SENSITIVITY OF PERFORMANCE IN THE ERLANG-A QUEUEING MODEL TO CHANGES IN THE MODEL PARAMETERS

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027-6699

Abstract

This paper studies the $M/M/s+M$ queue, i.e., the $M/M/s$ queue with customer abandonment, also called the Erlang-A model, having independent and identically distributed customer abandon times with an exponential distribution (the $+M$), focusing on the case in which the arrival rate and the number of servers are large. The goal is to better understand the sensitivity of performance to changes in the model parameters: the arrival rate, the service rate, the number of servers, and the abandonment rate. Elasticities are used to show the percentage change of a performance measure caused by a small percentage change in a parameter. Elasticities are calculated using an exact numerical algorithm and simple finite-difference approximations. Insight is gained by applying fluid and diffusion approximations. The analysis shows that performance is quite sensitive to small percentage changes in the arrival rate or the service rate, but relatively insensitive to small percentage changes in the abandonment rate.

Keywords: sensitivity analysis, elasticities, multiserver queues, queues with customer abandonment, Erlang queueing models, heavy traffic, diffusion approximation, fluid limit.

July 3, 2004; Revision: January 23, 2005

1. Introduction

Motivated by telephone call centers and more general customer contact centers, in this paper we study the multi-server queue with customer abandonments, focusing on the case in which the arrival rate and the number of servers are large; see Garnett et al. (2002), Gans et al. (2003) and Borst et al. (2004) for background. In particular, we consider the relatively elementary $M/M/s+M$ model, also known as the Erlang-A model, having a Poisson arrival process with arrival rate λ, independent and identically distributed (IID) service times (independent of the arrival process) with an exponential distribution having mean $1/\mu$, s homogeneous servers working in parallel, unlimited waiting room, IID customer abandon times (independent of the arrival and service processes) with an exponential distribution having mean $m_a = 1/\theta$ (the $+M$) and the first-come first-served service discipline. Abandonment is recognized as an important feature in call centers, and the IID assumption for the abandon times is natural for the invisible queues occurring in call centers.

In this paper we study the sensitivity of performance in the Erlang-A model to changes in the model parameters. In doing so, we were motivated by a statistical approach proposed by Pierson and Whitt (2005) to approximate the steady-state performance of the more general $M/GI/s + GI$ model, having general service-time and time-to-abandon distributions. Whitt (2005a) had previously proposed approximating the $M/GI/s+GI$ model by the purely Markovian $M/M/s + M(n)$ model, having IID exponential service times with the same mean and general state-dependent abandonment rates. The total abandonment rate when there are k customers waiting in queue, δ_k, is approximated by

$$\delta_k \approx \sum_{j=1}^{k} h(j/\lambda), \quad k \geq 0 \ , \tag{1.1}$$

where $h(x) \equiv f(x)/(1 - F(x))$ is the hazard function associated with the time-to-abandon cumulative distribution function (cdf) F, having probability density function f. Alternatively, one could use more complicated exact abandonment rates in the $M/M/s + GI$ model determined by Brandt and Brandt (2002). (The general approach of a state-dependent Markovian approximation was proposed by Brandt and Brandt (1999, 2002), but the specific approach to the $M/GI/s + GI$ model in Whitt (2005a) is different.)

Given that the $M/M/s+M(n)$ model can provide a good approximation to the $M/GI/s+GI$ model, it is natural to consider directly using the $M/M/s + M(n)$ model, without making direct reference to the exact time-to-abandon distribution. Pierson and Whitt (2005) investigate how such a direct $M/M/s+M(n)$ model fit can work by simulating various $M/GI/s+GI$ models, directly estimating state-dependent abandonment rates from the simulation output, and then using the $M/M/s+M(n)$ algorithm in Whitt (2005a) with these estimated abandonment rates. With ample data (long simulation runs), the exact state-dependent abandonment rates estimated in that way were found to be close to the approximation in (1.1), which provides additional support for the approximation in (1.1). The approximate performance was also close to the performance of the original $M/GI/s+GI$ model. The statistical procedure also performed quite well with only limited data.

The effectiveness of the statistical procedure with limited data clearly depends in part upon the sensitivity of performance in the $M/M/s+M(n)$ model to inaccuracies, or small changes, in the abandonment rates. That led to the present study: We wanted to see if performance in the $M/M/s+M(n)$ model is indeed relatively insensitive to small changes in the abandonment rates. In this paper we address that question for the $M/M/s+M$ special case. Our results here show that indeed the performance is remarkably insensitive to changes in the abandonment rate.

The first issue is how to evaluate the sensitivity. The natural direct approach is to calculate derivatives of performance measures with respect to the parameters, but it is difficult to interpret the derivatives. To aid interpretation, we follow the long tradition in economics and look at elasticities. Paralleling the price elasticity of demand, we look at elasticities such as the arrival-rate elasticity of the abandonment probability. The *elasticity* is the derivative of the performance measures (regarded as a function of the model parameter) multiplied by the parameter, divided by the performance measure itself. For example, if $f(\lambda)$ is the abandonment probability $P(Ab)$ as a function of the arrival rate λ, having derivative f', then the *arrival-rate elasticity of the abandonment probability* is

$$\mathcal{E}(f,\lambda) \equiv \frac{\lambda f'}{f} \equiv \frac{\lambda f'(\lambda)}{f(\lambda)} \ ; \qquad (1.2)$$

it shows the percentage change in the abandonment probability resulting from a small percentage change in the arrival rate. Very crudely, the sensitivity may be judged as large or small depending on whether the elasticity is greater than or less than 1.

The second issue is how to calculate the derivatives. The natural direct approach is to differentiate formulas for the performance measures, but we do not do that. When convenient formulas are available, it is natural to directly differentiate them, but the method is limited to those performance measures for which tractable formulas are available. Instead, we use the exact numerical algorithm for the $M/M/s+M$ model in Whitt (2005a) and simple finite-difference approximations; i.e., we approximate the derivative by

$$f'(\lambda) \approx \frac{f(\lambda+h)-f(\lambda)}{h} \tag{1.3}$$

for small positive h. We verify accuracy by performing the calculation for different intervals h, e.g., $h = 10^{-j}$ for $j = 3, 4, 5$. In this paper we show that the numerical algorithm is indeed effective for calculating the derivatives and the associated elasticities.

It is significant that the numerical algorithm has wider scope than we exploit here. First, it applies directly to more general $M/GI/s+GI$ models as an approximation. In that setting, it can be used to investigate sensitivity of performance to other parameters. For example, the same methods can be used to study sensitivity to variability. That can be accomplished in a variety of ways. One way is to work with two-parameter families of service-time or time-to-abandon distributions, such as gamma or lognormal, and differentiate with respect to the squared coefficient of variation (CSQ, variance divided by the square of the mean) of the distribution. It is convenient to work with CSQ's because they measure variability independent of scale (the mean).

We also show that useful insight can be gained from heavy-traffic diffusion and fluid approximations. Especially useful are the diffusion approximations arising in the Quality-and-Efficiency-Driven (QED) many-server heavy-traffic limiting regime developed by Garnett et al. (2002). Those approximations are easy to work with and remarkably accurate. Moreover, the QED approximations tell an interesting story: The arrival-rate and service-rate elasticities of the diffusion approximations for the standard performance measures are all of order $O(\sqrt{s})$ as $s \to \infty$, while the abandonment-rate elasticities of the diffusion approximations for the same performance measures are all of order $O(1)$. Analysis of the elasticities of the diffusion approximations shows that performance in the $M/M/s+M$ model for large s (in the QED regime) is remarkably sensitive to changes in the arrival rate or the service rate, but remarkably insensitive to changes in the abandonment rate.

We also investigate elasticities associated with deterministic fluid limits in the Efficiency-Driven (ED) many-server heavy-traffic limiting regime, drawing upon Whitt (2004, 2005b). In contrast to the QED regime, in the ED regime all the elasticities are of order $O(1)$ as $s \to \infty$. However the abandonment-rate elasticities approach 1, whereas the others approach a limit that explodes as the traffic intensity approaches 1, the critical value for stability without abandonment. So we see another view of the same phenomenon in the ED regime.

The different degrees of sensitivity have implications for our concern about the underlying model in applications. When the sensitivity of performance to a parameter is high, we should worry more about uncertainty about that model parameter, because the consequences from errors in specifying that parameter will be greater. Since the sensitivity to the arrival rate in the $M/M/s+M$ model is relatively large, we should be concerned about uncertainty about the arrival rate. That suggests that it may be wise to directly address uncertainty about the arrival rate in the analysis, as has been done in Whitt (2005c).

There is a substantial body of related literature: The sensitivity issue is closely related to the issue of *model continuity* or stability; the object there is to conclude that performance is a continuous function of a model parameter or a model distribution; e.g., see Whitt (1980), Chapter 5 of Kalashnikov and Rachev (1990), Rachev (1991) and the many references therein. Sensitivity goes beyond continuity to focus on derivatives. The sensitivity issue for Erlang models (A, B and C) is also related to the *convexity* issue for these models; see Harel (1990), Harel and Zipkin (1987), Jagers and van Doorn (1991) and references therein. Earlier papers that focus on many-server heavy-traffic *scaling* are Erlang (1924), Jagerman (1974) and Halfin and Whitt (1981). Sensitivity of performance in the $G/G/1/C$ model was studied via derivatives of the Brownian heavy-traffic approximation in Section 9 of Berger and Whitt (1992). Sensitivity of performance to the service-time distribution beyond its mean in the $M_t/GI/s/0$ loss model with time-varying arrival rate was studied by Davis et al. (1995). For more on the Erlang-A model and generalizations, see Brandt and Brandt (1999, 2002), Garnett et al. (2002), Mandelbaum and Zeltyn (2004), Whitt (2004, 2005a,b,c) and references therein.

Here is how the rest of this paper is organized: In Section 2 we start by applying the QED diffusion approximation in Garnett et al. (2002) to investigate the sensitivity. In Section 3 we apply the alternative ED fluid approximation from Whitt (2004, 2005b) to gain further insights. In Section 4 we conduct numerical experiments, applying the algorithm in Whitt (2005a) to calculate the elasticities in a range of cases. There we demonstrate that the scaling discussed in previous sections indeed provides valuable insight. In Section 5 we conduct subsequent experiments to study the sensitivity of state-dependent abandonments to the total-abandonment-rate function for large queue lengths. We show that performance tends to be quite insensitive to such changes as well.

2. Insights from the QED Many-Server Heavy-Traffic Limit

In this section we apply a diffusion approximation to investigate the sensitivity of the Erlang-A model to the model parameters: the arrival rate λ, the service rate μ, the number of servers s and the individual customer abandonment rate θ. Specifically, we apply the diffusion approximation obtained by Garnett et al. (2002) via the many-server heavy-traffic limit in the QED limiting regime, which is also known as the Halfin-Whitt limiting regime, because corresponding results for the Erlang-C model (without abandonments) were previously obtained by Halfin and Whitt (1981).

In the QED limiting regime, the arrival rate, λ, and the number of servers, s, are allowed to increase toward infinity, with the mean service time $1/\mu$ held fixed, so that the traffic intensity $\rho \equiv \lambda/s\mu$ approaches 1 and

$$(1-\rho)\sqrt{s} \to \beta \quad \text{for} \quad -\infty < \beta < \infty \; . \tag{2.1}$$

From a practical perspective, this means that λ and s both should be large and that λ should not be too different from s. In particular, the difference $(\lambda/\mu) - s$ should be of order $O(\sqrt{s})$. The limiting constant β in (2.1) is an indicator of the quality of service (QOS), capturing the impact of all parameters. The QOS improves as β increases.

As is usually the case with stochastic-process limits (e.g., see Section 5.5 of Whitt (2002)), the scaling leading to the stochastic-process limit is the most important part, assuming that the conditions of the limiting regime indeed prevail. From the scaling alone we will see that the elasticities of all the standard performance measures with respect to λ, μ and s (regarding s as a continuous variable) are all of order $O(\sqrt{s})$ as $s \to \infty$, whereas the elasticities with respect to the abandonment rate θ are all of order $O(1)$. The practical implication is that the performance in the QED regime is substantially less sensitive to small percentage changes in θ than to small percentage changes in the other parameters.

The importance of the QED many-server limiting regime specified by (2.1) is highlighted by the fact that the probability of delay approaches a limit strictly between 0 and 1 as $s \to \infty$ if and only if the limit in (2.1) holds; see Theorem 4 of Garnett et al. (2002). Let W denote the steady-state waiting time (with dependence on the parameters suppressed in the notation). If (2.1) holds, then

$$P(W > 0) \to w(-\beta, \sqrt{\mu/\theta}) \; , \tag{2.2}$$

where

$$w(x,y) \equiv \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1} , \tag{2.3}$$

and h is the standard-normal hazard function, defined by

$$h(x) \equiv \frac{\phi(x)}{\Phi^c(x)} \equiv \frac{\phi(x)}{\Phi(-x)}, \tag{2.4}$$

with ϕ being the probability density function (pdf), Φ the associated cumulative distributiuon function (cdf) and $\Phi^c \equiv 1 - \Phi$ the associated complementary cdf (ccdf) of a standard (mean 0 and variance 1) normal random variable.

The QED approximation for the probability of delay is obtained by replacing the limits in (2.1) and (2.2) by equality; i.e.,

$$P(W > 0) \approx w(-\beta, \sqrt{\mu/\theta}) \quad \text{where} \quad \beta = (1-\rho)\sqrt{s} \; ; \tag{2.5}$$

see Section 5.2 of Garnett et al. (2002). From a practical perspective, (2.5) provides a valuable simplification, because the two parameters λ and s have been replaced by the single parameter β.

From Section 5.2 of Garnett et al. (2002), we also obtain the following approximation for the conditional abandonment probability:

$$P(Ab|W > 0) \approx \frac{h(\beta\sqrt{\mu/\theta} + 1/\sqrt{s\mu/\theta}) - h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + 1/\sqrt{s\mu/\theta})} , \tag{2.6}$$

where h is again the standard-normal hazard function in (2.4). Note that the parameters μ and θ appear in (2.5) and (2.6) only via $\sqrt{\mu/\theta}$.

We now simplify the approximation in (2.6). For that purpose, it is convenient to define a family of functions associated with the normal ccdf Φ^c. For a real-valued function of a real variable, f, let f' denote its derivative. Then let

$$\begin{aligned} h_0 &\equiv \Phi^c, \\ h \equiv h_1 &\equiv \frac{-h_0'}{h_0} = \frac{\phi}{\Phi^c}, \\ h_k &\equiv \frac{h_{k-1}'}{h_{k-1}} \quad \text{for} \quad k \geq 2 \; . \end{aligned} \tag{2.7}$$

We apply a one-term Taylor-series expansion to obtain the following asymptotically equivalent (as $s \to \infty$) version of (2.6):

$$P(Ab|W > 0) \approx \frac{h_2(\beta\sqrt{\mu/\theta})}{\sqrt{s\mu/\theta}} , \tag{2.8}$$

where h_2 is defined in (2.7). In particular, the two approximations approach a common limit as $s \to \infty$ after multiplying by $\sqrt{s}$.

The family of functions in (2.7) is also convenient to express the elasticities. For example, combining (2.7) and (2.8), we obtain an expression for the arrival-rate elasticity of the conditional probability of abandonment given that a customer is delayed, namely,

$$\mathcal{E}(P(Ab|W > 0), \lambda) \approx \lambda h_3(\beta\sqrt{\mu/\theta})\frac{\partial \beta}{\partial \lambda} = -h_3(\beta\sqrt{\mu/\theta})\frac{\lambda}{\mu\sqrt{s}} \ . \tag{2.9}$$

Given results for $P(W > 0)$ and $P(Ab|W > 0)$, we obtain results for related quantities through the exact relations

$$\begin{aligned} P(Ab) &= P(W > 0)P(Ab|W > 0), \\ EW &= m_a P(Ab), \\ EQ &= \lambda EW \ . \end{aligned} \tag{2.10}$$

The following elementary proposition about elasticities explains the consequences of the relations in (2.10). It also shows that it does not matter whether we work with the given parameters or their reciprocals; e.g., we could work with either the abandonment rate θ or the mean time to abandon, $m_a \equiv 1/\theta$. Let $(f \circ g)(x) \equiv f(g(x))$ denote function composition.

Proposition 1. (basic elasticity properties)

Let f and g be positive differentiable functions of a real variable x and let c be a real number.

(a) $\mathcal{E}(cf, x) = \mathcal{E}(f, x)$ and $\mathcal{E}(x^c f, x) = \mathcal{E}(f, x) - c$.

(b) If $f(x) = cx$, then $\mathcal{E}(f, x) = 1$.

(c) $\mathcal{E}(fg, x) = \mathcal{E}(f, x) + \mathcal{E}(g, x)$.

(d) $\mathcal{E}(f \circ g, x) = \mathcal{E}(f, g)\mathcal{E}(g, x)$.

(e) If $g(x) = f(1/x)$, then $\mathcal{E}(g, x) = -\mathcal{E}(f, x)$.

(f) $\mathcal{E}(1/f, x) = -\mathcal{E}(f, x)$.

We now investigate how the diffusion approximations for the elasticities behave as $s \to \infty$ in the QED regime specified by (2.1). For that purpose, the key function in (2.5) is not w, but β: The key is the way β depends on the parameters λ and s, with the understanding that $\lambda/s \to \mu$ as $s \to \infty$. The performance measures are functions of β, but since β is not necessarily positive we cannot always apply Proposition 1 (d). Instead, we apply

$$\mathcal{E}(f \circ \beta, \eta) = \mathcal{E}(f, \beta)\eta\beta'_{\eta} \ , \tag{2.11}$$

where η is the parameter of interest.

For a differentiable real-valued function of two real variables, $g \equiv g(\beta, \gamma)$, let the partial derivative with respect to β, $\partial g / \partial \beta$, be denoted by g'_β. Note that

$$\begin{aligned} \beta'_\lambda &= -\frac{1}{\mu\sqrt{s}}, \quad \beta'_\mu = \frac{\lambda}{\mu^2\sqrt{s}}, \\ \beta'_s &= \frac{1}{2\sqrt{s}} + \frac{\lambda}{2\sqrt{\mu} s^{3/2}} \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} \lambda\beta'_\lambda &= -\mu\beta'_\mu = -\frac{\lambda}{\mu\sqrt{s}}, \\ s\beta'_s &= \frac{\sqrt{s}}{2} + \frac{\lambda}{2\sqrt{\mu}\sqrt{s}} \quad \text{and} \quad \theta\beta'_\theta = 0 \ . \end{aligned} \tag{2.13}$$

We say that two real-valued functions of a real variable, f and g, are asymptotically equivalent (at $+\infty$), and write $f \sim g$ as $x \to \infty$, if $f(x)/g(x) \to 1$ as $x \to \infty$. As a consequence of (2.13), in the QED limiting regime specified by (2.1) we have

$$\begin{aligned} \lambda\beta'_\lambda = -\mu\beta'_\mu &\sim -\frac{\sqrt{s}}{\mu} \quad \text{as} \quad s \to \infty, \\ s\beta'_s &\sim \frac{(1 + (1/\sqrt{\mu}))}{2}\sqrt{s} \quad \text{as} \quad s \to \infty \ . \end{aligned} \tag{2.14}$$

The asymptotic relations for the function β in (2.14) explain the asymptotic form of the elasticities of the performance measures with respect to the parameters λ, μ and s. When we let $\mu = 1$, all three are asymptotically equivalent except for the sign.

To be more concrete, we give additional details. We state results for an arbitrary function with certain properties. The assumed properties cover the diffusion approximations of all the performance measures above: $P(W > 0)$ in (2.5), $P(Ab|W > 0)$ in (2.8), $P(Ab)$ in (2.10), EW in (2.10) and EQ in (2.10). Below we regard s as a real variable, not restricted to integer values. We now state a general proposition, whose proof follows from elementary calculus, and will thus be omitted.

Proposition 2. (general form of elasticities in the QED limiting regime)

Consider differentiable real-valued functions of several real variables:

$$\begin{aligned} f &\equiv f(\beta, \gamma) > 0, \\ g &\equiv f/s^p \quad \textit{for} \quad p \geq 0, \\ \beta &\equiv \beta(\lambda, s, \mu) \equiv \left(1 - \frac{\lambda}{s\mu}\right)\sqrt{s}, \\ \gamma &\equiv \gamma(\theta, \mu) \ , \end{aligned} \tag{2.15}$$

where it is understood that the function f depends on the variables λ and s only through β, while the function γ does not depend on them at all. Then the elasticities take the form

$$\begin{aligned}
\mathcal{E}(g,\lambda) &= C_1 \frac{\lambda}{\sqrt{s}} \sim C_1\sqrt{s} \quad as \quad s \to \infty, \\
\mathcal{E}(g,\mu) &= -C_1 \frac{\lambda}{\sqrt{s}} + C_2 \sim -C_1\sqrt{s} \quad as \quad s \to \infty, \\
\mathcal{E}(g,s) &= C_3\sqrt{s} + C_4 \frac{\lambda}{\sqrt{s}} - p \sim (C_3 + C_4)\sqrt{s} \quad as \quad s \to \infty, \\
\mathcal{E}(g,\theta) &= C_5 \ , \qquad (2.16)
\end{aligned}$$

where C_i are functions of μ and θ, but not λ or s. In particular,

$$\begin{aligned}
C_1 &= \frac{g'_\beta}{g}\beta'_\lambda\sqrt{s} = -\frac{g'_\beta}{\mu g}, \qquad C_2 = \frac{g'_\gamma}{g}\mu\gamma'_\mu, \\
C_3 &= \frac{g'_\beta}{2g}, \qquad C_4 = \frac{g'_\beta}{2\mu g}, \qquad C_5 = \frac{g'_\gamma}{g}(\theta\gamma'_\theta) \ . \qquad (2.17)
\end{aligned}$$

If $\mu = 1$, then $C_3 + C_4 = -C_1$, and

$$\mathcal{E}(g,\lambda) \sim -\mathcal{E}(g,\mu) \sim -\mathcal{E}(g,s) \sim C_1\sqrt{s} \quad as \quad s \to \infty \ . \qquad (2.18)$$

Corollary 2.1. (consequences for performance measures)

Assume that the QED many-server heavy-traffic scaling in (2.1) *holds with $\mu = 1$. Let g denote the diffusion approximation for one of the performance measures $P(W > 0)$, $P(Ab|W > 0)$, $P(Ab)$, EW or EQ, combining* (2.5), (2.8) *and* (2.10)*. Then*

$$\mathcal{E}(g,\lambda) = -\mathcal{E}(g,\lambda^{-1}) \sim -\mathcal{E}(g,\mu) = \mathcal{E}(g,\mu^{-1}) \sim -\mathcal{E}(g,s) = O(\sqrt{s}) \quad as \quad s \to \infty \ , \qquad (2.19)$$

while

$$\mathcal{E}(g,\theta) = -\mathcal{E}(g,\theta^{-1}) = O(1) \quad as \quad s \to \infty \ . \qquad (2.20)$$

Corollary 2.1 substantiates the claim made earlier: When the number of servers gets large in the Erlang-A model, operating in the QED regime, the performance measures tend to be highly sensitive to changes in the arrival rate, the service rate or the number of servers, but relatively insensitive to changes in the abandonment rate, as measured by elasticities.

The right pictures should tell the story. However, plotting is a bit subtle. For example, it is difficult to see the effect of changing arrival rate by simply plotting the performance as a function of the arrival rate. That is so, because by using the arrival rate as the independent variable, we automatically fix the scale. If we instead plot the performance as a function of the traffic intensity, then the independent variable corresponds to *percentage change in arrival*

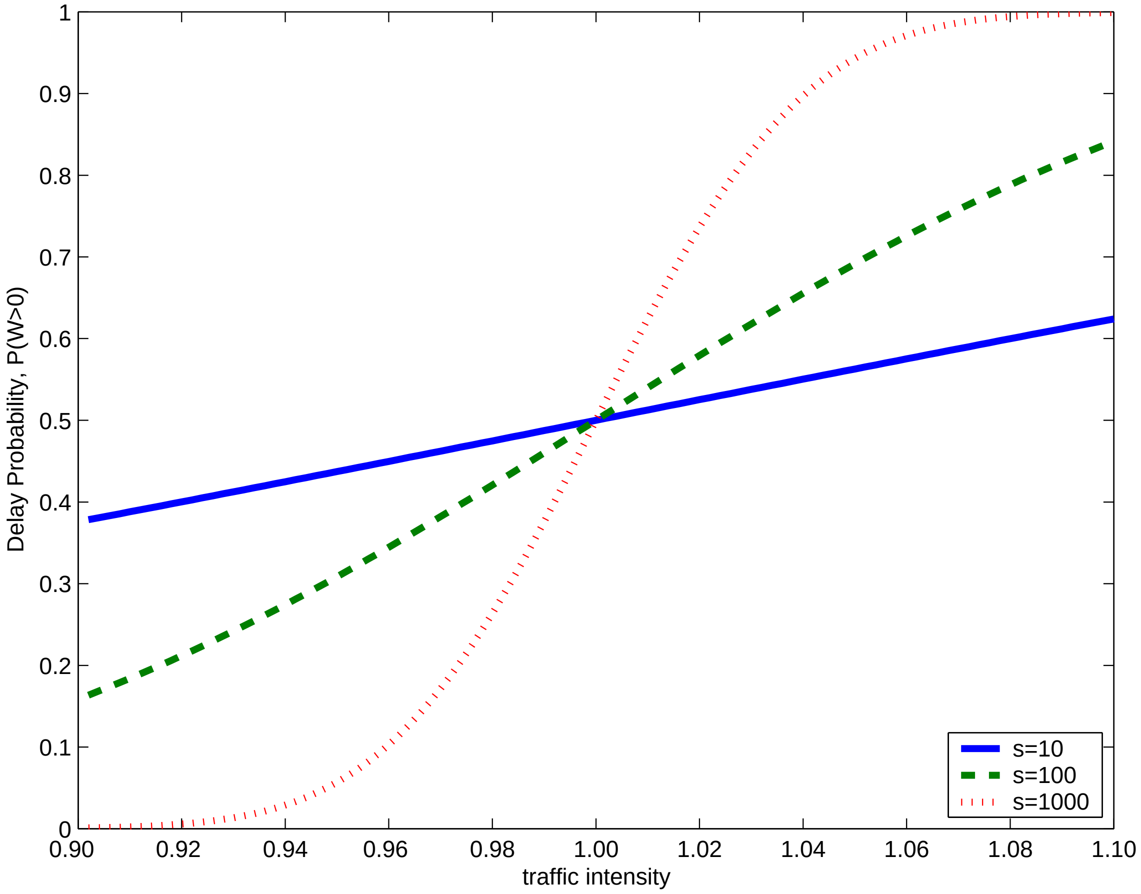

Figure 1: The diffusion approximation for the delay probability, $P(W > 0)$, in (2.5) as a function of the traffic intensity ρ for three different numbers of servers: $s = 10$, $s = 100$ and $s = 1000$. The other parameters are fixed at $\theta = \mu = 1$. Changes in ρ correspond to percentage changes in the arrival rate.

rate. To illustrate, we plot the diffusion approximation for the delay probability, $P(W > 0)$, in (2.5) as a function of the traffic intensity in Figure 1.

Figure 1 shows the growing sensitivity to the arrival rate that we have previously described in other ways. For example, the slope at $\rho = 1.00$, where $P(W > 0) \approx 0.50$ in all three cases, is 0.125, 0.400 and 1.25, respectively, when $s = 10$, $s = 100$ and $s = 1000$. The sensitivity can be understood from the basic QED scaling in (2.1) in the following way: For any given s, the relevant values of λ are $s + O(\sqrt{s})$. Thus λ is of order s, but changes in the delay probability, going from 0 to 1, take place over an interval of length $O(\sqrt{s})$. Consequently, the elasticity should be of order $O(\sqrt{s})$, because the derivative is roughly of order $O(1/\sqrt{s})$, the performance measure itself is roughly of order $O(1)$ and the parameter λ is of order $O(s)$.

Remark 2.1. *Contrast with Scaling in Direct Asymptotics.* It is useful to contrast the asymptotics of the elasticities with the asymptotics of the performance measures themselves. In the QED limiting regime, specified by (2.1), the performance measures themselves have limits with different scaling. In particular, in the QED regime, the following scaled performance measures converge to finite positive limits:

$$P(W > 0), \quad \sqrt{s}P(Ab), \quad \sqrt{s}EW, \quad \text{and} \quad \frac{EQ}{\sqrt{s}} \ . \tag{2.21}$$

The elasticities have meaning independent of this scaling. The elasticities being of order $O(\sqrt{s})$ implies that the sensitivity of performance to the parameters is growing as s increases, regardless of the direct scaling of the performance measures: An $x\%$ change in the parameter produces an increasing percentage change in the performance measure as s increases, given the QED condition (2.1).

Remark 2.2. *The Erlang B and C Models.* The results in this section also apply to the Erlang B and C models as special cases. We obtain the Erlang-B model from the Erlang-A model if we let $\theta \to \infty$; we obtain the Erlang-C model from the Erlang-A model if we let $\theta \to 0$, assuming that $\beta > 1$ in (2.1). When we let $\theta \to \infty$ to approach the Erlang-B model, the abandonment becomes the blocking, and the delay probability $P(W > 0)$ approaches the blocking probability $P(Bl)$; i.e., as $\theta \to \infty$,

$$P(W > 0) \approx w(-\beta, \sqrt{\mu/\theta}) \to [1 + (1/h(-\beta))]^{-1} \approx P(Bl) \ . \tag{2.22}$$

On the other hand, as $\theta \to 0$,

$$P(W > 0) \approx w(-\beta, \sqrt{\mu/\theta}) \to [1 + (\beta/h(-\beta))]^{-1} \ . \tag{2.23}$$

In these two cases, the two limits – on s and on θ – can be done in either order, although we do not verify that here. Direct QED many-server heavy-traffic limits for these two special cases are established in Erlang (1924), Jagerman (1974), Srikant and Whitt (1996) and Halfin and Whitt (1981). Then the abandonment rate θ ceases to be a relevant parameter. We deduce, either directly or by taking limits on θ of quantities here, that the arrival-rate, service-rate and number-of-server elasticities of the basic performance measures are again of order $O(\sqrt{s})$ as $s \to \infty$, assuming that (2.1) holds (with $\beta > 0$ for the Erlang-C model).

Remark 2.3. *Iterated Limits.* In this section we considered approximations generated from the heavy-traffic limits. Then we consider the derivatives of those approximations with respect to the parameters. We thus consider two limits, first, letting $s \to \infty$ with the associated condition, (2.1), and second, we take the derivative, which is tantamount to letting $h \to 0$ in the difference approximation in (1.3). It remains to interchange the order of the limits; i.e., it remains to consider the limits in the order $\lim_{s\to\infty} \lim_{h\to 0}$ instead of the order $\lim_{h\to 0} \lim_{s\to\infty}$. That is, it remains to establish heavy-traffic limits for the derivatives and elasticities themselves. It is intuitively clear that such heavy-traffic limits should be valid in the setting of the $M/M/s+M$ model, but it remains to provide proofs.

3. Insights from the ED Fluid Limit

In this section we apply the many-server heavy-traffic fluid limit in the ED limiting regime in order to obtain additional insights into the sensitivity of performance to model parameters. Here we draw upon Whitt (2004). The key theoretical results for the Erlang-A model follow from more general limits for state-dependent Markovian queues in Mandelbaum and Pats (1995). The fluid approximation in Whitt (2005b) makes it possible to perform similar analyses for the more general $G/GI/s+GI$ model.

In the ED many-server heavy-traffic limiting regime, we again let $\lambda \to \infty$ and $s \to \infty$, but now we let the traffic intensity approach a finite limit greater than 1. Here we will fix the traffic intensity, letting

$$\rho = \frac{\lambda}{s\mu} > 1 \ . \tag{3.1}$$

Because of the abandonments, a proper steady-state distribution exists for all $\rho > 1$. Now ρ plays the role of the QOS-parameter β in (2.1). Unlike β, the quality of service gets worse as ρ increases.

In the ED regime, there is both a fluid limit and a refined diffusion limit. Here we will focus on the elementary fluid limit. The ED fluid approximations for the basic performance measures are

$$\begin{aligned} P(W>0) &\approx 1, \\ P(Ab) \approx P(Ab|W>0) &\approx 1-\rho^{-1}, \\ EW = P(Ab)/\theta &\approx (1-\rho^{-1})/\theta, \\ EQ = \lambda EW &\approx (\lambda - s\mu)/\theta \ . \end{aligned} \tag{3.2}$$

Paralleling (2.13), from (3.1) we obtain

$$\lambda \rho'_\lambda = -\mu \rho'_\mu = -s \rho'_s = \rho \ . \tag{3.3}$$

Paralleling Proposition 2 and Corollary 2.1, we have

Proposition 3. (elasticities in the ED limiting regime)

In the ED limiting regime specified by (3.1)*, the fluid approximations in* (3.2) *satisfy*

$$\mathcal{E}(P(Ab),\lambda) \sim -\mathcal{E}(P(Ab),\mu) \sim -\mathcal{E}(P(Ab),s) \rightarrow \frac{1}{\rho-1} \quad as \quad s \rightarrow \infty \tag{3.4}$$

and $\mathcal{E}(P(Ab),\theta) = 0$*;*

$$\mathcal{E}(EQ,\mu) \sim \mathcal{E}(EQ,s) \sim \mathcal{E}(EW,\mu) \sim \mathcal{E}(EQ,s) \rightarrow -\frac{1}{\rho-1}; \tag{3.5}$$

$$\mathcal{E}(EW,\theta) = -\mathcal{E}(EW,1/\theta) = \mathcal{E}(EQ,\theta) = -\mathcal{E}(EQ,1/\theta) \rightarrow -1 \ ; \tag{3.6}$$

and

$$\mathcal{E}(EQ,\lambda) = 1 + \mathcal{E}(EW,\lambda) \rightarrow 1 + \frac{1}{\rho-1} = \frac{\rho}{\rho-1} \ . \tag{3.7}$$

From Proposition 3, we see that again the sensitivity of performance is much greater for the parameters λ, μ and s than for θ. However, the QED and ED stories are quite different: In the QED regime, the arrival-rate, service-rate and number-of-server elasticities are all of order $O(\sqrt{s})$ as $s \rightarrow \infty$, whereas in the ED regime they are all of order $O(1)$ as $s \rightarrow \infty$. That is not hard to understand, because the two regimes are quite different: In the QED regime, the traffic intensity ρ is much closer to the critical value $\rho = 1$. From Proposition 3, we also see that the arrival-rate, service-rate and number-of-server elasticities are greatest when ρ is close to the critical value $\rho = 1$; the elasticities decrease as ρ increases above that critical value.

Remark 3.1. *Contrast with the Standard Single-Server Queue.* It is well known that the many-server queue is quite different from the standard single-server queue. One way to see the difference is to look at the elasticities. Hence we now briefly discuss the standard $M/M/1$ queue. Sensitivity of the general single-server queue, $G/G/1/C$, to the key parameters was studied previously via the Brownian heavy-traffic approximation in Section 9 of Berger and Whitt (1992).

In the $M/M/1$ queue, the formulas for the basic performance measures are

$$P(W > 0) = \rho, \quad EQ = \lambda EW = \frac{\rho^2}{1-\rho} \; . \tag{3.8}$$

Since $\lambda \rho'_{\lambda} = -\mu \rho'_{\mu} = \rho$, we have

$$\mathcal{E}(EQ, \lambda) = -\mathcal{E}(EQ, \mu) = -\mathcal{E}(EW, \mu) = \frac{\rho}{1-\rho} \sim \frac{1}{1-\rho} \tag{3.9}$$

and

$$\mathcal{E}(EW, \lambda) = \mathcal{E}(EQ, \lambda) + \frac{1}{\lambda} = \frac{\rho}{1-\rho} + \frac{1}{\lambda} \sim \frac{1}{1-\rho} \tag{3.10}$$

as $\rho \rightarrow 1$. From (3.9) and (3.10), we see that the exact $M/M/1$ elasticities behave much like the fluid approximations for the elasticities in the $M/M/s+M$ model with $\rho > 1$. However, here we have $\rho < 1$ instead of $\rho > 1$. In both cases, the absolute values of the arrival-rate and service-rate elasticities are of the form $1/|1-\rho|$.

4. Numerical Calculations

In this section we show that it is also possible to calculate the *exact* values of the derivatives and the elasticities for the Erlang-A model by exploiting the exact numerical algorithm in Whitt (2005a). That algorithm was primarily intended to serve as an approximation for the more general $M/GI/s/r+GI$ model, but it also is yields an exact calculation for the $M/M/s/r+M$ special case. (Here we let the finite waiting room r be sufficiently large that the blocking is negligible.) We start with a base case, in which $\lambda = s = 100$ and $\theta = \mu = 1$, and then consider several variations of that base case, aiming to substantiate the main conclusions of Sections 2 and 3, and show the impact of key parameters, such as s and θ. Throughout all these numerical experiments we fix the service rate at $\mu = 1$ (which is without loss of generality, because we are free to choose the measuring units for time).

We calculate the derivatives by using the elementary method of finite differences, as indicated in (1.3). Given that the function f is indeed differentiable, the finite-difference approximation is asymptotically correct as $h \rightarrow 0$. Given that we can calculate $f(\lambda)$ with high accuracy, we should have no difficulty calculating the derivative. We can verify accuracy by performing successive calculations with different intervals h.

	elasticities				scaled second derivatives			
h	$P(W > 0)$	$P(Ab)$	EW	$SD(W)$	$P(W > 0)$	$P(Ab)$	EW	$SD(W)$
10^{-1}	0.2255	-0.2378	0.7384	0.5538	-0.1982	0.2091	-0.2248	-0.2483
10^{-2}	0.2355	-0.2483	0.7492	0.5660	-0.2319	0.2445	-0.2472	-0.2802
10^{-3}	0.2365	-0.2494	0.7503	0.5673	-0.2358	0.2487	-0.2497	-0.2838
10^{-4}	0.2366	-0.2496	0.7504	0.5674	-0.2362	0.2491	-0.2500	-0.2842
10^{-5}	0.2366	-0.2496	0.7504	0.5674	-0.2362	0.2493	-0.2500	-0.2842
10^{-6}	0.2366	-0.2496	0.7504	0.5674	-0.2349	0.2646	-0.2496	-0.2859

Table 1: The impact of the interval h upon estimates of the mean-time-to-abandon elasticities, $\mathcal{E}(f, m_a = 1/\theta)$, and associated scaled second derivatives, $\mathcal{S}(f, m_a)$, of several performance measures in the base case with $\lambda = s = 100$ and $\theta = \mu = 1$.

We also approximate the second derivative by

$$f''(\lambda) \approx \frac{f(\lambda + h) - 2f(\lambda) - f(\lambda - h)}{h^2} \ , \tag{4.1}$$

again for small positive h. With standard double precision, the second-derivative calculations will necessarily be less accurate. However, it is easy to determine the accuracy after performing the calculations by just repeating the calculation for several values of h.

Given the estimates of the derivatives, we scale to calculate the associated elasticities, as indicated in (1.2). We also calculate scaled second derivatives. We divide the second derivative by the performance measure itself and multiply by the *square* of the parameter. That coincides with the product of the arrival-rate elasticities of f and f'; i.e., the scaled second-derivative with respect to the arrival rate is

$$\mathcal{S}(f, \lambda) \equiv \frac{\lambda^2 f''_\lambda}{f} = \left(\frac{\lambda f''_\lambda}{f'_\lambda}\right)\left(\frac{\lambda f'_\lambda}{f}\right) = \mathcal{E}(f', \lambda)\mathcal{E}(f, \lambda) \ . \tag{4.2}$$

If, instead, we want the arrival-rate elasticity of f', then it can easily be obtained as the ratio $\mathcal{E}(f', \lambda) = \mathcal{S}(f, \lambda)/\mathcal{E}(f, \lambda)$.

We illustrate how the calculations perform by displaying the impact of the interval h upon the elasticities and scaled second derivatives for one case in Table 1. Specifically, in Table 1 we consider derivatives with respect to the mean time to abandon, $m_a = 1/\theta$, and thus display the mean-time-to-abandon elasticities and associated scaled second derivatives. (Recall that $\mathcal{E}(f, m_a) = -\mathcal{E}(f, \theta)$.) We consider the standard deviation of the steady-state waiting time, $SD(W)$, as well as $P(W > 0)$, $P(Ab)$ and EW in the base case with $\lambda = s = 100$ and $\theta = \mu = 1$.

From Table 1, we see that we obtain three-digit precision for all the elasticities using $h = 10^{-3}$ and at least four-digit precision using any value of h ranging from $h = 10^{-4}$ to

parameters		performance measures					
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EN	$SD(Q)$	$SD(N)$	$SD(W)$
10.0	10	0.320	0.186	8.33	0.53	2.08	0.041
	100	0.262	0.0605	94.6	1.50	6.8	0.014
	1000	0.247	0.0192	982.8	4.58	21.9	0.0045
1.0	10	0.542	0.125	10.0	1.96	3.16	0.183
	100	0.513	0.0399	100.0	5.95	10.0	0.058
	1000	0.504	0.0126	1000.	18.6	31.6	0.0185
0.1	10	0.779	0.0605	15.4	6.4	7.1	0.63
	100	0.766	0.0192	117.2	20.0	22.2	0.198
	1000	0.762	0.00605	1055.	62.8	69.9	0.063

Table 2: Several performance measures in the Erlang A model, as a function of the abandonment rate, θ and the number of servers, s, when $\lambda = s$ and $\mu = 1$.

$h = 10^{-6}$. For the scaled second derivatives, we also obtain three-digit precision using $h = 10^{-3}$, but we do not do much better as h increases, losing precision when h is very small. That should not be surprising, because there is division by h^2 in (4.1), and we have used standard double precision in matlab. Overall, we regard the simple finite-difference approach as providing ample precision for engineering purposes.

Our first set of experiments consists of 9 cases, with 3 values of s and 3 values of θ. We consider $s = 10$, $s = 100$ and $s = 1000$; and we consider $\theta = 10$, $\theta = 1$ and $\theta = 0.1$. Otherwise, we let $\lambda = s$, so that we are in the center of the QED limiting regime in (2.1) with $\beta = 0$, where $P(W > 0) \to 1/2$ as $s \to \infty$. The values of several basic performance measures in these 9 cases are given in Table 2. As noted in (2.10), the mean waiting time, EW, and the mean queue length, EQ, are constant multiples of the abandonment probability, $P(Ab)$, so they are omitted from Table 2. We include the expected steady-state number of customers in the system (waiting or in service), EN, as well as the standard deviations of the steady-state queue length, $QD(Q)$, number in system, $SD(N)$, and the waiting time, $SD(W)$, as well as previously discussed performance measures.

From the QED many-server heavy-traffic limits, we know how these performance measures should be scaled by s in order for the scaled performance measures to be nearly independent of s. We present the corresponding scaled performance measures in Table 3. After scaling by s in the indicated manner, the performance measures in Table 3 are approximately independent of s. From Table 3 we see that the scaling by s is the dominant effect, substantiating conclusions of Garnett et al. (2002). Table 3 also shows the remaining impact of the parameter θ.

Next, in Table 4 we present the abandonment-rate elasticities of the performance measures

parameters		scaled performance measures					
θ	$\lambda = s$	$P(W > 0)$	$\sqrt{s}P(Ab)$	$(EN - s)/\sqrt{s}$	$SD(Q)/\sqrt{s}$	$SD(N)/\sqrt{s}$	$\sqrt{s}SD(W)$
10.0	10	0.320	0.59	-0.53	0.168	0.66	0.13
	100	0.262	0.61	-0.54	0.150	0.68	0.14
	1000	0.247	0.61	-0.54	0.145	0.69	0.14
1.0	10	0.542	0.40	0.00	0.62	1.00	0.58
	100	0.513	0.40	0.00	0.60	1.00	0.58
	1000	0.504	0.40	0.00	0.59	1.00	0.59
0.1	10	0.779	0.19	1.71	0.20	2.2	2.0
	100	0.766	0.19	1.72	0.20	2.2	2.0
	1000	0.762	0.19	1.74	0.20	2.1	2.0

Table 3: Scaled versions of the performance measures in Table 2.

parameters		performance measures						
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	$EQ\&EW$	EN	$SD(Q)$	$SD(N)$	$SD(W)$
10.0	10	-0.22	0.103	-0.90	-0.04	-0.57	-0.103	-0.74
	100	-0.33	0.119	-0.88	-0.013	-0.63	-0.089	-0.66
	1000	-0.37	0.120	-0.88	-0.004	-0.65	-0.084	-0.66
1.0	10	-0.21	0.25	-0.75	-0.125	-0.54	-0.27	-0.58
	100	-0.24	0.25	-0.75	-0.04	-0.56	-0.26	-0.57
	1000	-0.23	0.37	-0.64	-0.011	-0.42	-0.18	-0.43
0.1	10	-0.108	0.38	-0.62	-0.26	-0.49	-0.42	-0.51
	100	-0.116	0.41	-0.62	-0.11	-0.50	-0.42	-0.50
	1000	-0.119	0.38	-0.62	-0.04	-0.50	-0.42	-0.50

Table 4: The abandonment-rate elasticities, $\mathcal{E}(f, \theta)$, of several performance measures (the f) in the setting of Table 2.

in Table 2 (without any additional scaling). Consistent with the conclusions in previous sections, these abandonment-rate elasticities are not large. Indeed, all are less than 1. So that an x percent change in the abandonment rate produces less than an x percent change in any of these performance measures.

Consistent with Sections 2 and 3, the abandonment-rate elasticities in Table 4 tend to be independent of s. The one exception is the mean number in system EN. As indicated in Table 3, the appropriate scaling for N is $(N - s)/\sqrt{s}$. Since we have not used that scaling in Table 4, it should not be surprising that we do not see elasticities of EN independent of s. Otherwise, the abandonment-rate elasticities are both small and largely independent of s. A similar conclusion holds for the scaled second derivatives, as shown in Table 5. (There the parameter has been taken to be the mean time to abandon, $m_a = 1/\theta$).

The arrival-rate and service-rate elasticities of the same performance measures are shown

parameters		performance measures						
m_a	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EQ & EW	EN	$SD(Q)$	$SD(N)$	$SD(W)$
0.1	10	-0.13	0.059	-0.15	-0.016	-0.24	-0.035	-0.30
	100	-0.24	0.086	-0.15	-0.0065	-0.25	-0.032	-0.26
	1000	-0.27	0.089	-0.15	-0.0020	-0.26	-0.030	-0.26
1.0	10	-0.20	0.24	-0.25	-0.061	-0.27	-0.12	-0.29
	100	-0.24	0.25	-0.25	-0.020	-0.28	-0.10	-0.28
	1000	-0.28	0.19	-0.54	-0.006	-0.28	-0.20	-0.28
10.0	10	-0.14	0.48	-0.28	-0.13	-0.26	-0.20	-0.27
	100	-0.15	0.48	-0.28	-0.054	-0.26	-0.20	-0.27
	1000	-0.15	0.48	-0.28	-0.019	-0.26	-0.20	-0.27

Table 5: The scaled second derivative of several performance measures with respect to the mean time to abandon, $m_a = 1/\theta$, in the Erlang A model, as a function of the mean time to abandon and number of servers. The scaling is as in (4.2).

parameters		performance measures					
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EQ	$SD(Q)$	EW	$SD(W)$
10.0	10	0.73	0.76	1.08	0.54	0.76	0.32
	100	0.77	0.88	0.98	0.48	0.88	0.40
	1000	0.80	0.92	0.95	0.46	0.92	0.44
1.0	10	0.73	1.04	1.36	0.70	1.04	0.44
	100	0.78	1.19	1.29	0.62	1.19	0.54
	1000	0.80	1.23	1.27	0.60	1.23	0.57
0.1	10	0.73	2.02	2.34	1.23	2.02	1.01
	100	0.78	2.17	2.27	1.15	2.17	1.09
	1000	0.80	2.22	2.25	1.13	2.24	1.11

Table 6: The arrival-rate elasticities, $\mathcal{E}(f, \lambda)$, of several performance measures in the setting of Table 2. The arrival-rate elasticities have been scaled by dividing by $\sqrt{s}$.

in Tables 6 and 7. These elasticities have been divided by $\sqrt{s}$. From Tables 6 and 7, we see that the arrival-rate and service-rate elasticities indeed become of order $O(1)$ after the additional scaling by $\sqrt{s}$. In fact, mostly, the absolute values of the arrival-rate and service-rate elasticities fall between $0.25\sqrt{s}$ and $2.5\sqrt{s}$. A rough approximation for any one of these elasticities is simply $\sqrt{s}$.

Paralleling Table 5, we present the scaled second derivatives of the performance measures with respect to the service rate in Table 8. The scaled second derivatives have been further scaled by dividing by s. Thus again, through the second derivatives, we see the strong sensitivity of performance to the service rate as s increases. A similar story holds for the arrival rate.

The initial experiment had $\lambda = s$ in all cases. We now want to modify the base case in

parameters		performance measures				
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	$SD(Q)$	$SD(N)$	$SD(W)$
10.0	10	-0.63	-0.79	-0.37	-0.15	-0.41
	100	-0.74	-0.89	-0.42	-0.22	-0.44
	1000	-0.78	-0.92	-0.44	-0.24	-0.45
1.0	10	-0.66	-1.14	-0.52	-0.073	-0.60
	100	-0.75	-1.21	-0.56	-0.024	-0.59
	1000	-0.78	-1.24	-0.58	-0.008	-0.59
0.1	10	-0.70	-2.15	-1.08	-0.79	-1.17
	100	-0.77	-2.21	-1.10	-0.79	-1.13
	1000	-0.79	-2.22	-1.11	-0.79	-1.12

Table 7: The service-rate elasticities, $\mathcal{E}(f, \mu)$, of several performance measures in the setting of Table 2. (The service-rate elasticities of EQ and EW coincide with the displayed service-rate elasticity of $P(Ab)$.) The service-rate elasticities have been scaled by dividing by $\sqrt{s}$.

parameters		performance measures					
m_a	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EN	$SD(Q)$	$SD(N)$	$SD(W)$
0.1	10	0.25	0.47	-0.011	0.034	-0.106	0.062
	100	0.20	0.45	-0.175	-0.023	-0.019	-0.0052
	1000	0.19	0.45	-0.007	-0.042	0.015	-0.036
1.0	10	0.11	1.05	0.085	0.025	-0.089	-0.16
	100	0.039	1.01	0.0095	-041	-0.036	0.0013
	1000	0.015	1.01	0.001	-0.063	-0.012	-0.05
10.0	10	-0.37	4.52	1.66	0.37	0.31	0.79
	100	-0.51	4.46	0.67	0.30	0.32	0.43
	1000	-0.56	4.43	0.23	0.29	0.33	0.33

Table 8: The scaled second derivative of several performance measures with respect to the service rate in the Erlang A model, as a function of the mean time to abandon and number of servers. There is extra scaling: The scaled second derivatives have been divided by s.

	elasticities			scaled second derivatives		
perf. meas.	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$
$P(W > 0)$	0.18	0.24	0.17	-0.219	-0.24	-0.23
$P(Ab)$	-0.50	-0.25	-0.08	0.59	0.25	0.10
$E[Q]$	0.50	0.75	0.92	-0.40	-0.25	-0.05
$SD(Q)$	0.41	0.56	0.60	-0.35	-0.28	-0.26
$E[N]$	0.0083	0.04	0.10	-0.008	-0.02	-0.015
$SD[N]$	0.079	0.26	0.42	-0.067	-0.10	-0.12
$E[W]$	0.50	0.75	0.92	-0.40	-0.25	-0.05
$SD(W)$	0.42	0.57	0.62	-0.35	-0.28	-0.24
$P(W \leq 0.05)$	-0.034	-0.24	-0.87	0.037	0.228	1.09
$P(W \leq 0.1)$	-0.022	-0.19	-0.74	0.017	0.110	0.67
$P(W \leq 0.2)$	-0.004	-0.06	-0.36	-0.002	-0.046	-0.18
$P(W \leq 0.4)$	-0.000	-0.0005	-0.010	0.000	-0.0026	-0.06

Table 9: The mean-time-to-abandon elasticities, $\mathcal{E}(f, m_a = 1/\theta)$, and scaled second derivatives, $\mathcal{S}(f, m_a)$, of steady-state performance measures as a function of the arrival rate. The number of servers is $s = 100$ and the abandonment rate is $\theta = 1$.

another way. We now fix the number of servers at $s = 100$ and the abandonment rate at $\theta = 1$, and vary the arrival rate. We consider three possible arrival rates: $\lambda = 90$, $\lambda = 100$ and $\lambda = 110$. These case correspond to $\beta = 1$, $\beta = 0$ and $\beta = -1$ in the QED regime specified by (2.1). For these examples, we also consider the steady-state waiting-time distribution (for all customers).

We display elasticities and scaled second derivatives with respect to the mean time to abandon (reciprocal of the abandonment rate), the arrival rate and the service rate, respectively, in Tables 9, 10 and 11. We have not scaled any of these elasticities. As before, the mean-time-to-abandon elasticities of the performance measures considered previously are of order $O(1)$, in fact less than 1, while the arrival-rate and service-rate elasticities of these performance measures are of order $O(\sqrt{s}) = 10$. In Tables 9–11 we also consider the steady-state waiting-time distribution. The derivative is consistently small for larger arguments, when the probability is already close to 1.

Finally, we consider a numerical example to evaluate the fluid approximation in Section 3. We fix the traffic intensity at $\rho = 1.1$ and consider three different values of s: $s = 100$, $s = 400$ and $s = 1600$, with $\theta = \mu = 1$. The performance measures and arrival-rate elasticities are compared to the fluid approximations in Table 12. We consider the performance measures discussed in Section 3, namely, $P(W > 0)$, $P(Ab) = EW$, and $EQ/s\mu$. In these case, with ED scaling, we see that the fluid approximations in Section 3 do indeed tell the main story. We also

		elasticities			scaled second derivatives	
perf. meas.	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$
$P(W > 0)$	14.7	7.8	2.96	132.	-7.8	-32.6
$P(Ab)$	18.0	11.9	7.5	243.	76.	8.3
$E[Q]$	19.0	12.9	8.5	280.	244.	25.1
$SD(Q)$	10.3	-7.6	2.86	59.	-7.1	-19.0
$E[N]$	1.00	1.00	1.00	0.0004	0.0002	-0.0023
$SD[N]$	0.50	0.50	0.50	-0.23	-0.15	-0.030
$E[W]$	18.0	11.9	7.5	243.	76.	10.2
$SD(W)$	9.5	5.4	2.4	43.7	13.9	18.7
$P(W \leq 0.05)$	-1.3	-4.8	-9.9	-17.2	-18.8	55.6
$P(W \leq 0.1)$	-0.51	-2.6	-6.5	-9.0	-22.4	3.2
$P(W \leq 0.2)$	-0.048	-0.51	-2.2	-1.2	-8.5	-19.8
$P(W \leq 0.4)$	-0.0001	-0.0024	-0.04	-0.18	-0.03	-0.84

Table 10: Arrival-rate elasticities and scaled second derivatives as a function of the arrival rate. The number of servers is $s = 100$ and the abandonment rate is $\theta = 1$.

		elasticities			scaled second derivatives	
perf. meas.	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$	$\lambda = 90$	$\lambda = 100$	$\lambda = 110$
$P(W > 0)$	-14.5	-7.6	-2.79	155.	3.9	-25.5
$P(Ab)$	-18.5	-12.1	-7.6	294.	102.	24.0
$E[Q]$	-18.5	-12.1	-7.6	294.	102.	24.0
$SD(Q)$	-9.9	-7.6	-2.26	67.9	-4.1	-20.6
$E[N]$	-1.00	-0.90	-0.09	1.7	-0.95	0.30
$SD[N]$	-0.42	-0.24	-0.08	-1.7	-3.6	-2.29
$E[W]$	-18.5	-12.1	-7.6	294.	102.	24.0
$SD(W)$	-10.1	-5.9	-2.8	71.7	0.13	-11.7
$P(W \leq 0.05)$	1.3	4.9	9.6	-20.4	-25.8	41.3
$P(W \leq 0.1)$	0.55	2.7	6.5	-11.1	-27.7	-4.0
$P(W \leq 0.2)$	0.054	0.56	2.3	-1.6	-11.2	-24.5
$P(W \leq 0.4)$	0.0001	0.0028	0.05	-0.003	-0.11	-1.4

Table 11: Service-rate elasticities and scaled second derivatives as a function of the arrival rate. The number of servers is $s = 100$ and the abandonment rate is $\theta = 1$.

	performance				elasticities			
perf. meas.	$s = 100$	$s = 400$	$s = 1600$	$fluid$	$s = 100$	$s = 400$	$s = 1600$	$fluid$
$P(W > 0)$	0.842	0.975	1.0000	1.0000	2.96	1.25	0.0088	0.0000
$P(Ab) = EW$	0.0992	0.0914	0.0909	0.0909	7.49	9.67	9.9994	10.0000
$E[Q]/s\mu$	0.109	0.101	0.1000	0.1000	8.49	10.67	10.9994	11.0000
$SD(Q)/\sqrt{s\mu}$	0.9091	1.025	1.049	0.0000	2.85	1.57	0.5083	0.0000
$\sqrt{s\mu}SD(W)$	0.84	0.98	1.06	–	2.43	2.17	4.48	–

Table 12: Performance measures and arrival-rate elasticities as a function of the number of servers, s, when the traffic intensity is fixed at $\rho = 1.1$. The other parameters are $\theta = \mu = 1$. For comparison, the fluid approximation from Section 3 is given too.

consider scaled standard deviations, $SD(Q)/\sqrt{s\mu}$ and $\sqrt{s/\mu}SD(W)$. The fluid approximation for the queue length is deterministic, so the fluid approximation for the standard deviation of the queue length is simply 0. (That is not true for the waiting time because of the random delay experienced by abandoning customers.) The observed regular behavior after scaling by $\sqrt{s}$ reflects refined diffusion approximations stemming from stochastic-process limits in the ED regime, as in Whitt (2004). We do not elaborate here.

5. Sensitivity to Abandonment Rates at Large Queue Lengths

In Pierson and Whitt (2005), $M/GI/s+GI$ models are approximated by purely-Markovian $M/M/s+M(n)$ models, having state-dependent arrival rates, by using the known mean arrival rate and service rate, and by statistically fitting the total state-dependent arrival rate to observed abandonment rates. An initial estimate for the total arrival rate when there are k customers waiting in queue is the observed number of abandonments when there are k customers waiting in queue, divided by the total time during which there are k customers waiting in queue. Subsequently, refined estimates for the total-abandonment-rate function can be obtained by fitting functions of k to the data, e.g., quadratic functions.

In that work, we observed that the performance is relatively insensitive to the state-dependent total-abandonment rates for large queue sizes, especially for large queue sizes that rarely occur (and for which the statistical estimates are unreliable). Similar insights are contained in Mandelbaum and Zeltyn (2004). In support of that conclusion, in this section we investigate the sensitivity to total-abandonment rates at large queue lengths in the $M/M/s+M$ model.

In the $M/M/s+M$ model the total abandonment rate when there are k customers waiting in queue is exactly $k\theta$ for all $k \geq 0$. We investigate the sensitivity to the total abandonment rate for large queue lengths by constructing loose lower and lower bounds. (Thus we are considering big changes in the total-arrival-rate function in a certain part of its domain.) A lower bound for the steady-state queue-length distribution is obtained by choosing an upper bound on the total-abandonment-rate function; we consider the associated $M/M/s/r+M$ model with a finite waiting room of size r. The finite waiting room of size r is equivalent to having an infinite total abandonment rate when the number in queue exceeds r. An upper bound for the steady-state queue-length distribution is obtained by choosing a lower bound on the total-abandonment-rate function; we consider the case in which the total abandonment rate is held constant at $c\theta$ after reaching the level $k = c$; i.e., the total abandonment rate function is $\delta_k = (k \wedge c)\theta$. The performance of the $M/M/s + M$ system is bounded between these two bounding systems. The ordering of performance can be formalized by stochastic-comparison concepts; e.g., see Whitt (1981), p. 196 of Müller and Stoyan (2002) and references there.

In Tables 13 and 14 below we see how these bounding systems behave as functions of the waiting-room size r and the cutoff level c. In these tables we let $s = 100$ and $\theta = \mu = 1$. In Table 13 we let $\lambda = 102$, while in Table 14 we let $\lambda = 110$. In Tables 13 and 14 we consider the conditional expected waiting time, given that the customer is eventually served, $E[W|S]$, or given that the customer eventually abandons, $E[W|A]$. We also consider the associated conditional standard deviations.

For the smaller arrival rate, $\lambda = 102$ in Table 13, we see that the two bounds are essentially equal when $r = c = 40$. For the larger arrival rate, $\lambda = 110$ in Table 14, we see that the two bounds are essentially equal when $r = c = 50$. The performance is not affected much if the parameters r and c are much smaller. Thus we see that the performance is indeed primarily determined by the total-abandonment-rate function only for relatively small queue lengths.

6. Acknowledgment

The author was supported by National Science Foundation Grant DMS-02-2340.

$M/M/100+M$ model, with $\lambda = 102$ and $m_A = 1/\theta = 1.0$

	lower bound: reduce wait. spaces				*upper bound: fixed rate after c*			
	no. wait. spaces, r			*both*	*cutoff level, c*			
perf. measure	20	25	30	40	30	25	20	15
$P(Loss)$	0.0084	0.0032	0.0010	$< 10^{-4}$				
$P(W = 0)$	0.424	0.413	0.410	0.408	0.408	0.407	0.404	0.393
$P(Aban)$	0.043	0.047	0.049	0.050	0.050	0.050	0.050	0.049
$E[Q]$	4.35	4.80	5.00	5.09	5.11	5.16	5.36	6.09
$SD(Q)$	5.5	6.2	6.5	6.7	6.7	6.8	7.2	8.5
$E[N]$	101	102	102	102	102	102	102	103
$SD[N]$	9.2	9.7	10.0	10.1	10.1	10.2	10.5	11.6
$E[W\|S]$	0.042	0.047	0.048	0.049	0.049	0.050	0.052	0.060
$SD(W\|S)$	0.056	0.061	0.063	0.065	0.065	0.067	0.071	0.085
$E[W\|A]$	0.057	0.062	0.065	0.067	0.066	0.066	0.063	0.058
$SD(W\|A)$	0.046	0.051	0.054	0.056	0.055	0.054	0.050	0.044
$P(W \leq 0.1\|S)$	0.830	0.809	0.801	0.799	0.798	0.797	0.791	0.768
$P(W \leq 0.1\|A)$	0.829	0.793	0.776	0.768	0.768	0.771	0.783	0.825
$P(W \leq 0.2\|S)$	0.987	0.975	0.968	0.965	0.964	0.962	0.954	0.928
$P(W \leq 0.2\|A)$	0.993	0.983	0.976	0.971	0.972	0.976	0.986	0.996

Table 13: Lower and upper bounds on the exact steady-state performance in the $M/M/100+M$ model when the arrival rate is $\lambda = 102$ and the mean time to abandon is $m_A = 1/\alpha = 1$. The lower bounds reduce the number of waiting spaces, r, while the upper bounds make the total abandonment rate constant after a cutoff level, c.

$M/M/100 + M$ model, with $\lambda = 110$ and $m_A = 1/\theta = 1.0$

	lower bound: reduce wait. spaces				*upper bound: fixed rate after c*			
	no. wait. spaces, r			*both*	*cutoff level, c*			
perf. measure	25	30	40	50	40	30	25	20
$P(Loss)$	0.0144	0.0065	0.00078	$< 10^{-4}$				
$P(W = 0)$	0.171	0.163	0.159	0.158	0.158	0.157	0.154	0.144
$P(Aban)$	0.085	0.093	0.098	0.099	0.099	0.099	0.099	0.098
$E[Q]$	9.4	10.2	10.8	10.9	10.9	11.1	11.7	13.6
$SD(Q)$	7.5	8.2	8.9	9.1	9.1	9.3	10.3	12.8
$E[N]$	108	109	110	110	110	110	111	113
$SD[N]$	9.1	9.7	10.1	10.5	10.5	10.8	11.6	13.9
$E[W\|S]$	0.088	0.095	0.100	0.101	0.101	0.103	0.109	0.129
$SD(W\|S)$	0.073	0.079	0.084	0.085	0.086	0.090	0.099	0.126
$E[W\|A]$	0.073	0.079	0.084	0.085	0.085	0.084	0.081	0.076
$SD(W\|A)$	0.056	0.063	0.066	0.068	0.067	0.065	0.062	0.056
$P(W \leq 0.1\|S)$	0.584	0.557	0.542	0.541	0.541	0.537	0.527	0.492
$P(W \leq 0.1\|A)$	0.719	0.684	0.660	0.656	0.656	0.659	0.665	0.689
$P(W \leq 0.2\|S)$	0.920	0.887	0.865	0.862	0.862	0.857	0.839	0.781
$P(W \leq 0.2\|A)$	0.971	0.953	0.934	0.930	0.931	0.937	0.950	0.974

Table 14: Lower and upper bounds on the exact steady-state performance in the $M/M/100+M$ model when the arrival rate is $\lambda = 110$ and the mean time to abandon is $m_A = 1/\alpha = 1$. The lower bounds reduce the number of waiting spaces, r, while the upper bounds make the total abandonment rate constant after a cutoff level, c.

References

Berger, A. W. and W. Whitt. 1992. The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue. *J. Discrete Event Dynamic Systems* 2, 7–60.

Borst, S., A. Mandelbaum and M. I. Reiman. 2004. Dimensioning large call centers. *Operations research* 52, 17-34.

Brandt, A., M. Brandt. 1999. On a two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability* 1, 191–210.

Brandt, A., M. Brandt. 2002. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems* 41, 73–94.

Davis, J. L., W. A. Massey and W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Science* 41, 1107–1116.

Erlang, A. K. 1924. On the rational determination of the number of circuits. In *The Life and Works of A. K. Erlang*, E. Brockmeyer, H. L. Halstrom and A. Jensen (eds.), Danish Academy of Technical Sciences, 1948, 216–221.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Opns. Mgmt.* 5, 79-141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Opns. Mgmt.*, 4, 208-227.

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

Harel, A. 1990. Convexity properties of the Erlang loss formula. *Operations Research* 38, 499–505.

Harel, A. and P. H. Zipkin. 1987. Strong convexity results for queueing systems. *Operations Research* 35, 405–418.

Jagerman, D. L. 1974. Some properties of the Erlang loss function. *Bell System Tech. J.* 53, 525–551.

Jagers, A. A. and E. A. van Doorn. 1991. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review* 33, 281-282.

Kalashnikov, V. V. and S. T. Rachev. 1990. *Mathematical Methods for Construction of Queueing Models*, Wadsworth & Brooks/Cole.

Mandelbaum, A. and G. Pats. 1995. State-dependent queues: approximations and applications. In *Stochastic Networks*, IMA Volumes in Mathematics, F. P. Kelly and R. J. Williams, eds., Springer, 239–282.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/N + G$ queue. *OR Spectrum* 26, 377–411.

Müller, A. and D. Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*, Wiley.

Pierson, M. P. and W. Whitt. 2005. A statistically-fit Markovian approximation of a basic call-center model. Department of Industrial Engineering and Operations Research, Columbia University. In preparation.

Rachev, S. T. 1991. *Probability Metrics and the Stability of Stochastic Models*, Wiley.

Whitt. W. 1980. Continuity of generalized semi-Markov processes. *Math. Oper. Res.* 5, 494–501.

Whitt, W. 1981. Comparing counting processes and queues. *Adv. Appl. Prob.* 13, 207–220.

Whitt, W. 2002. *Stochastic-Process Limits*, Springer, New York.

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

Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Science*, to appear.

Whitt, W. 2005b. Fluid models for multi-server queues with abandonments. *Operations Research*, to appear.

Whitt, W. 2005c. Staffing a call center with uncertain arrival rate and absenteeism. Department of Industrial Engineering and Operations Research, Columbia University. (Submitted to *Management Science*.) Available at http://columbia.edu/∼ww2040.