Economies of scale for long-range dependent traffic in short buffers[*]

N.G. Duffield
AT&T Laboratories,
Room 2C-323, 600 Mountain Avenue,
Murray Hill, NJ 07974, USA
E-mail: duffield@research.att.com

Consider the problem of predicting loss ratios for traffic streams sharing a buffer. Approximations based on the temporal statistical properties of single sources do not account for the economies of scale which can arise when there is statistical multiplexing gain across sources. These can occur whether the sources have long or short range dependence; in either case the economies may be positive or negative. In this paper we investigate this matter for sources described by fractional ARIMA processes, and show that their short-range structure can mean that a simple power-law tail based on the Hurst parameter alone can be over-optimistic when the buffer space allocated per source is not large.

Subject classification: AMS(MOS) 60K25, 60G18, 60F10

Keywords: Queueing theory, large deviations, fractional Brownian motion, fractional ARIMA processes, superpositions.

1 Introduction

Let an infinitely buffered queue be fed by a stationary arrival process. Denote by A_t the work arriving at the queue in the interval $[-t, 0)$, and suppose that the queue is served at rate s. The theory of large deviations can be used to relate the asymptotic properties of the arrival process to those of the distribution of the queue length Q. If the queue is stable (i.e. $\mathbf{E}A_1 < s$) and the following limit exists (possibly infinite) for all real θ:

$$\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta) \quad \text{with} \quad \lambda_t(\theta) = \log \mathbf{E}[e^{\theta A_t}] - s\theta, \tag{1}$$

then (modulo technical assumptions)

$$\lim_{b\to\infty} b^{-1} \log \mathbf{P}[Q > b] = -\delta, \tag{2}$$

where the decay rate δ is the positive root of $\lambda(\delta) = 0$. This result has been proved under various degrees of generality by several authors [4, 12, 15].

The existence of the limit λ is a statement about the shortness of the range of depen-

[*] To appear in Telecommunication Systems

dence of increments of the process A. For example, since for large t, $\lambda_t(\theta) \approx \lambda_{2t}(\theta)$, then $\mathbf{E}[e^{\theta A_{2t}}] \approx \mathbf{E}[e^{\theta(A_{2t}-A_t)}]\mathbf{E}[e^{\theta A_t}]$. Indeed, the limit λ as defined does not exist when A_t is the canonical example of a process with long-range dependence: fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$ (see [21] for definitions). In this case A_t is Gaussian with variance t^{2H} and so $t^{-1} \log \mathbf{E}[e^{\theta A_t}] = \theta^2 t^{2H-1} \to \infty$ as $t \to \infty$. However, the previous large deviation result can adapted through the introduction of a scaling function as follows. Suppose that the arrival process A_t is such that the limit

$$\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta) \quad \text{exists, with} \quad \lambda_t(\theta) = t^{2H-2} \log \mathbf{E}[\exp(\theta t^{2H-1} A_t)] - s\theta, \tag{3}$$

for some $H \in (1/2, 1)$, as is the case for the above fractional Brownian motion. Then the asymptotic behaviour of the queue length is that of a Weibull distribution:

$$\lim_{b\to\infty} b^{2H-2} \log \mathbf{P}[Q > b] = -\delta, \tag{4}$$

although generally δ is *not* given as the root of $\lambda(\delta) = 0$ when $H > 1/2$. A lower bound of this form was first obtained in [22]; subsequently the limit for these and more general long-range dependent sources was proved in [12]. Note that the limits (3) and (4) reduce to (1) and (2) when $H = 1/2$. In this paper we shall deal always deal with sources for which the limit (3) exists: if for $H = 1/2$ they will be called short-range dependent (SRD); if for $H \in (1/2, 1)$ then long-range dependent (LRD). Note that in both cases we assume that for each fixed t, the distribution of A_t has a sufficiently short tail for $\lambda_t(\theta)$ to be finite for some $\theta > 0$.

The asymptotics (2) and (4) can be used to approximate the queue length distributions:

$$\mathbf{P}[Q > b] \approx e^{-\delta b^p} \tag{5}$$

where $p = 1$ for SRD sources, while $p = 2 - 2H$ for LRD sources. For SRD sources this effective bandwidth approximation has been proposed as the basis for admission control in Asynchronous Transfer Mode (ATM) networks, based upon constraints on cell-loss ratios. (See [19, 26] and references therein).

However, numerical studies suggest that this approximation may not be sufficiently accurate for arrivals which are composed of a superposition of a large number of mixing sources, each with a high degree of autocorrelation. This is demonstrated in Figure 1 using log loss-curves obtained through simulation: these are seen to move away from a linear approximation based on (5) as the number of sources L is increased while the offered load and buffer allocation per source are held constant. In this case one should consider a joint scaling of the number of sources L, the service rate sL (so as to keep the load fixed) and queue level bL. This behaviour is accounted for by the theoretical result that

$$\lim_{L\to\infty} L^{-1} \log \mathbf{P}[Q > Lb] = -I(b), \tag{6}$$

where the *shape function* I is defined by a variational expression

$$I(b) = \inf_{t>0} I_t(b) \tag{7}$$

where the function I_t is determined from λ_t in a manner which we shall detail later. This result has been established in a variety of contexts: originally for Markov modulated

Poisson processes in [25], more recently for general short-range dependent (SRD) sources in [2], with the discrete-time case treated in [6], SRD alternating renewal processes in [24].

The limit (6) for LRD sources was established in [10]. The existence of this limit indicates that, although the temporal statistical properties and consequent tail distributions for large queue levels b of SRD and LRD sources are quite different, these tails scale in the same manner when the number of sources in the superposition becomes large while the buffer allocation per source is held constant. This can be seen in [10] where it is shown that the asymptotics of the shape function I for large b are determined to lowest order by those of the appropriate effective bandwidth approximation. Thus $I(b)/b^p \sim \delta$ as $b \to \infty$, with the appropriate p and δ as above. At this level of approximation, the shape function reflects the form of the large b asymptotic for single arrivals.

However, finer asymptotic corrections are also found: under very general conditions

$$I(b) - \delta b^p \sim \nu b^u, \qquad \text{as } b \to \infty \tag{8}$$

with p as before and $u < p$ [10], with $u = 0$ and $p = 1$ for SRD sources [2]. The difference $I(b) - \delta b^p$ determines the statistical multiplexing gain across the sources in the superposition, when compared with the effective bandwidth type formula. This gain may be positive or negative: if it is positive then we speak of the economies of scale which are to be obtained through the statistical multiplexing of many sources. For example, for SRD sources, (8) leads to the approximation

$$P[Q > b] \approx e^{-\delta b} e^{-\nu L} \tag{9}$$

when $b \gg L \gg 1$. The quantity ν can be shown to be positive for arrivals whose increments satisfy a positive autocorrelation property (see [2] for details).

The behaviour of $I(b)$ for small b is also of interest. For bursty SRD sources, the "two-slope" behaviour of the (typical) loss curves seen in Figure 1 find heuristic explanation in terms of the time-scales at which queueing occurs: the cell-level queue determined by the pattern of the arrival process at short time scales; the burst level queue by that at longer time-scale. However, is appears that little is known about such phenomena in queues of LRD traffic. In particular one can ask in what manner do short-term correlations within a LRD process manifest themselves in the small buffer behaviour of the queue-length distribution. The importance of finding the time-scale of the arrivals process which is most relevant for the buffer size in question has already been noted in [14] in the context of the problem of modelling arrival process with large-scale fluctuations. As is well-known, the large-deviation heuristic that "rare events occur in the most likely way" also tells us when (i.e. at which time-scales) they happen. In the shape-function scheme used in this paper, then for each queue level b the relevant time is that for which the infimum is achieved in (7). This observation has been used in [11] to examine the time-dependent behaviour of the queue length due to large superpositions of SRD sources, and recently in [23] to examine the extent to which it is possible to reproduce the loss curves of LRD arrivals by those of arrivals which are only SRD. Similar ideas have been used to explore the role of time-scales in modeling VBR video sources in [17].

In this paper, we investigate the effects of short range structure on loss ratios for queues with LRD arrivals. For Fractional Brownian motion with a given Hurst parameter $H \in (1/2, 1)$, the shape function is exactly that given by the effective bandwidth-type approximation: $I(b) = \delta b^{2-2H}$ for all b (see section 2.2). We model additional short-range

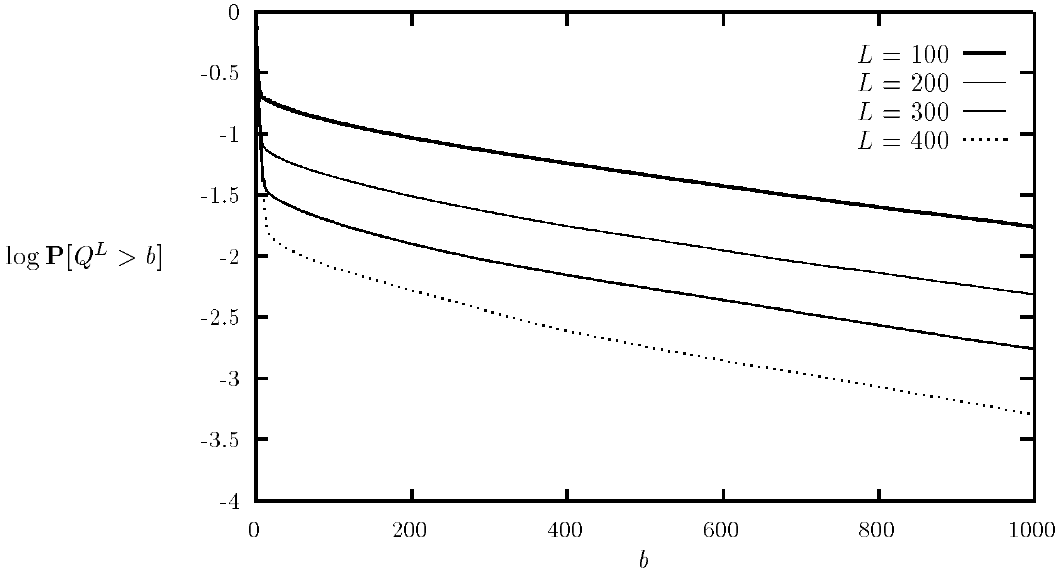

Figure 1: Simulated loss curves for increasing L.

structure by the use of fractional ARIMA processes. In section 2 we review the necessary analytical framework, unify the treatment of shape functions given previously for traffic whose dependence is short range [2] and long-range [10], and discuss the relevance for online estimation schemes based on the shape function. In section 3 we analyze fractional ARIMA processes by these means, and illustrate with specific examples. This establishes that when the buffer occupation per source b is small, the effective bandwidth type approximation $I(b) \approx \delta b^{2-2H}$ based on the Hurst parameter alone can be over-optimistic.

2 The asymptotics of large superpositions

We need to review in more detail the basis for the asymptotic results (6) and (8). For each positive integer L consider a queue with backward arrival process $(A_t^L)_{t \in \mathbf{Z}^+}$: i.e. A_t^L is the work arriving to be processed in the interval $[-t, 0)$. The work is served at rate sL for some fixed s. We have in mind the example that A_t^L is the arrival process due a superposition of L sources, although this is not necessary. Define the excess workload process $W_t^L = A_t^L - Lst$ and $W_0^L = 0$. Then the queue length at time zero is

$$Q^L = \sup_{t \geq 0} W_t^L. \tag{10}$$

2.1 The large L asymptotic.

To access the large deviation properties of LRD sources, a slightly more general scaling scheme will be used than in the introduction. Let v and a be scaling functions, i.e. increasing functions, each mapping the positive half-line onto itself. Define the log cumulant generating function (CGF) of the excess workload process by

$$\lambda_t^L(\theta) = (Lv(t))^{-1} \log \mathbf{E}[\exp(\theta W_t^L v(t)/a(t))]. \tag{11}$$

v and a are not arbitrary, for we will assume:

Hypothesis 1

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

$$\lambda_t(\theta) = \lim_{L\to\infty} \lambda_t^L(\theta) \qquad \text{and} \quad \lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta) \tag{12}$$

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

(ii) λ_t and λ are essentially smooth (see[9]).

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

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

$$\lim_{n\to\infty} \limsup_{L\to\infty} L^{-1} \log \sum_{n'\geq n} e^{\varepsilon L v(n')} = -\infty. \tag{13}$$

The role of the scaling functions a and v will be discussed further in section 2.3. The main theorem used here for the large L asymptotic is a slight adaptation of one from [10] (where it is proved also for continuous time processes under an additional regularity hypothesis).

Theorem 2
Under Hypothesis 1

$$\begin{aligned} -I(b) &\geq \limsup_{L\to\infty} L^{-1} \log \mathbf{P}[Q^L > Lb] && (14)\\ &\geq \liminf_{L\to\infty} L^{-1} \log \mathbf{P}[Q^L > Lb] \geq -I(b^+), && (15)\\ \text{with} \quad I(b) &= \inf_{t>0} \mu_t^*(b), && (16) \end{aligned}$$

where $^+$ denotes the limit from above, and f^* denotes the Legendre-Fenchel transform of a real function f, defined by

$$f^*(x) = \sup_\theta (x\theta - f(\theta)), \tag{17}$$

and specifically μ_t^* is the Legendre-Fenchel transform of μ_t defined by

$$\mu_t(\theta) := \lim_{L\to\infty} L^{-1} \log \mathbf{E}[e^{\theta W_t^L}]. \tag{18}$$

Proof
In [10] this theorem is proved with the variational expression

$$I(b) = \inf_{t>0} v(t)\lambda_t^*(b/a(t)) \tag{19}$$

in place of (16): so we just have to show that the two are equivalent. This follows from the fact that

$$\begin{aligned} v(t)\lambda_t^*(b/a(t)) &= v(t)\sup_\theta (\theta b/a(t) - \lambda_t(\theta)) && (20)\\ &= \sup_\theta(\theta b - \mu_t(\theta)) = \mu_t^*(\theta), && (21)\end{aligned}$$

since for fixed t we can rescale θ to $\theta v(t)/a(t)$ in (20). □

2.2 Example: Gaussian Processes

In the next section we will treat some examples of Gaussian processes with long-range dependence. Assume that A_t^L is a superposition of i.i.d Gaussian processes which we write as $A_t^L = \sum_L (A_t + mt)$, where A_t is a centered Gaussian processes and $m > 0$. In this scheme we regard A_t as describing fluctuations of arrivals in a single source about a mean rate m. Let σ_t be the variance of A_t. We write the service rate per source of $s + m$ with $s > 0$, so that m drops out of the expression for the excess workload process $W_t^L = \sum_L (A_t - st)$. Then

$$\mu_t(\theta) = \log \mathbf{E}[e^{\theta(A_t - st)}] = \theta^2\sigma_t^2/2 - st\theta, \tag{22}$$

and a simple calculation gives

$$\mu_t^*(b) = \frac{(b+st)^2}{2\sigma_t^2}. \tag{23}$$

An example of this is when A_t is fractional Brownian Motion (fBM) with Hurst parameter $H \in (0,1)$ [21]. In this case $\sigma_t^2 = t^{2H}$, and one establishes that

$$I(b) = \delta b^{2-2H} \quad \text{where} \quad \delta = \frac{1}{2}\left(\frac{s}{H}\right)^{2H}(1-H)^{-2(1-H)}. \tag{24}$$

Fractional Brownian Motion has been proposed as a model of Ethernet (and other) traffic: see [20].

2.3 The role of the scaling functions

It is worth remarking that the scaling functions a and v do not appear explicitly in the variational expression obtained from (16) and (18). Indeed, for each t, the existence of the limit μ_t (equivalent to the existence of the limit λ_t) is a reflection of the fact that the pair (W_t^L, L) is to satisfy a large deviation principle with rate function μ_t^*, a matter which is independent of the temporal structure of the excess workload process.

But the role of a and v is two-fold. Firstly, as discussed in the introduction, the form of the scaling functions determine the particular asymptotic from of $I(b)$ for large b, as displayed in eq. (8). When a and v are power-laws: $a(t) = t^a$ and $v(t) = t^v$ then $p = v/a$. Fractional Brownian motion provides an example of this, with $a(t) = t$, $v(t) = t^{2-2H}$. Secondly, the comparative simplicity of the formula (16) hides the fact that the proof of the upper bound (14) in [10] requires the existence of the limit λ.

2.4 Consequences for measurement schemes

The absence of the scaling functions from (18) has a useful consequence for resource allocation schemes based upon measurement of large deviation rate functions. Let us concentrate on the case of i.i.d. sources, in which case $\lambda_t^L = \lambda_t$ for all L, i.e. the CGF of the excess workload process of a single source served at rate s. Then μ_t and hence I could be estimated through measurement *without* making assumptions about the appropriate scaling functions a and v. Rather, one interprets the large L asymptotic (6) as saying that $L^{-1} \log \mathbf{P}[Q > Lb]$ is approximately independent of (sufficiently large) L. Thus measurements of loss probabilities for smaller L could be used to estimate those for larger L (see [3, 7] for further detail).

However, $a(t)$ and $v(t)$ *will* be required if one wishes to use large b asymptotic for fixed L: for single SRD sources by using (5); more generally using the large deviation asymptotic

$$\lim_{b\to\infty} \frac{1}{v(a^{-1}(b))} \log \mathbf{P}[Q > b] = \delta \tag{25}$$

from [12] of which (4) is a special case; or using (8) for many sources. This is seen already in schemes for online determination of δ for single sources proposed in [8, 13], where the estimation of δ for the linear asymptotics (2) is predicated on the assumption that processes are stationary and only SRD (at least during the period over which estimation is carried out).

In dealing with measurement schemes it is important to determine over what time scales measurements must be made to observe a given rare event. In the large L scheme outlined above, then for each buffer size b we can associated the time $t(b)$ at which the infimum in (16) is attained. The asymptotic properties of $t(b)$ are contained in the next result.

Proposition 3
Suppose Hypothesis 1 is satisfied with power-law scaling functions $v(t) = t^v$ and $a(t) = t^a$ with $a > v > 0$, and that for some $t_0 > 0$, $\liminf_{b\to\infty} \inf_{0<t<t_0} v(t)\lambda^*_{bt}(1/a(t)) \geq \delta$. Set $h(t) := v(a^{-1}(t)) = t^{v/a}$. Then $\delta = \inf_{c>0} h(1/c)\lambda^*(c)$ and

$$t(b) \sim b^{1/a} r^{-1/v}, \quad \text{where} \quad r = \left((\lambda^* \circ h^{-1})^*\right)'(\delta). \tag{26}$$

Proof
The formula for δ is from [12]. Following the proof of Theorem 2 in [10], then we can rewrite (19) by means of the time change $c = v(t)$ as

$$I(b) = J(h(b)), \text{ where } J(b) = \inf_{c>0} c f_c(b/c), \tag{27}$$

where $f_c = \lambda^*_{v^{-1}(c)} \circ h^{-1}$. In Theorem 4 of [10] it is shown that the infimum in (27), call it $\kappa(b)$, follows the asymptotic

$$\kappa(b) \sim b/r \qquad \text{as } b \to \infty. \tag{28}$$

Thus

$$t(b) \sim v^{-1}(\kappa(h(b))) = a^{-1}(b)/v^{-1}(r) \qquad \text{as } b \to \infty. \tag{29}$$

□

It is worth remarking that $a(t) = t$ in the models which we examine in this paper.

3 Fractional ARIMA processes

Let us now introduce some short-range structure into LRD processes. In this section we calculate the shape function for an arrival process whose increments are a Gaussian fractional ARIMA process. We review these briefly (following [18], to which we refer the reader for further details). The Gaussian random walk in discrete time is the ARIMA(0,1,0) process x_t defined by $\nabla x_t := (1-B)x_t = a_t$ where B is the backward shift operator $Bx_t = x_{t-1}$ and the a_t are i.i.d. standard Gaussian random variables: thus the *increments* of the process x_t are the stationary process x_t.

The fractional difference operator ∇^d is defined for real d by the power series

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k, \text{ where } \binom{d}{k} = \prod_{j=1}^{k} \frac{j+d-1}{j}. \tag{30}$$

The ARIMA(0,d,0) process is defined formally, by analogy with ARIMA(0,1,0), as the solution x_t to

$$\nabla^d x_t = a_t. \tag{31}$$

Setting $H = d+1/2$, then for $d \in (-1/2, 1/2)$ we can think of ARIMA(0,d,0) as a discrete time analogue of fractional Brownian motion with Hurst parameter H, the latter process being the $-d^{\text{th}}$ fractional difference of Brownian motion.

ARIMA(p,d,q) processes are formally defined as solutions x_t to

$$\phi(B)\nabla^d x_t = \theta(B) a_t \tag{32}$$

where ϕ and θ are polynomials of degree p and q respectively. The role of the polynomials ϕ and θ in modelling is (as it is for standard ARMA models) to introduce short-range structure (through autoregression and moving averages) into a process whose long-range structure is determined by d. $\phi = \theta = 1$ gives ARIMA(0,d,0) as a special case.

Proposition 4 [18]
Let x_t be an ARIMA(p,d,q) process with $-\frac{1}{2} < d < \frac{1}{2}$. Assume all the zeros of ϕ, θ lie outside the unit circle in the complex plane. Then x_t is stationary and has spectral density

$$f(\omega) = \gamma_0 + 2\sum_{t\geq 1} \gamma_t \cos(\omega t) = \frac{\theta(e^{i\omega})\theta(e^{-i\omega})}{\phi(e^{i\omega})\phi(e^{-i\omega})} (2\sin(\omega/2))^{-2d}, \tag{33}$$

for $0 < \omega \leq \pi$ where $\gamma_t = \mathbf{E}[x_k x_{k-t}]$.

Note: since x_t is stationary, $\gamma_t = \gamma_{-t}$, $f(\omega) = f(2\pi - \omega)$ is real, and so (33) extends to $(0, 2\pi)$.

3.1 Analysis of the Shape Function

Within the framework of section 2.2 we shall consider a queue whose excess workload process is an L-fold superposition of identical processes, $W_t^L = \sum_L (A_t - st)$. The arrivals per source processes at time $-t$ are $x_t + m$ where m is a drift and x_t is an ARIMA(p,d,q) process which satisfies the hypotheses of Prop. 4, with the additional requirement that $d > 0$ (correspondingly $H > 1/2$). Thus $A_t = \sum_{k=1}^t x_k$ Let σ_t^2 denote the variance of A_t. In our analysis of the shape function for ARIMA(p,d,q) processes, we shall need to establish the following result.

Proposition 5
(i)

$$\sigma_t^2 \sim \left(\frac{\theta(1)}{\phi(1)}\right)^2 \frac{t^{2d+1}}{\Gamma(2d+2)\sin(\pi d)}, \qquad t \to \infty. \tag{34}$$

(ii) $\sigma_1^2 > t^{-2}\sigma_t^2$ *for all* $t > 1$

Proof
(i)

$$\begin{aligned} \sigma_t^2 &= \sum_{1 \le k,j \le t} \gamma_{k-j} = \sum_{k=1}^t u_k \qquad \text{where} & (35) \\ u_k &= \gamma_0 + 2\sum_{j=1}^{k-1} \gamma_j . & (36) \end{aligned}$$

From Prop4 we have that $f(\omega) \sim (\theta(1)/\phi(1))^2 \omega^{-2d}$ as $\omega \to 0^+$. Thus by the Tauberian Theorem 4.10.1 of [1],

$$u_t \sim \left(\frac{\theta(1)}{\phi(1)}\right)^2 \frac{t^{2d}}{\Gamma(2d+1)\sin(\pi d)}, \qquad t \to \infty. \tag{37}$$

The stated result follows by summing this asymptotic relation using Theorem 1.5.8 of [1]
(ii) Since $f(\omega) > 0$,

$$\begin{aligned} t^2\sigma_1^2 - \sigma_t^2 &= (2\pi)^{-1} \int_0^{2\pi} d\omega \ f(\omega) \left(t^2 - \sum_{1 \le j,k \le t} \cos((j-k)\omega) \right) & (38) \\ &= (2\pi)^{-1} \int_0^{2\pi} d\omega \ f(\omega) \sum_{1 \le j,k \le t} (1 - \cos((j-k)\omega)) > 0. & (39) \end{aligned}$$

□

Theorem 6
(a) The excess workload process W_t^L *described above for ARIMA(p,d,q) arrivals with* $d \in (0, 1/2)$ *satisfies Hypothesis 1 with* $a(t) = t$ *and* $v(t) = t^{1-2d}$. *The corresponding shape function I has the following behaviour:*

(b) $I(b) \sim \delta b^{1-2d}$ *as* $b \to \infty$, *where*

$$\delta = 2^{-1}\Gamma(2d+2)\sin(\pi d)\left(\frac{\phi(1)}{\theta(1)}\right)^2\left(\frac{s}{d+1/2}\right)^{2d+1}(1/2-d)^{2d-1}. \tag{40}$$

(c) $I(0) = s^2/(2\sigma_1^2)$

(d) $I'(0) = s/\sigma_1^2$

Proof

(a) In Hypothesis 1(i), the first limit is trivial since the sources are identical and independent. For the second limit, observe that $\lambda_t(\theta) = v(t)\theta^2\sigma_t^2/(a(t))^2 - s\theta t/a(t) \sim \text{const.}\theta^2 - s\theta$ as $t \to \infty$; properties (ii) and (iii) follow straightforwardly, as does (iv) from the form of the power-law for $v(t)$.

(b) This follows from eq. (84) of [10] and Prop. 5(i).

(c) Recall $\mu_t^*(b) = (b+st)^2/(2\sigma_t^2)$. By (18), $I(0) = \inf_{t\geq 1}\mu_t^*(0)$. By Prop. 5(ii), $\mu_t^*(0) > \mu_1^*(0)$ for $t > 1$, and so the infimum is attained at $t = 1$.

(d) Since $\mu_t^*(0) > \mu_1^*(0)$ for $t > 1$, $b \to \mu_t^*(b)$ is increasing, and $b \to \mu_1^*(b)$ is continuous, then $\mu_t^*(b) > \mu_1^*(b)$ for all $t > 1$ and b in some neighbourhood $[0, \tilde{b})$. Thus $I(b) = \mu_1^*(b)$ for $b \in [0, \tilde{b})$ and the result follows by differentiation. □

3.2 Examples

We calculate the shape function for some examples of ARIMA(p,d,q) processes. For $p = q = 0$ we are able to calculate σ_t^2 explicitly from the spectral density f:

$$\sigma_t^2 = \frac{1}{2\pi}\int_0^{2\pi} d\omega f(w)\left(t + 2\sum_{k=1}^{t-1}(t-k)\cos(\omega k)\right) \tag{41}$$

$$= \frac{1}{2\pi}\int_0^{2\pi} d\omega f(w) g(t,w) \tag{42}$$

where

$$g(t,\omega) = \frac{1-\cos t\omega}{1-\cos\omega}. \tag{43}$$

By application of a standard result for $\int_0^\pi dx(\sin x)^{p-1}\cos(ax)$ (see p372 of [16]) this yields

$$\sigma_t^2 = \frac{\Gamma(-\frac{1}{2}-d)}{2^{2d+1}\sqrt{\pi}\,\Gamma(-d)} - \frac{2\cos(t\,\pi)\,\Gamma(-1-2\,d)}{\Gamma(-d-t)\,\Gamma(-d+t)} \tag{44}$$

For p, q not both zero, σ_t^2 is not generally available in closed form, and we have had to calculate σ_t^2 by numerical integration of (42) for particular choices of θ, ϕ and d. In both cases we calculate I by numerical minimization in (18) over a suitably large domain (the required size can be estimated from Prop. 3).

The results of these calculations are displayed for $d = \frac{1}{3}$ in Figure 2 and for $d = \frac{1}{4}$ in Figure 3. These shape functions were obtained after normalizing the spectral density by dividing by $(\theta(1)/\phi(1))^2$ in order that that the asymptotics of the variances σ_t^2 are the

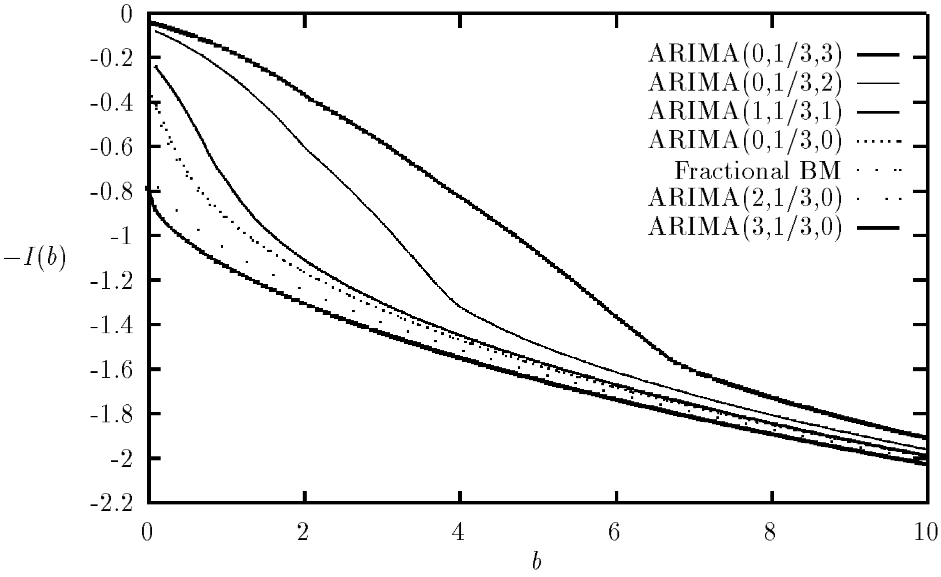

Figure 2: Shape functions for some ARIMA(p,$\frac{1}{3}$,q) processes

same for each process with a given d. Also plotted for comparison is the shape function for fBM with the same d and asymptotic variance.

Features common to both plots are as follows. (a) The shape function for ARIMA(0,d,0) is virtually identical with that of the corresponding fraction Brownian motion. (b) All curves approach the same asymptote for large b as expected from Theorem 6(i). (c) The difference between the shape functions and that for ARIMA(0,d,0) can take either sign. One can account for this at an intuitive level: for those processes for which the shape function lies below the ARIMA(0,d,0) curve have increments which are positively correlated at all lags. Those which lie above have increments which are negatively correlated for some finite set of initial lags, but which become positively correlated for all sufficiently large lags. (This can be related to the experience with the shape function for SRD sources: in the example of 2-state Markov sources, the shape function has the asymptotic form $I(b) - \delta b \sim \nu$ where the shift ν is positive (negative) when its increments are positively (negatively) correlated at lag 1; see [2]). However, we have not been able to establish a systematic connection between the sign of the autocorrelations and the sign of the shift between the curves.

4 Conclusions

Whereas it would be desirable to provide some further analysis to explain the features observed in Figures 2 and 3, this appears to be precluded by the complexity of the expressions involved. However, these numerical examples do demonstrate the short-range structure modifies the shape function $I(b)$ at small b from the corresponding pure power-

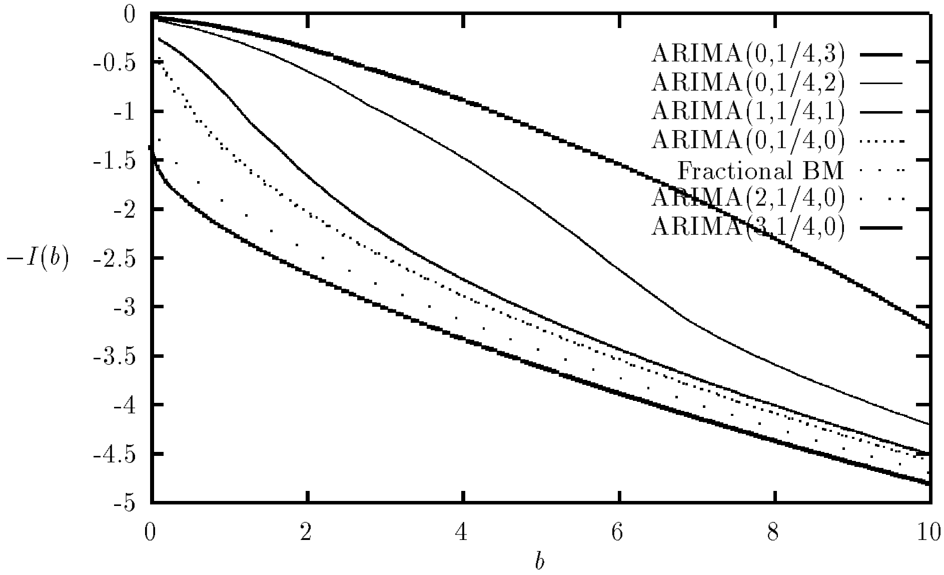

Figure 3: Shape functions for some ARIMA(p,$\frac{1}{4}$,q) processes

law for fBM. Thus schemes for buffer dimensioning based purely on determination of the Hurst parameter H could underestimate loss-ratios at smaller buffer sizes, since they ignore short-range structure.

Theorem 6 provides an estimate for the behavior of $I(b)$ for b in a neighbourhood of zero. Indeed, it appears from Figures 2 and 3 that the linear estimate $I(b) \approx I(0) + bI'(0)$ is conservative for the initial region of the graph in the case that it lies above the curve for fBM.

Acknowledgements

Thanks are due to Ashok Erramilli and Walter Willinger for a discussion which provided some motivation for this work.

References

[1] N.H. Bingham, C.M. Goldie and J.L. Teugels. *Regular Variation*, Encyclopedia of Mathematics and its Applications. Vol 27. Cambridge University Press, Cambridge, 1987.

[2] D.D. Botvich and N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Queueing Systems, 20 (1995) 293–320.

[3] D.D. Botvich, T.J. Corcoran, N.G. Duffield and P. Farrell. Economies of scale in long and short buffers of large multiplexers, Proceedings of 12^{th} IEE UK Teletraffic Symposium, Old Windsor, 15-17 March 1995.

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

[5] G.L. Choudhury, D.M. Lucantoni and W. Whitt. Squeezing the most out of ATM, IEEE Trans. Comm., to appear.

[6] C. Courcoubetis and R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, J. Appl. Prob., 1996, to appear.

[7] C. Courcoubetis, G. Fouskas and R. Weber, An On-line Estimation Procedure for Cell Loss Probabilities in ATM Links, Proceedings of the Third IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, 2-6 July, 1995

[8] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R.R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. IEEE Trans. Comm., 43 (1995) 1778–1784.

[9] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Jones and Bartlett, Boston, 1993.

[10] N.G. Duffield, Economies of scale in queues with sources having power-law large deviation scalings, J. Appl. Prob., 1996, to appear.

[11] N.G. Duffield, Conditioned asymptotics for tail probabilities in large multiplexers, Preprint, 1995.

[12] N.G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Math. Proc. Cam. Phil. Soc., 118 (1995) 363–374.

[13] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell and F. Toomey, Entropy of ATM traffic streams: a tool for estimating QoS parameters, IEEE JSAC, 13 (1995) 981–990.

[14] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell & Fergal Toomey. Predicting Quality of Service for traffic with long-range fluctuations, Proceedings of IEEE International Conference on Communications, Seattle, 18-22 June, 1995, pp473–477.

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

[16] I.S. Gradstheyn and I.W. Ryzhik, *Table of Integrals, Series and Products*, Academic, New York, 1965.

[17] D.P Heyman and T.V. Lakshman, What are the implications of long-range dependence on VBR-video traffic engineering. Preprint, 1995.

[18] J.R.M. Hosking, Fractional Differencing, Biometrika, 68 (1981) 165–176.

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

[20] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic, ACM SIGCOMM Computer Communications Review, 23 (1993) 183–193.

[21] B.B. Mandelbrot and J.W. Van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Review, 10 (1968) 422–437.

[22] I. Norros, A storage model with self-similar input, Queueing Systems, 16 (1994) 387–396.

[23] B.K. Ryu and A. Elwalid, The importance of Long-Range dependence of VBR Video Traffic in ATM Traffic Engineering. Preprint, 1996.

[24] A. Simonian and J. Guibert, Large deviations approximation for fluid queues fed by a large number of on-off sources, in *Proc. ITC 14*, 1994, pp.1013–1022.

[25] A. Weiss, A new technique for analyzing large traffic systems, J. Appl. Prob., 18 (1986) 506–532.

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