blend-of-flexible-workers-in-your-organization/
The rise of the blended workforce. According to an Intuit (2016) report, the number of contingent workers will constitute more than 40% of the American workforce by 2020. While the proportions of such workers are on the rise in all sectors of the economy, this is especially the case for the service sector (UKCES 2016), which we focus on in this paper. Indeed, multiple Fortune 500 companies in the service sector currently rely on a blended workforce model, e.g., Walmart (walmart.com), Time Warner (timewarnercable.com), and Netflix (netflix.com), to name a few. Similarly, several small to medium-size companies rely on such a model as well, e.g., R & M (renmmatrix.nl) is a market research bureau which has 40 full-time employees in addition to 300 freelance call-center agents who enjoy a great degree of flexibility in setting their own schedules (Feinberg et al. 2005).

Managerial challenges with a blended workforce. Several tradeoffs need to be weighted when managing a blended workforce. On one hand, fixed, full-time, workers are typically reliable and committed to the firm, and they usually have a number of required working hours. In other words, they are relatively easy to control. However, this control comes at the expense of a fixed and relatively steep labor cost. Moreover, a pool of regular full-time workers cannot be easily scaled to meet dynamic business needs, i.e., periods of high or low customer demand.

On the other hand, flexible, contingent, workers are recruited on a part-time basis. To be concrete, we are thinking here of contingent workers who are recruited to complete relatively low-skill tasks, such as those advertised on Wonolo (wonolo.com) or Zaarly (zaarly.com) which offer business-to-business services by matching qualified workers to client companies. Flexible workers have a legal right to various degrees of flexibility, e.g., in setting their own work schedules, at the expense of being deprived of benefits which are usually granted to regular employees, e.g., unemployment insurance, overtime compensation, etc. Contingent workers are also typically less reliable than regular employees: They tend to have high turnover rates and uncertain availabilities, mostly due to the flexibility that is inherent in their work contracts. However, a pool of flexible workers can be easily scaled to meet seasonal demand fluctuations.

A service-system manager who relies on a blended workforce must therefore decide, as a long-term business strategy in an initial planning stage, on the numbers of flexible and fixed agents to staff so as to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers.

Modelling framework and overview. In this paper, we study the problem of staffing a service system with a blended workforce in the context of a stylized queueing model. We assume that there are \( k \) working periods, customer demand rates are time-varying, and the agent pool comprises both fixed and flexible agents. (Hereafter, we use “agent” and “server” interchangeably.) A fixed

---

2 https://www.xero.com/uk/small-business-guides/business-management/independent-contractor-or-employee
server is available in each period, and is compensated $c_0$ per unit time. To capture the supply-side uncertainty associated with flexible agents, we assume that a flexible server may or may not be available for work in any given period. If a flexible server is available, she earns $c_1$ per unit time. That is, $c_0$ and $c_1$ are the staffing costs in the system. Because part of the agent pool is flexible, the total number of available servers in our queueing model is random.

In an initial planning stage, the system manager must decide on the number of fixed servers, $m$, and the expected number of flexible servers, $n$. Indeed, since flexible servers may not be available, the manager cannot enforce a realized number of flexible servers, and must plan on an expected number instead (the distribution of the random number of flexible agents who actually show up, which depends on that expected value, is assumed to be known to the manager). In selecting the respective pool sizes of fixed and flexible servers, the decision maker is effectively controlling the supply-side uncertainty in the system, i.e., the distribution of the number of available servers. As in Gurvich et al. (2017), we also allow for the system manager to impose a cap on the expected number of flexible agents that she allows in a period. That is, we allow the system manager to impose on a flexible server not to show up in a given period, even when the server is willing to do so. The imposition of a cap is an additional control lever available to the system manager, which allows her to dynamically scale the pool size of flexible agents to meet fluctuations (time variations) in customer demand. Customers are assumed to be both impatient and delay sensitive, as is usually the case in service systems.

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, many-server, mode of analysis. In particular, our modelling approach is close to the one in Bassamboo et al. (2010), who consider a single-period capacity-sizing problem with random arrival rates instead.

Here is a brief summary of our modelling approach. 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 formulation regimes. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. In this region, we derive the optimal staffing levels by solving a stochastic-fluid optimization problem which ignores stochastic variability; this problem is of a multi-period newsvendor type in our setting. 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 objective is to understand the properties of the alternative solutions which
are obtained under those two approximations, to rigorously justify their accuracy by quantifying their corresponding errors (asymptotically in large systems) and, most importantly, to characterize the relevant tradeoffs so as to draw insights into the effective staffing of systems with a blended workforce.

A key technical challenge in our analysis is that the ensuing randomness in the number of servers, i.e., the uncertain parameter in our queueing system, depends itself on the staffing levels, i.e., our decision variables. For example, the resulting multi-period newsvendor problem in the stochastic-fluid relaxation of our problem can be equivalently formulated as one where demand depends on the stocking quantity and, to the best of our knowledge, there do not exist simple closed-form solutions to this type of problems. More generally, this also implies that the asymptotic accuracy of our respective relaxations may depend on the specific solutions to those problems. To circumvent this difficulty, we derive approximate, asymptotically optimal (in a sense to be made more precise later), solutions to the stochastic-fluid problems as well, and quantify their corresponding accuracies.

Specific contributions of our paper. In this paper, we make the following contributions (our theoretical results are for exponentially-distributed patience and service times):

- We quantify the asymptotic order of magnitude of errors for the fluid and stochastic-fluid approximations in a multi-period setting, both with and without time-varying demand.
- We derive exact (in the form of an implicit relation) and approximate solutions of the stochastic-fluid formulation, which is difficult to solve in our setting, and quantify the accuracy of those approximate solutions.
- We show that with a random number of servers, as with random arrival rates (Bassamboo et al. 2010), one can distinguish between an uncertainty-dominated regime and a variability-dominated regime, depending on the magnitude of uncertainty in the number of servers. In particular, the stochastic-fluid optimal solution involves a base capacity and an uncertainty hedge which is “extremely” accurate in the uncertainty-dominated regime, i.e., the resulting staffing level is indistinguishable from the optimal solution to our original problem. However, there is no concrete benefit from that uncertainty hedge over the regular square-root staffing hedge (Halflin and Whitt 1981, Garnett et al. 2002) in the variability-dominated regime.
- We supplement our analytical results with detailed numerical studies throughout. As a robustness check, we consider a general patience distribution. There, we find that investigating the asymptotic accuracy of the respective relaxations, i.e., fluid and stochastic-fluid, boils down to determining the asymptotically optimal regime which is prescribed at fluid scale (depending on properties of the failure-rate function of the patience distribution). For example, if the

---

3 This challenge does not arise when considering random arrival rates as in Bassamboo et al. (2010): There, the stochastic-fluid optimal solution has a remarkably simple critical fractile form.
overloaded regime is prescribed at fluid scale, then there is generally no advantage in solving
the stochastic-fluid formulation over the fluid formulation, except possibly when the underly-
ing uncertainty in the number of servers is very large (coefficient of variation bounded away
from 0 in large systems). In other words, fluid approximations remain generally accurate in
the overloaded regime, as was observed with a deterministic number of servers in Bassamboo
and Randhawa (2010).

- We make several recommendations on staffing systems with a blended workforce by charac-
terising the tradeoffs between three factors: (i) operational costs, (ii) the time variation in
demand patterns, and (iii) the uncertainty ensuing from staffing flexible servers. For time-
varying demand, our analysis is based on a two-period setting with high or low demand.

  — When customer demand rates do not vary, we find that a manager should not use a
  blended workforce; instead, she should, in general, rely solely on the cheaper alternative,
  fixed or flexible. This remains true so long as the variability in the flexible resource is not
  too high. In the highly-variable case, i.e., where the coefficient of variation is bounded
  away from 0 in large systems, the manager could staff the more expensive fixed resource if
  the flexible resource is not much cheaper, i.e., there is a price to be paid for that variability.

  — When customer demand rates do vary, it may be cost effective for the manager to use
  a blended workforce. However, this is only the case if the fixed resource is cheaper
  and the disparity in pay between the fixed and flexible resources is not too great. If the fixed
  resource is more expensive, and the flexible resource is not too variable, then the manager
  should rely solely on the flexible resource.

  — Variation in demand patterns makes using the flexible resource more attractive
  over the fixed resource, because it can be easily scaled to dynamically meet variations in customer
  demand: A manager would always use a flexible resource (alone or through blending) so
  long as it is not “much more” expensive than the fixed resource.

  — When demand varies over the horizon, if it is optimal for a manager to use strictly one
  of the two resources, then she does so to match supply and demand in the high-demand
  period. If only the fixed resource is used, then the low-demand period will be overstaffed;
  if only the flexible resource is used, then a cap is used in the low-demand period. In
  contrast, if it is optimal to blend, a manager staffs the fixed resource supply to match
  the low-demand period, and uses the flexible resource to match supply up to demand in
  the high-demand period. This lends support to current business practices, e.g., whereby
  temporary flexible workers are called upon when there are unforeseen surges in demand.
The rest of this paper is organized as follows. In §2, we review the relevant literature. In §3, to build intuition, we formulate initial insights into the impact of randomness in capacity by considering a special case. In §4, we formulate our capacity-sizing problem, as well as its stochastic-fluid and fluid relaxations; we also quantify the corresponding optimality gaps. In §5, we derive the exact and asymptotically optimal solutions of the stochastic-fluid relaxation with stationary demand, and draw insights into staffing systems with a blended workforce. In §6, we extend our results to the two-period case with time-varying demand rates. In §7, we consider general abandonment, and in §8, we draw conclusions. We relegate all proofs to the e-companion.

2. Related Literature

Our modelling approach is close to the stream of literature initiated by Harrison and Zeevi (2005) and used in Bassamboo et al. (2010) who address the question of capacity planning under parameter uncertainty. Our paper is related to the extensive literature analyzing asymptotics of many-server queueing systems with impatient customers (Garnett et al. 2002, Zeltyn and Mandelbaum 2005, Whitt 2004, 2006a, Bassamboo and Randhawa 2010, Bassamboo et al. 2010), and to the large literature on optimal staffing decisions in service systems (Maglaras and Zeevi 2003, Borst et al. 2004, Harrison and Zeevi 2005, Bassamboo et al. 2005, 2010); for other references, see Gans et al. (2003) and Akşin et al. (2007). However, none of those papers considers a random number of servers. Whitt (2006b) considers many-server queues with an uncertain arrival rate, an uncertain number of servers, and a single period. Our focus here is on staffing multiple periods instead, and we go beyond the fluid approximation in that paper. 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 there.

There is a body of research within the queueing games literature which considers strategic servers that may select their service rates (Cachon and Harker 2002, Cachon and Zhang 2007). However, such papers do not consider staffing decisions, and the maximum number of servers considered is two. Recent exceptions are Gopalakrishnan et al. (2016) and Zhan and Ward (2017). Our work is 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.

This paper is most closely related to recent papers on queues with a self-scheduling capacity. Gurvich et al. (2017) 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 (2017) 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. (2017) and Ibrahim (2017) rely solely on fluid approximations to the system, and do not consider a blended workforce.

Taylor (2017) examines how two defining features of an on-demand service platform, delay sensitivity and agent independence, impact the platform’s optimal per-service price and wage. Ozkan and Ward (2017) study optimal matching decisions in a ride-sharing platform and demonstrate the need to go beyond the prevailing closed-driver matching policy. Braverman et al. (2017) model a ride-sharing system as a closed queueing network and rely on a fluid model to derive an optimal routing policy. Cachon et al. (2017) study optimal contracts in a platform with self-scheduling capacity; the agent pool size is large and assumed to be given a priori, i.e., the staffing question is not addressed. Riquelme et al. (2017) model a ride-sharing service using queueing theory and determine optimal platform pricing. They find that threshold-based dynamic pricing does not outperform a static pricing policy, but that dynamic pricing is more robust to fluctuations in system parameters.

3. A Random Number of Servers

In this section, we specify our queueing framework. To build intuition, we also formulate initial insights into the impact of randomness in capacity: In a special case, we demonstrate (Lemma 1) that randomness in capacity leads to a deterioration in the system’s performance. We then describe simulation results which illustrate that similar insights hold more generally as well.

3.1. Queueing Model

We consider a single-class $M/M/N + M$ queueing model $^4$. Customers arrive to the system according to a Poisson process with rate $\lambda$, and service times are independent and identically distributed (i.i.d.) exponential random variables with rate $\mu$. Customers are impatient, and their patience times are i.i.d. exponentially distributed with rate $\theta$. Customers are processed in the order in which they arrive, i.e., we use the first-come-first-served discipline.

The number of servers, $N$, is a nonnegative integer 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). Thus, a proper steady-state exists, and we focus on steady-state performances throughout.

$^4$ For now, we do not distinguish between fixed and flexible capacity; we will do so later when investigating optimal staffing decisions in the system.
3.2. Impact of Randomness in Capacity

Let $Q(N)$ denote the steady-state number of customers in a system with $N$ servers. For simplicity, we assume here that the abandonment rate, $\theta$, is equal to the service rate, $\mu$. Let $\sigma_1 > 0$, and $\sigma_2 > 0$, and consider two queueing systems with $N_1$ and $N_2$ servers, respectively, given by:

$$N_1 = n + \sigma_1 \epsilon \quad \text{and} \quad N_2 = n + \sigma_2 \epsilon,$$

for a proper random variable $\epsilon \geq \max\{-n/\sigma_1, -n/\sigma_2\}$ with $E[\epsilon] = 0$. That is, $E[N_1] = E[N_2] = n$, $\text{Var}[N_1] = \sigma_1^2 \text{Var}[\epsilon]$, and $\text{Var}[N_2] = \sigma_2^2 \text{Var}[\epsilon]$. All remaining parameters are assumed to be identical across the two systems. Then, the following lemma holds.

**Lemma 1.** If $\sigma_1 \leq \sigma_2$, then $E[Q(N_1)] \leq E[Q(N_2)]$.

Lemma 1 shows that, with all else held constant (including the expected number of servers), increased variability in the number of servers leads to worse system performance. In other words, a decision maker who ignores that variability, and approximates her available capacity by its expected value instead, would have an overoptimistic view of performance in her system.

3.3. Numerical Study

We now describe supporting results from a short simulation study (Table 1). Our objective here is two-fold: (i) to quantify the deterioration in system performance which results from an increased variability in capacity, and (ii) to show that the preliminary results of Lemma 1 hold under more general distributional assumptions as well.

We let the number of servers $N_\lambda$, in a system with arrival rate $\lambda$, have a truncated normal distribution; specifically, we let $N_\lambda = \lceil Z_\lambda^{\lambda} \rceil$ where $Z_\lambda \sim \text{Nor}(\lambda/\mu, \sigma_\lambda^2)$. We note that $E[Z_\lambda] = \lambda/\mu$ represents a “base capacity” to match mean demand. We consider alternative values of the arrival rate, $\lambda = 50, 100, 500, \text{and} 1000$, and hold the service rate $\mu = 1$ fixed. We also consider alternative functional forms for the variance, $\sigma_\lambda = \sqrt{\lambda}, \lambda^{3/4}$, and $0.25 \lambda$. As a benchmark, we consider $\sigma_\lambda = 0$, which represents the classical case where the number of servers is deterministic.

For patience times, we consider both the exponential ($\theta = 2$) and $H_2$ (hyperexponential with squared coefficient of variation equal to 4, balanced means, and mean equal to 2) distributions. For service times, we consider both the exponential ($\mu = 1$) and $LN$ (lognormal with mean and variance both equal to 1) distributions. We consider $H_2$ for abandonment and $LN$ for service times because there is empirical evidence suggesting good fits to those distributions in practice (Roubos and Jouini 2013, Brown et al. 2005). Our simulations estimates are based on 100 independent replications of length one million arrivals each, and we discard an initial transient period of length 20,000 arrivals from each replication.
\[ \text{Table 1} \quad \text{Expected value and variance of the steady-state queue length in the } M/GI/N + M \text{ queueing system with } N = Z^+ \text{ where } Z \sim \text{Nor}(\lambda/\mu, \sigma^2). \]

Inspecting the rows of Table 1 shows that as \( \sigma_\lambda \) increases for a fixed \( \lambda \), both \( \mathbb{E}[Q(N_\lambda)] \) and \( \text{Var}[Q(N_\lambda)] \) increase as well. This substantiates the preliminary results of Lemma 1. Inspecting the columns of Table 1 (for each model) allows for a more detailed understanding of the system’s performance. As expected, when the number of servers is constant, we observe that as \( \lambda \) increases by a factor \( l > 0 \), \( \mathbb{E}[Q(N_\lambda)] \) increases by a factor of \( \sqrt{l} \) and \( \text{Var}[Q(N_\lambda)] \) increases by that same factor \( l \) (Bassamboo et al. 2010). However, for \( \sigma_\lambda > 0 \), as \( \lambda \) increases by a factor \( l > 0 \), \( \mathbb{E}[Q(N_\lambda)] \) increases by a factor of \( \sigma_l \) instead, and \( \text{Var}[Q(N_\lambda)] \) continues to increase by that same factor \( l \). In other words, Table 1 suggests that, as \( \lambda \) increases, \( \mathbb{E}[Q(N_\lambda)] \) is on the order of magnitude of \( \sigma_\lambda \) (when \( \sigma_\lambda \geq \sqrt{\lambda} \)), and \( \text{Var}[Q(N_\lambda)] \) is on the order of magnitude of the maximum between \( \lambda \) and \( \sigma_\lambda^2 \). In §4.3, we will establish such results formally by considering an asymptotic many-server queueing framework where we let \( \lambda \) increase without bound.

Table 1 illustrates that the order of magnitude of “noise” in a system with a random number of servers may be higher than the usual square-root order of stochastic fluctuations. Thus, when
investigating appropriate staffing decisions with a blended workforce, there is a need to derive an appropriate hedge against such variability, which may be of a different order than the standard square-root safety capacity (Garnett et al. 2002). We describe our staffing problem in the following section; the solution to this problem allows for the specification of an appropriate hedge.

4. Capacity Sizing with a Blended Workforce

In this section, we formulate the capacity-sizing problem faced by the system manager. 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 work days. The arrival rate (of the Poisson arrival process) in period \( i \) is given by \( \lambda_i \). We fix \( \lambda > 0 \) and let \( \lambda_i = \lambda \xi_i \), where \( \xi_i \geq 0 \) for each \( i \). In formulating the capacity-sizing problem, we index all relevant quantities by \( \lambda \), to indicate dependence on the arrival rates. In our asymptotic analysis, we let \( \lambda \) grow without bound while keeping \( \xi_i \) constant for each \( i \).

The system manager must determine the staffing levels for the fixed and flexible agent pools so as to strike a balance between customer-related and staffing costs in the system. Specifically, consistently with Bassamboo and Randhawa (2010) and Bassamboo et al. (2010), we consider two quality-of-service 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 \( c_0 \) for a fixed server, and \( c_1 \) for a flexible server. In practice, it is usually the case that \( c_1 < c_0 \), but we do not impose this assumption here. Instead, we study how the solution to our capacity-sizing problem depends on the individual staffing costs, \( c_0 \) and \( c_1 \).

We let \( m_\lambda \) denote the number of fixed servers, and \( n_\lambda \) denote the expected number of flexible servers. Since our choices of \( m_\lambda \) and \( n_\lambda \) affect the distribution of the random number of servers, \( N(m_\lambda, n_\lambda) \), we need to make further specifications. In particular, we assume that the total number of servers can be expressed as:

\[
N(m_\lambda, n_\lambda) = m_\lambda + n_\lambda + \sigma_n \epsilon,
\]

for some random variable \(-1 \leq \epsilon \leq 1\) with \( \mathbb{E}[\epsilon] = 0 \). We assume that \( \epsilon \) has a strictly positive probability density function (pdf), \( f_\epsilon \), on \((-1, 1)\). Thus, its cumulative distribution function (cdf), \( F_\epsilon \), is invertible on that domain. We assume that \( \sigma_n \geq 0 \) is some function of \( n_\lambda \). The expression in (1) implies that \( \mathbb{E}[N(m_\lambda, n_\lambda)] = m_\lambda + n_\lambda \) and \( \text{Var}[N(m_\lambda, n_\lambda)] = \sigma_n^2 \text{Var}[\epsilon] \). We also make the following assumptions.

**Assumption 1.** For \( \sigma_n \) in (1), we assume that \( \sigma_n' > 0 \) and \( \sigma_n'' < 0 \); also, \( c_1, c_0 < (h/\theta + r)\mu \).
Assumption 1 guarantees that \( \sigma_n \) is strictly increasing and concave (which will be used later to ensure uniqueness of solutions). The assumption on costs 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. In (1), we ignore the integrality assumptions on \( m_\lambda, n_\lambda \), 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 notice that the expected queue length expression for Erlang-A model can be extended to real values of \( N \) (Mandelbaum and Zeltyn 2007). The staffing problem faced by the manager is defined for both integer and non-integer values of \( m_\lambda \) and \( n_\lambda \).

As in Gurvich et al. (2017), we also allow for the system manager to impose a cap on the expected number of flexible agents in a period. That is, we allow the system manager to impose on a flexible server not to show up in a given period, even when the server is willing to do so. In particular, we let \( \alpha_i^\lambda \) denote the cap in period \( i \), and \( \alpha_i^\lambda n_\lambda \) be the expected number of flexible servers that will be allowed in period \( i \). We let \( Q^i(m_\lambda, \alpha_i^\lambda n_\lambda) \) and \( \Xi^i(m_\lambda, \alpha_i^\lambda n_\lambda) \) denote the steady-state queue-length and steady-state rate of customer abandonment in period \( i \). We let \( X^i(m_\lambda, \alpha_i^\lambda n_\lambda) \) denote the steady-state number of customers in the system, in period \( i \), so that:

\[
Q^i(m_\lambda, \alpha_i^\lambda n_\lambda) = (X^i(m_\lambda, \alpha_i^\lambda n_\lambda) - N(m_\lambda, \alpha_i^\lambda n_\lambda))^+, \]

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

\[
\Xi^i(m_\lambda, \alpha_i^\lambda n_\lambda) = \theta \cdot \mathbb{E}[Q^i(m_\lambda, \alpha_i^\lambda n_\lambda)],
\]

where \( \theta \) is the rate of the patience-time distribution (Mandelbaum and Zeltyn 2007). The system manager’s staffing problem is:

\[
\begin{align*}
\min_{m_\lambda, n_\lambda, \alpha_i^\lambda} \Pi_\lambda(m_\lambda, n_\lambda, \alpha_i^\lambda) \\
= \sum_{i=1}^{k} T_i \left( c_0 m_\lambda + c_1 \alpha_i^\lambda n_\lambda + h \cdot \mathbb{E}[Q^i(m_\lambda, \alpha_i^\lambda n_\lambda)] + r \cdot \Xi^i(m_\lambda, \alpha_i^\lambda n_\lambda) \right), \\
= \sum_{i=1}^{k} T_i \left( c_0 m_\lambda + c_1 \alpha_i^\lambda n_\lambda + (h + r \theta) \mathbb{E}[Q^i(m_\lambda, \alpha_i^\lambda n_\lambda)] \right), \\
= \sum_{i=1}^{k} T_i \left( c_0 m_\lambda + c_1 \alpha_i^\lambda n_\lambda + (h + r \theta) \mathbb{E}[(X^i(m_\lambda, \alpha_i^\lambda n_\lambda) - N(m_\lambda, \alpha_i^\lambda n_\lambda))^+] \right), \\
= \sum_{i=1}^{k} T_i \left( c_0 m_\lambda + c_1 \alpha_i^\lambda n_\lambda + (h + r \theta) \mathbb{E}[(X^i(m_\lambda, \alpha_i^\lambda n_\lambda) - m_\lambda - \alpha_i^\lambda n_\lambda - \sigma \alpha_i^\lambda n_\lambda \cdot e)^+] \right).
\end{align*}
\]

The problem formulation in (2) is prohibitively difficult to solve in closed form, because our choices of \( m_\lambda \) and \( n_\lambda \) affect the distribution of the number of servers, which, in turn, affects the distributions of both the queue-length, \( Q \), and the rate of abandonment, \( \Xi \). Thus, we next turn to formulating stochastic-fluid (ignoring stochastic variability) and fluid (ignoring both stochastic variability and parameter uncertainty) relaxations of that problem.
4.1. Stochastic-Fluid Problem

For the stochastic-fluid relaxation of our problem, we ignore stochastic fluctuations in the system. In particular, customers arrive to period \( i \) at the rate of \( \lambda_i \) per unit of time. The processing capacity is \( N(m_\lambda, \alpha^i_\lambda n_\lambda) \) and, by conservation of flow, the resulting stochastic-fluid abandonment rate is given by \( (\lambda_i - N(m_\lambda, \alpha^i n_\lambda) \cdot \mu)^+ \). Thus, the resulting stochastic-fluid approximation to (2) is:

\[
\min_{m_\lambda, n_\lambda, \alpha^i_\lambda} \bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha^i_\lambda) \equiv \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha^i_\lambda n_\lambda + \left(\frac{h}{\theta} + r\right) (\lambda_i - N(m_\lambda, \alpha^i n_\lambda) \cdot \mu)^+ \right), \tag{3}
\]

\[
= \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha^i_\lambda n_\lambda + \left(\frac{h}{\theta} + r\right) (\lambda_i - m_\lambda \mu - \alpha^i_\lambda n_\lambda \mu - \sigma^i_\lambda \mu \epsilon)^+ \right). \tag{4}
\]

The stochastic-fluid formulation in (3) may be viewed as a multi-period newsvendor-type problem, albeit one where the “demand” in period \( i \), \( \lambda_i - \sigma^i \lambda \mu \epsilon \), is dependent on the overprocessing capacity, \( m_\lambda \mu + \alpha^i_\lambda n_\lambda \mu \), where \( m_\lambda \) and \( n_\lambda \) are the decision variables. There have been relatively few papers in the literature which study newsvendor-type problems where the distribution of the demand depends on the stocking quantity and these, to the best of our knowledge, are restricted to a single-period setting only; e.g., see Balakrishnan et al. (2008). Because it is difficult to derive a simple, closed-form, analytical solution to (3), we resort to deriving an asymptotically optimal solution instead, in a sense that will be made precise in §5.

4.2. Fluid Problem

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 \( (\lambda_i - m_\lambda \mu - \alpha^i_\lambda n_\lambda \mu)^+ \), which is obtained by substituting the random number of servers, \( N(m_\lambda, \alpha^i_\lambda n_\lambda) \), by its expected value, \( m_\lambda + \alpha^i_\lambda n_\lambda \). This leads to the following:

\[
\min_{m_\lambda, n_\lambda, \alpha^i_\lambda} \bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha^i_\lambda) \equiv \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha^i_\lambda n_\lambda + \left(\frac{h}{\theta} + r\right) (\lambda_i - m_\lambda \mu - \alpha^i_\lambda n_\lambda \mu)^+ \right). \tag{4}
\]

4.3. Optimality Gaps

We now study the accuracy of the staffing prescriptions in (3) and (4), in a regime where the arrival rate, \( \lambda_i \), is large. For the ease of exposition, we focus here on a single-period setting, and relegate generalization to the multi-period case to the e-companion. Thus, we drop dependence on the period’s index, \( i \). With a single period, we can equivalently consider the capped expected number of flexible servers, \( \alpha_\lambda n_\lambda \), to be our decision variable, i.e., we can eliminate \( \alpha_\lambda \) from the problem.

We begin with a statement of our main theorem. Let \((m_\lambda^*, n_\lambda^*)\) denote the optimal staffing levels to the original staffing problem in (2). Let \((\bar{m}_\lambda, \bar{n}_\lambda)\) and \((\tilde{m}_\lambda, \tilde{n}_\lambda)\) denote the optimal solutions to its stochastic-fluid relaxation in (3) and its fluid relaxation in (4), respectively. To interpret the results of Theorem 1, we need the following definitions.
Definition 1. Let $f$ and $g$ be two functions defined on some subset of $\mathbb{R}$. Then, as $n \to \infty$,
(a) $f(n) = \mathcal{O}(g(n))$ if there exists $M > 0$ and $C > 0$ such that $|f(n)| \leq M|g(n)|$ for $n \geq C$;
(b) $f(n) = o(g(n))$ if for all $\xi > 0$, there exists $N$ such that $|f(n)| \leq \xi|g(n)|$ for all $n \geq N$;
(c) $f(n) = \Theta(g(n))$ if there exists $M > 0$, $L > 0$ and $C > 0$ such that $L|g(n)| \leq |f(n)| \leq M|g(n)|$ for $n \geq C$.

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

Theorem 1. For large $\lambda$,

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

and

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

Furthermore, if $\bar{m}_\lambda + \bar{n}_\lambda = \lambda/\mu + \mathcal{O}(\sigma_\lambda)$ and $\bar{n}_\lambda = \mathcal{O}(\lambda)$, then:

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

The results of Theorem 1 are in line with those in Theorem 1 of Bassamboo et al. (2010), and indeed our proof proceeds in a similar fashion. In particular, Theorem 1 characterizes two operating modes, depending on the magnitude of variability in the random number of servers, as quantified by $\sigma_\lambda$. When $\sigma_\lambda$ is “large”, i.e., of an order which is larger than the square-root order of stochastic fluctuations in the system, the system can be considered to be in an uncertainty-dominated regime. In this regime, 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 $\sigma_\lambda$ is “small”, i.e., of an order which is smaller than the square-root order of stochastic fluctuations in the system, the system can be considered to be in a variability-dominated regime. In this case, the optimality gap for the stochastic-fluid solution is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. Since the fluid approximation to our problem ignores both uncertainty and variability effects, the corresponding optimality gap is dominated by either the order of stochastic fluctuations in the system, or the order of variability in the number of servers, whichever is larger. In other words, when the variability in the number of servers is small, there is no distinct (not asymptotically negligible) advantage from using stochastic-fluid approximations over fluid approximations to the system. Since the results of Theorem 1 assume specific forms for the solutions to (3) and (4), there remains to show that these are indeed the correct forms for those solutions: We show this in §5.
4.4. Supporting Numerical Study
We close this section with a numerical investigation of the performance of our stochastic-fluid and fluid solutions. In particular, we show that the asymptotic results of Theorem 1 describe small to moderate systems well, which validates their usefulness even further. Since including fixed servers does not affect our asymptotic accuracy results, we restrict attention to having only a flexible pool. We assume the following values for the cost parameters: \( c = 1/3, p = 1, \) and \( h = 1. \) We also assume that \( \mu = 1 \) and \( \theta = 3. \) We let \( \epsilon \) in (1) have a uniform distribution over \([-1,1]\) and vary the functional form for \( \sigma_n: \sigma_n = \sqrt{n}, n^{3/4}, \) and \( 0.25n. \) We report the optimal solutions of problems (3) and (4), \( \bar{n} \) and \( \hat{n}, \) and the corresponding relative and absolute errors.

We can make several observations based on our results in Table 2. First, as expected, fluid approximations are consistently less accurate than their stochastic-fluid counterparts. Second, stochastic-fluid approximations are extremely accurate, particularly when the variability in the number of servers, i.e., \( \sigma_n, \) is large. Third, our asymptotic results are useful in describing small systems, e.g., consisting of tens of servers. In particular, for \( \lambda = 20, \) the errors in both the stochastic-fluid and fluid approximations are remarkably small. Fourth, the results in Table 2 substantiate the orders of magnitude reported in Theorem 1. For example, focusing on the stochastic-fluid approximation and \( \sigma_n = n^{3/4}, \) we see that when \( \lambda \) is multiplied by a factor of \( l = 50 \) (in going from 20 to 1000, i.e., first to last row), the corresponding stochastic-fluid errors increase roughly by a factor of 18, which is approximately equal to \( l^{0.75} = 50^{0.75}. \) In other words, our numerical results suggest that those errors are on the order of magnitude of \( \sigma_{\lambda}, \) as given by Theorem 1.

5. Capacity Sizing with Stationary Demand
The asymptotic results of Theorem 1 quantify the optimality gaps for the fluid and stochastic-fluid prescriptions, provided that those prescriptions are consistent with their specifications in the theorem. There remains to show that this is indeed the case. We do so in this section by considering the stationary-demand case; this is equivalent to considering a single period, so we let \( k = 1 \) in what follows. We consider the case with time-varying demand in the following section.

5.1. Solution of the Fluid Problem
The solution to the fluid problem relaxation in (4) is straightforward, and given by the following lemma whose proof we omit.

**Lemma 2.** With a single period or, equivalently, with stationary demand, the solution to the fluid relaxation in (4) is to staff the cheaper resource only, i.e.,
(a) if \( c_0 \leq c_1, \) then \( \bar{m}_\lambda = \lambda/\mu \) and \( \bar{n}_\lambda = 0; \)
(b) if \( c_1 < c_0, \) then \( \bar{m}_\lambda = 0 \) and \( \bar{n}_\lambda = \lambda/\mu. \)
\[ \sigma_n = \sqrt{n} \]

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} |
|---|---|---|---|---|---|---|---|
| 20 | 24 | 22 | 20 | 0.173 | 0.700 | 1.86 | 7.20 |
| 50 | 57 | 53 | 50 | 0.333 | 1.13 | 1.60 | 5.43 |
| 100 | 110 | 105 | 100 | 0.360 | 1.64 | 0.921 | 4.20 |
| 150 | 162 | 156 | 150 | 0.474 | 2.04 | 0.83 | 3.58 |
| 500 | 522 | 511 | 500 | 0.894 | 3.82 | 0.498 | 2.13 |
| 1000 | 1031 | 1016 | 1000 | 1.18 | 5.40 | 0.337 | 1.54 |

\[ \sigma_n = n^{3/4} \]

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} |
|---|---|---|---|---|---|---|---|
| 20 | 25 | 24 | 20 | 0.010 | 0.556 | 0.101 | 5.38 |
| 50 | 59 | 58 | 50 | 0.015 | 1.12 | 0.0646 | 4.80 |
| 100 | 115 | 114 | 100 | 0.010 | 1.92 | 0.0231 | 4.39 |
| 150 | 171 | 169 | 150 | 0.0116 | 2.64 | 0.0182 | 4.15 |
| 500 | 551 | 550 | 500 | 6.19 \times 10^{-4} | 6.92 | 3.13 \times 10^{-4} | 3.50 |
| 1000 | 1086 | 1085 | 1000 | 2.50 \times 10^{-5} | 12.1 | 7.0 \times 10^{-6} | 3.14 |

\[ \sigma_n = 0.25n \]

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} | 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} |
|---|---|---|---|---|---|---|---|
| 20 | 24 | 22 | 20 | 0.143 | 0.595 | 1.51 | 6.30 |
| 50 | 57 | 55 | 50 | 0.0915 | 1.00 | 0.419 | 4.60 |
| 100 | 113 | 111 | 100 | 0.0211 | 1.61 | 0.0500 | 3.82 |
| 150 | 168 | 166 | 150 | 0.0212 | 2.23 | 0.0340 | 3.56 |
| 500 | 556 | 555 | 500 | 6.83 \times 10^{-4} | 6.70 | 3.35 \times 10^{-4} | 3.29 |
| 1000 | 1110 | 1109 | 1000 | 1.61 \times 10^{-3} | 13.2 | 3.96 \times 10^{-4} | 3.25 |

| Table 2 | Asymptotic accuracy of the stochastic-fluid and fluid solutions, \( \bar{n} \) and \( \tilde{n} \), to problems (3) and (4). |

5.2. Exact Solution for the Stochastic-Fluid Problem

For expositional ease, we define \( \beta \equiv (h/\theta + r)\mu \). Here is our first theorem.

**Theorem 2.** In an optimal solution to (3), we must have that:

\[
\tilde{n}_{\lambda} + \bar{m}_{\lambda} = \frac{\lambda}{\mu} - \sigma_{\bar{m}_{\lambda}} F^{-1} \left( \frac{\sigma_{\bar{n}_{\lambda}}}{\sigma_{\bar{n}_{\lambda}} + \sigma'_{\bar{n}_{\lambda}} (\lambda/\mu - \bar{n}_{\lambda} - \bar{m}_{\lambda})} \left( \frac{c_1}{\beta} - \sigma'_{\bar{n}_{\lambda}} \int_{-1}^{\lambda/\mu - \bar{n}_{\lambda} - \bar{m}_{\lambda}} F(x)dx \right) \right). \tag{5}
\]

That is,

\[
\bar{m}_{\lambda} + \bar{n}_{\lambda} = \lambda/\mu + \mathcal{O}(\sigma_{\bar{n}_{\lambda}}). \tag{6}
\]

The expression in (5) is an implicit relation between \( m_{\lambda} \) and \( n_{\lambda} \). It is insightful to contrast this with the remarkably simple critical fractile solution (to the stochastic-fluid problem with random arrival rates) in Bassamboo et al. (2010). The additional complexity in our setting is because when the number of servers itself is random, the choice of capacity levels impacts the distribution of the number of servers, which complicates the solution to the stochastic-fluid problem.
The expression in (6) is useful because it allows us to quantify the order of magnitude of the required safety hedge: It indicates that the optimal staffing level is in the form of a base capacity, $\lambda/\mu$, which matches the mean offered load, in addition to an uncertainty hedge, on the order of $O(\sigma\lambda/\mu)$. This additional capacity hedges against the uncertainty in the underlying parameter, i.e., the number of servers, much like the usual square-root additional capacity hedges against stochastic fluctuations in the system (Garnett et al. 2002). The asymptotic form in (6) coincides with the desired asymptotic form in Theorem 1, which implies that the solution to the stochastic-fluid formulation will indeed be “extremely” accurate, particularly when the uncertainty in the number of servers is large. Since the solution in (5) is given by an implicit relation, it cannot be generally used to derive the exact stochastic-fluid staffing level. Thus, we go further in §5.3, and derive asymptotically optimal solutions to that problem instead. We close this section by discussing a special case where it is easy to derive an exact, closed-form, solution to problem (3).

**Lemma 3.** If $\sigma_n = an$, for $a < 1$, then there exists $\omega < 0$ such that the solution to (3) is:

(a) If $c_1 > c_0 + \omega$, then:

$$\bar{m}_\lambda = \frac{\lambda}{\mu} \text{ and } \bar{n}_\lambda = 0;$$

(b) If $c_1 \leq c_0 + \omega$, then:

$$\bar{m}_\lambda = 0 \text{ and } \bar{n}_\lambda = \frac{\lambda}{\mu} \eta^*,$$

where $\eta^*$ denotes the solution to

$$c_1 + \beta a \int_{-1}^{1/(a\eta) - 1/a} F(u)du - \frac{\beta}{\eta} F \left( \frac{1}{a\eta} - \frac{1}{a} \right) = 0.$$

In particular, $\eta^* \geq (\leq) 1$ if we also have that $c_1 \leq (\geq) \beta F(0) - \beta a \int_{-1}^{0} F(u)du$.

Lemma 3 covers the case where $\sigma_\lambda$ is “very large”. In this case, it is possible for the manager to rely solely on the more expensive, yet less uncertain, supply i.e., on the fixed resource only. In other words, the manager incurs an additional cost for that uncertainty, and this cost is quantified by $|\omega|$ in the lemma. This remains true so long as the disparity in pay between the fixed and flexible resources is not too large. Otherwise, the difference in the staffing costs between the fixed and flexible resources is so large that it remains worthwhile for the system manager to staff only the flexible resource, despite the uncertainty that this entails (this is case (b) in the lemma). Thus, as a rule of thumb: When the flexible resource is highly uncertain, and demand is stationary, the manager relies exclusively on one of the two resources, but not necessarily the cheaper of the two.

Interestingly, Lemma 3 also shows that when staffing from the flexible pool, different possibilities emerge: The manager may either match supply and demand (as she would do when staffing from

$$5 \quad \omega \equiv c_0 a F^{-1}(\alpha_0) - a \beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx < 0 \text{ where } \alpha_0 \equiv \frac{\alpha}{\mu/\theta + r} = \frac{\alpha}{\beta}.$$
the fixed pool), or intentionally understaff or overstaff her system, depending on how small \( c_1 \) is. In other words, conventional wisdom for workforce management, whereby supply and demand are matched in expectation, no longer applies when the pool of flexible workers is highly uncertain.

5.3. Asymptotic Analysis

We let \( \Pi^*_\lambda \equiv \Pi_\lambda(m^*_\lambda, n^*_\lambda) \) denote the optimal objective value to the original problem in (2); we also let \( \hat{m}_\lambda \) and \( \hat{n}_\lambda \) denote the asymptotically optimal solutions to (3), where the exact solutions are given by the implicit relation in Theorem 2. Here is our main theorem.

**Theorem 3.** The approximate solution to the stochastic-fluid problem in (3), with stationary demand, is given by:

(a) If \( c_0 < c_1 \), then \( \hat{m}_\lambda = \lambda/\mu \) and \( \hat{n}_\lambda = 0 \) is \( \mathcal{O}(\sqrt{\lambda}) \) optimal, i.e.,

\[
\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda}).
\]

(b) If \( c_0 > c_1 \), then we consider two cases:

(b.I) If \( \sigma_n = \Theta(n^q) \), \( 0 \leq q \leq 1/2 \), then \( \hat{m}_\lambda = 0 \) and \( \hat{n}_\lambda = \lambda/\mu \) is \( \mathcal{O}(\sqrt{\lambda}) \) optimal, i.e.,

\[
\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda}).
\]

(b.II) If \( \sigma_n = \Theta(n^q) \), \( 1/2 < q < 1 \), then \( \hat{m}_\lambda = 0 \) and \( \hat{n}_\lambda = \lambda/\mu + \gamma^* \sigma_{\lambda/\mu} \) is \( o(\sigma_{\lambda}) \) optimal, i.e.,

\[
\lim_{\lambda \to \infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) - \Pi^*_\lambda}{\sigma_{\lambda}} = 0,
\]

where \( \gamma^* \) denotes the optimal solution to

\[
\min_{\gamma} c_1 \gamma + \beta \cdot \mathbb{E}[(-\gamma - \epsilon)^+].
\]

We can make several noteworthy observations based on Theorem 3. One main managerial insight which follows from our analysis is that, when demand is stationary, a manager should not blend fixed and flexible resources, i.e., using a blended workforce is not justified in this case. Intuitively, this is so because when customer demand rates do not vary substantially, there are only two effects which come into play: (i) the staffing cost (fixed and flexible), and (ii) the supply-side uncertainty (only flexible). The interaction between those two effects dictates the optimal staffing level in the system.

As a rule of thumb: When the flexible resource is moderately uncertain, and demand is stationary, the manager should staff only from the cheaper resource, but the additional safety-capacity hedge depends on whether the fixed or flexible resources are used. In particular, since the fixed resource does not entail any uncertainty (fixed servers are assumed to always show up), we must have that when \( c_0 < c_1 \), a manager should rely only on that fixed capacity.
It is interesting to inspect the solution for $c_0 > c_1$, i.e., when the fixed capacity is the more expensive. In this case, the supply-side uncertainty comes into play and the solution to the stochastic-fluid relaxation depends on its order of magnitude, as quantified by $\sigma_\lambda$. In particular, when $\sigma_\lambda$ is “small”, i.e., of an order that is at most equal to the order of stochastic fluctuations in the system, the manager should rely only on the flexible capacity. In this case, there is no advantage in using an uncertainty hedge, and the familiar square-root staffing hedge (Garnett et al. 2002) would yield the same optimality gap. On the other hand, when $\sigma_\lambda$ is “moderately large”, i.e., of an order that is larger than the order of stochastic fluctuations in the system but smaller than the order of magnitude of the arrival rate (Lemma 3), the manager would still rely solely on the flexible resource, but it is beneficial in this case to introduce an uncertainty hedge which is larger (in order of asymptotic magnitude) than the familiar square-root staffing hedge. The resulting solution is asymptotically optimal to the original problem in (2), and the corresponding optimality gap is $o(\sigma_\lambda)$. Our numerical results, described in the following section, indicate that this is only a loose upper bound, and that the optimality gap is actually smaller than this.

5.4. Numerical Study: On the Magnitude of the Appropriate Hedge

The results of Theorem 3 are asymptotic results; for completeness, we now provide supporting numerical results to test whether the optimality gaps of Theorem 3 are tight. We show that this is not the case, i.e., the accuracy of the approximate stochastic-fluid solutions is superior to the one specified by the theorem. In what follows, we assume that $c_1 < c_0$ and consider a variance term $\sqrt{\lambda} \leq \sigma_\lambda < \lambda$. Then, the optimal solution to (2) is to staff only flexible servers (cases b.I and b.II in Theorem 3). In Table 3, we let $c_1 = 1/3$, $p = 1$, $h = 1$, $\mu = 1$, and $\theta = 3$. We let $\epsilon$ in (1) have a uniform distribution over $[-1, 1]$, and we vary the value of $\lambda$ from 50 to 1,000.

We consider three functional forms for $\sigma_n$: $\sqrt{n}$, $n^{3/4}$, and $n^{0.9}$. In each case, we calculate (numerically): the optimal solution to (2), $n^*$, and the optimal solution to (3), $\bar{n}$. We also calculate the approximate solution, $\hat{n}$, as given by cases (b.I) and (b.II) in Theorem 3: The solution for case I is identical to the fluid-based solution, whereas the solution to case II involves an uncertainty hedge, i.e., $\hat{n} = \lambda/\mu + \gamma*\sigma_\lambda/\mu$. In each case, we calculate corresponding relative and absolute errors.

Table 3 shows that, in general, the approximate solution $\hat{n}$ of Theorem 3 is quite accurate. In particular, for $\sigma_n = \sqrt{n}$ there is no distinct advantage in using an uncertainty hedge. Indeed, when $\lambda$ increases from 50 to 1,000, i.e., is multiplied by a factor of 20, the optimality gaps for $\bar{n}$ and $\hat{n}$, in the fifth and sixth columns of the table, are almost identical, and both yield absolute errors on the order of magnitude of $\sqrt{\lambda}$.

For $\sigma_n = n^{3/4}$, there is still no noticeable difference in performance between the two solutions: While $\bar{n}$ yields a slightly smaller optimality gap, $\hat{n}$ remains asymptotically accurate. With a large
variance in the number of servers, i.e., for \( n = n^{0.9} \), Table 3 suggests that, while the optimality gap for \( \hat{n} \) grows with \( \lambda \) (unlike for \( \bar{n} \)), it does not increase by much. In particular, our numerical results suggest that the optimality gap in this case is on the order of magnitude of \( O(\sqrt{\lambda}) \).

\[
\sigma_n = \sqrt{n}
\]

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \hat{n} \) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|\) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|\) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|}{\Pi_\lambda(n^*)} \) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|}{\Pi_\lambda(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 57 | 53 | 54 | 0.333 | 0.140 | 1.60 | 0.674 |
| 100 | 110 | 105 | 105 | 0.360 | 0.360 | 0.921 | 0.921 |
| 300 | 317 | 309 | 309 | 0.604 | 0.604 | 0.550 | 0.550 |
| 500 | 522 | 511 | 511 | 0.895 | 0.895 | 0.499 | 0.499 |
| 700 | 726 | 713 | 713 | 1.07 | 1.07 | 0.431 | 0.431 |
| 900 | 929 | 915 | 915 | 1.18 | 1.18 | 0.371 | 0.371 |
| 1000 | 1031 | 1016 | 1016 | 1.21 | 1.21 | 0.344 | 0.344 |

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \hat{n} \) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|\) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|\) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|}{\Pi_\lambda(n^*)} \) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|}{\Pi_\lambda(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 59 | 58 | 59 | 0.0150 | 0.0 | 0.0646 | 0.0 |
| 100 | 115 | 114 | 116 | 0.0101 | 4.66 \times 10^{-3} | 0.0232 | 0.0107 |
| 300 | 334 | 333 | 336 | 6.75 \times 10^{-3} | 7.67 \times 10^{-3} | 5.55 \times 10^{-3} | 6.30 \times 10^{-3} |
| 500 | 551 | 550 | 553 | 6.17 \times 10^{-4} | 0.0137 | 3.12 \times 10^{-4} | 6.91 \times 10^{-3} |
| 700 | 765 | 764 | 768 | 3.07 \times 10^{-3} | 0.0144 | 1.13 \times 10^{-3} | 5.30 \times 10^{-3} |
| 900 | 979 | 978 | 982 | 1.49 \times 10^{-3} | 0.0154 | 4.29 \times 10^{-4} | 4.44 \times 10^{-3} |
| 1000 | 1086 | 1085 | 1089 | 0.0 | 0.0184 | 0.0 | 4.80 \times 10^{-3} |

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \hat{n} \) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|\) | \( |\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|\) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})|}{\Pi_\lambda(n^*)} \) | \( 100 \cdot \frac{|\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})|}{\Pi_\lambda(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 59 | 58 | 67 | 8.28 \times 10^{-3} | 0.346 | 0.0295 | 1.23 |
| 100 | 118 | 117 | 132 | 3.56 \times 10^{-3} | 0.593 | 6.60 \times 10^{-3} | 1.10 |
| 300 | 353 | 352 | 385 | 2.05 \times 10^{-3} | 1.16 | 1.34 \times 10^{-3} | 0.758 |
| 500 | 588 | 587 | 634 | 8.53 \times 10^{-4} | 1.57 | 3.41 \times 10^{-4} | 0.626 |
| 700 | 822 | 821 | 882 | 9.29 \times 10^{-4} | 1.99 | 2.69 \times 10^{-4} | 0.570 |
| 900 | 1056 | 1055 | 1128 | 6.24 \times 10^{-4} | 2.29 | 1.42 \times 10^{-4} | 0.520 |
| 1000 | 1173 | 1172 | 1251 | 4.11 \times 10^{-4} | 2.46 | 0.0 | 0.506 |

Table 3  Optimality gaps for the exact (\( \bar{n} \)) and approximate (\( \hat{n} \)) solutions of the stochastic-fluid problem in (3).

6. Capacity Sizing with Time-Varying Demand

Our analysis thus far has focused on the setting with stationary demand. We now consider a setting where demand varies according to predictable patterns, e.g., due to seasonality effects. We adopt this model for arrivals, although we recognize its shortcomings. For example, estimates of mean demand rates are often quite noisy; also, there may be unexpected surges in demand. As a result, the demand rates themselves may be considered to be random. However, since it is natural to begin
an investigation in a relatively tractable setting, we assume out such ambiguity in the arrival rates. Instead, we focus solely on the case with time-varying, yet deterministic, demand rates.

Since our aim is to formulate general insights, we consider a simple model with only two periods: A high-demand period with arrival rate $\lambda_H$, and a low-demand period with arrival rate $\lambda_L$; we assume that $\lambda_H > \lambda_L$. Consistently with our notation in §4, we fix $\lambda > 0$ and let $\lambda_i = \lambda \xi_i$, where $\xi_i \geq 0$ for $i \in \{H, L\}$. To derive asymptotic results, we let $\lambda$ increase without bound. Our original capacity-sizing problem in (2), reduced to a two-period setting, can be formulated as:

$$\max_{m_\lambda, n_\lambda, \alpha_\lambda^H, \alpha_\lambda^L} \Pi_\lambda(m_\lambda, n_\lambda, \alpha_\lambda^H, \alpha_\lambda^L) = T_H \Gamma_H(m_\lambda, n_\lambda, \alpha_\lambda^H) + T_L \Gamma_L(m_\lambda, n_\lambda, \alpha_\lambda^L),$$

(7)

where we define:

$$\Gamma_H(m_\lambda, n_\lambda, \alpha_\lambda^H) \equiv c_0 m_\lambda + c_1 n_\lambda + (h + \theta r)E[(X^H(m_\lambda, \alpha_\lambda^H n_\lambda) - N(m_\lambda, \alpha_\lambda^H n_\lambda))^+],$$

$$\Gamma_L(m_\lambda, n_\lambda, \alpha_\lambda^L) \equiv c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + (h + \theta r)E[(X^L(m_\lambda, \alpha_\lambda^L n_\lambda) - N(m_\lambda, \alpha_\lambda^L n_\lambda))^+].$$

Given (7), it is readily seen that we can normalize $\alpha^H = 1$, i.e., we can drop it from the problem formulation. We do so, hereafter, and assume that a cap $\alpha^L$ (which we denote by $\alpha_\lambda$, dropping dependence on the period) is applied to the flexible pool in the low-demand period.

### 6.1. Optimal solution

We now turn to formulating and describing the solutions of the two approximations, fluid and stochastic-fluid, of the original problem in (7).

#### 6.1.1. Fluid Problem

For our fluid approximation, we reduce the formulation in (4) to a two-period setting, and obtain the following:

$$\min_{m_\lambda, n_\lambda, \alpha_\lambda} \tilde{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) = T_H \left( c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r)(\lambda_H - m_\lambda \mu - n_\lambda \mu)^+ \right) + T_L \left( c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + (h/\theta + r)(\lambda_L - m_\lambda \mu - \alpha_\lambda n_\lambda \mu)^+ \right).$$

(8)

We are now ready to describe the optimal solution, $(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda)$, to the problem in (8). This solution depends on both the staffing costs and the lengths of the respective periods.

**Lemma 4.** The solution to the fluid problem in (8), with time-varying demand, is as follows:

1. If $c_0 \leq c_1$, then there are two subcases:
   - (a) If $c_0 \leq \frac{T_H}{T_H + T_L} c_1$, then:
     $$\tilde{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \tilde{n}_\lambda = 0, \quad \text{and} \quad \tilde{\alpha}_\lambda = 0;$$
   - (b) If $\frac{T_H}{T_H + T_L} c_1 < c_0 \leq c_1$, then:
     $$\tilde{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \tilde{n}_\lambda = \frac{\lambda_H}{\mu} - \frac{\lambda_L}{\mu}, \quad \text{and} \quad \tilde{\alpha}_\lambda = 0;$$

2. If $c_0 > c_1$, then:
   - (a) If $c_0 \leq \frac{T_H}{T_H + T_L} c_1$, then:
     $$\tilde{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \tilde{n}_\lambda = 0, \quad \text{and} \quad \tilde{\alpha}_\lambda = 0;$$
   - (b) If $\frac{T_H}{T_H + T_L} c_1 < c_0 \leq c_1$, then:
     $$\tilde{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \tilde{n}_\lambda = \frac{\lambda_H}{\mu} - \frac{\lambda_L}{\mu}, \quad \text{and} \quad \tilde{\alpha}_\lambda = 0;$$
ii) If $c_0 > c_1$, then:

\[ \tilde{m}_\lambda = 0, \quad \tilde{n}_\lambda = \frac{\lambda_H}{\mu}, \quad \text{and} \quad \tilde{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H}. \]

Lemma 4 coins the advantage of staffing a pool of flexible agents: Such a pool can be dynamically adjusted to meet time-fluctuations in customer demand, e.g., by setting a cap $\alpha_\lambda$ on supply in the low-demand period. Indeed, assuming out the uncertainty in staffing a flexible pool (fluid approximation), it is clear that if flexible servers are cheaper, i.e., $c_1 < c_0$, then the manager should staff only flexible servers and set a cap in the low-demand period; this is case (ii) in Lemma 4.

The only case where the manager would staff fixed servers is if $c_1 \geq c_0$. 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, as in case $i(a)$. The threshold $\frac{T_H}{T_H + T_L}$ in case (i) is the length of the high-demand period, $T_H$, relative to the length of the entire horizon, $T_L + T_H$. In words, as the high-demand period increases (decreases) in length, it becomes more (less) cost-effective for the manager to staff a pool of fixed servers to match demand in the high period, since the relative cost of overstaffing the low period decreases (increases).

6.1.2. Stochastic-Fluid Problem. The stochastic-fluid optimization problem is given by:

\[
\min_{m_\lambda, n_\lambda, \alpha_\lambda} \tilde{\Pi}(m_\lambda, n_\lambda, \alpha_\lambda) = T_H \left( c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r)E \left[ (\lambda_H - N(m_\lambda, n_\lambda)\mu)^+ \right] \right) + T_L \left( c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + (h/\theta + r)E \left[ (\lambda_L - N(m_\lambda, \alpha_\lambda n_\lambda)\mu)^+ \right] \right). \tag{9}
\]

For the solution of (9), we begin by treating the special case where $\sigma_n$ is a linear function of $n$. For ease of exposition, we define $\beta_0 \equiv c_0 \theta / ((h + r)\mu)$ and $\beta_1 \equiv \beta_0 (T_H + T_L)/T_H$. We also recall that $\beta \equiv \left(\frac{h}{\theta} + r\right) \mu$. Additionally, it will be convenient to define the constants $\zeta_0 < \zeta_1 < c_0$ given by:

\[ \zeta_0 \equiv c_0 + c_0 a F^{-1}(\beta_1) - a \beta \int_{-1}^{F^{-1}(\beta_1)} F(x)dx, \]

and

\[ \zeta_1 \equiv c_0 + c_0 a F^{-1}(\beta_1) - a \beta \int_{-1}^{F^{-1}(\beta_1)} F(x)dx. \]

The asymptotic accuracy of the following (exact) solution is given in Theorem 1.

**Lemma 5.** If $\sigma_n = an$ for $a < 1$, then $(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda)$ is given as follows.

(a) If $c_1 \geq \zeta_1$, then:

\[ \bar{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \bar{n}_\lambda = 0, \quad \text{and} \quad \bar{\alpha}_\lambda = 0. \]

(b) If $\zeta_0 \leq c_1 < \zeta_1$, then:

\[ \bar{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \bar{n}_\lambda = \frac{(\lambda_H - \lambda_L)\eta^*_2}{\mu}, \quad \text{and} \quad \bar{\alpha}_\lambda = 0; \]
If $c_1 < \zeta_0$, then:

$$\bar{m}_\lambda = 0, \quad \bar{n}_\lambda = \frac{\lambda_H}{\mu} \eta_2^*, \quad \text{and} \quad \bar{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H},$$

where $\eta_2^*$ is the solution of

$$c_1 + \beta a \int_{-1}^{(1/(a\eta))-1/a} F(x)dx - \frac{\beta}{\eta} F\left(\frac{1}{a\eta} - \frac{1}{a}\right) = 0.$$

In particular, $\eta_2^* \geq (\leq) 1$ if we also have that $c_1 \leq (\geq) \beta F(0) - \beta a \int_{-1}^{0} F(u)du$.

Consistently with Lemma 3, Lemma 5 shows that when the flexible pool is highly variable, then it may be possible to staff the more expensive but fixed pool; this is case (a) in the lemma. This is only true, however, when the flexible resource is expensive enough, i.e., $c_1 > \zeta_1$. Indeed, when $c_1$ is moderately cheaper than the fixed capacity (case (b)), then it is cost-effective for the manager to use a blended workforce: She uses the fixed pool to match demand in the low period, and the flexible pool to staff up to demand in the high period. This lends support to existing business practices where managers typically resort to hiring temporary help during periods of peak in demand (the flexible pool is capped in the low period). This is no longer the case, however, when the flexible capacity is cheap enough (case (c)), in which case the manager staffs only from the flexible resource.

As a rule of thumb: When the flexible resource is highly uncertain but cheaper, demand is time-varying, and the disparity in compensation is moderate, the manager relies on a blended workforce. She relies exclusively on the flexible resource only when it is much cheaper and, otherwise, relies exclusively on the possibly more expensive fixed resource. When she uses flexible servers, then she also caps that supply in the low-demand period.

We now derive asymptotically optimal solutions to problem (9), paralleling our analysis in §5.3. We denote those solutions by $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$, and let $\Pi^*_\lambda$ denote the optimal objective value for the original problem in (2).

**Theorem 4.** The approximate solution to the stochastic-fluid problem in (9), with time-varying demand, is given by:

(a) If $\sigma_n = \Theta(n^q)$, for $0 \leq q \leq 1/2$, then:

$$\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda}),$$

for the fluid-optimal solution $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$ as given by Lemma 4.

(b) If $\sigma_n = \Theta(n^q)$, for $1/2 < q < 1$, then:

$$\lim_{\lambda \to \infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda) - \Pi^*_\lambda}{\sigma_\lambda} = 0,$$

where $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$ are given as follows.
(b.I) If $c_0 \leq \frac{T_H}{T_H + T_L}c_1$, then:

$$\hat{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \hat{n}_\lambda = 0, \quad \text{and} \quad \hat{\alpha}_\lambda = 0;$$

(b.II) If $\frac{T_H}{T_H + T_L}c_1 < c_0 \leq c_1$, then:

$$\hat{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \hat{n}_\lambda = \left(\frac{\lambda_H}{\mu} - \frac{\lambda_L}{\mu}\right) + \gamma_2^* \sigma_{\lambda_H/\mu - \lambda_L/\mu}, \quad \text{and} \quad \hat{\alpha}_\lambda = 0,$$

where $\gamma_2^*$ denote the optimal solution of:

$$\min_{\gamma} c_1 \gamma + (h + r \theta) \frac{\mu}{\theta} \mathbb{E} \left[ (-\gamma - \epsilon)^+ \right];$$

(b.III) If $c_1 < c_0$, then:

$$\hat{m}_\lambda = 0, \quad \hat{n}_\lambda = \frac{\lambda_H}{\mu} + \gamma^*_3 \sigma_{\lambda_H/\mu}, \quad \text{and} \quad \hat{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H} + \nu^*_3 \frac{\mu}{\lambda_H} \sigma_{\lambda_H/\mu},$$

where $\gamma^*_3$ and $\nu^*_3$ denote the optimal solutions to:

$$\min_{\gamma, \nu} c_1 \left( T_H + T_L \frac{\lambda_L}{\lambda_H} \right) \gamma + c_1 T_L \nu + (h/\theta + r) \mu T_H \mathbb{E} \left[ (-\gamma - \epsilon)^+ \right]$$

$$+ (h/\theta + r) \mu T_L \mathbb{E} \left[ -\frac{\lambda_L}{\lambda_H} \gamma - \nu - \left(\frac{\lambda_L}{\lambda_H}\right)^q \epsilon \right]^+ \right].$$

Theorem 4 demonstrates that if the uncertainty in the flexible pool is “low”, i.e., of an order of magnitude which is smaller than the square-root order of stochastic fluctuations in the system, then the solution to the problem remains similar to the fluid-optimal solution in Lemma 4.

More generally, when supply is uncertain but not greatly so (smaller than that in Lemma 5), we find that the manager should generally rely on the flexible pool, either alone or through blending. Indeed, the only case where the manager would rely solely on the fixed pool is case (b.I), where the fixed supply is much cheaper than the flexible one. Otherwise, when the fixed servers are cheaper than flexible ones, but not by too much (case (b.II)) then the manager should blend her workforce, matching with the fixed pool up to demand in the low period, and utilizing flexible agents to match up to demand in the high period. Finally, when flexible servers are cheaper, then the manager should depend only on those servers, as in case (b.III). In this case, she matches up to demand in the high period, and caps her supply in the low period. As a rule of thumb: When the flexible resource is moderately uncertain, and demand is time-varying, the manager relies exclusively on the fixed resource only when it is much cheaper. Otherwise, she staffs a blended workforce when the fixed resource is cheaper but not greatly so, and staffs exclusively flexible servers when the flexible resource is cheaper. When she staffs from the flexible resource, then she also uses a cap in the low-demand period.
7. Generally-Distributed Abandonment

The results of the previous sections were restricted to exponentially-distributed patience times. Since there is statistical evidence indicating that patience times may not always be exponential (Brown et al. 2005), it is important to go beyond this assumption. We do so in this section by describing results from a numerical study quantifying the optimality gaps for problems (3) and (4) with a non-exponential abandonment distribution.

In Tables 4 and 5, we present our results for 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: \( c = 1, h = 1, p = 0.45, \) and \( \mu = 1, \) and restrict attention to a system with only flexible capacity. In each case, we report the fluid, \( \bar{n}, \) stochastic-fluid, \( \tilde{n}, \) and original, \( n^*, \) optimal solutions, as well as the corresponding optimality gaps, both relative and absolute.

We first discuss our numerical results with Pareto abandonment. In this case, the overloaded regime is asymptotically optimal at fluid scale (Bassamboo and Randhawa 2010). When \( \sigma_n = \omega(n) \) (first two sub-tables in Table 4), the system remains overloaded despite the uncertainty in the number of servers. Thus, we expect that fluid prescriptions should be extremely accurate, i.e., with absolute errors on the order of magnitude of \( O(1). \) In other words, we expect that stochastic-fluid prescriptions would not lead to a substantial improvement over their fluid counterparts; this is confirmed by Table 4. Of particular interest is the case where \( \sigma_n = \sqrt{n}, \) which shows that the respective errors are much smaller than with exponential abandonment (How do I compare here?).

It is unclear, a priori, how the fluid and stochastic-fluid solutions would perform when \( \sigma_n \) is large, i.e., of an order of magnitude equal to \( O(n) \) (last sub-table in Table 4), because the uncertainty in the number of servers is on the same order as the offered load in this case. Table 4 shows that, while stochastic-fluid approximations are more accurate in this case, the difference in performance is not too great.

With Weibull abandonment, the fluid solution prescribes a critically-loaded regime. Thus, we expect the optimality gaps of our respective solutions to be close to those with exponential abandonment (Theorem 1). Table 5 confirms that this is indeed the case. In particular, for all values of \( \sigma_n, \) the stochastic-fluid formulation is remarkably accurate, yielding an order of magnitude improvement over the fluid prescription (in most cases, \( n^* \) and \( \bar{n} \) are indistinguishable). Moreover, Table 5 shows that the optimality gaps obtained are consistent with those reported in Theorem 1.

8. Concluding Remarks

In this paper, we studied the problem of staffing a service system with a blended workforce. In particular, we showed that the capacity decision reduces to characterizing the interaction between
Pareto, \( \sigma_n = \sqrt{n} \)

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 35 | 41 | 41 | 0.388 | 0.388 | 0.770 | 0.770 |
| 100 | 75 | 83 | 83 | 0.315 | 0.315 | 0.315 | 0.315 |
| 300 | 242 | 248 | 248 | 0.0325 | 0.0325 | 0.0109 | 0.0109 |
| 500 | 411 | 413 | 413 | 2.06 \times 10^{-3} | 2.06 \times 10^{-3} | 4.13 \times 10^{-4} | 4.13 \times 10^{-4} |
| 650 | 536 | 537 | 537 | 1.89 \times 10^{-4} | 1.89 \times 10^{-3} | 2.90 \times 10^{-5} | 2.90 \times 10^{-5} |

Pareto, \( \sigma_n = n^{3/4} \)

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 33 | 36 | 41 | 0.152 | 1.025 | 0.300 | 2.01 |
| 100 | 70 | 75 | 83 | 0.196 | 1.375 | 0.194 | 1.36 |
| 300 | 226 | 240 | 248 | 0.411 | 1.15 | 0.137 | 0.383 |
| 500 | 389 | 409 | 413 | 0.526 | 0.789 | 0.105 | 0.158 |
| 650 | 512 | 537 | 537 | 0.583 | 0.583 | 0.0898 | 0.0898 |

Pareto, \( \sigma_n = 0.25n \)

| \( \lambda \) | \( n^* \) | \( \bar{n} \) | \( \tilde{n} \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})| \) | \( |\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})| \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\bar{n})|}{\Pi_{\lambda}(n^*)} \) | \( 100 \cdot \frac{|\Pi_{\lambda}(n^*) - \Pi_{\lambda}(\tilde{n})|}{\Pi_{\lambda}(n^*)} \) |
|---|---|---|---|---|---|---|---|
| 50 | 34 | 39 | 41 | 0.278 | 0.592 | 0.550 | 1.17 |
| 100 | 72 | 79 | 83 | 0.324 | 0.861 | 0.322 | 0.857 |
| 300 | 226 | 241 | 248 | 0.485 | 1.18 | 0.161 | 0.392 |
| 500 | 383 | 401 | 413 | 0.447 | 1.43 | 0.089 | 0.286 |
| 650 | 501 | 521 | 537 | 0.425 | 1.64 | 0.0655 | 0.252 |

Table 4 Performance of the stochastic-fluid and fluid optimal solutions with Pareto abandonment.

three competing factors: (i) operational costs in the system; (ii) the time-variation in customer demand; and (iii) the supply-side uncertainty which is associated with staffing flexible agents.

Our analysis suggests that the optimal staffing policy is not straightforward, in that it may be cost-effective to staff strictly one of the two resources, or to use a blended workforce instead. In each case, the resulting 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.

Part of our analysis provides support to some current business practices. For example, a manager who uses only flexible agents should staff enough to match peak loads, but will resort to capping the number of active agents under low loads; e.g., this is common in virtual call centers which hire work-from-home agents \(^6\), and is also consistent with previous results from the strategic-server literature (Gurvich et al. 2017). However, our analysis also yields results which challenge other current business trends. Indeed, one main insight is that it may not always be cost-effective to staff a blended workforce, i.e., that the modern shift in the business world towards staffing a blended

---

\(^6\) https://www.glassdoor.co.uk/Reviews/Arise-Reviews-E31617.htm
workforce should be handled with caution. For example, this is the case when demand rates are stationary, or approximately so. Then, a manager should generally staff only from the cheaper alternative, fixed or flexible. We also find that it may be cost-effective to, counter-intuitively, staff from a more expensive resource. For example, this is the case when the variability in the flexible supply is large, i.e., flexible agents are highly unreliable. This suggests that companies that rely solely on independent contractors, e.g., ride-sharing services such as Uber or Lyft, may benefit from staffing a core of full-time agents as well. Finally, even if the reliable fixed resource is cheaper, it may still be optimal to staff flexible agents because doing so allows for the flexibility of adjusting the agent pool size dynamically to meet fluctuations in incoming customer demand. This suggests that companies that currently rely only on full-time fixed workers, because they view them as cheaper and more reliable than independent contractors, may benefit from staffing a layer of independent contractors if they face significant variations in customer demand patterns. A cost-effective staffing strategy when blending is to use enough fixed workers to match demand in off-peak demand periods, and use flexible workers to match up to demand during peak periods.
References


Ibrahim, R. 2017. Staffing a service system with a random service capacity and impatient customers. University College London, working paper.


Electronic Companion:

EC.1. More on the impact of randomness in capacity

Let $X(N)$ denote the steady-state number in a system with $N$ servers. Since $\mu = \theta$, $X(N)$ has the same distribution as the steady-state number in system of an $M/M/\infty$ queue with arrival rate $\lambda$ and service rate $\mu$. That is, $X(N)$ has a Poisson distribution with rate $\lambda/\mu$, and we note that $X(N)$ is independent of $N$ and that $Q(N) = (X(N) - N)^+$. We recall the following definitions of variability ordering between two random variables $X$ and $Y$ with respective cdf’s $F_X$ and $F_Y$ (Shaked and Shanthikumar 2007). The inverse $F^{-1}$ of a cdf $F$ is defined as:

$$F^{-1}(t) = \inf\{x : F(x) > t\} \quad \text{for} \quad 0 < t < 1.$$  

**Definition EC.1.** $X$ is said to be *less dispersed than* $Y$, denoted by $X \leq_{disp} Y$, if $F_X^{-1}(\beta) - F_X^{-1}(\alpha) \leq F_Y^{-1}(\beta) - F_Y^{-1}(\alpha)$ for all $0 < \alpha < \beta < 1$.

**Proof of Lemma 1.** If $\sigma_1 < \sigma_2$, then $N_1 \leq_{disp} N_2$. Since $X \sim \text{Poisson}(\lambda/\mu)$ has a log-concave probability mass function, and is independent of both $N_1$ and $N_2$,

$$X - N_1 \leq_{disp} X - N_2.$$  

As $E[X - N_1] = E[X - N_2]$ and $u(x) = x^+ \equiv \max\{x, 0\}$ non-decreasing and convex,

$$E[(X - N_1)^+] \leq E[(X - N_2)^+] \quad \text{i.e.,} \quad E[Q(N_1)] < E[Q(N_2)].$$  

□

EC.2. Results on optimality gaps

Before proof Theorem 1, we first provide some auxiliary lemma on the accuracy of stochastic fluid and fluid approximations.

**Lemma EC.1.** When $\mu = \theta$,

$$E[(\lambda/\mu - m - N(n))^+] \leq E[(X(m + N(n)) - m - N(n))^+] \leq E[(\lambda/\mu - m - N(n))^+] + O(\sqrt{\lambda})$$

Moreover, when $m_\lambda + n_\lambda = \lambda/\mu + O(\sigma_\lambda/\mu)$,

$$E[(X(m_\lambda, n_\lambda) - m_\lambda - n_\lambda - \sigma_\lambda \epsilon)^+] \leq E[(\lambda/\mu - m_\lambda - n_\lambda - \sigma_\lambda \epsilon)^+] + O \left( \min \left\{ \sqrt{\lambda}, \lambda/\sigma_\lambda \right\} \right).$$
Proof. When $\mu = \theta$, $X(\lambda, \mu, N(n)) \sim \text{Poisson}(\lambda/\mu)$. By Lemma 3 of Bassamboo et al. (2010), we have

$$
\left(\frac{\lambda}{\mu} - m - N(n)\right)^+ \leq \mathbb{E}[(X(\lambda, \mu, m + N(n)) - m - N(n))^+ | N(n)]
$$

$$
\leq \left(\frac{\lambda}{\mu} - m - N(n)\right)^+ + \sqrt{\frac{4\pi \lambda}{\mu}} \exp \left(-\frac{\mu}{4\lambda} \left(\frac{\lambda}{\mu} - m - N(n)\right)^2\right) + \frac{1}{\log 2}
$$

Then as $\exp \left(-\frac{\mu}{4\lambda} \left(\frac{\lambda}{\mu} - m - N(n)\right)^2\right) \leq 1$,

$$
\mathbb{E} \left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] \leq \mathbb{E}[X(\lambda, \mu, m + N(n)) - m - N(n))^+]$

$$
\leq \mathbb{E} \left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] + O(\sqrt{\lambda}).
$$

When $m_\lambda + n_\lambda = \lambda/\mu + O(\sigma_{\lambda/\mu})$,

$$
\sqrt{\frac{4\pi \lambda}{\mu}} \exp \left(-\frac{\mu}{4\lambda} \left(\frac{\lambda}{\mu} - m_\lambda - N(n_\lambda)\right)^2\right) = \sqrt{\frac{4\pi \lambda}{\mu}} \exp \left(-\frac{\mu}{4\lambda} \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda - \sigma_{\lambda/\mu}\epsilon\right)^2\right)
$$

Following the same line of argument as in Lemma 1 of Bassamboo et al. (2010), we have

$$
\mathbb{E} \left[\sqrt{\frac{4\pi \lambda}{\mu}} \exp \left(-\frac{\mu}{4\lambda} \left(\frac{\lambda}{\mu} - m_\lambda - N(n_\lambda)\right)^2\right)\right] = \begin{cases} 
O(\sqrt{\lambda}) & \text{if } \sigma_{\lambda} \leq O(\sqrt{\lambda}) \\
O(\lambda/\sigma_{\lambda}) & \text{if } \sigma_{\lambda} > O(\sqrt{\lambda}).
\end{cases}
$$

\square

Lemma EC.2.

$$
\bar{\Pi}_\lambda(m, n) \leq \Pi_\lambda(m, n) \leq \bar{\Pi}_\lambda(m, n) + O(\sqrt{\lambda})
$$

Moreover, when $m_\lambda + n_\lambda = \lambda/\mu + O(\sigma_{\lambda/\mu})$,

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

Proof. We prove the general case only. The case $m_\lambda + n_\lambda = \lambda/mu + O(\sigma_{\lambda/\mu})$ follows exact the same line of arguments.

When $\mu = \theta$, the result follows directly from Lemma EC.1.

When $\mu > \theta$, we first consider an auxiliary “upper bound” system with abandonment rate $\mu$. Let $A_\lambda(m + N(n)) = \theta \mathbb{E}[(X(m + N(n); \lambda, \mu, \theta) - m - N(n))^+]$ and $A_\lambda^I(m + N(n)) = \mu \mathbb{E}[(X(m + N(n); \lambda, \mu, \mu) - m - N(n))^+]$. As $A_\lambda(m + N(n)) \leq A_\lambda^I(m + N(n))$ (Bassamboo et al. 2010),

$$
\Pi_\lambda(m, n) = \begin{cases} 
0 & \text{if } m = 0 \\
c_0 m + c_1 n + (h/\theta + r)A_\lambda^I(m + N(n)) & \text{if } m > 0
\end{cases}
$$

$$
\leq c_0 m + c_1 n + (h/\theta + r)A_\lambda^I(m + N(n))$

$$
\leq c_0 m + c_1 n + (h + r\theta)\frac{\mu}{\theta} \mathbb{E}[(\lambda/\mu - m - N(n))^+] + O(\sqrt{\lambda}) 	ext{ by Lemma EC.1}
$$

$$
= \bar{\Pi}_\lambda(m, n) + O(\sqrt{\lambda}).
$$
We then consider an auxiliary “lower bound” system with service rate $\theta$. Let $A^{ll}_\lambda(m + N(n)) = \theta \mathbb{E}[(X(m + N(n); \lambda, \theta, \theta) - m - N(n))^+]$. As $A_\lambda(m + N(n)) \geq A^{ll}_\lambda((m + N(n))\mu/\theta)$ (Bassamboo et al. 2010),

$$
\Pi_\lambda(m, n) = c_0 m + c_1 n + (h/\theta + r)A_\lambda(m + N(n)) \\
\geq c_0 m + c_1 n + (h/\theta + r)A^{ll}_\lambda\left(\frac{\mu}{\theta}(m + N(n))\right) \\
\geq c_0 m + c_1 n + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} m - \frac{\mu}{\theta} N(n)\right)^+\right] \text{ by Lemma EC.1} \\
= c_0 m + c_1 n + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] \\
= \tilde{\Pi}_\lambda(m, n).
$$

When $\mu < \theta$, the proof is similar to the case of $\mu > \theta$. We first consider an auxiliary “upper bound” system with service rate $\theta$. Let $A^{ul}_\lambda(m + N(n)) = \theta \mathbb{E}[(X(m + N(n); \lambda, \theta, \theta) - m - N(n))^+]$. As $A_\lambda(m + N(n)) \leq A^{ul}_\lambda((m + N(n))\mu/\theta)$ (Bassamboo et al. 2010),

$$
\Pi_\lambda(m, n) = c_0 m + c_1 n + (h/\theta + r)A_\lambda(m + N(n)) \\
\leq c_0 m + c_1 n + (h/\theta + r)A^{ul}_\lambda\left(\frac{\mu}{\theta}(m + N(n))\right) \\
\leq c_0 m + c_1 n + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} m - \frac{\mu}{\theta} N(n)\right)^+\right] + O(\sqrt{\lambda}) \text{ by Lemma EC.1} \\
= c_0 m + c_1 n + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] \\
= \bar{\Pi}_\lambda(m, n) + O(\sqrt{\lambda}).
$$

We then consider an auxiliary “lower upper” bound system with abandonment rate $\mu$. Let $A_\lambda(m + N(n)) = \theta \mathbb{E}[(X(m + N(n); \lambda, \mu, \theta) - m - N(n))^+]$ and $A_\lambda(m + N(n)) = \mu \mathbb{E}[(X(m + N(n); \lambda, \mu, \mu) - m - N(n))^+]$. As $A_\lambda(m + N(n)) \geq A_\lambda(m + N(n))$ (Bassamboo et al. 2010),

$$
\Pi_\lambda(m, n) = c_0 m + c_1 n + (h/\theta + r)A_\lambda(m + N(n)) \\
\geq c_0 m + c_1 n + (h/\theta + r)A_\lambda(m + N(n)) \\
\geq c_0 m + c_1 n + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] \text{ by Lemma EC.1} \\
= \tilde{\Pi}_\lambda(m, n).
$$

$\square$

**Lemma EC.3.**

$$
\left(\frac{\lambda}{\mu} - m - n\right)^+ \leq \mathbb{E}\left[\left(\frac{\lambda}{\mu} - m - N(n)\right)^+\right] \leq \left(\frac{\lambda}{\mu} - m - n\right)^+ + O(\sigma_n).
$$
Proof. We notice that by Jensen’s inequality,

\[ \mathbb{E} \left[ \left( \frac{\lambda}{\mu} - m - N(n) \right)^+ \right] \geq \left( \frac{\lambda}{\mu} - m \right)^+. \]

For the upper bound, as \(-1 < \epsilon < 1,\)

\[ \mathbb{E} \left[ \left( \frac{\lambda}{\mu} - m - N(n) \right)^+ \right] = \mathbb{E} \left[ \left( \frac{\lambda}{\mu} - m - n - \sigma_n \epsilon \right)^+ \right] =
\]

\[ \begin{cases} 0 & \text{for } \frac{\lambda/\mu - m - n}{\sigma_n} < -1, \\ \sigma_n \int_{-1}^{\frac{\lambda/\mu - m - n}{\sigma_n}} F(x) dx & \text{for } -1 \leq \frac{\lambda/\mu - m - n}{\sigma_n} \leq 1, \\ \frac{\lambda}{\mu} - m - n & \text{for } \frac{\lambda}{\mu} - m - n > \sigma_n. \end{cases} \]

Therefore,

\[ \mathbb{E} \left[ \left( \frac{\lambda}{\mu} - m - N(n) \right)^+ \right] = \left( \frac{\lambda}{\mu} - m - n \right)^+ \cdot \mathbb{1} \left( \frac{\lambda}{\mu} - m - n > \sigma_n \right) + \mathcal{O}(\sigma_n) \]

\[ \leq \left( \frac{\lambda}{\mu} - m - n \right)^+ + \mathcal{O}(\sigma_n). \]

□

Proof of Theorem 1 For \( \bar{m} + \bar{n} = \frac{\lambda}{\mu} + \mathcal{O}(\sigma_{n/\mu}) \), from Lemma EC.2, we have

\[ \Pi_{\lambda}(\bar{m}, \bar{n}) \leq \bar{\Pi}_{\lambda}(\bar{m}, \bar{n}) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) \]

\[ \leq \bar{\Pi}_{\lambda}(m^*, n^*) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) \]

\[ \leq \Pi_{\lambda}(m^*, n^*) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}), \]

and form Lemma EC.2 & EC.3, we have

\[ \Pi_{\lambda}(\bar{m}, \bar{n}) \leq \bar{\Pi}_{\lambda}(\bar{m}, \bar{n}) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) \]

\[ \leq \bar{\Pi}_{\lambda}(\bar{m}, \bar{n}) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) + \mathcal{O}(\sigma) \]

\[ \leq \bar{\Pi}_{\lambda}(m^*, n^*) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) + \mathcal{O}(\sigma) \]

\[ \leq \Pi_{\lambda}(m^*, n^*) + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\sigma_{\bar{n}}\}) + \mathcal{O}(\sigma), \]

EC.3. Proofs of the results in Section 5

Proof of Theorem 2 We begin by fixing \( m_\lambda \) and solving for \( n_\lambda \). We first notice that

\[ \bar{\Pi}'_{\lambda}(n_\lambda, m_\lambda) = c_1 - \beta \frac{\sigma_{n_\lambda} + \sigma_{n_\lambda}'}{\sigma_{n_\lambda}} F\left( \frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}} \right) + \beta \sigma_{n_\lambda} \int_{-1}^{\frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}}} F(x) dx \]
and

\[
\frac{1}{\beta} \Pi'(n_\lambda) = \sigma'' \int_{-1}^{\lambda/\mu - n_\lambda - m_\lambda} x f(x) dx + \frac{\sigma_{n_\lambda} + \sigma'_{n_\lambda} (\lambda/\mu - n_\lambda - m_\lambda)^2}{\sigma_{n_\lambda}^3} f \left( \frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}} \right) \geq 0
\]

Solve for \( \Pi'(n_\lambda, m_\lambda) = 0 \), we have \( \bar{n}_\lambda \) and \( \bar{m}_\lambda \) must satisfy

\[
\bar{n}_\lambda + \bar{m}_\lambda = \frac{\lambda}{\mu} - \sigma_{n_\lambda} F^{-1} \left( \frac{\sigma_{n_\lambda} + \sigma'_{n_\lambda} (\lambda/\mu - \bar{n}_\lambda - \bar{m}_\lambda)}{\sigma_{n_\lambda}^2} (c_1 - \sigma'_{n_\lambda} \int_{-1}^{\lambda/\mu - n_\lambda - m_\lambda} F(x) dx) \right)
\]

Since \( F^{-1}(x) \in [-1, 1] \) for all \( x \in [0, 1] \), we have

\[
-\sigma_{n_\lambda} \leq \bar{m}_\lambda + \bar{n}_\lambda - \lambda/\mu \leq \sigma_{n_\lambda}.
\]

We now show that \( \bar{n}_\lambda \leq \lambda/\mu \). For this, assume in contradiction that \( \bar{n}_\lambda > \lambda/\mu \). Then note that for any \( \bar{m}_\lambda \): \( \Pi'(\bar{m}_\lambda, \bar{n}_\lambda) > \Pi'(0, \lambda/\mu) \). Thus, we must have that \( \bar{n}_\lambda \leq \lambda/\mu \). This implies that

\[
-\sigma_{n_\lambda} \leq \bar{n}_\lambda + \bar{m}_\lambda - \lambda/\mu \leq \sigma_{n_\lambda} \leq \sigma_{\lambda/\mu}.
\]

Therefore, \( \bar{m}_\lambda + \bar{n}_\lambda = \lambda/\mu + \mathcal{O}(\sigma_{\lambda/\mu}) \), as desired. \( \square \)

**EC.3.1. Asymptotically optimal solutions to the stochastic fluid model with stationary demand**

**EC.3.1.1. Case I: The optimal solution when \( \sigma_n = \Omega(n^q) \) for \( 0 \leq q \leq 1/2 \)**

When \( c_1 < c_0 \), the optimal solution to \( \min_{m,n} \Pi_\lambda(m, n) \) is \( \bar{m} = 0 \), \( \bar{n} = \lambda/\mu \). We notice that, in this case, \( \bar{m} = \bar{m} \) and \( \bar{n} = \bar{n} \). From Theorem 1, we have

\[
\Pi_\lambda(\bar{m}, \bar{n}) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_{n_\lambda}) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda})
\]

**EC.3.1.2. Case II: The optimal solution when \( \sigma_n = \Omega(n^q) \) for \( 1/2 < q < 1 \)**

Let \( m = x \sigma_{\lambda/\mu} \), \( n = \lambda/\mu + y \sigma_{\lambda/\mu} \) for \( x \in \mathbb{R}^+ \) and \( y \in \mathbb{R} \). Then optimize \( \Pi_\lambda(m, n) \) is equivalent to optimize

\[
C_\lambda(x, y) = \frac{\Pi_\lambda(x \sigma_{\lambda/\mu}, \lambda/\mu + y \sigma_{\lambda/\mu}) - c_1 \lambda/\mu}{\sigma_{\lambda/\mu}^2} = c_0 x + c_1 y + (h + r \theta) \mu \mathbb{E} \left[ \left( -\left( \frac{x + y}{\sigma_{\lambda/\mu}^2} \right) \right)^+ \right]
\]

We denote the optimal solution of \( C_\lambda(x, y) \) as \( x^*_\lambda, y^*_\lambda \).

**Lemma EC.4.**

\[
\lim_{\lambda \to \infty} \sup x^*_\lambda < \infty, \quad \lim_{\lambda \to \infty} \sup y^*_\lambda < \infty \quad \text{and} \quad \lim_{\lambda \to \infty} \inf y^*_\lambda > -\infty.
\]
Proof. We first notice that $C_\lambda(0, 0) = (h + r \theta) \frac{\mu}{\theta} (-\epsilon)^\top$. We next prove the lemma by contradiction.

Assume if $\limsup_{\lambda \to \infty} y_\lambda^* = \infty$, then for $M = C_\lambda(0, 0)/c_1$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x_{\lambda_k}^* > M$. In this case,

$$C_{\lambda_k}(x_{\lambda_k}^*, y_{\lambda_k}^*) \geq c_1 y_{\lambda_k}^* > c_1 M = C_\lambda(0, 0).$$

Thus, we get a contradiction.

Assume if $\liminf_{\lambda \to \infty} y_\lambda^* = -\infty$, we divide the analysis into two cases.

i) If $\liminf_{\lambda \to \infty} x_{\lambda}^* + y_\lambda^* = -\infty$, Pick $L_1 = C_\lambda(0, 0)/(c_1 - (h + r \theta)) < 0$. For such $L_1$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x_{\lambda_k}^* + y_{\lambda_k}^* < -L_1$. In this case,

$$C_{\lambda_k}(x_{\lambda_k}^*, y_{\lambda_k}^*) > c_1 L_1 + (h + r \theta)(-L) = C_\lambda(0, 0).$$

Thus, we get a contradiction.

ii) If $\liminf_{\lambda \to \infty} x_{\lambda}^* + y_\lambda^* > -\infty$, we denote $L_2 = \min \{\liminf_{\lambda \to \infty} x_{\lambda}^* + y_\lambda^*, 0\}$. As $\liminf_{\lambda \to \infty} y_\lambda^* = -\infty$, we must have $\limsup_{\lambda \to \infty} x_{\lambda}^* = \infty$. Pick $M_2 = (C_\lambda(0, 0) - c_1 L)/(c_0 - c_1)$. For such $M_2$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x_{\lambda_k}^* > M_2$. In this case

$$C_{\lambda_k}(x_{\lambda_k}^*, y_{\lambda_k}^*) > (c_0 - c_1) M_2 + c_1 L_2 = C_\lambda(0, 0).$$

Thus, we get a contradiction.

We have already shown that $\limsup_{\lambda \to \infty} y_\lambda^* < \infty$ and $\liminf_{\lambda \to \infty} y_\lambda^* > -\infty$. Now let $L_3 = \min \{\liminf_{\lambda \to \infty} y_\lambda^*, 0\}$. Assume if $\limsup_{\lambda \to \infty} x_{\lambda}^* = \infty$, then for $M_3 = C_\lambda(0, 0)/c_0 - c_1 L_3/c_0$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x_{\lambda_k}^* > M_3$. In this case,

$$C_{\lambda_k}(x_{\lambda_k}^*, y_{\lambda_k}^*) \geq c_0 x_{\lambda_k}^* + c_1 y_{\lambda_k}^* > c_0 M_3 + c_1 L_3 = C_\lambda(0, 0).$$

Thus, we get a contradiction.

Let

$$\hat{C}(x, y) = c_0 x + c_1 y + (h + r \theta) \frac{\mu}{\theta} (-\epsilon)^\top.$$

Let $\hat{x}$ and $\hat{y}$ denote the optimal solution of $\hat{C}(x, y)$. As $c_1 < c_0$, $\hat{x} = 0$ and $\hat{y} = \gamma^*$

**Lemma EC.5.**

$$C_\lambda(x_{\lambda}^*, y_{\lambda}^*) \to C_\lambda(0, \gamma^*) \text{ as } \lambda \to \infty.$$

**Proof.** As $\sigma_{\lambda/\mu+y_{\lambda/\mu}}/\sigma_{\lambda/\mu} \to 1$ uniformly on compact sets (u.o.c.) as $\lambda \to \infty$,

$$C_\lambda(x, y) \to \hat{C}(x, y) \text{ u.o.c. as } \lambda \to \infty.$$

Then from Lemma EC.4, we have

$$C_\lambda(x_{\lambda}^*, y_{\lambda}^*) \to \hat{C}(x^*, y^*) \text{ as } \lambda \to \infty.$$
We are now ready to prove part II) of Theorem 4. We first notice that $\Pi_\lambda(\hat{m}, \hat{n}) > \Pi^*_\lambda$. We also have

$$
\Pi_\lambda(\hat{m}, \hat{n}) \leq \Pi_\lambda(\hat{m}, \hat{n}) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2}
$$

$$
\leq \Pi_\lambda(\hat{m}, \hat{n}) + o(\sigma_\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.5}
$$

$$
\leq \Pi^*_\lambda + o(\sigma_\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2}
$$

\[\leq \Pi^*_\lambda + o(\sigma_\lambda)\]

**EC.3.1.3. Case III: The optimal solution when $\sigma_n = an + o(n)$** Let $m = x\lambda/\mu, n = y\lambda/\mu$ for $x, y \in \mathbb{R}^+$. Then optimizing $\Pi_\lambda(m, n)$ is equivalent to optimize

$$
V_\lambda(x, y) = \frac{\Pi_\lambda(m, n)}{\lambda/\mu} = c_0 x + c_1 y + (h + r\theta) \frac{\mu}{\theta} E \left[ \left( 1 - x - y - \frac{\sigma_y\lambda/\mu}{\lambda/\mu} \epsilon \right) + \right].
$$

We denote the optimal solution to $V_\lambda(x, y)$ as $x^*_\lambda, y^*_\lambda$. We also let

$$
\hat{V}(x, y) = c_0 x + c_1 y + (h + r\theta) \frac{\mu}{\theta} E \left[ (1 - x - y - a\epsilon) + \right]
$$

We denote the optimal solution of $\hat{V}(x, y)$ as $x^*, y^*$. We thus get a contradiction.

**Proof.** We first show that $\limsup_{\lambda \to \infty} x^*_\lambda < \infty$ and $\limsup_{\lambda \to \infty} y^*_\lambda < \infty$. We first notice that $V_\lambda(1, 0) = c_0$. Suppose if $\limsup_{\lambda \to \infty} x^*_\lambda + y^*_\lambda = \infty$, then for $M = c_0/\min\{c_0, c_1\}$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x^*_{\lambda_k} + y^*_{\lambda_k} > M$. In this case

$$
V_{\lambda_k}(x^*_{\lambda_k}, y^*_{\lambda_k}) > \min\{c_0, c_1\} M = V_{\lambda_k}(1, 0).
$$

We thus get a contradiction.

As $\sigma_{y\lambda/\mu}/(\lambda/\mu) \to a$ uniformly u.o.c. as $\lambda \to \infty$,

$$
V_\lambda(x, y) \to \hat{V}(x, y) \text{ u.o.c. as } \lambda \to \infty.
$$

Then we have

$$
V_\lambda(x^*, y^*) \to \hat{V}(x^*, y^*) \text{ as } \lambda \to \infty.
$$

$\square$
Let \( \hat{m} = x^* \lambda / \mu \) and \( \hat{n} = y^* \lambda / \mu \). As \( \Pi_\lambda(\hat{m}, \hat{n}) > \Pi_\lambda^* \), and

\[
\Pi_\lambda(\hat{m}, \hat{n}) \leq \Pi_\lambda(\hat{m}, \hat{n}) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2}
\]

\[
\leq \Pi_\lambda(\hat{m}, \hat{n}) + o(\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.6}
\]

\[
\leq \Pi_\lambda^* + o(\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2}
\]

\[
= \Pi_\lambda^* + o(\lambda),
\]

We have \( \Pi_\lambda(\hat{m}, \hat{n}) = \Pi_\lambda^* + o(\lambda) \).

We next solve the optimization problem

\[
\min_{x, y} \hat{V}(x, y)
\]

where

\[
\hat{V}(x, y) = c_0 x + c_1 y + (h + r \theta) \frac{\mu}{\theta} ay \int_{-1}^{1-x-y/ay} F(u) du
\]

We first notice that for fixed \( y \), we have

\[
\frac{\partial \hat{V}(x, y)}{\partial x} = c_0 - (h + r \theta) \frac{\mu}{\theta} F\left(\frac{1-x-y}{ay}\right)
\]

\[
\frac{\partial^2 \hat{V}(x, y)}{\partial x^2} = (h + r \theta) \frac{\mu}{\theta} \frac{1}{ay} F\left(\frac{1-x-y}{ay}\right) \geq 0
\]

Let \( \alpha_0 = \frac{c_0 \theta}{(h + r \theta) \mu} \). Then we divide the analysis into two cases:

- **Case a**: If \( 1 - y - ayF^{-1}(\alpha_0) \geq 0 \), \( x^*(y) = 1 - y - ayF^{-1}(\alpha_0) \).
- **Case b**: If \( 1 - y - ayF^{-1}(\alpha) < 0 \), \( x^*(y) = 0 \).

In Case a, we have \( \frac{1-x^*(y)-y}{ay} = F^{-1}(\alpha_0) \). Then we find \( y \) that minimizes

\[
\Xi_1(y) := \hat{V}(x^*(y), y) = c_0 + (c_1 - c_0)y - c_0 ayF^{-1}(\alpha_0) + (h + r \theta) \frac{\mu}{\theta} ay \int_{-1}^{F^{-1}(\alpha_0)} F(u) du.
\]

If

\[
c_1 > c_0 + c_0 aF^{-1}(\alpha_0) - a(h + r \theta) \frac{\mu}{\theta} \int_{-1}^{F^{-1}(\alpha_0)} F(u) du,
\]

then \( x^* = 1, y^* = 0 \); Otherwise, \( x^* = 0, y^* = (1 + aF^{-1}(\alpha_0))^{-1} \).

In Case b, we find \( y \) that minimizes

\[
\Xi_2(y) := \hat{V}(0, y) = c_1 y + (h + r \theta) \frac{\mu}{\theta} ay \int_{-1}^{1/(ay) - 1/a} F(u) du.
\]

We notice that

\[
\Xi_2'(y) = c_1 + (h + r \theta) \frac{\mu}{\theta} a \int_{-1}^{1/(ay) - 1/a} F(x) dx - (h + r \theta) \frac{\mu}{\theta} \frac{1}{ay} F\left(\frac{1}{ay} - \frac{1}{a}\right)
\]
and
\[ \Xi''_2(y) = (h + \theta) \frac{\mu}{\theta} \frac{1}{a y^3} f \left( \frac{1}{a y} - \frac{1}{a} \right) > 0. \]

We also notice that for \( y = \mu/\lambda \) (i.e. \( n = 1 \)),
\[ \Xi_2(\mu/\lambda) - \Xi_1(0) = c_1 + (h + r\theta) \frac{\mu}{\theta} a \int_{-1}^{\lambda/(a\mu) - 1/a} F(u) du - (h + r\theta) \frac{\lambda}{\theta} > 0, \]
If
\[ c_1 + (h + r\theta) \mu a \int_{-1}^{1/(a\mu) - 1/a} F(u) du - (h + r\theta) \frac{\mu}{\theta} F \left( \frac{1}{a y} - \frac{1}{a} \right) = 0. \quad (EC.1) \]
In particular, we notice that if
\[ c_1 + (h + r\theta) \mu a \int_{-1}^{0} F(u) du - (h + r\theta) \frac{\mu}{\theta} F(0) < 0, \]
then \( y^* < 1 \); Otherwise, \( y^* \) is the solution of
\[ c_1 + (h + r\theta) \mu a \int_{-1}^{1/(a\mu) - 1/a} F(u) du - (h + r\theta) \frac{\mu}{\theta} F \left( \frac{1}{a y} - \frac{1}{a} \right) = 0. \quad (EC.1) \]

**EC.4. Asymptotically optimal solutions to the stochastic fluid model with nonstationary demand**

We denote \((m^*_\lambda, n^*_\lambda, \alpha^*_\lambda)\) as an optimal solution to \( \min_{m,n,\alpha} \Pi_\lambda(m, n, \alpha) \). We also write \((\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda)\) as an optimal solution to \( \min_{m,n,\alpha} \tilde{\Pi}_\lambda(m, n, \alpha) \).

**EC.4.1. Case I: \( \sigma_n = \Omega(n^q) \) for \( 0 \leq q \leq 1/2 \)**

By Lemma EC.2 and Lemma EC.3, we have:
\[ \Pi_\lambda(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda) \leq \Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda) \leq \tilde{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda) + O(\sqrt{\lambda}) \]
\[ \leq \Pi_\lambda(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda) + O(\sqrt{\lambda}) \]
\[ \leq \Pi_\lambda(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda) + O(\sqrt{\lambda}) \]
**EC.4.2. Case II: σ_n = Ω(n^q) for 1/2 < q < 1**

We would consider adding some refinement to the fluid optimal solution. Let

\[
\bar{\Gamma}_H(m,n) = c_0m + c_1n + \frac{h + r\theta}{\theta}E[(\lambda_H/\mu - m - N(n))^+]
\]

\[
\bar{\Gamma}_L(m,n,\alpha) = c_0m + c_1\alpha n + \frac{h + r\theta}{\theta}E[(\lambda_L/\mu - m - N(\alpha n))^+]
\]

Then \(\bar{\Pi}_\lambda(m,n,\alpha) = T_H\bar{\Gamma}_H(m,n,\alpha) + T_L\bar{\Gamma}_L(m,n,\alpha)\).

**When \(c_0 < c_1\) and \((T_H + T_L)c_0 - T_Hc_1 \leq 0\).** We first notice from Lemma EC.3, that

\[
\tilde{\Pi}_\lambda(m,n,\alpha) \leq \bar{\Pi}_\lambda(m,n,\alpha),
\]

and for \(n = 0\),

\[
\tilde{\Pi}_\lambda(m,0,\alpha) = \bar{\Pi}_\lambda(m,0,\alpha).
\]

We next show that \(\bar{n}_\lambda = 0\). We prove this by contradiction. Assume that \(\bar{n}_\lambda > 0\), then for any \(m_\lambda, \alpha_\lambda:\)

\[
\bar{\Pi}_\lambda(m_\lambda + \bar{n}_\lambda, 0, \alpha_\lambda) = \bar{\Pi}_\lambda(m_\lambda + \bar{n}_\lambda, 0, \alpha_\lambda)
\]

\[
< \tilde{\Pi}_\lambda(m_\lambda, \bar{n}_\lambda, \alpha_\lambda) \text{ as } c_0 < c_1
\]

\[
\leq \tilde{\Pi}_\lambda(m_\lambda, \bar{n}_\lambda, \alpha_\lambda) \text{ by Lemma EC.3 }
\]

We thus get a contradiction. Then we have \((m_\lambda, \bar{n}_\lambda, \alpha_\lambda) = (m_\lambda, \bar{n}_\lambda, \alpha_\lambda)\). From the solution to the fluid optimization problem, we can set \(m_\lambda = \lambda_L/\mu, \bar{n}_\lambda = 0\), and \(\alpha_\lambda = 0\), which are also the optimal solutions to the stochastic fluid optimization problem.

**When \(c_0 < c_1\) and \((T_H + T_L)c_0 - T_Hc_1 > 0\).** We first show that \(m_\lambda \geq \lambda_L/\mu\) and \(\alpha_\lambda = 0\). We prove this by contradiction. Suppose that \(m_\lambda < \lambda_L/\mu\) or \(\alpha_\lambda > 0\). Then set \(m' = \lambda_L/\mu, n' = (\bar{n}_\lambda - (\lambda_L/\mu - \bar{m}_\lambda))^+\) and \(\alpha' = 0\). We notice that \(\bar{m}_\lambda + \bar{n}_\lambda = m' + n'\) and \(\bar{n}_\lambda > n'\). For fixed value of \(\lambda\), \(\mu\), and \(s, \sigma[(\lambda/\mu - s - \sigma_\lambda \epsilon)^+\) is increasing in \(\lambda\). Thus, \(\bar{\Gamma}_H(m,\bar{n}) \geq \bar{\Gamma}_H(m',n')\). We also notice that \(\bar{\Gamma}_L(m,n,\alpha)\) is minimized at \(m' = \lambda_L/\mu, \alpha' = 0\), i.e. \(\bar{\Gamma}_L(m_\lambda, \bar{n}_\lambda, \alpha_\lambda) \geq \bar{\Gamma}_L(m, \bar{n}, \alpha)\) for fixed value of \(\bar{n}_\lambda\). Thus, \(\bar{\Pi}_\lambda(m_\lambda, \bar{n}_\lambda, \alpha_\lambda) \geq \bar{\Pi}_\lambda(m',n',\alpha')\), and we get a contradiction.

Now let \(m = \lambda_L/\mu + x\sigma_{\lambda H/\mu - \lambda_L/\mu}\) and \(n = (\lambda_H/\mu - \lambda_L/\mu) + y\sigma_{\lambda H/\mu - \lambda_L/\mu}\). Then we set

\[
C_{2,\lambda}(x,y) = \bar{\Pi}_\lambda(m,n) - c_0\lambda_L/\mu(T_H + T_L) - c_1(\lambda_H/\mu - \lambda_L/\mu)T_L
\]

\[
= c_0(T_H + T_L)x + c_1T_Hy
\]

\[
+ (h + r\theta)\frac{\mu}{\theta}T_HE\left[\left(-x - y - \frac{\sigma(\lambda_H/\mu - \lambda_L/\mu) + y\sigma_{\lambda H/\mu - \lambda_L/\mu} \epsilon}{\sigma_{\lambda H/\mu - \lambda_L/\mu}}\right)^+ight]
\]

Let \(x_\lambda, y_\lambda\) denote the optimal solution to \(\min_{x \geq 0, y \in \mathbb{R}} C_{2,\lambda}(x,y)\). We also set

\[
\hat{C}_2(x,y) := c_0(T_H + T_L)x + c_1T_Hy + (h + r\theta)\frac{\mu}{\theta}T_HE\left[\left(-x - y - \epsilon\right)^\ast\right]
\]
and denote $x^*$, $y^*$ as the optimal solution to $\min_{x,y} \hat{C}_2(x,y)$. For $c_0 < c_1$ and $c_0(T_H + T_L) > c_1 T_H$, we have $x^* = 0$ and $y^*$ solves $\hat{H}_2(y) = c_1 y + (h + r\theta) \frac{\mu}{\theta} E[(-y - \epsilon)^+]$.

**Lemma EC.7.**

$C_{2,\lambda}(x^*_\lambda, y^*_\lambda) \rightarrow \hat{C}_2(x^*, y^*)$ as $\lambda \rightarrow \infty$.

The proof Lemma EC.7 follows exactly the same line of analysis as Lemma EC.5. We shall omit it here.

Now for $\hat{m} = \lambda_L / \mu$, $\hat{n} = (\lambda_H / \mu - \lambda_L / \mu) + y^* \sigma_{\lambda_H / \mu} - \lambda_L / \mu$, and $\hat{\alpha} = 0$, we have

$$\Pi_\lambda(\hat{m}, \hat{n}, \hat{\alpha}) \leq \Pi_\lambda(\hat{m}, \hat{n}, \hat{\alpha}) + O(\sqrt{\lambda})$$

$$\leq \Pi_\lambda(\hat{m}, \hat{n}, \hat{\alpha}) + o(\lambda) + O(\sqrt{\lambda})$$

$$\leq \Pi_\lambda(\lambda) + o(\lambda).$$

**When** $c_0 > c_1$, let $m = x \sigma_{\lambda_H / \mu}$, $n = \lambda_H / \mu + y \sigma_{\lambda_H / \mu}$, and $\alpha = \lambda_L / \lambda_H + z \sigma_{\lambda_H / \mu} / (\lambda_H / \mu)$. Then

$$\Pi_\lambda(m, n, \alpha) - c_1 T_H \lambda_H / \mu - c_1 T_L \lambda_L / \mu$$

$$= T_H \left( K_0 x + c_1 y + (h + r\theta) \frac{\mu}{\theta} E \left[ (-x - y - \frac{\sigma_{\lambda_H / \mu} + y \sigma_{\lambda_H / \mu}}{\sigma_{\lambda_H / \mu}}) \epsilon \right] \right)$$

$$+ T_L \left( K_0 x + c_1 \left( \frac{\xi_L}{\xi_H} y + z \right) + c_1 y z \sigma_{\lambda_H / \mu} / \lambda_H / \mu + (h + r\theta) \frac{\mu}{\theta} \right.$$

$$\times \left. E \left[ (-x - \frac{\xi_L}{\xi_H} y - z - y z \sigma_{\lambda_H / \mu} / \lambda_H / \mu) \frac{\sigma_{\xi_L / \xi_H} (\lambda_H / \mu - y \sigma_{\lambda_H / \mu}^2 / (\lambda_H / \mu) \epsilon) \epsilon}{\sigma_{\lambda_H / \mu}} \right] \right)$$

$$= C_{3,\lambda}(x, y, z)$$

Let $x^*_\lambda$, $y^*_\lambda$, $z^*_\lambda$ denote the optimal solution to $\min_{x \geq 0, y \in \mathbb{R}, z \in \mathbb{R}} C_{3,\lambda}(x, y, z)$. We also define

$$\hat{C}_3(x, y, z) := T_H \left( K_0 x + c_1 y + (h + r\theta) \frac{\mu}{\theta} E \left[ (-x - y - \epsilon)^+ \right] \right)$$

$$+ T_L \left( K_0 x + c_1 \left( \frac{\xi_L}{\xi_H} y + z \right) + (h + r\theta) \frac{\mu}{\theta} E \left[ (-x - \frac{\xi_L}{\xi_H} y - z - \left( \frac{\xi_L}{\xi_H} \right)^q \epsilon \right] \right)$$

Let $x^*$, $y^*$, $z^*$ denote the optimal solution to $\min_{x,y,z} \hat{C}_3(x, y, z)$. For $c_0 > c_1$, we have $x^* = 0$, and $y^*$ and $z^*$ solves

$$\min_{y, z} \left( T_H + T_L \left( \frac{\xi_L}{\xi_H} \right) \right) y + c_1 T_L z + (h + r\theta) \frac{\mu}{\theta} T_H E \left[ (-y - \epsilon)^+ \right] + (h + r\theta) \frac{\mu}{\theta} T_L E \left[ \frac{\xi_L}{\xi_H} y - z - \left( \frac{\xi_L}{\xi_H} \right)^q \epsilon \right]$$

**Lemma EC.8.**

$C_{3,\lambda}(x^*_\lambda, y^*_\lambda, z^*_\lambda) \rightarrow \hat{C}_3(x^*, y^*, z^*)$ as $\lambda \rightarrow \infty$. 


The proof Lemma EC.7 follows exactly the same line of analysis as Lemma EC.5. We shall omit it here.

Now let $\hat{m} = 0$, $\hat{n} = \lambda_H/\mu + y^*\sigma_{\lambda_H/\mu}$, and $\hat{\alpha} = \frac{\xi_H}{\xi_H} + z^*\sigma_{\lambda_H/\mu}/(\lambda_H/\mu)$. Then we have

$$\Pi'(\hat{m}, \hat{n}, \hat{\alpha}) \leq \Pi'(\hat{m}, \hat{n}, \hat{\alpha}) + O(\sqrt{\lambda})$$

$$\leq \Pi'(\hat{m}, \hat{n}, \hat{\alpha}) + o(\sigma_{\lambda}) + O(\sqrt{\lambda})$$

$$\leq \Pi' + o(\sigma_{\lambda}).$$

**EC.4.3. Case III: $\sigma_n = an$**

In this case,

$$\Gamma_H(m, n) = c_0 m + c_1 n + (h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$$

$$\Gamma_L(m, n, \alpha) = c_0 m + c_1 an + (h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_L/\mu - m - an)} F(x)dx$$

We first look at $\Gamma_L(m, n, \alpha)$ alone. Let $\beta_0 = c_0\theta/((h + r\theta)\mu)$. From the analysis of the single period, we know when $c_1 \geq c_0 + c_0 aF^{-1} (\beta_0) - a(h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$, we will use the fixed capacity only. When $c_1 < c_0 + c_0 aF^{-1} (\beta_0) - a(h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$, we will use the flexible capacity only.

When $c_1 \geq c_0 + c_0 aF^{-1} (\beta_0) - a(h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$, we set $m = \lambda_L/\mu + x \lambda_L - \lambda_L/\mu$, $n = \lambda_H - \lambda_L/\mu$, and $\alpha = 0$ for $x, y \geq 0$. Then

$$\frac{\Pi'(m, n, 0) - c_0 (T_H + T_L)\lambda_L/\mu - c_1 T_H (\lambda_H - \lambda_L)/\mu}{(\lambda_H - \lambda_L)/\mu}$$

$$= c_0 (T_H + T_L)x + c_1 T_H y + (h + r\theta)\frac{\mu}{\theta} T_E [(1 - x - y - ay\epsilon)^+]$$

Let $\beta_1 = c_0 (T_H + T_L)\theta/((h + r\theta)T_H\mu) = \beta_0 (T_H + T_L)/T_H$. Then again from the analysis of the single period case, we have

If $c_1 \geq c_0 + c_0 aF^{-1} (\beta_1) - a(h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$, we shall set $x^* = 1$ and $y^* = 0$. In this case, $\hat{n}_\lambda = \lambda_H/\mu$ and $\hat{n}_\lambda = 0$ and $\hat{\alpha}_\lambda = 0$.

If $c_1 < c_0 + c_0 aF^{-1} (\beta_1) - a(h + r\theta)\frac{\mu}{\theta} \int_{-1}^{(\lambda_H/\mu - m - an)} F(x)dx$, we shall set $x^* = 0$ and $y^*$ solves

$$\min_y c_1 y + (h + r\theta)\frac{\mu}{\theta} T_E [(1 - y - ay\epsilon)^+]$$

or equivalently, $y^*$ solves

$$c_1 + (h + r\theta)\frac{\mu}{\theta} a \int_{-1}^{1/(ay)-1/a} F(x)dx - (h + r\theta)\frac{\mu}{\theta} \frac{1}{y} F\left(\frac{1}{ay} - \frac{1}{a}\right) = 0.$$ 

In this case, $\hat{m}_\lambda = \lambda_L/\mu$, $\hat{n}_\lambda = y^*(\lambda_H - \lambda_L)/\mu$, and $\hat{\alpha}_\lambda = 0$. 

When \( c_1 < c_0 + c_0 a F^{-1}(\beta_0) - a (h + r \theta) \frac{\mu}{\theta} \int_{-1}^{F^{-1}(\beta_0)} F(x) dx \), we shall only use the flexible capacity. In this case,

\[
\tilde{\Gamma}_H(0, n, \alpha) = c_1 n + (h + r \theta) \frac{\mu}{\theta} E[(\lambda_H / \mu - n - an \epsilon)^+] 
\]

and

\[
\tilde{\Gamma}_L(0, n, \alpha) = c_1 \alpha n + (h + r \theta) \frac{\mu}{\theta} E[(\lambda_L / \mu - \alpha n - a \alpha n \epsilon)^+] 
\]

Thus, if \( \tilde{n}_\lambda \) maximizes \( \tilde{\Gamma}_H(0, n, \alpha) \), then \( (\tilde{n}, \xi_L / \xi_H) \) maximizes \( \tilde{\Gamma}_L(0, n, \xi_L / \xi_H) \). Therefore, in this case, we set \( \tilde{n}_\lambda = 0, \tilde{n}_\lambda = \gamma \lambda_H / \mu \), where \( \gamma \) that solves

\[
c_1 + (h + r \theta) \frac{\mu}{\theta} a \int_{-1}^{(1/(\alpha y)) - 1/a} F(x) dx - (h + r \theta) \frac{\mu}{\theta} \frac{1}{\gamma y} F \left( \frac{1}{\alpha y} - \frac{1}{a} \right) = 0,
\]

and \( \tilde{\alpha}_\lambda = \xi_L / \xi_H \).

**EC.5. Impact on variance**

Let \( X^{N^\lambda} \) denote the number in system in an \( M/M/N^\lambda + M \) queue.

**Proposition EC.1.** Assume that \( \mu = \theta \). Assume that \( Y^\lambda \equiv \frac{N^\lambda - \lambda}{\sigma^\lambda} \Rightarrow Y \) where \( Y \) is a proper real-valued random variable. We assume that \( E[N^\lambda] = \lambda \).

(a) If \( \sigma^\lambda < \sqrt{\lambda} \) or \( \sigma^\lambda > \sqrt{\lambda} \), then

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \mathcal{O}(\max\{\lambda, \sigma^2_\lambda\}) = \mathcal{O}(\lambda) + \mathcal{O}(\sigma^2_\lambda).
\]

(b) If \( \sigma^\lambda = \sqrt{\lambda} \) and \( P(Y = y) = 1 \) for some \( y \in \mathbb{R} \) then

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \mathcal{O}(\lambda).
\]

Assume WLOG that \( \mu = 1 \). Note that \( X^{N^\lambda} \sim \text{Pois}(\lambda) \).

**Proof of (a).** If \( \sigma^\lambda \not= \mathcal{O}(\sqrt{\lambda}) \):

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \max\{\lambda, \sigma^2_\lambda\} \text{Var} \left[ \left( \frac{\text{Pois}(\lambda) - \lambda}{\max\{\sqrt{\lambda}, \sigma_\lambda\}} - \frac{N^\lambda - \lambda}{\max\{\sqrt{\lambda}, \sigma_\lambda\}} \right)^+ \right]
\]

\[
= \max\{\lambda, \sigma^2_\lambda\} \text{Var} \left[ (Z^\lambda - U^\lambda)^+ \right],
\]

where \( Z^\lambda \equiv \frac{\text{Pois}(\lambda) - \lambda}{\max\{\sqrt{\lambda}, \sigma_\lambda\}} \) and \( U^\lambda \equiv \frac{N^\lambda - \lambda}{\max\{\sqrt{\lambda}, \sigma_\lambda\}} \). If \( \sigma^\lambda < \sqrt{\lambda} \) then \( U^\lambda \Rightarrow 0 \) and \( Z^\lambda \Rightarrow Z \sim \text{Nor}(0,1) \). If \( \sigma^\lambda > \sqrt{\lambda} \) then \( U^\lambda \Rightarrow Y \) and \( Z^\lambda \Rightarrow 0 \). Thus, by the continuous-mapping theorem (CMT) and the fact that convergence in distribution to a constant is equivalent to convergence in probability, we obtain that there exists a constant \( k > 0 \) such that

\[
\frac{\text{Var}[(X^{N^\lambda} - N^\lambda)^+]}{\max\{\lambda, \sigma^2_\lambda\}} \to k \text{ as } \lambda \to \infty,
\]

and \( \text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \mathcal{O}(\lambda) + \mathcal{O}(\sigma^2_\lambda) \).
Proof of (b).

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \lambda \text{Var} \left( \frac{\text{Pois} \left( \lambda \right) - \lambda - \frac{N^\lambda - \lambda}{\sqrt{\lambda}}}{\sqrt{\lambda}} \right)^+
\]

= \lambda \text{Var} \left( (Z^\lambda - Y^\lambda)^+ \right),
\]

where \( Z^\lambda \equiv \frac{\text{Pois} \left( \lambda \right) - \lambda}{\sqrt{\lambda}} \). Since \( Y^\lambda \Rightarrow y \), we also have: \( Y^\lambda \overset{p}{\Rightarrow} y \). Thus, by the CMT: \( (Z^\lambda - Y^\lambda)^+ \Rightarrow (Z - y)^+ \) where \( Z^\lambda \Rightarrow Z \). Thus, there exists a constant \( k > 0 \) such that

\[
\frac{\text{Var}[(X^{N^\lambda} - N^\lambda)^+]}{\lambda} \rightarrow k \quad \text{as} \quad \lambda \rightarrow \infty,
\]

and \( \text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \mathcal{O}(\lambda) \).

Note that in an \( M/M/n \) queue, we have that \( \text{Var}[(X^\lambda - n^\lambda)^+] = \mathcal{O}(\lambda) \) if \( n^\lambda = \lambda + \mathcal{O}(\sqrt{\lambda}) \) or \( n^\lambda = \delta \lambda \) where \( \delta < 1 \), and have that \( \text{Var}[(X^\lambda - n^\lambda)^+] = \mathcal{O}(\lambda) \) if \( n^\lambda = \gamma \lambda \) where \( \gamma > 1 \).

**Proposition EC.2.** Assume that \( \mu \neq \theta \). Assume that \( Y^\lambda \equiv \frac{N^\lambda - \lambda}{\sigma^\lambda} \Rightarrow Y \) where \( Y \) is a proper real-valued random variable. Assume that \( \mathbb{E}[N^\lambda] = \lambda \). Then,

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = \mathcal{O}(\lambda) + \mathcal{O}(\sigma^\lambda).
\]

Consider a lower-bound system with \( \mu_L = \theta_L = \max\{\mu, \theta\} \equiv M_0 \) and an upper-bound system with \( \mu_H = \theta_H = \min\{\mu, \theta\} \equiv m_0 \). We assume that \( \mu > \theta \) and let \( \theta = 1 \) without loss of generality. Then, \( M_0 = \mu \) and \( m_0 = \theta = 1 \).

Let \( X^\lambda_L \) and \( X^\lambda_H \) denote the respective number in systems, where \( X^\lambda_L \sim \text{Poiss} \left( \frac{\lambda}{\mu} \right) \) and \( X^\lambda_H \sim \text{Poiss} \left( \frac{\lambda}{\theta} \right) \). Let \( Q_i \equiv (X^\lambda_i - N^\lambda)^+ \). Then, we have:

\[
Q_L \leq_{st} Q \leq_{st} Q_H.
\]

This implies (dropping indexing by \( \lambda \)):

\[
\mathbb{E}[Q^2_L] = \text{Var}[Q_L] + (\mathbb{E}[Q_L])^2 \leq \text{Var}[Q] + (\mathbb{E}[Q])^2 \leq \text{Var}[Q_H] + (\mathbb{E}[Q_H])^2.
\]

Note also that: \( \mathbb{E}[Q_L] \leq \mathbb{E}[Q] \leq \mathbb{E}[Q_H] \). We have that \( \mathbb{E}[Q_L] = \left( \frac{\lambda}{\mu} - \lambda \right)^+ + \mathcal{O}(\sigma^\lambda) + \mathcal{O}(\sqrt{\lambda}) \) and \( \mathbb{E}[Q_H] = \left( \frac{\lambda}{\theta} - \lambda \right)^+ + \mathcal{O}(\sqrt{\lambda}) \). Making use of Proposition EC.1, we can write:

\[
\text{Var}[Q] \geq \text{Var}[Q_L] + (\mathbb{E}[Q_L])^2 - (\mathbb{E}[Q])^2
\]

\[
\geq \text{Var}[Q_L] + (\mathbb{E}[Q_L])^2 - (\mathbb{E}[Q_H])^2
\]

\[
= \text{Var}[Q_L] + \left( \left( \frac{\lambda}{\mu} - \lambda \right)^+ + \mathcal{O}(\sigma^\lambda) + \mathcal{O}(\sqrt{\lambda}) \right)^2 - (\mathcal{O}(\sigma^\lambda) + \mathcal{O}(\sqrt{\lambda}))^2
\]

\[
= \mathcal{O}(\lambda) + \mathcal{O}(\sigma^2^\lambda) \quad \text{for} \quad \lambda \text{ large enough}.
\]
We can also write that:

\[
\text{Var}[Q] \leq \text{Var}[Q_H] + (\mathbb{E}[Q_H])^2 - (\mathbb{E}[Q])^2 \\
\leq \text{Var}[Q_H] + (\mathbb{E}[Q_H])^2 - (\mathbb{E}[Q_L])^2 \\
= \text{Var}[Q_H] + \left( O(\sigma_\lambda) + O(\sqrt{\lambda}) \right)^2 - \left( \left( \frac{\lambda}{\mu} - \lambda \right) + O(\sigma_\lambda) + O(\sqrt{\lambda}) \right)^2 \\
= O(\lambda) + O(\sigma_\lambda^2) \text{ for } \lambda \text{ large enough.}
\]

Thus,

\[
\text{Var}[(X^{N^\lambda} - N^\lambda)^+] = O(\lambda) + O(\sigma_\lambda^2).
\]

For \( \mu < \theta \), we repeat the argument above, reversing the roles of \( \theta \) and \( \mu \).