# First passage times to congested states of many-server systems in the Halfin-Whitt regime

# First passage times to congested states of many-server systems in the Halfin-Whitt regime

Brian Fralix* Charles Knessl† Johan S.H. van Leeuwaarden‡

February 14, 2013

**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 $\lambda$ grow large (taking the service rate as unity), with $\lambda = s - \beta\sqrt{s}$ and $\beta$ 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 first passage times of this hybrid diffusion process to levels in the state space that represent congested states in the original queueing system.

*Keywords*: $GI/M/s$ queue; Halfin-Whitt regime; queues in heavy traffic; diffusion process; asymptotic analysis; first passage times

*AMS* 2000 *Subject Classification*: 60K25, 60J60, 60J70, 34E05.

## 1 Introduction

Halfin and Whitt [8] 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*.

In [8] 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. 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 $\beta$ increases, capacity grows and the Halfin-Whitt diffusion will spend more time below zero.

*Clemson University, Department of Mathematical Sciences, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA. Email: bfralix@clemson.edu

*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

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

The diffusion process $(X(t))_{t\geq 0}$ can thus be employed to obtain simple approximations for the system behavior. The steady-state characteristics of the diffusion were studied in [8]. It is also of interest to study time-dependent characteristics like the mixing times, time-dependent distributions and first passage times to enhance our understanding of how the $GI/M/s$ system behaves over various time and space scales. The mixing time is closely related to the spectral gap, which for the Halfin-Whitt diffusion $(X(t))_{t\geq 0}$ has been identified by Gamarnik and Goldberg [6] building on the results of van Doorn [5] on the spectral gap of the $M/M/s$ queue. An alternative derivation of this spectral gap was presented in [12], along with expressions for the Laplace transform over time, and the large-time asymptotics for the time-dependent density. In this paper we derive results for first passage times to large levels. Such large levels typically correspond to highly congested states, in which users start receiving degraded service. An expression for the mean first passage time was derived in Maglaras and Zeevi [14]. We shall derive the Laplace transform of the first passage time density. From this Laplace transform, we can derive not only all moments, but also expressions for the first passage time density in various asymptotic regimes.

Mathematically, determining the Laplace transform of the first passage or time-dependent distributions for the present diffusion process involves analyzing a Schrödinger type equation with a piecewise parabolic potential function, or, equivalently, a Fokker-Planck equation with a piecewise linear drift. Such problems arise in a variety of other applications, such as linear systems driven by white noise [4, 3], the Kramers' problem [15] and escape over potential barriers [11]. Invariably, the solution involves the parabolic cylinder functions (see also [1, 13, 16] for more background on the parabolic cylinder function). The main results are presented in Section 2 and the proofs are given in Sections 3-5.

## 2 Main results

For the Halfin-Whitt diffusion process, define $T_x(b)$ as the first passage time out of the interval $(-\infty, b)$, starting at $x < b$ with $b > 0$. Define

$$\varphi_{x,b}(\theta) = \mathbb{E}[\mathrm{e}^{-\theta T_x(b)}], \tag{2.1}$$

so that if $P(x,t)\mathrm{d}t = \mathbb{P}\left(T_x(b) \in [t, t+\mathrm{d}t]\right)$

$$\varphi_{x,b}(\theta) = \int_0^\infty \mathrm{e}^{-\theta t} P(x,t)\mathrm{d}t, \quad \Re(\theta) > 0. \tag{2.2}$$

The first passage time density $P$ satisfies the backward Kolmogorov equation

$$P_t = A(x)P_x + \tfrac{1}{2}B(x)P_{xx}; \quad x < b,\ t > 0 \tag{2.3}$$

with $P(b,t) = \delta(t)$ (the Dirac function) and

$$A(x) = \begin{cases} -\beta, & x > 0, \\ -x - \beta, & x < 0. \end{cases} \tag{2.4}$$

We also require $P$ and $P_x$ to be continuous at $x = 0$. Here the diffusion coefficient is $B(x) = 1 + c^2$ where $c$ is the coefficient of variation for the interarrival distribution of the $GI/M/s$ system. For $GI$=$M$ we have $c = 1$, and in general we can rescale $x$ so as to make $B(x) = 2$, which we henceforth

assume.

Let $D_\nu(z)$ denote the parabolic cylinder function with index $\nu$ and argument $z$, which is defined, for example, by the integrals

$$D_\nu(z) = \frac{\mathrm{e}^{-z^2/4}}{\Gamma(-\nu)} \int_0^\infty \mathrm{e}^{-zu} \mathrm{e}^{-u^2/2} u^{-\nu-1} \mathrm{d}u, \quad \Re(\nu) < 0, \tag{2.5}$$

$$D_\nu(z) = \frac{\mathrm{e}^{z^2/4}}{i\sqrt{2\pi}} \int_{\mathcal{C}} u^\nu \mathrm{e}^{u^2/2} \mathrm{e}^{-uz} \mathrm{d}u. \tag{2.6}$$

Here, $\Gamma(\cdot)$ is the Gamma function, and the contour $\mathcal{C}$ in the second integral is a vertical Bromwich contour in the half-plane $\Re(u) > 0$. It is well known that $D_\nu(z)$ is an entire function of both index $\nu$ and argument $z$, and various properties of $D_\nu(z)$ are given in [1, Chapter 19] and [7, p. 1092-1095].

Define

$$M(\theta;\beta,b) = \cosh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right) - \frac{2D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} \frac{\sinh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right)}{\sqrt{\beta^2+4\theta}} \tag{2.7}$$

with $D'_{-\theta}(-\beta) = -\frac{d}{d\beta} D_{-\theta}(-\beta)$. Below we give expressions for $\varphi_{x,b}(\theta)$, where we must distinguish between the cases $x > 0$ and $x < 0$.

**Theorem 1.** *Let $x < 0$. Then, with $M(\theta;\beta,b)$ as defined in (2.7),*

$$\varphi_{x,b}(\theta) = \frac{1}{M(\theta;\beta,b)} \frac{D_{-\theta}(-\beta-x)}{D_{-\theta}(-\beta)} \exp\left(-\frac{\beta(b-x)}{2} + \frac{x^2}{4}\right). \tag{2.8}$$

**Theorem 2.** *Let $x > 0$. Then, with $M(\theta;\beta,b)$ as defined in (2.7),*

$$\varphi_{x,b}(\theta) = \exp\left(\frac{\beta(x-b)}{2}\right) \left[\frac{\sinh\left(\frac{x}{2}\sqrt{\beta^2+4\theta}\right)}{\sinh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right)} - \frac{1}{M(\theta;\beta,b)} \frac{\sinh\left(\frac{(x-b)}{2}\sqrt{\beta^2+4\theta}\right)}{\sinh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right)}\right]. \tag{2.9}$$

Using

$$\mathbb{E}\left[T_x(b)\right] = -\frac{\mathrm{d}}{\mathrm{d}\theta} \varphi_{x,b}(\theta)\Big|_{\theta=0}, \tag{2.10}$$

we obtain after tedious calculations the following result for the mean first passage time, which is in agreement with the result obtained in a different manner by Maglaras and Zeevi [14].

**Proposition 3.** [14, Proposition 3] *If $x > 0$ then*

$$\mathbb{E}\left[T_x(b)\right] = \frac{x-b}{\beta} + \left(\mathrm{e}^{\beta b} - \mathrm{e}^{\beta x}\right)\left[\frac{1}{\beta^2} + \frac{1}{\beta}\int_0^\infty \mathrm{e}^{\beta u - u^2/2} \mathrm{d}u\right]. \tag{2.11}$$

*If $x < 0$ then*

$$\mathbb{E}\left[T_x(b)\right] = \frac{\mathrm{e}^{\beta b} - 1 - \beta b}{\beta^2} + \frac{\mathrm{e}^{\beta b} - 1}{\beta}\int_0^\infty \mathrm{e}^{\beta u - u^2/2} \mathrm{d}u - \int_0^\infty \mathrm{e}^{\beta u - u^2/2}\left(\frac{\mathrm{e}^{ux} - 1}{u}\right) \mathrm{d}u. \tag{2.12}$$

Though (2.11) and (2.12) are already fairly simple, we give some asymptotic formulas below

that yields further insight on the magnitude of the mean passage time (the derivation is standard and therefore omitted).

**Proposition 4.** (a) *For $b \to \infty$ and $b - x \to \infty$,*

$$\mathbb{E}\left[T_x(b)\right] \sim \mathrm{e}^{\beta b}\left[\frac{1}{\beta^2} + \frac{1}{\beta}\int_0^\infty \mathrm{e}^{\beta u - u^2/2}\mathrm{d}u\right]. \tag{2.13}$$

*If $b - x = O(1)$ the above term should be multiplied by $1 - \mathrm{e}^{-\beta(b-x)}$. If $b, \beta \to \infty$ then (2.13) simplifies further to $\mathbb{E}\left[T_x(b)\right] \sim \sqrt{2\pi}\beta^{-1}\mathrm{e}^{\beta b}\mathrm{e}^{\beta^2/2}$.*
(b) *For $\beta \to -\infty$ and $x, b = O(|\beta|)$ (possibly $o(|\beta|)$),*

$$\mathbb{E}\left[T_x(b)\right] \sim \begin{cases} \frac{b-x}{-\beta}, & x \in [0, b), \\ \frac{b}{-\beta} + \log\left(1 + \frac{x}{\beta}\right), & x \in (-\infty, 0]. \end{cases} \tag{2.14}$$

(c) *For $\beta \to \infty$ with $\beta = O(b^{-1})$ and $x = O(b)$,*

$$\mathbb{E}\left[T_x(b)\right] \sim \begin{cases} \beta^{-2}[\mathrm{e}^{\beta b} - \mathrm{e}^{\beta x} + \beta(x - b)], & x/b \in [0, 1), \\ \beta^{-2}[\mathrm{e}^{\beta b} - 1 - \beta b], & x/b < 0. \end{cases} \tag{2.15}$$

We note that in (2.13) the mean first passage time is exponentially large and independent of the starting point $x$, in (2.14) it is asymptotically $O(1)$, and (2.15) represents the transition between these two cases, where $\mathbb{E}\left[T_x(b)\right] = O(b^2)$.

The Laplace transform $\varphi_{x,b}(\theta)$ is analytic in the entire $\theta$-plane, except for singularities in the range $\Re(\theta) < 0$. Hence, the asymptotic behavior of $T_x(b)$ is determined by the singularity $\theta_{\max}$ closest to the imaginary axis. In fact, from Theorems 1 and 2 it follows that $\theta_{\max}$ will be the largest negative solution to

$$D_{-\theta}(-\beta)M(\theta; \beta, b) = 0. \tag{2.16}$$

Note that (2.7) and (2.9) are invariant under the change $\sqrt{\beta^2 + 4\theta} \to -\sqrt{\beta^2 + 4\theta}$, so there is no branch point at $\theta = -\beta^2/4$. It seems impossible to find a closed-form solution to (2.16). We therefore consider several asymptotic regimes:

(i) Large levels: $b \to \infty$ and $\beta$ fixed.

(ii) Large levels and over/undercapacity: $\beta \to \pm\infty, b \to \infty$ at the same rate ($|\beta|/b$ fixed).

(iii) Small levels and undercapacity: $\beta \to -\infty$ and $b \to 0$.

Regime (i) represents the situation of reaching highly congested states, corresponding to large levels $b$. We have the following results:

**Proposition 5** (Regime (i))**.** *If $\beta < 0$ is fixed, and $b \to \infty$, then*

$$\theta_{\max} = -\frac{1}{4}\beta^2 - \frac{\pi^2}{b^2}\left[1 + \frac{2}{b}\frac{D_{\beta^2/4}(-\beta)}{D'_{\beta^2/4}(-\beta)} + O(b^{-2})\right]. \tag{2.17}$$

*If $\beta = 0$, and $b \to \infty$, then*

$$\theta_{\max} = -\frac{\pi^2}{4b^2}\left[1 - \frac{\sqrt{2\pi}}{b} + O(b^{-2})\right]. \tag{2.18}$$

*If $\beta > 0$ is fixed, and $b \to \infty$, then*

$$\theta_{\max} \sim -\frac{\beta^2 \mathrm{e}^{-\beta b}}{1 + \beta \mathrm{e}^{\beta^2/2} \int_{-\infty}^{\beta} \mathrm{e}^{-u^2/2} \mathrm{d}u}. \tag{2.19}$$

We can generalize (2.18) to the case where $b \to \infty$ and $\beta \to 0$ with $\beta b = \gamma$ fixed, where we have

$$\theta_{\max} \sim -\frac{1}{\beta^2}\left[\frac{\gamma^2}{4} + \omega(\gamma)\right] \tag{2.20}$$

where $\omega$ is the solution of the smallest absolute value to

$$\frac{\tan(\sqrt{\omega})}{\sqrt{\omega}} = \frac{\tanh(\sqrt{-\omega})}{\sqrt{-\omega}} = \frac{2}{\gamma}. \tag{2.21}$$

It follows that if $\gamma = 0$, $\omega(0) = \pi^2/4$ and then the leading term in (2.18) becomes a special case of (2.20). Also, if $\gamma = 2$, $\omega(2) = 0$, with $\omega > 0$ for $\gamma < 2$ and $\omega < 0$ for $\gamma > 2$.

From (2.19) we see that $|\theta_{\max}|$ is exponentially small, which implies exponentially large time scales. While the result (2.19) is established analytically in Section 4, it can also be seen as a consequence of the following result. Let $\Rightarrow$ denote convergence in distribution.

**Proposition 6** (Exponential limit law)**.** *Let $V$ be an exponential random variable with unit mean. Then,*

$$C\mathrm{e}^{-\beta b} T_x(b) \Rightarrow V, \quad \text{as } b \to \infty \tag{2.22}$$

*with*

$$C = \left(\frac{1}{\beta^2} + \frac{1}{\beta}\int_0^{\infty} e^{\beta u - u^2/2} du\right)^{-1} \tag{2.23}$$

As mentioned in [14], Proposition 6 can be established using a limit theorem from the theory of regenerative processes. In Section 5 we give two proofs of Proposition 6. The first proof is probabilistic and uses the theory of regenerative processes as pointed out in [14], and the second proof is analytic and uses the exact expressions for the Laplace transform.

We next give some results for the double limits $b \to \infty$ with $\beta \to \pm\infty$, and also $b \to 0$ with $\beta \to -\infty$.

**Proposition 7** (Regime (ii))**.** *If $\beta \to -\infty$ and $b \to \infty$, then*

$$\theta_{\max} \sim -\frac{1}{4}\beta^2 - \frac{\pi^2}{b^2}\left[1 + \frac{2}{b}\left(\frac{2}{-\beta}\right)^{1/3} \frac{\mathrm{Ai}(0)}{\mathrm{Ai}'(0)}\right], \tag{2.24}$$

*where* $\mathrm{Ai}(\cdot)$ *is the Airy function. If* $\beta \to \infty$ *and* $b \to +\infty$*, then*

$$\theta_{\max} \sim -\frac{\beta}{\sqrt{2\pi}} \mathrm{e}^{-b\beta} \mathrm{e}^{-\beta^2/2}. \tag{2.25}$$

**Proposition 8** (Regime (iii))**.** *If* $b \to 0$ *with* $\beta \to -\infty$*, we let* $B_* = b(-\beta)^{1/3}$ *and for* $B_*$ *fixed,*

$$\theta_{\max} \sim -\frac{1}{4}\beta^2 - \left(-\frac{\beta}{2}\right)^{2/3} \eta, \tag{2.26}$$

*where* $\eta = \eta(B_*)$ *is the minimal solution to*

$$\sqrt{-\eta} \cot\left[B_* 2^{-1/3} \sqrt{-\eta}\right] = \frac{\mathrm{Ai}'(\eta)}{\mathrm{Ai}(\eta)}. \tag{2.27}$$

We comment that (2.24) remains valid for $\beta \to -\infty$ with $b > 0$ fixed, (2.25) remains valid for $\beta \to \infty$ with $b > 0$ fixed and can be obtained as a limiting case of (2.19) (for $\beta \to \infty$).

When $B_* \to \infty$ it follows from (2.27) that $\eta \to 0$ with $\eta \sim -2^{2/3}\pi^2 B_*^{-2}$, while if $B_* \to 0^+$ we have $\eta \sim r_0 = -2.338\ldots$ where $r_0$ is the least negative root of $\mathrm{Ai}(z) = 0$. If $\eta = r_* < 0$ where $r_*$ is the least negative root of $\mathrm{Ai}'(z) = 0$, then $B_* = \frac{1}{2}\pi 2^{1/3}(-r_*)^{-1/2}$.

## 3 Proofs of Theorem 1 and Theorem 2

From (2.3) and (2.2) it follows that the Laplace transform $\varphi_{x,b}(\theta) = Q(x;\theta)$ satisfies the ODE

$$Q_{xx} + A(x)Q_x = \theta Q, \quad x < b, \tag{3.1}$$

with the boundary condition $Q(b;\theta) = 1$ and the interface conditions $Q(0^+;\theta) = Q(0^-;\theta)$ and $Q_x(0^+;\theta) = Q_x(0^-;\theta)$. For $x > 0$ we have $A(x) = -\beta$ and then (3.1) admits solutions of the form $\mathrm{e}^{\alpha x}$ where $\alpha = \alpha_\pm = \frac{1}{2}[\beta \pm \sqrt{\beta^2 + 4\theta}]$. For $x < 0$, $A(x) = -x - \beta$ and then (3.1) becomes the Hermite equation, and the only solution that decays as $x \to -\infty$ is proportional to $\mathrm{e}^{x^2/4}\mathrm{e}^{\beta x/2} D_{-\theta}(-x-\beta)$. It follows that

$$Q(x;\theta) = k_0 \exp\left(\frac{x^2}{4} + \frac{\beta x}{2}\right) \frac{D_{-\theta}(-\beta - x)}{D_{-\theta}(-\beta)}, \quad x < 0, \tag{3.2}$$

and

$$Q(x;\theta) = k_1 \mathrm{e}^{\alpha_+(\theta)(x-b)} + (1-k_1)\mathrm{e}^{\alpha_-(\theta)(x-b)}, \quad 0 < x < b, \tag{3.3}$$

where $k_0$ and $k_1$ are independent of $x$. Here we wrote the solution in (3.3) in such a way so to automatically satisfy the boundary condition $Q(b;\theta) = 1$. To determine $k_0$ and $k_1$ we can use the interface conditions at $x = 0$, which imply that

$$k_0 = k_1 \mathrm{e}^{-\alpha_+(\theta)b} + (1-k_1)\mathrm{e}^{-\alpha_-(\theta)b} \tag{3.4}$$

and

$$-k_0\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)}=\frac{1}{2}\sqrt{\beta^2+4\theta}\left[k_1\mathrm{e}^{-\alpha_+(\theta)b}-(1-k_1)\mathrm{e}^{-\alpha_-(\theta)b}\right]. \tag{3.5}$$

Solving the algebraic system in (3.4) and (3.5) for $k_0$ and $k_1$ leads to the expressions in Theorems 1 and 2.

## 4 Brief derivation of Propositions 5, 7 and 8

We discuss the various asymptotic formulas for $\theta_{\max}$. The expressions follow from routine manipulations of the parabolic cylinder functions that appear in (2.7) and (2.16). Consider the solution to $D_{-\theta}(-\beta)M(\theta;\beta,b)=0$ which is equivalent to

$$D_{-\theta}(-\beta)\cosh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right)=\frac{2}{\sqrt{\beta^2+4\theta}}D'_{-\theta}(-\beta)\sinh\left(\frac{b}{2}\sqrt{\beta^2+4\theta}\right). \tag{4.1}$$

We analyze this transcendental equation in various limits and find the least negative root $\theta_{\max}$, only sketching the main points in the calculations.

For $b\to\infty$ and $\beta>0$ we use the Taylor expansion, for $\theta\to 0$,

$$\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)}=\frac{\beta}{2}-\theta\mathrm{e}^{\beta^2/2}\int_{-\infty}^{\beta}\mathrm{e}^{-u^2/2}\mathrm{d}u+O(\theta^2) \tag{4.2}$$

and, for $\theta\to 0$ and $b\to\infty$,

$$\coth(\frac{b}{2}\sqrt{\beta^2+4\theta})=1+2\mathrm{e}^{-b\beta}[1+O(\theta,\mathrm{e}^{-b\beta})]. \tag{4.3}$$

Also, $\sqrt{\beta^2+4\theta}=(2/\beta)[1-2\theta/\beta^2+O(\theta^2)]$, so that (4.1) is equivalent to

$$\left[\frac{\beta}{2}-\theta\mathrm{e}^{\beta^2/2}\int_{-\infty}^{\beta}\mathrm{e}^{-u^2/2}\mathrm{d}u+O(\theta^2)\right]\frac{2}{\beta}\left[1-\frac{2\theta}{\beta^2}+O(\theta^2)\right]=1+2\mathrm{e}^{-b\beta}+O(\theta,\mathrm{e}^{-2b\beta}) \tag{4.4}$$

and hence

$$-2\theta\left[\frac{1}{\beta^2}+\frac{1}{\beta}\mathrm{e}^{\beta^2/2}\int_{-\infty}^{\beta}\mathrm{e}^{-u^2/2}\mathrm{d}u\right]\sim 2\mathrm{e}^{-b\beta}, \tag{4.5}$$

which leads to the exponentially small $\theta_{\max}$ in (2.19). A similar calculation applies for fixed $b>0$ and $\beta\to\infty$, which leads to (2.25) in Proposition 7.

For $b\to\infty$ with $\beta<0$ the solution $\theta_{\max}$ will be close to the apparent branch point at $\theta=-\beta^2/4$. From (4.1) if we set $\theta=-\beta^2/4-\Omega/b^2$ we obtain

$$\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)}+O(\theta+\beta^2/4)=\frac{b}{\sqrt{\Omega}}\tan(\sqrt{\Omega}). \tag{4.6}$$

If $b\to\infty$ with a fixed $\beta<0$, $\sqrt{\Omega}$ must be close to a zero of the tangent function, so that $\Omega\sim\pi^2/4$

(for the least negative solution $\theta_{\max}$). Then estimating the difference $\sqrt{\Omega} - \pi/2$ using (4.6) leads to (2.17).

If $b \to \infty$ with $\beta \to 0$ we again obtain (4.6) and for small $\beta$ we can further approximate

$$\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} \sim \frac{2}{\beta} = \frac{2b}{\gamma}. \tag{4.7}$$

Then (2.20) follows from (4.6) with $\Omega$ replaced by $\omega = b^2\theta_{\max} - \gamma^2/4$. If $\beta = 0$ we can express the parabolic cylinder functions $D_{-\theta}(0)$ and $D'_{-\theta}(0)$ in terms of Gamma functions, and (4.1) becomes

$$\frac{\tanh(b\sqrt{\theta})}{\sqrt{\theta}} = \frac{-\Gamma(\frac{\theta}{2})}{\sqrt{2}\Gamma(\frac{\theta+1}{2})}. \tag{4.8}$$

Then (2.18) follows by solving (4.8) for $\theta_{\max}$ with $b \to \infty$, where $\theta_{\max} \sim -\pi^2/(4b^2) = O(b^{-2})$. Then $\Gamma(\frac{\theta+1}{2}) \sim \sqrt{\pi}$ and $\Gamma(\frac{\theta}{2}) \sim \frac{2}{\theta}$.

If $\beta \to -\infty$ and $b \to \infty$ or $b = o(1)$, (2.24) follows by a calculation similar to (4.6), except now $D_{\beta^2/4}(-\beta)$ becomes proportional to $\mathrm{Ai}(0)$ in this limit. Finally, to obtain (2.26) we use the approximation

$$\frac{-\frac{\mathrm{d}}{\mathrm{d}\beta}D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} \sim \left(\frac{-\beta}{2}\right)^{1/3} \frac{\mathrm{Ai}'(\eta)}{\mathrm{Ai}(\eta)}, \quad \eta = \frac{\theta + \beta^2/4}{(-\beta/2)^{2/3}} \tag{4.9}$$

which applies for $-\beta \to \infty$ and $-\theta = \beta^2/4 + O(|\beta|^{2/3})$. With (4.9), (2.27) follows from (4.1). This concludes the sketched derivation of Propositions 5, 7 and 8.

# 5 Two proofs of Proposition 6

## 5.1 Probabilistic proof

The crucial observation is that the one-dimensional Halfin-Whitt diffusion process is an ergodic one-dimensional diffusion process that has, by the strong Markov property, the origin as a regeneration point. To make this formal, let $y$ be a fixed positive number, where $y < b$. This value can be used to construct a sequence of random times $\{T_k\}_{k\geq 1}$, where

$$T_1 = \inf\{t \geq 0 : X(t) = 0, \sup_{0\leq s\leq t} X(s) \geq y\} \tag{5.1}$$

and for each $n \geq 1$,

$$T_{n+1} = \inf\{t \geq T_n : X(t) = 0, \sup_{T_n\leq s\leq t} X(s) \geq y\}. \tag{5.2}$$

The Halfin-Whitt diffusion is a regenerative process with respect to these regeneration times, which form a delayed renewal process that, with probability one, has a finite number of points in each compact interval. To make matters simpler to state, we assume that $X(0) = 0$, but the procedure outlined here can be adjusted for any arbitrary initial condition.

Define $a(b)$ to be the probability that our process gets above level $b$ in the random interval $[0, T_1]$. Due to the Strong Markov property, along with the fact that $y < b$, we observe that this

probability is just the probability that a Brownian motion with drift $-\beta$ and diffusion coefficient 2 reaches level $b$ before level 0. Hence,

$$a(b) = \frac{e^{\beta y} - 1}{e^{\beta b} - 1}. \tag{5.3}$$

Moreover, the expected length of each regenerative cycle is just the expected amount of time it takes the Halfin-Whitt diffusion to reach level $y$, starting from 0, plus the expected amount of time it takes the diffusion to go from level $y$ back to 0. In other words,

$$\mathbb{E}\left[T_1\right] = (e^{\beta y} - 1)\left[\frac{1}{\beta^2} + \frac{1}{\beta}\int_0^\infty e^{\beta u - u^2/2} du\right] \tag{5.4}$$

Therefore, by [2, Theorem 4.2 on p. 181], we see that as $b \to \infty$,

$$\frac{a(b)}{\mathbb{E}\left[T_1\right]} T_x(b) \Rightarrow V \tag{5.5}$$

where $V$ is an exponential random variable with rate one. Moreover, observe that for each $y < b$

$$\frac{a(b)}{\mathbb{E}\left[T_1\right]} \quad = \quad \frac{1}{(e^{\beta b} - 1)\left[\frac{1}{\beta^2} + \frac{1}{\beta}\int_0^\infty e^{\beta u - u^2/2} du\right]} \tag{5.6}$$

which does not depend on $y$. Hence, as $b \to \infty$, we obtain (2.22). A very similar asymptotic result carries through as well for $\mathbb{E}\left[T_x(b)\right]$, due to Asmussen [2, Proposition 4.1 on p. 180]. From (2.22) it follows that the time it takes to reach a level $b$ is roughly exponential in $b$, which says that extreme congestion is not observed on relatively short time scales. For $\beta > 0$ the mean first passage time is $\mathbb{E}\left[T_x(b)\right] \sim 1/|\theta_{\max}|$ for $x$ bounded away from $b$ and $b \to \infty$.

## 5.2 Analytic proof

Consider the Laplace transforms in (2.8) and (2.9) on the scale $\theta = O(|\theta_{\max}|) = O(\mathrm{e}^{-\beta b})$ in (2.19), and then scale time as $t = T/|\theta_{\max}| = O(\mathrm{e}^{b\beta})$. Since $D_0(-\beta) = \mathrm{e}^{-\beta^2/4}$ we have

$$\frac{D'_{-\theta}(-\beta - x)}{D_{-\theta}(-\beta)} \sim \exp\left(-\frac{x\beta}{2} - \frac{x^2}{4}\right), \quad \theta \to 0. \tag{5.7}$$

Then we write $M$ in (2.7) as

$$M = \frac{\sinh\left(\frac{b}{2}\sqrt{\beta^2 + 4\theta}\right)}{\sqrt{\beta^2 + 4\theta}}\left[\sqrt{\beta^2 + 4\theta}\coth\left(\frac{b}{2}\sqrt{\beta^2 + 4\theta}\right) - \frac{2D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)}\right]. \tag{5.8}$$

For $\theta \to 0$ we use (4.3), and the recurrence relation for the parabolic cylinder function $D'_p(z) = -\frac{z}{2}D_p(z) + pD_{p-1}(z)$, which for $p \to 0$ yields

$$\frac{D'_p(z)}{D_p(z)} = -\frac{z}{2} + p\frac{D_{-1}(z)}{D_0(z)} + O(p^2). \tag{5.9}$$

Then (5.8) becomes

$$M = \frac{1}{2\beta}\mathrm{e}^{b\beta/2}\Big[(1+2\mathrm{e}^{-b\beta}+O(\theta\mathrm{e}^{-b\beta},\mathrm{e}^{-2b\beta}))(\beta+\frac{2\theta}{\beta}+O(\theta^2))-\beta+2\theta\mathrm{e}^{\beta^2/4}D_{-1}(-\beta)+O(\theta^2)\Big]$$
$$\sim \frac{1}{\beta}\mathrm{e}^{b\beta/2}\Big(\theta\Big[\frac{1}{\beta}+\mathrm{e}^{\beta^2/4}D_{-1}(-\beta)\Big]+2\beta\mathrm{e}^{-b\beta}\Big)$$
$$= \mathrm{e}^{-b\beta/2}\Big(\frac{\theta+|\theta_{\max}|}{|\theta_{\max}|}\Big), \tag{5.10}$$

$\theta = O(|\theta_{\max}|)$, as $D_{-1}(-\beta) = \mathrm{e}^{\beta^2/4}\int_{-\infty}^{\beta}\mathrm{e}^{-u^2/2}du = \mathrm{e}^{-\beta^2/4}\int_0^{\infty}\mathrm{e}^{\beta u}\mathrm{e}^{-u^2/2}du$. Then inverting the transform in (2.8), using (5.7) and (5.10), yields (with Br a vertical Bromwich contour in the half-plane $\Re(\xi) > 0$)

$$P(x,t) \sim \frac{|\theta_{\max}|}{2\pi i}\int_{\mathrm{Br}}\frac{1}{1+\xi}\mathrm{e}^{\xi T}d\xi = |\theta_{\max}|\mathrm{e}^{-T}, \tag{5.11}$$

since $\theta t = \xi T$ if $\theta = |\theta_{\max}|\xi$. This yields the exponential limit law for $x < 0$.

If $x > 0$ so that both $b \to \infty$ and $b - x \to \infty$, then an analogous calculation using (2.9) again leads to (5.11). However, if $x, b \to \infty$ in such a way that $b - x = O(1)$, then (2.9) leads to the limit

$$\varphi_{x,b}(\theta) \to \mathrm{e}^{(x-b)\beta} + \frac{|\theta_{\max}|}{\theta+|\theta_{\max}|}\Big[1-\mathrm{e}^{(x-b)\beta}\Big] \tag{5.12}$$

and this inverts to, on time scales $t = O(\mathrm{e}^{b\beta})$,

$$P(x,t) \sim \frac{|\theta_{\max}|}{2\pi i}\int_{\mathrm{Br}}\Big[\mathrm{e}^{(x-b)\beta}+\frac{1-\mathrm{e}^{(x-b)\beta}}{1+\xi}\Big]e^{T\xi}d\xi$$
$$= |\theta_{\max}|\Big[\mathrm{e}^{(x-b)\beta}\delta(T)+(1-\mathrm{e}^{(x-b)\beta})\mathrm{e}^{-T}\Big]. \tag{5.13}$$

The term proportional to $\delta(T)$ represents a probability mass on the large time scale, which corresponds to sample paths that hit $b$ in a short time (at least $t = o(e^{b\beta})$). The actual density of $P(x,t)$ does not have mass at $t = 0$, and finding the expansion of the density on shorter time ranges would require a different asymptotic analysis, and a different approximation to $M$. We refer the reader to [9, 10], where such problems were analyzed in detail for some simpler models. There both exponentially large and $t = O(1)$ time scales were considered, and they were related to one another by asymptotic matching.

## Acknowledgments

The work of Charles Knessl was supported partially by NSA grants H 98230-08-1-0102 and H 98230-11-1-0184. The work of Johan van Leeuwaarden was supported by an ERC Starting Grant.

## References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions* (10th printing), Dover, New York, 1972.

[2] S. Asmussen. Applied Probability and Queues (2nd edition), Springer-Verlag, New York, 2003.

[3] J.D. Atkinson and T.K. Caughley. Spectral density of piecewise linear first order systems excited by white noise. *Int. J. Non-Linear Mechanics*, 3:137–156, 1968.

[4] J.D. Atkinson. *Spectral density of first order piecewise linear systems excited by white noise*. PhD thesis, CalTech, 1967.

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

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

[7] I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series and Products*. 5th ed., Academic Press, New York, 1994.

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

[9] C. Knessl. Geometric optics approach to first-passage distributions: Caustic boundaries and exponentially small eigenvalues. *Studies in Applied Math* 105: 301-332, 2000.

[10] C. Knessl and Y. Yang. Asymptotic expansions for the congestion period for the $M/M/\infty$ queue. *Queueing Systems* 39: 213-256, 2001.

[11] J. Lehmann, P. Reimann, and P. Hänggi. Surmounting oscillating barriers: Path integral approach for weak noise. *Phys. Rev. E*, 62:6282–6303, 2000.

[12] J.S.H. van Leeuwaarden and C. Knessl. Transient analysis of the Halfin-Whitt diffusion. *Stochastic Processes and Their Applications* 121: 1524-1545, 2011.

[13] J.S.H. van Leeuwaarden and C. Knessl. Spectral gap of the Erlang A model in the Halfin-Whitt regime. Submitted, 2011.

[14] 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.

[15] A.N. Malakhov and A.L. Pankratov. Exact solution of Kramers' problem for piecewise parabolic potentials. *Physica A*, 229:109–126, 1996.

[16] N.M. Temme. Parabolic cylinder function. In R. F. Boisvert et al., editors, *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.