To appear in Queueing Systems in 1994

A storage model with self-similar input

Ilkka Norros
VTT Telecommunications
Otakaari 7 B, FI-02150 Espoo, Finland

Abstract

A storage model with self-similar input process is studied. A relation coupling together the storage requirement, the achievable utilization and the output rate is derived. A lower bound for the complementary distribution function of the storage level is given.

Keywords: Self-similar, fractional Brownian motion, Local Area Network traffic

1 Introduction

In a series of papers (e.g. Leland [8], Leland and Wilson [7], Fowler and Leland [4], Leland *et al.* [9]), researchers from Bellcore have reported and analyzed remarkable Local Area Network (LAN) traffic measurements challenging traditional data traffic modelling. The Bellcore data are both very accurate and extensive in time, and their most striking feature is the tremendous burstiness of LAN traffic at, practically, any timescale. More than that, the statistical analysis has shown that the traffic is *self-similar* with a surprising accuracy (see Leland *et al.* [9]).

Traditional traffic models based on the Poisson process or, more generally, on short-range dependent processes, cannot describe the behaviour of actual LAN traffic as observed in these measurements. Consequently, it becomes necessary to study storage systems with long-range dependent input processes. The system considered in this paper is perhaps the simplest of its kind.

In Section 2, we introduce a stochastic process $V(t)$ as a model for the content of a storage having self-similar input and being emptied at a constant rate. This is a continuous model which does not describe the movement of individual packets, but it gives already some new qualitative understanding of queueing phenomena in packet networks carrying traffic originating from a large number of LANs. The simplicity of the model makes it also mathematically attractive.

We present, essentially, two results for this model. A scaling law relating the storage requirement, the service rate and the server utilization is derived in Section 3. A lower bound for the complementary distribution function of the storage level is derived in Section 4.

Throughout this paper we denote by $Z(t)$, $t \in (-\infty, \infty)$, a normalized fractional Brownian motion with self-similarity parameter (Hurst parameter) $H \in [\frac{1}{2}, 1)$. This process is characterized by the following properties:

(i) $Z(t)$ has stationary increments;

(ii) $Z(0) = 0$, and $EZ(t) = 0$ for all t;

(iii) $EZ(t)^2 = |t|^{2H}$ for all t;

(iv) $Z(t)$ has continuous paths;

(v) $Z(t)$ is Gaussian, i.e. its finite-dimensional distributions are multivariate Gaussian distributions.

In the special case $H = \frac{1}{2}$, $Z(t)$ is the standard Brownian motion.

Most of the results in this paper do not depend on the Gaussian character of $Z(t)$ so that they can be immediately generalized by replacing $Z(t)$ by a more general self-similar process.

2 The fractional Brownian model

The object of our study is given in the following definition. The rest of this section is devoted to explaining and motivating this model.

Definition 2.1 The stationary storage model with fractional Brownian net input is the stochastic process $V(t)$, where

$$V(t) = \sup_{s \le t} \left(A(t) - A(s) - C(t-s) \right), \quad t \in (-\infty, \infty) \tag{2.1}$$

$A(t)$ being the process

$$A(t) = mt + \sqrt{am} Z(t), \quad t \in (-\infty, \infty), \tag{2.2}$$

and $Z(t)$ a normalized fractional Brownian motion. The system has four parameters m, a, H and C with the following interpretations and intervals of allowed values: $m > 0$ is the mean input rate, $a > 0$ is a variance coefficient, $H \in [\frac{1}{2}, 1)$ is the self-similarity parameter of $Z(t)$, and $C > m$ is the service rate.

We start with some explanatory remarks on the definition. First, it should be mentioned that although we have introduced a "traffic model" $A(t)$ and a constant leak rate C, it is

in fact mathematically relevant for $V(t)$ only that the *net input process* $X(t) = A(t) - Ct$ is of the form $c_1 Z(t) - c_2 t$ with $c_2 > 0$.

It is immediately seen that $V(t)$ is indeed a stationary process. Its a.s. finiteness is shown later in this section.

The formula (2.1) is similar to the well-known expression for the amount of work (or virtual waiting time) in a queueing system with service rate C and cumulative work arrival process $A(t)$. Beneš [1] calls it "Reich's formula", referring to Reich [11]. Cf. also Harrison [5], Section 2.2.

The choice of the process $A(t)$ can be understood as an analogue to and a generalization of a diffusion approximation for a Poisson process. Indeed, a Poisson process $N(t)$ with parameter m can be written in the form

$$N(t) = mt + M(t)$$

where $M(t)$ is the martingale $N(t) - mt$. It is well known that $(N(\alpha t) - \alpha m t)/\sqrt{\alpha m}$ converges towards the standard Brownian motion $W(t)$ as $\alpha \to \infty$ (e.g., Theorem VIII.3.11 of Jacod and Shiryaev [6]). This suggests the approximation of $N(t)$ by a diffusion process:

$$N(t) \approx mt + \sqrt{m} W(t). \tag{2.3}$$

In the Brownian case $H = \frac{1}{2}$ (with $a = 1$ and $C = 1$), $V(t)$ can thus be seen as a continuous approximation of the $M/D/1$ queue. It is well known that this diffusion approximation is asymptotically accurate as a heavy traffic limit. The definition (2.2) is obtained from (2.3) by replacing the Brownian motion by a fractional Brownian motion and adding a bit of flexibility with the constant coefficient a. The factor $\sqrt{m}$ is also motivated by the superposition property presented in Proposition 2.2 at the end of this section.

Let us then review some basic properties of the fractional Brownian motion $Z(t)$, defined at the end of Section 1. This process was used for modelling purposes already in the 1960's by Mandelbrot [10], who also gave its name. From the stationarity of the increments and the Gaussian property it follows that the finite-dimensional distributions of the process are determined by the mean and variance functions (properties (ii) and (iii)). The continuity assumption then completes the characterization of $Z(t)$.

It is now easy to see that $Z(t)$ is a *self-similar* process, i.e.

$$Z(\alpha t),\ t \in \mathbb{R}, \text{ is identical in distribution to } \alpha^H Z(t),\ t \in \mathbb{R},$$

for every $\alpha > 0$. Indeed, by the above remarks it is sufficient to note that $EZ(\alpha t)^2 = \alpha^{2H} EZ(t)^2$ for any t. As a general reference to self-similar processes, see, e.g., articles in the collection Eberlein and Taqqu [3].

For $t_1 < t_2 < t_3 < t_4$ we have

$$\begin{aligned} &\operatorname{Cov}\left(Z(t_2) - Z(t_1), Z(t_4) - Z(t_3)\right) \\ &= \frac{1}{2}\left((t_4 - t_1)^{2H} - (t_3 - t_1)^{2H} + (t_3 - t_2)^{2H} - (t_4 - t_2)^{2H}\right). \end{aligned}$$

In particular,

$$r(n) =_{\text{def}} \text{Cov}\,(Z(1), Z(n+1) - Z(n)) = H(2H-1)n^{-2(1-H)} + \mathcal{O}(n^{-(3-2H)}),$$

which shows that in the case $H > 1/2$ the increments of $Z(t)$ are positively correlated and the process possesses *long-range dependence* in the sense that $\sum_0^\infty r(n) = \infty$. (Note that in traditional traffic models the increments either are independent or have exponentially fast vanishing correlations.)

One might wonder whether the so strongly correlated stationary sequence $Z(n+1) - Z(n)$ (often called *fractional Gaussian noise*) is ergodic — non-ergodicity would be an unpleasant feature by a traffic model! In fact, the ergodicity of the fractional Gaussian noise follows from the general result that any stationary Gaussian sequence with continuous spectral measure is ergodic and weakly mixing — see, e.g., Cornfeld *et al.* [2], Theorem 14.2.1.

The a.s. finiteness of the process $V(t)$ defined by (2.1) follows now from Birkhoff's ergodic theorem. Indeed, we have $\lim_{t\to\infty} Z(t)/t = EZ(1) = 0$ a.s., which together with the assumption $m < C$ implies that

$$\lim_{s\to-\infty} (A(t) - A(s) - C(t-s)) = -\infty \quad \text{a.s.}$$

Many features of the fractional Brownian motion are different from those of most stochastic processes usually appearing in traffic models. It is indeed far from being a Markov process, and it is not even a semimartingale. Therefore, most of the standard methods of storage theory are not applicable. However, some insight into the properties of our model, being of great interest in teletraffic theory because of the empirical results mentioned in Section 1, can be obtained by simple means, as will be shown in the two remaining sections. We close this section by noting a superposition property of our traffic model, which is easily verified.

Proposition 2.2 Consider the processes $A_i(t)$, $i = 1, \ldots, K$, defined as

$$A_i(t) = m_i t + \sqrt{m_i a} Z_i(t), \quad t \in \mathbb{R},$$

where the m_is are arbitrary positive numbers, $a > 0$ and the processes $Z_i(t)$ are independent fractional Brownian motions with a common parameter H. Then the superposition

$$A(t) = \sum_{i=1}^{K} A_i(t)$$

can be written as $A(t) = mt + \sqrt{ma}Z(t)$, where $m = \sum_1^K m_i$ and $Z(t)$ is a fractional Brownian motion with parameter H.

Proposition 2.2 shows that the roles of the three parameters of the traffic model (2.2) can be separated so that H and a characterize the "quality" of the traffic in contrast to the long run mean rate m which characterizes its "quantity" alone.

3 Scaling laws

Some interesting properties of the fractional Brownian storage model $V(t)$ can be deduced from the self-similarity assumption alone. Let us first shortly consider $V(t)$ as a stochastic process at different time scales.

Theorem 3.1 Consider a process $V(t)$ with parameters m, H, a and C as in Definition 2.1, and let $\alpha > 0$ be an arbitrary number. Then the process $V(\alpha t)$ is distributed like α^H times the corresponding process arising from a fractional Brownian model with the original arrival process but with service rate $m + \alpha^{1-H}(C - m)$.

Proof We have

$$\begin{aligned} V(\alpha t) &= \sup_{s \le t} \Big(A(\alpha t) - A(\alpha s) - C \cdot (\alpha t - \alpha s) \Big) \\ &= \sup_{s \le t} \Big(m\alpha(t-s) + \sqrt{ma}(Z(\alpha t) - Z(\alpha s)) - C\alpha(t-s) \Big) \\ &=_{\text{(d)}} \alpha^H \sup_{s \le t} \Big(m\alpha^{1-H}(t-s) + \sqrt{ma}(Z(t) - Z(s)) - C\alpha^{1-H}(t-s) \Big) \\ &= \alpha^H \sup_{s \le t} \Big(A(t) - A(s) - (m + \alpha^{1-H}(C-m))(t-s) \Big), \end{aligned}$$

where $=_{\text{(d)}}$ means that the whole processes are similar in distribution. □

Let us then analyze the distribution of $V(0)$ (recall that $V(t)$ is a stationary process). A typical requirement in a telecommunications application would be that the probability that the amount of work in system exceeds a certain level x is required to be at most equal to a "Quality of Service parameter" ϵ. (The value x is the substitute for the size of the storage in our infinite storage model.) Thus, the following relation holds at the maximal allowed load:

$$\epsilon = P\,(V > x)\,. \tag{3.1}$$

Equation (3.1) can also be interpreted as defining a *storage requirement* x. Further, it defines a hypersurface in the space of system parameters, separating the allowed parameter combinations from unallowed ones. Now, the self-similarity of $Z(t)$ allows for deriving from (3.1) a more explicit relation between the design parameters x (storage requirement), C (service rate) and $\rho = m/C$ (utilization) at the critical boundary. As an application, this result gives us some insight into the management of teletraffic with long-range dependence.

Theorem 3.2 Assuming (3.1), the following equation holds:

$$\frac{1-\rho}{\rho^{1/(2H)}} \cdot C^{(H-\frac{1}{2})/H} \cdot x^{(1-H)/H} = const, \tag{3.2}$$

where the constant at the right hand side depends on H, a and ϵ but not on ρ, C or x.

Proof Note first that $V(0)$ is distributed like its time-reversed counterpart $\sup_{t\geq 0}(A(t) - Ct)$. Consider now the function

$$q(x,\beta) = P\left(\sup_{t\geq 0}(Z(t) - \beta t) > x\right).$$

By the self-similarity of $Z(t)$ we have

$$q(\alpha x,\beta) = P\left(\sup_{t\geq 0}[Z(\frac{t}{\alpha^{1/H}}) - \frac{\beta}{\alpha}t] > x\right) = q(x,\alpha^{\frac{1-H}{H}}\beta),$$

so that

$$q(x,\beta) = q(1,x^{\frac{1-H}{H}}\beta) = f(x^{\frac{1-H}{H}}\beta),$$

where the function

$$f(y) = q(1,y) = P\left(\sup_{t\geq 0}(Z(t) - yt) > 1\right)$$

is obviously strictly decreasing for $y \geq 0$, $f(0) = 1$ and $f(\infty) = 0$. Thus we may write

$$\begin{aligned}\epsilon &= P\left(V > x\right) = P\left(\sup_{t\geq 0}\left[Z(t) - \frac{C-m}{\sqrt{am}}t\right] > \frac{x}{\sqrt{am}}\right)\\ &= f\left(\left(\frac{x}{\sqrt{am}}\right)^{(1-H)/H}\cdot\frac{C-m}{\sqrt{am}}\right).\end{aligned} \tag{3.3}$$

Substituting $\rho = m/C$ we obtain the desired equation

$$\frac{1-\rho}{\rho^{1/(2H)}}\cdot C^{(H-\frac{1}{2})/H}\cdot x^{(1-H)/H} = a^{1/(2H)}\cdot f^{-1}(\epsilon). \tag{3.4}$$

□

In the Brownian case $H = \frac{1}{2}$, (3.4) reduces to

$$\frac{1-\rho}{\rho}\cdot x = const. \tag{3.5}$$

In this traditional case (a heavy traffic approximation of the $M/D/1$ queue) we may roughly say that reducing the relative free capacity $1-\rho$ by half costs doubling the storage size. The service rate C has disappeared from the equation, which means that it has nothing to do with relative utilization.

With $H > 1/2$ the situation is different. Let us first fix C and solve the storage requirement x as a function of ρ:

$$x = x(\rho) = const\cdot\rho^{1/(2(1-H))}\cdot(1-\rho)^{-H/(1-H)}. \tag{3.6}$$

It is seen that when H is high (the Bellcore measurements mostly give for H values in the region $(0.8, 0.9)$), a substantial increase in utilization, say again halving the free capacity,

requires a tremendous amount more storage space. Thus we have a new argument for the widely accepted view that for connectionless packet traffic the utilization factor cannot be practically improved by enlarging the buffers more and more.

Now, however, the absolute service rate C also effects the relative utilization ρ. Fixing x, we can solve C from (3.4) and get

$$C = C(\rho) = const \cdot \rho^{1/(2H-1)} \cdot (1-\rho)^{-H/(H-\frac{1}{2})}. \tag{3.7}$$

The important practical consequence of equation (3.7) is that transmission links with higher capacity can be used with higher utilization without increasing the buffers. The intuitive reason for this is the improved multiplexing efficiency. (Note that by Proposition 2.2, "increasing the amount of traffic" means increasing the parameter m, or $\rho = m/C$.)

4 A lower bound for the storage level

The following theorem gives a lower bound for the complementary distribution function of the fractional Brownian storage process. This bound turns out to be asymptotically (in a logarithmic sense) exact for the Brownian model, but we have, regrettably, no estimate for its accuracy in the general case. In this section we keep the service capacity fixed, choosing the units so that $C = 1$.

Theorem 4.1 Let $V(t)$ the stationary process of Definition 2.1. Then

$$P\left(V(t) > x\right) \geq \overline{\Phi}\left(\frac{1}{\sqrt{am}} \cdot \left(\frac{1-m}{H}\right)^{H} \left(\frac{x}{1-H}\right)^{1-H}\right), \tag{4.1}$$

where $\overline{\Phi}(y) = P\left(Z(1) > y\right)$ is the residual distribution function of the standard Gaussian distribution.

Proof Let us reverse the time as in the proof of Theorem 3.2, denote $V = V(0)$ and consider the lower bound

$$P\left(V > x\right) \geq \max_{t \geq 0} P\left(A(t) > t + x\right) = \max_{t \geq 0} \overline{\Phi}\left(\frac{(1-m)t + x}{\sqrt{am} \cdot t^{H}}\right), \tag{4.2}$$

where the equality follows from the self-similarity of $Z(t)$. It is seen by differentiating that the maximum in (4.2) is obtained at

$$t = \frac{Hx}{(1-H)(1-m)},$$

yielding the assertion of the theorem. □

Using further the approximation

$$\overline{\Phi}(y) \approx (2\pi)^{-1/2}(1+y)^{-1}\exp(-y^2/2) \sim \exp(-y^2/2), \tag{4.3}$$

we obtain the logarithmically asymptotical lower bound

$$\max_{t\geq 0} P\left(A(t) > t + x\right) \sim \exp\left(-\left[\frac{1}{2am(1-H)^2}\left(\frac{(1-m)(1-H)}{H}\right)^{2H}\right]\cdot x^{2(1-H)}\right). \tag{4.4}$$

The tail behaviour of the storage level in the fractional Brownian model is thus in the best case Weibullian: $P\left(V > x\right) \sim \exp(-\gamma x^{\beta})$ with $\beta \leq 1$. The value of the self-similarity parameter H has a tremendous significance for the storage requirement, showing how misleading the traditional models are when the real traffic is self-similar.

For the Brownian case $H = 1/2$, $a = 1$, the expression (4.4) reduces to the exponential distribution

$$\max_{t\geq 0} P(A(t) > t + x) \sim \exp\left(-2\frac{1-m}{m}\cdot x\right), \tag{4.5}$$

which is the well-known heavy traffic asymptotics of the $M/D/1$ system. It is interesting to note that the lower bound approximation (4.2) and approximation (4.3) happen to cancel out each other so that (4.5) gives in fact *exactly* $P\left(V > x\right)$ for the Brownian model. (See Takács [13], Chapter 6; in Roberts [12], this result is easily proven using the standard result on the maximum of a Brownian motion with negative drift.)

Acknowledgement I wish to thank Will Leland and Walter Willinger from Bellcore for kindly providing preprint material of their recent work (the work presented in this paper was originally based only on Fowler and Leland [4]). They also identify the fractional Gaussian noise as the simplest way to model the self-similarity of LAN traffic. Thanks are also due to the Editor and the referee for suggestions which considerably improved the presentation.

References

[1] V.E. Beneš. *General Stochastic Processes in the Theory of Queues.* Addison Wesley, 1963.

[2] I.P. Cornfeld, Ya.G. Sinai, and S.V. Fomin. *Ergodic Theory.* Springer Verlag, Berlin, 1982.

[3] E. Eberlein and M.S. Taqqu, editors. *Dependence in Probability and Statistics*, volume 11 of *Progress in Prob. and Stat.* Birkhäuser, Boston, 1986.

[4] H.J. Fowler and W.E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE J. Sel. Areas in Commun.*, 9(7):1139–1149, 1991.

[5] J.M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, New York, 1985.

[6] J. Jacod and A.N. Shiryaev. *Limit Theorems for Stochastic Processes.* Springer Verlag, 1987.

[7] W. Leland and D. Wilson. High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. In *IEEE INFOCOM'91*, 1991.

[8] W.E. Leland. LAN traffic behavior from milliseconds to days. In *7th ITC Specialists Seminar, Morristown*, October 1990.

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

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

[11] E. Reich. On the integrodifferential equation of Takács I. *Annals of Mathematical Statistics*, 29:563–570, 1958.

[12] J.W. Roberts, editor. *Performance Evaluation and Design of Multiservice Networks. COST 224 Final Report.* CEC, 1992.

[13] L. Takács. *Combinatorial Methods in the Theory of Stochastic Processes.* John Wiley & Sons, New York, 1967.