Large deviations and overflow probabilities for the general single-server queue, with applications

N.G. Duffield[*] Neil O'Connell[†]

July 11, 1994

DIAS-APG-93-30

[*] School of Mathematical Sciences, Dublin City University, Dublin 9, Ireland; Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. E-mail: duffieldn@dcu.ie

[†] Corresponding author. Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. E-mail: oconnell@stp.dias.ie

Abstract

We consider from a thermodynamic viewpoint queueing systems where the workload process is assumed to have an associated large deviation principle with arbitrary scaling: there exist increasing scaling functions $(a_t, v_t, t \in R_+)$ and a *rate function* I such that if $(W_t, t \in R_+)$ denotes the workload process, then

$$\lim_{t\to\infty} v_t^{-1} \log P(W_t/a_t > w) = -I(w)$$

on the continuity set of I. In the case that $a_t = v_t = t$ it has been argued heuristically, and recently proved in a fairly general context (for discrete time models) by Glynn and Whitt [8], that the queue-length distribution (that is, the distribution of supremum of the workload process $Q = \sup_{t\geq 0} W_t$) decays exponentially:

$$P(Q > b) \sim e^{-\delta b}$$

and the decay rate δ is directly related to the rate function I. We establish conditions for a more general result to hold, where the scaling functions are not necessarily linear in t: we find that the queue-length distribution has an exponential tail only if $\lim_{t\to\infty} a_t/v_t$ is finite and strictly positive; otherwise, provided our conditions are satisfied, the tail probabilities decay like

$$P(Q > b) \sim e^{-\delta v(a^{-1}(b))}.$$

We apply our results to a range of workload processes, including fractional Brownian motion (a model that has been proposed in the literature (see, for example, Leland *et al* [10] and Norros [15, 16]) to account for self-similarity and long range dependence) and, more generally, Gaussian processes with stationary increments. We show that the martingale upper bound estimates obtained by Kulkarni and Rolski [5], when the workload is modelled as an Ornstein-Uhlenbeck position process, are asymptotically correct. Finally we consider a non-Gaussian example, where the arrivals are modelled by a squared Bessel process.

1 Introduction

Consider a general single-server queue. For $s, t \in T$ ($T = Z_+$ or R_+), we denote by $A_{s,t}$ the amount of work that arrives to be processed in the time interval $[s, t)$, and by $S_{s,t}$ the amount of work that can be processed in the same time interval. (If more work arrives than can be processed, the surplus waits in the queue.) The *workload process* W is defined by

$$W_t := A_{-t,0} - S_{-t,0}, \tag{1}$$

and the *queue-length at time zero* is given by

$$Q = \sup_{t \geq 0} W_t. \tag{2}$$

There has been a recent flood of literature and discussion on the tail behaviour of the queue-length distribution, motivated by potential applications to the design and control of high-speed telecommunication networks. In this paper we generalise results of Glynn and Whitt [8] (see also references therein) on the problem of characterising the tails of the queue-length distribution in terms of the large deviation properties of the workload process. Our results are quite general and can, for example, be applied to processes that exhibit long range dependence, as have been proposed by Leland *et al* [10]. They can also be applied in other areas of probability theory.

We begin our discussion with the following easy inequality: for $b > 0$,

$$P(\sup_{t \geq 0} W_t > b) \geq \sup_{t \geq 0} P(W_t > b). \tag{3}$$

At first sight, this may seem a rather crude estimate. However, there are circumstances in which for large b we have

$$P(\sup_{t \geq 0} W_t > b) \simeq \sup_{t \geq 0} P(W_t > b). \tag{4}$$

The heuristics behind this claim are:

- *The principle of the largest term, or Laplace's method:* Suppose that $T = Z_+$, and consider the following calculation.

$$\begin{aligned} P(\sup_n W_n > b) &= P \bigcup_n \{W_n > b\} & (5) \\ &\leq \sum_n P(W_n > b). & (6) \end{aligned}$$

  Roughly speaking, if the probabilities $P(W_n > b)$ decay sufficiently fast in b, then the principle of the largest term applies, yielding (4).

- *Rare events occur in the most likely way:* this well known principle is the probabilitistic counterpart of the principle of the largest term. In this case it translates as "suppose that the process W is unlikely to ever reach a given level b; if W is conditioned to reach b, then it will do so at the time when this is most likely to occur".

This argument (namely that (4) can provide a good estimate) has been exploited and made rigorous in a variety of contexts: we refer the reader to the books of Dembo and Zeitouni [6] and Aldous [2] for details and references. (For a general introduction to large deviation theory, the paper of Lewis and Pfister [11] is an excellent source.) However, more relevant to our discussion here is the acute observation that using (4), the tail behaviour of the queue-length distribution can be derived from the large deviation behaviour of the workload process: this was originally proposed by Kesidis *et al* [9] and later made rigorous by Glynn and Whitt [8] (see also references therein and Chang [4] for related work).

Roughly speaking, Glynn and Whitt proved the following result. If $T = Z_+$ (discrete time) and the pair $(W_n/n, n)$ satisfy a large deviation principle with some well-behaved rate function, then

$$\lim_{b\to\infty} b^{-1} \log P(Q > b) = -\delta, \tag{7}$$

where

$$\delta = \sup\{\theta : \lambda(\theta) \leq 0\}, \tag{8}$$

and λ is the cumulant generating function defined by

$$\lambda(\theta) := \lim_{n\to\infty} n^{-1} \log E e^{\theta W_n}. \tag{9}$$

This is a very general result, and extremely useful for applications. In particular, it can be applied to any stable system with stationary arrivals and deterministic service rate, provided the arrivals process does not possess long range dependence. However, there has been some suggestion recently that real traffic can exhibit long range dependence, most notably by Leland *et al* [10], who propose *fractional Brownian motion* (with negative drift) as a canonical model for the workload process. See [13] for a historical survey and summary of the properties of this process. This has motivated us to generalise Glynn and Whitt's result. Roughly speaking, we show that if there exist increasing scaling functions (a_t, v_t) such that the pair $(W_t/a_t, v_t)$ satisfy a large deviation principle with rate function I, and if there exists a scaling function (h_t) such that the limit

$$g(c) = \lim_{t\to\infty} \frac{v(a^{-1}(c/t))}{h_t} \tag{10}$$

exists for each $c > 0$, then (under suitable hypotheses)

$$\lim_{b\to\infty} h_b^{-1} \log P(Q > b) = -\inf_{c>0} g(c) I(c). \tag{11}$$

It is not hard to check that this agrees with Glynn and Whitt's result in the case $a_t = v_t = t$ (see Lemma 2.1 below). For fractional Brownian motion (with negative drift), $a_t \equiv t$ and v_t is polynomial in t; in this case the appropriate scaling function is $h_b := v_b$. A natural generalisation of fractional Brownian motion as a model for the workload is to consider more general Gaussian processes with stationary increments: we demonstrate, as one might expect, that the tail behaviour of the queue-length distribution is governed by the asymptotic variance of the workload process.

We include two more examples. First, Kulkarni and Rolski [5] have justified the Ornstein-Uhlenbeck position process (with negative drift) as an approximation for the workload in a queue with a large number of independent bursty sources; we show that the upper bound estimates which they obtain via maringale methods are asymptotically correct. We also consider a non-Gaussian example, where the arrivals are modelled by the square of a Bessel process.

We present our results in §2, and the applications in §3.

2 Large deviations and overflow probabilities

Consider a stochastic process $(W_t,\ t \in T)$, where $T = Z_+$ or R_+, and set

$$Q = \sup_{t \geq 0} W_t. \tag{12}$$

Recall that for the general single-server queue, we are thinking of W as the workload and Q as the queue-length at time zero. In this section we characterise the tail behaviour of Q in terms of the large deviation properties of W, based on the heuristics described in the introduction. We use the terminology of Dembo and Zeitouni [6]. Our basic assumption is the following.

Hypothesis 2.1 *(i) There exists functions $a, v : T \to T$ that increase to infinity, such that for each $\theta \in R$, the cumulant generating function defined as the limit*

$$\lambda(\theta) := \lim_{t\to\infty} v_t^{-1} \log E e^{\theta v_t W_t / a_t} \tag{13}$$

exists as an extended real number. Moreover (note that λ is automatically convex) $\lambda(\cdot)$ is essentially smooth and lower semi-continuous.

(ii) There exists $\theta > 0$ for which $\lambda(\theta) < 0$.

(iii) There exists an increasing function $h : T \to T$ such that the limit

$$g(c) := \lim_{t\to\infty} \frac{v(a^{-1}(t/c))}{h_t} \tag{14}$$

exists for each $c > 0$, where

$$a^{-1}(x) := \sup\{s \in T : a(s) \leq x\}. \tag{15}$$

Hypothesis 2.1(i) ensures that the pair $(W_t/a_t, v_t)$ satisfy a large deviation principle with good rate function given by the Fenchel-Legendre transform of λ, which we denote by λ^*. This is a consequence of the Gärtner-Ellis Theorem (see, for example, [6, Theorem 2.3.6]). In other words, for any Borel set Γ,

$$\limsup_{t\to\infty} v_t^{-1} \log P(W_t/a_t \in \Gamma) \leq - \inf_{x\in\overline{\Gamma}} \lambda^*(x), \tag{16}$$

and

$$\liminf_{t\to\infty} v_t^{-1} \log P(W_t/a_t \in \Gamma) \geq - \inf_{x\in\Gamma^\circ} \lambda^*(x), \tag{17}$$

where

$$\lambda^*(x) := \sup_{\theta\in R}\{\theta x - \lambda(\theta)\}. \tag{18}$$

For $x > 0$, it follows that

$$\limsup_{t\to\infty} v_t^{-1} \log P(W_t/a_t > x) \leq -\lambda^*(x), \tag{19}$$

and

$$\liminf_{t\to\infty} v_t^{-1} \log P(W_t/a_t > x) \geq -\lambda^*(x^+). \tag{20}$$

Hypothesis 2.1(ii) is a stability condition (roughly speaking it ensures that $Q < \infty$ almost surely) and Hypothesis 2.1(iii) yields a suitable choice of scale for considering the asymptotics of $\log P(Q > b)$ when b is large.

The following lower bound result is essentially an immediate consequence of the basic inequality (3).

Theorem 2.1 *If Hypothesis 2.1 is satisfied, then*

$$\liminf_{b\to\infty} h_b^{-1} \log P(Q > b) \geq - \inf_{c>0} g(c)\lambda^*(c^+).$$

Proof. For each $c > 0$,

$$\begin{aligned}
\liminf_{b\to\infty} h_b^{-1} \log P(Q > b) &\geq \liminf_{b\to\infty} h_b^{-1} \log P(W_{a^{-1}(b/c)} > b) && (21)\\
&= g(c) \liminf_{t\to\infty} v_t^{-1} \log P(W_t/a_t > c) && (22)\\
&\geq -g(c)\lambda^*(c^+). && (23)
\end{aligned}$$

The result follows. □

To state a complimentary upper bound, we first record some hypotheses. The first is quite technical, and difficult to motivate without referring the reader to the proof of Theorem 2.2. Part (iii) is required for the application of the principle of the largest term to a countable sum. The reader may be relieved to note that Hypothesis 2.2 is automatically satisfied in the case where the scaling function v is polynomial, as is the case in most applications: this is established in the proof of Corollary 2.3.

Hypothesis 2.2 *There exists $d > 0$ such that*

(i)

$$\inf_{c>0} g(c)\lambda^*(c) = \inf_{c>d} g(c)\lambda^*(c) < \infty;$$

(ii)

$$\lim_{t\to\infty} \inf_{c>d} \frac{\lambda^*(c)v_t}{h(ca_t)} = \inf_{c>d} \lambda^*(c)g(c);$$

(iii) for each $\gamma > 0$,

$$\limsup_{b\to\infty} h_b^{-1} \log \sum_{k=[a^{-1}(b/d)]}^{\infty} e^{-\gamma v_k} \le -\inf_{c>0} g(c)\lambda^*(c);$$

(iv)

$$\limsup_{b\to\infty} h_b^{-1} \log a^{-1}(b/d) = 0.$$

The next hypothesis is not required if $T = Z_+$; in the case $T = R_+$ define, for $n \in Z_+$,

$$W_n^* := \sup_{0\le r<1} W_{n+r}. \tag{24}$$

Hypothesis 2.3 ($T = R_+$.) *Either*

$$\limsup_{n\to\infty} v_n^{-1} \log Ee^{\theta v_n(W_n^*-W_n)/a_n} = 0 \tag{25}$$

for all $\theta > 0$; or (25) holds for some $\theta > 0$, and

$$\limsup_{n\to\infty} v_n^{-1} \log P(W_n^* - W_n > xa_n) \le -\lambda^*(x) \tag{26}$$

for all $x > 0$.

Although we state the above hypothesis in terms of what happens over unit intervals, the length of the intervals can be arbitrarily small: this will be clear in the proof of Theorem 2.2. In particular—and we will make use of this fact in one of the examples—if W has stationary increments and $v_t = a_t = t$ then Hypothesis 2.3 can be replaced by the statement

There exists $\epsilon > 0$ such that either

$$Ee^{\theta \sup_{r \leq \epsilon} W_r} < \infty \tag{27}$$

for all $\theta > 0$, or (27) holds for some $\theta > 0$, and

$$\limsup_{n \to \infty} n^{-1} \log P(\sup_{r \leq \epsilon} W_r > xn) \leq -\lambda^*(x) \tag{28}$$

for all $x > 0$.

In words, the hypothesis ensures that the continuous-time process is *locally* well-behaved.

Theorem 2.2 *Suppose that Hypotheses 2.1 and 2.2 are satisfied and, if $T = R_+$, that Hypothesis 2.3 is also satisfied. Then*

$$\limsup_{b \to \infty} h_b^{-1} \log P(Q > b) \leq - \inf_{c>0} g(c)\lambda^*(c).$$

In most applications the scaling functions are polynomial: we present this case as a corollary. Note that in this case there is no need to check Hypothesis 2.2.

Corollary 2.3 *Suppose that Hypotheses 2.1 is satisfied and, if $T = R_+$, that Hypothesis 2.3 is also satisfied. If $a_t = t^a$ and $v_t = t^v$ for some $a,\ v > 0$, then*

$$\limsup_{b \to \infty} b^{-v/a} \log P(Q > b) \leq - \inf_{c>0} c^{-v/a} \lambda^*(c).$$

If λ^* is continuous we can combine this with Theorem 2.1 to get

$$\lim_{b \to \infty} b^{-v/a} \log P(Q > b) = - \inf_{c>0} c^{-v/a} \lambda^*(c).$$

The following lemma, due to Kesidis *et al* [9] and Duffield [7], reconciles our result with that of Glynn and Whitt [8] for the linear case.

Lemma 2.1 *In the above context, if Hypothesis 2.1 is satisfied and $\lambda(\theta_n) \nearrow \infty$ for any sequence of points θ_n in the interior of the effective domain of λ converging to a point on its boundary,*

$$\inf_{c>0} c^{-1}\lambda^*(c) = \sup\{\theta : \ \lambda(\theta) \leq 0\}.$$

Proof of Theorem 2.2. First suppose that $T = Z_+$. Let $d > 0$ be such that Hypothesis 2.2 is satisfied.

$$\begin{aligned} P(Q > b) &\leq P\left(\sup_{n<a^{-1}(b/d)} W_n > b\right) + P\left(\sup_{n\geq a^{-1}(b/d)} W_n > b\right) && (29)\\ &\leq a^{-1}(b/d)\sup_{c>d} P(W_{a^{-1}(b/c)} > b) + \sum_{n\geq a^{-1}(b/d)} P(W_n > b). && (30)\end{aligned}$$

By Hypothesis 2.1 we know that there exist $\theta, \epsilon > 0$ for which $\lambda(\theta) + \epsilon < 0$. Thus, by Chernoff's inequality and the definition of λ, we have for k sufficiently large,

$$\begin{aligned} P(W_k > b) &\leq e^{-\theta v_k b/a_k} E e^{-\theta v_k W_k/a_k} && (31)\\ &\leq e^{[\lambda(\theta)+\epsilon]v_k}. && (32)\end{aligned}$$

Combining this with Hypothesis 2.2(iii) we get

$$\limsup_{b\to\infty} h_b^{-1} \log \sum_{n\geq a^{-1}(b/d)} P(W_n > b) \leq -\inf_{c>0} g(c)\lambda^*(c). \qquad (33)$$

To treat the first term in (30) we appeal to the large deviation upper bound (19) and Hypothesis 2.2(ii, iv): for any $\delta > 0$,

$$\begin{aligned} &\limsup_{b\to\infty} h_b^{-1} \log \left[a^{-1}(b/d) \sup_{c>d} P(W_{a^{-1}(b/c)} > b)\right] \\ &= \limsup_{n\to\infty} \sup_{c>d} h(ca_n)^{-1} \log P(W_n/a_n > c) && (34)\\ &\leq \limsup_{n\to\infty} \sup_{c>d} h(ca_n)^{-1} v_n[\delta - \lambda^*(c)] && (35)\\ &\leq -\lim_{n\to\infty} \inf_{c>d} \frac{\lambda^*(c)v_n}{h(ca_n)} + \delta g(d) && (36)\\ &= -\inf_{c>d} g(c)\lambda^*(c) + \delta g(d) && (37)\end{aligned}$$

Since δ is arbitrary, we can combine the inequalities (30), (33) and (37), and Hypothesis 2.2(i), to obtain the result.

Now suppose $T = R_+$. If (25) holds for $\theta \le \theta^* \le \infty$, then by Hölder's inequality we have

$$\limsup_{n\to\infty} v_n^{-1} \log Ee^{\theta v_n W_n^*/a_n} \le \lambda(\theta/p), \tag{38}$$

for $0 < p < 1$, $\theta \le (1-p)\theta^*$. If (25) holds for all $\theta > 0$ we can let $p \nearrow 1$ and combine this with the trivial lower bound

$$\liminf_{n\to\infty} v_n^{-1} \log Ee^{\theta v_n W_n^*/a_n} \ge \lambda(\theta), \tag{39}$$

to get

$$\lim_{n\to\infty} v_n^{-1} \log Ee^{\theta v_n W_n^*/a_n} = \lambda(\theta). \tag{40}$$

We have thus shown that in this case Hypothesis 2.1 is satisfied by the (discrete time) process $(W_n^*,\ n \in Z_+)$ and the result follows.

Finally, if (25) holds only for some $\theta > 0$ we modify the proof of the discrete-time case: step (32) is justified by (38) and step (35) by the hypothesis (26); the rest of the proof is identical. □

Proof of Corollary 2.3. In this case we have $h_b = b^{v/a}$, $g(c) = c^{-v/a}$ and the statement follows from Theorem 2.2 provided Hypothesis 2.2 is satisfied. To check (i) we observe the following (given Hypothesis 2.1, this is standard convex analysis):

$$0 < \lambda^*(0) \equiv -\inf_{\theta\in R} \lambda(\theta) < \infty; \tag{41}$$

the interior of the effective domain of λ^* contains the origin, so we have

$$c^{-v/a}\lambda^*(c) < \infty \tag{42}$$

for $c > 0$ sufficiently close to zero. It follows that

$$\lim_{c\to 0^+} c^{-v/a}\lambda^*(c) = +\infty, \tag{43}$$

which, together with (42), implies (i).

Condition (ii) is trivial in this case.

We check (iii) by appealing to the asymptotic properties of incomplete gamma functions (see, for example, [1, p260]):

$$\begin{aligned}\limsup_{b\to\infty} b^{-v/a}\log\sum_{k=[(b/d)^{1/a}]}^{\infty} e^{-\gamma k^v} &= \limsup_{b\to\infty} b^{-v/a}\log\int_{(b/d)^{1/a}}^{\infty} e^{-\gamma t^v}\,dt &(44)\\ &= \limsup_{b\to\infty} b^{-v/a}\log\Gamma(v^{-1}, c(b/d)^{v/a}) &(45)\\ &= -\gamma d^{-v/a}, &(46)\end{aligned}$$

where

$$\Gamma(w,x) := \int_x^\infty r^{w-1}e^{-r}\,dr. \tag{47}$$

Now we can choose d sufficiently small to ensure that (iii) is satisfied.

Condition (iv) is trivial. □

3 Applications

3.1 Gaussian processes with stationary increments

Let $(Z_t,\ t \in R_+)$ be a zero-mean Gaussian process with stationary increments and covariance function

$$\Gamma(s,t) = EZ_sZ_t, \tag{48}$$

and set

$$W_t := Z_t - \mu t. \tag{49}$$

This is quite a general model for the workload process, and includes fractional Brownian motion; the practical generality is that we allow 'different levels of burstiness at different time-scales'. Gaussian processes may also be thought of as 'heavy traffic' approximations for a very large class of traffic models: for background on this topic see [3, Chapter 3] and references therein.

Setting

$$\sigma_t^2 := \Gamma(t,t), \tag{50}$$

we have

$$\begin{aligned} \lambda(\theta) &:= \lim_{t\to\infty} \frac{\sigma_t^2}{t^2} \log E e^{\theta t W_t/\sigma_t^2} && (51) \\ &= \frac{1}{2}\theta^2 - \theta\mu, && (52) \end{aligned}$$

and we see that Hypothesis 2.1 is satisfied with scaling functions $a_t = t$ and $v_t = t^2/\sigma_t^2$, provided the limit

$$g(c) := \lim_{t\to\infty} \frac{\sigma_t^2}{c^2\sigma_{t/c}^2} \tag{53}$$

exists for each $c > 0$. If σ_t^2 is such that Hypothesis 2.2 is satisfied, and if Hypothesis 2.3 is satisfied, then by Theorems 2.1 and 2.2 we have

$$\lim_{b\to\infty} \frac{\sigma_b^2}{b^2} \log P(\sup_t W_t > b) = -\inf_{c>0} g(c)(c+\mu)^2/2. \tag{54}$$

In particular, if the variance σ_t^2 is asymptotically linear in t:

$$\sigma_t^2/t \to \sigma^2 > 0 \tag{55}$$

say, as $t \to \infty$; then

$$\begin{aligned} \lim_{b\to\infty} b^{-1} \log P(Q > b) &= -\inf_{c>0} c^{-1}(c+\mu)^2/2\sigma^2 && (56) \\ &= -2\mu/\sigma^2. && (57) \end{aligned}$$

But Hypothesis 2.3 is automatically satisfied in this case: this is a consequence of a remarkable result in the theory of Gaussian processes, due to Marcus and Shepp [14], which states that for a bounded Gaussian process $X(t)$,

$$\lim_{x\to\infty} \frac{1}{x^2} \log P(\sup_t X(t) > x) = -1/2\sigma^2, \tag{58}$$

where σ^2 is the supremum of the variances of the individual $X(t)$.

3.2 Fractional Brownian motion

A special case of the above is where

$$2\Gamma(s,t) = s^{2H} + t^{2H} - |s-t|^{2H}, \tag{59}$$

for some $0 < H < 1$. In this case the process Z is called *fractional Brownian motion*. The parameter H is called the *Hurst parameter*, and when $H > 1/2$ the process exhibits long range dependence. This has been proposed as a model for the workload by Leland *et al* [10], based on observations of Ethernet traffic data. A lower bound for the tail of the queue-length distribution in this case was obtained by Norros [15, 16], using the inequality (3).

Since the scaling functions are polynomial in this case, we can apply Corollary 2.3 to get

$$\lim_{b\to\infty} b^{-2(1-H)} \log P(Q > b) = -\inf_{c>0} c^{-2(1-H)}(c+\mu)^2/2. \tag{60}$$

This is consistent with the lower bound estimate of Norros, and moreover demonstrates that it is asymptotically correct.

3.3 Ornstein-Uhlenbeck position process

Another example where the workload is modelled by a Gaussian process with stationary increments is the following. Consider a queue with constant service rate, for which the workload W_t is the position component of a stationary Ornstein-Uhlenbeck process with negative drift. Such an arrival process has been proposed by Norros *et al* [17] as a model of continuous correlated arrivals. It has been shown by Kulkarni and Rolski [5] that this arrival proceess occurs as the heavy traffic limit of superposed 2-state markov fluid sources under suitable rescaling of time and mean activity. Moreover, using martingale methods they obtain exponential upper bounds for the tail of the corresponding queue length distribution. Here we show that the exponential decay constant is equal to that obtained by the foregoing large deviation arguments, so that the upper bound estimate is asymptotically correct.

To be precise we consider the stationary Ornstein-Uhlenbeck process $(V_t,\ t \in R_+)$, defined to be the solution of the stochastic differential equation

$$dV_t = -V_t dt + \sqrt{2}(\mu/\nu)\, dB(t) \tag{61}$$

where V_0 is normally distributed with zero mean and variance $(\mu/\nu)^2$. Here B is standard Brownian motion, $\nu > 0$ is a load parameter (the case $\nu = 0$ corresponding to unit load), and μ is the service rate. The corresponding position process (with zero initial condition) is

$$Z_t = \int_0^t V_s ds, \tag{62}$$

and the workload is

$$W_t = Z_t - \mu t. \tag{63}$$

In [5] it is shown that

$$P(Q > b) \leq e^{-\nu^2/2 - \nu^2 b/\mu}. \tag{64}$$

In fact, as we now confirm,

$$\lim_{b \to \infty} b^{-1} \log P(Q > b) = -\nu^2/\mu. \tag{65}$$

To do this we note that Z is a centered Gaussian process with stationary increments, so we can use the arguments of §3.1. We begin by calculating the variance σ_t^2 of Z_t. The solution of (61) is

$$V_t = e^{-t} V_0 + e^{-t}\sqrt{2}(\mu/\nu) \int_0^t e^s \; dB(s), \tag{66}$$

and so the covariance function of V is (for $t \geq s \geq 0$)

$$\begin{aligned} \Xi(s,t) &:= EV_t V_s && (67) \\ &= e^{-(t+s)} \left(\mathrm{Var}(V_0) + 2(\mu/\nu)^2 E \left(\int_0^t e^{t'} dB(t') \int_0^s e^{s'} dB(s') \right) \right) && (68) \\ &= (\mu/\nu)^2 e^{-(t+s)} \left(1 + 2 \int_0^s e^{2s} \right) && (69) \\ &= (\mu/\nu)^2 e^{s-t}; && (70) \end{aligned}$$

hence

$$\begin{aligned} \sigma_t^2 &= 2 \int_{0 \leq s' \leq s \leq t} \Xi(s,s') \; ds ds' && (71) \\ &= 2(\mu/\nu)^2 (t + e^{-t} - 1). && (72) \end{aligned}$$

The result now follows from (57).

3.4 A non-Gaussian example

Consider the stationary diffusion X defined to be the solution of the stochastic differential equation

$$dX_t = 2\sqrt{X_t}dB_t + (d + 2\beta X_t)dt, \tag{73}$$

where $\beta < 0 < d$ and B is a standard Brownian motion. The stationary distribution of X is a gamma distribution with parameters δ and -2β. Set

$$W_t = \int_0^t X_s ds - \mu t \tag{74}$$

for some $\mu > 0$. It is not hard to check (using formulas in [18] for example) that

$$\begin{aligned} \lambda(\theta) \quad &:= \quad \lim_{t\to\infty} t^{-1} \log E e^{\theta W_t} && (75) \\ &= \begin{cases} -\frac{d}{2}[\sqrt{\beta^2 - 2\theta} + \beta] - \mu\theta & \text{if } \theta \leq \beta^2/2, \\ \infty & \text{otherwise,} \end{cases} && (76) \end{aligned}$$

and hence

$$\lambda^*(x) = \begin{cases} \frac{d\beta}{2} + \frac{\beta^2}{2}(x+\mu) + \frac{d^2}{8(x+\mu)} & \text{if } x > -\mu, \\ \infty & \text{otherwise.} \end{cases} \tag{77}$$

Note that if $\mu > -d/2\beta$ (this is required for stability) then Hypothesis 2.1 is satisfied with scaling functions $a_t = v_t = t$ and we deduce from Corollary 2.3 that provided Hypothesis 2.3 is satisfied,

$$\lim_{b\to\infty} b^{-1} \log P(\sup_t W_t > b) = d(d + 2\beta\mu)/2\mu^2. \tag{78}$$

Note that the hypothesis of Lemma 2.1 is not satisfied.

To check Hypothesis 2.3 we consider the solution Y of the stochastic differential equation

$$dY_t = 2\sqrt{Y_t}dB_t + ndt \tag{79}$$

with $Y_0 = X_0$, where $n = \lceil d \rceil$ and B is the same Brownian motion as the one in (73). Then $X_t < Y_t$ for all $t > 0$, so for any $\epsilon > 0$,

$$\sup_{r\leq\epsilon} W_r \leq \sup_{r\leq\epsilon} X_r < \sup_{r\leq\epsilon} Y_r. \tag{80}$$

The reason for introducing the process Y is that it can be conveniently represented as the square of the norm of a standard n-dimensional Brownian motion:

$$Y_t = \sum_{i=1}^{n} B_i(t)^2, \tag{81}$$

say, where B_i are independent one-dimensional Brownian motions. Without loss of generality we can assume that $B_1(0) = X_0$ and $B_i(0) = 0$ for $i > 1$: this ensures that $B_2^2, \ldots, B_n^2$ are stochastically dominated by B_1^2 (see, for example, [12, p216]). It follows that for $\theta > 0$,

$$E \exp[\theta \sup_{r \le \epsilon} X_r] \le \left[E \exp[\theta \sup_{r \le \epsilon} B_1(r)^2] \right]^n, \tag{82}$$

and since the supremum of a Brownian motion over a finite interval has Gaussian tail probabilities, its square has exponential tails and so (82) is finite for some $\theta > 0$. Finally,

$$\begin{aligned} \limsup_{t \to \infty} \frac{1}{t} \log P(\sup_{r \le \epsilon} W_r > xt) &\le \limsup_{t \to \infty} \frac{1}{t} \log P(\sup_{r \le \epsilon} B_1(r)^2 > xt) & (83) \\ &= -x/2\epsilon, & (84) \end{aligned}$$

and we can choose ϵ small enough so that $x/2\epsilon > \lambda^*(x)$ for all $x > 0$. Recalling the remark after Hypothesis 2.3 we are done.

Acknowledgements. The authors would like to thank John Lewis for many helpful discussions and comments, and for his encouragement. One of us (NO'C) was supported by grants from EOLAS and Mentec Computer Systems Ltd, under the Higher Education-Industry Cooperation Scheme. We would also like to thank the anonymous referee for helpful suggestions.

References

[1] Milton Abramowitz and Irene A. Stegun, eds. (1965). *Handbook of Mathematical Functions.* Dover, New York.

[2] David Aldous (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Applied Mathematical Scienes 77, Springer-Verlag.

[3] A.A. Borovkov (1984). *Asymptotic Methods in Queueing Theory.* Wiley, Chichester.

[4] Cheng-Shang Chang (1993). Stability, queue length and delay of deterministic and stochastic queueing networks. Preprint.

[5] V. Kulkarni and T. Rolski (1993). Fluid model driven by an Ornstein-Uhlenbeck process. Preprint.

[6] Amir Dembo and Ofer Zeitouni (1993). *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London.

[7] N.G. Duffield (1993). Exponential bounds for queues with Markovian arrivals. *Queueing Systems*, to appear.

[8] Peter W. Glynn and Ward Whitt (1993). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[9] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. Preprint.

[10] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson (1993). Ethernet traffic is self-similar: stochastic modelling of packet traffic data. Preprint.

[11] J.T. Lewis and C.-E. Pfister (1994). Thermodynamic probability theory: some aspects of large deviations. *Theor. Prob. Appl.*, to appear.

[12] Torgny Lindvall (1992). *Lectures on the Coupling Method.* Wiley.

[13] Benoit B. Mandelbrot and John W. Van Ness (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437.

[14] M.B. Marcus and L.A. Shepp (1972). Sample behaviour of Gaussian processes. Proceedings of the Sixth Berkeley Symposium.

[15] Ilkka Norros (1993). Studies on a model for connectionless traffic, based on fractional Brownian motion. Conference on Applied Probability in Engineering, Computer and Communication Sciences INRIA/ORSA/TIMS/SMAI, Paris 1993.

[16] Ilkka Norros (1994). A storage model with self-similar input. *Queueing Systems*, to appear.

[17] I. Norros, J.W. Roberts, A. Simonain and J. Virtamo (1991). The superposition of variable bitrate sources in ATM multiplexers. *IEEE J. Selected Areas in Commun.* 9:378–387.

[18] Jim Pitman and Marc Yor (1982). A decomposition of Bessel bridges. *Z. Wahr. verw. Gebeite* 59:425–457.