To appear in Proc. IEEE INFOCOM'00

Multiscale Queuing Analysis of Long-Range-Dependent Network Traffic

Vinay J. Ribeiro, Rudolf H. Riedi, Matthew S. Crouse, and Richard G. Baraniuk

Department of Electrical and Computer Engineering
Rice University
6100 South Main Street
Houston, TX 77005, USA

Abstract—**Many studies have indicated the importance of capturing scaling properties when modeling traffic loads; however, the influence of long-range dependence (LRD) and marginal statistics still remains on unsure footing. In this paper, we study these two issues by introducing a multiscale traffic model and a novel multiscale approach to queuing analysis. The multifractal wavelet model (MWM) is a multiplicative, wavelet-based model that captures the positivity, LRD, and "spikiness" of non-Gaussian traffic. Using a binary tree, the model synthesizes an N-point data set with only $O(N)$ computations.**

Leveraging the tree structure of the model, we derive a *multiscale queuing analysis* that provides a simple closed form approximation to the tail queue probability, valid for any given buffer size. The analysis is applicable not only to the MWM but to tree-based models in general, including fractional Gaussian noise. Simulated queuing experiments demonstrate the accuracy of the MWM for matching real data traces and the precision of our theoretical queuing formula. Thus, the MWM is useful not only for fast synthesis of data for simulation purposes but also for applications requiring accurate queuing formulas such as call admission control. Our results clearly indicate that the marginal distribution of traffic at different time-resolutions affects queuing and that a Gaussian assumption can lead to over-optimistic predictions of tail queue probability even when taking LRD into account.

I. INTRODUCTION

Traffic models play a significant rôle in the analysis and characterization of network traffic and network performance. Accurate models capture important characteristics of traffic and enhance our understanding of these complicated signals and systems by allowing us to study the effect of various model parameters on network performance through both analysis and simulation.

One key property of modern network traffic is the presence of *long-range dependence* (LRD) which was demonstrated convincingly in the landmark paper of Leland et. al. [1]. There, measurements of traffic load on an Ethernet were attributed to *fractal* behavior or *self-similarity*, i.e., to the fact that the data "looked statistically similar" (highly variable) on all time-scales. These features are inadequately described by classical traffic models such as Markov or Poisson models. In particular, the LRD of data traffic can lead to higher packet losses than that predicted by classical queuing analysis [1, 2].

These findings were immediately followed by the development of new fractal traffic models [3–5]. The most broadly applied fractal model is the *fractional Brownian motion* (fBm), $B(t)$, whose discrete increment process $G(k) := B((k+1)\Delta) - B(k\Delta)$, called *fractional Gaussian noise* (fGn), has an autocorrelation of the form

$$r_G[k] = \frac{\sigma^2}{2}|\Delta|^{2H}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \qquad (1)$$

with Δ a constant. Gaussianity and the strong scaling of fBm enable rigorous analytical studies of queuing behavior [6–10], thus increasing the popularity of the fBm/fGn models.

Though fGn is an appropriate traffic model in some cases [11, 12], it can only model real-world traces with the rigid, restrictive correlation structure (1). Indeed, convincing evidence has been produced establishing the importance of short-term correlations for buffering [13–15], and so-called relevant time scales have been discovered [14, 16, 17].

Generalizations of fBm/fGn with a more flexible correlation structure than (1) can be synthesized almost effortlessly using the powerful decorrelating capability of the *wavelet transform* [18–21]. There, independent Gaussian wavelet coefficients with variance decaying appropriately with scale form the building blocks for modeling both the long and short-term correlations of a target data set. Efficient $O(N)$ algorithms based on the tree structure of wavelet coefficients are available to synthesize N-point data sets [22, 23]. We will term all such models *wavelet-domain independent Gaussian* (WIG) models.

As a consequence of their Gaussian nature, the fBm/fGn/WIG models can produce unrealistic synthetic traffic traces in certain situations. In many networking applications, for instance, we are nowhere near the Gaussian limit, in particular on small time scales. Indeed, various authors have observed heavy-tailed

This work was supported by the NSF, grant no. CCR-9973188, by ONR, grant no. N00014-99-10813, by DARPA/AFOSR, grant no. F49620-97-1-0513, and by Texas Instruments. Email: {vinay, riedi, mcrouse, richb}@rice.edu. URL: www.dsp.rice.edu.

marginals in traffic [24, p. 364], [25]. More practically speaking, when the standard deviation of the data approaches or exceeds the mean, considerable portions of the fBm/fGn/WIG synthesis are negative (see Figure 1(a) and (b)).

Unlike the WIG model, the *multifractal wavelet model* (MWM), based on a multiplicative cascade in the wavelet domain, guarantees a positive output [27]. In its simplest form, the MWM is closely related to the wavelet-based construction of fBm/fGn, having the same short list of parameters (mean, variance, H). However, the MWM framework boasts the flexibility, if desired, to additionally match the short-term correlations like the WIG model. The superiority of the MWM at matching the qualitative visual appearance (see Figure 1(c)), the marginals (see Figure 4(c)), and the queuing behavior (see Figure 5) [27, 28] suggests that the multiplicative MWM approach is more appropriate than the additive Gaussian one.

The main contribution of this paper is a novel multiscale queuing analysis. For an infinite-length buffer with constant link capacity c, the queue length (assuming the queue was empty some time in the past) is given by

$$Q = \sup_r (K_r - rc), \tag{2}$$

where K_r is the total traffic that entered the queue in the past r instants. In other words, the queue size (2) is a supremum function of the traffic arrivals aggregated at multiple time-scales. For the WIG and MWM models, aggregates at dyadic time scales (i.e., K_r for $r = 2^m,\ m \in \mathbb{N}$) have simple expressions, and are related to each other by independent innovations. We exploit this fact to derive an approximation to the tail queue probability. The resulting multiscale queuing formula, which we call MSQ(b),

- is valid for any given queue length b,
- closely approximates the tail queue probability as experiments verify,
- requires statistics of traffic at only a few dyadic time-scales,
- is easy-to-use, and
- clearly demonstrates the importance of matching multiscale marginals (especially the tail of the marginal) in addition to the variance at different time-scales (i.e., the correlation structure), for accurate predictions of queuing behavior.

As a consequence, the MWM becomes viable for applications requiring models with accurate queuing formulas such as call admission control.

After introducing wavelets and explaining the WIG model in Section II, we describe the MWM and demonstrate its superiority over the WIG in capturing the marginals of traffic in Section III. After experimentally proving the importance of the non-Gaussian nature of traffic on queuing, we introduce the novel multiscale queuing analysis, which we apply to the WIG and MWM in Section IV. We provide empirical evidence for the accuracy of our theoretical queuing formulas in Section IV, use our queuing formulas to explain why marginals and LRD affect queuing in Section V and conclude in Section VI.

II. Classical Wavelet Models for LRD Processes

A. Long-range dependence

Consider a discrete-time, wide-sense stationary random process $\{X_t,\quad t \in \mathbb{Z}\}$ with auto-covariance function $r_X[k] = \mathrm{cov}(X_t, X_{t+k})$. A change in time scale can be represented by forming the aggregate process $X_t^{(m)}$, which is obtained by averaging X_t over non-overlapping blocks of length m and replacing each block by its mean

$$X_t^{(m)} \;=\; \frac{X_{tm-m+1} + \cdots + X_{tm}}{m}. \tag{3}$$

Denote the auto-covariance of $X_t^{(m)}$ by $r_X^{(m)}[k]$. The process X is said to exhibit LRD if its auto-covariance decays slowly enough to render $\sum_{k=-\infty}^{\infty} r_X[k]$ infinite [29]. Equivalently, $m\, r_X^{(m)}[0] \;\to\; \infty$ as $m \;\to\; \infty$, and the power spectrum $S_X(f)$ is singular near $f \;=\; 0$.

One example of an LRD process is fGn (1), whose LRD is captured by the Hurst parameter H (larger $H \;\Rightarrow$ stronger correlation or LRD). To estimate H by the *variance-time plot* method, we fit a straight line through the plot of an estimate of $\log \mathrm{var}(X^{(m)})$ against $\log m$. More reliable estimators of H have been devised [30], in particular an unbiased one based on wavelets [31, 32].

B. Wavelet transform

The discrete wavelet transform provides a multiscale signal representation of a one-dimensional random signal $C(t)$ in terms of shifted and dilated versions of a prototype bandpass wavelet function $\psi(t)$ and shifted versions of a lowpass scaling function $\phi(t)$ [33, 34]. For special choices of the wavelet and scaling functions, the atoms

$$\psi_{j,k}(t) := 2^{j/2}\,\psi\big(2^j t - k\big)\,,\ \phi_{j,k}(t) := 2^{j/2}\,\phi\big(2^j t - k\big)\,, \tag{4}$$

$j, k \;\in\; \mathbb{Z}$, form an orthonormal basis, and we have the signal representation [34]

$$C(t) \;=\; \sum_k U_{J_0,k}\,\phi_{J_0,k}(t) \;+\; \sum_{j=J_0}^{\infty} \sum_k W_{j,k}\,\psi_{j,k}(t). \tag{5}$$

Here the wavelet coefficients $W_{j,k}$ and the scaling coefficients $U_{j,k}$ are given by

$$W_{j,k} := \int C(t)\,\psi_{j,k}(t)\,\mathrm{d}t,\ U_{j,k} := \int C(t)\,\phi_{j,k}(t)\,\mathrm{d}t. \tag{6}$$

Without loss of generality, we will assume $J_0 = 0$.

In this representation, k indexes the spatial location of analysis and j indexes the *scale* or resolution of the wavelet analysis — larger j corresponds to higher resolution and $j = 0$ indicates the coarsest scale or lowest resolution of analysis. In practice, we work with a sampled or finite-resolution representation of $C(t)$, replacing the semi-infinite sum in (5) with a sum over a finite number of scales $0 \le j \le n-1,\ n \in \mathbb{N}$.

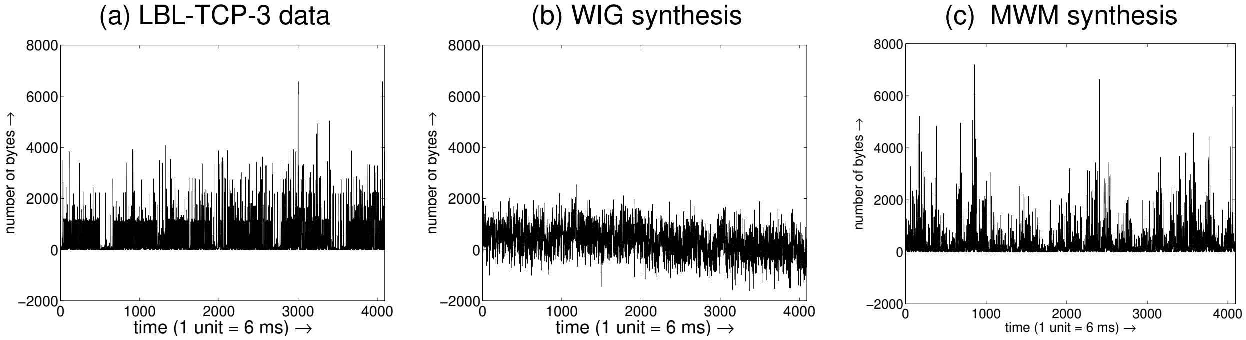


Fig. 1. Modeling bursty traffic data: *Bytes-per-time arrival process for (a) wide-area TCP traffic at the Lawrence Berkeley Laboratory (trace LBL-TCP-3) [26], (b) one realization of the state-of-the-art wavelet-domain independent Gaussian (WIG) model [22], and (c) one realization of the multifractal wavelet model (MWM) synthesis. The MWM trace closely resembles the real data, while the WIG trace does not.*

In this paper, we restrict our attention to the simplest wavelet system, that of *Haar*. The Haar scaling and wavelet functions are given by (see Figure 2(a))

$$\phi(t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & \text{else} \end{cases} \; ; \; \psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & \text{else.} \end{cases} \tag{7}$$

Since $\phi_{j,k}(t)$ is a rectangular function, the Haar scaling coefficients $U_{j,k}$ (6) represent the *local mean values* of the signal in the time intervals $[k2^{-j}, (k+1)2^{-j}]$ and thus form a discrete-time approximation of $C(t)$ at resolution j. By design, the supports of the $\phi_{j,k}(t)$ are nested within each other. This makes it natural to use a binary tree (see Figure 2(b)) to display the relationship between the coefficients $U_{j,k}$. Nodes at lower horizontal levels in the tree correspond to representations of the signal at finer resolutions.

The Haar wavelet transform of a signal can be computed recursively starting from its finest-scale scaling coefficients via [34]

$$\begin{aligned} U_{j-1,k} &= 2^{-1/2}(U_{j,2k} + U_{j,2k+1}), \\ W_{j-1,k} &= 2^{-1/2}(U_{j,2k} - U_{j,2k+1}). \end{aligned} \tag{8}$$

This corresponds to moving up the binary tree and storing in the Haar wavelet coefficients $W_{j,k}$ the detail information lost while going from fine to coarse resolutions (see Figure 2(b)).

The inverse Haar wavelet transform is computed via

$$\begin{aligned} U_{j,2k} &= 2^{-1/2}(U_{j-1,k} + W_{j-1,k}), \\ U_{j,2k+1} &= 2^{-1/2}(U_{j-1,k} - W_{j-1,k}) \end{aligned} \tag{9}$$

and is equivalent to moving down the scaling coefficient tree to finer representations of the signal (Figure 2(b)). It is easily seen that the forward and inverse Haar wavelet transforms of an N-point signal can be computed in $O(N)$ operations, using (8) and (9) respectively.

We introduce a new process $C^{(n)}[k]$, a discrete-time approximation to $C(t)$ defined by

$$C^{(n)}[k] \; := \; \int_{k2^{-n}}^{(k+1)2^{-n}} C(t)\, \mathrm{d}t. \tag{10}$$

For notational simplicity, we will assume that $C(t)$ lives on $[0,1]$ and that $C^{(n)}[k]$ is a length-2^n discrete-time signal. Thus, there is only one scaling coefficient $U_{0,0}$ in (5), that is, a single tree of scaling coefficients. (A more general case with multiple scaling coefficients at the coarsest scale is treated in [27].) We will focus on modeling the finest-scale scaling coefficients:

$$C^{(n)}[k] \;\; = \;\; 2^{-n/2} U_{n,k}, \;\; k = 0, 1, \ldots, 2^n - 1. \tag{11}$$

C. Wavelet-domain independent Gaussian (WIG) model

Wavelets serve as an approximate Karhunen-Loève or decorrelating transform for fBm [18], fGn, and more general LRD signals [23]. Hence, the difficult task of modeling these highly correlated signals in the time domain reduces to a simple one of modeling them approximately by an uncorrelated process in the wavelet domain.

The WIG model synthesizes a Gaussian process capturing both the long and short-term correlations, by generating the parent node $U_{0,0}$ of the scaling coefficient tree as a Gaussian random variable and by generating the wavelet coefficients as independent (uncorrelated), zero-mean Gaussian random variables identically distributed within scale according to $W_{j,k} \sim N(0, \sigma_j^2)$, with σ_j^2 the required wavelet-coefficient variance at scale j [18–22]. For example, a power-law decay for the σ_j^2's leads to approximate wavelet synthesis of fBm or fGn [18, 20]. Scaling coefficients at finer scales on the tree are then recursively computed through (9) until the finest scale scaling coefficients $U_{n,k}$ and hence the required signal $C^{(n)}_{\text{WIG}}[k]$ are obtained. The result is a fast $O(N)$ algorithm for generating a length-N signal, characterized by approximately $\log_2(N)$ (the number of time scales) parameters (see Figure 2(c)).

The WIG is an *additive model*, because we can express the signal $C^{(n)}_{\text{WIG}}[k]$ directly as a sum of independent random variables. First, we need some notation. Each shift k at scale n has a unique binary representation $k = \sum_{i=0}^{n-1} k_i' 2^{n-1-i}$ where each $k_i' \in \{0,1\}$. Letting $k_n = k$ and $k_{i-1} = k_i$ div 2 we have $k_i = 2 \cdot k_{i-1} + k_{i-1}' = \sum_{j=0}^{i-1} k_j' 2^{i-1-j}$. The shifts k_i correspond

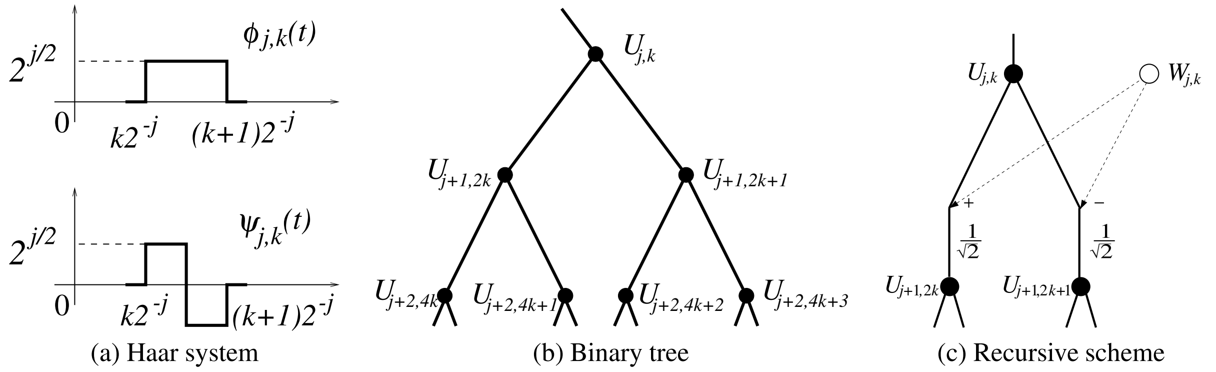


Fig. 2. The WIG model: *(a) Scaled and shifted Haar scaling and wavelet functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$. (b) Binary tree of scaling coefficients (local mean values of the signal). Nodes at each horizontal level in the tree provides a piecewise constant representation of the signal with lower levels corresponding to finer resolutions. (c) Recursive scheme for calculating the Haar scaling coefficients $U_{j+1,2k}$ and $U_{j+1,2k+1}$ at scale $j+1$ as sums and differences (normalized by $1/\sqrt{2}$) of the scaling and wavelet coefficients $U_{j,k}$ and $W_{j,k}$ at scale j. For the WIG model, the $W_{j,k}$'s are mutually independent and identically distributed within scale according to $W_{j,k} \sim N(0, \sigma_j^2)$.*

to the ancestors of k at scale i and so we can write

$$C_{\text{WIG}}^{(n)}[k] = 2^{-n} \left(U_{0,0} + \sum_{i=0}^{n-1} (-1)^{k'_i} 2^{i/2} W_{i,k_i} \right). \tag{12}$$

This result can be derived by iteratively applying (9).

The WIG model is Gaussian by construction, but network traffic signals (such as loads and interarrival times) can be highly "spiky" and non-Gaussian (recall from Figure 1). We seek a more accurate marginal characterization for these spiky, non-negative LRD processes, yet wish to retain the decorrelating properties of wavelets and the simplicity of the WIG model.

III. Multifractal Wavelet Model

A. Haar wavelet transform and positive signals

In order to model non-negative signals using the Haar wavelet transform, we must constrain the scaling and wavelet coefficient values to ensure that $C(t)$ in (5) is non-negative. While cumbersome for a general wavelet system,[1] these conditions are simple for the Haar system.

Since the Haar scaling coefficients $U_{j,k}$ represent the local mean values of the signal at different scales and shifts, they are non-negative if and only if the signal itself is non-negative, that is, $C(t) \geq 0 \Leftrightarrow U_{j,k} \geq 0, \ \forall\, j,k$. Combining (9) with the constraint $U_{j,k} \geq 0$, we obtain the condition

$$C(t) \geq 0 \ \Leftrightarrow\ |W_{j,k}| \leq U_{j,k}, \qquad \forall\, j,k. \tag{13}$$

B. MWM model

The positivity constraint (13) inspires a very simple multiscale, multiplicative signal model for positive processes. In the *multifractal wavelet model* (MWM) [27] we compute the wavelet coefficients recursively by

$$W_{j,k} = A_{j,k}\, U_{j,k}, \tag{14}$$

where the $A_{j,k}$'s are independent random variables supported on the interval $[-1,1]$.

The MWM synthesizes a data trace in a manner similar to the WIG. After generating the coarsest scale scaling coefficient $U_{0,0}$ and the multipliers $A_{j,k}$, the MWM generates scaling coefficients at finer scales of the scaling coefficient tree recursively using (9) and (14), that is (see Figure 3(a))

$$\begin{aligned} U_{j,2k} &= 2^{-1/2}(1 + A_{j+1,k})\, U_{j-1,k}, \\ U_{j,2k+1} &= 2^{-1/2}(1 - A_{j+1,k})\, U_{j-1,k}, \end{aligned} \tag{15}$$

until the finest scale has been reached.

The MWM is a *multiplicative model*, because we can express the signal $C_{\text{MWM}}^{(n)}[k]$ directly as a product (or cascade) of independent random multipliers $1 \pm A_{j,k}$. Using the notation introduced in Section II-C, we have

$$C_{\text{MWM}}^{(n)}[k] = 2^{-n}\, U_{0,0} \prod_{i=0}^{n-1} \left[1 + (-1)^{k'_i} A_{i,k_i} \right], \tag{16}$$

which should be compared with (12).

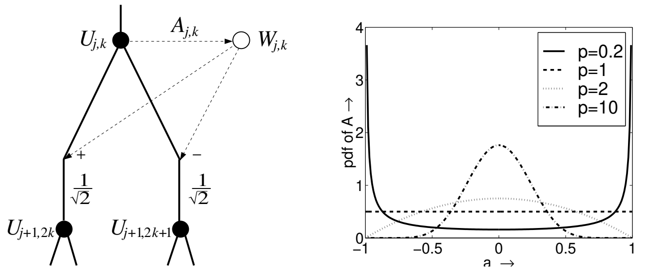
As a particular consequence of the multiplicative structure of (16), the process $C^{(n)}[k]$ will be positive, LRD (see Section III-C below), and have a "spiky" appearance. This matter is better explained in the framework of *multifractals*, which are beyond the scope of this paper (see [27, 28]).

It is easily shown that the total cost for computing N MWM signal samples is $O(N)$. In fact, synthesis of a trace of length 2^{18} data points takes just seconds of workstation cpu time. See [35] for a similar model to the MWM used as an intensity prior for wavelet-based image estimation.

We choose the multipliers $A_{j,k}$ to be symmetric about 0 and identically distributed within scale; it is easily shown that these two conditions are necessary for the $C_{\text{MWM}}^{(n)}$ process to be first-order stationary [27]. Due to its flexible shape, compact support and tractability to closed-form calculations, we choose the *symmetric beta distribution* [36],[2] $\beta_{-1,1}(p_j, p_j)$ (see Figure 3(b)) for

[1] The conditions are straightforward also for certain biorthogonal wavelet systems.

[2] We denote a beta random variable with support $[a,b]$ by $\beta_{a,b}$



(a) MWM construction (b) $\beta-$distribution

Fig. 3. The MWM model: *(a) Construction of multifractal wavelet model (MWM): At scale j, generate the multiplier $A_{j,k} \sim \beta_{-1,1}(p_j, p_j)$, and then form the wavelet coefficient as the product $W_{j,k} = A_{j,k}U_{j,k}$. At scale $j+1$, form the scaling coefficients in the same manner as the WIG model in Figure 2(c).*
(b) Probability density function of a $\beta_{-1,1}(p,p)$ random variable A. For $p = 0.2$, $\beta_{-1,1}(p,p)$ resembles a binomial distribution, and for $p = 1$ it has a uniform density. For $p > 1$ the density is close to a truncated Gaussian density with increasing resemblance as p increases.

the $A_{j,k}$'s, with p_j the beta parameter at scale j. In our simulation experiments we choose[3]

$$U_{0,0} \sim \beta_{0,M}(p_{-1}, p_{-1}) \tag{17}$$

with $M \geq 0$.

C. Correlation matching

Combining (14) and the fact that the variance of a random variable $A \sim \beta_{-1,1}(p,p)$ is given by var$[A] = 1/(2p+1)$, we obtain

$$\frac{\mathrm{var}(W_{j-1,k})}{\mathrm{var}(W_{j,k})} = \frac{2\,\mathrm{var}\,[A_{j-1,k}]}{\mathrm{var}\,[A_{j,k}]\ (1+\mathrm{var}\,[A_{j-1,k}])} = \frac{2p_j+1}{p_{j-1}+1}. \tag{18}$$

Thus, the parameters p_j control the wavelet-domain energy of the signal on all scales, hence also the variance on all time scales and in particular the LRD parameter H. Given training data, we set the p_j's using estimates of the variance of the trace's Haar wavelet coefficients at different scales using (18). With one parameter per wavelet scale, the MWM has approximately $\log_2 N$ parameters for a trace of length N. These could be reduced to, say, match only the variance *decay*, i.e., only the LRD parameter H. See Table I for a comparison of the WIG and MWM properties.

To complete the modeling, we must choose the parameters p_0, p_{-1}, and M of the model. Since

$$(2p_0+1)\mathrm{var}(W_{0,0}) = \mathbb{E}[U_{0,0}^2], \tag{19}$$

we calculate p_0 from estimates of $\mathbb{E}[U_{0,0}^2]$ and var$(W_{0,0})$. The parameters p_{-1} and M of $U_{0,0}$ are chosen using estimates of $\mathbb{E}(U_{0,0})$ and var$(U_{0,0})$.

[3]In general, any other distribution with positive support can be used for $U_{0,0}$. Even though it bounds $C^{(n)}[k]$ to a maximum value of M, we choose the β-distribution to facilitate approximations in the queuing analysis of the MWM in Section IV-F.

TABLE I

Comparison of the tree-based WIG and MWM models. For approximating a signal with a strict fGn covariance structure as in (1), both the WIG and MWM require only three parameters (mean, variance, and H).

	WIG	**MWM**
Building blocks	Independent wavelet coeffs.	Independent multipliers
Marginals	Gaussian	Asymp. Lognormal
LRD	matched	matched
Bursts	Monofractal	Multifractal
Parameters	$2 + \log_2 N$	$2 + \log_2 N$
Synthesis	$O(N)$	$O(N)$

D. Matching burstiness and LRD: WIG vs. MWM

To test the capability of both the WIG and MWM models, we use two real data traces:[4] the LBL-TCP-3 trace of Lawrence Berkeley Laboratory (1994) [26] and the BC-pAug89 trace of Bellcore (1989) [1]. To model the data, we use estimates of the var(W_j) at the 15 finest dyadic scales, where there is sufficient data to obtain good estimates.

Figure 1(c) demonstrates that the MWM produces positive "spiky" data akin to the real traffic, contrary to the WIG model.[5] Also, the marginals of the MWM traces match that of the LBL-TCP-3 trace much better than the WIG (see Figure 4). This seems surprising since we use all the MWM's parameters to match only the correlation structure just like the WIG. The superiority of the MWM indicates that both its multiplicative structure and the choice of β-distributions for the multipliers, are natural for modeling these data sets. However, the MWM can exactly match higher-order moments of training data by using multipliers with more parameters than the β-distribution. By design, both the WIG and MWM models match the second-order correlation structure (see Figure 4(d)).

IV. Multiscale Queuing Analysis

Queuing analysis is fundamental to network engineering. Buffer dimensioning in routers and call admission control are but two of the many crucial areas in networking research that rely on an accurate characterization of the queuing behavior of data traffic.

The discovery of LRD in traffic has created a challenging new area of research in queuing theory. Analytical studies have proven that an infinite-length buffer with constant service rate

[4]The traces contain traffic generated by closed-loop flow control algorithms (e.g., the transmission control protocol (TCP)). Such traffic is dependent on network parameters such as link capacities. Thus, using "open-loop" models such as the MWM to model TCP traffic for network design purposes (e.g., setting link capacities) can produce misleading results [37]. Open-loop models are more appropriate for traffic independent of the network (e.g., streaming video) and possibly for closed-loop traffic in applications other than network design.

[5]Additive models such as the WIG cannot possess *multifractal* properties similar to the MWM [27]. In order for an additive model to exhibit multifractal behavior the variances of the wavelet coefficients would have to depend not only on scale but also on location.

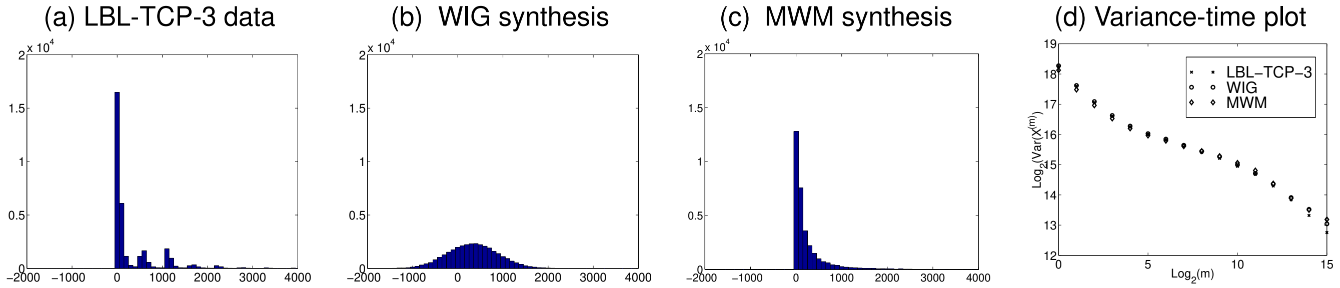


Fig. 4. *Histograms of the bytes-per-times process for (a) wide-area TCP traffic at the Lawrence Berkeley Laboratory (trace LBL-TCP-3) [26], (b) one realization of the WIG model, and (c) one realization of the MWM synthesis. Note the large probability mass over negative values for the WIG model. (d) Variance-time plots of the real data trace (LBL-TCP-3), and synthetic WIG and MWM data traces.*

fed with traffic loads from fGn-based models has a tail queue distribution that decays asymptotically like a Weibullian law

$$\mathrm{P}[Q > b] \simeq \exp(-\delta b^{2-2H}). \tag{20}$$

Here, δ is a positive constant that depends on the service rate of the queue [7, 8]. Clearly, (20) reveals that the decay of the tail queue distribution for fGn with $H > 1/2$ is much slower than the exponential decay predicted by short-range dependent (SRD) classical models [2] which correspond to the case $H = 1/2$. In spite of this result, there is still an ongoing discussion on the effect of LRD on queuing, with researchers arguing both for and against its importance [14–17, 38, 39].

The impact of multiscale marginals on queuing has been demonstrated experimentally in [28]. To better understand how marginals affect queuing, we develop a novel queuing analysis which is particularly adapted to *multiscale representations* of signals and processes. More precisely, exploiting the inherent binary tree structure of the Haar scaling coefficients of both the WIG and the MWM traffic models, we derive approximate formulas for their tail queue probability. Our queuing formulas:

1. are applicable to tree-based/multiresolution models such as the additive WIG and multiplicative MWM,
2. are valid for any queue size, unlike (20) which is an asymptotic result,
3. capture more complicated correlation structures than the mere asymptotic LRD exponent H, and
4. incorporate the entire distribution of the data at multiple time resolutions and not only the second-order statistics.

A. Analytic queuing for tree-based multiscale models

In this section, we develop a new multiscale approach to queuing analysis. We derive an approximate formula for the tail queue probability of tree-based multiscale models such as the WIG and MWM.[6]

Consider a discrete time random process $L_i,\ i \in \mathbb{Z}$, the traffic load, which we think of as entering an infinite buffer single server queue with constant link capacity c. Let Q_i represent the queue size at time instant i. Denote by K_r the aggregate traffic arriving between time instants $-r+1$ and 0, that is,

$$K_r := \sum_{i=-r+1}^{0} L_i. \tag{21}$$

In the sequel, we refer to K_r as representing the data at time-scale r. Set $K_0 = 0$. Using Lindley's equation [40], it is easily shown that

$$Q_0 = \max[Q_{-r} + K_r - rc, K_{r-1} - (r-1)c, \cdots, K_0]. \tag{22}$$

Since $Q_{-r} \geq 0$ for all r, we must have

$$Q_0 \geq \sup_{r \in \mathbb{N}} (K_r - rc). \tag{23}$$

Denoting by $-t$ the last instant that the queue was empty before time instant 0 (we set $-t = 0$ if $Q_0 = 0$), we obtain

$$Q_0 = K_t - tc \leq \sup_{r \in \mathbb{N}} (K_r - rc). \tag{24}$$

Thus if the queue was empty at some time in the past, then

$$Q_0 = \sup_{r \in \mathbb{N}} (K_r - rc). \tag{25}$$

In the remainder we will study exclusively Q_t at $t = 0$, and write $Q := Q_0$ for ease of notation.

Note that (25) provides a direct link between queue size Q and the aggregate of the traffic arrival process K_r at *multiple time scales* r. This and the fact that tree based models provide explicit and simple formulas of K_r for dyadic time scales (i.e., $r = 2^m$), are key to our analytical queuing formula. To this end, we make the following three assumptions which we will justify later:

A1. Dyadic time scales are representative of all time scales.
A2. Large arrivals at dyadic time scales are 'nearly' independent.
A3. The tail queue probability of tree based models at the last instant $2^n - 1$ are representative of the empirical tail queue probability of the fitted data.

In short, we claim that the following approximation is valid:

$$\begin{aligned} \mathrm{P}[Q < b] &\approx \mathrm{P}[\sup_{m \in \{0,\cdots,n\}} (K_{2^m} - c2^m) < b] \\ &= \mathrm{P}[(K_{2^m} - c2^m) < b,\ m \in \{0, \cdots, n\}] \\ &\approx \prod_{i=0}^{m} \mathrm{P}[K_{2^m} < b + c2^m]. \end{aligned}$$

[6] A similar analysis may be possible for models not based on trees, but with an explicit relationship between data at different time scales.

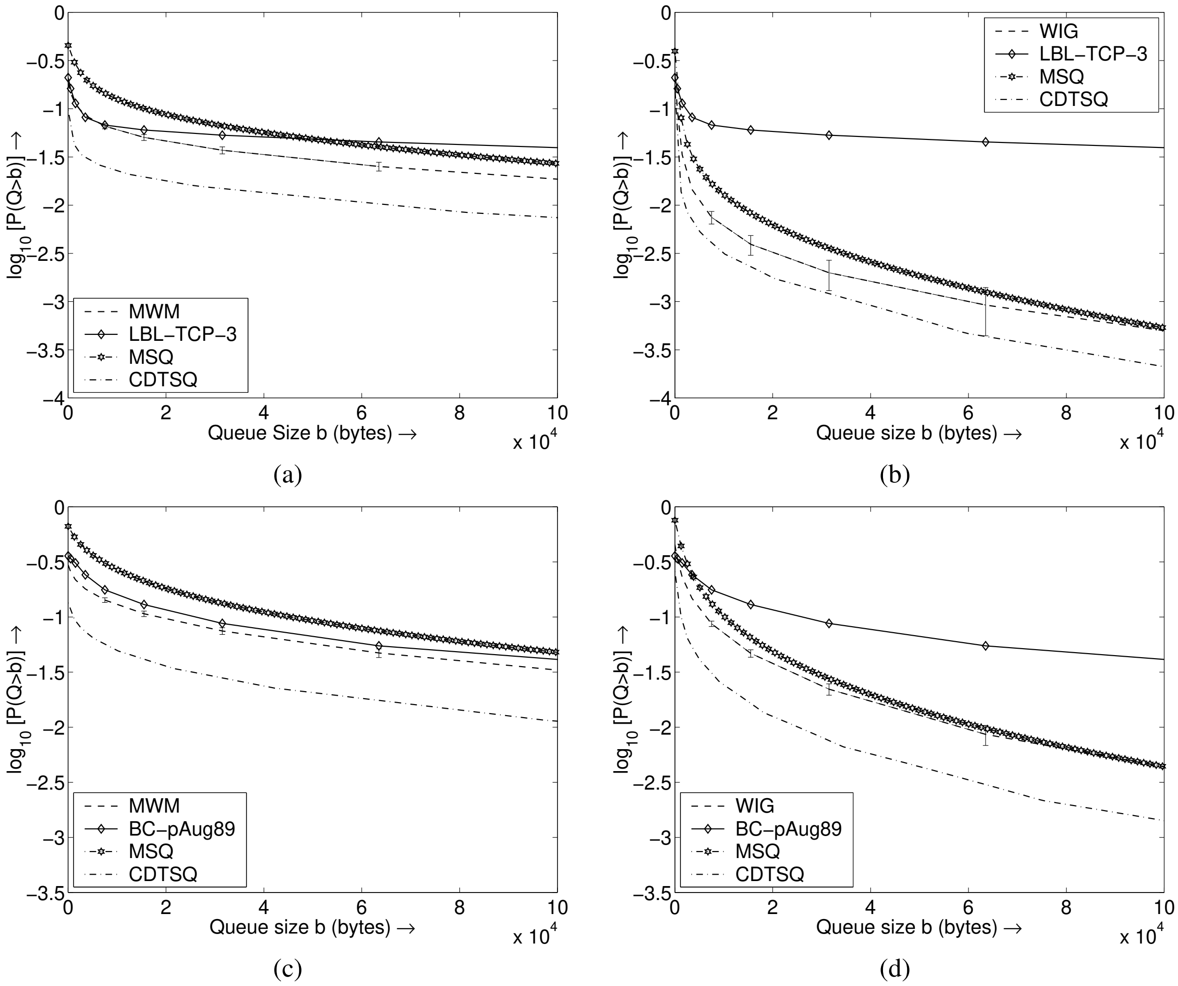


Fig. 5. Queuing performance of real data traces and synthetic WIG and MWM traces: *In (a), we observe that the MWM synthesis matches the queuing behavior of the LBL-TCP-3 data closely, while in (b) the WIG synthesis does not. In (c) and (d), we observe a similar behavior with the BC-pAug89 data. We also observe that the multiscale queue (MSQ) is a close approximation to the empirical queuing behavior for both synthetic traffic loads (both WIG and MWM) and that it is a closer than the critical dyadic time scale queue (CDTSQ). In all experiments in this paper, confidence intervals plotted correspond to a confidence level of* 95%.

This leads us to the following queuing approximation which we call the *multiscale queue* (MSQ):

$$\boxed{\mathrm{MSQ}(b) := 1 - \prod_{i=0}^{n} \mathrm{P}[K_{2^{n-i}} < b + c2^{n-i}].} \tag{26}$$

Note that *multiscale marginals* enter into (26) and not just the correlation structure (or variance-time plot) of the process.

Before going into a more detailed argument supporting this approximation we invite the reader to inspect Figures (5) and (6) for convincing numerical simulations which indicate that

$$\boxed{\mathrm{P}[Q > b] \approx \mathrm{MSQ}(b).} \tag{27}$$

B. *Restriction to dyadic scales (A1)*

To justify A1, we study the quantity Q_D, which is obtained by restricting the supremum in (25) to time scales which appear naturally in a multiscale representation, i.e., the *dyadic time scales*:

$$Q_D := \sup_{m \in \{0,\cdots,n\}} (K_{2^m} - c2^m). \tag{28}$$

The first approximation of our analysis reads then as

$$\mathbf{A1}: \qquad \mathrm{P}[Q > b] \approx \mathrm{P}[Q_D > b]. \tag{29}$$

Clearly, $Q_D \leq Q$ and $\mathrm{P}[Q > b] \geq \mathrm{P}[Q_D > b]$. We justify A1 using the notion of a *critical time scale* (CTS) [14, 16, 17]. The CTS is defined as

$$r^* = \arg \sup_{r \in \mathbb{N}} \mathrm{P}[K_r - cr > b] \tag{30}$$

and the *critical time-scale queue* (CTSQ) as

$$\mathrm{CTSQ}(b) := \mathrm{P}[K_{r^*} - cr^* > b]. \tag{31}$$

It has been shown that $\mathrm{CTSQ}(b) \approx \mathrm{P}[Q > b]$ [14, 16, 17].

Similarly, we introduce now the *critical dyadic time-scale* (CDTS) as

$$r_D^* = \arg \sup_{m \in \{0,\cdots,n\}} \mathrm{P}[K_{2^m} - c2^m > b] \tag{32}$$

and the *critical dyadic time-scale queue* (CDTSQ) as

$$\mathrm{CDTSQ}(b) := \mathrm{P}[K_{r_D^*} - cr_D^* > b]. \tag{33}$$

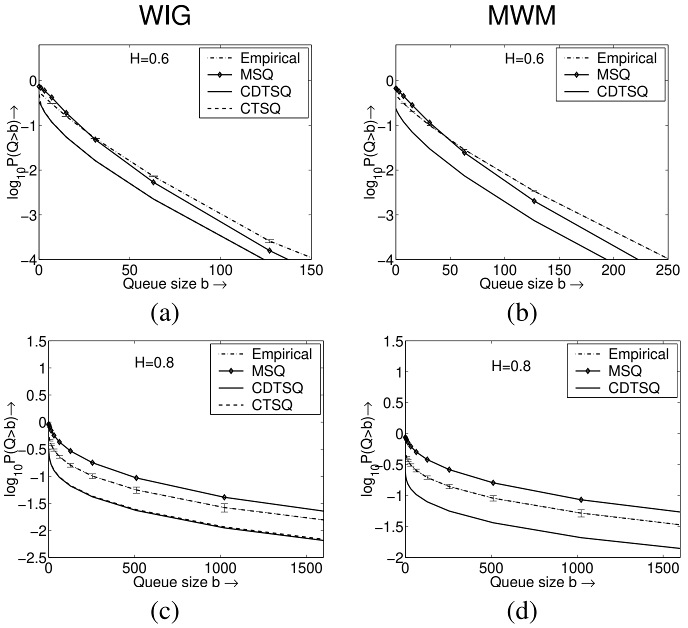


Fig. 6. Justification of the MSQ *We compare analytical queuing formulas (see text) and empirical queuing behavior of WIG (a and c) and MWM (b and d) traces with an fGn correlation structure as in (1). In all cases, the mean, standard deviation and link capacity were* 7, 7, *and* 10 *units respectively. In the case of WIG with (1), explicit formulas for the CTSQ exist [16]. The CDTSQ and CTSQ are almost identical indicating that dyadic scales alone can capture queuing behavior. The multiscale queue (MSQ) gives a good approximation to the empirical queuing behavior and is a closer approximation than the CDTSQ.*

Obviously, $\mathrm{CDTSQ}(b) \leq \mathrm{P}[Q_D > b] \leq \mathrm{P}[Q > b]$.

With the following two points we argue that an estimate of queue length distribution using critical time scales does not change much if we take into account only the distributions at dyadic time scales (CDTSQ) instead of all time scales (CTSQ). The arguments are as follows:

1. Dyadic time scales form only a small subset of $\mathbb{N}$ and so, (25) and (28) could be very different. However, our queuing experiments combined with the analytical expressions for CDTSQ (and CTSQ in the case of WIG where explicit formulas are available [16]) demonstrate convincingly that dyadic time scales are indeed representative for all time scales (see Figure 6).
2. The CDTS takes only time scales up to 2^n into account. However, this is valid if for a given queue size b and a given queue size distribution the bound n is chosen large enough. We will comment further on the dependence of the MSQ on n in a forthcoming paper [41].

Finally, note that the CDTSQ is a computationally efficient substitute for the CTSQ since it requires statistics at only a few dyadic time-scales.

C. Approximate independence of large arrivals on dyadic time scales (A2)

In our queuing analysis, we set

$$E_i := \{K_{2^{n-i}} < b + c2^{n-i}\}. \tag{34}$$

Since $E_i^c := \{K_{2^{n-i}} > b + c2^{n-i}\}$ corresponds to large values of $K_{2^{n-i}}$, we refer to the E_i^c's as *large arrival events*.

From (28) we see that

$$\mathrm{P}[Q_D > b] = 1 - \mathrm{P}[Q_D < b] \;=\; 1 - \mathrm{P}[\cap_{i=0}^{n} E_i]. \tag{35}$$

Thus, the MSQ (26) would equal $\mathrm{P}[Q_D]$ exactly if the events E_i were independent. However, the E_i's are highly probable events with $\mathrm{P}[E_i] \approx 1$. More precisely, most of the numbers $\mathrm{P}[E_i]$ are nearly indistinguishable from 1. Thus,

$$\mathrm{MSQ}(b) \approx \mathrm{P}[Q_D > b] \tag{36}$$

which is confirmed by our numerical experiments. This implies that the events E_i's (and equivalently the large arrival events E_i^c's) are "nearly" independent.

A more rigorous comparison of $\mathrm{MSQ}(b)$ and $\mathrm{P}[Q_D > b]$ can be obtained using the following Lemma which is proven in [41].

Lemma: *Assume that the events E_i are of the form $E_i = \{S_i < b_i\}$, where $S_i = X_0 + \ldots + X_{i-1}$ for $1 \leq i \leq n$ and where $X_0, \cdots, X_n$ are independent, otherwise arbitrary random variables. Then for $1 \leq i \leq n$*

$$P[E_i | E_{i-1}, \cdots, E_0] \geq P[E_i].$$

Given the Lemma and using (28) we have

$$\begin{aligned} \mathrm{P}[Q_D > b] &= 1 - \mathrm{P}[Q_D < b] \;=\; 1 - \mathrm{P}[\cap_{i=0}^{n} E_i] \\ &= 1 - \mathrm{P}[E_0] \prod_{i=1}^{n} \mathrm{P}[E_i | E_{i-1}, \cdots, E_0] \\ &\leq 1 - \prod_{i=0}^{n} \mathrm{P}[E_i] =: \mathrm{MSQ}(b). \end{aligned} \tag{37}$$

We conclude that the MSQ is a conservative approximation of the dyadic queue tail probability. Moreover, $\mathrm{P}[Q > b] \geq \mathrm{P}[Q_D > b] \leq \mathrm{MSQ}(b)$.

D. Stationarity assumption (A3)

Performing an exact queuing analysis of tree-based models such as the WIG and MWM at an arbitrary time instant is very complicated and will produce distributions of the queue size which are non-stationary and vary with time. For an illustration, note that in Figure 2(b) the neighboring nodes $U_{j+2,4k}$ and $U_{j+2,4k+1}$ share a parent node $U_{j+1,2k}$ at scale $j+1$ while the nodes $U_{j+2,4k+1}$ and $U_{j+2,4k+2}$ do not.

The queuing analysis at the last instant $2^n - 1$ of a tree-based model, on the other hand, is simple. Indeed, the Haar scaling coefficients on the branch linking $U_{0,0}$ and $U_{n,2^n-1}$ (the right edge of the tree of Figure 2(b)) are related to the quantities K_{2^m} in (21) by

$$K_{2^{n-i}} = 2^{-i/2} U_{i,2^i-1}, \quad \text{for } i = 0, \ldots, n, \tag{38}$$

and a queuing analysis is feasible. Choosing this particular time instant as the point of analysis is our third assumption which can be formulated in terms of the arriving traffic L_i as

A3: $L_i = C[2^n - 1 + i], \quad i = 0, -1, \ldots, -2^n + 1.$

Assuming stationarity of the data, a tree-based model will produce the same statistics no matter where its right most branch is placed. This justifies A3.

E. Multiscale queuing analysis (MSQ) of the WIG

For the WIG, on choosing

$$X_i := \begin{cases} U_{0,0} & \text{if } i = 0 \\ -2^{i/2} W_{i,2^i-1} & \text{otherwise} \end{cases} \tag{39}$$

we obtain from (38)

$$K_{2^{n-i}} = 2^{-i} \sum_{j=0}^{i-1} X_j = 2^{-i} S_i. \tag{40}$$

Now on setting

$$b_i = b2^i + c2^n, \tag{41}$$

we observe that the WIG satisfies the conditions of the Lemma.

Since for the WIG $K_{2^{n-i}}$ is Gaussian, the probability $\mathrm{P}[E_i]$ can be computed from a Gaussian cumulative distribution [36].

F. Multiscale queuing analysis (MSQ) of the MWM

Denoting $A_{j,2^j-1}$ by A_j, (38) reduces to

$$K_{2^{n-i}} = U_{0,0} \prod_{j=0}^{i-1} (1 - A_j)/2. \tag{42}$$

The event E_i is thus

$$\begin{aligned} E_i &= \{U_{0,0} \prod_{j=0}^{i-1} (1 - A_j) < b2^i + c2^n\} \qquad (43) \\ &= \left\{ \log(U_{0,0}) + \sum_{j=0}^{i-1} \log(1 - A_j) < \log(b2^i + c2^n) \right\}. \end{aligned}$$

By setting

$$X_i := \begin{cases} \log(U_{0,0}) & \text{if } i = 0 \\ \log(1 - A_i) & \text{otherwise} \end{cases} \tag{44}$$

and

$$b_i := \log(b2^i + c2^n), \tag{45}$$

we see that the Lemma applies to the MWM.

For the MWM, obtaining $\mathrm{P}[E_i]$ is not as straightforward as for the WIG. Recall from Section III-B that

$$U_{0,0} \sim \beta_{0,M}(p_{-1}, p_{-1}), \tag{46}$$

and

$$(1 - A_j)/2 \sim \beta_{0,1}(p_j, p_j). \tag{47}$$

Thus, (42) implies that $K_{2^{n-i}}$ is the product of $i+1$ independent β random variables. Using Fan's approximation [36, 42], we approximate the distribution of $K_{2^{n-i}}$ by a beta law supported on $[0, M]$ as follows

$$K_{2^{n-i}} \overset{d}{\approx} \beta_{0,M}(d_i, e_i). \tag{48}$$

The parameters d_i and e_i are given by

$$d_i = S(T - S^2)^{-1}(S - T), \qquad e_i = d_i(1 - S)/S, \tag{49}$$

where $S = 2^{-i}$ and

$$T = \prod_{j=-1}^{i-1} \frac{(p_j + 1)}{2(2p_j + 1)}. \tag{50}$$

This approximation matches the mean and variance of the actual distribution of $K_{2^{n-i}}$ exactly, and closely approximates the first 10 moments [42]. We thus use the cumulative distribution of the $\beta_{0,M}(d_i, e_i)$ random variable to calculate $\mathrm{P}[E_i]$ [36].

V. Multiscale marginals, LRD and queuing

The impact of the non-Gaussian nature of the real data on queuing is considerable, as we demonstrate in Figure 5. There, we observe that all traces exhibit Weibullian tail queue probabilities when input to an infinite-length single-server queue (link capacity 800 bytes/unit time), which is typical for LRD traffic (compare (20)). However, apart from this asymptotic match, the MWM is much closer to the queuing behavior of the real data traces.

The MSQ uses not just the variance (or LRD) of the data, but its entire distribution at multiple time scales. It is thus a tool fit to assess the influence of marginals and LRD on queuing and hence the difference in queuing behavior of the Gaussian WIG and the approximately log-normal MWM models.

In particular, observe from (26) that the MSQ increases as the distribution of data at different scales becomes more heavy-tailed. Thus, a Gaussian LRD process will have a higher MSQ than a Gaussian SRD process. However, Gaussian LRD models cannot capture the tail distribution of non-Gaussian processes and hence can lead to optimistic predictions of queuing behavior. In this sense, the MSQ reveals the limitations of Gaussian modeling.

VI. Conclusions

The importance of capturing scaling properties when modeling traffic loads has now been well recognized [1, 27]. In our work, we rely on multiscale models such as the Gaussian WIG and non-Gaussian MWM models. Both models are built on binary trees, which allow fast $O(N)$ algorithms for synthesis of an $N-$point data set. By matching the variance of a given traffic trace on all dyadic scales, both models capture the correlation structure with only about $\log N$ parameters.

The main contribution of this paper is our *multiscale queuing* (MSQ) approach, which provides a closed form queuing formula for tree-based models. Unlike earlier work on queuing of LRD traffic [8,9], our formula takes into account the entire cumulative distribution of the traffic at different time scales and not just their variances.

The implications are manifold. First, the MSQ is applicable to multiscale models such as the WIG and the MWM. As a consequence, the versatile MWM model is now viable for numerous networking applications, including call admission control.

Second and most importantly, the MSQ is to our knowledge the first tool for assessing the impact of multiscale marginals on queuing. Earlier queuing experiments with synthetic traffic produced using the WIG and the MWM have already suggested that marginals have an influence on the queue length distributions of LRD traffic [28]. Confirming these findings with the marginal-sensitive MSQ, we are now able to conclude that indeed modeling heavy-tailed spiky data with Gaussian models can lead to over-optimistic predictions of tail queue probability.

Thirdly, since the MSQ captures the queuing behavior of training data while using statistics from just the dyadic time-scales, we conclude that dyadic time-scales though few in number, efficiently capture the queuing behavior of traffic.

Our future research will aim at making the MWM practicable for prediction. The parameters of the MWM could also be used to capture the effect of different protocols on shaping data flow. In short, the use of the MWM in real-time network protocols and control algorithms seems very promising.

References

[1] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, pp. 1–15, 1994.

[2] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent traffic," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 209–223, April 1996.

[3] F. Brichet, J. Roberts, A. Simonian, and D. Veitch, "Heavy traffic analysis of a fluid queue fed by a superposition of ON/OFF sources," *COST*, vol. 242, 1994.

[4] N. Likhanov, B. Tsybakov, and N. Georganas, "Analysis of an ATM buffer with self-similar input traffic," *Proc. IEEE, Info com '95 (Boston 1995)*, pp. 985–992, 1995.

[5] M. Taqqu and J. Levy, *Using renewal processes to generate LRD and high variability.* In: Progress in probability and statistics, E. Eberlein and M. Taqqu eds., vol. 11. Birkhaeuser, Boston, 1986. pp 73–89.

[6] J. Choe and N. Shroff, "Supremum distribution of gaussian processes and queueing analysis including long-range dependence and self-similarity," *Stochastic Models* submitted, 1997.

[7] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *COST*, vol. 242, 1994.

[8] N. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. Cambr. Phil. Soc.*, vol. 118, pp. 363–374, 1995.

[9] I. Norros, "Four approaches to the fractional Brownian storage," *Fractals in Engineering*, pp. 154–169, 1997.

[10] G. Gripenberg and I. Norros, "On the prediction of fractional Brownian motion," *J. Applied Probability*, vol. 33, pp. 400–410, 1996.

[11] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic. Evidence and possible causes," in *Proceedings of SIGMETRICS '96*, May 1996.

[12] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking (Extended Version)*, vol. 5, pp. 71–86, Feb. 1997.

[13] N. Duffield, "Economies of scale for long-range dependent traffic in short buffers," *Telecommunication Systems,* to appear, 1998.

[14] B. K. Ryu and A. Elwalid, "The Importance of Long-range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities," *Proc. ACM SIGCOMM Conf.*, vol. 26, no. 4, pp. 3–14, 1996.

[15] D. P. Heyman and T. V. Lakshman, "What are the implications of long-range dependence for VBR-video traffic engineering?," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 301–317, June 1996.

[16] A. L. Neidhardt and J. L. Wang, "The concept of relevant time scales and its application to queuing analysis of self-similar traffic," in *Proc. SIGMETRICS '98/PERFORMANCE '98*, pp. 222–232, 1998.

[17] M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *Computer-Communication-Review*, vol. 26, pp. 15–24, October 1996.

[18] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, pp. 910–916, Mar. 1992.

[19] L. Kaplan and C.-C. Kuo, "Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Haar basis," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3554–3562, Dec. 1993.

[20] G. W. Wornell, "A Karhunen-Loève like expansion for $1/f$ processes via wavelets," *IEEE Trans. Inform. Theory*, vol. 36, pp. 859–861, Mar. 1990.

[21] A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE transactions on information theory*, vol. 38, pp. 904–909, March 1992.

[22] S. Ma and C. Ji, "Modeling video traffic in the wavelet domain," in *Proc. of 17th Annual IEEE Conf. on Comp. Comm., INFOCOM*, pp. 201–208, Mar. 1998.

[23] L. Kaplan and C.-C. Kuo, "Extending self-similarity for fractional Brownian motion," *IEEE Trans. Signal Proc.*, vol. 42, pp. 3526–3530, Dec. 1994.

[24] J. Roberts, U. Mocci, and J. V. (eds.), "Broadband network teletraffic," in *Lecture Notes in Computer Science, No 1155*, Springer, 1996.

[25] S. Bates and S. McLaughlin, "The estimation of stable distribution parameters from teletraffic data," *preprint*, 1998.

[26] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, 1995.

[27] R. H. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. Info. Theory, (Special issue on multiscale statistical signal analysis and its applications)*, vol. 45, pp. 992–1018, April 1999. Available at `www.dsp.rice.edu`.

[28] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Simulation of non-Gaussian long-range-dependent traffic using wavelets," *Proc. SigMetrics*, pp. 1–12, May 1999.

[29] D. Cox, "Long-range dependence: A review," *Statistics: An Appraisal*, pp. 55–74, 1984.

[30] M. Taqqu, V. Teverovsky, and W. Willinger, "Estimators for long-range dependence: An empirical study," *Fractals.*, vol. 3, pp. 785–798, 1995.

[31] P. Abry, P. Gonçalvés, and P. Flandrin, "Wavelets, spectrum analysis and $1/f$ processes," in *Lecture Notes in Statistics: Wavelets and Statistics* (A. Antoniadis and G. Oppenheim, eds.), vol. 103, pp. 15–29, 1995.

[32] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation and synthesis of scaling data," in *Self-similar Network Traffic and Performance Evaluation*, Wiley, June 1999.

[33] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, 1998.

[34] I. Daubechies, *Ten Lectures on Wavelets*. New York: SIAM, 1992.

[35] K. E. Timmerman and R. D. Nowak, "Multiscale Bayesian estimation of Poisson intensities," in *Proc. 31st Asilomar Conf.*, (Pacific Grove, CA), Nov. 1997.

[36] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1-2. New York: John Wiley & Sons, 1994.

[37] Y. Joo, V. Ribeiro, A. Feldmann, A. C. Gilbert, and W. Willinger, "On the impact of variability on the buffer dynamics in IP networks," *Proc. of the 37th Annual Allerton Conference on Communication, Control, and Computing, Allerton, IL*, Sept. 22-24 1999. Available at `www.dsp.rice.edu/publications`.

[38] M. Paulekar and A. M. Makowski, "Tail probabilities for a multiplexer with self-similar traffic," *Proc. IEEE INFOCOM*, pp. 1452–1459, 1996.

[39] B. V. Rao, K. R. Krishnan, and D. P. Heyman, "Performance of Finite-Buffer Queues under Traffic with Long-Range Dependence," *Proc. IEEE GLOBECOM*, vol. 1, pp. 607–611, November 1996.

[40] D. V. Lindley, "The theory of queues with a single server," *Proceedings of the Cambridge Philosophical Society*, vol. 48, pp. 277–289, 1952.

[41] V. J. Ribeiro, R. Riedi, M. Crouse, and R. Baraniuk, "Multiscale modeling and queuing analysis of long-range-dependent network traffic," *preprint*, 1999. Available at `www.dsp.rice.edu/publications`.

[42] D.-Y. Fan, "The distribution of the product of independent beta variables," *Commun. Statist.-Theory Meth.*, vol. 20, no. 12, pp. 4043–4052, 1991.