The Concept of Relevant Time Scales and Its Application to Queuing Analysis of Self-Similar Traffic (or Is Hurst Naughty or Nice?)

Arnold L. Neidhardt and Jonathan L. Wang
Bellcore
331 Newman Springs Road
Red Bank, NJ 07701
{arnie, jwang}@bellcore.com

Abstract

Recent traffic analyses from various packet networks have shown the existence of long-range dependence in bursty traffic. In evaluating its impact on queuing performance, earlier investigations have noted how the presence of long-range dependence, or a high value of the Hurst parameter H, is often associated with surprisingly large queue sizes. As a result, a common impression has been created of expecting queuing performance to be worse as H increases, but this impression can be misleading. In fact, there are examples in which larger values of H are associated with smaller queues. So the question is how can one tell whether queuing performance would improve or degrade as H rises? In this paper, we show that the relative queuing performance can be assessed by identifying a couple of time scales. First, in comparing a high-H process with a low-H process, there is a unique time scale t_m at which the variances of the two processes match (assuming exact, second-order self similarity for both processes). Second, there are time scales t_{qi} that are most relevant for queuing the arrivals of process i. If both of the queuing scales t_{qi} exceed the variance-matching scale t_m, then the high-H queue is worse; if the queuing scales are smaller, then the low-H queue is worse. However, no firm prediction can be made in the remaining case of t_m falling between the two queuing scales. Numerical examples are given to demonstrate our results.

1 Introduction

In the past 5 years, large amounts of traffic measurements from working packet networks (including Ethernet LANs, wide-area TCP/IP, CCSN/SS7, ISDN, VBR video, Frame Relay, and ATM) have been collected and analyzed. The results reported in [1, 3, 4, 10, 11, 13, 16, 17, 18, 21] have been striking for two reasons: (i) these studies demonstrate that it is possible to distinguish clearly between actual packet-network traffic and traffic generated by traditional Markovian models, and (ii) in sharp contrast to the traditional packet-traffic models, aggregate packet streams are statistically *self-similar* or *fractal* in nature; that is, realistic network traffic looks the same when measured over time scales ranging from milliseconds to minutes and hours.

From a modeling viewpoint, [16, 20] emphasize that finding self similarity in traffic measurements from modern packet networks does not necessitate complicated and highly parameterized traffic models. In fact, [16, 20] suggest compact and parsimonious modeling based on *fractional Brownian motion* (FBM, for short) processes as a feasible and attractive alternative to traditional modeling approaches. More recent results obtained in [6] provide conditions under which FBM-based models can be expected to describe packet traffic in modern packet networks realistically and predict their performance accurately. At the same time, there is mounting evidence that, beyond its omnipresence and statistical significance in measured data, long-range dependence is a traffic characteristic that (i) has a measurable and practical impact on queuing behavior, (ii) is of crucial importance for a number of packet-traffic-engineering problems (e.g., traffic measurements [8], buffer sizing [6], admission control [9], and rate control [7]), and (iii) if ignored, typically results in overly optimistic performance predictions and inadequate network-resource allocations. These earlier investigations (see for example, [5, 6, 9, 16, 20]) have noted how the presence of long-range dependence, or a high value of the Hurst parameter H, is often associated with surprisingly large queue sizes. As a result, a common impression has been created of expecting queue performance to be worse as H rises, but this impression can be misleading. In fact, there are examples in which larger values of H are associated with smaller queues [14, 15].

In this paper, we show that the relative queuing performance can be assessed by identifying a couple of time scales: i) a time scale in the arrival process, and ii) a relevant time scale in the queuing. The paper is organized as follows. In the next section, we will describe self-similar traffic and the FBM model. It is described to the extent that is relevant to our discussion for the rest of the paper. Section 3 derives and discusses the relevant times in the queuing as well as arrival processes. Specifically, there are time scales t_{qi} that are most relevant for queuing the arrivals of process i, usually identifiable as $\frac{B}{C-m_i} \cdot \frac{H_i}{1-H_i}$, where B is the queue size, C is the capacity of the queue server, and m_i is the mean arrival rate of process i. Further, in comparing a high-H process with a low-H process, there is a unique time scale t_m at which the variances of the two processes match. Section 4 establishes elementary conditions for determining which queuing performance is worse, with illustrative examples demonstrating the results. Finally, Section 5 concludes this paper by summarizing the results.

2 Fractional Brownian Motion and Hurst Parameter

With recent analyses of traffic measurements of various packet-based networking technologies and services, it is now widely accepted that real traffic has variations over many time scales ("fractal") and exhibits scaling relations ("self-similar"). This is in contrast to the simple traffic patterns that are currently assumed in most engineering practices and load testing.

Quite generally, an arrival process refers to a family of random variables $A(s,t)$, with the interpretation as the amount of traffic arriving in the time interval $(s,t]$. In support of this interpretation, the family must satisfy

$$A(r,t) = A(r,s) + A(s,t). \tag{1}$$

Usually, the greatest interest is in stationary processes, so stationarity is assumed here.[1] The arrival process can also be identified with the singly indexed process $A(t) = A(0,t)$ for $t \geq 0$ (and $A(t) = -A(t,0)$ for $t < 0$), because the doubly indexed process can be recovered from it: $A(s,t) = A(t) - A(s)$. (The singly indexed process is not stationary, however; instead, it is described as having stationary increments.) A consequence of stationarity, plus assumptions of finite means and of positive average arrivals in positive time, is the existence of a mean arrival rate $m > 0$:

$$\mathrm{E}[A(s,t)] = m \cdot (t-s). \tag{2}$$

Now self similarity of an arrival process actually refers to the fluctuations about the mean, and *not* to the process itself. In detail, first let X be the process of fluctuations:

$$X(s,t) = A(s,t) - m \cdot (t-s). \tag{3}$$

Self similarity means that for any (time) factor $\alpha > 0$, letting X_α be the time-scaled process $X_\alpha(s,t) = X(\alpha s, \alpha t)$, there must be a corresponding (space) factor $\beta(\alpha) > 0$ for which X_α matches $\beta(\alpha)X$. Strict self similarity means the match must be in the sense of distributions: X_α and $\beta(\alpha)X$ must have the same distribution. Second-order self similarity means the match is in the sense of the first two moments: the means and covariances of X_α must be those of $\beta(\alpha)X$. In particular, the moments exist and are finite. Regardless of the version of self similarity, if a process is self-similar, then the function β satisfies $\beta(\alpha_1\alpha_2) = \beta(\alpha_1)\beta(\alpha_2)$. Hence, β is given by a power law, at least within the set of rational powers of a generating α value. Indeed, if β is measurable (which would follow from the technical condition that A be jointly measurable as a function of both the explicit index (s,t) and the implicit random sample), then the function β is given by a power law without restrictions: for some $H \in \mathcal{R}$, for all $\alpha > 0$,

[1] An arrival process A is said to be *stationary*, if for any $r \in \mathcal{R}$, letting A_r be the translated process defined by $A_r(s,t) = A(s-r, t-r)$, it turns out that A_r has the same distribution as A.

$$\beta(\alpha) = \alpha^H. \tag{4}$$

Moreover, the Hurst parameter H must satisfy $H \geq 0$ in general, and if second moments are finite, then $H \leq 1$. The assumption is made in this paper that the first two moments are finite.

The assumptions on arrival processes alone imply that just a few parameters can determine the first two moments of all the random variables in the process. Specifically, the means and covariances of all the random variables in a self-similar arrival process are determined by three parameters: a mean rate $m > 0$, a Hurst parameter $H \in [0, 1]$, and some variance parameter σ^2 or peakedness parameter $a = \sigma^2/m$ to fix the size of the fluctuations at unit time. Explicitly,

$$\mathrm{E}[A(s,t)] = m \cdot (t - s), \tag{5}$$

$$\mathrm{Var}[A(s,t)] = V(t-s) = \sigma^2|t-s|^{2H}, \tag{6}$$

and

$$\begin{aligned} & \mathrm{Cov}[A(p,q), A(s,t)] \\ = \; & \frac{1}{2}\{V(t-p) + V(s-q) - V(t-q) - V(s-p)\} \\ = \; & \frac{\sigma^2}{2}\left\{|t-p|^{2H} + |s-q|^{2H} - |t-q|^{2H} - |s-p|^{2H}\right\} \end{aligned} \tag{7}$$

For definiteness, the further assumption will be made that the "arrival" process is Gaussian. Then the full distribution of A is determined by the three parameters m, H, and a (or σ^2), and is a *fractional Brownian motion* (FBM). An unrealistic aspect of this assumption is that negative values of $A(s,t)$ can occur (corresponding to negative number of arrivals) with positive probability. Indeed, so long as $a > 0$, it follows that with probability 1, there is a time $t \in (0,1)$ with $A(t) < 0$. In this sense, A cannot be an arrival process. Nevertheless, by the central-limit theorem, any real traffic is approximately FBM, if it is the aggregate of independent, identically distributed streams that are approximately second-order self-similar. Accordingly, we will simply proceed under the assumption that the "arrival" process is FBM. Note that for $H = 0.5$, FBM becomes ordinary Brownian motion. The increment process $Y = (Y(k) = A(k, k+1) : k \geq 0)$ is called *fractional Gaussian noise* (FGN) and is a stationary (discrete-time) Gaussian process with autocorrelation function $r(k) = 1/2(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H})$, $k \geq 1$. It is easy to see that, asymptotically, $r(k) \sim H \cdot (2H-1)|k|^{2H-2}$, for $1/2 < H < 1$, i.e., Y exhibits *long-range dependence* (see [16]). Also, simple calculations show that the *aggregated processes* $A^{(n)} = (A^{(n)}(k) = n^{-H}(Y(kn-n+1)+\ldots+Y(kn)) : k \geq 1)$, $n > 0$, all have the same distribution as Y, i.e., Y is *exactly self-similar* in the sense of [2].

From the packet-traffic-modeling viewpoint, the FBM traffic model is a reasonable representation of *aggregate* data traffic (i.e., formed by multiplexing a large number of independent data sources). This has recently been observed and validated in traffic analyses of various packet-network technologies and services, see for example [1, 10, 11, 16, 18, 21]. In these comparisons of FBM models to real traffic, it has been noted that the match is poor at the smallest time scales where physical limitations govern traffic generation, but that the match at intermediate time scales seems to extend to the longest time scales for which the data allow a comparison. Thus, these comparisons indicate that while real traffic is not exactly self-similar, it does seem to be asymptotically self-similar. To emphasize the correspondence of real traffic with the abstract FBM model, note that the three FBM parameters do capture significant features of the real traffic, as follows. m is the mean rate (or equivalently, resource utilization) that measures the *volume* or "quantity" of traffic. The other two parameters refer to the *burstiness* or "quality" of traffic. a (the peakedness) measures the magnitude of fluctuations about the mean rate. *At the unit time scale*, it is the ratio of the variance of packet counts to the mean value. H (the Hurst parameter) is an indication of the rate of decay of correlations in the traffic. As noted earlier, the combination of $\{m,\ a,\ H\}$ is a complete description of the model for an aggregate data-traffic stream. Thus, to the extent that FBM models do capture the traffic features relevant for performance, any two traffic streams with the same $\{m,\ a,\ H\}$ will result in the same performance.

The parameter H is obviously important, for it describes the existence and, for the case of $0.5 < H < 1$,[2] the intensity of any long-range dependence (LRD). Moreover, earlier investigations (see for example, [6, 9, 20]) have noted that a high value of the Hurst parameter H is often associated with surprisingly large queue sizes. As a result, a common impression has been created of expecting queue performance to be worse as H rises, but this impression can be misleading. In fact, there are examples in which larger values of H are

[2] Traffic with $H < 0.5$ has also been observed [12, 13], in which case a high traffic period is likely to be followed by a low traffic period and is referred to as *anti-persistent*.

associated with smaller queues as demonstrated in [14, 15], and was referred to as the *cross-over effect*. So the question is how can one tell whether queuing performance would improve or degrade as H rises? In the following section, we introduce the concept of relevant time scales both in the arrival and queuing processes and show its application in comparing the queuing performance driven by self-similar arrival processes.

3 Relevant Time Scales

To address the posed question concerning two arrival processes, we will in this section, describe the relevant time scales in comparing the arrival processes to the queuing system and to each other. Comparison of these relevant time scales determines whether the arrival process with the higher Hurst parameter encounters poorer performance.

3.1 Relevant Queuing Time Scales

In this paper, we assume that the queuing system is characterized merely by a service rate or capacity C and a buffer size B. Indeed, the "jobs" (or packets) on which the server works are conceived as a fluid, with the server passing fluid at rate C, with the queue size measuring the amount of fluid that has arrived without being passed, and with the arrival process $A(s,t)$ giving the amount of fluid arriving in the interval $(s,t]$. To simplify calculations, we will estimate probabilities associated with the hypothetical, infinite-capacity queue Q that never overflows, as opposed to the "real" queue that is bounded by B. Thus, the hypothetical, unbounded queue Q can be expressed in terms of the arrivals A as

$$Q(t) = \sup_{s:s\leq t} [A(s,t) - C \cdot (t-s)], \tag{8}$$

which can be verified as follows. To see that $Q(t)$ is at least as large as the supremum, note that for any $s \leq t$, each fluid particle in the total amount $Q(s) + A(s,t)$ either leaves the system or stays to contribute to $Q(t)$. Since the amount leaving the system in $(s,t]$ is bounded by $C\cdot(t-s)$, $Q(t)$ is at least as large as $Q(s)+A(s,t)-C\cdot(t-s) \geq A(s,t)-C\cdot(t-s)$. Thus, $Q(t)$ is at least as large as the supremum. For the converse, let s be the last time before t with $Q(s) = 0$. Because the queue was nonempty since that time, the server must have been busy in (s,t), so $C \cdot (t-s)$ really did leave, and $Q(t)$ is exactly $Q(s)+A(s,t)-C\cdot(t-s) = A(s,t)-C\cdot(t-s)$. Thus, Equation (8) is correct. Note that the stationarity of A implies the stationarity of the process Q. Hence, all the $Q(t)$ have the same distribution. As a surrogate for queue overflows, we will estimate $\Pr(Q > B)$, the probability that the hypothetical, unbounded queue exceeds the real buffer size B.

With these background assumptions, the question of a comparison of two processes should be a bit more concrete. Recall that the question was which arrival process would create worse queuing performance, given two self-similar arrival processes "1" and "2" with different Hurst parameters H_1 and H_2 (without loss of generality, we assume $H_1 < H_2$). Thus, the two arrival processes under consideration are A_1 and A_2, with mean rates m_1 and m_2, and with variance functions V_1 and V_2, and the corresponding queues are Q_1 and Q_2, where the queuing systems serving the arrivals are the same in the two cases (i.e., $C_1 = C_2 = C$ and $B_1 = B_2 = B$). The question is whether $\Pr(Q_1 > B)$ is larger than $\Pr(Q_2 > B)$.

First, the capacity C can be separated into a mean rate m_i and an "excess" (or spare) capacity $c_i = C - m_i$. The reason is that there is no difference between the queue formed by arrivals A_i at a server with rate C and the queue formed by the reduced arrivals $X_i(s,t) = A_i(s,t) - m_i \cdot (t-s)$ at a server with the reduced capacity c_i. (This observation can be justified mathematically from the definition Equation (8) of the queues in the two cases: the suprema are over exactly the same differences, whether $m_i \cdot (t-s)$ is subtracted from both terms or not.[3]) In other words, the queue is most intuitively imagined as being driven entirely by the *fluctuations* in the arrivals, with these fluctuations being served just by the "excess" capacity c_i. This excess capacity c_i can be regarded as having the role of controlling the size of the queue. Thus, c_i determines the queue-size distribution, with larger values of c_i tending to make the queue smaller, i.e., to concentrate the mass of the queue-size distribution onto smaller neighborhoods of zero. Then the buffer size B is merely a threshold, which is high enough (one hopes) that the queue-size distribution assigns sufficiently small mass to the interval (B, ∞). (The conception of B as merely a threshold is especially relevant in our conception of Q_i as referring to hypothetical, unbounded queues.)

[3] There is a conceptual difference, however, in that A_i in general can be an arrival process in the sense of being always positive, but X_i cannot be an arrival process in this sense. Of course, our assumption of FBM arrivals has already invalidated the positivity constraint.

This conception of queues being driven by arrival fluctuations supports the idea of a distinction in the service quality between the first-order effect of ensuring that capacity exceeds the mean, and the second-order effect of handling the queue. In one sense, this distinction is correct, in that the mean rate m_i depends only on the first moments of the arrival process, while the queue is driven by the fluctuations X_i, which have no first-order dependence whatsoever. Quantitatively, however, this distinction can be misleading if one jumps to the conclusion that second-order effects are always "smaller" than first-order effects. Specifically, the arriving traffic can be so variable that it can be necessary to install much more "excess" capacity c_i than the "first-order" capacity m_i. Nevertheless, for our comparison of queuing performance with different arrival streams distinguished by different Hurst parameters, which refers to different second-order behavior, we will assume that the two streams share the same first-order behavior. In other words, we assume that the mean rates are the same: $m_1 = m_2 = m$. With this assumption, it follows that the excess capacities are the same: $c_1 = c_2 = c = C - m$.

Large queues can form in a variety of ways. Specifically, when a queue forms that is large enough to exceed B, the period of formation can be long or short. Nevertheless, it cannot be extremely long, because while the fluctuations X_i in the arrivals have mean zero, the excess capacity c is strictly positive ($C > m$ for a stable queue), which means that the server will eventually "catch up" to any burst in arrivals. Similarly, unless instantaneous bursts of size B are common, the period of formation cannot be too short. Indeed, if arrivals are FBM with $H > 0$, then arrivals are continuous with no instantaneous bursts. Similarly, if arrivals are merely approximately FBM, so long as B is large enough that the FBM approximation is relevant, the instantaneous bursts of size B are rare enough to be neglected. Mathematically,

$$\{Q_i > B\} = \bigcup_{f>0} \{X_i(-f,0) > B + cf\}, \tag{9}$$

and the point is that within this union, the events corresponding to intermediate periods f of formation have the largest probabilities.

For simplicity, we consider just one of the events in the union of Equation (9), corresponding to one period f of formation.[4] Let

$$D_i(f) = \sqrt{V_i(f)} \tag{10}$$

and

$$Z_i(f) = \frac{X_i(-f,0)}{D_i(f)}, \tag{11}$$

so $Z_i(f)$ is a zero-mean, unit-variance random variable. In terms of this normalized random variable, the event under consideration is

$$E_i(f) = \{Z_i(f) > z_i(f)\}, \tag{12}$$

where

$$z_i(f) = \frac{B + cf}{D_i(f)} \tag{13}$$

is the normalized fluctuation corresponding to a queue overflow for the period f. Intuitively, since the random variables $Z_i(f)$ are all normalized, the value of $z_i(f)$ determines at least a first approximation to the improbability of the event $E_i(f)$. Indeed, under our assumption of Gaussian arrivals, $z_i(f)$ determines the improbability exactly:

$$\Pr(E_i(f)) = \overline{\Phi}(z_i(f)), \tag{14}$$

where

$$\overline{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-z^2/2} dz. \tag{15}$$

One consequence of these considerations is a formula for the "most relevant" period of formation of queues larger than B. Thus, since $z_i(f)$ gives the improbability of the event $E_i(f)$, the event with the greatest probability is the one with the smallest value of $z_i(f)$. In other words, the task is to find f to minimize $z_i(f)$. According to the self-similarity assumptions, $V_i(f) = {\sigma_i}^2 f^{2H_i}$, so $D_i(f) = \sigma_i f^{H_i}$, and

$$z_i(f) = \frac{B + cf}{\sigma_i f^{H_i}}. \tag{16}$$

So long as $0 < H_i < 1$, $z_i(f) \to \infty$ as $f \to 0$ and as $f \to \infty$, so any minima of z_i would appear as solutions of

$$0 = {z_i}'(f) = f^{-1-H_i}\left[-H_i \frac{B}{\sigma_i} + f \cdot (1 - H_i)\frac{c}{\sigma_i}\right]. \tag{17}$$

Note that this equation for f has only the one solution

$$t_{qi} = \frac{B}{c}\frac{H_i}{1-H_i} = \frac{B}{C-m}\frac{H_i}{1-H_i} = \frac{B}{C\cdot(1-\rho)}\frac{H_i}{1-H_i}, \tag{18}$$

where ρ $(= m/C)$ is the utilization factor. It follows that this value t_{qi} is the "most relevant" period of formation for

[4] Our later numerical experiments will check the validity of this simplification, e.g., Figure 5.

queues larger than B. Note that t_{qi} increases with H_i, and $t_{qi} = B/(C-m)$ for $H_i = 0.5$.

This formula (18) can be interpreted intuitively, as follows. We can imagine the queue as currently having size B and wonder how long it will take for the server (e.g., link) with capacity C to drain the queue. Without arrivals continuing to show up, the time required would be B/C. In reality, however, traffic continues to arrive. On average, traffic arrives at rate m, which suggests that the drain time would be about $B/(C-m)$. This occasion, however, on which the queue has reached the large value of B, is not average. Given the correlations of arrivals, and the fact that a current large value for Q implies an excess of arrivals in the recent past, one should expect the excess to persist in the near future, at least for $H > 0.5$. Thus, the drain time should increase with H, and the formula (18) may be understood as expressing this increase quantitatively. (Although the formula estimates a likely time of formation of a large queue, while the intuition deals with a likely time of erosion. Intuition also suggests that these times should be close to each other.) Figure 1 shows the $z_i(f)$ versus time for two cases: $H = 0.7$ and $H = 0.9$ ($B = 1,000$, $C = 625$, $m = 62.5$, and $a = 6.25$) with the dotted lines represent the relevant queuing time t_{qi} (Equation (18)), i.e., the time scale when $z_i(f)$ is minimized.

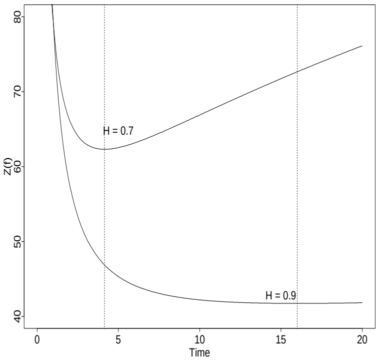

Figure 1: $z_i(f)$ vs. time

This "most relevant" period t_{qi} corresponds to an "easiest" damaging (normalized) fluctuation $z_i(t_{qi})$. Thus,

$$z_i(t_{qi}) = \frac{B \ + \ ct_{qi}}{\sigma_i {t_{qi}}^{H_i}} \quad = \quad \frac{B/(1-H_i)}{\sigma_i \left(\frac{B}{c}\frac{H_i}{1-H_i}\right)^{H_i}}$$
$$= \quad \frac{B^{1-H_i}c^{H_i}}{\sigma_i {H_i}^{H_i} \cdot (1-H_i)^{1-H_i}}. \quad (19)$$

This value determines the probability of the most likely of the events $E_i(f)$ through $\Pr(E_i(t_{qi})) = \overline{\Phi}(z_i(t_{qi}))$.

Figures 2 and 3 show $z_i(t_{qi})$ versus H for two different sets of parameters. The first set of parameters is intended to model the ATM environment with $C = 365,000$ cells/second (roughly an OC-3c link), while the second set of parameters is intended to model the Frame Relay environment with $C = 625$ frames/second (with an average frame size of 300 bytes, this roughly models a DS-1 link). In each case, the mean rate m is 10% of the capacity and the peakedness a ($\sigma = \sqrt{am}$) is 10% of the mean rate, when the peakedness is evaluated at the same time scale of 1 second as the time unit employed in reporting rates. Figure 2 corresponds to the ATM environment and Figure 3 corresponds to the Frame Relay environment. The dotted lines denote where $z_i(t_{qi})$ is maximized (or the best performance), which is around 0.88 in Figure 2 and 0.53 in Figure 3. We see that for the most part, $z_i(t_{qi})$ increases with rising Hurst values in the ATM environment while the other way is true in the Frame Relay environment.

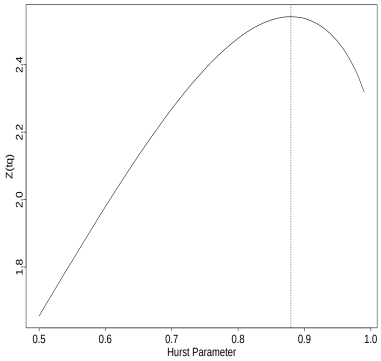

Figure 2: $z_i(t_{qi})$ vs. H for the ATM setting

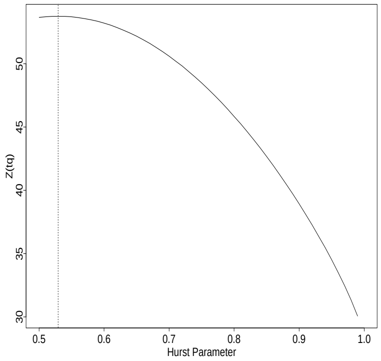

Figure 3: $z_i(t_{qi})$ vs. H for the Frame Relay setting

Strictly speaking, our interest should be in the probability of the union: $\Pr(Q_i > B) = \Pr(\bigcup_{f>0} E_i(f))$. Instead, we will pay attention simply to $z_i(t_{qi})$, with the following excuses. When $f \neq t_{qi}$, $z_i(f) > z_i(t_{qi})$, and because $\overline{\Phi}(z)$ decays so swiftly with z (roughly as $e^{-z^2/2}$), it follows correspondingly that $\Pr(E_i(f))$ is much smaller than $\Pr(E_i(t_{qi}))$. This suggests that the probability of the union really should be close to the largest individual probability $\Pr(E_i(t_{qi}))$. Indeed, a version of this claim can be established rigorously in large-deviations calculations of the asymptotic behavior of $\Pr(Q_i > B)$ [5]. The point is that the error should not be horribly large if one replaces the union $\bigcup_{f>0} E_i(f)$ by the single event $E_i(t_{qi})$. Since $z_i(t_{qi})$ does determine the improbability of this single event, it follows that it can also be used as an indicator of the improbability of the union. Thus, the question, of which process leads to a greater probability $\Pr(Q_i > B)$, is practically the same as the question, of which process has a smaller value for $z_i(t_{qi})$. This will be discussed further in Section 4.

3.2 Relevant Arrival Time Scales

Self-similar arrival processes, or their distributions, need more than just their Hurst parameters to be specified. The Hurst parameter merely determines the ratios

$$\frac{V_i(\alpha \cdot t)}{V_i(t)} = \alpha^{2H_i} \tag{20}$$

(self similarity yielding the fact that these ratios are independent of t); the Hurst parameter tells nothing of the magnitude of an individual variance. For a process with Hurst parameter H_i and finite second moments, the variance function must have the form

$$V_i(t) = {\sigma_i}^2 t^{2H_i} \tag{21}$$

for some $\sigma_i \geq 0$. For any two processes with $H_1 < H_2$, it follows that the variance function of the first process approaches infinity more slowly than the second. Indeed, it follows that there is a unique time t_m for which the variances match, as follows. Thus, $V_1(t_m) = V_2(t_m)$ means that

$${\sigma_1}^2 {t_m}^{2H_1} = {\sigma_2}^2 {t_m}^{2H_2}. \tag{22}$$

Taking square roots, and then rearranging the equation with powers of t on the left, one finds

$${t_m}^{H_2 - H_1} = \frac{\sigma_1}{\sigma_2}. \tag{23}$$

Solving this equation for t_m yields

$$t_m = \left(\frac{\sigma_1}{\sigma_2}\right)^{1/(H_2 - H_1)} \tag{24}$$

as the one time for which the variances match: $V_1(t_m) = V_2(t_m)$. For pairs of more realistic processes that are merely asymptotically self-similar, the variance-matching time need not be unique, or even exist. Yet, because exactly self-similar models are often used to estimate performance features of these processes, our interest here is confined to the case of exact self similarity in which there is a unique variance-matching time.

Figure 4 shows the standard deviation of the number of arrivals $(\sqrt{V(t)})$ versus the time scale t with $\sigma_1/\sigma_2 = 1/2$. The dotted line denotes t_m (≈ 0.177 in the example), and we can clearly see that when the time scale is greater than t_m, the arrival process with the higher Hurst value ($= 0.9$) has higher variance. On the other hand, the arrival process with the lower Hurst value ($= 0.7$) has higher variance when the time scale is less than t_m.

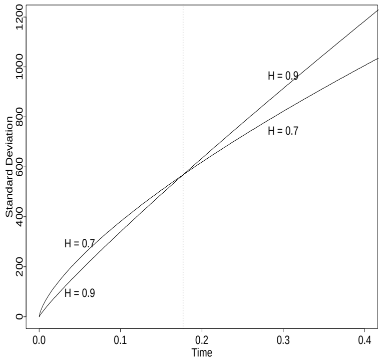

Figure 4: Standard deviation vs. time

A comment should be made on our procedure of introducing an explicit matching time t_m, instead of assuming that, in the different cases when H varied, all other parameters would remain constant. It is true that in abstract studies of the effect of different parameters, it is easiest to fix all but the varied parameter. In particular, when specifying an arrival process by the three parameters m, H, and σ^2, it would be natural, in an abstract study of cases with different values of H, to keep m and σ^2 fixed. It is also true that whenever there are scaling laws, the report of an analysis with one convention of units can be translated and usually generalized to other scales or units. Here, however, one of the crucial scaling laws of interest varies in the two cases being compared, so the translation is not particularly direct. Indeed, if the cases meet the condition of identical parameters with respect to one choice of time unit, they will not meet this condition for any other choice of unit. For example, consider how to translate, into a time unit of milliseconds, a report on two cases that is originally given with respect to a time unit of seconds. Let primes refer to milliseconds, so the lack of primes will refer to seconds. Thus, a time of τ milliseconds is the same as $t = \tau/1000$ seconds, so

$$A_i'(\xi,\eta) = A_i(\frac{\xi}{1000}, \frac{\eta}{1000}). \qquad (25)$$

Taking means and variances, it follows that

$$m_i' = \mathrm{E}[A_i'(0,1)] = \mathrm{E}[A_i(0, \frac{1}{1000})] = \frac{m_i}{1000}, \qquad (26)$$

and

$$V_i'(t) = \mathrm{Var}[A_i'(0,t)] = \mathrm{Var}[A_i(0, \frac{t}{1000})] = V_i(t)1000^{-2H_i}. \qquad (27)$$

From $V(t) = \sigma^2 t^{2H}$, it now follows that $H_i' = H_i$ and

$$\sigma_i'^2 = \sigma_i^2 1000^{-2H_i}. \qquad (28)$$

In particular, if $\sigma_1 = \sigma_2$, then $\sigma_1' \neq \sigma_2'$. The point is that a requirement for σ_1 to match σ_2 is not a requirement on the processes, but on the time unit with respect to which the processes are characterized. Naturally, we do not want our results to be limited to a particular time unit of choice, and that is why we do not insist on matching all non-Hurst parameters and why we introduce explicitly the matching time t_m.

4 Are Higher Hurst Values Good or Bad?

Once the matching time t_m (derived in the previous section) has been identified, comparisons of variances or deviations at other time scales are determined simply by whether the other time scale is bigger or smaller than t_m. First, recalling that $D_i(f) = \sqrt{V_i(f)}$, it follows that $D_i(\alpha \cdot f)/D_i(f) = \alpha^{H_i}$. Let

$$d = D_1(t_m) = D_2(t_m). \qquad (29)$$

Hence, $D_i(f)/d = (f/t_m)^{H_i}$. Now for any ratio $r > 1$, $r^{H_2} > r^{H_1}$, while $r^{H_2} < r^{H_1}$ for $r < 1$. Therefore,

$$D_2(f) > D_1(f) \quad \text{for } f > t_m, \qquad (30)$$

and

$$D_2(f) < D_1(f) \quad \text{for } f < t_m. \qquad (31)$$

In other words, the high-Hurst process is smoother at small time scales with smaller variances and deviations, but at larger time scales, the stronger correlations of the high-Hurst process produces larger fluctuations. So if one is concerned with a particular time scale, the process with the larger fluctuations at this scale will depend on whether the time scale is larger than t_m.

In particular, these comparisons translate into corresponding comparisons of the queuing indicators $z_i(t_{qi})$, assuming clear comparisons are available of the queuing times t_{qi} with the matching time t_m. The keys to this translation are the general formula $z_i(f) = (B + cf)/D_i(f)$ and the defining property of t_{qi} as the value that minimizes z_i. First, if $t_{q1} \geq t_m$, then $D_1(t_{q1}) \leq D_2(t_{q1})$, so

$$z_1(t_{q1}) = \frac{B + ct_{q1}}{D_1(t_{q1})} \geq \frac{B + ct_{q1}}{D_2(t_{q1})} = z_2(t_{q1}) \geq z_2(t_{q2}). \qquad (32)$$

Given our use of $z_i(t_{qi})$ as the indicator of service quality, this comparison says that the low-Hurst process has better queuing performance in this case of $t_{q1} \geq t_m$. Next, if $t_{q2} \leq t_m$, then $D_2(t_{q2}) \leq D_1(t_{q2})$, so

$$z_2(t_{q2}) = \frac{B + ct_{q2}}{D_2(t_{q2})} \geq \frac{B + ct_{q2}}{D_1(t_{q2})} = z_1(t_{q2}) \geq z_1(t_{q1}). \qquad (33)$$

This comparison says that the high-Hurst process has better queuing performance in this case of $t_{q2} \leq t_m$.

These two comparisons (in the cases t_{q1} high and t_{q2} low) do not deal with a clean partition of all cases. First, the formula

$$t_{qi} = \frac{B}{c}\frac{H_i}{1 - H_i} \qquad (34)$$

shows that t_{qi} increases with H_i, so $t_{q2} > t_{q1}$. Hence, the two cases above, defined by $t_{q1} \geq t_m$ and by $t_{q2} \leq t_m$, cannot occur simultaneously. Finally, in the remaining untreated case of $t_{q1} < t_m < t_{q2}$, except for the direct calculation of the $z_i(t_{qi})$, there is no simple way to determine which queue performs better. Now from the treated cases, it is clear that among intermediate values of H (while keeping t_m fixed), the best queuing behavior occurs (in our sense of a large value for $z_H(t_{qH})$) when H satisfies $t_m = (B/c)H/(1 - H)$. Unfortunately, this observation merely implies that processes 1 and 2 both perform worse than the intermediate process, so it cannot be used to compare the given processes 1 and 2.

Figures 5 and 6 show the $z_i(t_{qi})$ versus buffer size for the ATM and the Frame Relay settings much like Figures 2 and 3. Again, in each case, the mean rate is 10% of the capacity, two traffic streams with Hurst parameters of 0.7 and 0.9 are input to the queue, and the ratio of the standard deviation of the two streams is 1/2 at the time scale of one second. In Figure 5, in addition to $z_i(t_{qi})$, we also plot the asymptotic results of $\Pr(Q_i > B)$ recently suggested by Narayan [19]. The two curves below the 0 on the y-axis depict the the cell-loss ratio based on Narayan's results on a $\log_{10}$ scale. We can see that $z_i(t_{qi})$ is a good surrogate for comparing the cell loss ratio performance. Furthermore, in Figure 5, we can clearly identify the three regions as discussed. In Region I (where $t_m \geq t_{q2} > t_{q1}$), the stream with higher Hurst value receives better performance; in Region III (where $t_{q2} > t_{q1} \geq t_m$), the stream with lower Hurst value receives better performance; and in Region II, the cross-over occurs.

In Figure 6, only Region III exists, that is, with the parameter settings we have for the Frame Relay case,[5] the lower-Hurst-parameter stream always performs better. This may be one of the reasons that creates the common misconception of higher Hurst values always incurring worse queuing performance since it is relatively unlikely for the cross-over scenarios to occur in lower-speed networks.

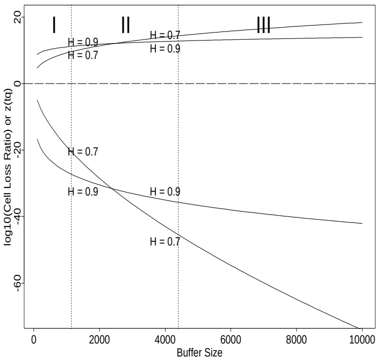

Figure 5: $\log_{10}(\Pr(Q_i > B))$ or $z_i(t_{qi})$ vs. B for the ATM setting

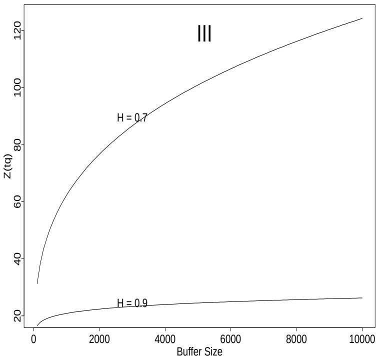

Figure 6: $z_i(t_{qi})$ vs. B for the Frame Relay setting

Figures 7 and 8 further demonstrate that the performance comparison between high-Hurst and low-Hurst processes depends on a combination of traffic and system parameters, and not just the Hurst values. Figure 7 plots $z_i(t_{qi})$ vs. server capacity with various parameter settings. Figure 7(a) is the base case with $B = 1,000$, $H_1 = 0.7$, $H_2 = 0.9$, $\rho = 0.1$, and the ratio of the peakedness of process 2 to process 1 is 4 at the time scale of one second. We see that Region I occurs when the capacity is greater than 3.2×10^5 cells/second and Region III occurs when the capacity is less than about 82,962 cells/second. Figure 7(b) is the same as the base case (a) except that the buffer size is increased to 5,000. We see that only Region III exists with the range of capacity values we choose. The boundaries between Regions II and III and Regions I and II are around $C \sim 414,814$ and $C \sim 1.6 \times 10^6$, respectively. Figure 7(c) is the same as the base case (a) except that the utilization is increased to 0.5. We see that Region II and III exist with the range of capacity values we choose. The boundary between Regions I and II is around $C \sim 5.76 \times 10^5$. Finally, Figure 7(d) is the same as the base case (a) except that the ratio of the peakedness of process 2 to process 1 is now 1/4 as opposed to 4, that is, the higher Hurst process has lower peakedness at the unit of time. In this case, only Region I exists with the range of capacity values we choose. The boundary between Regions II and III (Regions I and II) occurs around $C \sim 81$ ($C \sim 312$).

Figure 8 plots $z_i(t_{qi})$ vs. utilization with various parameter settings. Figure 8(a) is the base case with $B = 1,000$, $H_1 = 0.7$, $H_2 = 0.9$, $C = 365,000$ (e.g., an ATM link), and the ratio of the peakedness of process 2 to process 1 is 4 at the 1-second time scale. We see that Region I occurs when the utilization is less than 0.21 and Region III occurs when the utilization is greater than about 0.80. Figure 8(b) is the same as the base case (a) except that the buffer size is increased to 5,000. We see that only Region III exists. In this case, the boundary between Regions II and III as well as the boundary between Regions I and II occur at utilization

[5] Actually, the boundary between Regions I and II occurs at buffer size of roughly 1, while the boundary between Regions II and III occurs at buffer size of roughly 8. It is highly unlikely that existing Frame Relay equipment has such small buffer sizes.

less than 0, which are not feasible. Figure 8(c) is the same as the base case (a) except that the capacity is reduced to 625 (e.g., a Frame Relay link). Again, only Region III is feasible. Finally, Figure 8(d) is the same as the base case (a) except that the ratio of the peakedness of process 2 to process 1 is now 1/4 as opposed to 4, that is, the higher Hurst process has lower peakedness at the unit of time. In this case, only Region I exists with the range of utilization values we choose. In this case, Regions II and III only exist when the utilization is very close to 1. The boundary between Regions II and III is around $\rho \sim 0.9998$, while the boundary between Regions I and II is around $\rho \sim 0.9992$.

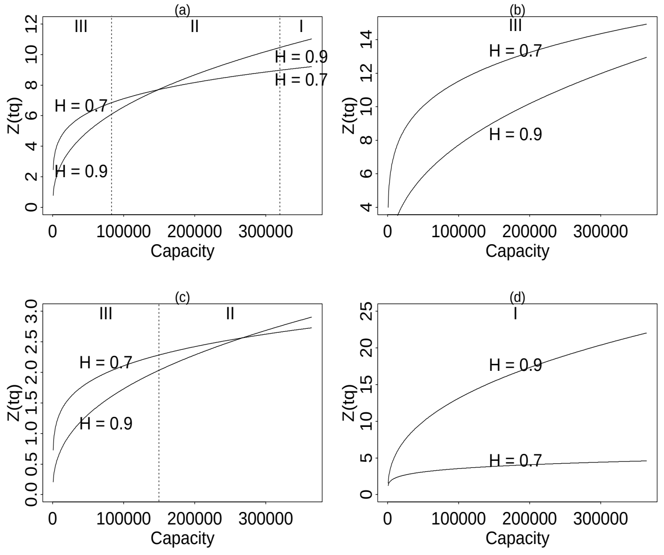

Figure 7: $z_i(t_{qi})$ vs. capacity

There is a sense in which the untreated case (Region II) (of $t_{q1} < t_m < t_{q2}$) should be expected to be rare. Specifically, the buffer sizes contemplated for different systems vary by more than an order of magnitude, with some buffers designed to hold less than 10 ms of work, while others are designed to hold almost a second. In contrast, the range of Hurst parameters commonly contemplated, often from $H = 0.5$ to $H = 0.9$, corresponds to values for the expression $H/(1-H)$ ranging from 1 to 9, all within single order of magnitude. No matter what scale t_m turns out to be, the wide possibilities for B/c imply that it will be quite rare for t_{q1} and t_{q2} to turn out to be on opposite sides of any given value of interest. In Figure 5 (Figure 6), only buffers with sizes between roughly 1140 and 4400 (1 and 8) fall within this untreated case.

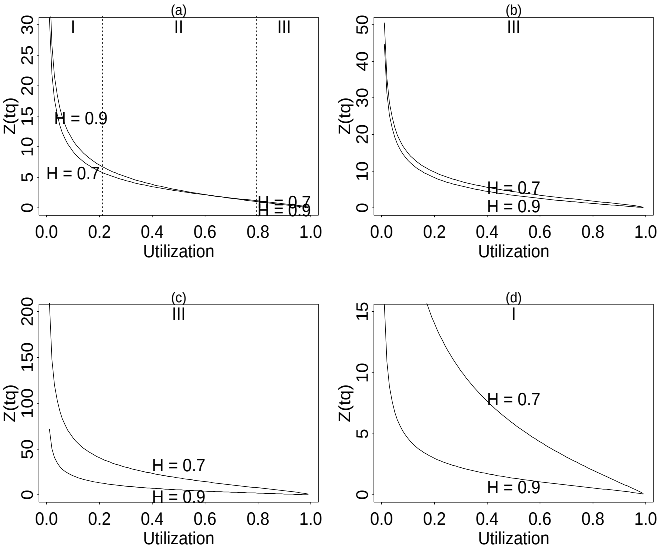

Figure 8: $z_i(t_{qi})$ vs. utilization

5 Conclusion

Earlier performance analyses of self-similar traffic have noted how the presence of long-range dependence, or a high value of the Hurst parameter H, is often associated with surprisingly large queue sizes. As a result, a common impression has been created of expecting queue performance to be worse as H rises, but this impression can be misleading. In fact, the queuing performance depends on a combination of various traffic and system parameters.

In this paper, we have shown and derived the concept of relevant time scales both in arrival as well as queuing and its application in performance comparison between self-similar traffic streams. First, in comparing a high-H process with a low-H process, there is a unique time scale t_m at which the variances of the two processes match. Second, there are time scales t_{qi} that are most relevant for queuing the arrivals of process i, usually identifiable as $(B/(C-m))H_i/(1-H_i)$ and both close in order of magnitude to $B/(C-m)$.

We have derived the conditions under which the performance is better based on these identifiable time scales: if the queuing scales t_{qi} both exceed the variance-matching scale t_m, then the high-H queue is worse; if the queuing scales are smaller, then the low-H queue is worse. No firm prediction can be made in the remaining case of t_m falling between the two queuing scales.

References

[1] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger, "Long-Range Dependence in Variable-Bit-Rate Video Traffic," *IEEE Trans. Commun.*, Vol. 43, No. 2/3/4, pp. 1566-1579, 1995.

[2] D.R. Cox, "Long-Range Dependence: A Review", in: *Statistics: An Appraisal*, H. A. David and H. T. David (Eds.), The Iowa State University Press, Ames, Iowa, 55-74, 1984.

[3] M.E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *Proc. ACM Sigmetrics*, pp. 160-169, May 1996.

[4] D.E. Duffy, A.A. McIntosh, M. Rosenstein and W. Willinger, "Statistical Analysis of CCSN/SS7 Traffic Data from Working Subnetworks," *IEEE JSAC*, Vol. 12, No. 3, pp. 544-551, 1994.

[5] N.G. Duffield and N. O'Connell, "Large Deviation and Overflow Probabilities for the General Single-Server Queue, with Applications", *Math. Proc. Cam. Phil. Soc.*, Vol. 118, 363-374, 1995.

[6] A. Erramilli, O. Narayan and W. Willinger, "Experimental Queuing Analysis with Long-Range Dependent Packet Traffic," *IEEE/ACM Trans. on Networking*, Vol. 4, No. 2, pp. 209-223, April 1996.

[7] A. Erramilli and A. Neidhardt, "The Role of Shaping and Policing in ATM Networks," *Proc. ITC Specialist Seminar*, Lund, Sweden, 1996.

[8] A. Erramilli and J.L. Wang, "Monitoring Packet Traffic Levels," *Proc. IEEE Globecom*, pp. 274-280, San Francisco, CA, 1994.

[9] A. Erramilli, W. Willinger and J.L. Wang, "Modeling and Management of Self-Similar Traffic Flows in High-Speed Networks," to appear *State-of-the-Art in Performance Modeling and Simulation*, K. Bagchi (ed.), Gordon and Breach.

[10] J.L. Jerkins, M. Pucci, J.L. Wang and J. Monroe, "Carrying Internet Traffic Over Frame Relay Links: Traffic Analysis and Characterization," preprint 1997.

[11] J.L. Jerkins and J.L. Wang, "A Measurement Analysis of ATM Cell-Level Aggregate Traffic," *Proc. IEEE Globecom*, pp. 1589-1595, Phoenix, AZ, 1997.

[12] J.L. Jerkins and J.L. Wang, "Traffic Analysis and Engineering for CCS Links Carrying 800 or AIN Service," *Proc. ISCOM*, pp. 15-19, Hsinchu, Taiwan, 1997.

[13] J.L. Jerkins and J.L. Wang, "Establishing Broadband Application Signatures through ATM Network Traffic Measurement Analyses," to appear *IEEE ICC*, 1998.

[14] K.R. Krishnan, "A New Class of Performance Results for a Fractional Brownian Traffic Model," *Queueing Systems*, Vol. 22, pp. 277-285, 1996.

[15] K.R. Krishnan, A.L. Neidhardt and A. Erramilli, "Scaling Analysis in Traffic Management of Self-Similar Processes," *Proc. 15th ITC*, Washington, D.C., pp. 1087-1096, June 1997.

[16] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. on Networking*, Vol. 2, No. 1, pp. 1-15, 1994.

[17] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements," *Statistical Science*, Vol. 10, pp. 67-85, 1995.

[18] K. Meier-Hellstern, P.E. Wirth, Y.-L. Yan and D.A. Hoeflin, "Traffic Models for ISDN Data Users: Office Automation Application," *Proc. 13th ITC*, Copenhagen, pp. 167-172, 1991.

[19] O. Narayan, "New Results in Fractional Brownian Storage", presented at the Workshop on *Traffic Characterization and Performance Analysis for Modern High-Speed Networks: New Developments in Self-Similar Performance Modeling*, Tsukuba, Japan, November 1997.

[20] I. Norros, "A Storage Model with Self-Similar Input," *Queueing Systems*, Vol. 16, pp. 387-396, 1994.

[21] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *Proc. ACM Sigcomm*, pp. 257-268, London, UK, 1994.