Transient Behavior of the Halfin-Whitt Diffusion

Johan S.H. van Leeuwaarden[◦] & Charles Knessl[•]

October 24, 2008

Abstract: We consider the heavy-traffic approximation to the $GI/M/s$ queueing system in the Halfin-Whitt regime, where both the number of servers s and the arrival rate λ grow large (taking the service rate as unity), with $\lambda = s - \beta\sqrt{s}$ and β some constant. In this asymptotic regime, the queue length process can be approximated by a diffusion process that behaves like a Brownian motion with drift above zero and like an Ornstein-Uhlenbeck process below zero. We analyze the transient behavior of this hybrid diffusion process, including the transient density, approach to equilibrium, and spectral properties. The transient behavior is shown to depend on whether β is smaller or larger than the critical value $\beta_* \approx 1.85722$, which confirms the recent result of Gamarnik and Goldberg [9].

2000 Mathematics Subject Classification: 60K25, 60J60, 60J70, 34E05.
Keywords & Phrases: $GI/M/s$ queue; $M/M/s$ queue; Halfin-Whitt regime; queues in heavy traffic; diffusion; asymptotic analysis.

1 Introduction

Halfin and Whitt [13] introduced in their 1981 paper a new heavy-traffic limit theorem for the $GI/M/s$ system. They demonstrated how under certain conditions a sequence of normalized queue-length processes converges to a process that behaves like a Brownian motion with drift above zero and like an Ornstein-Uhlenbeck process below zero. We refer to this hybrid diffusion process as the *Halfin-Whitt diffusion*. Our concern is with the transient behavior of this diffusion.

What is nowadays known as the Halfin-Whitt regime refers to the scaling of the arrival rate λ and the numbers of servers s such that, while both λ and s increase toward infinity, the traffic intensity $\rho = \lambda/s$ approaches one and

$$(1-\rho)\sqrt{s} \to \beta, \quad \beta \in (-\infty, \infty). \tag{1.1}$$

This type of scaling was already proposed by Erlang (see [3]) for the $M/M/s/s$ system, and by Pollaczek [22], p. 28, for the $M/D/s$ system. Halfin and Whitt [13] presented a formal limit theorem for the $GI/M/s$ system. Then, some two decades later, the regime got immensely popular due to its application to call centers (see [2, 8, 14]). The scaling (1.1) combines large capacity with high utilization such that the probability of delay converges to a non-degenerate limit away from both zero and one; cf. (2.21). Limit theorems for other, more general systems were obtained in [10, 11, 15, 20, 21, 24]. For delay systems like $M/D/s$ and $GI/M/s$ one should impose $\beta \in (0, \infty)$ to guarantee stability.

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

[•]University of Illinois at Chicago, Department of Mathematics, Statistics and Computer Science, 815 South Morgan Street, Chicago, IL 60607-7045, USA. Email address: knessl@uic.edu

In [13] it is established that by setting the traffic intensity $\rho = 1 - \beta/\sqrt{s}$, $\beta \in (0, \infty)$, the number of customers in the $M/M/s$ system can be roughly expressed as $s + \sqrt{s}X(t)$ for s sufficiently large and $(X(t))_{t\geq 0}$ the Halfin-Whitt diffusion. It is further shown that properties of the limiting diffusion process for the $GI/M/s$ system can be obtained from $(X(t))_{t\geq 0}$ as well. The boundary between the Brownian motion and the Ornstein-Uhlenbeck process can be thought of as the number of servers, and $(X(t))_{t\geq 0}$ will keep fluctuating between these two regions. The process mimics a single server queue above zero, and an infinite server queue below zero, for which Brownian motion and the Ornstein-Uhlenbeck process are indeed the respective heavy-traffic limits. As β increases, capacity grows and the Halfin-Whitt diffusion will spend more time below zero.

The diffusion process $(X(t))_{t\geq 0}$ can thus be employed to obtain simple approximations for the system behavior. The steady-state properties of the diffusion are well studied, but less is known about the transient behavior. Transient results enhance our understanding of how the $GI/M/s$ system behaves over various time and space scales. Results for the mean hitting time were presented in Maglaras and Zeevi [20]. We shall derive explicit results for the transient density of the diffusion, both exact and asymptotic.

We first derive the Laplace transform over time, which leads to a representation of the density as a contour integral, from which a spectral expansion may be obtained by analyzing the complex singularities of the integrand. The spectral expansion can be interpreted as a large-time expansion in which the first term, corresponding to the singularity at zero, gives the steady-state density (which exists if $\beta > 0$). The other singularities of the Laplace transform provide finite-time corrections to the steady-state density. This facilitates us to study how, and in what time (relaxation time), the process converges to its steady state.

The approach to equilibrium is governed by the singularity in the left half-plane with the largest real part. This dominant singularity turns out to be either a branch point or a pole, depending on whether β is smaller or larger than the critical value $\beta_* \approx 1.85722$. This confirms the recent result of Gamarnik and Goldberg [9] who identified β_* using the framework of Karlin and McGregor [16] for birth-death processes, and the result of van Doorn [6] on the spectral gap of the $M/M/s$ queue. We shall also show how the branch point and the pole each give rise to different large-time asymptotics for the density. The main results are presented in Section 2 and the proofs are presented in Section 3.

2 Main results

The Halfin-Whitt diffusion is a Markov process on the real line with continuous paths and density $p = p(x, t)$ that satisfies the forward Kolmogorov equation

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[A(x)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[B(x)p(x,t)]. \tag{2.1}$$

Here $B(x) = 2$,

$$A(x) = \begin{cases} -\beta, & x > 0, \\ -x - \beta, & x < 0, \end{cases} \tag{2.2}$$

and there are the initial condition $p(x, 0) = \delta(x - x_0)$ (the Dirac function) and the boundary conditions $p(\infty, t) = p(-\infty, t) = 0$. This diffusion process applies directly to the $M/M/s$ system. For the $GI/M/s$ system we would need to first take the diffusion coefficient $B(x) = (1 + c^2)$, with $c^2 > 0$, and scale x so as to make $B(x) = 2$, and then scale β by the same factor as x (see [13], Theorem 4).

Define the Laplace transform over time $\hat{p}$ by

$$\hat{p}(x;\theta) = \int_0^\infty e^{-\theta t} p(x,t)\mathrm{d}t, \quad \Re(\theta) > 0. \tag{2.3}$$

Let

$$R_\beta(\theta) = \frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} \tag{2.4}$$

with $D_\nu(z)$ the parabolic cylinder function with index ν and argument z. Below we give expressions for $\hat{p}$, where we must distinguish the cases $x_0 > 0$ and $x_0 < 0$.

Theorem 1. *Consider* $x_0 > 0$.

(i) *For* $x > 0$,

$$\begin{aligned}\hat{p}(x;\theta) &= \frac{e^{\frac{1}{2}\beta(x_0-x)}}{\sqrt{\beta^2+4\theta}}\left(e^{-|x-x_0|\sqrt{\theta+\beta^2/4}} - e^{-(x+x_0)\sqrt{\theta+\beta^2/4}}\right) \\ &\quad + \frac{e^{\frac{1}{2}\beta(x_0-x)}e^{-(x+x_0)\sqrt{\theta+\beta^2/4}}}{\sqrt{\theta+\beta^2/4}-R_\beta(\theta)}.\end{aligned} \tag{2.5}$$

(ii) *For* $x < 0$,

$$\hat{p}(x;\theta) = e^{-\frac{1}{4}x^2}e^{-\frac{1}{2}\beta x}\frac{D_{-\theta}(-\beta-x)}{D_{-\theta}(-\beta)}\frac{e^{\frac{1}{2}x_0\beta - x_0\sqrt{\theta+\beta^2/4}}}{\sqrt{\theta+\beta^2/4}-R_\beta(\theta)}. \tag{2.6}$$

Theorem 2. *Consider* $x_0 < 0$.

(i) *For* $x > 0$,

$$\hat{p}(x;\theta) = e^{\frac{1}{4}x_0^2}e^{\frac{1}{2}\beta x_0}\frac{D_{-\theta}(-\beta-x_0)}{D_{-\theta}(-\beta)}\frac{e^{-\frac{1}{2}x\beta - x\sqrt{\theta+\beta^2/4}}}{\sqrt{\theta+\beta^2/4}-R_\beta(\theta)}. \tag{2.7}$$

(ii) *For* $x < 0$,

$$\begin{aligned}\hat{p}(x;\theta) &= A(\theta)e^{\frac{1}{4}(x_0^2-x^2)}e^{\frac{1}{2}\beta(x_0-x)}D_{-\theta}(-\beta-x) \\ &\quad + \mathbf{1}\{x_0 < x < 0\}e^{\frac{1}{4}(x_0^2-x^2)}e^{\frac{1}{2}\beta(x_0-x)}\frac{\Gamma(\theta)}{\sqrt{2\pi}} \\ &\quad \times \left[D_{-\theta}(-\beta-x_0)D_{-\theta}(\beta+x) - D_{-\theta}(\beta+x_0)D_{-\theta}(-\beta-x)\right],\end{aligned} \tag{2.8}$$

where

$$\begin{aligned}A(\theta) = \frac{\Gamma(\theta)}{\sqrt{2\pi}}\Big[&D_{-\theta}(\beta+x_0) \\ &- \frac{D_{-\theta}(\beta)D_{-\theta}(-\beta-x_0)}{D_{-\theta}(-\beta)}\frac{\sqrt{\theta+\beta^2/4}+R_{-\beta}(\theta)}{\sqrt{\theta+\beta^2/4}-R_\beta(\theta)}\Big].\end{aligned} \tag{2.9}$$

In (2.8) $\mathbf{1}\{\cdot\}$ is the indicator function. Theorems 1 and 2 coincide if $x_0 \to 0$, and yield the Laplace transform if we start the process at the origin.

We can rewrite (2.8) in the following alternate form:

$$\begin{aligned}\hat{p}(x;\theta) &= \frac{\Gamma(\theta)}{\sqrt{2\pi}} e^{\frac{1}{4}(x_0^2-x^2)} e^{\frac{1}{2}\beta(x_0-x)} \Big[D_{-\theta}(\beta + x_>) D_{-\theta}(-\beta - x_<) \\ &\quad - \frac{D_{-\theta}(\beta)}{D_{-\theta}(-\beta)} D_{-\theta}(-\beta - x) D_{-\theta}(-\beta - x_0) \Big] \\ &\quad + e^{\frac{1}{4}(x_0^2-x^2)} e^{\frac{1}{2}\beta(x_0-x)} \frac{D_{-\theta}(-\beta - x) D_{-\theta}(-\beta - x_0)}{D_{-\theta}^2(-\beta)[\sqrt{\theta + \beta^2/4} - R_\beta(\theta)]}, \end{aligned} \tag{2.10}$$

where $x_> = \max\{x, x_0\}$ and $x_< = \min\{x, x_0\}$. The equivalence of (2.8) and (2.10) follows from the Wronskian identity

$$-\frac{\sqrt{2\pi}}{\Gamma(\theta)} = D_{-\theta}(z) D'_{-\theta}(-z) + D_{-\theta}(-z) D'_{-\theta}(z), \tag{2.11}$$

which is independent of z.

While it does not seem possible to invert the Laplace transforms in Theorems 1 and 2 to get the density $p(x,t)$ explicitly, parts of $\hat{p}$ can be inverted. For $x_0 > 0$ we note that the first part of $\hat{p}$ in the right-hand side of (2.5) inverts to

$$\frac{1}{2\sqrt{\pi t}} e^{-\frac{1}{4}\beta^2 t} e^{\frac{1}{2}\beta(x_0-x)} \Big(e^{-\frac{1}{4}(x-x_0)^2/t} - e^{-\frac{1}{4}(x+x_0)^2/t} \Big), \tag{2.12}$$

which corresponds to a Brownian motion with absorption at $x = 0$. The inversion of the second part of $\hat{p}$ in (2.5) seems less straightforward.

For $x_0 < 0$ we can invert the first term in the right-hand side of (2.10). Since $\Gamma(\theta)$ has simple poles at $\theta = -n$, $n = 0, 1, 2, \ldots$, with residues $(-1)^n/n!$, and $D_{-\theta}(\cdot)$ is an entire function of θ, the first term inverts to

$$e^{\frac{1}{4}(x_0^2-x^2)} e^{\frac{1}{2}\beta(x_0-x)} \sum_{n=0}^{\infty} D_n(-\beta - x_0) D_n(\beta + x) \frac{(-1)^n e^{-nt}}{n!\sqrt{2\pi}}. \tag{2.13}$$

This corresponds to the transient solution of an Ornstein-Uhlenbeck process, starting at x_0 at time $t = 0$ (see e.g. [17]). The remaining two terms in (2.10) represent the effects of the "interface" at $x = 0$, where the form of the drift changes. As $t \to \infty$ (2.13) approaches $\exp(-(x+\beta)^2/2)/\sqrt{2\pi}$, as only the term $n = 0$ remains, and $D_0(z) = e^{-z^2/4}$.

2.1 Relaxation time

In queueing theory, the *relaxation time* is a notion that measures the time it takes for the system to approach its steady-state behavior. There are various ways to define relaxation time, but we use the definition

$$\tau = \inf\{T : p(x,t) - p(x,\infty) = O(e^{-t/T})\}, \tag{2.14}$$

in the spirit of [1, 4, 19]. The Laplace transform $\hat{p}$ is analytic in the entire θ-plane, except for singularities in the range $\Re(\theta) \le 0$. Hence, the asymptotic behavior of $p(x,t)$ (for large

t) is determined by the singularity $\hat{\theta}$ closest to the imaginary axis. In fact, from (2.14) it follows that

$$\tau^{-1} = -\Re(\hat{\theta}). \tag{2.15}$$

The dominant singularity $\hat{\theta}$ will either be the branch point $\theta_B = -\frac{1}{4}\beta^2$ or the largest negative solution θ_P to

$$\varphi_\beta(\theta) := \sqrt{\theta + \beta^2/4} - R_\beta(\theta) = 0. \tag{2.16}$$

We have the following result.

Theorem 3. *Let $\beta_* = 1.85722...$ represent the smallest positive real solution to*

$$R_\beta(-\tfrac{1}{4}\beta^2) = 0, \quad \text{or} \quad D'_{\beta^2/4}(-\beta) = 0. \tag{2.17}$$

The dominant singularity $\hat{\theta}$ of the Laplace transform $\hat{p}(x;\theta)$ is then given by

$$\hat{\theta} = \begin{cases} \theta_B = -\frac{1}{4}\beta^2, & 0 < \beta \le \beta_*, \\ \theta_P, & \beta \ge \beta_*. \end{cases} \tag{2.18}$$

This completely determines the relaxation time as defined in (2.14). More detailed information on the distance to steady state can be obtained from investigating $\hat{p}$ in the vicinity of the dominant singularity; see Theorems 4 and 5. When $\beta \le 0$ the process is transient and the large-time behavior is still determined by θ_B.

Using the recurrence relations for parabolic cylinder functions it follows that (2.17) is equivalent to

$$\frac{-2D_{\beta^2/4}(-\beta)}{\beta D_{\beta^2/4-1}(-\beta)} = 1. \tag{2.19}$$

The left-hand side of (2.19) can be written as (see [12], p. 1064)

$$\frac{2}{\beta}\frac{\int_0^\infty x^{1-\beta^2/4}e^{-(\beta-x)^2/2}\mathrm{d}x}{\int_0^\infty x^{-\beta^2/4}e^{-(\beta-x)^2/2}\mathrm{d}x}, \quad \beta^2 < 4, \tag{2.20}$$

which is the expression derived by Gamarnik and Goldberg [9].

2.2 Limiting density

Let $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-u^2/2}\mathrm{d}u$, be the density and the distribution function of a standard normal random variable. Then we define

$$C(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \tag{2.21}$$

which is the non-degenerate limit of the delay probability. The limiting distribution of the diffusion process is (see [13])

$$p(x,\infty) = \begin{cases} C(\beta)\beta e^{-\beta x}, & x > 0, \\ C(\beta)\beta e^{-\frac{1}{2}x^2}e^{-\beta x}, & x < 0. \end{cases} \tag{2.22}$$

This also follows from our expression for the Laplace transform $\hat{p}$. Since $D_0(\beta) = e^{-\beta^2/4}$, we have $R_\beta(0) = \frac{1}{2}\beta$, and the function $\hat{p}$ has a pole at $\theta = 0$ if $\beta > 0$ (the stable case). Calculating the residue yields

$$p(x,\infty) = \frac{1}{1 - \beta R'_\beta(0)} \begin{cases} \beta e^{-\beta x}, & x > 0, \\ \beta e^{-\frac{1}{2}x^2}e^{-\beta x}, & x < 0, \end{cases} \tag{2.23}$$

and some further algebra shows that indeed $R'_\beta(0) = -\Phi(\beta)/\phi(\beta)$.

2.3 Large-time asymptotics

We give the approach to equilibrium, distinguishing the cases x, x_0 positive or negative. We note that $p(x,\infty) = 0$ if $\beta \le 0$.

Theorem 4. *Consider* $x_0 > 0$.

(i) *For* $x > 0$, $\beta < \beta_*$, *and* $\beta \neq 0$,

$$p(x,t) - p(x,\infty) \sim \frac{1}{2\sqrt{\pi}t^{3/2}} e^{-\frac{1}{4}\beta^2 t} e^{\frac{1}{2}\beta(x_0-x)} \times \left[x x_0 - \frac{x+x_0}{R_\beta(-\beta^2/4)} + \frac{1}{R_\beta^2(-\beta^2/4)} \right]. \tag{2.24}$$

(ii) *For* $x < 0$, $\beta < \beta_*$, *and* $\beta \neq 0$,

$$p(x,t) - p(x,\infty) \sim \frac{1}{2\sqrt{\pi}t^{3/2}} e^{-\frac{1}{4}\beta^2 t} e^{-\frac{1}{4}x^2} e^{\frac{1}{2}\beta(x_0-x)} \times \frac{\left[1 - x_0 R_\beta(-\beta^2/4)\right]}{R_\beta^2(-\beta^2/4)} \frac{D_{\beta^2/4}(-\beta - x)}{D_{\beta^2/4}(-\beta)}. \tag{2.25}$$

(iii) *For* $x > 0$ *and* $\beta > \beta_*$,

$$p(x,t) - p(x,\infty) \sim e^{\theta_P t} \frac{e^{\frac{1}{2}\beta(x_0-x)} e^{-(x+x_0)\sqrt{\theta_P + \beta^2/4}}}{\varphi'_\beta(\theta_P)}. \tag{2.26}$$

(iv) *For* $x < 0$ *and* $\beta > \beta_*$,

$$p(x,t) - p(x,\infty) \sim e^{\theta_P t} e^{-\frac{1}{4}x^2} e^{-\frac{1}{2}\beta x} \frac{D_{-\theta_P}(-\beta - x)}{D_{-\theta_P}(-\beta)} \frac{e^{\frac{1}{2}x_0\beta - x_0\sqrt{\theta_P + \beta^2/4}}}{\varphi'_\beta(\theta_P)}. \tag{2.27}$$

(v) *For* $x > 0$ *and* $\beta = \beta_*$,

$$p(x,t) - p(x,\infty) \sim \frac{1}{\sqrt{\pi t}} e^{-\frac{1}{4}\beta_*^2 t} e^{\frac{1}{2}\beta_*(x_0-x)}. \tag{2.28}$$

(vi) *For* $x < 0$ *and* $\beta = \beta_*$,

$$p(x,t) - p(x,\infty) \sim \frac{1}{\sqrt{\pi t}} e^{-\frac{1}{4}\beta_*^2 t} e^{-\frac{1}{4}x^2} e^{\frac{1}{2}\beta_*(x_0-x)} \frac{D_{\beta_*^2/4}(-\beta_* - x)}{D_{\beta_*^2/4}(-\beta_*)}. \tag{2.29}$$

Theorem 5. *Consider* $x_0 < 0$.

(i) *For* $x > 0$, $\beta < \beta_*$, *and* $\beta \neq 0$,

$$p(x,t) - p(x,\infty) \sim \frac{1}{2\sqrt{\pi}t^{3/2}} e^{-\frac{1}{4}\beta^2 t} e^{\frac{1}{2}\beta(x_0-x)} e^{\frac{1}{4}x_0^2} \times \frac{D_{\beta^2/4}(-\beta - x_0)}{D_{\beta^2/4}(-\beta)} \left[\frac{x}{R_\beta(-\beta^2/4)} - \frac{1}{R_\beta^2(-\beta^2/4)} \right]. \tag{2.30}$$

(ii) *For $x<0$, $\beta<\beta_*$, and $\beta\neq 0$,*

$$p(x,t)-p(x,\infty)\sim\frac{1}{2\sqrt{\pi}t^{3/2}}e^{-\frac{1}{4}\beta^2 t}e^{\frac{1}{2}\beta(x_0-x)}e^{\frac{1}{4}(x_0^2-x^2)} \times\frac{1}{R_\beta^2(-\beta^2/4)}\frac{D_{\beta^2/4}(-\beta-x)D_{\beta^2/4}(-\beta-x_0)}{D^2_{\beta^2/4}(-\beta)}. \quad (2.31)$$

(iii) *For $x>0$ and $\beta>\beta_*$,*

$$p(x,t)-p(x,\infty)\sim e^{\theta_P t}e^{\frac{1}{4}x_0^2}e^{\frac{1}{2}\beta x_0}\frac{D_{-\theta_P}(-\beta-x_0)}{D_{-\theta_P}(-\beta)}\frac{e^{-\frac{1}{2}x\beta-x\sqrt{\theta_P+\beta^2/4}}}{\varphi'_\beta(\theta_P)}. \quad (2.32)$$

(iv) *For $x<0$ and $\beta>\beta_*$,*

$$p(x,t)-p(x,\infty)\sim e^{\theta_P t}e^{\frac{1}{4}(x_0^2-x^2)}e^{\frac{1}{2}\beta(x_0-x)}\frac{D_{-\theta_P}(-\beta-x_0)D_{-\theta_P}(-\beta-x)}{D^2_{-\theta_P}(-\beta)\varphi'_\beta(\theta_P)}. \quad (2.33)$$

(v) *For $x>0$ and $\beta=\beta_*$,*

$$p(x,t)-p(x,\infty)\sim\frac{1}{\sqrt{\pi t}}e^{-\frac{1}{4}\beta_*^2 t}e^{\frac{1}{4}x_0^2}e^{\frac{1}{2}\beta_*(x_0-x)}\frac{D_{\beta_*^2/4}(-\beta_*-x_0)}{D_{\beta_*^2/4}(-\beta_*)}. \quad (2.34)$$

(vi) *For $x<0$ and $\beta=\beta_*$,*

$$p(x,t)-p(x,\infty)\sim\frac{1}{\sqrt{\pi t}}e^{-\frac{1}{4}\beta_*^2 t}e^{\frac{1}{4}(x_0^2-x^2)}e^{\frac{1}{2}\beta_*(x_0-x)} \times\frac{D_{\beta_*^2/4}(-\beta_*-x)D_{\beta_*^2/4}(-\beta_*-x_0)}{D^2_{\beta_*^2/4}(-\beta_*)}. \quad (2.35)$$

Here $\varphi'_\beta(\theta_P)=(4\theta_P+\beta^2)^{-1/2}-R'_\beta(\theta_P)$, as in (2.16). When $\beta=0$ the result is independent of x_0 and we have

$$p(x,t)\sim\frac{1}{\sqrt{\pi t}}\begin{cases}1, & x>0,\\ e^{-x^2/2}, & x<0.\end{cases} \quad (2.36)$$

2.4 Spectral properties

We now examine some properties of the spectrum of the Halfin-Whitt diffusion.

Theorem 6. *While keeping $y=x+\beta$ and $y_0=x_0+\beta$ fixed, and letting $\beta\to\infty$, the Halfin-Whitt diffusion converges to the free-space Ornstein-Uhlenbeck process with density $q(y,t)$ satisfying $q(y,0)=\delta(y-y_0)$,*

$$\frac{\partial}{\partial t}q(y,t)=\frac{\partial}{\partial y}[yq(y,t)]+\frac{\partial^2}{\partial y^2}q(y,t),\quad y\in\mathbb{R}, \quad (2.37)$$

and with solution

$$q(y,t)=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{1-e^{-2t}}}\exp\left[-\frac{(y-y_0e^{-t})^2}{2(1-e^{-2t})}\right],\quad y\in\mathbb{R}. \quad (2.38)$$

This Ornstein-Uhlenbeck process is well known to have a purely discrete spectrum, with the corresponding Laplace transform having poles at $0, -1, -2, ...$ and the eigenfunctions being Hermite polynomials. That is, (2.38) can be written as (see [17])

$$q(y,t) = \frac{e^{\frac{1}{4}(y_0^2 - y^2)}}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{D_n(y_0) D_n(y)}{n!} e^{-nt}, \quad y \in \mathbb{R}. \tag{2.39}$$

Here $D_n(y) = e^{-y^2/4} 2^{-n/2} H_n(y/\sqrt{2})$ where $H_n(\cdot)$ is the nth Hermite polynomial. Thus the spectrum of the Halfin-Whitt diffusion should approach the set $\{0, -1, -2, \ldots\}$ as β increases toward infinity. As β increases through 0 we see the appearance of a pole at 0, as β increases through $\beta_* = \beta_{*,1} \approx 1.85722$ a second pole appears in the range $(-\beta^2/4, 0)$, and this pole rapidly settles to -1 as β increases further. Further poles appear in the range $(-\beta^2/4, 0)$ at the critical values $\beta_{*,2} \approx 2.72133$, $\beta_{*,3} \approx 3.37465$, $\beta_{*,4} \approx 3.92155$, and so on. The critical values are solutions of $D'_{\beta^2/4}(-\beta) = 0$; see Figure 1.

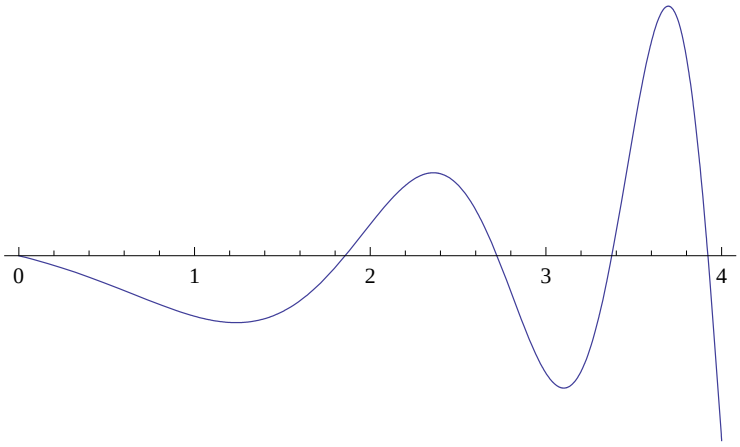

Figure 1: The function $D'_{\beta^2/4}(-\beta)$ for $\beta \in [0, 4]$.

The pole that first appears when $\beta = \beta_{*,N}$ then settles to the value $-N$ quickly, and the spectrum thus approaches that of the Ornstein-Uhlenbeck process in this manner. Let $\theta_{P,N}$ denote the location of the N-th pole (so $\theta_P = \theta_{P,1}$). Then we have the following estimate of the approach of $\theta_{P,N}$ to $-N$.

Theorem 7. *As $\beta \to \infty$ we have*

$$\theta_{P,N} + N \sim \frac{1}{(N-1)!} \frac{e^{-\beta^2/2}}{\sqrt{2\pi}} \beta^{2N-3}, \quad N = 1, 2, \ldots. \tag{2.40}$$

3 Proofs

3.1 Laplace transforms

We now present the proofs of Theorems 1 and 2. If p satisfies (2.1) its Laplace transform satisfies

$$\theta \hat{p}(x;\theta) - \delta(x - x_0) = -\frac{d}{dx}[A(x)\hat{p}(x;\theta)] + \frac{d^2 \hat{p}(x;\theta)}{dx^2}, \tag{3.1}$$

where

$$-\frac{d}{dx}[A(x)\hat{p}(x;\theta)] = \begin{cases} \beta\frac{d}{dx}\hat{p}(x;\theta), & x > 0, \\ (x+\beta)\frac{d}{dx}\hat{p}(x;\theta) + \hat{p}(x;\theta), & x < 0. \end{cases} \tag{3.2}$$

First we take $x_0 > 0$ so that $\delta(x - x_0) = 0$ in the range $x < 0$. For $x < 0$ we write $\hat{p} = e^{-x^2/4}e^{-\beta x/2}v$ and then (3.1) reduces to the differential equation (Erdelyi [7], p. 116)

$$v'' + [\tfrac{1}{2} - \theta - \tfrac{1}{4}(x+\beta)^2]v = 0, \tag{3.3}$$

whose solution is $v(x;\theta) = \alpha_1(\theta)D_{-\theta}(-\beta - x)$, where α_1 is still to be determined. Note that $D_{-\theta}(-z)$ has Gaussian decay as $z \to -\infty$, while $D_{-\theta}(z)$, which is a second solution to the parabolic cylinder equation (3.3), grows like $O(e^{z^2/4})$ as $z \to -\infty$.

For $x > 0$ the function $\hat{p} = e^{wx}$ satisfies the homogeneous version of (3.1) if

$$w^2 + \beta w - \theta = 0, \tag{3.4}$$

with solutions $w = \frac{1}{2}[-\beta - \sqrt{\beta^2 + 4\theta}]$ and $w_* = \frac{1}{2}[-\beta + \sqrt{\beta^2 + 4\theta}]$. It thus follows that

$$\hat{p}(x;\theta) = \begin{cases} \alpha_2(\theta)e^{wx} + \alpha_3(\theta)e^{w_*x}, & 0 < x < x_0 \\ \alpha_4(\theta)e^{wx}, & x > x_0, \end{cases} \tag{3.5}$$

where α_2, α_3 and α_4 still need to be determined. Continuity at $x = x_0$ yields $\hat{p}(x_0^+;\theta) = \hat{p}(x_0^-;\theta)$ and the derivative has a jump at x_0, with

$$\hat{p}(x_0^+;\theta) - \hat{p}(x_0^-;\theta) = -\int_{x_0^-}^{x_0^+} \delta(x - x_0)\mathrm{d}x = -1, \tag{3.6}$$

which translates into

$$\alpha_2 e^{wx_0} + \alpha_3 e^{w_*x_0} = \alpha_4 e^{wx_0}, \tag{3.7}$$

$$w\alpha_4 e^{wx_0} - w\alpha_2 e^{wx_0} - w_*\alpha_3 e^{w_*x_0} = -1. \tag{3.8}$$

Continuity at $x = 0$ of $\hat{p}$ and $\hat{p}_x$ yields the additional relations

$$\alpha_2 + \alpha_3 = \alpha_1 D_{-\theta}(-\beta), \tag{3.9}$$

$$w\alpha_2 + w_*\alpha_3 = -\alpha_1[D'_{-\theta}(-\beta) + \tfrac{1}{2}\beta D_{-\theta}(-\beta)]. \tag{3.10}$$

(3.7)-(3.10) give four equations for the four unknowns α_1, α_2, α_3 and α_4. Some further algebra and the definition $R_\beta(\theta) = D'_{-\theta}(-\beta)/D_{-\theta}(-\beta)$ yields

$$\alpha_1(\theta) = -\frac{1}{D_{-\theta}(-\beta)}\frac{\alpha_3(\theta)\sqrt{\beta^2 + 4\theta}}{R_\beta(\theta) + w + \beta/2}, \tag{3.11}$$

$$\alpha_2(\theta) = -\alpha_3(\theta) - \frac{\alpha_3(\theta)\sqrt{\beta^2 + 4\theta}}{R_\beta(\theta) + w + \beta/2}, \tag{3.12}$$

$$\alpha_3(\theta) = \frac{1}{\sqrt{\beta^2 + 4\theta}}e^{-x_0 w_*}, \tag{3.13}$$

$$\alpha_4(\theta) = \alpha_2(\theta) + \alpha_3(\theta)e^{x_0(w_* - w)}. \tag{3.14}$$

We thus obtain Theorem 1. Using the absolute value $|x - x_0|$ allows us to write the solution as a single formula that applies for all $x > 0$ (cf. (2.5)).

To establish Theorem 2 we note that now $\delta(x-x_0)=0$ in the range $x>0$. Thus we write

$$\hat{p}(x;\theta)=\gamma_4(\theta)e^{wx},\quad x>0, \tag{3.15}$$

and we need $\hat{p}$ to decay for $x\to-\infty$ so we write

$$\hat{p}(x;\theta)=\gamma_1(\theta)e^{-\frac{1}{4}x^2}e^{-\frac{1}{2}\beta x}D_{-\theta}(-\beta-x),\quad x<x_0<0. \tag{3.16}$$

But in the range $x_0<x<0$ the solution will involve both of the parabolic cylinder functions $D_{-\theta}(-\beta-x)$ and $D_\theta(\beta+x)$, hence

$$\hat{p}(x;\theta)=e^{-\frac{1}{4}x^2}e^{-\frac{1}{2}\beta x}\left[\gamma_2(\theta)D_{-\theta}(-\beta-x)+\gamma_3(\theta)D_{-\theta}(\beta+x)\right]. \tag{3.17}$$

The functions $\gamma_j(\theta)$ are determined by continuity of $\hat{p}$ and $\frac{d}{dx}\hat{p}$ at $x=0$, which leads to

$$\begin{aligned}\gamma_4 &= \gamma_2D_{-\theta}(-\beta)+\gamma_3D_{-\theta}(\beta), &(3.18)\\ w\gamma_3 &= -\tfrac{1}{2}\beta\gamma_4-\gamma_2D'_{-\theta}(-\beta)+\gamma_3D'_{-\theta}(\beta), &(3.19)\end{aligned}$$

continuity of $\hat{p}$ at $x=x_0$,

$$\gamma_1D_{-\theta}(-\beta-x_0) = \gamma_2D_{-\theta}(-\beta-x_0)+\gamma_3D_{-\theta}(\beta+x_0), \tag{3.20}$$

and the jump condition of $\frac{d}{dx}\hat{p}$ at $x=x_0$

$$\begin{aligned}-1=e^{-\frac{1}{4}x_0^2}e^{-\frac{1}{2}\beta x_0}\Big[&-\gamma_2D'_{-\theta}(-\beta-x_0)\\ &+\gamma_3D'_{-\theta}(\beta+x_0)+\gamma_1D'_{-\theta}(-\beta-x_0)\Big].\end{aligned} \tag{3.21}$$

Equations (3.18)-(3.21) give a 4×4 linear system whose solution leads to Theorem 2. The Wronskian identity (2.11) allows us to simplify some of the final expressions. In Theorem 2, $A(\theta)$ is the same as $\gamma_1(\theta)e^{-x_0^2/4}e^{-\beta x_0/2}$.

3.2 Asymptotic results

We now briefly derive the asymptotic results that appear in Theorems 4-7. We merely sketch the relaxation asymptotics that appear in Theorems 4-5. Consider a contour integral

$$I(t)=\frac{1}{2\pi\mathrm{i}}\int_{\mathrm{Br}}\frac{g(z)}{\sqrt{z}+f(z)}e^{zt}\mathrm{d}z. \tag{3.22}$$

Here Br is a vertical Bromwich contour in the z-plane, with the integrand analytic to the right of Br. First we assume that f and g are analytic functions of z in the half-plane $\Re(z)<-\varepsilon_0$ for some $\varepsilon_0>0$ with $g(0)\neq0$ and $f(0)\neq0$. Then the asymptotics as $t\to\infty$ are governed by the branch point at $z=0$, if $\sqrt{z}+f(z)=0$ has no solutions in the range $\Re(z)>0$. Under these assumptions we can obtain the asymptotics of (3.22) simply by expanding the analytic functions f and g about $z=0$:

$$\begin{aligned}I(t) &= \frac{1}{2\pi\mathrm{i}}\int_{\mathrm{Br}}\frac{g(0)}{f(0)}\left[1-\frac{\sqrt{z}}{f(0)}+O(z)\right]e^{zt}\mathrm{d}z\\ &= \frac{g(0)}{f(0)}\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{2\pi\mathrm{i}}\int_{\mathrm{Br}}\left[1-\frac{\sqrt{z}}{f(0)}+O(z)\right]\frac{e^{zt}}{z}\mathrm{d}z\right\}\\ &\sim -\frac{g(0)}{f^2(0)}\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathcal{L}^{-1}\left(z^{-1/2}\right)(t)\right]\\ &= -\frac{g(0)}{f^2(0)}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\sqrt{\pi t}}\right) = \frac{1}{2\sqrt{\pi}}\frac{g(0)}{f^2(0)}t^{-3/2}.\end{aligned} \tag{3.23}$$

Here $\mathcal{L}^{-1}(F(z))$ is the inverse Laplace transform of $F(z)$.

If $g(0) \neq 0$ but $f(0) = 0$ then again expanding about $z = 0$ leads to

$$I(t) \;=\; \frac{1}{2\pi \mathrm{i}} \int_{\mathrm{Br}} g(0) \left[\frac{1}{\sqrt{z}} + O(\sqrt{z}) \right] e^{zt} \mathrm{d}z \;\sim\; \frac{g(0)}{\sqrt{\pi t}}. \tag{3.24}$$

If $f(z)+\sqrt{z} = 0$ has a solution at $z = z_*$ in the range $\Re(z) > 0$, with $f'(z_*)+\frac{1}{2}z_*^{-1/2} \neq 0$ then the simple pole at z_* determines the behavior of $I(t)$ and we obtain

$$I(t) \sim \frac{g(z_*)}{f'(z_*) + \frac{1}{2} z_*^{-1/2}} e^{z_* t}. \tag{3.25}$$

We can also consider the case where the branch point and pole are close to each other. Then $f(0)$ would be small so we set $f(0) = \varepsilon$. By expanding the integrand about $z = 0$ and introducing the (large) time scale $t = \varepsilon^{-2} T$ we have

$$\begin{aligned} I(t) \;&\sim\; \frac{1}{2\pi \mathrm{i}} \int_{\mathrm{Br}} \frac{g(0)}{\sqrt{z} + \varepsilon} e^{zt} \mathrm{d}z \\ &=\; g(0) \left\{ \frac{|\varepsilon| \mathrm{sgn}(\varepsilon)}{\sqrt{\pi T}} - \frac{2\varepsilon}{\sqrt{\pi}} e^{T} \int_{\sqrt{T}\mathrm{sgn}(\varepsilon)}^{\infty} e^{-u^2} \mathrm{d}u \right\}. \end{aligned} \tag{3.26}$$

For $\varepsilon > 0$ and $T \to \infty$ we recover the behavior in (3.23), as the right-hand side of (3.26) becomes $O(T^{-3/2})$. For $\varepsilon < 0$ and $T \to \infty$ (3.26) behaves as an exponential, as in (3.25). Finally, if $\varepsilon = 0$ (3.26) becomes $g(0)/\sqrt{\pi t}$, so that (3.24) is recovered as a special case.

Since $D_{-\theta}(\cdot)$ is an entire function of θ, we immediately obtain Theorems 4 and 5. When $\beta = 0$ or $\beta = \beta_*$ the asymptotics follow from (3.24), when $\beta > \beta_*$ (3.25) applies, while for $\beta < \beta_*$ (with $\beta \neq 0$) (3.23) holds. We must simply identify $f(z)$ and $g(z)$ from Theorems 1 and 2, which necessitates that we distinguish between x, x_0 positive and negative.

To establish Theorem 6 we consider $x, x_0 < 0$, where (2.10) applies. As $\beta \to \infty$ $D_{-\theta}(-\beta)$ grows roughly as $e^{\beta^2/4}$, so that for fixed $y = x + \beta$ and $y_0 = x_0 + \beta$, and $\beta \to \infty$, the second and third terms in the right-hand side of (2.10) rapidly decay, as they contain reciprocal factors of $D_{-\theta}(-\beta)$. The first term then inverts to (2.13) which is the same as (2.39) since $D_n(z) = (-1)^n D_n(-z)$.

To derive Theorem 7 we study asymptotically, as $\beta \to \infty$, the equation

$$\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} = \sqrt{\theta + \beta^2/4}. \tag{3.27}$$

For $\beta \to \infty$ the right-hand side becomes

$$\frac{\beta}{2} \left[1 + \frac{2\theta}{\beta^2} - \frac{2\theta^2}{\beta^4} + O(\beta^{-6}) \right]. \tag{3.28}$$

In this limit the parabolic cylinder functions have the expansion

$$\begin{aligned} D_{-\theta}(-\beta) \;&=\; (-\beta)^{-\theta} e^{-\beta^2/4} \left[1 - \frac{\theta(\theta+1)}{2\beta^2} + O(\beta^{-4}) \right] \\ &\quad + \frac{\sqrt{2\pi}}{\Gamma(\theta)} \beta^{\theta-1} e^{\beta^2/4} \left[1 + O(\beta^{-2}) \right]. \end{aligned} \tag{3.29}$$

The second term is exponentially large ($O(e^{\beta^2/4})$) while the first term is exponentially small ($O(e^{-\beta^2/4})$), unless $\theta = 0, -1, -2, \ldots$. In that case $1/\Gamma(\theta)$ vanishes and then $D_n(-\beta)$ is exponentially small, and proportional to the nth Hermite polynomial. Our analysis of (3.27) will show that θ must be very close to a negative integer if (3.27) holds. If this were not the case then the second term in (3.29) would dominate and $D'_{-\theta}(-\beta)/D_{-\theta}(-\beta) \sim -\beta/2$ which could not equals (3.28) for $\beta \to \infty$.

For $\theta \to -N$ we have

$$\Gamma(\theta) = \frac{(-1)^N}{N!}\frac{1}{\theta+N} + O(1), \tag{3.30}$$

which is just the Laurent expansion of $\Gamma(\theta)$ near a pole. To balance the two parts of the right-hand side of (3.29) we need to scale $\theta + N$ to be roughly $O(e^{-\beta^2/2})$, so we define ω_N by

$$\theta + N = \omega_N e^{-\beta^2/2}. \tag{3.31}$$

Then (3.29) becomes

$$\begin{aligned} D_{-\theta}(-\beta) &= e^{-\beta^2/4}\Big\{(-\beta)^N\left[1 - \frac{N(N-1)}{2\beta^2} + O(\beta^{-4})\right] \\ &\quad +(-1)^N N!\omega_N\beta^{-N-1}\left[1 + O(\beta^{-2})\right]\Big\}, \end{aligned} \tag{3.32}$$

where θ could be replaced by $-N$ in all factors except $1/\Gamma(\theta)$. Up to an exponentially small error, (3.28) becomes

$$\frac{\beta}{2} - \frac{N}{\beta} - \frac{N^2}{\beta^3} + O(\beta^{-5}). \tag{3.33}$$

Computing the logarithmic derivative of (3.29), with the scaling (3.31), and equating the result to (3.33) leads to

$$\begin{aligned} &\frac{\beta}{2} - \frac{N}{\beta} - \frac{N^2}{\beta^3} + O(\beta^{-5}) \\ &\quad \sim \frac{-\Delta'(\beta) + \beta\Delta(\beta)/2 - \sqrt{2\pi}N!\omega_N(-\beta)^{-N}/2}{\Delta(\beta) - \sqrt{2\pi}N!\omega_N(-\beta)^{-N-1}}, \end{aligned} \tag{3.34}$$

where

$$\Delta(\beta) \;=\; e^{\beta^2/4}D_N(-\beta) \;=\; (-\beta)^N[1 - \tfrac{1}{2}N(N-1)\beta^{-2} + O(\beta^{-4})], \tag{3.35}$$

so that $\Delta'(\beta)/\Delta(\beta) = N/\beta + N(N-1)/\beta^3 + O(\beta^{-5})$ as $\beta \to \infty$. Thus the right-hand side of (3.34), after some further expansion, becomes

$$\frac{\beta}{2} - \frac{N}{\beta} - \frac{N(N-1)}{\beta^3} - \frac{\sqrt{2\pi}}{\Delta(\beta)}N!(-\beta)^{-N}\omega_N[1 + o(1)]. \tag{3.36}$$

Comparing this to (3.33) we see that the first two terms agree automatically, and agreement of the $O(\beta^{-3})$ terms forces

$$\omega_N \;\sim\; \frac{-1}{\sqrt{2\pi}N!}(-\beta)^{N-3}\Delta(\beta)N \;\sim\; \frac{\beta^{2N-3}}{\sqrt{2\pi}(N-1)!}. \tag{3.37}$$

We also see that this analysis would predict that $\omega_0 = 0$, and indeed $\theta = 0$ is a solution of (3.27) (exactly) when $\beta > 0$.

Acknowledgments

We would like to thank Ton Dieker, David Gamarnik and David Goldberg for stimulating discussions, and for sharing their unpublished work. The work of Charles Knessl was supported partially by NSF grant DMS-05-03745. The work of Johan van Leeuwaarden was supported by a VENI grant from The Netherlands Organization for Scientific Research (NWO).

References

[1] Blanc, J.P.C. and van Doorn, E.A. (1984). Relaxation times for queueing systems. In *Mathematics and Computer Science* (eds. J.W. de Bakker, M. Hazewinkel, J.K. Lenstra), North-Holland, Amsterdam, 139-162.

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

[3] Brockmeyer, E., Halstrøm, H.L. and Jensen, A. (1948). *The Life and Works of A.K. Erlang*, Trans. Danish Acad. Tech. Sci. **2**, Denmark.

[4] Cohen, J.W. (1982). *The Single Server Queue*. North Holland, Amsterdam.

[5] Darling, D.A. and Siegert, A.J.F. (1953). The first passage problem for a continuous Markov process. *Ann. Math. Statist.* **24**, 624-639.

[6] van Doorn, E.A. (1985). Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Advances in Applied Probability* **17**, 514-530.

[7] Erdelyi, A. (1953). *Higher Transcendental Functions*. Vol. 2. MacGraw-Hill, New York.

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

[9] Gamarnik, D. and Goldberg, D.A. (2008). On the exponential rate of convergence to stationarity in the Halfin-Whitt regime I: The spectral gap of the $M/M/n$ queue. Preprint.

[10] Gamarnik, D. and Momčilovic, P. (2007). Queues with many servers: The virtual waiting-time process in the QED regime. Preprint.

[11] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**, 208-227.

[12] Gradshteyn, I.S. and Ryzhik, I.M. (1980). *Table of Integrals, Series and Products*. 4th ed., Academic Press, New York.

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

[14] Janssen, A.J.E.M., van Leeuwaarden, J.S.H. and Zwart, B. (2008). Refining square root staffing by expanding Erlang C. Second revision *Operations Research*.

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

[16] Karlin, S, and McGregor, J.L. (1958). Many server queueing processes with Poisson input and exponential service times. *Pacific J. Math.* **8**, 87-118.

[17] Keilson, J. (1964). A review of transient behavior in regular diffusion and birth-death processes. *J. Applied. Prob.* **1**, 247-266.

[18] Kingman, J.F.C. (1962). On queues in which customers are served in random order. *Proc. Cambridge Philos. Soc.* **58**, 79-91.

[19] Kingman, J.F.C. (1965). The heavy traffic approximation in the theory of queues. In: *Proc. Symp. on Congestion Theory*, eds. W.L. Smith and W. Wilkinson (University of North Carolina Press, Chapel Hill) pp. 137-169.

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

[21] Mandelbaum, A. and Momčilovic, P. (2007). Queues with many servers: The virtual waiting-time process in the QED regime. Preprint.

[22] Pollaczek, F. (1931). Über zwei Formeln aus der Theorie des Wartens vor Schaltergruppen. *Elektrische Nachrichtentechnik* **8**, 256-268.

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

[24] Reed, J. (2008). The G/GI/N queue in the Halfin-Whitt regime. Preprint.