Version of December 20, 2008

Refining square root safety staffing by expanding Erlang C

A.J.E.M. Janssen [1] J.S.H. van Leeuwaarden [2] Bert Zwart [3]

Abstract

We apply a new corrected diffusion approximation for the Erlang C formula to determine staffing levels in cost minimization and constraint satisfaction problems. These problems are motivated by large customer contact centers that are modeled as an $M/M/s$ queue with s the number of servers or agents. The proposed staffing levels are refinements of the celebrated square root safety staffing rule, and have the appealing property that they are as simple as the conventional square root safety staffing rule. In addition, we provide theoretical support for the empirical fact that square root safety staffing works well for moderate-sized systems.

1 Introduction

Customer contact centers, in particular call centers, play a dominant role in society. Customer contact centers can be of any size and appear in a variety of places. Many call centers are being managed by economic principles. In such a setting, it is desirable that agents are highly utilized, answering calls almost 100 percent of the time; on the other hand, a large fraction of customers should receive no or just a small amount of waiting. In their pioneering paper, Halfin & Whitt (1981) showed that when the offered load R is high, and an appropriate number of agents are employed, a system can achieve a high agent utilization and yet deliver a good service level by choosing the number of servers as $R+\beta\sqrt{R}+\mathrm{o}(\sqrt{R})$. If we omit the small order term, we call this *square-root safety staffing.* Since, under square-root safety staffing and large R, the system operates both in heavy traffic and can serve a significant fraction of the customers immediately, the system is said to operate in the Quality-and-Efficiency-Driven (QED) regime. See for example Borst *et al.* (2004), and the survey paper Gans *et al.* (2003).

The emergence of large systems like customer contact centers makes the QED regime practically relevant. This has generated an extensive research effort. Some studies focusing on obtaining limiting approximations for the steady-state distribution or for the time-dependent process are Gamarnik & Momčilovic (2007), Jelenkovic *et al.* (2004), Mandelbaum & Momčilovic (2007), Mandelbaum & Zeltyn (2005), Puhalskii & Reiman (2000), Reed

[1]Philips Research. Digital Signal Processing Group, HTC-36, 5656 AE Eindhoven, The Netherlands. Email: a.j.e.m.janssen@philips.com.

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

[3]Georgia Institute of Technology. H. Milton Stewart School of Industrial and Systems Engineering, 765 Ferst Drive, 30332 Atlanta, USA. Email: bertzwart@gatech.edu.

(2007a,b) and Whitt (2005). Another body of work is concerned with optimization issues and developing asymptotically control policies, see for example Atar (2005), Borst *et al.* (2004), Dai & Tezcan (2007) and Mandelbaum & Zeltyn (2007).

The general idea behind square-root safety staffing is as follows: a finite server system is modeled as a system in heavy-traffic, where the number of servers is large, while at the same time, the system is critically loaded. This can be achieved by setting $s = R + \beta\sqrt{R}$, and letting $R \to \infty$, while keeping $\beta = (s - R)/\sqrt{R}$ fixed. In this way the system reaches the QED regime.

For the $M/M/s$ queue, it is shown in Borst *et al.* (2004) that this procedure has certain asymptotic optimality properties. To illustrate the type of results they obtain, we consider the case of linear costs: waiting costs are w per customer per time unit, and staffing costs are q per time unit per server. Let μ be the service rate, and let the total number of costs be $K(s, R)$. It can be shown that

$$K(s,R) = qs + \frac{w}{\mu}\frac{\lambda}{s-\lambda}C(s,R) = qr + \sqrt{r}\left(q\beta + \frac{w}{\mu\beta}C(s,R)\right),$$

with $C(s, R)$ representing the probability that a customer is delayed, which is known as the Erlang C formula. The main difficulty in optimizing $K(s, R)$ over s is the complexity of $C(s, R)$. Halfin & Whitt (1981) showed that, under square-root safety staffing, $C(s, R)$ converges to a nondegenerate limit $C_*(\beta)$ as $R \to \infty$.

Borst *et al.* (2004) replace $C(s, R)$ by the simpler function $C_*(\beta)$, with $\beta = (s - R)/\sqrt{R}$, reducing the problem to finding an optimal value of β, which we denote by β_*. This leads to an approximation $s_* = R + \beta_*\sqrt{R}$ for the optimal staffing level. Borst *et al.* (2004) show that this procedure yields costs which are optimal up to a level $\mathrm{o}(\sqrt{R})$.

Based on the Halfin-Whitt limiting regime, one expects the approximation s_* to be accurate for large values of R, and in particular relevant for large scale service systems such as customer contact centers. Errors or inaccuracies are expected to arise from the fact that the actual system is finite-sized.

Despite these potential inaccuracies, Borst *et al.* (2004) show by numerical experiments that the approximation s_* performs exceptionally well in almost all regimes. That is, s_* usually differs not more than one agent from the true optimum, even for systems with moderate values of R. The results in Borst *et al.* (2004) suggest that any staffing rule of the form $R + \beta_*\sqrt{R} + o(\sqrt{R})$ is asymptotically optimal. It would therefore be useful to examine what $o(\sqrt{R})$ really means.

In this paper we explore refinements of the square-root safety staffing principle by utilizing a new asymptotic expansion for the Erlang C formula. In particular, we characterize the above-mentioned $\mathrm{o}(\sqrt{R})$ small order term for two staffing problems: we develop staffing rules of the form

$$s_\bullet = R + \beta_*\sqrt{R} + \beta_\bullet, \tag{1.1}$$

with $\beta_\bullet$ a (non-negative) constant. An intriguing finding is that, for the staffing problems we consider, the constant $\beta_\bullet$ is as easy to compute as β_*. It is possible to evaluate the behavior of $\beta_\bullet$ in cases where the bulk of the costs is due to waiting costs (i.e., the quality driven regime), and in the efficiency driven regime, where most costs are related to staffing costs. We refer Sections 1,7 and 8 of Borst *et al.* (2004) for a detailed description of the quality driven regime and the efficiency driven regime.

These refinements also provide theoretical support for the above-mentioned experiments in Borst *et al.* (2004): the precise value of the constant $\beta_\bullet$ turns out to be smaller than one in a large number of cases. In Section 3, we examine staffing under the constraint that the delay probability should be smaller than ϵ. The correction term $\beta_\bullet$ turns out to be smaller than one for values of $\epsilon > 0.1$. Only for very small values of ϵ, in the range of 10^{-3} and smaller, it makes sense to include a correction term. We find that square-root safety staffing is off by about two servers if $\epsilon = 10^{-3}$ and by three to four servers if $\epsilon = 10^{-5}$. The corrected staffing level $s_\bullet$ is accurate well within one server in all cases.

Similar insights are obtained for the scenario with linear waiting and staffing costs, as investigated in Section 4. In that section we establish that a suitable refinement of the form $s_\bullet$ is strongly optimal. In particular, we show that the associated costs are optimal up to a factor $\mathcal{O}(1/\sqrt{R})$ with respect to the optimal value of the continuous relaxation. This is stronger than the result in Borst *et al.* (2004), who obtained optimality up to a factor $o(\sqrt{R})$.

The results in this paper are related to our recent work on bounds and corrected diffusion approximations in the Halfin-Whitt regime for the delay probability in the $M/D/s$ queue and the Erlang B queue, see Janssen *et al.* (2007, 2008). These papers do not consider optimal staffing problems; we state some preliminary results from these papers in Section 2.1. In that section we present bounds for the Erlang delay formula that are valid for all parameter combinations, and are particularly sharp in the Halfin-Whitt regime. We derive a new corrected diffusion approximation in Section 2.2, which can be used to determine staffing levels.

The rest of this paper is organized as follows. Several performance results are developed in Section 2. The staffing problem with a delay constraint is analyzed in Section 3. Section 4 considers the staffing problem with linear waiting and staffing costs. We make several concluding remarks in Section 5 and present additional proofs in Section 6.

2 Bounds and expansions for Erlang C

The objective of this section is to present several performance results that are necessary in this paper. Consider the Erlang C ($M/M/s$) queueing model with Poisson arrival rate λ, exponential service times with mean 1, and s servers. Let $\rho = \lambda/s < 1$ be the system load. The probability that an arriving customer experiences delay is denoted by $C(s,\lambda)$. The Erlang delay formula $C(s,\lambda)$ in its basic form is only defined for integer values of s. An extension of this formula that is well defined for all real $s > \lambda$ is given by (see for example Jagers & Van Doorn (1986))

$$C(s,\lambda) = \left[\lambda \int_0^\infty t e^{-\lambda t}(1+t)^{s-1} dt\right]^{-1}. \tag{2.1}$$

Throughout the paper, we treat this expression as a definition of $C(s,\lambda)$. As explained in Borst *et al.* (2004), determining the optimal number of agents can be done by first solving a continuous optimization problem involving the right-hand side of (2.1). Throughout this paper, $s_{\rm opt}$ is the optimizing value of this continuous relaxation. Jagers & Van Doorn (1986) have shown that $C(s, \lambda)$ is convex in s. By convexity, the optimal number of agents is then determined by a round-up, or round-down of s_{opt}, whichever leads to the most beneficial feasible solution. Therefore, we can (and will) always focus on the continuous relaxation of a staffing problem.

Although (2.1) is fairly explicit, and other explicit expressions for $C(s, \lambda)$ exist (see for example Gross & Harris (1998)), these are not very insightful and tractable if λ or s is large. This motivates us to consider approximations that are sharp for large systems.

To describe these approximations, we introduce the following key parameters:

$$\alpha = \sqrt{-2s(1-\rho+\ln\rho)}, \tag{2.2}$$
$$\beta = (s-\lambda)/\sqrt{\lambda}, \tag{2.3}$$
$$\gamma = (s-\lambda)/\sqrt{s} = (1-\rho)\sqrt{s} = \beta\sqrt{\rho}. \tag{2.4}$$

It has been shown in Lemma 7 of Janssen *et al.* (2008) that $\alpha < \beta$. By expanding $\frac{1}{2}\alpha^2$ in powers of $(1-\rho)$, it easily follows that $\gamma < \alpha$, so we have $\gamma < \alpha < \beta$.

Let $\Phi(u)$ be the distribution function of the standard normal random variable, and let $\phi(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$ be its density. The Halfin-Whitt approximation of the delay probability $C(s, \lambda)$, which is asymptotically exact if $\lambda \to \infty$ and β fixed, reads

$$C_*(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}. \tag{2.5}$$

Sometimes the approximation $C_*(\gamma)$ is used, see for example Whitt (1992). In Janssen *et al.* (2008) it is shown that the usage of α in the Halfin-Whitt type approximation for the Erlang B formula leads to a better approximation than the usage of β or γ.

In Section 2.1 we present upper and lower bounds for the Erlang C formula which have similar structure as the Halfin-Whitt approximation. These bounds are based on our results in Janssen *et al.* (2008) and are shown to hold for the continuous extension (2.1) of the Erlang C formula. Section 2.2 presents a new corrected diffusion approximation for the Erlang C formula (2.1). Proofs of the results in this section are presented in Section 6.

2.1 Bounds for the Erlang C formula

The next result provides bounds for the probability $C(s, \lambda)$ that a customer has to wait in an $M/M/s$ queue as described above.

Theorem 1. *For $s > \lambda$,*

$$C(s,\lambda) \le \left[\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s}}\right)\right]^{-1}, \tag{2.6}$$

and

$$C(s,\lambda) \ge \left[\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s}} + \frac{1}{\phi(\alpha)}\frac{1}{12s-1}\right)\right]^{-1}. \tag{2.7}$$

The proof of this theorem is deferred to Section 6.1. As mentioned in the introduction, the structure of the bounds (2.6), (2.7) is quite similar to the Halfin-Whitt approximation, which is obtained by taking $\lambda \to \infty$ while keeping β fixed. In this asymptotic regime $s \to \infty$, one can see that α and γ both converge to β. With the above theorem at hand, convergence of $C(s,\lambda)$ towards the Halfin-Whitt function $C_*(\beta)$ is obvious. In particular, our bounds are sharp in the Halfin-Whitt regime. The difference between the lower and upper bound is only $\mathcal{O}(1/s)$ (in fact, it is approximately $1/(12s-1)$). We take the opportunity to illustrate the quality of these bounds in Table 1. In Table 1, we keep $\beta = 1$ fixed, and vary s. The load λ is chosen such that $s = \lambda + \beta\sqrt{\lambda}$. The quality of the bounds is apparent, even for small systems.

s	λ	α	(2.7)	$C(s,\lambda)$	(2.6)	$\frac{(2.6)-(2.7)}{C(s,\lambda)}$
1	$3.8197{\cdot}10^{-1}$	$8.2993{\cdot}10^{-1}$	$3.6571{\cdot}10^{-1}$	$3.8197{\cdot}10^{-1}$	$3.9437{\cdot}10^{-1}$	$7.5040{\cdot}10^{-2}$
2	$1.0000{\cdot}10^{0}$	$8.7897{\cdot}10^{-1}$	$3.2678{\cdot}10^{-1}$	$3.3333{\cdot}10^{-1}$	$3.3936{\cdot}10^{-1}$	$3.7727{\cdot}10^{-2}$
5	$3.2087{\cdot}10^{0}$	$9.2364{\cdot}10^{-1}$	$2.8886{\cdot}10^{-1}$	$2.9097{\cdot}10^{-1}$	$2.9328{\cdot}10^{-1}$	$1.5181{\cdot}10^{-2}$
10	$7.2984{\cdot}10^{0}$	$9.4624{\cdot}10^{-1}$	$2.6937{\cdot}10^{-1}$	$2.7030{\cdot}10^{-1}$	$2.7142{\cdot}10^{-1}$	$7.6160{\cdot}10^{-3}$
20	$1.6000{\cdot}10^{1}$	$9.6215{\cdot}10^{-1}$	$2.5565{\cdot}10^{-1}$	$2.5608{\cdot}10^{-1}$	$2.5663{\cdot}10^{-1}$	$3.8180{\cdot}10^{-3}$
50	$4.3411{\cdot}10^{1}$	$9.7618{\cdot}10^{-1}$	$2.4361{\cdot}10^{-1}$	$2.4377{\cdot}10^{-1}$	$2.4398{\cdot}10^{-1}$	$1.5310{\cdot}10^{-3}$
100	$9.0488{\cdot}10^{1}$	$9.8320{\cdot}10^{-1}$	$2.3761{\cdot}10^{-1}$	$2.3769{\cdot}10^{-1}$	$2.3779{\cdot}10^{-1}$	$7.6654{\cdot}10^{-4}$
200	$1.8635{\cdot}10^{2}$	$9.8815{\cdot}10^{-1}$	$2.3340{\cdot}10^{-1}$	$2.3344{\cdot}10^{-1}$	$2.3349{\cdot}10^{-1}$	$3.8365{\cdot}10^{-4}$
500	$4.7813{\cdot}10^{2}$	$9.9252{\cdot}10^{-1}$	$2.2969{\cdot}10^{-1}$	$2.2970{\cdot}10^{-1}$	$2.2972{\cdot}10^{-1}$	$1.5360{\cdot}10^{-4}$
1000	$9.6887{\cdot}10^{2}$	$9.9472{\cdot}10^{-1}$	$2.2783{\cdot}10^{-1}$	$2.2783{\cdot}10^{-1}$	$2.2784{\cdot}10^{-1}$	$7.6836{\cdot}10^{-5}$

Table 1: Results for the bounds on $C(s,\lambda)$ for $\beta = 1$.

2.2 A corrected diffusion approximation

The goal of this section is to obtain a two-term corrected diffusion approximation of the delay probability in the case that $\lambda \to \infty$ and β is bounded. In this case set $s = \lambda + \beta\sqrt{\lambda}$ and define

$$C_\lambda(\beta) := C(s,\lambda). \tag{2.8}$$

The results of Halfin & Whitt (1981) imply that $C_\lambda(\beta) \to C_*(\beta)$. The theorem in this section is a refinement of this result and appears to be new. For corrected diffusion approximations for single-server queues, we refer to Blanchet & Glynn (2006) and Siegmund (1979).

We need the following notation. A function $f(\beta,\lambda)$ is said to be of $\mathcal{U}O(1/\lambda)$ if for any $0 < \beta_g < \beta_d < \infty$,

$$\sup_{\lambda>0,\beta\in[\beta_g,\beta_d]} \lambda|f(\beta,\lambda)| < \infty. \tag{2.9}$$

This is a useful notion, since it allows one to vary β with λ, which we will do in the next section, where we will optimize over β. All functions we will consider that are of $\mathcal{O}(1/\lambda)$ will be $\mathcal{U}O(1/\lambda)$ as well (recall that a function f is said to be $\mathcal{O}(x)$ if $\sup f(x)/x < \infty$).

Theorem 2. *As* $\lambda \to \infty$,

$$C_\lambda(\beta) = C_*(\beta) + C_\bullet(\beta)\frac{\beta}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda), \tag{2.10}$$

with

$$C_\bullet(\beta) = C_*(\beta)^2 \left[\frac{1}{3} + \frac{\beta^2}{6} + \frac{\Phi(\beta)}{\phi(\beta)}\left(\frac{\beta}{2} + \frac{\beta^3}{6}\right)\right]. \tag{2.11}$$

Note that $C_\bullet(\cdot)$ is a smooth function. Although it is possible to derive Theorem 2 from Theorem 1, we follow the suggestion of a referee and give a direct proof of Theorem 2 in Section 6.2 which is based on (2.1). Although Theorem 2 may be interesting in itself, its main purpose in this paper is to serve as departure point for determining refined staffing levels of the form (1.1). To evaluate the performance, we recommend using the bounds in Theorem 1, or the series expansion given in Theorem 8 of Janssen *et al.* (2008). Refining the square-root safety staffing levels will be the topic of the next two sections.

3 Corrected staffing under a delay constraint

A classical problem is to determine the number of servers necessary to ensure that the fraction of customers that are delayed before entering service is below a certain threshold, say ϵ.

Borst *et al.* (2004) propose to determine the number of servers as a round-up of $s_* = \lambda + \beta_*(\epsilon)\sqrt{\lambda}$, with $\beta = \beta_*$ the solution of $C_*(\beta(\epsilon)) = \epsilon$. A natural question is how well this approximation performs. To obtain more insight, we propose to replace $\beta_*(\epsilon)$ with $\beta_*(\epsilon) + \beta_\bullet(\epsilon)/\sqrt{\lambda}$, giving rise to the corrected staffing level $s_\bullet = \lambda + \beta_*(\epsilon)\sqrt{\lambda} + \beta_\bullet(\epsilon)$, where $\beta_\bullet(\epsilon)$ needs to be determined. Surprisingly, as will be shown below, $\beta_\bullet(\epsilon)$ can be written explicitly in terms of $\beta_*(\epsilon)$ so that the additional computational requirement for this staffing level is negligible.

The goal is to determine $\tilde{\beta}$ such that

$$C_\lambda(\tilde{\beta}) = C_*(\tilde{\beta}) + C_\bullet(\tilde{\beta})\frac{\tilde{\beta}}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) = \epsilon. \tag{3.1}$$

It therefore makes sense to consider the equation

$$C_*(\tilde{\beta}) + C_\bullet(\tilde{\beta})\frac{\tilde{\beta}}{\sqrt{\lambda}} = \epsilon. \tag{3.2}$$

We fix ϵ and write $\beta_* = \beta_*(\epsilon)$. Replace $\tilde{\beta}$ by $\beta_* + g(\lambda)$ in (3.2). Apply a Taylor approximation for $C_*(\tilde{\beta})$ to obtain

$$C_*(\beta_*) + g(\lambda)C_*'(\beta_*) + \mathcal{O}(g(\lambda)^2) + C_\bullet(\beta_*)\frac{\beta_*}{\sqrt{\lambda}} + \mathcal{O}(g(\lambda)/\sqrt{\lambda}) = \epsilon. \tag{3.3}$$

By definition, the first term equals ϵ, which yields

$$g(\lambda) = -\frac{C_\bullet(\beta_*)}{C_*'(\beta_*)}\frac{\beta_*}{\sqrt{\lambda}} + \mathcal{O}(1/\lambda). \tag{3.4}$$

We thus define

$$\beta_\bullet(\epsilon) = -\frac{C_\bullet(\beta_*(\epsilon))}{C'_*(\beta_*(\epsilon))}\beta_*(\epsilon). \tag{3.5}$$

This expression can be simplified by using the identity $C_*(\beta_*(\epsilon)) = \epsilon$, which implies

$$\frac{\beta_*(\epsilon)\Phi(\beta_*(\epsilon))}{\phi(\beta_*(\epsilon))} = \frac{1}{\epsilon} - 1. \tag{3.6}$$

In addition, observe that for any $\beta > 0$,

$$C'_*(\beta) = -C_*(\beta)^2\left(\frac{\Phi(\beta)}{\phi(\beta)} + \frac{\beta}{C_*(\beta)}\right). \tag{3.7}$$

Applying these results several times we obtain the following theorem.

Theorem 3.

$$\beta_\bullet(\epsilon) = \beta_*(\epsilon)\frac{(1-\epsilon)\left(\frac{1}{2}\beta_*(\epsilon) + \frac{1}{6}\beta_*(\epsilon)^3\right) + \epsilon\left(\frac{1}{3}\beta_*(\epsilon) + \frac{1}{6}\beta_*(\epsilon)^3\right)}{1-\epsilon+\beta_*(\epsilon)^2}. \tag{3.8}$$

If $\epsilon \downarrow 0$, it can be shown by taking logarithms in (3.6) that $\beta_*(\epsilon) \sim \sqrt{-2\ln\epsilon}$. We therefore see that

$$\beta_\bullet(\epsilon) \sim \frac{1}{6}\beta_*(\epsilon)^2 \sim \frac{1}{3}\ln(1/\epsilon). \tag{3.9}$$

Based on this expansion, one could conclude that standard square-root safety staffing may produce rather optimistic estimates of the required number of servers when ϵ is small. Since $\ln 10 \approx 2.3 < 3$, a safe choice is to add n more servers to s_* if the delay requirement is 10^{-n}. If $\epsilon \to 1$, then $\beta_*(\epsilon) \sim (1-\epsilon)\sqrt{2/\pi}$, implying that

$$\beta_\bullet(\epsilon) \sim \frac{2}{3\pi}(1-\epsilon). \tag{3.10}$$

This suggests that the conventional square-root safety staffing algorithm is sharp when the delay constraint is not too severe. This is further illustrated by Figure 1, which plots $\beta_\bullet(\epsilon)$.

Figure 1 clearly shows why square-root safety staffing works so well: the next term in the expansion indicates that square-root safety staffing is off by less than one server if the delay requirement is not too stringent. If the delay requirement becomes very strict, it seems worthwhile to add a correction term.

We now compare the staffing levels s_* and $s_\bullet$ with the optimal staffing level s_{opt} by the equation $C(s_{\text{opt}}, \lambda) = \epsilon$; recall that s_{opt} is the solution of the continuous relaxation of the server staffing problem. For values of ϵ around 0.5 we find that the difference between the optimal staffing level and s_* (or $s_\bullet$) is well within one server. It is therefore more interesting to show the results for smaller values of ϵ. Tables 2, 3 and 4 report results for $\epsilon = 10^{-1}, 10^{-3}$ and 10^{-5}. In all of these cases, the corrected staffing level produces very accurate results; in all cases $s_\bullet$ is within one server of s_{opt}. If the desired probability of delay is 10 percent, s_* is not off by more than one server. As the desired delay probability gets smaller, the square-root safety staffing level s_* underestimates the correct staffing level - up to 3 servers for $\epsilon = 10^{-5}$. In all cases $\beta_\bullet$ accurately predicts the deviation of s_* from the optimal staffing level.

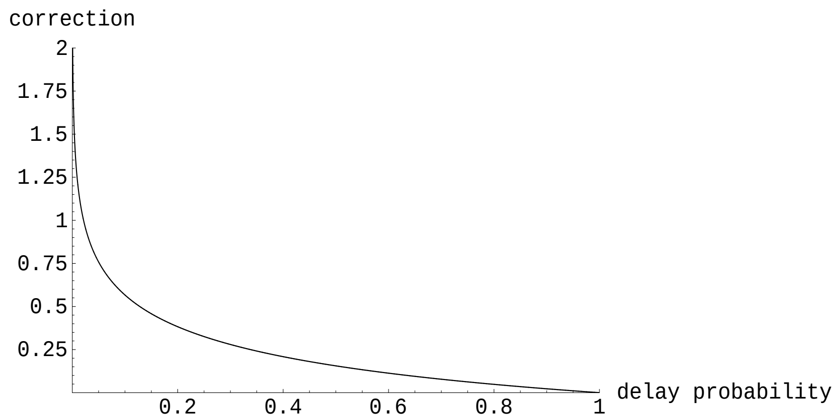

Figure 1: The correction term $\beta_\bullet$ as function of the delay probability ϵ

4 Corrected staffing under a linear cost structure

The second staffing problem we consider determines the number of servers s in such a way that a certain cost function is minimized. As in the previous section, this is equivalent to choosing β. It is obvious that optimizing total costs using $C_*(\beta)$ is more tractable than using $C_\lambda(\beta)$. Extensive numerical experiments conducted in Borst *et al.* (2004) show that optimal agent staffing based on $C_*(\beta)$ rather than $C_\lambda(\beta)$ lead to staffing levels which are usually not more off than a single agent. The cost structure we consider is a special case of that in Borst *et al.* (2004), and is as follows. Waiting costs are assumed to be w per customer per time unit, and service costs are assumed to equal q per agent per time unit. The expected waiting time is equal to $C(s,\lambda)/(s(1-\rho))$. The expected total costs $K(s,\lambda)$ per unit of time becomes

$$\begin{aligned} K(s,\lambda) &= w\frac{\lambda}{(1-\rho)s}C_\lambda(\beta) + qs & (4.1)\\ &= w\sqrt{\lambda}\frac{C_\lambda(\beta)}{\beta} + q\lambda + q\beta\sqrt{\lambda} & (4.2)\\ &=: q\lambda + \sqrt{\lambda}K_\lambda(\beta), & (4.3)\end{aligned}$$

with $K_\lambda(\beta) = \frac{wC_\lambda(\beta)}{\beta} + q\beta$. The structure of this cost function is illuminating, since it can be decomposed into two terms. The first term, $q\lambda$, is the amount of costs necessary to keep the system stable, and it is independent of β. To optimize (that is, minimize) the second term over β, the idea is to replace $K_\lambda(\beta)$ by a simpler cost function. To explain the general procedure outlined in Borst *et al.* (2004) in the present case, let β_* correspond to the optimal staffing level found by optimizing the function

$$K_*(\beta) = \frac{w}{\beta}C_*(\beta) + q\beta. \tag{4.4}$$

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$2.9315{\cdot}10^{0}$	$2.4202{\cdot}10^{0}$	$-5.1134{\cdot}10^{-1}$	$2.9868{\cdot}10^{0}$	$5.5299{\cdot}10^{-2}$
2	$4.5328{\cdot}10^{0}$	$4.0084{\cdot}10^{0}$	$-5.2435{\cdot}10^{-1}$	$4.5751{\cdot}10^{0}$	$4.2293{\cdot}10^{-2}$
5	$8.7134{\cdot}10^{0}$	$8.1756{\cdot}10^{0}$	$-5.3775{\cdot}10^{-1}$	$8.7423{\cdot}10^{0}$	$2.8892{\cdot}10^{-2}$
10	$1.5036{\cdot}10^{1}$	$1.4491{\cdot}10^{1}$	$-5.4534{\cdot}10^{-1}$	$1.5058{\cdot}10^{1}$	$2.1304{\cdot}10^{-2}$
20	$2.6902{\cdot}10^{1}$	$2.6351{\cdot}10^{1}$	$-5.5110{\cdot}10^{-1}$	$2.6918{\cdot}10^{1}$	$1.5537{\cdot}10^{-2}$
50	$6.0599{\cdot}10^{1}$	$6.0042{\cdot}10^{1}$	$-5.5653{\cdot}10^{-1}$	$6.0609{\cdot}10^{1}$	$1.0109{\cdot}10^{-2}$
100	$1.1476{\cdot}10^{2}$	$1.1420{\cdot}10^{2}$	$-5.5939{\cdot}10^{-1}$	$1.1477{\cdot}10^{2}$	$7.2537{\cdot}10^{-3}$
200	$2.2065{\cdot}10^{2}$	$2.2008{\cdot}10^{2}$	$-5.6146{\cdot}10^{-1}$	$2.2065{\cdot}10^{2}$	$5.1831{\cdot}10^{-3}$
500	$5.3232{\cdot}10^{2}$	$5.3176{\cdot}10^{2}$	$-5.6333{\cdot}10^{-1}$	$5.3232{\cdot}10^{2}$	$3.3089{\cdot}10^{-3}$
1000	$1.0455{\cdot}10^{3}$	$1.0449{\cdot}10^{3}$	$-5.6429{\cdot}10^{-1}$	$1.0455{\cdot}10^{3}$	$2.3509{\cdot}10^{-3}$

Table 2: Results for $\epsilon = 10^{-1}$; $\beta_* = 1.4202$ and $\beta_\bullet = 0.5666$.

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$5.7408{\cdot}10^{0}$	$4.1153{\cdot}10^{0}$	$-1.6256{\cdot}10^{0}$	$6.0350{\cdot}10^{0}$	$2.9412{\cdot}10^{-1}$
2	$8.0910{\cdot}10^{0}$	$6.4056{\cdot}10^{0}$	$-1.6854{\cdot}10^{0}$	$8.3253{\cdot}10^{0}$	$2.3433{\cdot}10^{-1}$
5	$1.3718{\cdot}10^{1}$	$1.1966{\cdot}10^{1}$	$-1.7516{\cdot}10^{0}$	$1.3886{\cdot}10^{1}$	$1.6811{\cdot}10^{-1}$
10	$2.1643{\cdot}10^{1}$	$1.9851{\cdot}10^{1}$	$-1.7917{\cdot}10^{0}$	$2.1771{\cdot}10^{1}$	$1.2796{\cdot}10^{-1}$
20	$3.5756{\cdot}10^{1}$	$3.3932{\cdot}10^{1}$	$-1.8239{\cdot}10^{0}$	$3.5852{\cdot}10^{1}$	$9.5831{\cdot}10^{-2}$
50	$7.3884{\cdot}10^{1}$	$7.2028{\cdot}10^{1}$	$-1.8556{\cdot}10^{0}$	$7.3948{\cdot}10^{1}$	$6.4077{\cdot}10^{-2}$
100	$1.3303{\cdot}10^{2}$	$1.3115{\cdot}10^{2}$	$-1.8730{\cdot}10^{0}$	$1.3307{\cdot}10^{2}$	$4.6688{\cdot}10^{-2}$
200	$2.4594{\cdot}10^{2}$	$2.4406{\cdot}10^{2}$	$-1.8859{\cdot}10^{0}$	$2.4598{\cdot}10^{2}$	$3.3751{\cdot}10^{-2}$
500	$5.7156{\cdot}10^{2}$	$5.6966{\cdot}10^{2}$	$-1.8979{\cdot}10^{0}$	$5.7158{\cdot}10^{2}$	$2.1783{\cdot}10^{-2}$
1000	$1.1004{\cdot}10^{3}$	$1.0985{\cdot}10^{3}$	$-1.9041{\cdot}10^{0}$	$1.1004{\cdot}10^{3}$	$1.5565{\cdot}10^{-2}$

Table 3: Results for $\epsilon = 10^{-3}$; $\beta_* = 3.1153$ and $\beta_\bullet = 1.9197$.

Let $s_* = \lambda + \beta_* \sqrt{\lambda}$ and let K_{opt} be the optimal cost level of the continuous relaxation. It is obvious that $K_{\text{opt}} \leq K(s_*, \lambda)$. In Borst *et al.* (2004) it is shown that the staffing level s_* is asymptotically optimal in the sense that

$$K(s_*, \lambda) = K_{\text{opt}} + \mathrm{o}(\sqrt{\lambda}). \tag{4.5}$$

This brings us to the goal of the present section. Our aim is to find a staffing level $s_\bullet$ such that the stronger result

$$K(s_\bullet, \lambda) = K_{\text{opt}} + \mathcal{O}(1/\sqrt{\lambda}) \tag{4.6}$$

holds. Again, this staffing level is of the form

$$s_\bullet = \lambda + \beta_* \sqrt{\lambda} + \beta_\bullet = s_* + \beta_\bullet. \tag{4.7}$$

As in the previous section, we shall give an explicit characterization of $\beta_\bullet$.

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$8.0194{\cdot}10^{0}$	$5.2758{\cdot}10^{0}$	$-2.7436{\cdot}10^{0}$	$8.6388{\cdot}10^{0}$	$6.1943{\cdot}10^{-1}$
2	$1.0907{\cdot}10^{1}$	$8.0468{\cdot}10^{0}$	$-2.8602{\cdot}10^{0}$	$1.1410{\cdot}10^{1}$	$5.0281{\cdot}10^{-1}$
5	$1.7555{\cdot}10^{1}$	$1.4561{\cdot}10^{1}$	$-2.9937{\cdot}10^{0}$	$1.7924{\cdot}10^{1}$	$3.6935{\cdot}10^{-1}$
10	$2.6598{\cdot}10^{1}$	$2.3521{\cdot}10^{1}$	$-3.0773{\cdot}10^{0}$	$2.6884{\cdot}10^{1}$	$2.8571{\cdot}10^{-1}$
20	$4.2268{\cdot}10^{1}$	$3.9122{\cdot}10^{1}$	$-3.1460{\cdot}10^{0}$	$4.2485{\cdot}10^{1}$	$2.1703{\cdot}10^{-1}$
50	$8.3450{\cdot}10^{1}$	$8.0234{\cdot}10^{1}$	$-3.2157{\cdot}10^{0}$	$8.3597{\cdot}10^{1}$	$1.4735{\cdot}10^{-1}$
100	$1.4601{\cdot}10^{2}$	$1.4276{\cdot}10^{2}$	$-3.2547{\cdot}10^{0}$	$1.4612{\cdot}10^{2}$	$1.0833{\cdot}10^{-1}$
200	$2.6375{\cdot}10^{2}$	$2.6047{\cdot}10^{2}$	$-3.2842{\cdot}10^{0}$	$2.6383{\cdot}10^{2}$	$7.8861{\cdot}10^{-2}$
500	$5.9892{\cdot}10^{2}$	$5.9561{\cdot}10^{2}$	$-3.3118{\cdot}10^{0}$	$5.9897{\cdot}10^{2}$	$5.1241{\cdot}10^{-2}$
1000	$1.1385{\cdot}10^{3}$	$1.1352{\cdot}10^{3}$	$-3.3263{\cdot}10^{0}$	$1.1386{\cdot}10^{3}$	$3.6746{\cdot}10^{-2}$

Table 4: Results for $\epsilon = 10^{-5}$; $\beta_* = 4.2758$ and $\beta_\bullet = 3.3631$.

4.1 A corrected optimization problem

Theorem 2 provides an approximation for $C_\lambda(\beta)$ that is correct up to $\mathcal{U}O(1/\lambda)$. It is clear that we can write

$$K_\lambda(\beta) = K_*(\beta) + \frac{w}{\sqrt{\lambda}} C_\bullet(\beta) + \mathcal{U}O(1/\lambda). \tag{4.8}$$

This motivates us to consider the cost function

$$K_\bullet(\beta) = K_*(\beta) + \frac{w}{\sqrt{\lambda}} C_\bullet(\beta). \tag{4.9}$$

Let $\beta_*(\lambda)$ be the optimal point of this cost function. It is clear that $\beta_*(\lambda)$ is the solution of the equation

$$K_*'(\beta) = -\frac{w}{\sqrt{\lambda}} C_\bullet'(\beta). \tag{4.10}$$

Write $\beta_*(\lambda) = \beta_* + \epsilon(\lambda)$. From the last equation, and since $C_\bullet'$ is continuous, it follows that $\epsilon(\lambda) \downarrow 0$. Since $K_*'(\beta_*) = 0$, and by invoking Taylor's theorem, we observe that

$$K_*'(\beta_*(\lambda)) = \epsilon(\lambda) K_*''(\beta_*) + \mathcal{O}(\epsilon(\lambda)^2). \tag{4.11}$$

In addition, we have

$$\frac{w}{\sqrt{\lambda}} C_\bullet'(\beta_*(\lambda)) + \mathcal{O}(1/\lambda) = \frac{w}{\sqrt{\lambda}} C_\bullet'(\beta_*) + \mathcal{O}(\frac{\epsilon(\lambda)}{\sqrt{\lambda}}) + \mathcal{O}(1/\lambda). \tag{4.12}$$

Combining the last three displays, we conclude that

$$\epsilon(\lambda) = -\frac{w C_\bullet'(\beta_*)}{K_*''(\beta_*)} \frac{1}{\sqrt{\lambda}} + \mathcal{O}(1/\lambda), \tag{4.13}$$

Define $\beta_\bullet = -\frac{w C_\bullet'(\beta_*)}{K_*''(\beta_*)}$. This formula for $\beta_\bullet$ can be simplified by noting that

$$K_*''(\beta) = \frac{w}{\beta} \left[C_*''(\beta) - \frac{2}{\beta} C_*'(\beta) + \frac{2}{\beta^2} C_*''(\beta) \right]. \tag{4.14}$$

This yields

$$\beta_\bullet = -\frac{\beta_* C'_\bullet(\beta_*)}{C''_*(\beta_*) - \frac{2}{\beta_*}C'_*(\beta_*) + \frac{2}{\beta_*^2}C_*(\beta_*)} = -\frac{\beta_* C'_\bullet(\beta_*)}{C''_*(\beta_*) + 2q/w}. \tag{4.15}$$

We are now ready to prove the main result of this section.

Theorem 4. *Let* $s_\bullet = s_* + \beta_\bullet$ *with* $\beta_\bullet$ *defined by* (4.15). *Then*

$$K(s_\bullet, \lambda) = K_{\text{opt}} + \mathcal{O}(1/\sqrt{\lambda}). \tag{4.16}$$

Proof. Let $\bar{\beta}(\lambda)$ be the optimizing value of $K_\lambda(\beta)$ and observe that

$$\begin{aligned} K_{\text{opt}} &= \lambda q + \sqrt{\lambda} K_\lambda(\bar{\beta}(\lambda)) \\ &= \lambda q + \sqrt{\lambda}\left(K_\bullet(\bar{\beta}(\lambda)) + \mathcal{U}O(1/\lambda)\right). \\ &\geq \lambda q + \sqrt{\lambda} K_\bullet(\beta_*(\lambda)) + \mathcal{O}(1/\sqrt{\lambda}). \end{aligned}$$

The third step follows from the property $K_\bullet(\bar{\beta}(\lambda)) \geq K_\bullet(\beta_*(\lambda))$ and the result $\bar{\beta}(\lambda) \to \beta_*$, which is shown below. From the relation between $\beta_*(\lambda)$ and $\beta_* + \beta_\bullet/\sqrt{\lambda}$, it follows that

$$K_\bullet(\beta_*(\lambda)) = K_\bullet(\beta_* + \beta_\bullet/\sqrt{\lambda}) + \mathcal{O}(1/\lambda). \tag{4.17}$$

This yields

$$K(s_\bullet, \lambda) \leq K_{\text{opt}} + \mathcal{O}(1/\sqrt{\lambda}). \tag{4.18}$$

The proof is completed by noting that $K(s_\bullet, \lambda) \geq K_{\text{opt}}$. □

In the proof above we have used the following fact for the optimizing value $\bar{\beta}(\lambda)$ of $K_\lambda(\beta)$.

Lemma 1. $\bar{\beta}(\lambda) \to \beta_*$.

We could not find a proof of this result in the literature, therefore we include it for completeness.

Proof. We first note that $\limsup_{\lambda\to\infty} \bar{\beta}(\lambda) < \infty$ as shown in Borst *et al.* (2004). Next we show that $\liminf_\lambda \bar{\beta}(\lambda) > 0$. For this, we derive a lower bound on the delay probability that is explicit in β, using the lower bound in Theorem 1, and using that $\gamma < \beta$ and also $\alpha < \beta$, cf. [12], Lemma 7. Combining these bounds yields

$$C_\lambda(\beta)^{-1} \leq 1 + \frac{12}{11}\sqrt{2\pi}\beta \exp\{\frac{1}{2}\beta^2\} =: \hat{C}_\lambda(\beta)^{-1}.$$

Replace $C_\lambda(\beta)$ with $\hat{C}_\lambda(\beta)$ in the cost optimization problem, and call the corresponding cost function $\hat{K}_\lambda(\beta)$. If there would exist a subsequence (λ_n) such that $\beta(\lambda_n) \to 0$ as $n \to \infty$, this would imply that the cost $K_\lambda(\beta(\lambda_n))$ would be lower bounded by $\hat{K}_\lambda(\beta(\lambda_n))$, which diverges along the chosen subsequence. This violates the fact that $K_\lambda(\beta(\lambda_n)) \to K_*(\beta_*)$ which is shown in Borst *et al.* (2004).

Now assume that $\lambda \to \infty$ along a subsequence such that $\beta(\lambda) \to \tilde{\beta}$ for some $\tilde{\beta}$. By the above considerations, $\tilde{\beta} \in (0, \infty)$. By Theorem 2, $C_\lambda(\beta)$ converges to $C_*(\beta)$ uniformly in a neighborhood of $\tilde{\beta}$, which implies that $K_\lambda(\beta(\lambda)) \to K_*(\tilde{\beta})$. Since also $K_\lambda(\beta(\lambda)) \to K_*(\beta_*)$, we conclude that $K_*(\tilde{\beta}) = K_*(\beta_*)$. Since C_* is strictly convex and decreasing (Borst *et al.* (2004), Lemma B.1), we conclude from (4.14) that K_* is strictly convex. Thus, we arrive at $\beta_* = \tilde{\beta}$. This holds for any converging subsequence, from which the statement follows. □

Note that both β_* and $\beta_\bullet$ are a function of $t = q/w$, so we can write $\beta_* = \beta_*(t)$ and $\beta_\bullet = \beta_\bullet(t)$.

Like in the previous section, we can estimate the behavior of the correction term $\beta_\bullet(t)$ as the ratio t becomes small or large, although the analysis is more involved here. Remark 6.4 of Borst *et al.* (2004) implies that $\beta_*(t) \sim \sqrt{-2\ln t}$ as $t \downarrow 0$ and $\beta_*(t) \sim 1/\sqrt{t}$ as $t \to \infty$. Note that small values of t correspond to the quality driven regime, where most customers enter service immediately after arrival. Similarly, large values of t correspond to the efficiency driven regime, where most customer are delayed before entering service, cf. Borst *et al.* (2004).

Proposition 1. *In the quality driven regime (i.e. as $t \downarrow 0$):*

$$\beta_\bullet(t) \sim \frac{1}{9}\ln(1/t). \tag{4.19}$$

In the efficiency driven regime (as $t \uparrow \infty$):

$$\beta_\bullet(t) \sim \frac{1}{3\sqrt{2\pi}} t^{-3/2}. \tag{4.20}$$

The proof of this result is presented in Section 6.3.

The asymptotic estimates are illustrated by Figure 2, which plots $\beta_\bullet$ as function of q/w. Again, the moderate size of our correction term shows why square-root safety staffing works so well for almost every value of q/w. Only for large values of w (with respect to q), it is necessary to include a correction term.

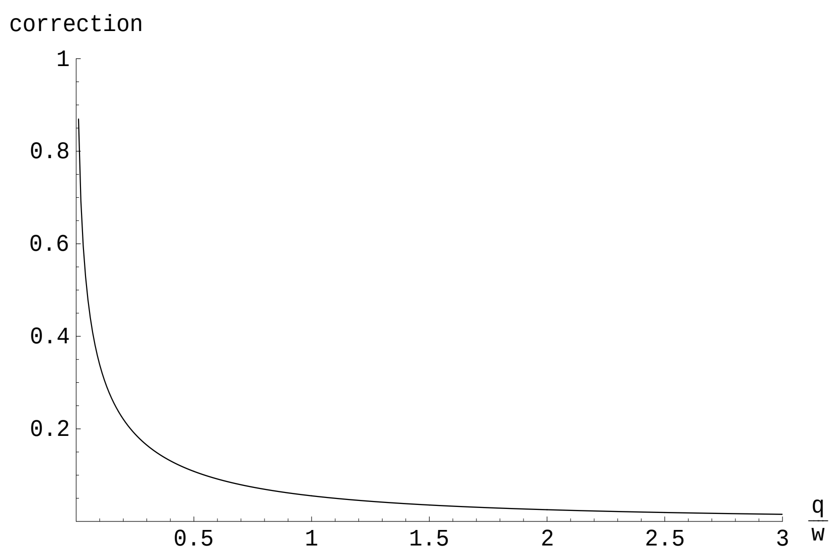

Figure 2: The correction term $\beta_\bullet$ as function of q (for $w = 1$)

We close this section with a numerical illustration of our results. For values of q/w that are moderate, the difference between the optimal staffing level, square-root safety staffing, and corrected staffing is within a single agent. Here we restrict to presenting numerical results

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$2.9239{\cdot}10^0$	$2.6674{\cdot}10^0$	$-2.5645{\cdot}10^{-1}$	$3.0059{\cdot}10^0$	$8.2069{\cdot}10^{-2}$
2	$4.6328{\cdot}10^0$	$4.3581{\cdot}10^0$	$-2.7469{\cdot}10^{-1}$	$4.6966{\cdot}10^0$	$6.3828{\cdot}10^{-2}$
5	$9.0226{\cdot}10^0$	$8.7284{\cdot}10^0$	$-2.9411{\cdot}10^{-1}$	$9.0670{\cdot}10^0$	$4.4410{\cdot}10^{-2}$
10	$1.5578{\cdot}10^1$	$1.5273{\cdot}10^1$	$-3.0540{\cdot}10^{-1}$	$1.5611{\cdot}10^1$	$3.3117{\cdot}10^{-2}$
20	$2.7771{\cdot}10^1$	$2.7457{\cdot}10^1$	$-3.1415{\cdot}10^{-1}$	$2.7795{\cdot}10^1$	$2.4369{\cdot}10^{-2}$
50	$6.2113{\cdot}10^1$	$6.1790{\cdot}10^1$	$-3.2252{\cdot}10^{-1}$	$6.2129{\cdot}10^1$	$1.5994{\cdot}10^{-2}$
100	$1.1700{\cdot}10^2$	$1.1667{\cdot}10^2$	$-3.2698{\cdot}10^{-1}$	$1.1701{\cdot}10^2$	$1.1533{\cdot}10^{-2}$
200	$2.2391{\cdot}10^2$	$2.2358{\cdot}10^2$	$-3.3018{\cdot}10^{-1}$	$2.2392{\cdot}10^2$	$8.3396{\cdot}10^{-3}$
500	$5.3762{\cdot}10^2$	$5.3728{\cdot}10^2$	$-3.3322{\cdot}10^{-1}$	$5.3762{\cdot}10^2$	$5.2957{\cdot}10^{-3}$
1000	$1.0531{\cdot}10^3$	$1.0527{\cdot}10^3$	$-3.3476{\cdot}10^{-1}$	$1.0531{\cdot}10^3$	$3.7525{\cdot}10^{-3}$

Table 5: Results for $q/w = 10^{-1}$; $\beta_* = 1.6674$ and $\beta_\bullet = 0.3385$.

for $q/w = 10^{-1}$, 10^{-3} and 10^{-5}, see Tables 5–7. The conclusions are similar to those in the previous section: while the corrected staffing algorithm is accurate for all cost structures, conventional square-root safety staffing tends to underestimate the optimal number of agents as waiting costs become higher.

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$5.3309{\cdot}10^0$	$4.1678{\cdot}10^0$	$-1.1631{\cdot}10^0$	$5.6809{\cdot}10^0$	$3.4999{\cdot}10^{-1}$
2	$7.7131{\cdot}10^0$	$6.4800{\cdot}10^0$	$-1.2331{\cdot}10^0$	$7.9931{\cdot}10^0$	$2.8000{\cdot}10^{-1}$
5	$1.3395{\cdot}10^1$	$1.2083{\cdot}10^1$	$-1.3111{\cdot}10^0$	$1.3597{\cdot}10^1$	$2.0196{\cdot}10^{-1}$
10	$2.1376{\cdot}10^1$	$2.0018{\cdot}10^1$	$-1.3588{\cdot}10^0$	$2.1531{\cdot}10^1$	$1.5430{\cdot}10^{-1}$
20	$3.5564{\cdot}10^1$	$3.4167{\cdot}10^1$	$-1.3967{\cdot}10^0$	$3.5680{\cdot}10^1$	$1.1638{\cdot}10^{-1}$
50	$7.3835{\cdot}10^1$	$7.2400{\cdot}10^1$	$-1.4351{\cdot}10^0$	$7.3913{\cdot}10^1$	$7.7988{\cdot}10^{-2}$
100	$1.3313{\cdot}10^2$	$1.3168{\cdot}10^2$	$-1.4560{\cdot}10^0$	$1.3319{\cdot}10^2$	$5.7085{\cdot}10^{-2}$
200	$2.4627{\cdot}10^2$	$2.4480{\cdot}10^2$	$-1.4715{\cdot}10^0$	$2.4631{\cdot}10^2$	$4.1617{\cdot}10^{-2}$
500	$5.7232{\cdot}10^2$	$5.7083{\cdot}10^2$	$-1.4855{\cdot}10^0$	$5.7235{\cdot}10^2$	$2.7606{\cdot}10^{-2}$
1000	$1.1017{\cdot}10^3$	$1.1002{\cdot}10^3$	$-1.4921{\cdot}10^0$	$1.1017{\cdot}10^3$	$2.1011{\cdot}10^{-2}$

Table 6: Results for $q/w = 10^{-3}$; $\beta_* = 3.1678$ and $\beta_\bullet = 1.5131$.

5 Concluding remarks

This paper established a corrected diffusion approximation for the Erlang delay formula, that yields refinements of square-root safety staffing levels as considered by Borst *et al.* (2004). These refinements enable an analytical assessment of the accuracy of square-root safety staffing. If the fraction of customers that have to wait is not too small (say 0.05 or higher), then the correction term $\beta_\bullet$ is well within one server. This indicates that the speed

λ	s_{opt}	s_*	$s_* - s_{\text{opt}}$	$s_\bullet$	$s_\bullet - s_{\text{opt}}$
1	$7.5224{\cdot}10^0$	$5.2985{\cdot}10^0$	$-2.2239{\cdot}10^0$	$8.2139{\cdot}10^0$	$6.9140{\cdot}10^{-1}$
2	$1.0432{\cdot}10^1$	$8.0790{\cdot}10^0$	$-2.3525{\cdot}10^0$	$1.0994{\cdot}10^1$	$5.6280{\cdot}10^{-1}$
5	$1.7112{\cdot}10^1$	$1.4612{\cdot}10^1$	$-2.4998{\cdot}10^0$	$1.7527{\cdot}10^1$	$4.1549{\cdot}10^{-1}$
10	$2.6186{\cdot}10^1$	$2.3593{\cdot}10^1$	$-2.5929{\cdot}10^0$	$2.6508{\cdot}10^1$	$3.2238{\cdot}10^{-1}$
20	$4.1894{\cdot}10^1$	$3.9224{\cdot}10^1$	$-2.6702{\cdot}10^0$	$4.2139{\cdot}10^1$	$2.4509{\cdot}10^{-1}$
50	$8.3146{\cdot}10^1$	$8.0395{\cdot}10^1$	$-2.7505{\cdot}10^0$	$8.3311{\cdot}10^1$	$1.6480{\cdot}10^{-1}$
100	$1.4578{\cdot}10^2$	$1.4299{\cdot}10^2$	$-2.7962{\cdot}10^0$	$1.4590{\cdot}10^2$	$1.1915{\cdot}10^{-1}$
200	$2.6358{\cdot}10^2$	$2.6079{\cdot}10^2$	$-2.7904{\cdot}10^0$	$2.6371{\cdot}10^2$	$1.2491{\cdot}10^{-1}$
500	$5.9897{\cdot}10^2$	$5.9612{\cdot}10^2$	$-2.8528{\cdot}10^0$	$5.9903{\cdot}10^2$	$6.2537{\cdot}10^{-2}$
1000	$1.1388{\cdot}10^3$	$1.1359{\cdot}10^3$	$-2.8635{\cdot}10^0$	$1.1388{\cdot}10^3$	$5.1818{\cdot}10^{-2}$

Table 7: Results for $q/w = 10^{-5}$; $\beta_* = 4.2985$ and $\beta_\bullet = 2.9153$.

of convergence of the optimal safety staffing factor to its limiting value is fast, which explains why square-root safety staffing works so well for moderate-sized systems. If the costs of delay are more stringent, then including a correction term makes sense.

We are currently carrying out a similar program for the Erlang model with abandonments. Mandelbaum & Zeltyn (2007) report less favorable numerical results on conventional square-root safety staffing in this setting; it therefore makes sense to include a correction term in this case.

6 Additional proofs

6.1 Proof of Theorem 1

The bounds in Theorem 1 are based on similar bounds for the Erlang B blocking formula, which in turn are based on a continuous extension of the Erlang B formula, derived in Janssen *et al.* (2008), and the well-known identity

$$C(s,\lambda)^{-1} = \rho + (1-\rho)B(s,\lambda)^{-1}, \tag{6.1}$$

with

$$B(s,\lambda)^{-1} = \lambda \int_0^\infty e^{-\lambda t}(1+t)^s dt \tag{6.2}$$

(it is actually not hard to show this identity directly by combining (6.1) and (2.1)). The continuous extension used in Janssen *et al.* (2008) reads

$$B(s,\lambda)^{-1} = \frac{1}{\phi(\alpha)\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{x^2}{2}} y'(x/\sqrt{s})dx, \tag{6.3}$$

where y is the function that solves the equation

$$y(x) + \ln(1-y(x)) = -\frac{1}{2}x^2, \qquad y(0) = 0.$$

To prove Theorem 1, it suffices to show that the continuous extensions (6.3) and (6.2) are identical for all values of s. In Janssen *et al.* (2008), the following identity is shown:

$$\frac{1}{\phi(\alpha)\sqrt{2\pi}}\int_{-\infty}^{\alpha} e^{-\frac{x^2}{2}} y'(x/\sqrt{s})dx = \frac{se^s}{\phi(\alpha)\sqrt{2\pi}}\int_{\rho}^{\infty} u^s e^{-us}du. \tag{6.4}$$

Note that, using the definition of ϕ and α in the first step and the change of variables $v = u/\rho$ in the second,

$$\begin{aligned}\frac{se^s}{\phi(\alpha)\sqrt{2\pi}}\int_{\rho}^{\infty} u^s e^{-us}du &= se^{\lambda}\rho^{-s}\int_{u=\rho}^{\infty} e^{-us}u^s du \\ &= \lambda e^{-\lambda}\int_{v=1}^{\infty} e^{-v\lambda}v^s dv \\ &= \lambda\int_{t=0}^{\infty} e^{-t\lambda}(t+1)^s dt.\end{aligned}$$

We conclude that (6.3) and (6.2) are indeed identical. The proof of Theorem 1 now follows by combining (6.1) with Theorem 1 of Janssen *et al.* (2008).

6.2 Proof of Theorem 2

As a point of departure, we use the expression (2.1) with $s = \lambda + \beta\sqrt{\lambda}$, and assume throughout this section that $\beta = (s-\lambda)/\sqrt{\lambda}$ is bounded. Consider the function

$$h(t) = \lambda t - (\lambda + \beta\sqrt{\lambda})\ln(1+t), \tag{6.5}$$

with derivative

$$h'(t) = \lambda\frac{t - \beta/\sqrt{\lambda}}{1+t}, \tag{6.6}$$

vanishing at $t = \beta/\sqrt{\lambda} =: t_0$. Let $\alpha > 0$ and $v \in \mathbb{R}$ be given by (see (2.2))

$$-\frac{1}{2}\alpha^2 = h(t_0), \quad \frac{1}{2}(v^2 - \alpha^2) = h(t), \quad t \geq 0. \tag{6.7}$$

We take v such that the sign of v is equal to the sign of $t_0 - t$. We write $t = t(v)$ for the inverse function, noting that this inverse function is well-defined on all of $(-\infty, \alpha]$. From the definition of h and t_0 we have that

$$\alpha = \sqrt{-2h(t_0)} = \beta - \frac{\beta^2}{6\sqrt{\lambda}} + \mathcal{O}(1/\lambda). \tag{6.8}$$

It then follows that

$$[C_{\lambda}(\beta)]^{-1} = -\lambda\int_{-\infty}^{\alpha}\frac{t(v)}{1+t(v)}e^{-\frac{1}{2}(v^2-\alpha^2)}t'(v)dv. \tag{6.9}$$

Combining $h'(t(v))t'(v) = v$ with (6.7) and (6.6) we see that

$$\frac{-\lambda t'(v)}{1+t(v)} = \frac{v}{t_0 - t(v)}, \tag{6.10}$$

and so

$$[C_\lambda(\beta)]^{-1} = e^{\frac{1}{2}\alpha^2} \int_{-\infty}^{\alpha} \frac{vt(v)}{t_0 - t(v)} e^{-\frac{1}{2}v^2} dv. \tag{6.11}$$

We aim to estimate $[C_\lambda(\beta)]^{-1}$ with an accuracy of $\mathcal{O}(1/\lambda)$ while keeping β bounded, so it is enough to consider integration ranges $[v(\lambda), \alpha]$ in (6.11) such that $\lambda \exp -\frac{1}{2}v^2(\lambda) \to 0$ as $\lambda \to \infty$. For our purposes, it is sufficient to consider v's such that $v/\sqrt{\lambda} \to 0$ as $\lambda \to \infty$. For such v's, we approximate $t(v)$ by first writing

$$\begin{aligned} h(t) &= h(t_0) + \frac{1}{2}h''(t_0)(t_0 - t)^2 - \frac{1}{6}h'''(t_0)(t_0 - t)^3 + \dots \\ &= -\frac{1}{2}\alpha^2 + \frac{\lambda}{2(1+t_0)}(t_0 - t)^2 + \frac{\lambda}{3(1+t_0)^2}(t_0 - t)^3 + \dots . \end{aligned} \tag{6.12}$$

Solving t (such that the sign of $t_0 - t$ equals the sign of v) from $h(t) = \frac{1}{2}v^2 - \frac{1}{2}\alpha^2$, we then find

$$t_0 - t(v) = \frac{v}{\sqrt{\lambda}} + \frac{\beta v}{2\lambda} - \frac{v^2}{3\lambda} + \mathcal{O}\left(\frac{v^3}{\lambda^{3/2}}\right). \tag{6.13}$$

For the v's we are considering, we have

$$\begin{aligned} \frac{vt(v)}{t_0 - t(v)} &= \frac{\beta - v - \frac{\beta v}{2\sqrt{\lambda}} + \frac{v^2}{3\sqrt{\lambda}} + \mathcal{O}\left(\frac{v^3}{\lambda}\right)}{1 + \frac{\beta}{2\sqrt{\lambda}} - \frac{v}{3\sqrt{\lambda}} + \mathcal{O}\left(\frac{v^2}{\lambda}\right)} \\ &= \beta - v + \frac{\beta v}{3\sqrt{\lambda}} - \frac{\beta^2}{2\sqrt{\lambda}} + \mathcal{O}\left(\frac{1+|v|^5}{\lambda}\right). \end{aligned} \tag{6.14}$$

Thus, we get, cf. (6.11),

$$[C_\lambda(\beta)]^{-1} = e^{\frac{1}{2}\alpha^2} \int_{-\infty}^{\alpha} \left(\beta - v + \frac{\beta}{\sqrt{\lambda}}\left(\frac{v}{3} - \frac{\beta}{2}\right)\right) e^{-\frac{1}{2}v^2} dv + \mathcal{O}(1/\lambda), \tag{6.15}$$

where we have restored the lower integration limit from $-v(\sqrt{\lambda})$ to $-\infty$ at the expense of exponentially small error. Using the definition of Φ and ϕ, and

$$\int_{-\infty}^{\alpha} ve^{-\frac{1}{2}v^2} dv = -e^{-\frac{1}{2}\alpha^2}, \tag{6.16}$$

it then follows that

$$[C_\lambda(\beta)]^{-1} = 1 + \frac{\beta\Phi(\alpha)}{\phi(\alpha)} - \frac{\beta}{\sqrt{\lambda}}\left(\frac{1}{3} + \frac{\beta\Phi(\beta)}{2\phi(\beta)}\right) + \mathcal{O}(1/\lambda). \tag{6.17}$$

Finally, noting that $\alpha = \beta - \beta^2/(6\sqrt{\lambda}) + \mathcal{O}(1/\lambda)$, we obtain

$$\begin{aligned} \beta\frac{\Phi(\alpha)}{\phi(\alpha)} &= \beta\frac{\Phi(\beta)}{\phi(\beta)} + \beta\left(\frac{\Phi(\beta)}{\phi(\beta)}\right)'(\alpha - \beta) + \mathcal{O}(1/\lambda) \\ &= \beta\frac{\Phi(\beta)}{\phi(\beta)} + \beta\left(1 + \beta\frac{\Phi(\beta)}{\phi(\beta)}\right)\frac{-\beta^2}{6\sqrt{\lambda}} + \mathcal{O}(1/\lambda). \end{aligned} \tag{6.18}$$

Theorem 2 now follows by noting that the above estimates are all valid when β is bounded, and by inserting (6.18) into (6.17).

Alternative proof of Theorem 2

As an alternative, Theorem 2 can be proven from Theorem 1, by expanding α, γ, ρ in powers of β, by developing Taylor series expansions for quantities like $\Phi/(\alpha)/\phi(\alpha)$: From Theorem 1 it can be seen that

$$C_\lambda(\beta)^{-1} = \rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s}}\right) + \mathcal{U}O(1/\lambda). \tag{6.19}$$

Simple computations show that

$$\begin{aligned}
\rho &= 1 - \frac{\beta}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda), &(6.20)\\
1/\sqrt{s} &= 1/\sqrt{\lambda} + \mathcal{U}O(1/\lambda), &(6.21)\\
\alpha^2 &= \beta^2 - \frac{1}{3}\beta^3\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda), &(6.22)\\
\alpha &= \beta - \frac{1}{6}\beta^2\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda), &(6.23)\\
\gamma &= \beta - \frac{1}{2}\beta^2\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda). &(6.24)
\end{aligned}$$

These relations will be used several times. The property in (2.9) for the remainder terms follows from the estimate for the remainder term in the corresponding Taylor series expansions, combined with the fact that all second derivatives are continuous (and therefore locally bounded) functions. A similar argument holds for the computations below. Next, we write

$$\frac{\Phi(\alpha)}{\phi(\alpha)} = \frac{\Phi(\beta)}{\phi(\beta)} + \frac{\Phi(\alpha) - \Phi(\beta)}{\phi(\beta)} + \Phi(\alpha)\left(\frac{1}{\phi(\alpha)} - \frac{1}{\phi(\beta)}\right). \tag{6.25}$$

Number the terms on the right hand side by I, II, III. We see that, using $\Phi(\alpha) - \Phi(\beta) = (\alpha - \beta)\phi(\beta) + \mathcal{U}O(1/\lambda)$,

$$\begin{aligned}
\mathrm{II} &= \phi(\beta)^{-1}(\alpha-\beta)\phi(\beta) + \mathcal{U}O(1/\lambda)\\
&= \alpha - \beta + \mathcal{O}^*(1/\lambda) = -\frac{1}{6}\beta^2\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda).
\end{aligned}$$

For the third term, observe that the derivative of $1/\phi(x)$ equals $x/\phi(x)$. Therefore,

$$\begin{aligned}
\frac{1}{\phi(\alpha)} - \frac{1}{\phi(\beta)} &= (\alpha-\beta)\frac{\beta}{\phi(\beta)} + \mathcal{U}O((\alpha-\beta)^2)\\
&= -\frac{1}{\phi(\beta)}\frac{1}{6}\beta^3\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda).
\end{aligned}$$

This yields

$$\mathrm{III} = -\frac{\Phi(\alpha)}{\phi(\beta)}\frac{1}{6}\beta^3\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) = -\frac{\Phi(\beta)}{\phi(\beta)}\frac{1}{6}\beta^3\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda). \tag{6.26}$$

Inserting these estimates for II and III in (6.19), we obtain

$$
\begin{aligned}
C_\lambda(\beta)^{-1} &= \rho + \gamma\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\gamma\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) \\
&= \rho + \gamma\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\beta\frac{1}{\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) \\
&= 1 - \frac{\beta}{3\sqrt{\lambda}} + \gamma\frac{\Phi(\alpha)}{\phi(\alpha)} + \mathcal{U}O(1/\lambda) \\
&= 1 - \frac{\beta}{3\sqrt{\lambda}} + \gamma\frac{\Phi(\beta)}{\phi(\beta)} - \frac{\gamma\beta^2}{6\sqrt{\lambda}} - \gamma\frac{\Phi(\beta)}{\phi(\beta)}\frac{\beta^3}{6\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) \\
&= C_*(\beta)^{-1} - \frac{\beta}{3\sqrt{\lambda}} - \frac{\Phi(\beta)}{\phi(\beta)}\frac{\beta^2}{2\sqrt{\lambda}} - \frac{\beta^3}{6\sqrt{\lambda}} - \frac{\Phi(\beta)}{\phi(\beta)}\frac{\beta^4}{6\sqrt{\lambda}} + \mathcal{U}O(1/\lambda) \\
&= C_*(\beta)^{-1} - \frac{1}{\sqrt{\lambda}}\left[\frac{\beta}{3} + \frac{\beta^3}{6} + \frac{\beta\Phi(\beta)}{\phi(\beta)}\left(\frac{\beta}{2} + \frac{\beta^3}{6}\right)\right] + \mathcal{U}O(1/\lambda).
\end{aligned}
$$

The expansion for $C_\lambda(\beta)$ then easily follows.

6.3 Proof of Proposition 1

We first need to work out the expressions for C_*' and C_*''. Since the derivative of $\Phi(\beta)/\phi(\beta)$ equals $1/C_*(\beta)$, it follows that

$$
C_*'(\beta) = -C_*(\beta)^2\frac{\Phi(\beta)}{\phi(\beta)} - \beta C_*(\beta). \tag{6.27}
$$

To get a convenient form for the second derivative, note that

$$
\begin{aligned}
C_*''(\beta) &= -2C_*(\beta)\frac{\Phi(\beta)}{\phi(\beta)}C_*'(\beta) - 2C_*(\beta) - \beta C_*'(\beta) \\
&= 2C_*(\beta)^3\left(\frac{\Phi(\beta)}{\phi(\beta)}\right)^2 + 2\beta C_*(\beta)^2\frac{\Phi(\beta)}{\phi(\beta)} - 2C_*(\beta) + \beta C_*(\beta)^2\frac{\Phi(\beta)}{\phi(\beta)} + \beta^2 C_*(\beta) \\
&= C_*(\beta)\left[2\left(\frac{\Phi(\beta)}{\phi(\beta)}\right)^2 + 2\beta C_*(\beta)\frac{\Phi(\beta)}{\phi(\beta)} - 2 + C_*(\beta)\frac{\beta\Phi(\beta)}{\phi(\beta)} + \beta^2\right] \\
&= C_*(\beta)\left[\frac{2}{\beta^2}\left(1 - C_*(\beta)\right)^2 + 1 - 3C_*(\beta) + \beta^2\right].
\end{aligned}
$$

In the last step we used the identity $C_*(\beta)\frac{\beta\Phi(\beta)}{\phi(\beta)} = 1 - C_*(\beta)$.

We now use these expressions to find the limiting behavior of these quantities at 0 and ∞. Using $C(\beta) \sim \phi(\beta)/\beta$ as $\beta \to \infty$, it follows that

$$
\begin{aligned}
C_*'(\beta) &\sim -\phi(\beta), \\
C_*''(\beta) &\sim \beta\phi(\beta),
\end{aligned}
$$

as $\beta \to \infty$. When $\beta \downarrow 0$, observe that $1 - C(\beta) \sim 1$, which implies

$$
\begin{aligned}
C_*'(\beta) &\to -\sqrt{\pi/2} \\
C_*''(\beta) &\to \pi - 2.
\end{aligned}
$$

Finally, rewrite $C_\bullet(\beta)$ into

$$C_\bullet(\beta) = C_*(\beta)\left(\frac{1}{2}+\frac{\beta^2}{6}\right) - \frac{1}{6}C_*(\beta)^2, \tag{6.28}$$

which implies that

$$C'_\bullet(\beta) = C'_*(\beta)\left(\frac{1}{2}+\frac{\beta^2}{6}\right) + C_*(\beta)\frac{\beta}{3} - \frac{1}{3}C_*(\beta)C'_*(\beta). \tag{6.29}$$

From this expression and the above results, it easily follows that

$$C'_\bullet(\beta) \sim -\frac{\beta^2}{6}\phi(\beta) \tag{6.30}$$

as $\beta \to \infty$, and that

$$C'_\bullet(\beta) \to -\frac{1}{6}\sqrt{\pi/2} \tag{6.31}$$

as $\beta \downarrow 0$.

Replace now β with $\beta_*(t)$ in the above expressions. Suppose first that $t \to 0$, in which case $\beta_*(t) \to \infty$. Combining all the above we see that

$$\begin{aligned} \beta_\bullet(t) &\sim \frac{1}{6}\beta_*(t)\frac{\beta_*(t)^2\phi(\beta_*(t))}{\beta_*(t)\phi(\beta_*(t))+2t} \\ &= \frac{1}{6}\beta_*^2(t)\frac{1}{1+2\frac{t}{\beta_*(t)\phi(\beta_*(t))}}. \end{aligned}$$

Since $\beta_*(t)$ satisfies the first order condition

$$t = \frac{C_*(\beta)}{\beta_*(t)^2} + C_*(\beta_*(t))^2\frac{\Phi(\beta_*(t))}{\phi(\beta_*(t))} + \beta_*(t)C_*(\beta_*(t)), \tag{6.32}$$

we conclude that $\frac{t}{\beta_*(t)\phi(\beta_*(t))} \to 1$ as $t \downarrow 0$. This implies

$$\beta_\bullet(t) \sim \frac{1}{18}\beta_*(t)^2 \sim \frac{1}{9}\ln(1/t). \tag{6.33}$$

Next, consider the case that $t \to \infty$, in which case $\beta_*(t) \to 0$. Combining the above results once more we arrive at

$$\beta_\bullet(t) \sim \beta_*(t)\frac{1}{6\sqrt{\pi/2}}\frac{1}{\pi-2+2t} \sim \frac{1}{3\sqrt{2\pi}}t^{-3/2}. \tag{6.34}$$

Acknowledgments
The research of Johan van Leeuwaarden is supported by an NWO VENI grant. The research of Bert Zwart is supported in part by NSF grants CMMI-0727400 and CNS-0718701. We would like to thank a referee for providing a large number of helpful comments that led to a significant improvement in the presentation. In addition, we would like to thank another referee for suggesting the direct proof of Theorem 2 that is now followed in Section 6.2.

References

[1] Atar, R. (2005b). Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. *Annals of Applied Probability* **15**: 2606-2650.

[2] Borst S., A. Mandelbaum, M. Reiman (2004). Dimensioning large call centers. *Operations Research* **52**: 17–34.

[3] Blanchet, J., P. Glynn (2006). Complete corrected diffusion approximations for the maximum of a random walk. *Annals of Applied Probability* **16**, 951-983.

[4] De Bruijn, N.G. (1981). *Asymptotic Methods in Analysis.* Dover Publications, New York.

[5] Dai, J.G., T. Tezcan (2007). Optimal Control of Parallel Server Systems with Many Servers in Heavy Traffic. Submitted for publication.

[6] Gamarnik, D., P. Momčilovic (2007). Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. Submitted for publication.

[7] Gans, N., G. Koole, A. Mandelbaum (2003). Telephone Call Centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Operations Management* **5**: 79-141.

[8] Gross, M. , J. Harris (1998). *Fundamentals of Queueing Theory.* Wiley, New York.

[9] Halfin, S., W. Whitt (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**: 567–588.

[10] Harel, A. (1988). Sharp bounds and approximations for the Erlang delay and loss formulas. *Management Science* **34**: 959–972.

[11] Jagers, A.A., E.A. van Doorn (1986). On the continued Erlang loss function. *Operations Research Letters* **5**:43-46.

[12] Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B. Zwart (2008). Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. *Queueing Systems* **58**: 261-301.

[13] Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B. Zwart (2008). Normal approximations for the Poisson distribution and the Erlang B formula. *Advances in Applied Probability* **40**: 122-143.

[14] Jagerman, D. (1974). Some properties of the Erlang loss function. *Bell System Technical Journal* **53**: 525–551.

[15] Jelenković, P., A. Mandelbaum, P. Momčilovic (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47**: 53–69.

[16] Mandelbaum, A., P. Momčilovic (2007). Queues with many servers: The virtual waiting-time process in the QED regime. To appear in *Mathematics of Operations Research.*

[17] Mandelbaum, A., Z. Zeltyn (2005). Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* **51**: 361–402.

[18] Mandelbaum, A., Z. Zeltyn (2007). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Submitted for publication.

[19] Puhalskii, A., M. Reiman (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability* **32**: 564–595.

[20] Reed, J. (2007b). The $G/GI/N$ queue in the Halfin-Whitt regime. Submitted for publication.

[21] Reed, J. (2007). The $G/GI/N$ queue in the Halfin-Whitt regime II: Idle time System Equations. Submitted for publication.

[22] Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems. *Advances in Applied Probability* **11**: 701–719.

[23] Szegö, G. (1922). Über eine Eigenschaft der Exponentialreihe. *Sitzungsberichte der Berliner Math. Gesellschaft* **22**: 50–64.

[24] Whitt, W. (1992). Understanding the Efficiency of Multi-Server Service Systems. *Management Science* **38**: 708–723.

[25] Whitt, W. (2005). Heavy traffic limit theorems for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research* **30**:1–27.