Refined square-root staffing for call centers with impatient customers

Bo Zhang[*], Johan S.H. van Leeuwaarden[†], Bert Zwart[‡]

December 18, 2009

Abstract

In call centers it is crucial to staff the right number of agents so that the targeted service levels are met. These staffing problems typically lead to constraint satisfaction problems that are hard to solve. During the last decade, a beautiful asymptotic theory has been developed to solve such problems for large call centers operating in the quality-and-efficiency-driven (QED) regime. In this asymptotic regime, optimal staffing rules are known to obey the square-root staffing principle. This paper presents refinements to this principle that take into account the effect of impatient customers and work well for small systems.

1 Introduction

A key challenge in managing call centers is to balance the trade-off between operational costs and quality-of-service offered to customers. Most operational costs involve staffing costs, which makes it essential to develop adequate models of call center operations that relate operational performance to staffing levels; see Garnett et al. (2002), Gans et al. (2003), and Borst et al. (2004) for background.

Due to recent theoretical studies, backed up by assessments of empirical data, it is by now widely accepted that the phenomenon of impatient customers (the fact that waiting customers may abandon the system before receiving service) is one of the driving factors for call center performance (see Garnett et al. (2002) for a thorough discussion). Among different queueing models for call centers with impatient customers, the simplest, yet widely used one is the completely Markovian $M/M/s+M$ model, also referred to as the Erlang A model. Its performance analysis has been an important subject of study in the literature (see for example Garnett et al. (2002) and Whitt (2006b)), not only because the Erlang A model is worthy of being used in practice (see Mandelbaum and Zeltyn (2007)), but also because it delivers valuable approximations for more general abandonment models (see Whitt (2005a,b)).

There is by now a vast literature on the asymptotic analysis of call center models, which has proven to provide useful managerial insights. In these asymptotic studies, a finite-size queueing system is perceived as one in a sequence of queues and then the limiting behavior of this sequence is used to approximate the performance of this finite-size system. Depending on how this sequence is parameterized, its limiting behavior is different, giving rise to different approximations (see Borst et al. (2004) and Mandelbaum and Zeltyn (2008)). More specifically, queues with abandonments have been analyzed through fluid approximations (see for example Whitt (2005b, 2006a), Kang and Ramanan (2008), and Zhang (2009)) and diffusion approximations (see, e.g., Dai et al. (2009) and Mandelbaum and Momcilovic (2009)).

[*] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive NW, Atlanta, Georgia 30332-0205, USA. Email address: bozhang@gatech.edu

[†] Eindhoven University of Technology and EURANDOM, P.O. Box 513 - 5600 MB Eindhoven, The Netherlands. Email address: j.s.h.v.leeuwaarden@tue.nl

[‡] CWI, PNA2, Science Park 123, Amsterdam, North-Holland 1098XG, The Netherlands. Email address: bert.zwart@cwi.nl

One of the most effective approximations arises in the Quality-and-Efficiency-Driven (QED) regime, in which the number of servers s and the offered workload R are related according to a square-root principle, namely $s = R + \beta\sqrt{R}$, for a constant β. Square-root staffing and the QED limiting regime for multi-server queues without abandonments were brought to the center of attention by the work of Halfin and Whitt (1981). Garnett et al. (2002) study the steady-state performance approximation (as well as a process-level approximation) for the Erlang A model in the QED regime, and Zeltyn and Mandelbaum (2005) extend the asymptotic steady-state performance analysis to the $M/M/s + G$ model in the QED regime (as well as in other regimes).

Based on the QED diffusion approximations developed by Halfin and Whitt (1981), Borst et al. (2004) provide a rigorous justification, in an asymptotic framework, of applying the square-root staffing principle to two classes of problems: constraint satisfaction and cost minimization. They observe that square-root staffing is accurate over a wide range of system parameters for the Erlang C (or $M/M/s$) model without abandonments. Mandelbaum and Zeltyn (2008) apply the results in Zeltyn and Mandelbaum (2005) to the constraint satisfaction problem for the $M/M/s + G$ model, and find that square-root staffing is not as robust as in models without abandonments. In particular, for the Erlang A model, they observe from numerical experiments that square-root staffing is far from optimal for satisfying loose constraints on the tail of the waiting time distribution, and recommend staffing based on a new type of limiting approximation referred to as ED+QED (cf. (35)).

Therefore, for queueing models with abandonments, it is of great interest to understand why the inaccuracy of square-root staffing arises, and to develop performance approximations and staffing rules that are accurate in all circumstances. One approach towards accomplishing this goal, which is taken in the present paper, is to explicitly characterize, and subsequently correct, the errors of conventional QED diffusion approximation and square-root staffing. Correcting the error of the diffusion approximation, thus obtaining what is known as corrected diffusion approximation, has previously been studied by Blanchet and Glynn (2006) and Siegmund (1979) in the random walk or $GI/G/1$ queue setting and by Janssen et al. (2008a,b) for the Erlang B and C models. Yet, the explicit characterization of the error of a staffing prescription is more challenging, because both the approximative staffing level and the exact (optimal) one are typically defined implicitly, i.e., characterized as a solution to some equation. The only study in this regard is the work by Janssen et al. (2008b), which develops refined square-root staffing rules for the Erlang C model. The present paper extends this approach to the Erlang A model. In comparison to the Erlang C model, the Erlang A model brings about additional mathematical challenges. This increased level of technicality gives in return valuable insights for a much more realistic model for call centers than the Erlang C model (see Mandelbaum and Zeltyn (2007)). Our main results are captured in Theorems 2, 4, and 6, which formally establish the staffing refinements as a characterization of the optimality gap of conventional square-root staffing; we believe that the implication of these results holds in more generality (e.g., for cost minimization, capacity allocation among multi-class customers, or staffing multi-skill call centers): a refinement of performance approximation by an order of $\sqrt{R}$ can be used to yield a refinement of staffing prescription by an order of $\sqrt{R}$. Also, the findings in this paper are completely different: unlike in the Erlang C model, the refinements are significant in many cases, due to different system parameters. This makes the refined staffing rules particularly relevant for practical purposes.

Another motivation for this study is to assess analytically the accuracy of square-root staffing and its underlying QED approximations in the presence of abandonments. Although the development of accurate and usable performance approximations has been the primary motivation for a large body of research over recent decades, there has been little work or success on the analytical assessment of the accuracy (or equivalently the error) of various asymptotic performance approximations. Most approximations are justified by proving limit theorems, while the assessment of their accuracy is usually performed empirically or via simulation. The explicit characterization of the error that we develop allows us to perform this task analytically. One related study in this regard is the work by Bassamboo and Randhawa (2009), which investigates the error in the fluid approximations of the steady-state expected queue-length and abandonment probability and shows that the fluid approximations are $\mathcal{O}(1)$ accurate in the overloaded regime under some regularity conditions.

In short, this paper makes the following contributions. First, for a useful call center model taking abandonments into account, namely the Erlang A model, we develop corrected diffusion approximations for several main steady-state performance measures that are of independent interest. Second, we apply the corrected approximations to develop refined square-root staffing rules for several constraint satisfaction problems with respect to these performance measures. The refined staffing rules are as easy to implement as the conventional square-root staffing principle, and yet the error of the refined rules is smaller, as is shown both analytically and numerically. Also, the explicit form of the refinement yields important insights into the appropriateness of the conventional square-root staffing for call centers with different demand volumes and staffing objectives, and enables us to provide practical recommendations on when to use the refined or the conventional square-root staffing rules.

The remainder of this paper is organized as follows. Section 2 provides a technical overview of the asymptotic dimensioning framework and our refined staffing approach, as well as a discussion on the influence of abandonments. In Sections 3, 4, and 5, based on corrected diffusion approximations, we develop the refined square-root staffing rules for three constraint satisfaction problems. Section 6 contains concluding remarks.

2 The Erlang A model and refined staffing

Let us first introduce the Erlang A model, also referred to as the $M/M/s+M$ queue. Customers arrive according to a Poisson process with rate λ and require service times that are independent and exponentially distributed with mean $1/\mu$. There are s homogeneous servers working in parallel, and there is unlimited waiting space. Customers that are waiting in the queue abandon the system after an exponentially distributed time with mean $1/\theta$. Without loss of generality, we assume $\mu = 1$ throughout this paper. Therefore, the traffic intensity is $\rho = \lambda/s$. Let W denote the steady-state waiting time of a customer before receiving service or abandoning the system. We denote $P\{W > 0\}$ by $A(s, \lambda, \theta)$, and henceforth refer to $A(s, \lambda, \theta)$ as the Erlang A formula, naturally generalizing the Erlang B and C formulas through $\lim_{\theta\to\infty} A(s, \lambda, \theta) = B(s, \lambda)$ and $\lim_{\theta\downarrow 0} A(s, \lambda, \theta) = C(s, \lambda)$. Let $P\{\text{Ab}\}$ denote the steady-state probability that a customer abandons the system.

2.1 Asymptotic dimensioning

The core of staffing problems in call centers is to determine the right trade-off between quality and capacity. Quality is formulated in terms of some targeted service level. Take as an example the delay probability $A(s, \lambda, \theta)$. A large delay probability is perceived as negative, and the targeted service level could be to keep the delay probability below some value ϵ. The smaller ϵ, the higher the target, and the better the offered service. Once the targeted service level is set, the objective from the call center's perspective is to determine the lowest staffing level s such that the target $A(s, \lambda, \theta) \le \epsilon$ is met. This is what we have referred to as a constraint satisfaction problem.

For simplicity, we assume throughout that staffing levels can take on non-integer values. The delay probability is a function of the three model parameters s, λ and θ, and the analytic extension of $A(s, \lambda, \theta)$ to all positive real s is a continuous and monotone decreasing function in s. Therefore, the constraint satisfaction problem is equivalent to finding the s_{opt} such that $A(s_{\text{opt}}, \lambda, \theta) = \epsilon$. To solve this inverse problem, we shall invoke the theory of asymptotic dimensioning introduced in Borst et al. (2004) and extended in Mandelbaum and Zeltyn (2008) to abandonments. This theory fully exploits the QED regime for large call centers, in a way that reduces considerably the complexity of the inverse problem. That is, under square-root staffing $s = \lambda + \beta\sqrt{\lambda}$ with β some fixed constant, and in the QED regime (when $s \to \infty$), the performance measures in the Erlang A model can be approximated by their diffusion limit counterparts. For instance, $A(s, \lambda, \theta)$ can be approximated by some function $A_*(\beta)$ that only depends on β and θ (and no longer on s or λ). Hence, the inverse problem can then be approximatively solved by searching for the β_* such that $A_*(\beta_*) = \epsilon$, and then setting the staffing level according to $s_* = \lambda + \beta_*\sqrt{\lambda}$. In this asymptotic approach, it should be intuitively obvious that the better the approximation $A(s,\lambda,\theta) \approx A_*(\beta)$, the smaller the error $|s_{\mathrm{opt}} - s_*|$. Based on the QED regime, one expects the approximation s_* to be accurate for large values of λ, and in particular for large-scale service systems such as call centers.

2.2 Refined staffing

Mandelbaum and Zeltyn (2008) show that any staffing rule of the form $\lambda + \beta_*\sqrt{\lambda} + o(\sqrt{\lambda})$ is asymptotically optimal under the $M/M/s+G$ model assumption, where a function $f(\lambda) = o(g(\lambda))$ if $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) = 0$. The main technical contribution of this paper is to develop a stronger form of optimality by characterizing the $o(\sqrt{\lambda})$ small order term. Specifically, we shall develop refined staffing rules for the Erlang A model. These refined staffing rules should be capable of dealing with the effects of abandonments, thus extending the work of Janssen et al. (2008b). Our approach consists of first developing corrected diffusion approximations for the objective functions, and then characterizing the approximative solutions to the constraint satisfaction problems. The refined staffing rules are of the form

$$s_\bullet = \lambda + \beta_*\sqrt{\lambda} + \beta_\bullet, \tag{1}$$

with $\beta_\bullet$ some function of β_*, θ, λ, and the constraint target level ϵ that depends on the staffing problem under consideration. For three different constraint satisfaction problems, we shall uniquely identify $\beta_\bullet$, and prove that the refined staffing level in (1) yields

$$s_{\mathrm{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}), \tag{2}$$

where a function $f(\lambda) = \mathcal{O}(g(\lambda))$ if $\limsup_{\lambda\to\infty} |f(\lambda)/g(\lambda)| < \infty$. We refer to the order term that expresses the difference between the exact optimal staffing level and the approximate staffing level as the *optimality gap*. Hence, the optimality gap of $s_\bullet$ is $\mathcal{O}(\lambda^{-1/2})$, which suggests that the staffing level $s_\bullet$ not only becomes accurate in the QED regime ($\lambda \to \infty$), but is also more accurate in situations away from the limit. Note that $s_\bullet = s_* + \beta_\bullet$. We shall prove that the optimality gap of the conventional staffing level s_* equals $\mathcal{O}(1)$, which indicates that $s_\bullet$ is a clear improvement. In addition, because $\beta_\bullet$ in fact describes the optimality gap of s_*, it allows us to perform an analytical assessment of the accuracy of conventional square-root staffing and its underlying QED approximations, and to make some practical recommendations for call center staffing.

2.3 The influence of abandonments

We consider three different constraint satisfaction problems: (i) zero delay constraint $P\{W > 0\} \leq \epsilon$, (ii) excess delay constraint $P\{W > T\} \leq \epsilon$ with $T > 0$, and (iii) abandonment constraint $P\{\mathrm{Ab}\} \leq \epsilon$. In each problem, we search for the lowest staffing level such that the constraint is met. Clearly, all three performance measures decrease as a function of s, and higher staffing levels are required when ϵ becomes smaller.

The influence of abandonments on the accuracy of conventional square-root staffing or the magnitude of $\beta_\bullet$ is less obvious. By deriving and examining its explicit expression, we find that for the first two problems, due to the presence of customer abandonments, $\beta_\bullet$ is significant if ϵ, λ, and/or θ are large. This is in stark contrast to the fact that in the absence of abandonments, as reported in Janssen et al. (2008b), $\beta_\bullet$ is mostly negligible and only becomes slightly larger than one if ϵ is extremely small. Another intriguing observation is that $\beta_\bullet$ is especially significant if the staffing problem leads to an overloaded system, i.e., $\beta_* < 0$ and hence $s_* < \lambda$. For the third problem (which is not applicable without abandonments), $\beta_\bullet$ shows a clear insensitivity to θ and λ.

3 Zero delay constraint

The objective of the zero delay constraint satisfaction problem is to determine the number of servers that are required to ensure that $A(s,\lambda,\theta) = P\{W > 0\}$ is below a threshold ϵ. The conventional square-root staffing rule is to use the approximation $A(s,\lambda,\theta) \approx A_*(\beta)$, obtain the solution to $A_*(\beta) = \epsilon$, say β_*, and then prescribe the staffing level as $s_* = \lambda + \beta_*\sqrt{\lambda}$. Now, according to our scheme for refined staffing described in Section 2, we shall first derive a corrected diffusion approximation for the objective function, and then solve the asymptotic inverse problem. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution function and density function, respectively.

Theorem 1 (Refined approximation for delay probability). *Let $A_{\lambda,\theta}(\beta) = A(s,\lambda,\theta)$ with $\beta = (s-\lambda)\lambda^{-1/2}$ assumed fixed. Then,*

$$A_{\lambda,\theta}(\beta) = A_*(\beta) + A_\bullet(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{3}$$

where

$$A_*(\beta) = \Big(1 + \sqrt{\theta}G(\beta)H_\theta(\beta)\Big)^{-1}, \tag{4}$$

$$A_\bullet(\beta) = A_*(\beta)^2\left(\frac{1}{3}\sqrt{\theta}H_\theta(\beta)A_*(\beta)^{-1} - h_\theta(\beta)\right), \tag{5}$$

$$h_\theta(\beta) = -\frac{1}{6}\sqrt{\theta}\beta^2 H_\theta(\beta)\Big(G(\beta)H_\theta(\beta)\theta^{-1/2} - \beta G(\beta)\theta^{-1} + 1 + \beta G(\beta)\Big), \tag{6}$$

$$G(\beta) = \frac{\Phi(\beta)}{\phi(\beta)}, \quad H_\theta(\beta) = \frac{\phi(\beta/\sqrt{\theta})}{\Phi(-\beta/\sqrt{\theta})}. \tag{7}$$

Our proof of Theorem 1 is based on the following relation between the Erlang A and Erlang B formulas (e.g., equation (A.1) in Mandelbaum and Zeltyn (2007))

$$A(s,\lambda,\theta)^{-1} = 1 + \frac{B(s,\lambda)^{-1} - 1}{(s/\theta)e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta,\lambda/\theta)}, \tag{8}$$

with γ the incomplete gamma function (cf. (45)) and $B(s,\lambda)$ the Erlang B formula, or the blocking probability in the corresponding $M/M/s/s$ queue. First, a power series approximation in terms of $s^{-1/2}$ is derived for the denominator of the second term in (8), which involves the incomplete gamma function. Then, we combine this result with an approximation of $B(s,\lambda)^{-1}$ developed in Janssen et al. (2008b) to obtain a series approximation of $A(s,\lambda,\theta)^{-1}$ with respect to $s^{-1/2}$. Finally, we derive the desired power series expansion of the Erlang A formula in $\lambda^{-1/2}$ using the square-root relation between λ and s. We include the full proof in Section A.

Relation (8) can be further exploited to derive a set of upper and lower bounds for the Erlang A formula. Specifically, the incomplete gamma function term $\gamma(\cdot)$ in (8) can be expressed in terms of the (complete) gamma function $\Gamma(\cdot)$, for which sharp bounds are derived in Spira (1971). This, combined with the bounds for the Erlang B formula developed in Janssen et al. (2008a), immediately yields bounds for $A(s,\lambda,\theta)$.

The corrected diffusion approximation for the delay probability is thus given by the two terms on the right-hand side of (3), where we ignore the order term. If the second term is also ignored, we retrieve the conventional first-order diffusion approximation $A_{\lambda,\theta}(\beta) \approx A_*(\beta)$ that was derived in Garnett et al. (2002). An additional check follows from the case without abandonments. Indeed, by letting $\theta \to 0$ in (3) and using $H_\theta(\beta) \sim \beta/\sqrt{\theta}$, we retrieve Theorem 2 of Janssen et al. (2008b).

Despite the complicated expression of the corrected diffusion approximation, its computation is as easy as the conventional approximation, because the additional computation of the higher-order term only involves simple algebraic operations on quantities which are already required for evaluating the first-order diffusion approximation (e.g., $G(\beta)$ and $H_\theta(\beta)$).

We shall now use the corrected diffusion approximation to derive a refined staffing level.

Theorem 2 (Refined staffing level for zero delay constraint). *Let $s_{\text{opt}} \in (0,\infty)$ be the solution to $A(s,\lambda,\theta) = \epsilon$. Let β_* be the solution to $A_*(\beta) = \epsilon$, $s_* = \lambda + \beta_*\sqrt{\lambda}$, and $s_\bullet = s_* + \beta_\bullet$ with*

$$\beta_\bullet = \frac{\beta_*^2}{6}\left(1 - \frac{\sqrt{\theta}H_\theta(\beta_*)}{3h_\theta(\beta_*)\epsilon}\right). \tag{9}$$

Then,

$$s_{\text{opt}} - s_* = \mathcal{O}(1), \tag{10}$$

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{11}$$

Proof. Proof. Define β_λ as the solution to

$$A_*(\beta_\lambda) + A_\bullet(\beta_\lambda)\lambda^{-1/2} = \epsilon. \tag{12}$$

Let $g(\lambda) := \beta_\lambda - \beta_*$, and then (12) can be rewritten as

$$A_*(\beta_* + g(\lambda)) + A_\bullet(\beta_* + g(\lambda))\lambda^{-1/2} = \epsilon. \tag{13}$$

A first-order Taylor expansion of (13) yields

$$A_*(\beta_*) + \mathcal{O}(g(\lambda)) + A_\bullet(\beta_*)\lambda^{-1/2} + \mathcal{O}(g(\lambda)\lambda^{-1/2}) = \epsilon. \tag{14}$$

Because $A_*(\beta_*) = \epsilon$, it immediately follows that

$$g(\lambda) = \mathcal{O}(\lambda^{-1/2}). \tag{15}$$

Then, we apply a second-order Taylor expansion to (13) to have

$$A_*(\beta_*) + A'_*(\beta_*)g(\lambda) + \mathcal{O}(g(\lambda)^2) + A_\bullet(\beta_*)\lambda^{-1/2} + \mathcal{O}(g(\lambda)\lambda^{-1/2}) = \epsilon. \tag{16}$$

Using (15) and $A_*(\beta_*) = \epsilon$, we solve (16) and obtain that

$$g(\lambda) = -\frac{A_\bullet(\beta_*)}{A'_*(\beta_*)}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{17}$$

Therefore, β_λ is well approximated by $\beta_* + \beta_\bullet\lambda^{-1/2}$, up to $\mathcal{O}(\lambda^{-1})$, where

$$\beta_\bullet = -\frac{A_\bullet(\beta_*)}{A'_*(\beta_*)}. \tag{18}$$

By using (4), (5), and $A_*(\beta_*) = \epsilon$, (18) can be further simplified as (9).

We next turn to proving the optimality gap results in (10) and (11). Let $\beta_{\text{opt}} = (s_{\text{opt}} - \lambda)\lambda^{-1/2}$. The desired result is equivalent to

$$\beta_{\text{opt}} - \beta_* = \mathcal{O}(\lambda^{-1/2}), \tag{19}$$

$$\beta_{\text{opt}} - \left(\beta_* + \beta_\bullet\lambda^{-1/2}\right) = \mathcal{O}(\lambda^{-1}). \tag{20}$$

From Theorem 1, we have that

$$\epsilon = A_{\lambda,\theta}(\beta_{\text{opt}}) = A_*(\beta_{\text{opt}}) + \mathcal{O}(\lambda^{-1/2}). \tag{21}$$

Let $g_*(\lambda) := \beta_{\text{opt}} - \beta_*$. Then applying a first-order Taylor expansion to (21), we obtain that

$$\epsilon = A_*(\beta_*) + \mathcal{O}(g_*(\lambda)) + \mathcal{O}(\lambda^{-1/2}). \tag{22}$$

Since $A_*(\beta_*) = \epsilon$, $g_*(\lambda) = \mathcal{O}(\lambda^{-1/2})$ or (19) holds. Because the derivation of $\beta_\bullet$ implies that

$$\beta_\lambda - \left(\beta_* + \beta_\bullet\lambda^{-1/2}\right) = \mathcal{O}(\lambda^{-1}), \tag{23}$$

in order to conclude (20), it suffices to prove that

$$\beta_{\text{opt}} - \beta_\lambda = \mathcal{O}(\lambda^{-1}). \tag{24}$$

Let $g_\bullet(\lambda) := \beta_{\text{opt}} - \beta_\lambda$. The rest of the proof is similar as above:

$$\epsilon = A_{\lambda,\theta}(\beta_{\text{opt}}) = A_*(\beta_{\text{opt}}) + A_\bullet(\beta_{\text{opt}})\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}) \tag{25}$$

$$= A_*(\beta_\lambda) + \mathcal{O}(g_\bullet(\lambda)) + A_\bullet(\beta_\lambda)\lambda^{-1/2} + \mathcal{O}(g_\bullet(\lambda)\lambda^{-1/2}) + \mathcal{O}(\lambda^{-1}). \tag{26}$$

Since $A_*(\beta_\lambda) + A_\bullet(\beta_\lambda)\lambda^{-1/2} = \epsilon$, we find that $g_\bullet(\lambda) = \mathcal{O}(\lambda^{-1})$, which proves the assertion in (24). □

For the zero delay constraint satisfaction problem, we recommend the refined staffing level $s_\bullet = s_* + \beta_\bullet$, with $\beta_\bullet$ defined in (9). Note that $\beta_\bullet$ is just a simple function of β_*, θ, and ϵ. Since the classical staffing scheme already requires solving for β_*, which is the hardest task, adapting the refined scheme using $\beta_\bullet$ requires hardly any additional computation. Therefore, we claim that obtaining $s_\bullet$ is as easy as s_*, while $s_\bullet$ achieves a stronger asymptotic optimality than s_*. One interpretation of Theorem 2 is that $\beta_\bullet$, as defined by (9), exactly captures the dominating term of the error of s_*, or the $\mathcal{O}(1)$ term in (10). By adding the refinement $\beta_\bullet$, the optimality gap of $s_\bullet$ decreases at the rate of $\lambda^{-1/2}$. We remark that it is proved in Mandelbaum and Zeltyn (2008) that $s_{\text{opt}} - s_* = o(\sqrt{\lambda})$, whereas our refined staffing approach enables us to show that the $o(\sqrt{\lambda})$ gap is actually $\mathcal{O}(1)$.

3.1 Numerical experiments

In our extensive numerical experiments, $|s_{\text{opt}} - s_\bullet|$ is almost always less than 1. As an indication of the error made by the conventional square-root staffing, $\beta_\bullet$ becomes more significant as the abandonment rate θ increases. Also, with the increase of θ, $\beta_\bullet$ gradually becomes a monotone increasing function of the targeted delay probability.

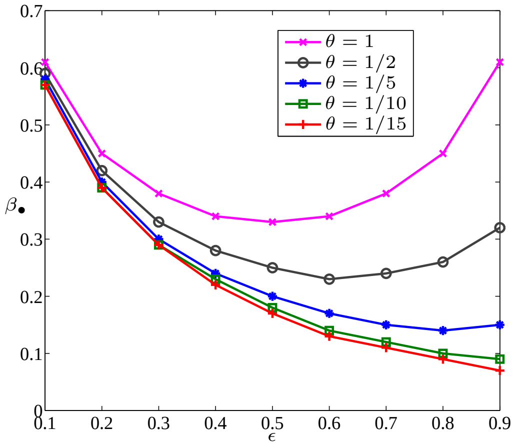

Figure 1: The refinement $\beta_\bullet$ as a function of ϵ, with $\theta \leq 1$.

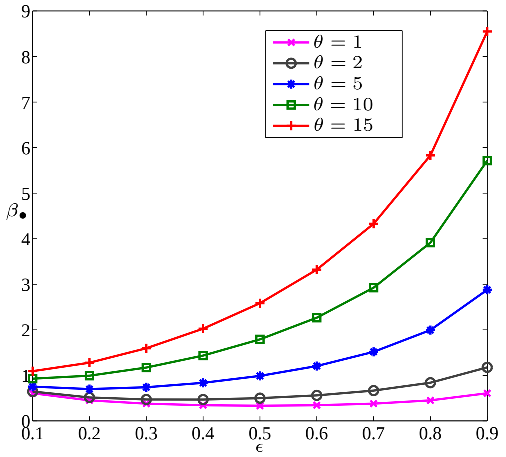

Figure 2: The refinement $\beta_\bullet$ as a function of ϵ, with $\theta \geq 1$.

Figure 1 shows that, when $\theta \leq 1$, $\beta_\bullet$ is always less than 1 and its curve gradually turns to symmetrically bowl-shaped from monotone decreasing in ϵ, as θ increases to 1. In Figure 2, as θ further increases from 1 to 15, $\beta_\bullet$ becomes more significant. In particular, when $\theta \geq 5$, $\beta_\bullet$ is always larger under a looser delay constraint (i.e., a greater ϵ value). For example, as ϵ increases from 0.1 to 0.9, $\beta_\bullet$ increases from about 1 to 6, for $\theta = 10$, and from 1 to nearly 9, for θ=15. Because $\beta_\bullet$ does not depend on λ, such errors are rather severe for a small or moderate size system. For instance, Tables 1 and 2 display the case of $\lambda = 30$, in which the rather large errors are almost completely corrected by $\beta_\bullet$.

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$
0.1	35.6364	0.8568	34.6932	0.9432	0.9267	35.6199	0.0165
0.2	32.2059	0.2161	31.1838	1.0222	0.9927	32.1764	0.0295
0.3	29.5538	-0.3028	28.3416	1.2123	1.1717	29.5132	0.0406
0.4	27.1519	-0.7918	25.6630	1.4889	1.4348	27.0978	0.0541
0.5	24.7924	-1.2909	22.9292	1.8632	1.7898	24.7190	0.0734
0.6	22.3326	-1.8324	19.9637	2.3690	2.2654	22.2291	0.1036
0.7	19.6159	-2.4580	16.5368	3.0791	2.9241	19.4609	0.1550
0.8	16.3821	-3.2471	12.2151	4.1669	3.9130	16.1281	0.2540
0.9	11.9658	-4.4276	5.7491	6.2167	5.7145	11.4636	0.5022

Table 1: $P\{W > 0\} = \epsilon, \theta = 10, \lambda = 30$

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$
0.1	35.1431	0.7333	34.0162	1.1269	1.0923	35.1085	0.0346
0.2	31.5051	0.0320	30.1751	1.3299	1.2773	31.4524	0.0527
0.3	28.6506	-0.5502	26.9865	1.6642	1.5944	28.5809	0.0697
0.4	26.0387	-1.1095	23.9231	2.1156	2.0238	25.9469	0.0918
0.5	23.4556	-1.6893	20.7473	2.7083	2.5836	23.3309	0.1247
0.6	20.7546	-2.3265	17.2570	3.4976	3.3201	20.5771	0.1776
0.7	17.7769	-3.0710	13.1796	4.5973	4.3282	17.5078	0.2691
0.8	14.2667	-4.0185	7.9900	6.2767	5.8296	13.8196	0.4471
0.9	9.6101	-5.4473	0.1639	9.4462	8.5488	8.7128	0.8973

Table 2: $P\{W > 0\} = \epsilon, \theta = 15, \lambda = 30$

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$
0.1	2996.8250	-0.1231	2993.2590	3.5659	3.5069	2996.7660	0.0591
0.2	2933.3450	-1.3225	2927.5630	5.7820	5.6995	2933.2620	0.0824
0.3	2874.1970	-2.4526	2865.6640	8.5331	8.4198	2874.0840	0.1133
0.4	2812.8280	-3.6347	2800.9180	11.9103	11.7517	2812.6700	0.1586
0.5	2745.7460	-4.9359	2729.6470	16.0990	15.8728	2745.5200	0.2263
0.6	2669.3000	-6.4292	2647.8580	21.4421	21.1122	2668.9700	0.3299
0.7	2577.8430	-8.2299	2549.2310	28.6124	28.1160	2577.3470	0.4964
0.8	2459.8590	-10.5766	2420.6950	39.1638	38.3702	2459.0650	0.7936
0.9	2281.4960	-14.1803	2223.3110	58.1854	56.7158	2280.0270	1.4696

Table 3: $P\{W > 0\} = \epsilon, \theta = 100, \lambda = 3000$

For large systems, if the customer patience level is low, $\beta_\bullet$ can be quite substantial. For example, Table 3 shows that, when $\theta = 100$, s_* can be off by as many as 20 to 60 servers, while $s_\bullet$ provides an extremely accurate approximation of s_{opt}. We note that $\beta_\bullet$ tends to be significant when $\beta_* < 0$, as illustrated in Tables 1, 2, and 3. For a number of other cases, especially when $\beta_* > 0$, the refinement $|\beta_\bullet|$ turns out to be less than one, which provides theoretical support for the adequacy of square-root staffing or QED approximation in those parameter regions. Therefore, we recommend that the refined square-root staffing rule should be adopted for any small to moderate size call center and any large size call center with impatient customers, especially if it operates under a moderate or loose zero delay constraint. In other cases, the conventional staffing rule can be followed without running the risk of substantial inaccuracies.

4 Excess delay constraint

We now turn to the constraint satisfaction problem in which the objective function is the steady-state probability that the delay exceeds a certain level T. Specifically, we want to determine the minimum number of servers required to meet the constraint $P\{W > T\} \leq \epsilon$. We start by deriving a corrected diffusion approximation for this performance measure.

Theorem 3 (Refined approximation for excess delay)**.** *Let $\beta = (s - \lambda)\lambda^{-1/2}$ assumed fixed.*

$$P\{W > t\lambda^{-1/2}\} = A_*(\beta)d_*(\beta, t) + \left[A_*(\beta)d_\bullet(\beta, t) + A_\bullet(\beta)d_*(\beta, t)\right]\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{27}$$

where

$$d_*(\beta, t) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\Phi(-\beta\theta^{-1/2})}, \tag{28}$$

$$d_\bullet(\beta, t) = d_*(\beta, t)\left(\frac{1}{6}I_\bullet(\beta, \theta/2, t)\frac{\theta^{5/2}\phi(\beta\theta^{-1/2})}{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})} - \frac{1}{6}I_\bullet(\beta, \theta/2, 0)\theta^{5/2}H_\theta(\beta) - \theta t\right), \tag{29}$$

$$I_\bullet(a, b, t) = \int_t^\infty \exp\{-ay - by^2\}y^3 dy, \quad \forall a > 0, b > 0, t \geq 0. \tag{30}$$

The main step in the proof of Theorem 3 is to show that

$$P\{W > t\lambda^{-1/2} | W > 0\} = d_*(\beta, t) + d_\bullet(\beta, t)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{31}$$

We prove (31) by deriving and combining corrected approximations for two integral-form building blocks of the exact expression for $P\{W > t\lambda^{-1/2} | W > 0\}$. In particular, we apply the Laplace method to analyze their asymptotic behavior and refine the results presented in Section 10 and Theorem 4.1(g) in Zeltyn and Mandelbaum (2005). The detailed proof is included in Section B.

The right-hand side of (27), excluding the order term, serves as the corrected diffusion approximation for $P\{W > t\lambda^{-1/2}\}$, while the conventional diffusion approximation is given by the first term only, i.e., $P\{W > t\lambda^{-1/2}\} \approx A_*(\beta)d_*(\beta, t)$. Again, the evaluation of the correction term only involves simple algebra on known quantities from the computation of the conventional diffusion approximation, and in particular $I_\bullet(a, b, t)$ can be calculated fast using (68), where it is expressed explicitly in terms of $\Phi(\cdot)$.

Now we first consider the constraint of the form $P\{W > t\lambda^{-1/2}\} \leq \epsilon$. Because the (corrected) diffusion approximations for $P\{W > t\lambda^{-1/2}\}$ in (27) and $P\{W > 0\}$ in (3) have exactly the same order in each corresponding term, the staffing procedure in Section 3 and, in particular, the expression (18) can be directly applied here with proper substitutions, leading to the following result:

Theorem 4 (Refined staffing level for excess delay constraint)**.** *Let $s_{\text{opt}} \in (0, \infty)$ be the solution to $P\{W > t\lambda^{-1/2}\} = \epsilon$, for some $t > 0$. Let β_* be the solution to $A_*(\beta)d_*(\beta, t) = \epsilon$, $s_* = \lambda + \beta_*\sqrt{\lambda}$, and $s_\bullet = s_* + \beta_\bullet$ with*

$$\beta_\bullet = -\frac{A_*(\beta_*)d_\bullet(\beta_*, t) + A_\bullet(\beta_*)d_*(\beta_*, t)}{A_*'(\beta_*)d_*(\beta_*, t) + A_*(\beta_*)d_*'(\beta_*, t)}, \tag{32}$$

where $d_*'(\cdot, \cdot)$ denotes the derivative of $d_*(\cdot, \cdot)$ with respect to the first argument. Then,

$$s_{\text{opt}} - s_* = \mathcal{O}(1), \tag{33}$$

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{34}$$

Proof. Proof. We follow the same procedure as for Theorem 2, by replacing $P\{W > 0\}$ with $P\{W > t\lambda^{-1/2}\}$, $A_*(\cdot)$ with $A_*(\cdot)d_*(\cdot)$, and $A_\bullet(\cdot)$ with $A_*(\cdot)d_\bullet(\cdot) + A_\bullet(\cdot)d_*(\cdot)$. We omit further details. □

For staffing in practice, when the constraint has the form $P\{W > T\} \leq \epsilon$, for a fixed T, we let $t = T\sqrt{\lambda}$. Then the constraint to satisfy becomes $P\{W > t\lambda^{-1/2}\} \leq \epsilon$, and the above staffing rule applies. In this case, $\beta_\bullet$ depends on θ, ϵ, λ, and T (through β_* and t).

4.1 Numerical experiments

In this subsection, we investigate numerically the gain of refined staffing. We also compare square-root staffing, both conventional and refined, with ED+QED staffing, which is a staffing principle developed for satisfying the excess delay constraint in Mandelbaum and Zeltyn (2008). Specifically, for the constraint $P\{W > T\} \leq \epsilon$, Theorem 4.4 in Mandelbaum and Zeltyn (2008) prescribes the staffing level

$$s_{\text{EQ}} = e^{-\theta T}\lambda + \delta^*\sqrt{\lambda}, \tag{35}$$

where

$$\delta^* = \Phi^{-1}(1 - \epsilon \cdot e^{\theta T})\sqrt{\theta e^{-\theta T}}. \tag{36}$$

Note that, if $\epsilon \geq e^{-\theta T}$, $s_{\text{opt}} = 0$. We do not consider such cases.

First, we focus on the constraints with small T values, which describes some of the key performance measures for call centers. For example, extremely small T and ϵ values may correspond to emergency call centers, such as 911 in the U.S., and $P\{W >$20 seconds$\} \leq \epsilon$, for some ϵ at the order of 10%, is the rule of thumb for many other types of call centers. Note that, in the following analysis, $T = 0.05$ is equivalent to 20 seconds if the average service time is 400 seconds.

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$	s_{EQ}	$s_{\text{opt}} - s_{\text{EQ}}$
0.001	47.001	2.845	45.585	1.417	1.501	47.086	-0.085	41.051	5.951
0.002	45.688	2.637	44.444	1.244	1.316	45.760	-0.072	40.238	5.451
0.003	44.890	2.510	43.745	1.144	1.209	44.954	-0.065	39.738	5.152
0.004	44.307	2.416	43.233	1.074	1.134	44.367	-0.060	39.371	4.937
0.005	43.846	2.342	42.826	1.020	1.076	43.902	-0.056	39.078	4.768
0.006	43.463	2.280	42.487	0.976	1.029	43.516	-0.053	38.834	4.629
0.007	43.134	2.226	42.194	0.939	0.990	43.184	-0.051	38.624	4.510
0.008	42.845	2.179	41.937	0.907	0.956	42.893	-0.049	38.438	4.407
0.009	42.587	2.137	41.708	0.880	0.926	42.634	-0.047	38.272	4.315
0.010	42.354	2.099	41.499	0.855	0.900	42.399	-0.045	38.121	4.233

Table 4: $P\{W > 0.05\} = \epsilon, \theta = 0.5, \lambda = 30, \epsilon = 0.001$ to 0.01

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$	s_{EQ}	$s_{\text{opt}} - s_{\text{EQ}}$
0.1	36.429	1.110	36.080	0.349	0.366	36.445	-0.017	34.106	2.322
0.2	34.118	0.712	33.898	0.219	0.230	34.129	-0.011	32.410	1.708
0.3	32.528	0.434	32.375	0.153	0.161	32.535	-0.008	31.182	1.346
0.4	31.219	0.202	31.108	0.111	0.117	31.224	-0.006	30.128	1.090
0.5	30.035	-0.009	29.953	0.082	0.086	30.039	-0.004	29.138	0.897
0.6	28.886	-0.214	28.826	0.060	0.062	28.888	-0.002	28.139	0.747
0.7	27.685	-0.429	27.648	0.037	0.037	27.685	-0.000	27.056	0.629
0.8	26.301	-0.675	26.303	-0.002	-0.003	26.300	0.001	25.754	0.547
0.9	24.336	-1.007	24.486	-0.150	-0.135	24.351	-0.015	23.812	0.523

Table 5: $P\{W > 0.05\} = \epsilon, \theta = 0.5, \lambda = 30, \epsilon = 0.1$ to 0.9

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$	s_{EQ}	$s_{\text{opt}} - s_{\text{EQ}}$
0.001	45.791	2.680	44.678	1.113	1.226	45.904	-0.113	54.599	-8.808
0.002	44.360	2.452	43.429	0.931	1.031	44.460	-0.099	52.459	-8.099
0.003	43.479	2.310	42.653	0.826	0.918	43.571	-0.092	51.141	-7.662
0.004	42.831	2.205	42.079	0.751	0.838	42.917	-0.086	50.173	-7.342
0.005	42.313	2.121	41.619	0.694	0.776	42.396	-0.082	49.400	-7.087
0.006	41.880	2.051	41.233	0.647	0.726	41.959	-0.079	48.755	-6.875
0.007	41.506	1.990	40.898	0.608	0.684	41.582	-0.076	48.198	-6.692
0.008	41.175	1.936	40.602	0.573	0.647	41.249	-0.074	47.707	-6.531
0.009	40.879	1.887	40.336	0.544	0.615	40.951	-0.072	47.267	-6.387
0.010	40.610	1.843	40.093	0.517	0.587	40.680	-0.070	46.867	-6.257

Table 6: $P\{W > 0.05\} = \epsilon, \theta = 4, \lambda = 30$

In this case, if the abandonment rate is low, the conventional square-root staffing is extremely accurate, regardless of the system size or the targeted service level. Tables 4 and 5 illustrate the cases for small λ values; similar findings hold for other λ and ϵ values. ED+QED staffing tends to prescribe staffing levels that are too low, especially under tight constraints, as shown in Table 4. This parameter region is of particular interest to the staffing of emergency call centers, having relatively patient customers and tight delay constraints.

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$	s_{EQ}	$s_{\text{opt}} - s_{\text{EQ}}$
0.05	909.683	-3.046	903.683	6.000	6.535	910.218	-0.535	907.195	2.488
0.10	887.412	-3.766	880.908	6.504	7.086	887.994	-0.582	885.363	2.049
0.15	872.193	-4.254	865.473	6.720	7.341	872.814	-0.621	870.418	1.775
0.20	859.959	-4.643	853.183	6.775	7.437	860.621	-0.662	858.366	1.592
0.25	849.332	-4.976	842.630	6.702	7.414	850.044	-0.713	847.863	1.468
0.30	839.652	-5.276	833.147	6.504	7.283	840.430	-0.779	838.265	1.387
0.35	830.530	-5.554	824.358	6.173	7.041	831.399	-0.869	829.190	1.340
0.40	821.696	-5.818	816.015	5.682	6.677	822.691	-0.995	820.372	1.324
0.45	812.932	-6.073	807.942	4.990	6.168	814.110	-1.178	811.593	1.339
0.50	804.026	-6.325	799.996	4.030	5.480	805.476	-1.450	802.642	1.384

Table 7: $P\{W > 0.05\} = \epsilon, \theta = 4, \lambda = 1000$

If the abandonment rate is high, the conventional square-root staffing is still very accurate for small systems (or small λ's), while ED+QED staffing tends to overstaff, especially under tight constraints (see Table 6). For large λ's, when the constraint can be satisfied with the system being overloaded, $\beta_\bullet$ becomes substantial and s_{EQ} also becomes more accurate than s_*. Table 7 shows such an example.

Next, we consider the constraints with moderate or large T values. As illustrated in Mandelbaum and Zeltyn (2008), s_* is accurate when the load is small, but not so when the load is moderate or large. In the latter case, the refinement significantly improves the accuracy. Table 8 displays the same example as considered in Section 5.3 of the online appendix of Mandelbaum and Zeltyn (2008). For $P\{W > \frac{1}{3}\}$, $\theta = 0.5$, and $\lambda = 1000$, s_* always underestimates s_{opt} by nearly 10 servers, while the difference between s_{opt} and $s_\bullet$ is less than 1.

The fact that s_*, as an asymptotic approximation, is less accurate for larger λ values might seem counterintuitive, but it can be easily explained with the aid of the explicit $\beta_\bullet$ expression. Again, we consider the above example, i.e., $T = \frac{1}{3}$ and $\theta = 0.5$. In Figure 3, with ϵ fixed at different values, we plot the $\beta_\bullet$, as a function of λ, calculated by (32). The plot clearly shows the growth of $\beta_\bullet$ with λ. It is interesting to note that the increase is approximately linear and that the five lines corresponding to different ϵ values do not differ much.

In summary, for the excess delay constraint satisfaction problem, we recommend that refined staffing should always be adopted. Also, the experimental results show that the accuracy im-

ϵ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$	s_{EQ}	$s_{\text{opt}} - s_{\text{EQ}}$
0.05	878.999	-4.107	870.113	8.885	9.409	879.523	-0.524	878.630	0.369
0.10	871.130	-4.364	861.990	9.140	9.681	871.671	-0.540	870.847	0.283
0.15	865.771	-4.538	856.509	9.263	9.816	866.325	-0.554	865.534	0.238
0.20	861.469	-4.675	852.153	9.317	9.884	862.037	-0.567	861.260	0.209
0.25	857.737	-4.794	848.415	9.322	9.905	858.320	-0.583	857.547	0.191
0.30	854.343	-4.900	845.059	9.283	9.886	854.945	-0.602	854.165	0.178
0.35	851.150	-4.998	841.949	9.200	9.828	851.778	-0.628	850.979	0.171
0.40	848.066	-5.091	838.998	9.067	9.730	848.729	-0.663	847.899	0.167
0.45	845.017	-5.182	836.143	8.874	9.586	845.729	-0.712	844.850	0.167
0.50	841.936	-5.270	833.333	8.602	9.385	842.718	-0.782	841.764	0.171

Table 8: $P\{W > \frac{1}{3}\} = \epsilon, \theta = 0.5, \lambda = 1000$

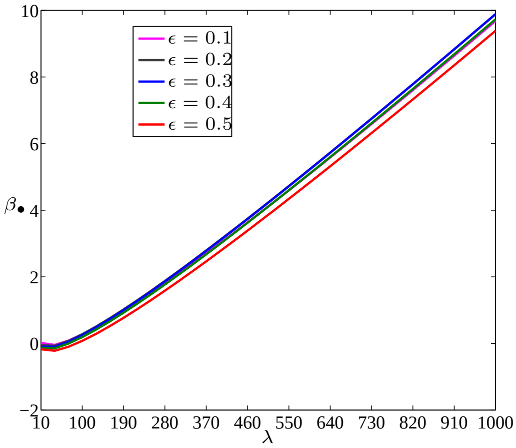

Figure 3: The refinement $\beta_\bullet$ as a function of λ, for $P\{W > \frac{1}{3}\} = \epsilon$ with $\theta = 0.5$. The five lines corresponding to different ϵ values are either indistinguishable or very close.

provement due to the refinement is especially significant if $\beta_* < 0$; this is the same as in Section 3.

5 Abandonment constraint

In this section, we develop the refined staffing rule for satisfying the constraint on the steady-state abandonment probability. Again, we start with a refined diffusion approximation.

Theorem 5 (Refined approximation for abandonment probability)**.** *Let* $\beta = (s-\lambda)\lambda^{-1/2}$ *assumed fixed.*

$$P\{\text{Ab}\} = b_*(\beta)\lambda^{-1/2} + b_\bullet(\beta)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}), \tag{37}$$

where

$$b_*(\beta) = (\sqrt{\theta}H_\theta(\beta) - \beta)A_*(\beta), \qquad b_\bullet(\beta) = u_\theta(\beta)b_*(\beta), \tag{38}$$

$$u_\theta(\beta) = -h_\theta(\beta)A_*(\beta) - \frac{1}{6}\beta^2 H_\theta(\beta)\theta^{-1/2} + \frac{1}{6}\beta H_\theta(\beta)\sqrt{\theta}\Big(\sqrt{\theta}H_\theta(\beta) - \beta\Big)^{-1}. \tag{39}$$

We prove Theorem 5 by first deriving a power series approximation of $P\{\text{Ab}|W > 0\}$ in terms of $s^{-1/2}$, then combining this with the refined approximation of $P\{W > 0\}$ to get the series expansion of $P\{\text{Ab}\}$ in terms of $s^{-1/2}$, and finally obtaining (37) by exploiting the square-root relation between λ and s. The full proof can be found in Section C.

We consider the constraint of the form $P\{\mathrm{Ab}\} \le \epsilon\lambda^{-1/2}$, and refined staffing again strengthens the asymptotic optimality:

Theorem 6 (Refined staffing level for abandonment constraint)**.** *Let $s_{\mathrm{opt}} \in (0,\infty)$ be the solution to $P\{\mathrm{Ab}\} = \epsilon\lambda^{-1/2}$. Let β_* be the solution to $b_*(\beta)\lambda^{-1/2} = \epsilon\lambda^{-1/2}$ or $b_*(\beta) = \epsilon$, $s_* = \lambda + \beta_*\sqrt{\lambda}$, and $s_\bullet = s_* + \beta_\bullet$ with*

$$\beta_\bullet = -\frac{b_\bullet(\beta_*)}{b_*'(\beta_*)}. \tag{40}$$

Then,

$$s_{\mathrm{opt}} - s_* = \mathcal{O}(1), \tag{41}$$

$$s_{\mathrm{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{42}$$

The proof of Theorem 6 is similar to Theorem 2 and is included in Section D. Furthermore, simple calculations show that

$$b_\bullet(\beta_*) = u_\theta(\beta_*)\epsilon\sqrt{\lambda} \tag{43}$$

and

$$b_*'(\beta_*) = \left(6A_*(\beta_*)h_\theta(\beta_*)\beta_*^{-2} - \beta_*\theta^{-1}\right)\epsilon\sqrt{\lambda} + \left(H_\theta(\beta_*)^2 - \beta_*^2\theta^{-1} - 1\right)A_*(\beta_*). \tag{44}$$

Therefore, one may use (43) and (44) to evaluate (40). In practice, the constraint of the form $P\{\mathrm{Ab}\} \le \epsilon$ can be first translated into $P\{\mathrm{Ab}\} \le \epsilon_\lambda\lambda^{-1/2}$, where $\epsilon_\lambda := \epsilon\sqrt{\lambda}$, and then one can apply (40), (43), and (44), in which ϵ is replaced by ϵ_λ and β_* is the solution to $b_*(\beta) = \epsilon_\lambda$ (see Remark 4.3 in Mandelbaum and Zeltyn (2008) on the scaling).

5.1 Numerical experiments

When the abandonment probability constraint becomes very tight (ϵ= 0.1% or even smaller), $\beta_\bullet$ becomes non-negligible and its magnitude is not sensitive to the abandonment rate or the offered load. For example, Tables 9 and 10 show that, for $\epsilon = 10^{-5}$, s_* is always off by a couple of servers, for a wide range of θ and λ values. For loose or moderate constraints, $|\beta_\bullet|$ is mostly less than 1. Again, in all cases, the refined square-root staffing rule yields an accurate approximation of s_{opt}. Therefore, we recommend that, for call centers with a tight abandonment constraint, the refined staffing procedure should be followed, regardless of the customer patience level, and s_* can be used otherwise.

λ	s_{opt}	β_*	s_*	$s_{\mathrm{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\mathrm{opt}} - s_\bullet$
1	7.0643	3.9236	4.9236	2.1407	2.7156	7.6392	-0.5749
2	9.6022	3.8434	7.4355	2.1668	2.6114	10.0468	-0.4446
5	15.5222	3.7354	13.3526	2.1696	2.4741	15.8267	-0.3045
10	23.6967	3.6519	21.5485	2.1482	2.3707	23.9191	-0.2225
20	38.0604	3.5669	35.9518	2.1086	2.2677	38.2195	-0.1591
50	76.4422	3.4520	74.4093	2.0329	2.1323	76.5416	-0.0994
100	135.5921	3.3630	133.6302	1.9620	2.0304	135.6605	-0.0684
200	248.1577	3.2722	246.2752	1.8825	1.9290	248.2042	-0.0465
500	572.1810	3.1490	570.4127	1.7683	1.7959	572.2086	-0.0276
1000	1098.2300	3.0533	1096.5520	1.6775	1.6959	1098.2480	-0.0184

Table 9: $P\{\mathrm{Ab}\} = 10^{-5}, \theta = 1$

Finally, note that, since $\theta \cdot E[W] = P\{\mathrm{Ab}\}$, the result in this section also holds for staffing with respect to the mean waiting time.

λ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\beta_\bullet$	$s_\bullet$	$s_{\text{opt}} - s_\bullet$
1	7.8970	4.4461	5.4461	2.4510	3.4560	8.9021	-1.0051
2	10.6991	4.3699	8.1800	2.5191	3.3441	11.5241	-0.8250
5	17.1268	4.2673	14.5419	2.5849	3.1961	17.7381	-0.6113
10	25.8574	4.1880	23.2437	2.6137	3.0843	26.3280	-0.4706
20	40.9903	4.1073	38.3684	2.6219	2.9726	41.3410	-0.3507
50	80.8694	3.9982	78.2715	2.5978	2.8250	81.0966	-0.2272
100	141.6912	3.9137	139.1373	2.5539	2.7135	141.8508	-0.1596
200	256.6201	3.8275	254.1288	2.4913	2.6021	256.7309	-0.1107
500	585.3574	3.7105	582.9701	2.3874	2.4549	585.4250	-0.0676
1000	1116.7620	3.6197	1114.4640	2.2974	2.3437	1116.8080	-0.0463

Table 10: $P\{\text{Ab}\} = 10^{-5}, \theta = 50$

6 Conclusions

The analytical assessment and numerical experiments in Sections 3 and 4 clearly suggest that the first-order diffusion approximations and conventional square-root staffing with respect to the tail probability of the customer delay are less accurate for overloaded systems. It is shown that significant $\beta_\bullet$ values arise when $\beta_* < 0$ (especially when β_* is relatively small or more negative), while $\beta_* > 0$ is typically associated with a small $\beta_\bullet$. In these two types of constraint satisfaction problems, $\beta_* < 0$ can be due to different system parameters, such as a large ϵ (i.e., a loose constraint), a large λ (due to economy of scale), and/or a large θ (more “contribution” from customer abandonment). In these cases, the refinement term (in either the approximation or staffing) significantly improves the accuracy, and such an improvement leads to the right staffing level in most cases of practical interest to call center staffing.

Although ED+QED staffing is more accurate than conventional square-root staffing when the system is more overloaded, refined square-root staffing is the most accurate in all cases (in particular, as accurate as ED+QED in the overloaded case) and thus overall the most reliable method, at least under our model assumptions.

As for staffing under the abandonment constraint (or equivalently the mean waiting time constraint), we observe in Section 5 that the refinement is significant when the constraint is tight, regardless of the customer patience level or the system size. In all our experiments, the refined square-root staffing rule yields satisfactory results.

A Proof of Theorem 1

We denote the incomplete gamma functions by

$$\gamma(s,a) = \int_0^a t^{s-1} e^{-t} dt, \quad \Gamma(s,a) = \int_a^\infty t^{s-1} e^{-t} dt, \tag{45}$$

and the gamma function by $\Gamma(s) = \gamma(s,a) + \Gamma(s,a)$. Using the relations

$$\gamma(s,\lambda) = \frac{\lambda^s e^{-\lambda}}{s} + \Gamma(s) \left(1 - \frac{\Gamma(s+1,\lambda)}{\Gamma(s+1)}\right) \tag{46}$$

and

$$B(s,\lambda) = \frac{e^{-\lambda}\lambda^s}{\Gamma(s+1,\lambda)} \tag{47}$$

yields

$$\frac{se^\lambda}{\lambda^s}\gamma(s,\lambda) = 1 + \frac{\Gamma(s+1)e^\lambda}{\lambda^s} - B(s,\lambda)^{-1}. \tag{48}$$

In Janssen et al. (2008b), it is shown that

$$B(s,\lambda)^{-1} = \frac{\Phi(\alpha)}{\phi(\alpha)}s^{1/2} + \frac{2}{3} + \mathcal{O}(s^{-1/2}), \tag{49}$$

where

$$\alpha = \sqrt{-2s(1-\rho+\ln\rho)}, \quad \operatorname{sign}(\alpha) = \operatorname{sign}(1-\rho), \tag{50}$$

a simple function of λ and s with $\alpha \to \beta$ as $s \to \infty$. By letting $p(s) := s^s e^{-s}\sqrt{2\pi s}\,\Gamma(s+1)^{-1}$, we rewrite the second term in (48) as

$$\frac{\Gamma(s+1)e^{\lambda}}{\lambda^s} = \frac{s^{1/2}}{\phi(\alpha)p(s)}. \tag{51}$$

Applying (49) and (51) to (48) yields

$$\frac{se^{\lambda}}{\lambda^s}\gamma(s,\lambda) = \frac{\Phi(-\alpha)}{\phi(\alpha)}s^{1/2} + \frac{1}{3} + \mathcal{O}(s^{-1/2}), \tag{52}$$

which, upon inversion, becomes

$$\frac{\lambda^s e^{-\lambda}}{s\gamma(s,\lambda)} = \frac{\phi(\alpha)}{\Phi(-\alpha)}s^{-1/2} - \frac{1}{3}\left(\frac{\phi(\alpha)}{\Phi(-\alpha)}\right)^2 s^{-1} + \mathcal{O}(s^{-3/2}). \tag{53}$$

We now restate (8) in the main paper, the relation between Erlang A and Erlang B formulas

$$A_{\lambda,\theta}(\beta)^{-1} = A(s,\lambda,\theta)^{-1} = 1 + \frac{B(s,\lambda)^{-1} - 1}{(s/\theta)e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta,\lambda/\theta)}. \tag{54}$$

Substituting (53) and (49) into (54) then yields

$$A_{\lambda,\theta}(\beta)^{-1} = A_*(\alpha)^{-1}\left(1 - \frac{1}{3}\sqrt{\theta}H_\theta(\alpha)s^{-1/2}\right) + \mathcal{O}(s^{-1}). \tag{55}$$

Simple computations show that

$$G(\alpha) = G(\beta) - \frac{1}{6}\beta^2\left(1+\beta G(\beta)\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{56}$$

$$\phi(\alpha)^{-1} = \phi(\beta)^{-1} - \frac{1}{6}\beta^3\phi(\beta)^{-1}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{57}$$

and $s^{-1/2} = \lambda^{-1/2} + \mathcal{O}(\lambda^{-1})$. Subtracting (56) from (57) yields

$$\frac{\Phi(-\alpha)}{\phi(\alpha)} = \frac{\Phi(-\beta)}{\phi(\beta)} + \frac{1}{6}\beta^2\left(1 - \frac{\beta\Phi(-\beta)}{\phi(\beta)}\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{58}$$

Inverting (58) gives

$$\frac{\phi(\alpha)}{\Phi(-\alpha)} = \frac{\phi(\beta)}{\Phi(-\beta)} - \frac{1}{6}\beta^2\left(\frac{\phi(\beta)^2}{\Phi(-\beta)^2} - \frac{\beta\phi(\beta)}{\Phi(-\beta)}\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{59}$$

Using (56) and (59) in (55), we arrive at

$$A_*(\alpha)^{-1} = A_*(\beta)^{-1} + h_\theta(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}) \tag{60}$$

and

$$1 - \frac{1}{3}\sqrt{\theta}H_\theta(\alpha)s^{-1/2} = 1 - \frac{1}{3}\sqrt{\theta}H_\theta(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{61}$$

Therefore, by multiplying (60) and (61), we obtain that

$$A_{\lambda,\theta}(\beta)^{-1} = A_*(\beta)^{-1} + \left(h_\theta(\beta) - \frac{1}{3}\sqrt{\theta}H_\theta(\beta)A_*(\beta)^{-1}\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{62}$$

Finally, inverting (62) yields (3).

B Proof of Theorem 3

We first show a technical lemma, which is needed in the later proof.

Lemma 1. *Let*

$$v_\lambda(x) = \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\}(1 + b_3\lambda x^3), \tag{63}$$

$$w_\lambda(x) = \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2 + b_3\lambda x^3\}, \tag{64}$$

where $b_i > 0$, $i = 1, 2, 3$, are constants. Let $t \geq 0$ and $\delta \in (t\lambda^{-1/2}, b_2/b_3)$ be a constant, and define

$$I(\lambda) = \int_{t\lambda^{-1/2}}^{\delta} w_\lambda(x)dx, \qquad I_A(\lambda) = \int_{t\lambda^{-1/2}}^{\infty} v_\lambda(x)dx. \tag{65}$$

Then,

$$I(\lambda) = I_A(\lambda) + \mathcal{O}(\lambda^{-3/2}), \tag{66}$$

$$I_A(\lambda) = \frac{\Phi\Big(-\sqrt{2b_2}t - \frac{1}{\sqrt{2}}b_1 b_2^{-1/2}\Big)}{\phi\Big(\frac{1}{\sqrt{2}}b_1 b_2^{-1/2}\Big)\sqrt{2b_2}}\lambda^{-1/2} + I_\bullet(b_1, b_2, t)b_3\lambda^{-1}, \tag{67}$$

where, $\forall a > 0, b > 0, t \geq 0$,

$$\begin{aligned} I_\bullet(a, b, t) &= \int_t^\infty \exp\{-ay - by^2\}y^3 dy \\ &= \frac{1}{16}b^{-7/2}e^{-t(a+bt)}\Big[2\sqrt{b}\left(a^2 - 2abt + 4b\left(1 + bt^2\right)\right) \\ &\quad + a\left(a^2 + 6b\right)e^{(a+2bt)^2/4b}\sqrt{\pi}\Big(\mathrm{Erf}\Big[\frac{1}{2}b^{-1/2}(a + 2bt)\Big] - 1\Big)\Big]. \end{aligned} \tag{68}$$

Proof. Proof. We have that

$$\begin{aligned} I(\lambda) &= \int_{t\lambda^{-1/2}}^{\delta} w_\lambda(x)dx \\ &= \lambda^{-1/2}\int_t^{\delta\sqrt{\lambda}} \exp\{-b_1 y - b_2 y^2 + b_3 y^3\lambda^{-1/2}\}dy \\ &= z\int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2 + b_3 y^3 z\}dy \end{aligned} \tag{69}$$

with $y = x\sqrt{\lambda}$ and $z = \lambda^{-1/2}$. By Taylor series expansion, for some $\xi \in (0, z)$,

$$\exp\{-b_1 y - b_2 y^2 + b_3 y^3 z\} = \exp\{-b_1 y - b_2 y^2\}\Big(1 + b_3 y^3 z + \frac{1}{2}b_3^2 y^6 e^{b_3 y^3 \xi} z^2\Big).$$

Therefore,

$$I(\lambda) = I_1(\lambda) + I_2(\lambda), \tag{70}$$

where

$$I_1(\lambda) = z\int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\}(1 + b_3 y^3 z)dy, \tag{71}$$

$$I_2(\lambda) = \frac{1}{2}z^3\int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\}b_3^2 y^6 e^{b_3 y^3 \xi}dy. \tag{72}$$

Let $I_{A_1}(\lambda) = \int_{t\lambda^{-1/2}}^{\delta} v_\lambda(x)dx$ and $I_{A_2}(\lambda) = \int_\delta^\infty v_\lambda(x)dx$, and then we have

$$I_A(\lambda) = I_{A_1}(\lambda) + I_{A_2}(\lambda). \tag{73}$$

It is easy to verify that

$$I_1(\lambda) = I_{A_1}(\lambda). \tag{74}$$

The fact that

$$I_2(\lambda) = \mathcal{O}(\lambda^{-3/2}) \tag{75}$$

follows from

$$\begin{aligned}
I_2(\lambda) &= \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\} b_3^2 y^6 e^{b_3 y^3 \xi} dy \\
&\leq \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\} b_3^2 y^6 e^{b_3 y^2 \frac{\delta}{z} z} dy \quad \text{(because } y \leq \delta/z \text{ and } \xi \leq z) \\
&= \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2 + b_3 y^2 \delta\} b_3^2 y^6 dy \\
&\leq \frac{1}{2}z^3 \int_0^{\infty} \exp\{-b_1 y - (b_2 - b_3\delta) y^2\} b_3^2 y^6 dy \\
&= C_0 z^3 = C_0 \lambda^{-3/2},
\end{aligned} \tag{76}$$

for some constant $C_0 > 0$, because we assume $\delta < b_2/b_3$ or $b_2 - b_3\delta > 0$.

Next we show that

$$I_{A_2}(\lambda) = o(e^{-\lambda^{\nu_2}}), \quad \text{for some } \nu_2 > 0. \tag{77}$$

For an arbitrarily chosen $C_1 \in (0,1)$, $\exists \lambda_{b_1,b_2,b_3,C_1} > 0$ such that, for any $\lambda > \lambda_{b_1,b_2,b_3,C_1}$, $v_\lambda(x) < \exp\{-b_2 C_1 \lambda x^2\}$. After integration, $I_{A_2}(\lambda) \leq \int_\delta^\infty \exp\{-b_2 C_1 \lambda x^2\} dx$, for any $\lambda > \lambda_{b_1,b_2,b_3,C_1}$. Then by Lemma 4.3 in the Internet supplement to Zeltyn and Mandelbaum (2005), we have $\int_\delta^\infty \exp\{-b_2 C_1 \lambda x^2\} dx = o(e^{-\lambda^{\nu_2}})$, for some $\nu_2 > 0$, and thus (77) follows.

Using (74), (75), and (77), we subtract (73) from (70) and arrive at

$$I(\lambda) - I_A(\lambda) = \mathcal{O}(\lambda^{-3/2}) + o(e^{-\lambda^{\nu_2}}) = \mathcal{O}(\lambda^{-3/2}). \tag{78}$$

Expression (67) follows from straightforward calculations. □

Define

$$u_\lambda(x) = \lambda\theta^{-1}(1 - e^{-\theta x}) - \lambda x - \beta\sqrt{\lambda}x, \tag{79}$$

$$J(y) = \int_y^\infty \exp\{u_\lambda(x)\} dx, \qquad \forall y \geq 0. \tag{80}$$

Next, we use Lemma 1 to derive the refined asymptotic expansions for $J(t\lambda^{-1/2})$ and $J(0)$, which are key components of the expression of $P\{W > t\lambda^{-1/2} | W > 0\}$.

Lemma 2.

$$J(t\lambda^{-1/2}) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}} \lambda^{-1/2} + \frac{1}{6} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \theta^2 \lambda^{-1} + \mathcal{O}(\lambda^{-\frac{3}{2}}), \tag{81}$$

$$J(0) = H_\theta(\beta)^{-1} \theta^{-1/2} \lambda^{-1/2} + \frac{1}{6} I_\bullet\Big(\beta, \frac{1}{2}\theta, 0\Big) \cdot \theta^2 \lambda^{-1} + \mathcal{O}(\lambda^{-\frac{3}{2}}). \tag{82}$$

Proof. Proof. We start from

$$e^{-\theta x} = 1 - \theta x + \frac{1}{2}\theta^2 x^2 - \frac{1}{6}\theta^3 x^3 + o(x^3). \tag{83}$$

Therefore, $\forall \epsilon > 0$, $\exists \delta(\epsilon) > 0$, such that, for any $x \in [0, \delta(\epsilon)]$

$$\frac{|e^{-\theta x} - (1 - \theta x + \frac{1}{2}\theta^2 x^2 - \frac{1}{6}\theta^3 x^3)|}{x^3} \leq \epsilon. \tag{84}$$

Combining (84) with (79), we have that

$$-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3 \le u_\lambda(x) \le -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3. \tag{85}$$

In particular, we only consider those $\epsilon \in (0, \frac{1}{6}\theta^3)$ (so that the coefficient $\frac{1}{6}\theta^3 - \epsilon$ in the lower bound part of (85) is positive) and choose $\delta(\epsilon)$ such that

$$\delta(\epsilon) \in \left(0, \frac{1}{2}\theta^2\left(\frac{1}{6}\theta^3 + \epsilon\right)^{-1}\right). \tag{86}$$

With fixed ϵ and $\delta(\epsilon)$, let $\lambda(\epsilon, \delta) = t^2/\delta(\epsilon)^2$. Then, for any $\lambda > \lambda(\epsilon, \delta)$, we have

$$t\lambda^{-1/2} < \delta(\epsilon), \tag{87}$$

and thus, by (85), we have that

$$\begin{aligned}\int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3\Big\}dx &\le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\{u_\lambda(x)\}dx \\ \le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3\Big\}dx.\end{aligned} \tag{88}$$

From (11.10) on p. 33 of Zeltyn and Mandelbaum (2005), $J(t\lambda^{-1/2}) = \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\{u_\lambda(x)\}dx + o(e^{-\nu_1\lambda})$, for some $\nu_1 > 0$. Substituting this into (88) yields

$$\begin{aligned}\int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3\Big\}dx + o(e^{-\nu_1\lambda}) &\le J(t\lambda^{-1/2}) \\ \le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3\Big\}dx + o(e^{-\nu_1\lambda}).\end{aligned} \tag{89}$$

Now, (86) and (87) allow us to apply Lemma 1 to (89) (with δ replaced by $\delta(\epsilon)$, b_1 by β, b_2 by $\frac{1}{2}\theta$, and b_3 by $\frac{1}{\theta}(\frac{1}{6}\theta^3 \pm \epsilon)$), and it follows that

$$\begin{aligned}\frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}) &\le J(t\lambda^{-1/2}) \\ \le \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}).\end{aligned} \tag{90}$$

From (90), we have that, for fixed $\epsilon > 0$, $\exists\lambda_2(\epsilon) > \lambda(\epsilon, \delta)$ such that for any $\lambda > \lambda_2(\epsilon)$,

$$\begin{aligned}\frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon)\lambda^{-1} &\le J(t\lambda^{-1/2}) \\ \le \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon)\lambda^{-1}\end{aligned} \tag{91}$$

or

$$\begin{aligned}\boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon)\lambda^{-1} &\le J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} \\ &\le \boldsymbol{I}_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon)\lambda^{-1}.\end{aligned} \tag{92}$$

Letting $\lambda \to \infty$, we have that

$$\begin{aligned} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon) &\le \liminf_{\lambda\to\infty}\left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) \\ &\le \limsup_{\lambda\to\infty}\left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) \\ &\le I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon). \end{aligned} \tag{93}$$

Letting $\epsilon \to 0$ yields

$$\lim_{\lambda\to\infty}\left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) = \frac{1}{6}I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\theta^2. \tag{94}$$

This implies that

$$J(t\lambda^{-1/2}) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \frac{1}{6}I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\theta^2\lambda^{-1} + o(\lambda^{-1}), \tag{95}$$

and then, from (90), we know that this $o(\lambda^{-1})$ is indeed $\mathcal{O}(\lambda^{-3/2})$. This yields the desired result (81), and (82) follows by letting $t = 0$. □

Finally, we complete the proof of Theorem 3.

Proof. Proof of Theorem 3. From equations (9.7) and (9.15) in Zeltyn and Mandelbaum (2005), we have that, for $\forall t > 0$,

$$P\{W > t\lambda^{-1/2} | W > 0\} = \frac{e^{-\theta t\lambda^{-1/2}} J(t\lambda^{-1/2})}{J(0)}. \tag{96}$$

A straightforward Taylor series expansion yields

$$e^{-\theta t\lambda^{-1/2}} = 1 - \theta t\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{97}$$

Substituting (97), (81), and (82) into (96), we obtain that

$$P\{W > t\lambda^{-1/2} | W > 0\} = d_*(\beta, t) + d_\bullet(\beta, t)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{98}$$

Multiplying (3) with (98) yields (27).

□

C Proof of Theorem 5

From equation (A.2) in Mandelbaum and Zeltyn (2007), we have that

$$P\{\text{Ab}|W > 0\} = \Big(\rho s\theta^{-1}e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta, \lambda/\theta)\Big)^{-1} + 1 - \rho^{-1}, \tag{99}$$

where we note that

$$\rho^{-1} = \mathcal{O}(1), \tag{100}$$

$$1 - \rho^{-1} = \mathcal{O}(\lambda^{-\frac{1}{2}}) = \mathcal{O}(s^{-\frac{1}{2}}). \tag{101}$$

Substituting (53) into (99), we obtain that

$$P\{\text{Ab}|W > 0\} = 1 - \rho^{-1} + \sqrt{\theta}H_\theta(\alpha)\rho^{-1}s^{-1/2} - \frac{1}{3}\rho^{-1}H_\theta(\alpha)^2\theta s^{-1} + \mathcal{O}(s^{-3/2}). \tag{102}$$

Inverting (55) yields

$$P\{W>0\} = A(s,\lambda,\theta) = A_*(\alpha) + \frac{1}{3}\sqrt{\theta}A_*(\alpha)H_\theta(\alpha)s^{-1/2} + \mathcal{O}(s^{-1}). \tag{103}$$

By noting (100), (101), $A_*(\alpha) = \mathcal{O}(1)$, and $H_\theta(\alpha) = \mathcal{O}(1)$, we multiply (102) and (103) to arrive at

$$P\{\text{Ab}\} = A_*(\alpha)(1-\rho^{-1}) + \frac{1}{3}(2+\rho)A_*(\alpha)H_\theta(\alpha)\sqrt{\theta}\rho^{-1}s^{-1/2} + \mathcal{O}(s^{-3/2}). \tag{104}$$

We then just need to derive the series expansion of (104). Inverting (60) yields

$$A_*(\alpha) = A_*(\beta) - h_\theta(\beta)A_*(\beta)^2\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{105}$$

Then, using $1-\rho^{-1} = -\beta\lambda^{-1/2}$, we have that

$$A_*(\alpha)(1-\rho^{-1}) = -A_*(\beta)\beta\lambda^{-1/2} + h_\theta(\beta)A_*(\beta)^2\beta\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}). \tag{106}$$

To expand the second term of (104), we first note that

$$s^{-1/2} = \lambda^{-1/2} - \frac{1}{2}\beta\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}). \tag{107}$$

Combining (107) with (59) and (105), we obtain that

$$\begin{aligned}\frac{1}{3}(2+\rho)A_*(\alpha)H_\theta(\alpha)\sqrt{\theta}\rho^{-1}s^{-1/2} &= A_*(\beta)\sqrt{\theta}H_\theta(\beta)\lambda^{-1/2} + \Big[-\frac{1}{6}\beta A_*(\beta)H_\theta(\beta) \\ &\Big(\beta H_\theta(\beta) - \beta^2\theta^{-1/2} - \sqrt{\theta}\Big) - h_\theta(\beta)A_*(\beta)^2H_\theta(\beta)\sqrt{\theta}\Big]\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}).\end{aligned} \tag{108}$$

Summing (106), (108) and $\mathcal{O}(s^{-3/2}) = \mathcal{O}(\lambda^{-3/2})$ yields the desired result.

D Proof of Theorem 6

Let β_λ be the solution to

$$b_*(\beta_\lambda)\lambda^{-1/2} + b_\bullet(\beta_\lambda)\lambda^{-1} = \epsilon\lambda^{-1/2}, \tag{109}$$

or equivalently

$$b_*(\beta_\lambda) + b_\bullet(\beta_\lambda)\lambda^{-1/2} = \epsilon. \tag{110}$$

Then (40) can be derived the same way as (18) in the proof of Theorem 2. Let $\beta_{\text{opt}} = (s_{\text{opt}} - \lambda)\lambda^{-1/2}$, and $b(\beta) := P\{\text{Ab}\}$. The desired result on the optimality gaps is equivalent to

$$\beta_{\text{opt}} - \beta_* = \mathcal{O}(\lambda^{-1/2}), \tag{111}$$

$$\beta_{\text{opt}} - \Big(\beta_* + \beta_\bullet\lambda^{-1/2}\Big) = \mathcal{O}(\lambda^{-1}). \tag{112}$$

It follows from Theorem 5 that

$$\epsilon\lambda^{-1/2} = b(\beta_{\text{opt}}) = b_*(\beta_{\text{opt}})\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{113}$$

Let $g_*(\lambda) := \beta_{\text{opt}} - \beta_*$. Applying a first-order Taylor expansion, we have that

$$\epsilon\lambda^{-1/2} = b_*(\beta_*)\lambda^{-1/2} + \mathcal{O}(g_*(\lambda)\lambda^{-1/2}) + \mathcal{O}(\lambda^{-1}). \tag{114}$$

Since $b_*(\beta_*)\lambda^{-1/2} = \epsilon\lambda^{-1/2}$, $g_*(\lambda) = \mathcal{O}(\lambda^{-1/2})$ or (111) holds. Because the derivation of $\beta_\bullet$ implies that

$$\beta_\lambda - \Big(\beta_* + \beta_\bullet\lambda^{-1/2}\Big) = \mathcal{O}(\lambda^{-1}), \tag{115}$$

in order to conclude (112), it suffices to prove

$$\beta_{\mathrm{opt}} - \beta_\lambda = \mathcal{O}(\lambda^{-1}). \tag{116}$$

Let $g_\bullet(\lambda) := \beta_{\mathrm{opt}} - \beta_\lambda$. The rest of the proof is again similar as above:

$$\begin{aligned}
\epsilon\lambda^{-1/2} &= b(\beta_{\mathrm{opt}}) = b_*(\beta_{\mathrm{opt}})\lambda^{-1/2} + b_\bullet(\beta_{\mathrm{opt}})\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}) && (117)\\
&= b_*(\beta_\lambda)\lambda^{-1/2} + \mathcal{O}(g_\bullet(\lambda)\lambda^{-1/2}) + b_\bullet(\beta_\lambda)\lambda^{-1} + \mathcal{O}(g_\bullet(\lambda)\lambda^{-1}) + \mathcal{O}(\lambda^{-3/2}). && (118)
\end{aligned}$$

Since $b_*(\beta_\lambda)\lambda^{-1/2} + b_\bullet(\beta_\lambda)\lambda^{-1} = \epsilon\lambda^{-1/2}$, $g_\bullet(\lambda) = \mathcal{O}(\lambda^{-1})$ or (116) holds.

References

Bassamboo, A., R. S. Randhawa. 2009. On the accuracy of fluid models for capacity planning in queueing systems with impatient customers. *Preprint* .

Blanchet, J., P. Glynn. 2006. Complete corrected diffusion approximations for the maximum of a random walk. *Ann. Appl. Probab.* **16**(2) 951–983.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52** 17–34.

Dai, J. G., S. He, T. Tezcan. 2009. Many-server diffusion limits for G/Ph/n + GI queues. *Preprint* .

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

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

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B. Zwart. 2008a. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. in Appl. Probab.* **40**.

Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B. Zwart. 2008b. Refining square root safety staffing by expanding Erlang C. *To appear in Operations Research* .

Kang, W., K. Ramanan. 2008. Fluid limits of many-servers queues with reneging. *Preprint* .

Mandelbaum, A., P. Momcilovic. 2009. Queues with many servers and impatient customers. *Preprint* .

Mandelbaum, A., S. Zeltyn. 2007. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. *Advances in Services Innovations*. Springer-Verlag, 17–48.

Mandelbaum, A., S. Zeltyn. 2008. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Under revision for Operations Research* .

Siegmund, D. 1979. Corrected diffusion approximations in certain random walk problems. *Adv. Appl. Prob.* **11**(4) 701–719.

Spira, R. 1971. Calculation of the Gamma function by Stirling's formula. *Math. Comp.* **25**(114) 317–322.

Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Sci.* **51**(2) 221–235.

Whitt, W. 2005b. Two fluid approximations for multi-server queues with abandonments. *Oper. Res. Lett.* **33** 363–372.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Whitt, W. 2006b. Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Oper. Res.* **54**(2) 247–260.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Syst. Theory Appl.* **51** 361–402.

Zhang, J. 2009. Fluid models of many-server queues with abandonment. *Preprint* URL `http://arxiv.org/abs/0909.1671v1`.