Economies of scale in queues with sources having power-law large deviation scalings.*

N.G. Duffield

School of Mathematical Sciences,
Dublin City University, Dublin 9, Ireland;
and
School of Theoretical Physics,
Dublin Institute for Advanced Studies,
10 Burlington Road, Dublin 4, Ireland.

September 29, 1994. Revision, March 13, 1995

AMS Classifications: Primary 60K25, Secondary 60F10, 60G18

Abstract

We analyse the queue Q^L at a multiplexer with L sources which may display long-range dependence. This includes, for example, sources modelled by fractional Brownian Motion (fBM). The workload processes W due to each source are assumed to have large deviation properties of the form $P[W_t/a(t) > x] \approx e^{-v(t)K(x)}$ for appropriate scaling functions a and v, and rate-function K. Under very general conditions,

$$\lim_{L\to\infty} L^{-1} \log P[Q^L > Lb] \quad = \quad -I(b)$$

provided the offered load is held constant, where the shape function I is expressed in terms of the cumulant generating functions of the input traffic. For power-law scalings $v(t) = t^v$, $a(t) = t^a$ (such as occur in fBM) we analyse the asymptotics of the shape function:

$$\lim_{b\to\infty} b^{-u/a} \left(I(b) - \delta b^{v/a} \right) \quad = \quad \nu_u$$

for some exponent u and constant ν depending on the sources. This demonstrates the economies of scale available through the multiplexing of a large number of such sources, by comparison with a simple approximation $P[Q^L > Lb] \approx e^{-\delta L b^{v/a}}$ based on the asymptotic decay rate δ alone. We apply this formula to Gaussian processes, in particular fBM, both alone, and also perturbed by an Ornstein-Uhlenbeck process. This demonstrates a richer potential structure than occurs for sources with linear large deviation scalings.

Keywords: Large deviations, scaling limits, ATM multiplexers, fractional Brownian Motion, effective bandwidth approximation

*To appear in Journal of Applied Probability, September 1996

1 Introduction.

In this paper we analyse the distribution of the queue length in a queue which serves the superposition of a large number L of independent streams of customers, asymptotically as L becomes large, and in the case that the arrival patterns of each stream may display long-range dependence. The motivation for this comes from the field of telecommunications, specifically the design and dimensioning of buffers in ATM (Asynchronous Transfer Mode) multiplexers. In this application the streams are packets (or cells) of data. It is projected that large numbers of streams will be multiplexed for transmission over high-speed communication links. Recent observations by Leland *et al* [16] have been interpreted as indicating the presence of long-range dependence within such streams in some cases.

Here we bring together two recent strands of work applying the theory of large deviations to queue length asymptotics: namely, large buffer asymptotics for single streams, and large L asymptotics for superposed streams. Let us review both of these.

Large buffer asymptotics: Consider a general single server queue. For $t \in T$ (here $T = \mathbf{R}_+$ or $\mathbf{Z}_+$) denote by A_t the amount of work which arrives to be processed in the interval $[-t, 0)$ and by S_t the amount which can be processed in the same interval. ($S_t = st$ for service at constant rate s). If more work arrives than can be processed, the surplus waits in the queue. The workload process W is defined by $W_0 = 0$ and

$$W_t = A_t - S_t, \tag{1}$$

and the queue of unprocessed work at time zero is

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

We recall from Duffield and O'Connell [9] the relation between the large deviation properties of the workload W and those of the queue length Q. Suppose there are increasing positive functions a and v on $\mathbf{R}_+$ which diverge at $+\infty$ such that the pair $(W_t/a(t), v(t))$ satisfies a large deviation principle with rate function K: informally

$$P[W_t/a(t) > x] \approx e^{-v(t)K(x)} \tag{3}$$

for t large. If there exists a scaling function h such that the limit

$$g(c) = \lim_{t \to \infty} \frac{v(a^{-1}(t/c))}{h(t)} \tag{4}$$

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

$$-\lim_{b \to \infty} \frac{1}{h(b)} \log P[Q > b] = \delta := \inf_{c > 0} g(c)K(c). \tag{5}$$

This result generalizes one due to Glynn and Whitt [12], who proved such a result for sources obeying (3) with the linear scaling $a(t) = v(t) = h(t) = t$. (See also Kesidis *et al*[15] and Chang [5] for related work). The above generalization allows the treatment of workload processes which have long-range dependence, such as fractional Brownian Motion: in this case the scaling functions are power laws. (We mention also the large deviation lower bound for the queue length for fBM by Norros [19]).

In all cases the proof relies on a standard method in large deviation theory. Defining the log cumulant generating function of W by $\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta)$ where $\lambda_t(\theta) = \frac{1}{v(t)} \log E[e^{W_t v(t)/a(t)}]$ then K is the Legendre-Fenchel transform λ^* of λ: $\lambda^*(x) = \sup_\theta\{x\theta - \lambda(\theta)\}$. (At this point we refer the reader to [8] for a comprehensive treatment of large deviation theory, and to [17] for a general introduction). For linear large-deviation scalings and suitably well-behaved rate-functions then δ is the unique positive solution of the equation

$$\lambda(\delta) = 0. \tag{6}$$

Large superposition asymptotics. For linear scaling functions, (5) gives rise to the effective bandwidth approximation

$$P[Q > b] \approx e^{-\delta b}. \tag{7}$$

In fact, this formula is invariant under L-fold superpositions of identical sources, provided we scale the service rate proportionately. The idea generalizes to heterogeneous superpositions. It has been widely examined as a basis for admission control in ATM networks. (See [5, 11, 13, 14, 24] for more details).

However, there is already theoretical and numerical work indicating that (7) can be inaccurate when applied to streams which have a high degree of auto-correlation. Numerical studies by Choudhury *et al* [6] have found in examples that $P[Q > b] \approx \eta e^{-\nu L} e^{-\delta b}$ where η is close to unity and ν is positive for streams with positive correlations, negative with negative correlations. Buffet and Duffield [3] have obtained the bound $P[Q > b] \leq e^{-\nu L} e^{-\delta b}$ with ν positive for positively correlated 2-state Markovian arrivals. Motivated by work of Weiss [23], Botvich and Duffield [4] have recently obtained asymptotics for the queue length Q^L for L sources with linear large deviation scalings:

$$\lim_{b\to\infty} L^{-1} \log P[Q^L > Lb] = -I(b), \tag{8}$$

where I is the *shape function* defined by

$$I(b) = \inf_{t>0} t\lambda_t^*(b/t). \tag{9}$$

(This variational formula has also been found independently by Courcoubetis and Weber [7] for $T = \mathbf{Z}_+$, and by Simonian and Guibert [22] for on-off fluid sources). The result rests on the following observations:

(a) Let W_t^L be the workload process for the L-fold superposition. Then $P[Q^L > Lb] = P[\cup_{t>0}\{W_t^L > Lb\}]$. When L is large, each term in the union has exponentially small probability, so the probability of the union is dominated by the largest probability of its terms.

(b) For fixed t, then under our hypotheses (W_t^L, L) satisfies a large deviation principle with rate function $I_t = (t\lambda_t)^*$. (For independent sources this is just Cramér's Theorem). In particular, the law of large numbers holds for W_t^L/L as $L \to \infty$.

The analysis in [4] shows that, subject to technical conditions, the asymptotics of the shape function are

$$\lim_{b\to\infty} (I(b) - \delta b) = \nu, \tag{10}$$

where

$$\nu = -\lim_{t\to\infty} t\lambda_t(\delta), \tag{11}$$

when this limit exists and is finite. These equations give rise to a modification to the effective bandwidth formula for suitably large b and L:

$$P[Q^L > b] \approx e^{-LI(b/L)} \approx e^{-\delta b} e^{-\nu L}. \tag{12}$$

This formula, and ν in particular, captures the *economies of scale* which are available through the statistical multiplexing of large numbers of sources with linear large deviation scalings. Whereas the first term in this product represent the multiplexing gain in large buffers b due to the statistical properties of the individual sources, the second reflects the gain due to the L-fold superposition. For if ν is positive then (12) indicates that statistical multiplexing becomes more advantageous (in the sense that loss ratios become smaller) as large numbers of sources are multiplexed together. ν is shown to be positive for sources with positive autocorrelations, while is it 0 for sources whose arrivals are independent. ν exists in general for Markovian sources, and can be calculated in terms of the (Laplace transform of) the Markov transition operator.

In this paper we examine the economies of scale available to sources with general scaling functions. In section 2 we generalize the basic large deviation result (8) to include sources with asymptotics as in (3). The essential reason this can be done is that under our hypotheses, the large L asymptotics described in (b) above hold irrespective of the nature of the correlations of the traffic. These are manifested rather in the form of the shape function $I(b)$, in particular through its asymptotic behaviour for large b (see below).

In section 3 we specialize to the case of power-law scaling functions: $a(t) = t^a$, $v(t) = t^v$ with $0 < v < a$. Such sources can be regarded as being more bursty than (for example) any Markov source, since they possess correlations at *all* time scales. Thus we should expect economies of scale to exist for these sources. Indeed, the first-order asymptotics (in b) are

$$P[Q^L > b] \approx e^{-\delta L (b/L)^{v/a}}. \tag{13}$$

This is the analog of the effective bandwidth approximation for linear large-deviation scalings. In distinction with (7), this approximation is *not* invariant with respect to scalings of L. Since $v/a < 1$ one sees that the approximation is decreasing in L at fixed b. Thus even at the first-order approximation there is statistical advantage in multiplexing larger numbers of sources with power-law large deviation scalings, an advantage which is absent (at this level of approximation) for linear scalings.

In section 4 we show that the asymptotics of the shape function I turn out to have a richer possible structure than for linear scalings:

$$\lim_{b\to\infty} b^{-u/a} \left(I(b) - \delta b^{v/a} \right) = \nu_u \tag{14}$$

where $u < v$ is an exponent such that $\lim_{t\to\infty} t^{v-u}(\lambda_t^* \circ h^{-1})^*(\delta)$ is finite. This yields an improvement on (13) for large b and L:

$$P[Q^L > b] \approx e^{-L[\delta(b/L)^{v/a} + \nu_u (b/L)^{u/a}]}, \tag{15}$$

the power-law analog of (12). If ν_u is positive, then there are further economies of scale: since $u < a$, then for large L, (13) will over-estimate loss probabilities.

The possibility of different values of the exponent u for power-law scalings, and compared with linear scalings ($u = 0$) can be understood as follows. The existence of the log-cumulant generating functional in the linear scaling depends the exponential decay of correlations within the arrival streams: the fastest possible. The value of ν is determined by the transient behaviour of λ_t. For power-law large deviations, these transients can decay at any faster time-scale (e.g exponential, or a faster power law): hence a range of exponents u are possible. We demonstrate this with some examples in section 5. We analyse I for general Gaussian processes with power law scalings. We apply the analysis to fractional Brownian Motion with Hurst parameter $H > 1/2$ (see [18] for terminology): its variance grows as t^{2H}, and its increments have long-range dependence. We also consider the sum of this fBM with an Ornstein-Uhlenbeck position process: the latter is the integral of a Markov process of exponentially decaying autocorrelation. In the first case there are no economies of scale (beyond those of the first-order approximation) at large buffer sizes in the sense that

$$I(b) = \delta b^{2-2H}, \tag{16}$$

while in the second,

$$\lim_{b\to\infty} b^{4H-3}\left(I(b) - \delta b^{2-2H}\right) = -2\delta\left(\frac{2s(1-H)}{2H-1}\right)^{2H-1}, \tag{17}$$

where s is the service rate.

2 A Large Deviation Principle.

We begin by stating our hypotheses concerning the workload processes, then give some examples which satisfy the hypotheses. For each $L \in \mathbf{N}$, $(W_t^L)_{t\in T}$ (where $T = \mathbf{Z}_+$ or $\mathbf{R}_+$) is a stochastic process, and $W_0^L = 0$. The queue length at time zero is

$$Q^L = \sup_{t\in T} W_t^L. \tag{18}$$

(Note that if the increments of W^L are stationary, then the distribution of Q^L is also stationary). Let v and a be increasing functions $\mathbf{R}_+ \to \mathbf{R}_+$. For $\theta \in \mathbf{R}$ define the cumulant generating function

$$\lambda_t^L(\theta) = (Lv(t))^{-1} \log E[e^{\theta W_t^L v(t)/a(t)}]. \tag{19}$$

Hypothesis 1

(i) For each $\theta \in \mathbf{R}$, the limits

$$\lambda_t(\theta) = \lim_{L\to\infty} \lambda_t^L(\theta) \qquad \text{and} \quad \lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta) \tag{20}$$

exist as extended real numbers. Moreover, the first limit exists uniformly for all t sufficiently large.

(ii) λ_t and λ are essentially smooth. (Both are automatically convex by Hölder's inequality).

(iii) There exists $\theta > 0$ for which $\lambda_t(\theta) < 0$ for all $t \in T$.

(iv) For all $\varepsilon < 0 < \xi$,

$$\lim_{n\to\infty} \limsup_{L\to\infty} L^{-1} \log \sum_{n' \geq n} e^{\varepsilon L v(n'\xi)} = -\infty. \tag{21}$$

(v) ($T = \mathbf{R}_+$) For all $t \geq r \geq 0$ define $\tilde{W}_{t,r}^L = \sup_{0<r'<r} W_{t-r'}^L - W_t^L$. Then for all $\theta \geq 0$

$$\limsup_{r\to 0} \limsup_{L\to\infty} L^{-1} \sup_{t\geq 0} \log E[e^{\theta v(t) \tilde{W}_{t,r}^L / a(t)}] \leq 0. \tag{22}$$

Remarks: if Hypotheses 1(i),(ii) are satisfied, then by the Gärtner-Ellis theorem, for each t the pair $(W_t^L/L, L)$ satisfies a large deviation principle with good rate function given by the Legendre-Fenchel transform of $\theta \mapsto v(t)\lambda_t(\theta a(t)/v(t))$ In other words, for any Borel set Γ,

$$\limsup_{L\to\infty} L^{-1} \log P(W_t^L/L \in \Gamma) \leq - \inf_{x \in \overline{\Gamma}} (v(t)\lambda_t(\cdot a(t)/v(t)))^*(x), \tag{23}$$

and

$$\liminf_{L\to\infty} L^{-1} \log P(W_t^L/L \in \Gamma) \geq - \inf_{x \in \Gamma^\circ} (v(t)\lambda_t(\cdot a(t)/v(t)))^*(x). \tag{24}$$

Here the Legendre-Fenchel transform of a function f is

$$f^*(x) := \sup_\theta \{\theta x - f(\theta)\}. \tag{25}$$

From this it follows that

$$\left(v(t)\lambda_t\left(\cdot a(t)/v(t)\right)\right)^*(x) = v(t)\lambda_t^*\left(x/a(t)\right). \tag{26}$$

By Hypothesis 1(iii), for $x \geq 0$,

$$\limsup_{L\to\infty} L^{-1} \log P(W_t^L/L > x) \leq -v(t)\lambda_t^*\left(x/a(t)\right), \tag{27}$$

and

$$\liminf_{L\to\infty} L^{-1} \log P(W_t^L/L > x) \geq -v(t)\lambda_t^*\left(x^+/a(t)\right). \tag{28}$$

Hypothesis 1(iv) is a technical growth condition. A sufficient condition for it to be satisfied is $v(t) = t^v$ with $v \geq 0$. Hypothesis 1(v) is a local regularity condition on the sample paths of the workload.

Examples. The simplest example we have in mind is where W^L is a superposition of L independent identical sources, served at rate sL. In this case

$$\lambda_t^L(\theta) = \lambda_t(\theta) = \frac{1}{v(t)} \log E[e^{\theta v(t)(A_t - st)/a(t)}] \tag{29}$$

where A is the arrival process of a single source. However, the result is not restricted to homogeneous superpositions. (See the remarks preceding Theorem 1 of [4] concerning heterogeneous superpositions the "linear" large deviation scaling $a(t) = v(t) = t$).

Theorem 1 . LOWER BOUND. *Under Hypothesis 1(i,ii), for each $b > 0$*

$$\limsup_{L\to\infty} L^{-1} \log P[\sup_{t>0} W_t^L > Lb] \geq -\inf_{t>0}\ v(t)\lambda_t^*\left(b^+/a(t)\right). \tag{30}$$

UPPER BOUND. *With the addition of Hypotheses 1(iii,iv) and also Hypotheses 1(v) for $T = \mathbf{R}_+$, then for each $b > 0$*

$$\limsup_{L\to\infty} L^{-1} \log P[\sup_{t>0} W_t^L > Lb] \leq -\inf_{t>0}\ v(t)\lambda_t^*\left(b/a(t)\right). \tag{31}$$

Proof of Theorem 1 : The proof is a modification of that of Theorem 1 of [4].

LOWER BOUND. By (28),

$$\begin{aligned}
\liminf_{L\to\infty} L^{-1} \log P[\sup_{t>0} W_t^L > Lb] &\geq \liminf_{L\to\infty} L^{-1} \sup_{t>0} \log P[W_t^L > Lb] & (32)\\
&\geq \sup_{t>0} \liminf_{L\to\infty} L^{-1} \log P[W_t^L > Lb] & (33)\\
&= \sup_{t>0} -v(t)\lambda_t^*(b^+/a(t)). & (34)
\end{aligned}$$

UPPER BOUND, $T = \mathbf{Z}_+$. For any $t, \theta > 0$ and $\theta_{t'} > 0, 0 < t' < t$,

$$\begin{aligned}
P[\sup_{t'>0} W_{t'}^L > Lb] &\leq t \max_{0<t'<t} P[W_{t'}^L > Lb] + \sum_{t'\geq t} P[W_{t'}^L > Lb] & (35)\\
&\leq t \max_{0<t'<t} e^{-Lv(t')\left(b\theta_{t'}/a(t') - \lambda_{t'}^L(\theta_{t'})\right)} + \sum_{t''\geq t} e^{Lv(t'')\lambda_{t''}^L(\theta)} & (36)
\end{aligned}$$

by Chebychev's inequality. Since $\lambda_t^L(\theta) \to \lambda_t(\theta)$ uniformly in t, $\lambda_t(\theta) \to \lambda(\theta)$ and $\lambda(\theta) < 0$ on $(0, \delta)$, we can find $\theta > 0$ and $\varepsilon < 0$ such that $\lambda_t^L(\theta) < \varepsilon$ for all L, t sufficiently large. Taking logarithms, dividing by L, taking the lim sup as $L \to \infty$ and finally taking the infimum over the $\theta_{t'}$ we obtain

$$\limsup_{L\to\infty} L^{-1} \log P[\sup_{t'>0} W_{t'}^L > Lb] \leq \max\left(\max_{0<t'<t} \left(-v(t')\lambda_{t'}^*(b/a(t'))\right), \limsup_{L\to\infty} L^{-1} \log \sum_{t''\geq t} e^{\varepsilon L v(t'')}\right) \tag{37}$$

By Hypothesis 1(iv) we can take the limit $t \to \infty$ and obtain the stated result.

Upper Bound, $T = \mathbf{R}_+$. For any $\epsilon > 0$ and $n \in \mathbf{N}$ define

$$\hat{W}_n^L = \sup_{n\epsilon \leq t < (n+1)\epsilon} W_t^L \quad \text{and} \quad \hat{\lambda}_n^L = (v(n\epsilon)L)^{-1} \log E[e^{\theta v(n\epsilon)\hat{W}_n^L/a(n\epsilon)}]. \tag{38}$$

By Hölder's inequality then for any p in $(0,1)$:

$$\hat{\lambda}_n^L(\theta) \leq p\lambda_{n\epsilon}^L(\theta/p) + (1-p)(v(n\epsilon)L)^{-1} \log E[e^{\theta v(n\epsilon)\tilde{W}_{n\epsilon,\epsilon}^L/((1-p)a(n\epsilon))}], \tag{39}$$

with $\tilde{W}^L$ as in Hypothesis 1(v). According to this, for any $p \in (0,1)$ we can make the second term of the right hand side of (39) as small as we like by choosing ϵ sufficiently small then L sufficiently large. Thus we can repeat the steps (35) and (36) with ϵ and p fixed, take the limits $t \to \infty$ then $\epsilon \to 0$ to obtain

$$\begin{aligned} \limsup_{L\to\infty} L^{-1} \log P[\sup_{t>0} W_t^L > Lb] &\leq \limsup_{L\to\infty} L^{-1} \log P[\sup_{n>0} \hat{W}_n^L > Lb] & (40)\\ &\leq p \sup_{t>0} -v(t)\lambda_t^* \left(b/a(t)\right), & (41) \end{aligned}$$

since for any function f, $(pf(\cdot/p))^*(x) = pf^*(x)$, and finally let $p \nearrow 1$ to get the stated result. ■

3 The shape function for power-law scalings.

With the assumptions of the previous section, define the **shape function**, I on $\mathbf{R}_+$ by

$$I(b) = \inf_{t>0} v(t)\lambda_t^* \left(b/a(t)\right). \tag{42}$$

Then according to Theorem 1 we have, for suitably large b and L,

$$P[Q^L > b] \approx e^{-LI(b/L)}. \tag{43}$$

In this section we investigate the form of I for the case of **power-law** large deviation scalings, which we define to occur when Hypotheses 1(i,ii,iii) hold with

$$v(t) = t^v, \quad a(t) = t^a, \quad \text{with } a > v > 0. \tag{44}$$

For simplicity we present only with the case $T = \mathbf{R}_+$. (As in [4], the case $T = \mathbf{Z}_+$ can be treated under additional technical assumptions).

Define $h : \mathbf{R}_+ \to \mathbf{R}_+$ by $h = v \circ a^{-1}$.

Hypothesis 2 *(i) There exists a unique $\delta > 0$ in the interior of the domain of $(\lambda^* \circ h^{-1})^*$ such that $(\lambda^* \circ h^{-1})^*(\delta) = 0$. (Such a δ is automatically unique).*

(ii) The limit $\nu := -\lim_{t\to\infty} t^v(\lambda_t^* \circ h^{-1})^*(\delta)$ exists and is finite.

(iii) λ_t and λ are closed.

Note that 2(iii) follows from 1(ii) if in addition $\lambda(\theta_n) \to \infty$ for any sequence (θ_n) converging to a point on the boundary of the effective domain of λ (and similarly for λ_t)

Theorem 2 *($T = \mathbf{R}_+$). Assume power-law large deviation scalings. Under Hypothesis 2*

$$\lim_{b\to\infty} \left(I(b)) - \delta b^{v/a}\right) = \nu. \tag{45}$$

Proof of Theorem 2: . It is convenient to define $J = I \circ h^{-1}$, change the time variable to $c = v(t)$ and set

$$f_c(x) = \begin{cases} \lambda^*_{v^{-1}(c)}(h^{-1}(x)) & x \in [0,\infty) \cap h(\operatorname{dom} \lambda^*_{v^{-1}(c)}) \\ +\infty & \text{otherwise} \end{cases}, \tag{46}$$

and define f similarly in terms of λ. (Note that $0 \in \operatorname{dom} f_c$ and $\operatorname{dom} f$ by Hypothesis 1(iii)). With these changes

$$J(b) = \inf_{c>0} c f_c(b/c), \tag{47}$$

and δ is the unique positive solution of the equation

$$f^*(\delta) = 0. \tag{48}$$

Now observe that

$$\lim_{c\to\infty} f^*_c = f^* \quad \text{pointwise on} \quad \operatorname{int}\operatorname{dom} f^*, \tag{49}$$

and that f_c and f are essentially smooth. This follows from Lemma 1 in Appendix A and the the duality Theorem 26.3 of [20], if we can show that $\lim_{c\to\infty} f_c = f$, pointwise on $\operatorname{int}\operatorname{dom} f$, and that f_c and f are essentially strictly convex. This latter convergence follows from the fact that $\lim_{t\to\infty} \lambda^*_t = \lambda^*$ pointwise on $\operatorname{int}\operatorname{dom}\lambda^*$, by Hypothesis 1(i) and Lemma 1. The convexity follows from the fact that λ^*_t and λ^* are essentially strictly convex (by the duality theorem and Hypothesis 1(ii)) and increasing (by Hypothesis 1(iii)) and that h^{-1} is strictly convex on $\mathbf{R}^+$ (by (44)).

Let

$$\beta(c) = c(f^*_c)'(\delta). \tag{50}$$

By Hypothesis 1(iii), λ^* is increasing and bounded away from 0 on $\mathbf{R}_+$, and hence $f^*(0)$ is negative. So by convexity of f^*, $(f^*)'(\delta)$ is positive. From Lemma IV.6.3 of [10] $\lim_{c\to\infty}(f^*_c)'(\delta) = (f^*)'(\delta)$ and so $\beta(c)$ is asymptotically linear with finite limiting positive gradient $(f^*)'(\delta)$ as $c \to \infty$. Denote by κ the left-inverse of β:

$$\kappa(b) = \sup\{c \mid \beta(c) \le b\}. \tag{51}$$

Then $\kappa(b) \to +\infty$ as $b \to \infty$.

We now obtain the following *upper bound*:

$$\inf_{c>0} c f_c(b/c) - \delta b \le \kappa(b) f_{\kappa(b)}(b/\kappa(b)) - \delta b = -\kappa(b) f^*_{\kappa(b)}(\delta). \tag{52}$$

Since by Hypothesis 2(iii) λ_t is closed, so is λ^*_t and hence also f_c; the upper bound follows from

$$\begin{aligned} f_{\kappa(b)}(b/\kappa(b)) &= f^{**}_{\kappa(b)}(b/\kappa(b)) && \text{(convexity and closedness of } f_c) && (53) \\ &= b\delta/\kappa(b) - f^*_{\kappa(b)}(\delta) && \text{(by (50) and (51))}. && (54) \end{aligned}$$

Thus $\limsup_{b\to\infty}(J(b)-\delta b)\le\nu$ since $\lim_{b\to\infty}\kappa(b)=+\infty$.

To obtain the corresponding *lower bound*, suppose $\inf_{c>0} cf_c(b/c)$ is attained at $\hat\kappa(b)$. (If the infimum is not achieved, then one can work with $\hat\kappa_\varepsilon(b)$ for which the infimum is approximated to within $\varepsilon>0$, then take $\varepsilon\searrow 0$ at the end). Then

$$J(b)-\delta b=\hat\kappa(b)f_{\hat\kappa(b)}(b/\hat\kappa(b))-\delta b\ge-\hat\kappa(b)f^*_{\hat\kappa(b)}(\delta) \tag{55}$$

and so

$$\liminf_{b\to\infty}(J(b)-\delta b)\ge\nu \tag{56}$$

provided $\lim_{b\to\infty}\hat\kappa(b)=+\infty$. But if this is not the case then $\hat\kappa(b)$ is bounded, and so we obtain a contradiction with the upper bound if we can show that $\lim_{b\to\infty}(cf_c(b/c)-\delta b)=+\infty$ for every fixed $c>0$. But this is true since $b\mapsto cf_c(b/c)-\delta b$ is essentially strictly convex, and, by (50) achieves its infimum at $\beta(c)<\infty$. ■

We finish this section by making contact with a result on the asymptotic decay constant for the queue length distribution for a single source, as given in Theorems 2.1 and 2.2 of [9]. In the simple example before the proof of our Theorem 1, then subject to appropriate technical conditions the asymptotics for the distribution of the queue due to a single source with power-law large deviation scalings are

$$-\lim_{b\to\infty}\frac{1}{h(b)}\log P[\sup_{t\ge0}W_t>b]=\hat\delta:=\inf_{c>0}h(c^{-1})\lambda^*(c). \tag{57}$$

Theorem 3 *For power-law large deviations, under Hypothesis 2(i), $\delta=\hat\delta$.*

Proof of Theorem 3: Rewrite $\hat\delta=\inf_{b>0}b^{-1}f(b)$. First we show $\delta\le\hat\delta$. From (48)

$$0=f^*(\delta)=\sup_{c\ge0}(c\delta-f(c))\ge b\delta-f(b) \tag{58}$$

for any $b>0$. Thus $\delta\le b^{-1}f(b)$ for any $b>0$, and hence $\delta\le\hat\delta$. We complete the proof by showing that $\delta\ge\hat\delta$. Suppose the supremum in (58) is attained at $\tilde b$. Then $\delta=\tilde b^{-1}f(\tilde b)\ge\hat\delta$. Otherwise, for all $\varepsilon>0$ there exists $\tilde b_\varepsilon$ such that $0=(\delta+\varepsilon)\tilde b_\varepsilon-f(\tilde b_\varepsilon)$. Thus $\delta+\varepsilon=\tilde b_\varepsilon^{-1}f(\tilde b_\varepsilon)\ge\hat\delta$. But ε is arbitrary, so the result follows. ■

The large-buffer asymptotics for the L-fold superposition in our example (with proportionally scaled service rate) are found be observing that the corresponding log-cumulant generating function is $L\lambda$. Thus

$$\begin{aligned}
-\lim_{b\to\infty}\frac{1}{h(b)}\log P[Q^L>b] &= \hat\delta_L && (59)\\
&:= \inf_{c>0}h(c^{-1})(L\lambda)^*(c) && (60)\\
&= L\inf_{c>0}h(c^{-1})\lambda^*(c/L) && (61)\\
&= L^{1-v/a}\inf_{c>0}h(c^{-1})\lambda^*(c) \qquad \text{since } h(t)=t^{v/a} && (62)\\
&= L^{1-v/a}\hat\delta. && (63)
\end{aligned}$$

This is the basis of the approximation (13).

4 Finer asymptotics for the shape function.

In fact, as we shall see in section 5.3, it is not too difficult to construct simple examples in which Hypothesis 2(ii) is not satisfied for finite ν. This is is contrast to the case of linear large deviation scalings, where ν is finite in many common cases—Markovian arrivals for example (see section 4 of [4]). This motivates us to generalize the hypothesis and obtain finer details on the asymptotics of I.

Hypothesis 3 *(i) For some* $t_0 > 0$, $\liminf_{b\to\infty} \inf_{0<t<t_0} v(t)\lambda^*_{bt}(1/a(t)) \geq \delta$.

(ii) $\nu_u := -\lim_{t\to\infty} t^{v-u}(\lambda^*_t \circ h^{-1})^*(\delta)$ *exists and is finite for some* $u \in (0, v)$.

Theorem 4 *Assume Hypotheses 1, 2(i,iii), 3 and power-law large deviation scalings. Then*

$$\lim_{b\to\infty} b^{-u/a}\left(I(b) - \delta b^{v/a}\right) = \hat{\nu}_u := \nu_u \left[\left((\lambda^* \circ h^{-1})^*\right)'(\delta)\right]^{-u/v}. \tag{64}$$

Proof of Theorem 4: Using the change of time variables as in (46), then $t^{v-u}\lambda^*_t(h^{-1}(\delta)) = c^{1-u/v} f^*_c(\delta)$. Below we shall prove that Hypothesis 3(i) implies that $\lim_{b\to\infty} \kappa(b)/\hat{\kappa}(b) = 1$. Then by Hypothesis 3(ii), (50) and (51),

$$\lim_{b\to\infty} b/\kappa(b) = \lim_{b\to\infty} b/\hat{\kappa}(b) = (f^*)'(\delta) = \left((\lambda^*_t \circ h^{-1})^*\right)'(\delta). \tag{65}$$

Combining these with the bounds (52) and (55), taking the limit $b \to \infty$ and using Hypothesis 3(ii) we get

$$\lim_{b\to\infty} b^{-u/v}\left(J(b) - \delta b\right) = \hat{\nu}_u \tag{66}$$

from which the statement of the theorem follows since $J(b) = I(b^{a/v})$.

It remains to prove that Hypothesis 3(i) implies that $\lim_{b\to\infty} \kappa(b)/\hat{\kappa}(b) = 1$. Write $\inf_{c>0} cf_c(b/c) = b \inf_{z>0} zf_{bz}(1/z)$. Note by Hypothesis 1(i), $\lim_{b\to\infty} zf_{zb}(1/z) = zf(1/z)$ pointwise, so by Theorem 3, and Hypothesis 3(i),

$$\delta = \inf_{z>0} zf(1/z) \geq \lim_{b\to\infty} \inf_{z\geq d} zf_{zb}(1/z), \tag{67}$$

for d sufficiently small. But by Hypothesis 1(iii), $f_c(0)$ and $f(0)$ are strictly positive, so

$$\lim_{z\to\infty} zf_{bz}(1/z) = +\infty. \tag{68}$$

Hence $\inf_{z\geq d} zf_{zb}(1/z)$ is achieved at some $\hat{z}(b) = \hat{\kappa}(b)/b$.

Now $z \mapsto zf(1/z)$ is convex on since for $\mu = 1 - \mu' \in [0,1]$, $z, z' > 0$

$$\begin{aligned}
(\mu z + \mu' z')f(1/(\mu z + \mu' z')) &= \left((\mu z + \mu' z')f^*(\cdot)\right)^*(1) && (69)\\
&= \inf_{s\in[0,1]} \left\{(\mu z f^*)^*(s) + (\mu' z' f^*)^*(1-s)\right\} && (70)\\
&= \inf_{s\in[0,1]} \left\{\mu z f\left(s/(\mu z)\right) + \mu' z' f\left((1-s)/(\mu' z')\right)\right\} && (71)\\
&\leq \mu z f(1/z) + \mu' z' f(1/z'), && (72)
\end{aligned}$$

and furthermore essentially strictly convex, having an affine portion only if f does: but we have shown the latter to be essentially strictly convex. Thus choosing $d \leq d_0 := 1/(f^*)'(\delta)$, we see from Theorem 3 that $\inf_{z \geq d} zf(1/z)$ is achieved at $\hat{z} := 1/(f^*)'(\delta) = \lim_{b\to\infty} \kappa(b)/b$. Thus we need only show that $\lim_{b\to\infty} \hat{z}(b) = \hat{z}$. But this follows, since by Hypothesis 1(i), $f_{bz}(1/z)/f(1/z)$ converges uniformly on $[d_0, \infty)$ to 1 as $b \to \infty$, while (67) and (68) together preclude $\hat{z}(b)$ having limit points at $+\infty$ as $b \to \infty$. ■

Remark: Suppose $f(1/z)$ and all $f_b(1/z)$ (for sufficiently large b) are finite as $z \to \infty$. Then a sufficient condition for Hypothesis 3(i) is that for some $d > 0$

$$(0, d) \ni z \mapsto \frac{f_{bz}(1/z)}{f(1/z)} \qquad \text{is non-increasing,} \tag{73}$$

(or equivalently: $(0, t_0) \ni t \mapsto \lambda^*_{bt}(1/a(t))/\lambda^*(1/a(t))$ is non-increasing for some t_0), for then

$$\liminf_{b\to\infty} \inf_{z\in(0,d)} zf_{bz}(1/z) \geq \inf_{z\in(0,d)} zf(1/z) \liminf_{b\to\infty} \inf_{z\in(0,d)} \frac{f_{bz}(1/z)}{f(1/z)} \tag{74}$$

$$\geq \delta \liminf_{b\to\infty} f_{bd}(1/d)/f(1/d) = \delta. \tag{75}$$

5 Examples.

5.1 Gaussian Processes with Stationary Increments.

Let $(Z_t,\ t \in \mathbf{R}_+)$ be a zero-mean Gaussian process with stationary increments and covariance function

$$\Gamma(s,t) = EZ_sZ_t, \tag{76}$$

and set

$$W_t := Z_t - \mu t. \tag{77}$$

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 [2, Chapter 3] and references therein. However, we are not excluding the use of non-Gaussian models in principle: some non-Gaussian processes with long-range dependence are presented in [21].

We set

$$\sigma_t^2 := \Gamma(t,t), \tag{78}$$

and make the following assumptions:

Hypothesis 4 *(i) For some $H \in (1/2, 1)$ and $\sigma \in \mathbf{R}_+$, $\lim_{t\to\infty} \sigma_t^2 t^{-2H} = \sigma^2$;*

(ii) Set $n_t = \sigma_t^2/(\sigma^2 t^{2H})$. Then for some $u < v = 2 - 2H$, the limit $\rho := \lim_{t\to\infty} t^{v-u}(n_t - 1)$ exists in $\mathbf{R}$;

(iii) For some function $k(t)$ increasing to $+\infty$ as $t \to 0$: $\ell(t)m_t$ and $\ell(t)\sigma_t$ remain bounded as $t \searrow 0$, where

$$m_t = E[\sup_{0<r<t} |Z_t|] \qquad \text{and} \qquad \ell(t) = k(t)v(t)/a(t). \tag{79}$$

As before, we take $W_t^L = \sum_L W_t$, the workload due to an L-fold superposition of independent processes Z_t served at fixed rate Ls. We now verify the various hypotheses of the previous two sections in order to apply the results therein.

Hypothesis 1:. With $a(t) = t$, $v(t) = t^{2-2H}$ then

$$\lambda_t(\theta) = \frac{1}{2} n_t \sigma^2 \theta^2 - s\theta \qquad \text{and} \qquad \lambda(\theta) = \frac{1}{2}\sigma^2\theta^2 - s\theta, \tag{80}$$

by Hypothesis 4(i). Then Hypothesis 1(i,ii,iii,iv) follows easily.

To get Hypothesis 1(v), it suffices, since $t \mapsto v(t)/a(t)$ is decreasing and Z has stationary increments, to show that for all θ

$$\lim_{t\to 0} E[\exp(\theta \sup_{0<r<t} |Z_r|)] = 1. \tag{81}$$

Note that by Borell's inequality (see [1], Theorem 5.2),

$$P[\sup_{0<r<t} |Z_r| \geq x] \leq 1 - \Phi\left((x - 2m_t)/\bar{\sigma}_t\right) \tag{82}$$

for any $x \geq 2m_t$, where $\bar{\sigma}_t = \sup_{0<r<t} \sigma_t$ and Φ is the canonical Gaussian distribution function. From this is follows that for any θ,

$$E[\exp(\theta\ell(t) \sup_{0<r<t} |Z_r|)] \leq e^{2\theta\ell(t)m_t} \left(1 + e^{(\theta\ell(t)\bar{\sigma}_t)^2/2}\left(1 - \Phi(-\theta\ell(t)\bar{\sigma}_t)\right)\right). \tag{83}$$

With Hypothesis 4(iii) we get that $\ell(t)m_t$, $\ell(t)\sigma_t$ and hence also $\ell(t)\bar{\sigma}_t$ remain bounded as $t \to 0$, so that
$E[\exp(\theta\ell(t)\sup_{0<r<t} |Z_r|)]$ is also bounded as $t \to 0$. Hence

$$E[\exp(\theta(v(t)/a(t)) \sup_{0<r<t} |Z_r|)] \leq E[\exp(\theta\ell(t) \sup_{0<r<t} |Z_r|)]^{1/k(t)} \to 1 \tag{84}$$

as $t \to 0$.

Hypotheses 2 and 3: Clearly λ_t and λ are closed.

$$\lambda^*(x) = \frac{1}{2\sigma^2}(x+s)^2 \qquad \text{and} \qquad \lambda_t^*(x) = \frac{1}{n_t}\lambda^*(x), \tag{85}$$

and $h^{-1}(t) = t^{1/(2-2H)}$. One verifies through Theorem 3 that

$$\delta = \delta_H := \frac{1}{2\sigma^2}\left(\frac{s}{H}\right)^{2H} (1-H)^{-2(1-H)}. \tag{86}$$

Using the convexity of $\lambda^* \circ h^{-1}$, and Hypothesis 4(ii), then

$$\begin{aligned}\nu_u &= -\lim_{t\to\infty} t^{v-u}(n_t^{-1}\lambda^* \circ h^{-1})^*(\delta) && (87)\\ &= -\lim_{t\to\infty} t^{v-u} n_t^{-1}(\lambda^* \circ h^{-1})^*(n_t\delta) && (88)\\ &= -\delta((\lambda^* \circ h^{-1})^*)'(\delta)\lim_{t\to\infty} t^{v-u}(n_t-1) && (89)\\ &= -\delta((\lambda^* \circ h^{-1})^*)'(\delta)\rho && (90)\end{aligned}$$

for some $u > 0$. By explicit calculation $((\lambda^* \circ h^{-1})^*)'(\delta) = (sv/(1-v))^v$, and so finally, if Hypothesis 3(i) is satisfied, then (64) holds with

$$\hat{\nu}_u = -\delta\rho\left(\frac{s2(1-H)}{2H-1}\right)^{2-2H-u}. \tag{91}$$

By (73), a sufficient condition for Hypothesis 3(i) is that

$$t \mapsto n_t \quad \text{is non-decreasing on some interval} \quad (0, t_0). \tag{92}$$

5.2 Fractional Brownian Motion.

A special case of the above is where

$$2\Gamma(s,t) = s^{2H} + t^{2H} - |s-t|^{2H}, \tag{93}$$

for some $0 < H < 1$. In this case the process Z is called *fractional Brownian motion.* The parameter H is called the *Hurst parameter.* When $H > 1/2$ the process exhibits long range dependence. This process has been proposed as a model for the workload by Leland *et al* [16], based on observations of Ethernet traffic data. For a queue processing a single stream of such arrivals, a large deviation lower bound on the queue-length distribution was obtained by Norros [19]. The corresponding upper bound was found by Duffield and O'Connell [9].

In accordance with Hypothesis 4(i) we take $H > 1/2$. $\sigma_t^2 = t^{2H}$ and in Lemma 2 of Appendix B we show that $m_t \sim t^H$ as $t \to 0$. Thus Hypothesis 4(iii) is satisfied with $k(t) = t^{H-1}$. From the point of view of economies of scale our first example is trivial: $\sigma_t^2 = t^{2H}$ and since $\sigma^2 = 1$ and $n_t = 1$ for all t we have $\nu = \rho = 0$. In fact,

$$\begin{aligned}I(b) &= \inf_{t>0} t^{2(1-H)}\lambda^*(b/t) && (94)\\ &= b^{2(1-H)}\inf_{t>0} t^{-2(1-H)}\lambda^*(t) && (95)\\ &= b^{2(1-H)}\delta_H \qquad \text{by Theorem 3.} && (96)\end{aligned}$$

5.3 FBM with Ornstein-Uhlenbeck Perturbation.

Let Z_t be a sum of fractional Brownian motion(with Hurst parameter $H \in (1/2,1)$) with an independent Ornstein-Uhlenbeck position process for which the corresponding velocity process is stationary with unit variance. Then (see (72) in [9]),

$$\sigma_t^2 = t^{2H} + 2(t + e^{-t} - 1). \tag{97}$$

The Ornstein-Uhlenbeck process has correlations which decay exponentially fast. It can be viewed as a short-range perturbation to the fractional Brownian motion. We shall see that the Z_t still has the power-law large deviations of its fBM component, but that the Ornstein-Uhlenbeck perturbation gives rise to modification which decrease the economies of scale since $\nu_u < 0$.

We check the condition of Hypothesis 4. (i) is satisfied since $\lim_{t\to\infty} \sigma_t^2 t^{-2H} = 1$. Observe $n_t \sim 1 + 2t^{1-2H}$ for t large, thus (ii) is satisfied with $u = 4H - 3$ and $\rho = 2$. In section 4.3 of [4] it is shown that for the Ornstein-Uhlenbeck arrivals *alone* one has $E[\sup_{0<r<t} |Z_r|] \sim t^{3/2}$ for small t: so for the combined arrivals we get (using Lemma 2 in Appendix B) $m_t \sim t^H + t^{3/2}$ for small t, while from (97) $\sigma_t^2 \sim t^{2H} + t^2$ for t small. Since $v(t)/a(t) = t^{1-2H}$ we find that Hypothesis 4 is satisfied with $k(t) = t^{\bar{k}}$ with $\bar{k} = \max\{\ H - 1\ ,\ 2H - 5/2\ \} < 0$. Furthermore, $n_t \sim 1 + t^{2-2H}$ for t small, so by (92), Hypothesis 3(i) is satisfied.

To summarize from (64) and (91):

$$\lim_{b\to\infty} b^{4H-3} \left(I(b) - \delta_H b^{2(1-H)} \right) = -2\delta_H \left(\frac{2s(1-H)}{2H-1} \right)^{2H-1}. \tag{98}$$

From this one sees, as might be expected, that the corrections to the first order approximation $I(b) \approx \delta_H b^{2(1-H)}$ become less pronounced as H increases, i.e. as the relative difference in the time scales of the fBM and Ornstein-Uhlenbeck components increases. This is seen strikingly if one tries to apply Theorem 2. It holds with

$$\nu = \begin{cases} -2\delta_{3/4}\sqrt{s} & \text{if} \quad H = 3/4 \\ 0 & \text{if} \quad H \in (3/4, 1) \end{cases} \tag{99}$$

while ν is not finite if $H \in (1/2, 3/4)$.

Finally we note from (85)

$$I(0) = \inf_{t>0}\ t^v/n_t = \inf_{t>0}\ \frac{t^2}{t^{2H} + 2(t + e^{-t} - 1)} = 0. \tag{100}$$

Generally, since $n_t \geq 1$, $\lambda_t \leq \lambda$ and thus $I(b) \leq b^{2(1-H)}\delta_H$.

A Appendix: Convergence of Legendre-Fenchel Transforms.

Lemma 1 *Let $(f_n)_{n\in\mathbf{N}}$ and f be convex functions on $\mathbf{R}$. If $f = \lim_{n\to\infty} f_n$ pointwise on* $\operatorname{int}\operatorname{dom} f$, *then $f^* = \lim_{n\to\infty} f_n^*$ pointwise on* $\operatorname{int}\operatorname{dom} f^*$.

Proof: By Theorem 24.1 of [20] the left and right derivatives f'_- and f'_+ of f are defined throughout $\operatorname{int}\operatorname{dom} f$ (and similarly for each f_n). Furthermore, these derivatives are non-decreasing functions, and for all $x_1 < x < x_2$ in $\operatorname{int}\operatorname{dom} f$

$$f'_+(x_1) \leq f'_-(x) \leq f'_+(x) \leq f'_-(x_2). \tag{101}$$

Let $t \in \operatorname{int}\operatorname{dom} f^*$, and define

$$x^+ = \sup\{x \mid f'_+(x) = t\} < \infty \quad \text{and} \quad x^- = \inf\{x \mid f'_-(x) = t\} > -\infty, \tag{102}$$

in other words, $[x^-, x^+]$ is the subdifferential $\partial f^*(t)$ of f^* at t. (Note that $x^\pm \in \operatorname{int}\operatorname{dom} f$ since $\operatorname{ran}\partial f^* = \operatorname{dom}\partial f$ and in the 1-dimensional case $\operatorname{int}\operatorname{dom} f = \operatorname{int}\operatorname{dom}\partial f$). Suppose we can find sequences $x_n^+ \searrow x^+$ and $x_n^- \nearrow x^-$ such that $f'_{n,-}(x_n^-) \leq t \leq f'_{n,+}(x_n^+)$ for all n sufficiently large. Then by (101) there exists a sequence (x_n) in $\operatorname{int}\operatorname{dom} f$ with $x_n^- \leq x_n \leq x_n^+$ such that $t \in [f'_{n,-}(x_n), f'_{n,+}(x_n)]$ (and hence since we work in $\mathbf{R}$, $t \in \operatorname{dom} f_n^*$) for all n sufficiently large, and x_n has limit points in $[x^-, x^+]$. Specialize to a subsequence converging to such a limit point x. Then for sufficiently large n

$$f_n^*(t) - f^*(t) = f(x) - f_n(x_n), \tag{103}$$

which goes to 0 as $n \to \infty$ by Theorem 10.8 of [20].

We establish the existence of the sequences $(x_n^\pm)$ with the desired properties. By Theorem 24.1 of [20] and (102), for any $x > x^+$ in $\operatorname{int}\operatorname{dom} f$ then

$$\forall \varepsilon > 0 \exists \delta : \quad f'_-(y) > f'_+(x) - \varepsilon \quad \forall y \in (x, x+\delta). \tag{104}$$

Now let $y_- < y < y_+$ in $\operatorname{int}\operatorname{dom} f$. By hypothesis, y_-, y, y_+ are also in $\operatorname{int}\operatorname{dom} f_n$ for n sufficiently large. Thus for such n, we have by convexity

$$\frac{f_n(y_-) - f_n(y)}{y_- - y} \leq f'_{n,-}(y) \leq f'_{n,+}(y) \leq \frac{f_n(y_+) - f_n(y)}{y_+ - y}. \tag{105}$$

Taking $n \to \infty$ then $y_- \nearrow y$ and $y_+ \searrow y$ we find

$$f'_-(y) \leq \liminf_{n\to\infty} f'_{n,-}(y) \leq \limsup_{n\to\infty} f'_{n,+}(y) \leq f'_+(y). \tag{106}$$

Thus

$$\forall y \in \operatorname{int}\operatorname{dom} f,\ \forall \varepsilon' > 0 \quad \exists n_0 : \quad f'_{n,-}(y) > f'_-(y) - \varepsilon' \quad \forall n > n_0. \tag{107}$$

Thus letting $x \searrow x^+$ and for each x setting $\varepsilon = f'_+(x) - f'_+(x^+) > 0$, $y = x + \min[\delta, x - x^+]/2$ and $\varepsilon' = f'_-(y) - f'_+(x^+) > 0$ we can construct a sequence $x_n^+ \searrow x^+$ such that $t \leq f'_{n,-}(x_n^+)$. Similarly we can construct a sequence $x_n^- \nearrow x^-$ such that $t \geq f'_{n,+}(x_n^-)$. The desired properties of $x_n^\pm$ then follow from (101). ■

B Appendix: Estimates for fractional Brownian motion.

Lemma 2 *Let Z_t be fBM with Hurst parameter $H > 1/2$. Then $E[\sup_{0<r<t} |Z_r|] \sim t^H$ as $t \to 0$.*

Proof: We use the stochastic integral representation of fractional Brownian motion (see, for example, [18]): if B_1 and B_2 are two independent, one-dimensional Brownian motions started at zero, then

$$Z_t := \int_0^\infty \left[(t+s)^{H-\frac{1}{2}} - s^{H-\frac{1}{2}}\right] dB_1(s) + \int_0^t (t-s)^{H-\frac{1}{2}} dB_2(s), \tag{108}$$

where B_1 and B_2 are two independent, one-dimensional Brownian motions started at zero.

By Itô's formula we have

$$\begin{aligned} \frac{Z_r}{H-1/2} &= \int_0^\infty \left[s^{H-\frac{3}{2}} - (r+s)^{H-\frac{3}{2}}\right] B_1(s)ds + \int_0^r (r-s)^{H-\frac{3}{2}} B_2(s)ds && (109) \\ &\leq \int_0^\infty \left[s^{H-\frac{3}{2}} - (r+s)^{H-\frac{3}{2}}\right] B_1^+(s)ds + \int_0^t s^{H-\frac{3}{2}} \sup_{s<r<t} B_2(r-s)ds, && (110) \end{aligned}$$

for $r < t$, and so

$$\begin{aligned} \frac{E[\sup_{0<r<t} Z_t]}{H-1/2} &\leq \int_0^\infty \left[s^{H-\frac{3}{2}} - (t+s)^{H-\frac{3}{2}}\right] E[B_1^+(s)]ds + \int_0^t s^{H-\frac{3}{2}} E[\sup_{s<r<t} B_2(r-s)]ds && (111) \\ &\leq \int_0^\infty \left[s^{H-\frac{3}{2}} - (1+s)^{H-\frac{3}{2}}\right] s^{1/2}ds + \int_0^t s^{H-\frac{3}{2}}(r-s)^{\frac{1}{2}}ds. && (112) \end{aligned}$$

Both these integrals are $\mathbf{O}(t^H)$ as $t \searrow 0$. By symmetry, $E[\sup_{0<r<t}|Z_r|] \leq 2E[\sup_{0<r<t} Z_r]$ and so we are done. ■.

Acknowledgements. The author thanks Wayne Sullivan for discussions on Lemma 1, and Neil O'Connell for supplying some arguments used in Lemma 2.

References

[1] C. Borell (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30:205–216.

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

[3] E. Buffet and N.G. Duffield (1994) Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *J. Appl. Prob.* 31:1049–1061.

[4] D.D. Botvich and N.G. Duffield (1994). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Submitted to *Queueing Systems.*

[5] C.S. Chang (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control.* 39:913-931.

[6] G.L. Choudhury, D.M. Lucantoni and W. Whitt (1993). Squeezing the most out of ATM. *IEEE Transactions on Communications,* to appear.

[7] C. Courcoubetis and R. Weber (1994). Buffer overflow asymptotics for a switch handling many traffic sources. Preprint.

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

[9] N.G. Duffield and Neil O'Connell (1993). Large deviations and overflow probabilities for the general single-server queue, with applications. *Proc. Cam. Phil. Soc.*, to appear.

[10] R.S. Ellis (1985). *Entropy, Large Deviations, and Statistical Mechanics*, Springer, New York.

[11] R.J. Gibbens and P.J. Hunt (1991). Effective Bandwidths for the multi-type UAS channel *Queueing Systems*, 9:17–28

[12] P.W. Glynn and W. Whitt (1993). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31A:131–159

[13] J.Y. Hui (1988). Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.* 6:1598–1608

[14] F.P. Kelly (1991). Effective bandwidths at multi-type queues. *Queueing Systems* 9:5–16

[15] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Networking* 1:424-428.

[16] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson (1993). On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communications Review* 23:183-193.

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

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

[19] Ilkka Norros (1994). A storage model with self-similar input. *Queueing Systems,* 16:387–396

[20] R. Tyrrell Rockafellar (1970) *Convex Analysis.* Princeton University Press, Princeton.

[21] Gennady Samorodnitsky and Murad S. Taqqu (1993). Linear models with long-range dependence and with finite or infinite variance. In: New directions in time–series analysis, part II. *IMA Vol. Math. Appl.,* 46:325-340. Springer, New York.

[22] A. Simonian and J. Guibert (1994). Large deviations approximation for fluid queues fed by a large number of on-off sources. *Proceedings of ITC 14, Antibes, 1994* 1013–1022.

[23] Alan Weiss (1986). A new technique for analysing large traffic systems. *J. Appl. Prob.* 18:506–532

[24] W. Whitt (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunications Systems.* 2:71-107