ON THE MAXIMUM WORKLOAD OF A QUEUE FED BY FRACTIONAL BROWNIAN MOTION

By Assaf J. Zeevi[1] and Peter W. Glynn[2]

Stanford University

Consider a queue with a stochastic fluid input process modeled as fractional Brownian motion (fBM). When the queue is stable, we prove that the maximum of the workload process observed over an interval of length t grows like $\gamma(\log t)^{1/(2-2H)}$ where $H > 1/2$ is the self similarity index (aka the Hurst parameter) characterizing the fBM, and can be explicitly computed. Consequently, one has also that the typical time required to reach a level b grows like $\exp\{b^{2(1-H)}\}$. We also discuss the implication of these results for statistical estimation of the tail probabilities associated with the steady-state workload distribution.

1. Introduction. Triggered by measurements and statistical analysis of traffic in high-speed networks, recent research has focused on stochastic models of network traffic having the properties of *long-range dependence*, and *self-*

---

[1]Research partially supported by JSEP Grant #DAAH04-94-G-0058 and National Science Foundation Grant #NCR-9628193

[2]Research partially supported by Army Research Office Grant #DAAG55-97-1-0377-P0001 and National Science Foundation Grant #DMS-9704732-001

AMS 1991 *subject classifications.* Primary 60K25; secondary 60G70

Key words and phrases. long-range dependence, queues, fractional Brownian motion, extreme values

similarity; see for example Leland, Taqqu, Willinger and Willson (1993), Beran, Sherman, Taqqu and Willinger (1995), and Erramilli, Narayan and Willinger (1997) for a discussion of the statistical evidence favoring such models. Perhaps the most theoretically important traffic model that exhibits these properties is fractional Brownian motion (fBM). In fact, just as Brownian motion is supported as a model of short-range dependency on the basis of Donsker's theorem and its generalizations, fBM can be viewed as a natural limiting approximation to a broad class of more physically plausible models, that describe how traffic in a network is generated from its individual sources; see Heath, Resnick and Samorodnitsky (1997), Heath, Resnick and Samorodnitsky (1998), Konstantopoulos and Lin (1996), and Whitt (1998). Because of both theoretical and statistical evidence supporting fBM as a possible traffic model, there is significant interest in trying to reach an understanding of the implications of such long-range dependent traffic for the performance of queues (because queueing effects occur at the buffers to the switches in a network).

Because fBM is highly non-Markovian (i.e., there is no finite-dimensional supplementary variable representation of fBM that makes it Markov), it is a challenging process to analyze as an input to a queue. At this point in time, we are aware of only two sets of results that describe the performance of a queue fed by fBM. The first such result is an asymptotic, due to Norros (1994) and Duffield and O'Connell (1995), for the tail probabilities of the steady-state workload in such a queue. For a description of the result, see Section 2 of this paper. These results have been refined recently by Massoulie and Simonian (1999), and Hüsler and Piterbarg (1999) to give the exact tail asymptotic [see also Narayan (1998)]. The second set of results, due to Krishnan (1996), takes advantage of the self-similarity of fBM to obtain a family of parameter scaling relationships that hold for the steady-state distribution of a queue fed by fBM.

Our objective in this paper is to establish several additional results that serve to enhance our understanding of queues fed by fBM. Our main focus here is on studying the maximum of the workload process over an interval of length t. The analysis of such maximum r.v.'s has a long history within the queueing literature; see, for example, Cohen (1968), Iglehart (1972), and the recent survey by Asmussen (1998). The principal results in this paper are:

1. The derivation of the asymptotic behavior of the maximum of the workload process over an interval of length t as $t \to \infty$; see Theorem 1 and Proposition 2.

2. The development of an asymptotic approximation for the time required by the workload process to first hit level b when $b \to \infty$; see Theorem 2.

3. Some remarks on estimating buffer loss probabilities from observed traffic and buffer dynamics, when the input is fBM; see Proposition 3 (the convergence rate of the associated estimators is very slow, however).

In Section 2 of this paper, we describe the precise model being considered here and discuss the main results. Section 3 contains the proofs.

2. Main Results. A real-valued process $B_H = (B_H(t) : t \geq 0)$ is said to be a fBM with *self-similarity* index $H \in [1/2, 1)$ if $B_H(0) = 0$, B_H has continuous sample paths, and B_H is a zero-mean stationary increments Gaussian process with

$$\mathrm{Cov}(B_H(t), B_H(s)) = \frac{1}{2}\left\{t^{2H} + s^{2H} - |t-s|^{2H}\right\}$$

for $s,t \geq 0$. The case $H = 1/2$ corresponds to standard Brownian motion. When $H > 1/2$, the correlation between two unit-length intervals separated by time t decays as t^{2H-2}, so that the autocorrelation sequence for $(B_H(k) - B_H(k-1) : k \geq 1)$ is non-summable. Thus, B_H describes a long-range dependent process. Additional properties and constructions of fBM are described in Samorodnitsky and Taqqu (1994, §7.2), and Mandelbrot and Van Ness (1968).

Let $\Gamma(t)$ be the cumulative amount of work input to the system over $[0,t]$. We shall assume that $\Gamma(t) = \lambda t + \sigma B_H(t)$ for $t \geq 0$ and $\sigma, \lambda > 0$, so that the input to the system is fBM. Given that $\Gamma = (\Gamma(t) : t \geq 0)$ has continuous sample paths, we view Γ as a fluid inflow to the queue. If the service mechanism deterministically serves work at rate $\mu > 0$, then the workload present in the system at time t is given by

$$W(t) = \Gamma(t) - \mu t - \min_{0 \leq s \leq t} [\Gamma(s) - \mu s] \quad ; \tag{2.1}$$

see Harrison (1985) for additional details on this representation of the workload. Put $X(t) = \Gamma(t) - \mu t$. Because $X = (X(t) : t \geq 0)$ evolves freely of any boundary behavior, we call X the "free process". On the other hand, $W = (W(t) : t \geq 0)$ is non-negative. The mapping that carries the free process X into the non-negative process W is called the "regulator mapping". We will therefore call the workload process W that appears here a *regulated* fBM process. We prefer the term "regulator mapping" over "reflection mapping" to differentiate this map from the Skorohod mapping that appears in the theory of reflected diffusions; see Lions and Sznitman (1984).

Let $\rho := \lambda/\mu$ and suppose that the "traffic intensity" $\rho < 1$. Then,

$$W(t) \Rightarrow W(\infty) \tag{2.2}$$

as $t \to \infty$, where $W(\infty) = \max\{X(s) : s \geq 0\}$. The tail asymptotics of the steady-state workload are known.

PROPOSITION 1. *Suppose $\rho < 1$. Then:*

i.) $\mathbb{P}(W(\infty) > b) \geq \mathbb{P}(X(t^*) > b)$, *where* $t^* := Hb/((1-H)\mu(1-\rho))$;

ii.) $b^{2H-2} \log \mathbb{P}(W(\infty) > b) \to -\theta^*$ *as* $b \to \infty$, *where*

$$\theta^* := \frac{(\mu(1-\rho))^{2H}}{2\sigma^2} \frac{1}{H^{2H}(1-H)^{2(1-H)}}$$

The lower bound i.) is due to Norros (1994), while the asymptotic ii.) can be found in Duffield and O'Connell (1995). Related results appear in Chang, Yao and Zajic (1996), and O'Connell and Procissi (1999). Because $W(\infty)$ is expressed easily in terms of X, these types of tail asymptotics can be attacked directly in terms of the free process alone. It is also known (see Konstantopoulos, Zazanis, and De Veciana (1996)) that if $\rho < 1$, one can construct a probability space supporting both the process X and a stationary process $W^* = (W^*(t) : t \geq 0)$ such that:

i.) $W^*(t) \stackrel{\mathcal{D}}{=} W(\infty)$ for $t \geq 0$, where '$\stackrel{\mathcal{D}}{=}$' denotes equality in distribution;

ii.) $W^*(t) = \Gamma(t) - \mu t + \{W^*(0) \vee L^*(t)\}$ for $t \geq 0$

where $L^*(t) = -\min\{\Gamma(s) - \mu s : s \in [0,t]\}$ is the non-decreasing process "regulating" the fBM. Thus, W^* is a stationary version of the workload process for our system, in which the input process is fBM.

Our main focus in this paper is to study the two maximum r.v.'s

$$M(t) = \max_{0 \leq s \leq t} W(s) \quad ,$$

$$M^*(t) = \max_{0 \leq s \leq t} W^*(s) \quad .$$

It is natural to expect that these two r.v.'s behave in an asymptotically identical fashion for large t. Furthermore, it is known, in substantial generality, that the asymptotic behavior of $M^*(t)$ is closely related to the tail behavior of $W^*(t)$. In particular, under suitable mixing conditions on W^*, $M^*(t)$ should behave asymptotically like the maximum of $\lfloor at \rfloor$ i.i.d. copies of $W(\infty)$, for some $a \in (0,1)$; see Leadbetter, Lindgren and Rootzén (1983) for such results. In principle, this then yields the asymptotic behavior of $M^*(t)$ (because the maximum behavior for i.i.d. sequences is well known).

The difficulty with this approach is the verification of the requisite mixing properties in our present setting. Such a methodology is particularly effective when W^* is regenerative; see Asmussen (1998) for many examples of queueing-related maximum processes that can be readily studied by taking advantage of the regenerative cycle structure of W^*. Such regenerations are easily identified for many common used short-range dependent input processes (e.g., Markov-modulated arrivals). However, because of the non-Markov nature of fBM, it is unclear that any regenerative structure is present in regulated fBM. In addition, the more general mixing conditions permitting one to view $M^*(t)$ as the maximum of i.i.d. r.v's seem difficult to verify directly, given that W^* is non-Markov and has long-range dependent input. However, our main result (based on a different style of argument) proves that the asymptotics suggested above are indeed correct in a suitable asymptotic sense.

THEOREM 1. *If $\rho < 1$, then*

$$\frac{M^*(t)}{(\log t)^{\frac{1}{2(1-H)}}} \Rightarrow \left(\frac{1}{\theta^*}\right)^{1/(2-2H)} \qquad \text{as } t \to \infty$$

$$\frac{M(t)}{(\log t)^{\frac{1}{2(1-H)}}} \Rightarrow \left(\frac{1}{\theta^*}\right)^{1/(2-2H)} \qquad as\ t \to \infty$$

In fact, the above convergence is actually in L_p, for all $p \in [1, \infty)$.

The cases of "heavy-traffic" and unstable queues are the subject of

PROPOSITION 2.

i.) If $\rho > 1$, then

$$t^{-H}\left(M(t) - \mu(\rho - 1)t\right) \Rightarrow \sigma \mathcal{N}(0,1) \qquad as \quad t \to \infty$$

ii.) If $\rho = 1$, then

$$t^{-H} M(t) \Rightarrow \sigma\xi \qquad as \quad t \to \infty$$

where $\xi := \max_{0 \le r \le 1} \max_{0 \le v \le r}[B_H(r) - B_H(v)]$.

Let $T(b) := \inf\{t \ge 0 : W(t) \ge b\}$ and note that $\{T(b) \le t\} = \{M(t) \ge b\}$. Because of this relationship between $T(b)$ and $M(t)$, Theorem 1 and Proposition 2 together yield asymptotic approximations for $T(b)$ that are valid for large b.

THEOREM 2.

i.) If $\rho < 1$, then

$$\frac{\log T(b)}{b^{2(1-H)}} \Rightarrow \theta^* \qquad as\ b \to \infty$$

ii.) If $\rho = 1$, then

$$\frac{T(b)}{b^{1/H}} \Rightarrow (\sigma\xi)^{-1/H} \qquad as\ b \to \infty$$

iii.) If $\rho > 1$, then

$$b^{-H}\left(T(b) - \frac{b}{\mu(\rho-1)}\right) \Rightarrow \left(\frac{1}{\mu(\rho-1)}\right)^{1+H} \sigma \cdot \mathcal{N}(0,1) \qquad as\ b \to \infty$$

If W were regenerative, i.) could be obtained by appealing, for example, to the regenerative approach described in Glasserman and Kou (1995). However, as discussed earlier, it is unclear how to implement this idea in the current fBM setting.

We conclude this section with some discussion of the implications of the results for estimation of loss probabilities in finite buffer queues based on real-time measurement of traffic. Such loss-probability estimators could potentially be useful in admission control for high-speed networks. To connect the infinite capacity model considered so far in this paper to a finite buffered system, we shall view the exceedence probability, $\mathbb{P}(W(\infty) > b)$ as a surrogate for the loss probability in a buffer of size b fed by fBM. Recall that if b is large, Proposition 1 asserts that $\mathbb{P}(W(\infty) > b)$ is essentially determined by θ^* and H. Thus, $\mathbb{P}(W(\infty) > b)$ can be roughly (i.e., in a logarithmic scale) estimated once estimators for θ^* and H are determined. In the short-range dependent context, several authors have proposed estimating θ^* from observed traffic using the maximum workload r.v. $M(t)$; see Berger and Whitt (1995) and Hsu and Walrand (1996). Theorem 1 states that if H is known, then θ^* can also be successfully estimated from long-range dependent traffic using $M(t)$.

Of course, in general, H itself would also need to be estimated from incoming traffic. Assume that the process Γ is observable, so that we can form the r.v. $H(t)$ from the traffic observed over $[0, t]$, where

$$H(t) := \left(\log \max_{0 \le s \le t} \left[X(s) - s\frac{X(t)}{t} \right] \right) / \log t \quad .$$

PROPOSITION 3.

i.) For $0 \le t \le 1$, *let* $B_H^0(t) = B_H(t) - tB_H(1)$ *be the fractional Brownian* *bridge process. Then,*

$$(\log t)\,(H(t) - H) \Rightarrow \zeta$$

as $t \to \infty$, *where* $\zeta := \log \max\{\sigma B_H^0(s) : 0 \le s \le 1\}$.

ii.) If $\rho < 1$, *then*

$$\frac{(M(t))^{2(1-H(t))}}{\log t} \Rightarrow \frac{1}{\theta^*}$$

as $t \to \infty$.

Proposition 3 asserts that the parameters H and θ^* can be consistently estimated, using the maximum workload r.v. $M(t)$, when the input process is fractional Brownian motion.

REMARK 1. Another natural estimator to consider in this context is the so-called moving average estimator, constructed as follows. Fix $a(t)$ and $m(t) := \lfloor t/a(t) \rfloor$. Chop up the the observation window $[0,t]$ into $m(t)$ sub-windows of length $a(t)$ each, and a remainder which is a fraction of $a(t)$ in length. Let $M_i(t) := \sup\{W(s) : s \in [(i-1)a(t), ia(t)]\}$ for $i = 1, 2, \ldots, m(t)$, and let $H(t)$ be as in Proposition 3. Then, the moving average estimator is defined to be

$$\gamma(t) := \frac{1}{m(t)} \sum_{i=1}^{m(t)} \frac{(M_i(t))^{2(1-H(t))}}{(\log a(t))} \quad .$$

Using the results of Theorem 1 and Proposition 3, one can establish that $\gamma(t) \Rightarrow (\theta^*)^{-1}$.

3. Proofs. For brevity, set

$$\beta := \frac{1}{2(1-H)} \quad .$$

Otherwise, notation and definition follow Section 2.

3.1. *Proof of Theorem reftheo:main1.* The proof of the theorem involves establishing the usual 'upper' and 'lower' bounds, i.e., our goal is to prove that for arbitrary positive constant $\delta > 0$ we have

$$\mathbb{P}\left(\frac{M^*(t)}{(\log t)^\beta} \geq \left(\frac{1-\delta}{\theta^*}\right)^\beta\right) \stackrel{t\to\infty}{\longrightarrow} 1 \quad , \tag{3.3}$$

$$\mathbb{P}\left(\frac{M^*(t)}{(\log t)^\beta} \geq \left(\frac{1+\delta}{\theta^*}\right)^\beta\right) \stackrel{t\to\infty}{\longrightarrow} 0 \quad , \tag{3.4}$$

from which the convergence in probability follows. We will then argue that essentially the same proof yields convergence for $M(t)$ as well. Finally, we prove the uniform integrability of the family $\left\{M^*(t)/(\log t)^\beta\right\}_{t\geq 2}$, which establishes the L^p convergence.

I. *Proof of the lower bound (3.3):* Again, we break up the proof into several steps.

1^0. Fix $\Delta \in (0,t)$, so that

$$\begin{aligned} W^*(t) &\geq X(t) - \inf_{0\leq s\leq t} X(s) \\ &\geq X(t) - X(t-\Delta) \quad . \end{aligned} \tag{3.5}$$

This gives

$$\begin{aligned} M^*(t) &= \max_{0\leq s\leq t} W^*(s) \\ &\geq \max_{k=1,2,\ldots,\lfloor t/\Delta\rfloor} W^*(k\Delta) \\ &\geq \max_{k=1,2,\ldots,\lfloor t/\Delta\rfloor} [X(k\Delta) - X((k-1)\Delta)] \\ &:= \max_{1\leq k\leq\lfloor t/\Delta\rfloor} Y_k^{(\Delta)} \end{aligned}$$

and note that

$$\begin{aligned} Y_k^{(\Delta)} &:= X(k\Delta) - X((k-1)\Delta) \\ &\stackrel{\mathcal{D}}{=} \sigma B_H(\Delta) - \mu(1-\rho)\Delta \\ &\stackrel{\mathcal{D}}{=} \sigma \Delta^H B_H(1) - \mu(1-\rho)\Delta \end{aligned}$$

using the properties of stationary increments and self-similarity of fBM. Set

$$Z_i = \frac{Y_i^{(\Delta)} + \mu(1-\rho)\Delta}{\Delta^H \sigma}$$

with $\sigma^2 = \mathrm{Var} B_1$. Then, since fBM is a Gaussian process, we have that the sequence $\{Z_i\}_{i=1}^{\lfloor t/\Delta \rfloor}$ is a stationary sequence of standardized Gaussian r.v.'s, the so-called *fractional Gaussian noise*. In addition,

$$\begin{aligned} \rho_Z(\ell) &:= \frac{\mathbb{E}[(Y_1^{(\Delta)} + \mu(1-\rho)\Delta)(Y_{1+\ell}^{(\Delta)} + \mu(1-\rho)\Delta)]}{\sigma^2 \Delta^{2H}} \\ &= \frac{1}{\sigma^2 \Delta^{2H}} \mathbb{E}[\sigma B_H(\Delta)][\sigma B_H((1+\ell)\Delta) - \sigma B_H(\ell\Delta)] \\ &= \frac{1}{2}\left[(\ell+1)^{2H} - (2\ell)^{2H} + (\ell-1)^{2H}\right] \end{aligned}$$

where the last step follows from the covariance structure of the fractional Gaussian noise sequence (cf. Proposition 7.2.9 in Samorodnitsky and Taqqu (1994)). Thus

$$\rho_Z(\ell) \sim H(2H-1)\ell^{2H-2}$$

as $\ell \to \infty$.

2^0. Note that $\{Z_i\}_{i=1}^{\lfloor t/\Delta \rfloor}$ is a stationary sequence of standard Gaussian r.v.'s, with $\rho_Z(\ell)\log \ell \to 0$. Thus, we can appeal to Theorem 4.3.3 in Leadbetter *et al.* (1983) which states that for a sequence of real numbers u_m and a standardized

Gaussian sequence, say $\{Z_i\}_{i=1}^m$, with the above properties, we have

$$\mathbb{P}\left(\bigvee_{i=1}^{m} Z_i \geq u_m\right) \to 1 + e^{-\tau}$$

if and only if

$$m\mathbb{P}(Z > u_m) \to \tau$$

as $m \to \infty$, with $\tau \in [0, \infty]$. To apply the theorem, we will let $\Delta = \Delta(t)$ depend on t, and choose $\Delta(t)$ carefully. Fix $\delta \in (0,1)$, and set

$$\alpha(t) := \left(\frac{1-\delta}{\theta^*}\log t\right)^{\beta}$$

$$u(t) := \frac{\alpha(t) + \mu(1-\rho)\Delta(t)}{\sigma\Delta^H(t)}$$

$$\tau(t) := m(t)\mathbb{P}(Z_1 > u(t))$$

with $m(t) := \lfloor t/\Delta(t) \rfloor$. Noting that

$$\mathbb{P}\left(\bigvee_{i=1}^{m(t)} Y_i^{(\Delta)} \geq \alpha(t)\right) = \mathbb{P}\left(\bigvee_{i=1}^{m(t)} Z_i \geq u(t)\right) \quad ,$$

we are left with the task of showing that $\tau(t) \to \infty$, which will establish the lower bound. The crucial step is to choose $\Delta(t)$ carefully, so that for $\alpha(t)$ as above, $\tau(t) \to \infty$. The only feasible choice turns out to be

$$\Delta(t) = \left(\frac{2(1-\epsilon)\sigma^2 H^2}{(\mu(1-\rho))^2}\log t\right)^{1/(2-2H)}$$

for any $\epsilon \in (0, \delta]$. Plugging this value into $u(t)$, straightforward algebra verifies that indeed $\tau(t) \to \infty$ as required. In verifying this, we use the standard bound on the Gaussian tail probability, namely, $\mathbb{P}(Z_1 > u(t)) \geq c(u(t))^{-1}\exp\{-u^2(t)/2\}$. Unfolding the arguments, we have that

$$\mathbb{P}\left(\frac{M^*(t)}{(\log t)^{\beta}} \geq \left(\frac{1-\delta}{\theta^*}\right)^{\beta}\right) \xrightarrow{t\to\infty} 1$$

as required.

II. *Proof of the upper bound (3.4).* The proof will be broken up into several steps.

1^0. Consider the following discretization of the continuous time problem. Recall, $M^*(t) = \sup\{W^*(s) : s \in [0,t]\}$. Set,

$$Y_i = \sup_{s\in[i-1,i)} W^*(s) \quad .$$

Then, trivially, $M^*(t) \le \bigvee_{i=1}^{\lceil t\rceil} Y_i$. Observe that by stationarity of W^* it follows that (Y_i) form a sequence of identically distributed random variables. Fix $\delta > 0$. Then, using the union bound, we have

$$\mathbb{P}\left(M^*(t) \ge \left(\frac{1+\delta}{\theta^*}\right)^\beta (\log t)^\beta\right) \le \lceil t\rceil \mathbb{P}\left(Y_1 > \left(\frac{1+\delta}{\theta^*}\right)^\beta (\log t)^\beta\right) \quad .$$

2^0. The goal here is to estimate the tail behavior of Y_1. Start by observing that

$$\begin{aligned} Y_1 &\le W^*(0) + \max_{0\le s\le 1}\left(X(s) - \min_{0\le\tau\le s} X(\tau)\right) \\ &\le W^*(0) + \max_{0\le s\le 1} X(s) - \min_{0\le s\le 1} X(s) \quad . \end{aligned} \tag{3.6}$$

Thus,

$$\begin{aligned} P_1 &:= \mathbb{P}\left(Y_1 > \left(\frac{1+\delta}{\theta^*}\right)^\beta (\log t)^\beta\right) \\ &\le \mathbb{P}\left(W^*(0) + \max_{0\le s\le 1} X(s) - \min_{0\le s\le 1} X(s) \ge \left(\frac{1+\delta}{\theta^*}\right)^\beta (\log t)^\beta\right) \\ &\le \underbrace{\mathbb{P}\left(W^*(0) \ge \left(\frac{1+\delta/2}{\theta^*}\right)^\beta (\log t)^\beta\right)}_{Q_1} \end{aligned}$$

$$+ \underbrace{\mathbb{P}\left(\max_{0\le s\le 1} X(s) \ge \frac{1}{2}\left(\frac{\delta/2}{\theta^*}\right)^{\beta} (\log t)^{\beta}\right)}_{Q_2}$$

$$+ \underbrace{\mathbb{P}\left(-\min_{0\le s\le 1} X(s) \ge \frac{1}{2}\left(\frac{\delta/2}{\theta^*}\right)^{\beta} (\log t)^{\beta}\right)}_{Q_3} \quad ,$$

where we used the fact that $\beta > 1$ implies $(1+\delta)^{\beta} > (1+\delta/2)^{\beta} + (\delta/2)^{\beta}$.

3^0. Our goal is to prove that $\lceil t \rceil Q_i \to 0$ for $i = 1, 2, 3$. First consider Q_2 and note

$$\max_{0\le s\le 1} X(s) = \max_{0\le s\le 1} (\sigma B_H(s) - \mu(1-\rho)s) \le \max_{0\le s\le 1} \sigma B_H(s) \quad .$$

To clinch the result, we need an estimate on the tail behavior of the maximum of standard fBM on a fixed interval. We appeal to Theorem 5.5 of Adler (1990), applied as in his Corollary 5.6, which when specialized to the case of fBM yields

$$\mathbb{P}\left(\max_{0\le s\le 1} B_H(s) > x\right) \sim \mathbb{P}(Z > x) \tag{3.7}$$

where $Z \sim \mathcal{N}(0,1)$. This implies that

$$\lceil t \rceil Q_2 \to 0$$

since for $\beta > 1$ standard estimates on the Gaussian tail give $\lceil t \rceil \mathbb{P}(Z > c(\log t)^{\beta}) \to 0$, and $Q_2 \sim \mathbb{P}(Z > c(\log t)^{\beta})$ using (3.7), where $c > 0$ is a generic constant premultiplying the logarithmic term. We next consider Q_3. The key is to note that

$$\begin{aligned} -\min_{0\le s\le 1} X(s) &\stackrel{\mathcal{D}}{=} \max_{0\le s\le 1} (\mu(1-\rho)s + \sigma B_H(s)) \\ &\le \mu(1-\rho) + \sigma \max_{0\le s\le 1} B_H(s) \end{aligned}$$

since $(B_H(s) : s \in [0,1]) \stackrel{\mathcal{D}}{=} (-B_H(s) : s \in [0,1])$. Therefore, upper bounding Q_3 by the sum of the probabilities involving the maximum and the minimum respectively, the previously established result for Q_2 can be applied (with different constants premultiplying $(\log t)^\beta$) to give that $\lceil t \rceil Q_3 \to 0$. It remains to show that $\lceil t \rceil Q_1 \to 0$ or equivalently $a(t) := \log\lceil t \rceil + \log Q_1 \to -\infty$. The result in Proposition 1 implies that

$$\lim_{t\to\infty} \frac{\log P\left\{W^*(0) > \left(\frac{1+\delta/2}{\theta^*}\right)^\beta (\log t)^\beta\right\}}{\log t} = -(1+\delta/2) \quad .$$

Consequently we have that

$$\begin{aligned} a(t) &= \log(t) + \log Q_1 \\ &= \log(t)\left(1 + \frac{\log P\left\{W^*(0) > \left(\frac{1+\delta/2}{\theta^*}\right)^\beta (\log t)^\beta\right\}}{\log t}\right) \\ &\to -\infty \end{aligned}$$

Thus, $\lceil t \rceil Q_1 \to 0$. Going back to Step 1^0 we see that $\lceil t \rceil \mathbb{P}(Y_1 > x(t)) \to 0$, thus

$$\mathbb{P}\left(\frac{M^*(t)}{(\log t)^\beta} \geq \left(\frac{1+\delta}{\theta^*}\right)^\beta\right) \to 0$$

establishing the upper bound.

III. *Proof sketch for $M(t)$.*

Note that for the process $M(t)$, $W(0) = 0$ as the free-process $X(t)$ starts at 0. The proof of the lower bound then holds, with equality replacing the first inequality in (3.5). The upper bound on the tail probability of Y_1 in (3.6) holds with $W(t)$ replacing $W^*(t)$, and the bounds on P_1 still holds as $W^*(0) \geq 0$ a.s. The rest of the arguments deal with estimates on the tails of the free-process and carry through without change.

IV. *Proof of L_p convergence.* Fix $p \in [1,\infty)$. It suffices to show that the sequence $(M^*(t)/(\log t)^\beta)^p$ is uniformly integrable. A sufficient condition for this is

$$\sup_{t\geq 2} \mathbb{E}\left[\frac{M^*(t)}{(\log t)^\beta}\right]^{p+1} < \infty \tag{3.8}$$

where the estimator can be arbitrarily defined as 0 for $t \leq 2$. To this extent, define

$$K_1 = \inf\left\{y > 0 : \text{such that } \frac{\log P(W^*(0) > x)}{x^{1/\beta}} \leq -\frac{\theta^*}{2}, \quad \forall x \geq y\right\}$$

and note that $K_1 < \infty$ follows from Proposition 1, i.e.,

$$\limsup_{x\to\infty} \frac{\log P\{W^*(0) > x\}}{x^{1/\beta}} \leq -\frac{\theta^*}{2} \quad .$$

Let

$$K_2 = \inf\left\{x \geq 0 : \mathbb{P}\left(\max_{0\leq s\leq 1} B_H(s) \geq y\right) \leq 2\mathbb{P}(Z \geq y), \quad \forall y > x\right\} \quad ,$$

where $Z \sim \mathcal{N}(0,1)$. The finiteness of K_2 follows from Step 3^0 of the proof of II. Then, setting $K = \max\{K_1, K_2, 4/\theta^*\}$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{M^*(t)}{(\log t)^\beta}\right]^{p+1} &= \int_0^\infty (p+1)y^p \mathbb{P}\left(M^*(t) > y(\log t)^\beta\right) dy \\
&= \int_0^K (p+1)y^p \mathbb{P}\left(M^*(t) > y(\log t)^\beta\right) dy \\
&\qquad + \int_K^\infty (p+1)y^p \mathbb{P}\left(M^*(t) > y(\log t)^\beta\right) dy \\
&\leq K^{p+1} + \underbrace{\int_K^\infty (p+1)y^p \lfloor t\rfloor \mathbb{P}\left(W^*(0) > y(\log t)^\beta/2\right) dy}_{I_t} \\
&\quad + \underbrace{\int_K^\infty (p+1)y^p \lfloor t\rfloor \mathbb{P}\left(\max_{0\leq s\leq 1} X(s) - \min_{0\leq s\leq 1} X(s) > y(\log t)^\beta/2\right) dy}_{R_t} \quad ,
\end{aligned} \tag{3.9}$$

where the inequality follows from Step 1^0 of II. Now,

$$
\begin{aligned}
I_t &= \int_K^\infty (p+1)y^p \lfloor t \rfloor \mathbb{P}\left(W^*(0) > y(\log t)^\beta/2\right) dy \\
&\leq \int_K^\infty (p+1)y^p \exp\left\{\frac{(\log\lfloor t\rfloor)^\beta}{2}\left(1 + y\frac{\log \mathbb{P}(W^*(0) > (y(\log t)^\beta)/2)}{y\log\lfloor t\rfloor)^\beta/2}\right)\right\} dy \\
&\overset{(a)}{\leq} \int_K^\infty (p+1)y^p \exp\left\{((\log\lfloor t\rfloor)^\beta/2)\left(1 - y\frac{\theta^*}{2}\right)\right\} dy \\
&\overset{(b)}{\leq} \int_K^\infty (p+1)y^p e^{-\frac{\theta^*}{8}y(\log\lfloor t\rfloor)^\beta} dy \\
&\leq \left(\frac{8}{\theta^*}\right)^{p+1} \frac{(p+1)!}{(\log\lfloor t\rfloor)^{(p+1)\beta}}
\end{aligned}
$$

where (a) and (b) follow from the definition of K. Using the bound on Q_3 in Step 4^0 of II, and by definition of K, it is clear that there exists a $C < \infty$ such that $R_t \leq C/(\log t)^{(p+1)\beta}$ (in fact, it is clear that $R_t = o(I_t)$). Combining (3.9), the upper bound on I_t above, and the bound on R_t we have proved (3.8). Thus, the sequence $(M^*(t)/(\log t)^\beta)^p$ is uniformly integrable. Putting parts I–IV together, the proof is complete. ■

3.2. *Proof of Theorem 2 and Propositions 2 and 3.* We start with

Proof of Proposition 2. For $\rho = 1$,

$$
\begin{aligned}
M(t) &\overset{\mathcal{D}}{=} \max_{0\leq s\leq t}\max_{0\leq u\leq s}[\sigma B_H(s) - \sigma B_H(u)] \\
&= \sigma \max_{0\leq r\leq 1}\max_{0\leq v\leq r}[B_H(rt) - B_H(vt)] \\
&\overset{\mathcal{D}}{=} \sigma t^H \max_{0\leq r\leq 1}\max_{0\leq v\leq r}[B_H(r) - B_H(v)] \\
&= \sigma t^H \xi \quad .
\end{aligned}
$$

For $\rho > 1$, consider the sequence of unit increments of fBM (sometimes referred to as fractional Gaussian noise), which is a stationary Gaussian sequence. From the properties of its spectral density (cf. Proposition 7.2.9 and Proposition 7.2.10 in Samorodnitsky and Taqqu (1994)), one easily sees that its spectral measure does not have point masses, which in turn, for stationary Gaussian processes, is a necessary and sufficient condition for ergodicity. Details of this argument can be found, for example, in Rozanov (1967, p. 163). Consequently, the pointwise ergodic theorem gives

$$\frac{B_H(t)}{t} \to 0 \quad \text{a.s.}$$

so $X(t) = \mu(\rho-1)t + \sigma B_H(t) \to +\infty$ a.s. Set $\Psi := \inf\{X(t) : t \geq 0\}$. Hence, there exists $T_0 < \infty$ s.t. for $t \geq T_0$,

$$W(t) = X(t) - \Psi \quad ,$$

so for $t \geq T_0$,

$$\max_{0 \leq s \leq t} X(s) - \Psi \leq \max_{0 \leq s \leq t} W(s) \leq \max_{0 \leq s \leq T_0} W(s) + \max_{0 \leq s \leq t} X(s) - \Psi \quad .$$

Since $\mu(\rho-1) > 0$,

$$\begin{aligned}
\max_{0 \leq s \leq t} X(s) - X(t) &= \max_{0 \leq s \leq t} [\sigma B_H(s) + \mu(\rho-1)s - \sigma B_H(t) - \mu(\rho-1)t] \\
&\stackrel{\mathcal{D}}{=} \max_{0 \leq s \leq t} [\sigma B_H(t) - \sigma B_H(s) - \mu(\rho-1)(t-s)] \\
&\stackrel{\mathcal{D}}{=} \max_{0 \leq s \leq t} [\sigma B_H(s) - \mu(\rho-1)s] \\
&\Rightarrow \tilde{W}(\infty) \quad ,
\end{aligned}$$

with $\tilde{W}(\infty) < \infty$ almost surely. So,

$$t^{-H}(M(t) - X(t)) \Rightarrow 0$$

which concludes the proof. ■

Proof of Theorem 2. Claim i.) follows from the relation $\{M^*(t) \geq b\} = \{T(b) \leq t\}$, and take a sequence

$$t_b := \exp\left\{ b^{1/\beta} \frac{\theta^*}{1+\delta} \right\} \tag{3.10}$$

so that $t_b \to \infty$ as $b \uparrow \infty$. Then, by Theorem 1

$$\mathbb{P}\left(M^*(t) > \left(\frac{1+\delta}{\theta^*} \log t \right)^{\beta} \right) \to 0 \tag{3.11}$$

and the convergence holds also along the subsequence t_b. In particular, substituting (3.10) into (3.11), we have

$$\mathbb{P}\left(\frac{T(b)}{\theta^* b^{1/\beta}} \leq 1 - \delta' \right) \to 0$$

with $\delta' := \delta/(1+\delta)$. The upper bound follows similarly.

ii.) Follows from

$$\begin{aligned} \mathbb{P}(b^{-1/H} T(b) \leq t) &= \mathbb{P}(M(tb^{1/H}) \geq b) \\ &= \mathbb{P}\left((tb^{1/H})^{-H} M(tb^{1/H}) \geq t^{-H} \right) \\ &\to \mathbb{P}(\sigma\xi \geq t^{-H}) \\ &= \mathbb{P}((\sigma\xi)^{-1/H} \leq t) \end{aligned}$$

iii.) The argument follows along the lines of Theorem 6 of Glynn and Whitt (1988). Fix $x \in \mathbb{R}$, then

$$\begin{aligned} &\mathbb{P}\left(b^{-H} \left(T(b) - \frac{b}{\mu(\rho-1)} \right) \leq x \right) = \\ &\quad = \mathbb{P}\left(T(b) \leq \frac{b}{\mu(\rho-1)} + x b^H \right) \\ &\quad = \mathbb{P}\left(M(t_b) \geq b \right) \quad ; \quad t_b := b/(\mu(\rho-1)) + x b^H \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(t_b^{-H}\left(M(t_b) - \mu(\rho-1)t_b\right) \geq (b - \mu(\rho-1)t_b)t_b^{-H}\right) \\
&\to \mathbb{P}\left(Z \geq -x(\mu(\rho-1))^{1+H}\right) \quad \text{as} \quad b \to \infty \\
&= \mathbb{P}\left((\mu(\rho-1))^{-(1+H)} Z \leq x\right)
\end{aligned}
$$

where $Z \sim \sigma\mathcal{N}(0,1)$ by Theorem 1. Since x is arbitrary, the convergence holds for all continuity points of the limiting (Gaussian) cdf, and the proof is complete. ■

Proof of Proposition 3. Claim i.) follows from

$$
\begin{aligned}
\tilde{M}(t) &:= \sup_{0 \leq s \leq t} \left(X(s) - s\frac{X(t)}{t} \right) \\
&= \sup_{0 \leq s \leq 1} \left(\sigma B_H(st) - st\frac{\sigma B_H(t)}{t} \right) \\
&\stackrel{\mathcal{D}}{=} \sigma t^H \sup_{0 \leq s \leq 1} \left(B_H(s) - sB_H(1) \right)
\end{aligned}
$$

taking log's from both sides and rearranging gives the result.

ii.) The trick is to reduce the problem to that studied in Theorem 1, i.e., replace $H(t)$ with H. Write

$$
\begin{aligned}
\frac{(M(t))^{2(1-H(t))}}{\log t} &= \frac{(M(t))^{2(1-H)}}{\log t} \frac{(M(t))^{2(1-H(t))}}{(M(t))^{2(1-H)}} \\
&= \underbrace{\left(\frac{M(t)}{(\log t)^{\frac{1}{2(1-H)}}} \right)^{2(1-H)}}_{I_t} \underbrace{(M(t))^{2(H-H(t))}}_{II_t}
\end{aligned}
$$

and observe that

$$
I_t \Rightarrow \frac{1}{\theta^*}
$$

by Theorem 1, and the continuous mapping theorem. As for II_t, observe that

$$\log II_t = 2(H - H(t)) \log M(t) \quad .$$

Now, by part i.) we have

$$|H - H(t)| = O_p\left(\frac{1}{\log t}\right)$$

and taking log's in Theorem 1 we have also

$$\frac{\log M(t)}{\log \log t} \Rightarrow \frac{1}{2(1-H)} \quad .$$

Thus, an application of the continuous mapping theorem gives

$$\log II_t \Rightarrow 0$$

which concludes the proof. ■

Acknowledgments. The authors wish to thank an anonymous referee for careful reading and helpful comments.

REFERENCES

Adler, R.J. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes.* Institute of Mathematical Statistics, Hayward, CA, 1990.

Asmussen, S. Extreme value theory for queues via cycle maxima. *Extremes*, 1(2):137–168, 1998.

Berger, A.W. and Whitt, W. Maximum values in queueing processes. *Prob. Engrg. and Info. Sci.*, 9:375–409, 1995.

Beran, J., Sherman, R., Taqqu, M., and Willinger, W. Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. on Commun.*, 43:1566–1579, 1995.

Chang, C.-S., Yao, D.D. and Zajic, T. Moderate deviations for queues with long-range dependent input. In *Stochastic Networks*, volume 117 of *Lecture Notes in Statist.*, pages 275–298. Springer-Verlag, New York, 1996.

COHEN, J.W. Extreme value distribution for the $M/G/1$ and the $G/M/1$ queueing systems. *Ann. Inst. H. Poincaré Sect. B*, 4:83–98, 1968.

DUFFIELD, N.G. AND O'CONNELL, N. Large deviations and overflow probabilities with general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.*, 118:363–374, 1995.

ERRAMILLI, A., NARAYAN, O., AND WILLINGER, W. Fractal queueing models. In *Frontiers in queueing*, Probab. Stochastics Ser., pages 245–269. CRC, Boca Raton, FL, 1997.

GLASSERMAN, P. AND KOU, S. Limits of first passage times to rare sets in regenerative processes. *Ann. Appl. Probab.*, 5:424–445, 1995.

GLYNN, P. AND WHITT, W. Ordinary CLT and WLLN versions of L= λW. *Math. Oper. Res.*, 13:674–692, 1988.

HARRISON, M.J. *Brownian Motion and Stochastic Flow Systems*. Wiley, NY, 1985.

HEATH, D., RESNICK, S., AND SAMORODNITSKY, G. Patterns of buffer overflow in a class of queues with long memory in the input stream. *Ann. Appl. Prob.*, 7:1021–1057, 1997.

HEATH, D., RESNICK, S., AND SAMORODNITSKY, G. Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.*, 23:145–165, 1998.

HSU, I. AND WALRAND, J. Dynamic bandwidth allocation for ATM switches. *J. Appl. Probab.*, 33:758–771, 1996.

HÜSLER, J AND PITERBARG, V. Extremes of a certain class of Gaussian processes. *Stochastic Process. Appl.*, 83:257–271, 1999.

IGLEHART, D.L. Extreme values in the GI/G/1 queue. *Ann. Math. Statist.*, 43:627–635, 1972.

KONSTANTOPOULOS, T. AND LIN, S.J. Fractional Brownian approximations of queueing networks. In *Stochastic Networks*, volume 117 of *Lecture Notes in Statist.*, pages 257–273. Springer-Verlag, New York, 1996.

KONSTANTOPOULOS, T., ZAZANIS, M., AND DE VECIANA, G. Conservation laws and reflection mappings with an application to multiclass mean value analysis for stochastic fluid queues. *Stochastic Processes and their Applications*, 65:139–146, 1996.

Krishnan, K.R. A new class of performance results for a fractional Brownian traffic model. *Queueing Systems*, 22:277–285, 1996.

Leadbetter, M.R., Lindgren, G., and Rootzén, H. *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York, 1983.

Leland, W.E., Taqqu, M.S., Willinger, W., and Willson, D.V. On the self-similar nature of Ethernet traffic. In *SIGCOMM93*, 1993.

Lions, P.-L. and Sznitman, A.-S. Stochastic differential equations with reflecting boundary conditions. *Comm. Pure Appl. Math.*, 37:511–537, 1984.

Mandelbrot, B.B. and Van Ness, J.W. Fractional Brownian motion, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

Massoulie, L. and Simonian, A. Large buffer asymptotics for the queue with FBM input *J. Appl. Probab.*, 36, to appear, 1999.

Narayan, O. Queue length distribution for fractional Brownian traffic. *Adv. Perf. Analysis*, 1:39–65, 1998.

Norros, I. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

O'Connell, N. and Procissi, G. A variational approach to performance evaluation for a queue loaded by fractional Brownian motion. HP Labs technical report HPL-BRIMS-98-18, 1998.

Rozanov, Y.A. *Stationary Random Processes.* Holden-Day, SF, 1967.

Samorodnitsky, G. and Taqqu, M.S. *Stable Non-Gaussian Random Processes.* Chapman & Hall, NY, 1994.

Whitt, W. Non-Brownian FCLTS for the single server queue. Technical report, AT&T Labs, 1998.

INFORMATION SYSTEMS LAB
STANFORD UNIVERSITY
STANFORD, CA, 94305-9510
assaf@isl.stanford.edu

DEPARTMENT OF EES & OR
STANFORD UNIVERSITY
STANFORD, CA, 94305
glynn@leland.stanford.edu