Scaled control in the QED regime

Scaled control in the QED regime

A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and Jaron Sanders*

Department of Mathematics & Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

July 4, 2013

Abstract

We develop many-server asymptotics in the Quality-and-Efficiency-Driven (QED) regime for models with admission control. The admission control, designed to reduce the incoming traffic in periods of congestion, scales with the size of the system. For a class of Markovian models with this scaled control, we identify the QED limits for two stationary performance measures. We also derive corrected QED approximations, generalizing earlier results for the Erlang B, C and A models. These results are useful for the dimensioning of large systems equipped with an active control policy. In particular, the corrected approximations can be leveraged to establish the optimality gaps related to square-root staffing and asymptotic dimensioning with admission control.

Keywords: scaled control, QED regime, Halfin-Whitt regime, queues in heavy traffic, diffusion process, asymptotic analysis
2010 MSC: 60K25, 60J60, 60J70, 34E05

1 Introduction

Many-server systems have the capability of combining large capacity with high utilization while maintaining satisfactory system performance. This potential for achieving economies of scale is perhaps most pronounced in the QED regime, or Halfin-Whitt regime. Halfin and Whitt [HW81] were the first to study the QED regime for the $GI/M/s$ system. Assuming that customers require an exponential service time with mean 1, the QED regime refers to the situation that the arrival rate of customers λ and the numbers of servers s are increased in such a way that the traffic intensity $\rho = \lambda/s$ approaches one and

$$(1-\rho)\sqrt{s} \to \gamma, \quad \gamma \in \mathbb{R}. \tag{1.1}$$

The scaling (1.1) is effective because the probability of delay converges to a non-degenerate limit away from both zero and one. Limit theorems for other, more general systems are obtained in [GMR02, JMM04, MZ04, MM08, Ree09], and in all these cases, the limiting probability of delay remains in the interval $(0,1)$. In fact, not only the probability of delay, but many other performance characteristics or objective functions are shown to behave (near) optimally in the QED regime, see for example [BMR04, GKM03]. An important reason for this near optimal behavior are the relatively small fluctuations of the queue-length process.

This can be understood in the following way. Let $X_s(t) = (Q_s(t) - s)/\sqrt{s}$ denote a sequence of normalized processes, with $Q_s(t)$ the process describing the number of customers in the system over time. When $X_s(t) > 0$, it is equal to the scaled total number of customers in the queue, whereas when $X_s(t) < 0$, it is equal to the scaled number of idle servers. Halfin and Whitt showed for the $GI/M/s$ system how under (1.1), $X_s(t)$ converges to a diffusion process $X(t)$ on $\mathbb{R}$, that behaves like a Brownian motion with drift above zero and like an Ornstein-Uhlenbeck (OU) process below zero, and that has a non-degenerate stationary distribution. This shows that the natural scale of $Q_s(t) - s$ is of the order $\sqrt{s}$. More precisely, the queue length is of the order $\sqrt{s}$, as well as the number of idle servers.

*Electronic addresses: a.j.e.m.janssen@tue.nl, j.s.h.v.leeuwaarden@tue.nl, and jaron.sanders@tue.nl

This paper adds to the QED regime the feature of state-dependent control, by considering a control policy that lets an arriving customer enter the system according to some probability depending on the queue length. In particular, a customer meeting upon arrival k other waiting customers is admitted with probability $p_s(k)$, and we allow for a wide range of such control policies characterized by $\{p_s(k)\}_{k\in\mathbb{N}_0}$ with $\mathbb{N}_0 = \{0, 1, \dots\}$. An important property of this control is that it is allowed to scale with the system size s. Consider for example *finite-buffer control*, in which new customers are rejected when the queue length equals N, so that $p_s(k) = 1$ for $k < N$ and $p_s(k) = 0$ for $k \geq N$. A finite-capacity effect in the QED regime that is neither dominant nor negligible occurs when $N \approx \eta\sqrt{s}$ with $\eta > 0$, because the natural scale of the queue length is $\sqrt{s}$. A similar threshold in the context of many-server systems in the QED regime has been considered in [AM04, MW04, Whi04, Whi05].

We introduce a class of QED-specific control policies $\{p_s(k)\}_{k\in\mathbb{N}_0}$ designed, like the finite-buffer control, to control the fluctuations of $Q_s(t)$ around s. To this end, we consider control policies for which $p_s(x\sqrt{s}) \approx 1 - a(x)/\sqrt{s}$ when $x > 0$. Here, a denotes a non-negative and non-decreasing function. While almost all customers are admitted as $s \to \infty$, this control is specifically designed for having a decisive influence on the system performance in the QED regime. We also provide an in-depth discussion of two canonical examples. The first is *modified-drift control* given by $a(x) = \vartheta > -\gamma$, which is shown to effectively change the QED parameter γ in (1.1) into $\gamma + \vartheta$. The second example is *Erlang A control* given by $a(x) = \vartheta x$, for which the system behavior is shown to be intimately related with that of the Erlang A model in which waiting customers abandon the system after an exponential time with mean $1/\vartheta$.

Our class of QED-specific control policies stretches much beyond these two examples. In principle, we can choose the control such that, under Markovian assumptions, the stochastic-process limit for the normalized queue-length process changes the Brownian motion in the upper half plane (corresponding to the system without control), into a diffusion process with drift $-\gamma - a(x)$ in state $x \geq 0$. We give a formal proof of this process-level result.

We next consider the controlled QED system in the stationary regime and derive the QED limits for the probability of delay and the probability of rejection. Typically, such results can be obtained by using the central limit theorem and case-specific arguments, see for example [GMR02, HW81, JMM04, MW04]. However, we take a different approach, aiming for new asymptotic expansions for the probability of delay and the probability of rejection. The first terms of these expansions are the QED limits, and the higher-order terms are refinements to these QED limits for finite s. This generalizes earlier results on the Erlang B, C and A models [JvLZ08, JvLZ11, ZvLZ12].

Conceptually, we develop a unifying approach to derive such expansions for these control policies. A crucial step in our analysis is to rewrite the stationary distribution in terms of a Laplace transform that contains all specific information about the control policy. Mathematically, establishing the expansions requires an application of Euler-Maclaurin (EM) summation, essentially identifying the error terms caused by replacing a series expression in the stationary distribution by the Laplace transform. In this paper we focus on the probability of delay and the probability of rejection, but it is fairly straightforward using the same approach to obtain similar results for other characteristics of the stationary distribution, such as the mean and the cumulative distribution function.

The paper is structured as follows. In §2 we introduce the many-server system with admission control and derive the stability condition under which the stationary distribution exists. In §3 we discuss in detail the QED scaled control. We introduce a *global* control for managing the overall system fluctuations, and a *local* control that entails a precise form of $p_s(k)$. For both the global and local control, we derive the stability condition and the stochastic-process limit for the normalized queue-length process in terms of a diffusion process. In §4 we derive QED approximations for systems with global control. Hereto, we enroll our concept of describing the stationary distribution in terms of a Laplace transform and using EM summation to derive the expansions. In §5 we derive QED approximations for local control, making heavy use of the intimate connection with global control and the tools developed in §4. For demonstrational purposes we also provide some numerical results for the Erlang A control. In §6 we discuss the potential applications of the results obtained in this paper.

2 Many-server systems with admission control

Consider a system with s parallel servers to which customers arrive according to a Poisson process with rate λ. The service times of customers are assumed exponentially distributed with mean 1. A control policy dictates whether or not a customer is admitted to system. A customer that finds upon arrival k other waiting customers in the system is allowed to join the queue with probability $p_s(k)$ and is rejected with probability $1 - p_s(k)$. In this way, the sequence $\{p_s(k)\}_{k\in\mathbb{N}_0}$ defines the control policy. Since we are interested in large, many-server systems, working at critical load and hence serving many customers, the probability $p_s(k)$ should be interpreted as the fraction of customers admitted in state $s + k$.

Under these Markovian assumptions, and assuming that all interarrival times and service times are mutually independent, this gives rise to a birth–death process $Q_s(t)$ describing the number of customers in the system over time. The birth rates are λ for states $k = 0, 1, \ldots, s$ and $\lambda \cdot p_s(k - s)$ for states $k = s, s + 1, \ldots$. The death rate in state k equals $\min\{k, s\}$ for states $k = 1, 2, \ldots$. Assuming the stationary distribution to exist, with $\pi_k = \lim_{t\to\infty} \mathbb{P}(Q_s(t) = k)$, it follows from solving the balance equations that

$$\pi_k = \begin{cases} \pi_0 \frac{(s\rho)^k}{k!}, & k = 1, 2, \ldots, s, \\ \pi_0 \frac{s^s \rho^k}{s!} \prod_{i=0}^{k-s-1} p_s(i), & k = s + 1, s + 2, \ldots. \end{cases} \tag{2.1}$$

Here

$$\rho = \frac{\lambda}{s}, \quad \pi_0^{-1} = \sum_{k=0}^{s} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!} F_s(\rho) \tag{2.2}$$

with

$$F_s(\rho) = \sum_{n=0}^{\infty} p_s(0) \cdot \ldots \cdot p_s(n)\, \rho^{n+1}. \tag{2.3}$$

2.1 Stability

The wide class of allowed control policies renders it necessary to carefully investigate the precise conditions under which the controlled system is stable. From (2.1)–(2.3), we conclude that the stationary distribution exists if and only if $\{p_s(k)\}_{k\in\mathbb{N}_0}$ and ρ are such that $F_s(\rho) < \infty$. Let

$$P_s := \limsup_{n\to\infty} \left(p_s(0) \cdot \ldots \cdot p_s(n)\right)^{\frac{1}{n+1}}, \tag{2.4}$$

and set $1/P_s = \infty$ when $P_s = 0$. We then see that $F_s(\rho) < \infty$ when

$$0 \le \rho < \frac{1}{P_s}. \tag{2.5}$$

For convenience, we henceforth assume that

$$\lim_{\rho\uparrow 1/P_s} F_s(\rho) = \infty, \tag{2.6}$$

so that the stationary distribution exists if and only if (2.5) holds. The case $\lim_{\rho\uparrow 1/P_s} F_s(\rho) < \infty$ (as considered for example in [FA95, JvL12]) can also be considered in the present context, but leads to some complications that distract attention from the bottom line of the exposition.

2.2 Performance measures

We consider in this paper two performance measures, viz. the stationary probability $D_s(\rho)$ that an arriving customer finds all servers occupied, and the stationary probability $D_s^R(\rho)$ that an arriving customer is rejected. In terms of π_k and $p_s(k)$, these stationary probabilities are given by $D_s(\rho) = \sum_{k=s}^{\infty} \pi_k$ and $D_s^R(\rho) = \sum_{k=s}^{\infty} \pi_k(1 - p_s(k))$. Denoting the Erlang B formula by

$$B_s(\rho) = \frac{(s\rho)^s/s!}{\sum_{k=0}^{s} (s\rho)^k/k!}, \tag{2.7}$$

we express $D_s(\rho)$ and $D_s^R(\rho)$ in terms of $B_s(\rho)$ and $F_s(\rho)$ as

$$D_s(\rho) = \frac{1 + F_s(\rho)}{B_s^{-1}(\rho) + F_s(\rho)} \tag{2.8}$$

and

$$D_s^R(\rho) = \frac{1 + (1 - \rho^{-1})\, F_s(\rho)}{B_s^{-1}(\rho) + F_s(\rho)}. \tag{2.9}$$

There are two extreme control policies. The first is the control that denies all customers access whenever all servers are occupied, i.e. $p_s(k) = 0$ for $k \in \mathbb{N}_0$. This is in fact the Erlang B system. Then, $F_s(\rho) = 0$ and (2.8), (2.9) indeed give $D_s(\rho) = D_s^R(\rho) = B_s(\rho)$. The other is the control that allows all customers access, i.e. $p_s(k) = 1$ for $k \in \mathbb{N}_0$, known as the Erlang C system. Equation (2.3) gives $F_s(\rho) = \rho/(1-\rho)$ for $0 \le \rho < 1$. Subsequently, (2.9) gives $D_s^R(\rho) = 0$ (a customer is never rejected) and expression (2.8) reduces to the Erlang C formula

$$C_s(\rho) = \frac{\frac{(s\rho)^s}{s!(1-\rho)}}{\sum_{k=0}^{s-1} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!\,(1-\rho)}}. \tag{2.10}$$

3 QED scaled control

To enforce the QED regime in (1.1) we henceforth couple λ and s according to

$$\rho = \frac{\lambda}{s} = 1 - \frac{\gamma}{\sqrt{s}} \Leftrightarrow \lambda = s - \gamma\sqrt{s}, \quad \gamma \in \mathbb{R}. \tag{3.1}$$

We next introduce two types of control, referred to as *global* and *local* control, both designed to reduce the incoming traffic in periods of congestion.

3.1 Global control

Recall (2.3) and let

$$q_s(n) := p_s(0) \cdot \ldots \cdot p_s(n), \quad n \in \mathbb{N}_0 \tag{3.2}$$

be the coefficient of ρ^{n+1} in $F_s(\rho)$. For $n \in \mathbb{N}_0$, $q_s(n)$ is roughly equal to the probability that a (fictitious) batch arrival of n customers is allowed as a whole to enter the system, given that all servers are busy and that the waiting queue is empty. Since in the QED regime queue lengths are of the order $\sqrt{s}$, it is natural to consider control policies such that $q_s(n)$ scales with s in a $\sqrt{s}$-manner as well. One way to achieve this is by choosing $q_s(n)$ of the form

$$q_s(n) = f\Big(\frac{n+1}{\sqrt{s}}\Big), \quad n \in \mathbb{N}_0 \tag{3.3}$$

for $s \ge 1$, where $f(x)$, henceforth referred to as *scaling profile*, is a non-negative, non-increasing function of $x \ge 0$ with $f(0) = 1$. With global control we mean that the admission control is defined through $q_s(n)$ in (3.3). A key example is what we have called modified-drift control, in which case

$$p_s(k) = p^{\frac{1}{\sqrt{s}}},\ p \in (0,\infty), \quad q_s(n) = p^{\frac{n+1}{\sqrt{s}}} = f\Big(\frac{n+1}{\sqrt{s}}\Big) \text{ with } f(x) = p^x. \tag{3.4}$$

It appears that many practical admission policies fit into the Ansatz (3.3), or do so in a limit sense as $s \to \infty$. This is the case for the class of local control, as discussed next.

3.2 Local control

While the global control is defined via $q_s(n)$, we also introduce a local control, that for each state k, defines the probability of admitting a new customer as

$$p_s(k) = \frac{1}{1 + \frac{1}{\sqrt{s}} a(\frac{k+1}{\sqrt{s}})}, \quad k \in \mathbb{N}_0, \tag{3.5}$$

with $a(x)$ a non-negative, non-decreasing function of $x \geq 0$. A special case is Erlang A control $a(x) = \vartheta x$, which gives $p_s(k) = 1/(1 + (k+1)\vartheta/s)$. In this case the stationary distribution is identical to that of an $M/M/s + M$ system (or Erlang A model), with the feature that customers that are waiting in the queue abandon the system after exponentially distributed times with mean $1/\vartheta$. Garnett et al. [GMR02] obtained the diffusion limit for the Erlang A model in the QED regime, and the limiting diffusion process turned out to be a combination of two OU processes with different restraining forces, depending on whether the process is below or above zero.

Note that setting $a(x) = 0$ leads to the ordinary $M/M/s$ system considered in [HW81] with in the QED regime as limiting process a Brownian motion in the upper half plane. Depending on a, i.e. the type of control, one gets a specific limiting behavior in the upper half plane, described by Brownian motion, an OU process, or some other type of diffusion process with drift $-\gamma - a(x)$ in state $x \geq 0$. We give a formal proof of this process-level convergence in §3.5.

3.3 Connection between local and global control

There is a fundamental relation between local and global control. By substituting (3.5) into (3.2), rewriting the product and using Taylor expansion, we see that

$$\begin{aligned} q_s(n) = p_s(0) \cdot \ldots \cdot p_s(n) &= \exp\Big(-\sum_{k=0}^{n} \ln\Big(1 + \frac{1}{\sqrt{s}}\, a\Big(\frac{k+1}{\sqrt{s}}\Big)\Big)\Big) \\ &= \exp\Big(\frac{-1}{\sqrt{s}} \sum_{k=0}^{n} a\Big(\frac{k+1}{\sqrt{s}}\Big) + O\Big(\frac{1}{s} \sum_{k=0}^{n} a^2\Big(\frac{k+1}{\sqrt{s}}\Big)\Big)\Big), \quad n \in \mathbb{N}_0. \end{aligned} \tag{3.6}$$

For large s and under mild conditions on a, the last expression in (3.6) can be approximated by

$$\exp\Big(-\int_0^{\frac{n+1}{\sqrt{s}}} a(y)\, dy + O\Big(\frac{1}{\sqrt{s}} \int_0^{\frac{n+1}{\sqrt{s}}} a^2(y)\, dy\Big)\Big), \tag{3.7}$$

which will be discussed in more detail in §5.1. We get the approximation

$$q_s(n) \approx f\Big(\frac{n+1}{\sqrt{s}}\Big), \quad n \in \mathbb{N}_0, \tag{3.8}$$

where

$$f(x) = \exp\Big(-\int_0^x a(y)\, dy\Big), \quad x \geq 0. \tag{3.9}$$

The validity range and the approximation error in (3.8) depend on the particular form of a, which will be discussed in detail in §5. Also, (3.9) implies that f and a are related as

$$a(x) = -\frac{f'(x)}{f(x)}, \quad x \geq 0. \tag{3.10}$$

From here onwards we assume that f in (3.3) and a in (3.5) are indeed related according to (3.9) and (3.10). We can then show that both local and global control have a similar impact on a system, characterized by

$$p_s(k) \approx 1 - \frac{1}{\sqrt{s}} a\Big(\frac{k+1}{\sqrt{s}}\Big), \quad k \in \mathbb{N}_0. \tag{3.11}$$

In §2.2 we discussed how our class of control policies can cover the entire range between the Erlang B model and the Erlang C model. Let us demonstrate that for the modified-drift control described in (3.4) that admits a customer when all servers are busy with probability $p^{1/\sqrt{s}}$, where $p \in (0, 1)$. Figure 1 shows for fixed $\rho = 0.99$ the delay probability $D_s(\rho)$ as a function of p. Here we show both global control $f(x) = p^x$ and the local control counterpart $a(x) = -\ln p$. Notice the relatively small difference between global and local control, which would be even smaller for larger values of s.

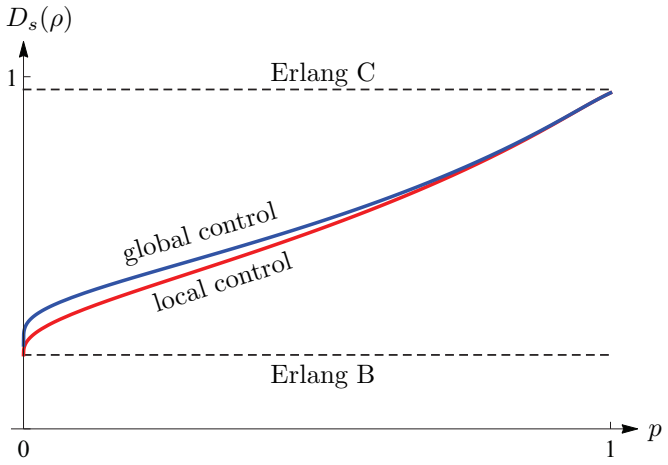

Figure 1: The stationary probability of delay for global control $f(x) = p^x$ and local control $a(x) = -\ln p$ for $s = 10$ and $\rho = 0.99$.

3.4 Stability with control

Now that we have established the connection between global and local control via the relations (3.9) and (3.10), we next show that the stability conditions for the systems with these respective controls are similar as well.

Define the Laplace transform of the scaling profile f as

$$\mathcal{L}(\gamma) = \int_0^\infty e^{-\gamma x} f(x) dx, \quad \gamma > \gamma_{\min}, \tag{3.12}$$

where $\gamma_{\min} = \inf\{\gamma \in \mathbb{R} | \mathcal{L}(\gamma) < \infty\}$. From (3.9), it follows that $\gamma_{\min} = -\lim_{x\to\infty} a(x)$, and since a is non-decreasing, we have

$$\lim_{\gamma \downarrow \gamma_{\min}} \mathcal{L}(\gamma) = \infty. \tag{3.13}$$

In A, we derive the following stability condition for global and local control in terms of $\gamma_{\min}$. For large s, the two stability conditions are almost identical.

Proposition 3.1 (Stability conditions)**.** *Assume* (3.9) *and* (3.10)*. The stationary distribution* (2.1) *exists for*

(i) *the global control* (3.3) *if and only if* $0 \le \rho < e^{-\gamma_{\min}/\sqrt{s}} = 1 - \gamma_{\min}/\sqrt{s} + O(1/s)$*;*

(ii) *the local control* (3.5) *if and only if* $0 \le \rho < 1 - \gamma_{\min}/\sqrt{s}$*.*

3.5 Stochastic-process limit

We now derive using the local control in (3.5) a stochastic-process limit, which provides additional insight into the roles of the function a and the Laplace transform $\mathcal{L}$.

Let $Q_s(t)$ denote the process describing the number of customers present in the system over time. The subscript s is attached to all relevant quantities to denote their dependence on the size of the system. We obtain a scaling limit for the sequence of normalized processes $X_s(t) = (Q_s(t) - s)/\sqrt{s}$. Let "$\Rightarrow$" denote weak convergence in the space $D[0, \infty)$ or convergence in distribution. The next result is proved in B.

Proposition 3.2 (Weak convergence to a diffusion process)**.** *Assume* (1.1) *and* (3.5)*. If a is continuous and bounded on every compact subinterval I of $\mathbb{R}$, and $X_s(0) \Rightarrow X(0) \in \mathbb{R}$, then for every $t \ge 0$, as $s \to \infty$,*

$$X_s(t) \Rightarrow X(t), \tag{3.14}$$

where the limit $X(t)$ is the diffusion process with infinitesimal drift $m(x)$ given by

$$m(x) = \begin{cases} -\gamma - x, & x < 0, \\ -\gamma - a(x), & x \geq 0, \end{cases} \tag{3.15}$$

and constant infinitesimal variance $\sigma^2(x) = 2$.

Proposition 3.2 sheds light on the effect of the control $p_s(k)$ as s becomes large. It shows that for local control (3.5), which is asymptotically of the form (3.11), the process $Q_s(t)$ approximately behaves as $s + X(t)\sqrt{s}$, where $X(t)$ is a diffusion process with drift $-\gamma - a(x)$ for $x \geq 0$ and an OU process with drift $-\gamma - x$ for $x < 0$.

The stationary distribution of $X(t)$ is easy to derive. Denote the probability density function of the standard normal distribution by $\phi(x)$, and its cumulative distribution function by $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$.

Proposition 3.3 (Stationary distribution of the diffusion process)**.** *The density function $\omega(x)$ of the stationary distribution for $X(t)$ is given by*

$$\omega(x) = \begin{cases} C(\gamma)\frac{\phi(x+\gamma)}{\phi(\gamma)}, & x < 0, \\ C(\gamma)\exp(\int_0^x m(u)du), & x \geq 0, \end{cases} \tag{3.16}$$

with $C(\gamma) = \Big(\int_0^\infty \exp(\int_0^x m(u)du)dx + \frac{\Phi(\gamma)}{\phi(\gamma)}\Big)^{-1}$. Moreover,

$$\int_0^\infty \omega(x)dx = \frac{\mathcal{L}(\gamma)}{\mathcal{L}(\gamma) + \frac{\Phi(\gamma)}{\phi(\gamma)}}. \tag{3.17}$$

Proof. Since the diffusion process $X(t)$ has piecewise continuous parameters, we can apply the procedure developed in [BW95] to find the stationary distribution. This procedure consists of composing the density function as in (3.16) based on the density function of a diffusion process with drift $-\gamma - a(x)$ for $x > 0$ and of an OU process with drift $-\gamma - x$ for $x < 0$. The function $C(\gamma)$ normalizes the distribution.

Equation (3.17) follows after substituting (3.15) with $a(x) = -f'(x)/f(x)$ into (3.16) and evaluating

$$\int_0^\infty \exp\Big(\int_0^x m(u)du\Big)dx = \int_0^\infty e^{-\gamma x} f(x)\,dx = \mathcal{L}(\gamma), \tag{3.18}$$

proving (3.17). □

A natural approach now is to approximate the distribution of $Q_s(t)$ by the distribution of $s + X(t)\sqrt{s}$ when s is large. In §4, specifically Theorem 4.3, we show that (3.17) equals $\lim_{s\to\infty} D_s(\rho)$, as expected, and we also derive the most relevant correction terms for finite s. Important here is that $D_s(\rho)$ converges to a value in the interval $(0, 1)$ as $s \to \infty$, which confirms that the local control in (3.5) leads to a non-degenerate limit. In [JvL12], s-independent control policies have been considered for which $D_s(\rho)$ has $1/\sqrt{s}$-behavior for large s.

4 QED approximations for global control

In this section we focus on global control defined by the scaling profile f and Ansatz (3.3). For this type of control there is a convenient manner of approximating $F_s(\rho)$ as $s \to \infty$ in terms of the Laplace transform of f. Define

$$\gamma_s = -\sqrt{s}\ln(1 - \gamma/\sqrt{s}), \quad \gamma \in \mathbb{R}. \tag{4.1}$$

Utilizing (3.2) and (3.3) and recalling that $\rho = 1 - \gamma/\sqrt{s}$, we can write (2.3) as

$$F_s(\rho) = \sum_{n=0}^{\infty} e^{-\frac{n+1}{\sqrt{s}}\gamma_s} f\Big(\frac{n+1}{\sqrt{s}}\Big), \tag{4.2}$$

This expression for $F_s(\rho)$ is instrumental for our analysis. We apply EM summation to (4.2), in order to replace the summation over n by an integral and an appropriate number of error terms. The approach is explained in §4.1, and leads to the QED approximations for the stationary delay and rejection probability presented in §4.2. These approximations are demonstrated in §4.3 for several types of global control.

4.1 EM summation

We assume that the function $f(x)$ in (3.3) is non-negative, non-increasing and smooth, that is $f \in C^4([0,\infty))$, and that $f(0)=1$. Furthermore, we assume that for any $\gamma > \gamma_{\min}$, $e^{-\gamma x} f^{(j)}(x) \in L_1([0,\infty))$ and $e^{-\gamma x} f^{(j)}(x) \to 0$ as $x \to \infty$ for $j = 0,1,2,3,4$.

We shall use the following form of the EM summation formula about which more details are collected in C. Assume that $g : [0,\infty) \to \mathbb{R}$ with $g \in C^2([0,\infty))$ and $g^{(j)} \in L^1([0,\infty))$, $j = 0,1,2$. Then

$$\sum_{n=0}^{\infty} g\Big(\frac{n+1}{\sqrt{s}}\Big) = \sqrt{s}\int_{\frac{1}{2\sqrt{s}}}^{\infty} g(x)dx + \frac{1}{24\sqrt{s}} g'\Big(\frac{1}{2\sqrt{s}}\Big) + R, \tag{4.3}$$

where

$$|R| \le \frac{1}{12\sqrt{s}} \int_0^{\infty} |g^{(2)}(x)|dx. \tag{4.4}$$

When also $g^{(4)} \in L^1([0,\infty))$, we have the asymptotically tighter bound

$$|R| \le \frac{1}{384 s\sqrt{s}} \int_0^{\infty} |g^{(4)}(x)|dx. \tag{4.5}$$

By setting $g(x) = e^{-\gamma x} f(x)$ for $x \ge 0$ and $\gamma > \gamma_{\min}$, and using these formulas, we will now obtain several QED approximations.

4.2 Corrected QED approximations

We first present a result for $F_s(\rho)$.

Theorem 4.1. *With $\rho = 1 - \gamma/\sqrt{s}$ and $\gamma_{\min} < \gamma \le \sqrt{s}$,*

$$F_s(\rho) = \sqrt{s}\mathcal{L}(\gamma_s) - \frac{1}{2} + O\Big(\frac{1}{\sqrt{s}}\Big), \tag{4.6}$$

where $\gamma_s = -\sqrt{s}\ln(1-\gamma/\sqrt{s})$ and where $O(1/\sqrt{s})$ holds uniformly in any compact set of γ's contained in $(\gamma_{\min},\infty)$.

Proof. We have from (4.2), (4.3) and (4.4) that

$$F_s(\rho) = \sqrt{s}\int_{\frac{1}{2\sqrt{s}}}^{\infty} e^{-\gamma_s x} f(x)dx + \frac{1}{24\sqrt{s}} (e^{-\gamma_s x} f(x))'\Big(\frac{1}{2\sqrt{s}}\Big) + R \tag{4.7}$$

with

$$|R| \le \frac{1}{12\sqrt{s}} \int_0^{\infty} |(e^{-\gamma_s x} f(x))^{(2)}(x)|dx. \tag{4.8}$$

Assume that γ is restricted to a compact set C contained in $(\gamma_{\min},\infty)$. Then γ_s is restricted to a compact set D contained in $(\gamma_{\min},\infty)$ for all $s \ge 1$ with $\sqrt{s} \ge 2\max\{|\gamma| \mid \gamma \in C\}$. Hence

$$e^{-\gamma_s x} f(x) - 1 = O\Big(\frac{1}{\sqrt{s}}\Big), \quad 0 \le x \le \frac{1}{2\sqrt{s}}, \tag{4.9}$$

where $O(1/\sqrt{s})$ holds uniformly in $\gamma \in C$. Therefore, we can replace the integral at the right-hand side of (4.7) by $\mathcal{L}(\gamma_s) - 1/(2\sqrt{s})$, at the expense of an error $O(1/s)$ uniformly in $\gamma \in C$. Furthermore,

$$(e^{-\gamma_s x} f(x))'\Big(\frac{1}{2\sqrt{s}}\Big) = O(1), \quad s \ge 1, \tag{4.10}$$

uniformly in $\gamma \in C$ by smoothness of f. Finally, $R = O(1/\sqrt{s})$ uniformly in $\gamma \in C$ since there is the bound

$$|R| \le \frac{1}{12\sqrt{s}} \int_0^{\infty} e^{-\gamma_s x}(|\gamma_s|^2 + 2|\gamma_s||f'(x)| + |f''(x)|)dx \tag{4.11}$$

in which $\gamma_s \in D$ with f satisfying the assumptions made at the beginning of §4.1. □

Theorem 4.1 yields a simple and often accurate approximation of $F_s(\rho)$, which we illustrate using examples in §4.3. However, in the leading term $\sqrt{s}\,\mathcal{L}(\gamma_s)$, the dependence on the number of servers s and the parameter γ is combined into the single quantity γ_s. A more insightful result is given in Theorem 4.2 below, where the dependence of the approximating terms on s and γ is separated.

Theorem 4.2. *With $\rho = 1-\gamma/\sqrt{s}$ and $\gamma_{\min} < \gamma \le \sqrt{s}$,*

$$F_s(\rho) = \sqrt{s}\,\mathcal{L}(\gamma) + \mathcal{M}(\gamma) + O\Big(\frac{1}{\sqrt{s}}\Big), \tag{4.12}$$

where

$$\mathcal{M}(\gamma) = \frac{1}{2}\gamma^2\,\mathcal{L}'(\gamma) - \frac{1}{2}, \tag{4.13}$$

and where $O(1/\sqrt{s})$ holds uniformly in any compact set of γ's contained in $(\gamma_{\min},\infty)$. In leading order, the $O(1/\sqrt{s})$ is given as $\mathcal{N}(\gamma)/\sqrt{s}$, where

$$\mathcal{N}(\gamma) = \frac{1}{3}\gamma^3\,\mathcal{L}'(\gamma) + \frac{1}{8}\gamma^4\,\mathcal{L}''(\gamma) + \frac{1}{12}(\gamma - f'(0)). \tag{4.14}$$

Proof. This result is obtained from (4.3) in a similar way as Theorem 4.1, using now the estimate (4.5) of R, and approximating $\int_0^{1/(2\sqrt{s})} e^{-\gamma_s x} f(x)dx$ and $(e^{-\gamma_s x} f(x))'(1/(2\sqrt{s}))$ more carefully. In particular, we have

$$\int_0^{1/(2\sqrt{s})} e^{-\gamma_s x} f(x)dx = \frac{1}{2\sqrt{s}} + \frac{1}{8s}(f'(0) - \gamma_s) + O\Big(\frac{1}{s\sqrt{s}}\Big) \tag{4.15}$$

and

$$\frac{d}{dx}\left(e^{-\gamma_s x}\, f(x)\right)\Big(\frac{1}{2\sqrt{s}}\Big) = \gamma_s - f'(0) + O\Big(\frac{1}{\sqrt{s}}\Big). \tag{4.16}$$

Furthermore, because $\gamma_s = \gamma + \gamma^2/(2\sqrt{s}) + \gamma^3/(3s) + \ldots$ for $|\gamma| < \sqrt{s}$, we can approximate $\mathcal{L}(\gamma_s)$ by

$$\begin{aligned}\mathcal{L}(\gamma_s) - \mathcal{L}(\gamma) &= (\gamma_s - \gamma)\mathcal{L}'(\gamma) + \frac{1}{2}(\gamma_s - \gamma)^2\,\mathcal{L}''(\gamma) + \frac{1}{6}(\gamma_s - \gamma)^3\,\mathcal{L}'''(\gamma) + \ldots \\ &= \frac{\gamma^2}{2\sqrt{s}}\,\mathcal{L}'(\gamma) + \frac{1}{s}\Big(\frac{1}{3}\gamma^3\mathcal{L}'(\gamma) + \frac{1}{8}\gamma^4\,\mathcal{L}''(\gamma)\Big) + O\Big(\frac{1}{s\sqrt{s}}\Big).\end{aligned} \tag{4.17}$$

The O's in (4.15)–(4.17) hold uniformly in γ in any compact set contained in $(\gamma_{\min},\infty)$. □

We use the following short-hand notations for approximations of $F_s(\rho)$ and $B_s(\rho)$ as they occur in the performance measures $D_s(\rho)$ and $D_s^R(\rho)$ in (2.8) and (2.9). We write (4.12) using (4.14) as

$$F_s(\rho) = \sqrt{s}\,\mathcal{L} + \mathcal{M} + \frac{1}{\sqrt{s}}\mathcal{N} = \sqrt{s}\,\mathcal{L} + \mathcal{M} + O\Big(\frac{1}{\sqrt{s}}\Big), \tag{4.18}$$

and we write the Jagerman approximation of $B_s(\rho)$, see [JvL12] and [Jag74, Theorem 14], as

$$B_s(\rho) = \frac{1}{\sqrt{s}}\,g + \frac{1}{s}\,h + O\Big(\frac{1}{s\sqrt{s}}\Big). \tag{4.19}$$

Here, $\rho = 1-\gamma/\sqrt{s}$, and

$$g(\gamma) = \frac{\phi(\gamma)}{\Phi(\gamma)}, \quad h(\gamma) = -\frac{1}{3}\big(\gamma^2 + (\gamma^2+2)g(\gamma)\big)g(\gamma). \tag{4.20}$$

The following results for $D_s(\rho)$ and $D_s^R(\rho)$ are proved in D using (4.18) and (4.19).

Theorem 4.3 (Corrected QED approximations)**.** *The stationary probability of delay satisfies*

$$D_s(\rho) = T_1(\gamma) + \frac{1}{\sqrt{s}}\,T_2(\gamma) + O\Big(\frac{1}{s}\Big), \tag{4.21}$$

where

$$T_1 = \frac{g\mathcal{L}}{1+g\mathcal{L}}, \quad T_2 = \frac{(h+g^2)\,\mathcal{L} + g(\mathcal{M}+1)}{(1+g\mathcal{L})^2}, \tag{4.22}$$

and where $O(1/s)$ holds uniformly in any compact set of γ's contained in $(\gamma_{\min}, \infty)$. The stationary rejection probability satisfies

$$D_s^R(\rho) = \frac{1}{\sqrt{s}}\,T_1^R(\gamma) + \frac{1}{s}\,T_2^R(\gamma) + O\Big(\frac{1}{s\sqrt{s}}\Big), \tag{4.23}$$

where

$$T_1^R = \frac{1-\gamma\mathcal{L}}{1+g\mathcal{L}}\,g, \qquad T_2^R = \frac{1-\gamma\mathcal{L}}{1+g\mathcal{L}}\Big(h - \gamma g\,\frac{\gamma\mathcal{L}+\mathcal{M}}{1-\gamma\mathcal{L}} - g\,\frac{h\mathcal{L}+g\mathcal{M}}{1+g\mathcal{L}}\Big), \tag{4.24}$$

and where $O(1/s\sqrt{s})$ holds uniformly in any compact set of γ's contained in $(\gamma_{\min}, \infty)$.

4.3 Examples

We now present several examples to illustrate Theorems 4.1, 4.2 and 4.3.

4.3.1 Modified-drift control (global)

Consider $f(x) = p^x$ for $x \geq 0$, with $p \in (0,1)$ fixed. Then, $\gamma_{\min} = \ln p$, $P_s = p^{1/\sqrt{s}}$ and

$$F_s(\rho) = \frac{p^{1/\sqrt{s}}(1-\gamma/\sqrt{s})}{1-p^{1/\sqrt{s}}(1-\gamma/\sqrt{s})}, \quad \sqrt{s}(1-p^{-1/\sqrt{s}}) < \gamma \leq \sqrt{s}. \tag{4.25}$$

Theorem 4.2 gives the approximation

$$F_s(\rho) \approx \frac{\sqrt{s}}{\gamma - \ln p} - \frac{\gamma^2}{2(\gamma-\ln p)^2} - \frac{1}{2}, \quad \gamma > \ln p. \tag{4.26}$$

4.3.2 Erlang A control (global)

Let $f(x) = p^{x^2}$ for $x \geq 0$, with $p \in (0,1)$ fixed. In this case, $\gamma_{\min} = -\infty$ and $P_s = 0$. Also,

$$\mathcal{L}(\gamma) = \frac{1}{\sqrt{\alpha}}\chi(\gamma/(2\sqrt{\alpha})), \quad \gamma \in \mathbb{R}, \tag{4.27}$$

where $\alpha = -\ln p$ and χ is Mills' ratio, defined as $\chi(\delta) = e^{\delta^2}\int_\delta^\infty e^{-y^2}dy$ for $\delta \in \mathbb{R}$ [OLBC10, §7.8]. Taking the derivative, we find that

$$\mathcal{L}'(\gamma) = \frac{\gamma}{2\alpha\sqrt{\alpha}}\chi(\gamma/(2\sqrt{\alpha})) - \frac{1}{2\alpha}, \quad \gamma \in \mathbb{R}, \tag{4.28}$$

and we then obtain from Theorem 4.2 the approximation

$$F_s(\rho) \approx \frac{\sqrt{s}}{\sqrt{\alpha}}\chi(\gamma/(2\sqrt{\alpha})) + \frac{1}{4}\Big(\frac{\gamma}{\sqrt{\alpha}}\Big)^3\chi(\gamma/(2\sqrt{\alpha})) - \frac{1}{4}\Big(\frac{\gamma}{\sqrt{\alpha}}\Big)^2 - \frac{1}{2}, \quad \gamma \in \mathbb{R}. \tag{4.29}$$

Note that the approximation $F_s(\rho) \approx \sqrt{s}\mathcal{L}(\gamma_s) - 1/2$ as described in Theorem 4.1 can also be computed from (4.27) after recalling that $\gamma_s = -\sqrt{s}\ln(1-\gamma/\sqrt{s})$.

4.3.3 Scaled buffer control (global)

Take a fixed $\eta > 0$ and set $p_s(k) = \mathbb{1}[k+1 < \eta\sqrt{s}]$ for $k \in \mathbb{N}_0$. Thus for $n \in \mathbb{N}_0$, $q_s(n) = p_s(n) = f((n+1)/\sqrt{s})$, with $f(x) = \mathbb{1}[x \in [0,\eta)]$ for $x \geq 0$. It follows that $P_s = 0$, $\gamma_{\min} = -\infty$ and

$$F_s(\rho) = \Big(\frac{\sqrt{s}}{\gamma} - 1\Big)\Big(1 - \Big(1 - \frac{\gamma}{\sqrt{s}}\Big)^{\lfloor\eta\sqrt{s}\rfloor}\Big), \quad -\infty < \gamma \leq \sqrt{s}. \tag{4.30}$$

The function f is not smooth, and strictly speaking, Theorems 4.1 and 4.2 do not apply. Still, we can calculate

$$\mathcal{L}(\gamma) = \frac{1}{\gamma}(1 - e^{-\gamma\eta}), \quad \mathcal{L}'(\gamma) = -\frac{1}{\gamma^2}(1 - (1+\gamma\eta)e^{-\gamma\eta}), \quad \gamma \in \mathbb{R}, \tag{4.31}$$

and use the approximation that Theorem 4.1 would give, i.e.

$$F_s(\rho) \approx \sqrt{s}\mathcal{L}(\gamma_s) - \frac{1}{2} = \frac{1 - (1 - \gamma/\sqrt{s})^{\eta\sqrt{s}}}{-\ln(1 - \gamma/\sqrt{s})} - \frac{1}{2}. \tag{4.32}$$

Alternatively, Theorem 4.2 gives the approximation

$$F_s(\rho) \approx \sqrt{s}\mathcal{L}(\gamma) + \frac{1}{2}\gamma^2\mathcal{L}'(\gamma) - \frac{1}{2} = \Big(\frac{\sqrt{s}}{\gamma} - 1\Big)(1 - e^{-\gamma\eta}) - \frac{1}{2}e^{-\gamma\eta}(1 - \gamma\eta), \quad \gamma \in \mathbb{R}. \tag{4.33}$$

While (4.30) has a jump as a function of s at all s where $\eta\sqrt{s}$ is integer, its approximations in (4.32) and (4.33) are smooth functions of s, if we consider $s \geq 1$ as a continuous variable. The averages of the approximations over s-intervals $[(k/\eta)^2, ((k+1)/\eta)^2]$ with integer k agree well with the average of (4.30) over these intervals. Thus, while Theorems 4.1 and 4.2 do not apply, they yield approximations that perform well in an appropriate average sense. This is also illustrated in Figure 2.

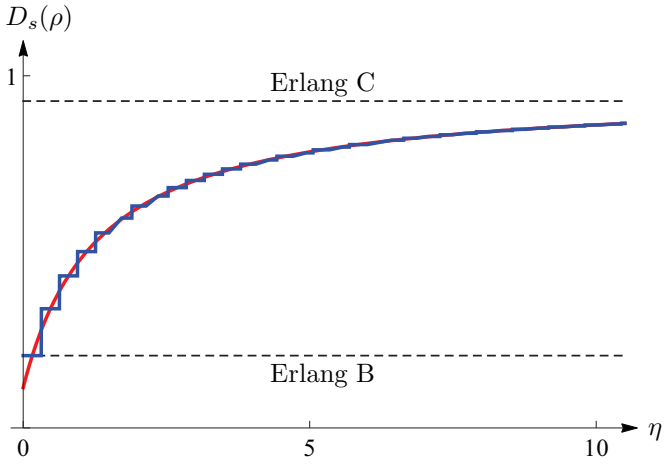

Figure 2: The stationary probability of delay for $s = 10$ and scaled buffer control. The jagged, blue curve gives the exact value (4.30) and the red, smooth curve pertains to approximation (4.33).

5 QED approximations for local control

A clear technical advantage of global control is that it leads to infinite-series expressions for the performance measures $D_s(\rho)$ and $D_s^R(\rho)$ that are directly amenable to asymptotic analysis based on EM summation. This approach was followed in §4 and led to Theorem 4.3. As argued in §3, in some practical cases it is more natural to work with the local control defined in (3.5). In this section we show that Theorem 4.3 also gives sharp approximations for local control. Indeed, in §3.3 it was argued that for local control, $q_s(n) \approx f((n+1)/\sqrt{s})$ with f defined as in (3.9). Consider for instance the modified-drift control in (3.4), in which case

$$\begin{aligned} F_s(\rho) &= \sum_{n=0}^{\infty} p^{\frac{n+1}{\sqrt{s}}}\Big(1 - \frac{\gamma}{\sqrt{s}}\Big)^{n+1} \\ &= \frac{\sqrt{s}}{\gamma - \ln p} - \frac{\gamma}{\gamma - \ln p} - \frac{1}{2}\Big(\frac{\ln p}{\gamma - \ln p}\Big)^2 + O\Big(\frac{1}{\sqrt{s}}\Big). \end{aligned} \tag{5.1}$$

Here, the second equality follows from the QED approximation in (4.12). The local counterpart follows from $a(x) = -f'(x)/f(x) = -\ln p$ and (3.5), for which

$$F_s(\rho) = \sum_{n=0}^{\infty} \Big(\frac{1}{1-\frac{1}{\sqrt{s}}\ln p}\Big)^{n+1}\Big(1-\frac{\gamma}{\sqrt{s}}\Big)^{n+1} = \frac{\sqrt{s}}{\gamma-\ln p} - \frac{\gamma}{\gamma-\ln p}. \tag{5.2}$$

The second equality in (5.2) follows from summation of a geometric series. Hence, for the example of modified-drift control, it can be seen from the close resemblance of the last members of (5.1) and (5.2) that approximating local control by global control yields sharp estimates in the QED regime.

In §5.1 we make formal the accuracy of the approximation $q_s(n) \approx f((n+1)/\sqrt{s})$ for a wide range of local controls. As it turns out, the approximation becomes asymptotically correct in the QED regime. Therefore, for local controls for which the Ansatz $q_s(n) = f((n+1)/\sqrt{s})$ does not hold precisely, it will give sharp approximations for the performance measures in the QED regime. For the example of Erlang A control, with $a(x) = \vartheta x$, this is demonstrated in §5.2.

5.1 Approximating local by global control

We first present a general result for all functions a considered in this paper. The proof is given in E.

Proposition 5.1 (Relation between global and local control)**.** *Assume that $a(x)$ is non-negative and non-decreasing in $x \geq 0$. There is an increasing function $\psi(s)$ of $s \geq 1$ with $\psi(s) \to \infty$, $s \to \infty$, such that*

$$q_s(n) = f\Big(\frac{n+1}{\sqrt{s}}\Big)\Big(1+O(s^{-\frac{1}{4}})\Big), \quad 0 \leq n+1 \leq \sqrt{s}\psi(s), \tag{5.3}$$

where f is given as in (3.9).

We next illustrate Proposition 5.1 for a special case of increasing a. Let $\vartheta > 0$ and $\alpha \geq 0$, and let $a(x) = \vartheta\, x^\alpha$ for $x \geq 0$. Inspecting the proof of Proposition 5.1, case $\delta = 1/4$, it is seen that ψ is found by requiring

$$a\Big(\frac{n+1}{\sqrt{s}}\Big) \leq \frac{1}{2}s^{\frac{1}{4}}, \quad \int_0^{\frac{n+1}{\sqrt{s}}} a^2(x)dx \leq s^{\frac{1}{4}}. \tag{5.4}$$

Proposition 5.1 yields

$$q_s(n) = \exp\Big(\frac{-\vartheta}{\alpha+1}\Big(\frac{n+1}{\sqrt{s}}\Big)^{\alpha+1}\Big)\Big(1+O(s^{-\frac{1}{4}})\Big), \quad 0 \leq n+1 \leq \sqrt{s}\psi(s), \tag{5.5}$$

with

$$\psi(s) = \min\Big\{\Big(\frac{1}{2\vartheta}\Big)^{\frac{1}{\alpha}} s^{\frac{1}{4\alpha}}, \Big(\frac{2\alpha+1}{\vartheta^2}\Big)^{\frac{1}{2\alpha+1}} s^{\frac{1}{4(2\alpha+1)}}\Big\}. \tag{5.6}$$

5.2 Erlang A control (local)

We now consider in detail Erlang A control, in order to demonstrate our obtained QED approximations. Erlang A control gives rise to a birth–death process that is identical to the classical Erlang A model [GMR02, ZvLZ12]. It allows us to express $F_s(\rho)$ in terms of the confluent hypergeometric function and to subsequently show that an asymptotic expansion of the confluent hypergeometric function leads to a QED approximation similar to (4.29).

We thus consider the example

$$p_s(k) = \frac{1}{1+(k+1)\frac{\vartheta}{s}}, \quad k \in \mathbb{N}_0, \tag{5.7}$$

which corresponds to $\alpha = 1$ when $a(x) = \vartheta x^\alpha$ for $x \geq 0$, so $f(x) = \exp(-\vartheta x^2/2)$ for $x \geq 0$.

Proposition 5.2. *Assume that $p_s(k)$ is given by (5.7). Then*

$$F_s(\rho) = \frac{1}{\rho}\big(M(1, s/\vartheta, s\rho/\vartheta) - 1 - \rho\big), \quad \rho \geq 0, \tag{5.8}$$

in which

$$M(a,b,z)=\sum_{n=0}^{\infty}\frac{(a)_n}{(b)_n}\frac{z^n}{n!},\quad z\in\mathbb{C} \tag{5.9}$$

is the confluent hypergeometric function [OLBC10, Chapter 13], *with* $(x)_l$ *the Pochhammer symbol, i.e.* $(x)_l=0$ *for* $l=1$ *and* $(x)_l=x(x+1)\cdot\ldots\cdot(x+l-1)$ *for* $l\geq 1$.

Proof. For $n\in\mathbb{N}_0$,

$$q_s(n)=\prod_{k=0}^{n}\frac{1}{1+(k+1)\,\frac{\vartheta}{s}}=\frac{(s/\vartheta)^{n+2}}{(s/\vartheta)_{n+2}}. \tag{5.10}$$

Therefore

$$F_s(\rho)=\sum_{n=0}^{\infty}q_s(n)\,\rho^{n+1}=\sum_{n=0}^{\infty}\frac{(s/\vartheta)^{n+2}}{(s/\vartheta)_{n+2}}\,\rho^{n+1}, \tag{5.11}$$

and (5.8) follows after some rearrangements. □

In [OLBC10, 13.8(ii)] the asymptotics of $M(a,b,z)$ is considered when b and z are large while a is fixed and b/z is in a compact set contained in $(0,\infty)$. With

$$a=1\,,\quad b=s/\vartheta\,,\quad z=s\rho/\vartheta \tag{5.12}$$

and $s\to\infty$, while $\rho=1-\gamma/\sqrt{s}$ is close to 1, this is precisely the situation we are interested in. Temme [Tem78] gives a complete asymptotic series, and this leads to the following result.

Proposition 5.3. *As* $s\to\infty$,

$$F_s(\rho)\sim\Big(\frac{2}{\vartheta}\Big)^{\frac12}\chi(\gamma/\sqrt{2\vartheta})\sqrt{s}+\frac{\gamma^3\sqrt{2}}{3\vartheta^{\frac32}}\chi(\gamma/\sqrt{2\vartheta})-\frac{\gamma^2}{3\vartheta}-\frac23. \tag{5.13}$$

Proof. The first two terms of Temme's asymptotic series [Tem78] are as follows. Let $\zeta=\sqrt{2(\rho-1-\ln\rho)}$, where $\operatorname{sgn}(\zeta)=\operatorname{sgn}(\rho-1)$. Then

$$\begin{aligned}M\Big(1,\frac{s}{\vartheta},\frac{s\rho}{\vartheta}\Big)\sim\Big(\frac{s}{\vartheta}\Big)^{\frac12}\exp\Big(\frac{\zeta^2 s}{4\vartheta}\Big)\Big\{&\rho U\Big(\frac12,-\zeta\Big(\frac{s}{\vartheta}\Big)^{\frac12}\Big)\\&+\Big(\rho-\frac{\zeta}{\rho-1}\Big)\frac{1}{\zeta\big(\frac{s}{\vartheta}\big)^{\frac12}}\,U\Big(-\frac12,-\zeta\Big(\frac{s}{\vartheta}\Big)^{\frac12}\Big)\Big\}\end{aligned} \tag{5.14}$$

with U the parabolic cylinder function of [OLBC10, Ch. 12]. In this particular case [OLBC10, §12.5.1, §12.7.1],

$$U\Big(\frac12,z\Big)=\mathrm{e}^{-\frac14 z^2}\int_0^\infty\mathrm{e}^{-\frac12 t^2-zt}dt,\quad U\Big(-\frac12,z\Big)=\mathrm{e}^{-\frac14 z^2}. \tag{5.15}$$

Since $\rho=1-\gamma/\sqrt{s}$,

$$\zeta=-\frac{\gamma}{\sqrt{s}}-\frac{\gamma^2}{3s}+O\Big(\frac{\gamma^3}{s\sqrt{s}}\Big),\quad \rho-\frac{\zeta}{\rho-1}=-\frac43\frac{\gamma}{\sqrt{s}}+O\Big(\frac{\gamma^2}{s}\Big). \tag{5.16}$$

Substituting in (5.14), together with (5.15), we get

$$\begin{aligned}F_s(\rho)\sim\,-1-\frac{1}{\rho}+\Big(\frac{s}{\vartheta}\Big)^{\frac12}\exp\Big(\frac{\zeta^2 s}{4\vartheta}\Big)\Big\{&\exp\Big(-\frac{\zeta^2 s}{4\vartheta}\Big)\int_0^\infty\mathrm{e}^{-\frac12 t^2+\zeta(\frac{s}{\vartheta})^{\frac12}t}dt\\&+\frac{\rho-\zeta/(\rho-1)}{\zeta\rho(s/\vartheta)^{\frac12}}\exp\Big(-\frac{\zeta^2 s}{4\vartheta}\Big)\Big\},\end{aligned} \tag{5.17}$$

so that

$$F_s(\rho)\sim-\frac23+\Big(\frac{s}{\vartheta}\Big)^{\frac12}\int_0^\infty\mathrm{e}^{-\frac12 t^2+\zeta(\frac{s}{\vartheta})^{\frac12}t}dt+O\Big(\frac{\gamma}{\sqrt{s}}\Big). \tag{5.18}$$

Next

$$\zeta\Big(\frac{s}{\vartheta}\Big)^{\frac{1}{2}} = -\frac{\gamma}{\sqrt{\vartheta}} - \frac{\gamma^2}{3\sqrt{\vartheta}s} + O\Big(\frac{\gamma^3}{s}\Big), \tag{5.19}$$

and using this in (5.18), we get

$$\begin{aligned} F_s(\rho) = &-\frac{2}{3} + \Big(\frac{s}{\vartheta}\Big)^{\frac{1}{2}} \int_0^\infty \mathrm{e}^{-\frac{1}{2}t^2 - \frac{\gamma t}{\sqrt{\vartheta}}} dt \\ &- \frac{1}{3}\Big(\frac{s}{\vartheta}\Big)^{\frac{1}{2}} \frac{\gamma^2}{\sqrt{\vartheta}s} \int_0^\infty \mathrm{e}^{-\frac{1}{2}t^2 - \frac{\gamma t}{\sqrt{\vartheta}}} t dt + O\Big(\frac{\gamma}{\sqrt{s}}\Big). \end{aligned} \tag{5.20}$$

Finally, using that

$$\int_0^\infty \mathrm{e}^{-\frac{1}{2}t^2 - \frac{\gamma t}{\sqrt{\vartheta}}} dt = \sqrt{2}\chi(\gamma/\sqrt{2\vartheta}), \tag{5.21}$$

$$\int_0^\infty \mathrm{e}^{-\frac{1}{2}t^2 - \frac{\gamma t}{\sqrt{\vartheta}}} t dt = -\chi'(\gamma/\sqrt{2\vartheta}) = -\Big(\frac{2\gamma}{\sqrt{2\vartheta}}\chi(\gamma/\sqrt{2\vartheta}) - 1\Big), \tag{5.22}$$

we obtain the result. □

It is instructive to rewrite the asymptotics (4.29) of $F_s(\rho)$ in Example 4.3.2 for the case that $q_n = f((n+1)/\sqrt{s})$ and $f(x) = p^x$, in terms of $\vartheta = 2\alpha = -2\ln p$. In doing so, (4.29) becomes

$$F_s(\rho) \sim \Big(\frac{2}{\vartheta}\Big)^{1/2} \chi(\gamma/\sqrt{2\vartheta})\sqrt{s} + \frac{\gamma^3}{\sqrt{2}\vartheta^{3/2}}\chi(\gamma/\sqrt{2\vartheta}) - \frac{\gamma^2}{2\vartheta} - \frac{1}{2}. \tag{5.23}$$

Observe the close resemblance between (5.23) and (5.13).

5.2.1 Numerical comparison

For Erlang A control, for which $f(x) = p^{x^2}$ and $-\ln p = \alpha = \vartheta/2$, we have now determined an exact expression and an asymptotic expression for $F_s(\rho)$, given in Proposition 5.2 and Proposition 5.3, respectively. Through (2.8) we then obtain exact and asymptotic expressions for D_s. Furthermore, we can obtain approximate values for D_s using the first-order and second-order approximation in Theorem 4.3. Table 1 shows a numerical comparison when using these different expressions for D_s.

	$\vartheta = 1$			$\vartheta = 10$			$\vartheta = 100$		
s	D_s^{exact}	D_s^{asymp}	D_s^{approx}	D_s^{exact}	D_s^{asymp}	D_s^{approx}	D_s^{exact}	D_s^{asymp}	D_s^{approx}
1	0.59343	0.57277	0.62582	0.49415	0.39305	0.48528	0.47591	0.29172	0.41076
2	0.55437	0.54342	0.57730	0.41389	0.34704	0.40797	0.38093	0.23525	0.31498
4	0.52652	0.52092	0.54300	0.35137	0.31225	0.35330	0.29862	0.19283	0.24726
8	0.50691	0.50410	0.51874	0.30732	0.28658	0.31465	0.23226	0.16172	0.19938
16	0.49313	0.49172	0.50158	0.27830	0.26792	0.28731	0.18229	0.13925	0.16552
32	0.48343	0.48273	0.48946	0.25956	0.25448	0.26798	0.14717	0.12315	0.14157
64	0.47660	0.47625	0.48088	0.24735	0.24487	0.25432	0.12407	0.11169	0.12464
128	0.47178	0.47160	0.47481	0.23924	0.23802	0.24465	0.10961	0.10354	0.11267
256	0.46837	0.46828	0.47053	0.23375	0.23316	0.23782	0.10068	0.09776	0.10421
512	0.46597	0.46592	0.46749	0.23000	0.22970	0.23299	0.09506	0.09367	0.09822
1024	0.46427	0.46425	0.46535	0.22740	0.22725	0.22957	0.09146	0.09774	0.09399

Table 1: Numerical comparison of different expressions of D_s for $f(x) = \exp(-\vartheta x^2/2)$ and $\gamma = 0.1$. Here, D_s^{exact} is calculated using Proposition 5.2, D_s^{asymp} using Proposition 5.3 and $D_s^{\text{approx}} = T_1 + T_2/\sqrt{s}$ using Theorem 4.3. Furthermore, for all s, $T_1 \approx 0.46017, 0.22132$ and 0.08377 for $\vartheta = 1, 10$ and 100, respectively.

From Table 1 we see that the precision of all approximations increase with s. We also see that the second-order approximation $T_1 + T_2/\sqrt{s}$ is more accurate than the first-order approximation T_1, particularly for moderate values of s.

6 Conclusions and outlook

We have introduced QED scaled control, designed to reduce the incoming traffic in periods of congestion, in such a way that the controlled many-server system remains within the domain of attraction of the favorable QED regime. The scaled control is chosen such that it affects the typical $O(\sqrt{s})$ queue lengths that arise in the QED regime. The class of many-server systems with QED control introduced in this paper contains the Erlang B, C and A models as special cases. For all cases we have derived so-called corrected QED approximations, which not only identify the QED limits as leading terms, but also provide corrections through higher order terms for finite system sizes s. As a key example we took the stationary probability of delay, for which the corrected QED approximation reads $D_s(\rho) \approx T_1(\gamma) + T_2(\gamma)/\sqrt{s}$, as stated in Theorem 4.3. The technique developed in this paper to obtain the corrected diffusion approximations can be easily applied to other characteristics of the stationary distribution, such as the mean and the cumulative distribution function.

Our corrected QED approximations pave the way for obtaining optimality results for dimensioning systems [BMR04]. Consider for instance the basic problem of determining the largest load ρ such that $D_s(\rho) \leq \varepsilon$ with $\varepsilon \in (0,1)$. The delay probability is a function of the two model parameters s and λ, and of the control policy. Denote this unique solution by $\rho = \rho_{\rm opt}$ and define $\gamma_{\rm opt} = \sqrt{s}(1-\rho_{\rm opt})$. Asymptotic dimensioning would approximate $D_s(\rho)$ by the QED limit $T_1(\gamma)$ that only depends on γ (and no longer on both s and λ). Hence, the inverse problem can then be approximatively solved by searching for the $\gamma = \gamma_*$ such that $T_1(\gamma) = \varepsilon$, and then setting the load according to $\rho_* = 1 - \gamma_*/\sqrt{s}$. This procedure is referred to as square-root staffing, and the error $|\gamma_{\rm opt} - \gamma_*|$ is called the *optimality gap*. In future work, we will leverage the corrected QED approximations derived in the present paper to characterize the optimality gaps for a large class of dimensioning problems.

Acknowledgments

This research was financially supported by The Netherlands Organization for Scientific Research (NWO) in the framework of the TOP-GO program and by an ERC Starting Grant.

References

[AM04] M. Armony and C. Maglaras. Customer contact centers with a call-back option. *Oper. Res.*, 52:271–292, 2004.

[BMR04] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Oper. Res.*, 52:17–34, 2004.

[BW95] S. Browne and W. Whitt. *Advances in Queueing: Theory, Methods, and Open Problems - Piecewise-linear diffusion processes.* CRC Press, Boca Raton, FL, 1995. ed. J. Dshalalow.

[Ell98] D. Elliott. The Euler-Maclaurin formula revisited. *J. Austral. Math. Soc. B*, 40(E):E27–E76, 1998.

[FA95] G. I. Falin and J. R. Artalejo. Approximations for multi-server queues with balking/retrial discipline. *OR Spektrum*, 17:239–244, 1995.

[GKM03] N. Gans, G. Koole, and A. Mandelbaum. Telephone Call Centers: Tutorial, Review and Research Prospects. *M&SOM-Manuf. Serv. Op.*, 5:79–141, 2003.

[GMR02] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *M&SOM-Manuf. Serv. Op.*, 4:208–227, 2002.

[Hil56] F. B. Hildebrand. *Introduction to Numerical Analysis.* McGraw-Hill, New York, USA, 1956.

[HW81] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29:567–588, 1981.

[Igl74] D. L. Iglehart. Weak convergence in applied probability. *Stoch. Proc. Appl.*, 2(3):211–241, 1974.

[Jag74] D. Jagerman. Some properties of the Erlang loss function. *Bell Syst. Tech. J.*, 53:525–551, 1974.

[JMM04] P. Jelenković, A. Mandelbaum, and P. Momčilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Syst.*, 47:53–69, 2004.

[JvL12] A. J. E. M. Janssen and J. S. H. van Leeuwaarden. Staffing many-server systems with admission control and retrials. *Submitted*, 2012.

[JvLZ08] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. in Appl. Probab.*, 40(1):122–143, 2008.

[JvLZ11] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and B. Zwart. Refining square root staffing by expanding Erlang C. *Oper. Res.*, 59:1512–1522, 2011.

[Lyn85] J. N. Lyness. The Euler Maclaurin expansion for the Cauchy principal value integral. *Numer. Math.*, 46:611–622, 1985.

[MM08] A. Mandelbaum and P. Momčilovic. Queues with many servers: The virtual waiting-time process in the QED regime. *Math. Oper. Res.*, 33:561–586, 2008.

[MW04] W. A. Massey and R. B. Wallace. An asymptotically optimal design of the $M/M/c/k$ queue. *Unpublished report*, 2004.

[MZ04] C. Maglaras and A. Zeevi. Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.*, 29:786–813, 2004.

[OLBC10] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, Cambridge, United Kingdom, 2010.

[Ree09] J. Reed. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 19:2211–2269, 2009.

[Tem78] N. M. Temme. Uniform asymptotic expansions of confluent hypergeometric functions. *J. Inst. Math. Appl.*, 22:215–223, 1978.

[Whi04] W. Whitt. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.*, 30:1–27, 2004.

[Whi05] W. Whitt. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.*, 52:922–941, 2005.

[ZvLZ12] B. Zhang, J. S. H. van Leeuwaarden, and B. Zwart. Refining square-root staffing for call centers with impatient customers. *Oper. Res.*, 60:461–474, 2012.

A Proof of Proposition 3.1

A.1 Proof of Proposition 3.1(i)

We assume that $F_s(\rho)$, $\rho = 1 - \gamma/\sqrt{s}$, is of the form (2.3) with

$$p_s(0) \cdot \ldots \cdot p_s(n) = q_s(n) = f\Big(\frac{n+1}{\sqrt{s}}\Big), \quad n = 0, 1, \ldots, \tag{A.1}$$

where $s \geq 1$ and $f(x)$ is a non-negative and non-increasing function in $x \geq 0$ with $f(0) = 1$. Furthermore, we recall the definition of $\gamma_s = -\sqrt{s}\ln(1-\gamma/\sqrt{s})$ in (4.1). The stability result of Proposition 3.1(i) is a consequence of the following inequality.

Proposition A.1. *For* $\sqrt{s}(1-\exp(-\gamma_{\min}/\sqrt{s})) < \gamma \le 0$,

$$e^{\gamma_s/\sqrt{s}} \int_{1/\sqrt{s}}^{\infty} e^{-\gamma_s x} f(x)\,dx \le \frac{1}{\sqrt{s}} F_s(\rho) \le e^{-\gamma_s/\sqrt{s}} \int_0^{\infty} e^{-\gamma_s x} f(x)\,dx. \tag{A.2}$$

Proof. We start by noting that

$$\gamma > \sqrt{s}(1-e^{-\gamma_{\min}/\sqrt{s}}) =: \gamma_{\min,s} \Leftrightarrow \gamma_s > \gamma_{\min}. \tag{A.3}$$

We consider formula (4.2) for $F_s(\rho)$. We have for $\gamma \le 0$ and $n = 0, 1, \dots$ from monotonicity of f that

$$f(x) \ge f\Big(\frac{n+1}{\sqrt{s}}\Big), \quad e^{-\gamma_s x} \ge e^{\gamma_s/\sqrt{s}}\, e^{-\frac{n+1}{\sqrt{s}}\gamma_s}, \quad \frac{n}{\sqrt{s}} \le x \le \frac{n+1}{\sqrt{s}} \tag{A.4}$$

and

$$f(x) \le f\Big(\frac{n+1}{\sqrt{s}}\Big), \quad e^{-\gamma_s x} \le e^{-\gamma_s/\sqrt{s}}\, e^{-\frac{n+1}{\sqrt{s}}\gamma_s}, \quad \frac{n+1}{\sqrt{s}} \le x \le \frac{n+2}{\sqrt{s}}. \tag{A.5}$$

Hence, from (A.4) for $n = 0, 1, \dots$

$$\int_{n/\sqrt{s}}^{(n+1)/\sqrt{s}} e^{-\gamma_s x} f(x)\,dx \ge \frac{1}{\sqrt{s}}\, e^{\gamma_s/\sqrt{s}}\, e^{-\gamma_s \frac{n+1}{\sqrt{s}}} f\Big(\frac{n+1}{\sqrt{s}}\Big), \tag{A.6}$$

and from (A.5) for $n = 0, 1, \dots$

$$\int_{(n+1)/\sqrt{s}}^{(n+2)/\sqrt{s}} e^{-\gamma_s x} f(x)\,dx \le \frac{1}{\sqrt{s}}\, e^{-\gamma_s/\sqrt{s}}\, e^{-\gamma_s \frac{n+1}{\sqrt{s}}} f\Big(\frac{n+1}{\sqrt{s}}\Big). \tag{A.7}$$

From (A.6) and (A.7) the two inequalities in (A.2) readily follow. □

Proposition A.1 shows that $F_s(\rho) < \infty$ if and only if $\mathcal{L}(\gamma_s) < \infty$, where it is used that f is non-negative and bounded. By the definition of $\gamma_{\min}$ and the assumption in (3.13) it follows that $F_s(\rho) < \infty$ if and only if $\gamma_s > \gamma_{\min}$. Then from (A.3) the equivalence in Proposition 3.1(i) follows.

A.2 Proof of Proposition 3.1(ii)

Let $s = 1, 2, \dots$ and consider the case $-\gamma_{\min} = \lim_{x\to\infty} a(x) =: L < \infty$. Then, by monotonicity of a,

$$q_s(n) = \prod_{k=0}^{n} \frac{1}{1+\frac{1}{\sqrt{s}}a(\frac{k+1/2}{\sqrt{s}})} \ge \Big(\frac{1}{1+\frac{1}{\sqrt{s}}L}\Big)^{n+1}, \tag{A.8}$$

and so $F_s(\rho) = \infty$ when $\rho \ge 1+\frac{1}{\sqrt{s}}L$. Next, take $0 \le \rho < 1+\frac{1}{\sqrt{s}}L$. From $a = a(\frac{k+1/2}{\sqrt{s}}) \le L$, $\rho < 1+\frac{1}{\sqrt{s}}L$,

$$\begin{aligned}\frac{\rho}{1+\frac{1}{\sqrt{s}}a} &= 1 - \frac{1+\frac{1}{\sqrt{s}}L-\rho}{1+\frac{1}{\sqrt{s}}a} + \frac{1}{\sqrt{s}}\frac{L-a}{1+\frac{1}{\sqrt{s}}a} \\ &< 1 - \Big(1-\frac{\rho}{1+\frac{1}{\sqrt{s}}L}\Big) + \frac{1}{\sqrt{s}}(L-a).\end{aligned} \tag{A.9}$$

Hence, we can find a $K = 1, 2 \dots$ such that

$$\frac{\rho}{1+\frac{1}{\sqrt{s}}a(\frac{k+1/2}{\sqrt{s}})} \le 1 - \frac{1}{2}\Big(1-\frac{\rho}{1+\frac{1}{\sqrt{s}}L}\Big), \quad k > K. \tag{A.10}$$

Therefore,

$$\begin{aligned}F_s(\rho) &\le \sum_{n=0}^{K} \rho^{n+1} + \rho^{K+1} \sum_{n=K+1}^{\infty} \rho^{n-K} \prod_{k=K+1}^{n} \frac{1}{1+\frac{1}{\sqrt{s}}a(\frac{k+1/2}{\sqrt{s}})} \\ &< \sum_{n=0}^{K} \rho^{n+1} + \rho^{K+1} \sum_{n=K+1}^{\infty} \Big(1-\frac{1}{2}\Big(1-\frac{\rho}{1+\frac{1}{\sqrt{s}}L}\Big)\Big)^{n-K} < \infty.\end{aligned} \tag{A.11}$$

This proves the result for the case $L < \infty$. The proof for the case $L = \infty$ is similar.

B Proof of Proposition 3.2

We will use Stone's theorem [Igl74], for which we need to verify that (a) the state space of the normalized process converges to a limit that is dense in $\mathbb{R}$ and (b) the infinitesimal mean and variance of $X_s(t)$ converge uniformly to $m(x)$ and $\sigma^2(x)$, respectively.

Condition (a) is readily verified. The state space of $X_s(t)$ is given by $\Omega_s = \{(k-s)/\sqrt{s} | k \in \mathbb{N}_0\}$, and we see that for every $x \in \mathbb{R}$ and every $\epsilon > 0$, there exists an $s > 0$ such that $\min_{y \in \Omega_s} |x - y| < \epsilon$.

To verify condition (b), we recall that for any birth–death process $Y(t)$ with birth–death parameters $\lambda^{(k)}$ and $\mu^{(k)}$ and associated states

$$y^{(0)} < y^{(1)} < y^{(2)} < \dots, \tag{B.1}$$

the infinitesimal mean and variance are defined as

$$m(y) = \lambda^{(e(y))}(y^{(e(y)+1)} - y^{(e(y))}) - \mu^{(e(y))}(y^{(e(y))} - y^{(e(y)-1)}) \tag{B.2}$$

and

$$\sigma^2(y) = \lambda^{(e(y))}(y^{(e(y)+1)} - y^{(e(y))})^2 + \mu^{(e(y))}(y^{(e(y))} - y^{(e(y)-1)})^2, \tag{B.3}$$

respectively. Here,

$$e(y) = \arg\sup_{k \in \mathbb{N}_0} \{y^{(k)} | y^{(k)} \le y\} \tag{B.4}$$

is the label of the state closest to, but never above, $y \in [y^{(0)}, y^{(\infty)})$.

For each birth–death process $X_s(t)$, we have that

$$\lambda_s^{(k)} = \lambda_s \mathbb{1}[k < s] + \lambda_s p_s(k-s)\mathbb{1}[k \ge s], \quad k \in \mathbb{N}_0, \tag{B.5}$$

$$\mu_s^{(k)} = \min\{k, s\}, \quad k \in \mathbb{N}_0, \tag{B.6}$$

$$x_s^{(e_s(x)+1)} - x_s^{(e_s(x))} = 1/\sqrt{s}, \quad x \in [-\sqrt{s}, \infty), \tag{B.7}$$

and

$$e_s(x) = \lfloor s + x\sqrt{s} \rfloor, \quad x \in [-\sqrt{s}, \infty). \tag{B.8}$$

Because γ is fixed, (3.1) prescribes that we are scaling the arrival rate as $\lambda_s = s - \gamma\sqrt{s}$. This yields

$$m_s(x) = \begin{cases} -\gamma + \frac{s - \lfloor s + \sqrt{s}x \rfloor}{\sqrt{s}}, & x < 0, \\ -\gamma p_s(\lfloor s + \sqrt{s}x \rfloor - s) + (p_s(\lfloor s + \sqrt{s}x \rfloor - s) - 1)\sqrt{s}, & x \ge 0, \end{cases} \tag{B.9}$$

and

$$\sigma_s^2(x) = \begin{cases} \frac{s + \lfloor s + \sqrt{s}x \rfloor)}{s} - \frac{\gamma}{\sqrt{s}}, & x < 0, \\ p_s(\lfloor s + \sqrt{s}x \rfloor - s) + 1 - \frac{\gamma p_s(\lfloor s + \sqrt{s}x \rfloor - s)}{\sqrt{s}}, & x \ge 0. \end{cases} \tag{B.10}$$

By first Taylor expanding (3.5),

$$p_s(k) = \frac{1}{1 + \frac{1}{\sqrt{s}} a(\frac{k+1}{\sqrt{s}})} = 1 - \frac{1}{\sqrt{s}} a\Big(\frac{k+1}{\sqrt{s}}\Big) + O(s^{-1}), \quad k \in \mathbb{N}_0, \tag{B.11}$$

and then substituting (B.11) into (B.9) and (B.10), we conclude that

$$m_s(x) = \begin{cases} -\gamma - \frac{\lfloor s + \sqrt{s}x \rfloor - s}{\sqrt{s}}, & x < 0, \\ -\gamma - a\big(\frac{\lfloor s + \sqrt{s}x \rfloor - s + 1}{\sqrt{s}}\big) + O(s^{-\frac{1}{2}}), & x \ge 0, \end{cases} \tag{B.12}$$

and $\sigma_s^2(x) = 2 + O(s^{-\frac{1}{2}})$ for all x. Because $(\lfloor s + \sqrt{s}x \rfloor - s + 1)/\sqrt{s} \to x$ as $s \to \infty$ and a is continuous and bounded on every compact subinterval I of $\mathbb{R}$, we have that for every compact subinterval I of $\mathbb{R}$, $\lim_{s\to\infty} m_s(x) = m(x)$ and $\lim_{s\to\infty} \sigma_s^2(x) = \sigma^2(x) = 2$ uniformly for $x \in I$. This concludes the proof.

C EM summation

Let $m = 1, 2, \ldots$, $N = 1, 2, \ldots$, and let $h \in C^{2m}([0, N+1])$. Then

$$\begin{aligned}
&\sum_{n=0}^{N} h(n+1/2) - \int_0^{N+1} h(x)dx \\
&= \sum_{k=1}^{m} \frac{B_{2k}(1/2)}{(2k)!}(h^{(2k-1)}(N+1) - h^{(2k-1)}(0)) - \int_0^{N+1} \frac{\tilde{B}_{2m}(x-1/2)}{(2m)!} h^{(2m)}(x)dx \\
&= \sum_{k=1}^{m-1} \frac{B_{2k}(1/2)}{(2k)!}(h^{(2k-1)}(N+1) - h^{(2k-1)}(0)) - \int_0^{N+1} \frac{\tilde{B}_{2m}(x-1/2) - B_{2m}(1/2)}{(2m)!} h^{(2m)}(x)dx. \qquad \text{(C.1)}
\end{aligned}$$

Here

$$B_{2k}(1/2) = -(1 - 2^{-2k+1})B_{2k}, \quad k = 1, 2, \ldots, \qquad \text{(C.2)}$$

with B_{2k} the Bernoulli numbers of positive, even order, see [OLBC10, §24.2 (i)] and $\tilde{B}_{2m}(x) = B_{2m}(x - \lfloor x \rfloor)$ with $B_{2m}(x)$ the Bernoulli polynomial of degree $2m$. Moreover, the two integrals in (C.1) involving $\tilde{B}_{2m}$ can be estimated as

$$\Big| \int_0^{N+1} \frac{\tilde{B}_{2m}(x-1/2)}{(2m)!} h^{(2m)}(x)dx \Big| \le \frac{|B_{2m}|}{(2m)!} \int_0^{N+1} |h^{(2m)}(x)|dx \qquad \text{(C.3)}$$

and

$$\Big| \int_0^{N+1} \frac{\tilde{B}_{2m}(x-1/2) - B_{2m}(1/2)}{(2m)!} h^{(2m)}(x)dx \Big| \le 2(1 - 2^{-2m}) \frac{|B_{2m}|}{(2m)!} \int_0^{N+1} |h^{(2m)}(x)|dx, \qquad \text{(C.4)}$$

respectively. These formulas follow from [Lyn85, Theorem 1.3] or [Ell98, Theorem 2.1] (the latter reference containing a proof) for EM summation of h sampled at points $n+\nu$, $n = 0, 1, \ldots, N$. For the special case that $\nu = 1/2$, a simplification occurs due to $B_j(1/2) = 0$ for $j = 1, 3, \ldots$. The bounds in (C.3) and (C.4) follow from [OLBC10, §24.12 (i) and §24.4.34], with a special consideration for $B_2(x) = x^2 - x + 1/6$. We have $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = 1/42, \ldots$. When $m = 1$, the series over k in the second form in (C.1) is absent.

The formula (C.1) is sometimes called the second EM summation formula, see [Hil56, (5.8.18–19) on p. 154]. It distinguishes itself from the first EM summation formula, as appears in [OLBC10, §2.10 (i)], in that (i) half-integer, rather than integer samples of h are used at the left-hand side, (ii) absence of a term $\frac{1}{2}(h(N+1/2) + h(1/2))$ at the right-hand side, and (iii) smaller coefficients $B_{2k}(1/2)$ in the series over k at the right-hand side. Hence, also see the comment in [Hil56, (5.8.18–19)], the second EM formula is somewhat simpler in form and slightly more accurate when the remainder terms R are dropped than the first EM formula.

In the main text, this formula is used for $h(x) = g((x+1/2)/\sqrt{s})$, with $g \in C^{2m}([0, \infty))$ and $m = 1$ and 2 while assuming that $\int_0^\infty |g^{(2m)}(x)|dx < \infty$ and that $g^{(2k-1)}(x) \to 0$, $x \to \infty$, for $k = 1$ and $k = 1, 2$, respectively. Subsequently, the formula is used for functions $g(x)$ of the form $\exp(-\gamma_s x)f(x)$, $x \ge 0$.

D Proof of Theorem 4.3

We can write

$$D_s(\rho) = \frac{1 + F_s(\rho)}{B_s^{-1}(\rho) + F_s(\rho)} = H_1(F_s(\rho), B_s(\rho)) \qquad \text{(D.1)}$$

and

$$D_s^R(\rho) = \frac{1 + (1 - \rho^{-1})F_s(\rho)}{B_s^{-1}(\rho) + F_s(\rho)} = H_{1-\rho^{-1}}(F_s(\rho), B_s(\rho)), \qquad \text{(D.2)}$$

where

$$H_a(x, y) = \frac{1 + ax}{y^{-1} + x}. \qquad \text{(D.3)}$$

Error propagation in D_s and D_s^R when both $F_s(\rho)$ and $B_s(\rho)$ are approximated can be assessed using the following result.

Proposition D.1. *For $a \in \mathbb{R}$ and $x \geq 0$, $x + \Delta x \geq 0$ and $0 \leq y \leq 1$, $0 \leq y + \Delta y \leq 1$, it holds that*

$$|H_a(x + \Delta x, y + \Delta y) - H_a(x, y)| \leq |y(a - y)||\Delta x| + |1 + ax||\Delta y|. \tag{D.4}$$

Proof. We have

$$\frac{\partial H_a}{\partial x} = \frac{y(a-y)}{(1+xy)^2}, \quad \frac{\partial H_a}{\partial y} = \frac{1+ax}{(1+xy)^2}. \tag{D.5}$$

Therefore

$$\max_{x \geq 0} \Big| \frac{\partial H_a}{\partial x} \Big| = |y(a-y)|, \quad 0 \leq y \leq 1, \tag{D.6}$$

and

$$\max_{0 \leq y \leq 1} \Big| \frac{\partial H_a}{\partial y} \Big| = |1 + ax|, \quad x \geq 0, \tag{D.7}$$

and the result follows. □

We insert the approximations (4.18) and (4.19) into (D.1), and we get

$$D_s(\rho) = \Big(\frac{1}{\sqrt{s}} g + \frac{1}{s} h \Big) \frac{1 + \sqrt{s}\mathcal{L} + \mathcal{M}}{1 + (\frac{1}{\sqrt{s}} g + \frac{1}{s} h)(\sqrt{s}\mathcal{L} + \mathcal{M})} + O\Big(\frac{1}{s}\Big). \tag{D.8}$$

The approximation error $O(1/s)$ here is obtained from using (D.4) with $a = 1$ and with

$$x = \sqrt{s}\,\mathcal{L} + \mathcal{M} = O(\sqrt{s}), \quad \Delta x = O\Big(\frac{1}{\sqrt{s}}\Big), \tag{D.9}$$

$$y = \frac{1}{\sqrt{s}}\, g + \frac{1}{s}\, h = O\Big(\frac{1}{\sqrt{s}}\Big), \quad \Delta y = O\Big(\frac{1}{s\sqrt{s}}\Big), \tag{D.10}$$

where the O's in (D.9–D.10) hold uniformly in any compact set of γ's contained in $(\gamma_{\min}, \infty)$. Consequently, the $O(1/s)$ in (D.8) holds uniformly in any compact set of γ's contained in $(\gamma_{\min}, \infty)$. Expanding the expression at the right-hand side of (D.8), retaining only the terms $O(1)$ and $O(1/\sqrt{s})$, then yields (4.21–4.22).

In a similar fashion (4.23–4.24) is shown, although the computations are rather involved. We must be a bit careful with T_2^R because of the denominator $1 - \gamma\mathcal{L}$ that appears in (4.24). Recall that in the case that $f(x) = 1$, $x \geq 0$, we have that $1 - \gamma\mathcal{L} = 0 = \gamma\mathcal{L} + \mathcal{M}$, and so $T_1^R = T_2^R = 0$. In the case that $f(x_0) < 1$ for some $x_0 \geq 0$, it is easy to show from $f(0) = 1$, non-negativity and decreasingness of $f(x)$ that for any compact $C \subset (\gamma_{\min}, \infty)$

$$\max_{\gamma \in C} \gamma\mathcal{L}(\gamma) < 1. \tag{D.11}$$

This yields uniform validity of the $O(1/s\sqrt{s})$ in (4.23) when γ is restricted to a compact subset of $(\gamma_{\min}, \infty)$.

E Proof of Proposition 5.1

We assume that $a(x)$ is non-negative and non-decreasing. Define

$$a^{\leftarrow}(y) = \sup\{x \geq 0 \,|\, a(x) \leq y\} \tag{E.1}$$

for $y \geq a(0)$. This generalized inverse function is continuous from the right at all y such that $a^{\leftarrow}(y)$ is finite. Furthermore note that $a(x) \leq y$ when $x < a^{\leftarrow}(y)$.

Lemma E.1. *Let $s = 1, 2, \ldots$, and denote for $n = 1, 2, \ldots$*

$$S_s(n) = \sum_{k=0}^{n} \ln\Big(1 + \frac{1}{\sqrt{s}} a\Big(\frac{k+1}{\sqrt{s}}\Big)\Big). \tag{E.2}$$

Also, let $\delta \in (0, 1/2)$. Then

$$0 \le S_s(n) - \sqrt{s} \int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx \le \frac{1}{2} s^{\delta - \frac{1}{2}}. \tag{E.3}$$

when $n + 1 < \sqrt{s} a^{\leftarrow}(s^{\delta}/2)$. Furthermore, define

$$A(x) = \int_0^x a^2(u) du, \quad x \ge 0. \tag{E.4}$$

Then, except in the trivial case $a \equiv 0$, $A(x)$ is continuous and strictly increasing from 0 at $x_0 := \sup\{x \ge 0 \,|\, a(x) = 0\}$ to ∞ at $x = \infty$. Furthermore,

$$0 \le \int_0^{\frac{n+1}{\sqrt{s}}} a(x) dx - \sqrt{s} \int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx \le \frac{1}{2} s^{\delta - \frac{1}{2}} \tag{E.5}$$

when $n + 1 < \sqrt{s} A^{\leftarrow}(s^{\frac{1}{2} - \delta})$.

Proof. Since $a(x)$ is non-decreasing in $x \ge 0$, we have that $S_s(n)/\sqrt{s}$ is an upper Riemann sum for

$$\int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx, \tag{E.6}$$

while

$$\frac{1}{\sqrt{s}} S_s(n-1) = \frac{1}{\sqrt{s}} S_s(n) - \frac{1}{\sqrt{s}} \ln\Big(1 + \frac{1}{\sqrt{s}} a\Big(\frac{n+1}{\sqrt{s}}\Big)\Big) \tag{E.7}$$

is a lower Riemann sum for

$$\int_{\frac{1}{\sqrt{s}}}^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx. \tag{E.8}$$

It follows that

$$\begin{aligned}
&\sqrt{s} \int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx \le S_s(n) \\
&\le \sqrt{s} \int_{\frac{1}{\sqrt{s}}}^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx + \ln\Big(1 + \frac{1}{\sqrt{s}} a\Big(\frac{n+1}{\sqrt{s}}\Big)\Big) \\
&\le \sqrt{s} \int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx + \ln\Big(1 + \frac{1}{\sqrt{s}} a\Big(\frac{n+1}{\sqrt{s}}\Big)\Big),
\end{aligned} \tag{E.9}$$

where in the last inequality $a(x) \ge 0$ has been used. Now

$$\ln\Big(1 + \frac{1}{\sqrt{s}} a\Big(\frac{n+1}{\sqrt{s}}\Big)\Big) \le \frac{1}{\sqrt{s}} a\Big(\frac{n+1}{\sqrt{s}}\Big) \le \frac{1}{2} s^{\delta - \frac{1}{2}} \tag{E.10}$$

when $n + 1 < \sqrt{s} a^{\leftarrow}(s^{\delta}/2)$. This yields (E.3).

As for (E.5), we note that

$$a(x) - \frac{1}{2\sqrt{s}} a^2(x) \le \sqrt{s} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) \le a(x). \tag{E.11}$$

Hence,

$$0 \le \int_0^{\frac{n+1}{\sqrt{s}}} a(x) dx - \sqrt{s} \int_0^{\frac{n+1}{\sqrt{s}}} \ln\Big(1 + \frac{1}{\sqrt{s}} a(x)\Big) dx \le \frac{1}{2\sqrt{s}} \int_0^{\frac{n+1}{\sqrt{s}}} a^2(x) dx \le \frac{1}{2} s^{-\delta} \tag{E.12}$$

when $n + 1 < \sqrt{s} A^{\leftarrow}(s^{\frac{1}{2} - \delta})$. □

The proof of Proposition 5.1 follows now from Lemma E.1 with $\delta = 1/4$ and taking $\psi(s) = \min\{a^{\leftarrow}(s^{\delta}/2), A^{\leftarrow}(s^{\frac{1}{2} - \delta})\}$.