SUBMITTED TO IEEE TRANSACTION ON NETWORKING

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— **This paper develops a novel approach to queuing analysis tailor-made for multiscale long-range-dependent (LRD) traffic models. We review two such traffic models, the wavelet-domain independent Gaussian model (WIG) and the multifractal wavelet model (MWM). The WIG model is a recent generalization of the ubiquitous fractional Brownian motion process. Both models are based on a multiscale binary tree structure that captures the correlation structure of traffic and hence its LRD. Due to its additive nature, the WIG is inherently Gaussian, while the multiplicative MWM is non-Gaussian. The MWM is set within the framework of multifractals, which provide natural tools to measure the multiscale statistical properties of traffic loads, in particular their burstiness.**

Our queuing analysis leverages the tree structure of the models and provides a simple closed-form approximation to the tail queue probability for any given queue size. This makes the WIG and MWM suitable for numerous practical applications, including congestion control, admission control, and cross-traffic estimation. The queuing analysis reveals that the marginal distribution and, in particular, the large values of traffic at different time scales strongly affect queuing. This implies that merely modeling the traffic variance at multiple time scales, or equivalently, the second-order correlation structure, can be insufficient for capturing the queuing behavior of real traffic. We confirm these analytical findings by comparing the queuing behavior of WIG and MWM traffic through simulations.

I. Introduction

Traffic models play a significant rôle in network engineering. First, they help identify the key properties of traffic affecting performance through both analysis and simulation. Second, they are essential to on-line prediction and estimation algorithms. Third, they help unravel the causes of complex network traffic dynamics.

One crucial property of high-speed network traffic is the presence of *long-range dependence* (LRD), which was demonstrated convincingly in the landmark paper of Leland et al [2]. There, measurements of traffic load on an Ethernet displayed *fractal* behavior or *self-similarity*, i.e., the traffic "looked statistically similar" (highly variable) on all time scales. These features are inadequately described by classical traffic models such as homogeneous Poisson or Markov models.

The LRD of data traffic has important performance implications. A single-server FIFO queue fed with LRD traffic as input leads to much larger queue sizes than when fed with traffic from classical models [2–8]. This implies that LRD affects packet loss and delay and hence must be taken into account carefully when designing networks, providing quality-of-service (QoS) guarantees, etc.

Among the numerous LRD traffic models that have been proposed [7, 9], *fractional Brownian motion* (fBm) has received the most attention, mainly because its Gaussian nature and its strong scaling properties facilitate analytical studies of its queuing behavior [5, 6, 8]. Moreover, fBm has helped reveal that client behavior is the cause of self-similarity in traffic over time scales from a few hundreds of milliseconds and larger [10, 11].

Though fBm is an appropriate traffic model in some cases [3, 10, 11], it can only model real-world traces with a rigid, restrictive correlation structure. Indeed, the importance of *short-term correlations* has been convincingly demonstrated for queuing in finite-length buffers [12–14], and so-called *critical time scales* that affect queuing have been discovered [14–16].

Wavelet-based multiscale models provide generalizations of fBm with more flexible correlation structures [17–20]. Using efficient multiscale tree structures, these models provide fast $O(N)$ synthesis algorithms to synthesize N-point data sets [21, 22]. Due to their additive nature, these models are inherently Gaussian, and so we will term them *wavelet-domain independent Gaussian* (WIG) models.

As a consequence of its Gaussian nature, unfortunately, a WIG model can produce unrealistic synthetic traffic traces in certain situations. First, Gaussian traffic can take negative values while real traffic is inherently positive. Second, a Gaussian

[$]This work was supported by NSF grants CCR–99–73188 and ANI–99–79465, ONR grant N00014–99–10813, DARPA/AFRL grant F30602–00–2–0557, and Texas Instruments.
Email: {vinay, riedi, mcrouse, richb}@rice.edu.
URL: www.ece.rice.edu/INCITE.
A conference version of this paper was presented at INFOCOM 2000 [1].

marginal cannot capture burstiness on small time scales (see Figure 1 and Figure 5). Various authors have observed heavy-tailed marginals in traffic [23, p. 364], [24], in particular on small time scales from a few hundred milliseconds and smaller. However, the WIG model may be appropriate for modeling traffic at time scales on the order of seconds and larger, where traffic appears more Gaussian.

In [26], we proposed a simple multiplicative traffic model called the *multifractal wavelet model* (MWM). The non-Gaussian MWM captures the "spiky" bursts of measured traffic better than Gaussian models (see Figure 1(c)) and matches the tail queue probability of measured traffic more accurately than Gaussian LRD traffic models (see Figure 7).

The MWM is a multiplicative cascade. Cascades are naturally associated with a powerful tool called *multifractal analysis*, which provides a statistical language and calculus to characterize burstiness. Using multifractal scaling and the simple concept of cascades, the TCP layer has been identified as the most likely component in the present hierarchy of information transfer where the multifractal bursts of data traffic are formed [27–29].

The primary contribution of this paper is a novel multiscale approach to queuing analysis that applies whenever a multiscale representation with independent inter-scale innovations of the traffic load (such as in the WIG or the MWM) is an acceptable approximation of reality. The queue length of an infinite-length buffer with constant link capacity c (assuming the queue was empty some time in the past) obeys the identity [30]

$$Q = \sup_r (K[r] - rc). \tag{1}$$

Here $K[r]$ is the total traffic that entered the queue in the past r time instants. In other words, the queue size Q is a function of the traffic arrivals aggregated at multiple time scales corresponding to r time units. In the multiscale representation of the WIG and MWM models, such aggregates appear explicitly at dyadic time scales, i.e., for $r = 2^m,\ m \in \mathbb{N}$, and these dyadic scale aggregates are related to each other by independent random innovations. We exploit this fact to derive an approximation to the tail queue probability.

The resulting queuing formula, which we call the *multiscale queuing formula* (MSQ),

- is non-asymptotic, i.e., it is valid for any queue length,
- closely approximates the tail queue probability, as the experiments in Section IV verify (see Figure 7),
- requires traffic statistics at only a few dyadic time scales, and
- is easy-to-use.

As a consequence, the WIG and MWM become viable for applications requiring models with accurate queuing formulas [31].

The MSQ reveals that the *tail behavior* of the multiscale marginals of the traffic load significantly impacts queuing. Since LRD captures only the *variance* of traffic at multiple time scales, it inadequately captures queuing behavior and leads to poor predictions of the tail queue probability. We will demonstrate that the multiplicative structure of the MWM synthesizes multiscale marginals much more similar to that of training traffic traces than the WIG.

In this paper we will restrict our attention to "open-loop" traffic models. Network traffic loads are determined by a number of factors, including user behavior, the network topology, and the interaction of numerous protocols from the application to physical layers. Open-loop models treat traffic traces as random processes, capturing important statistical properties. Such models ignore the closed-loop feedback used by the transmission control protocol (TCP) for congestion control. They are useful models for user datagram protocol (UDP) traffic as well as for TCP traffic in situations where TCP's closed loop is not significantly affected, for example applications like cross-traffic estimation [31]. For certain network design applications like setting link capacities or router buffer sizes, open-loop models with fixed parameters can give erroneous results [32].

In Section II we introduce fBm and the WIG. Section III describes the MWM and demonstrates its superiority over the WIG model in capturing the marginals of traffic. We introduce our multiscale queuing formula (MSQ) in Section IV and apply it to both the WIG and MWM. We also experimentally demonstrate the importance of the non-Gaussian nature of traffic on queuing and provide empirical evidence for the accuracy of our theoretical queuing formulas. We use our queuing formulas to explain why marginals and LRD affect queuing in Section V and conclude in Section VI. Appendices A and B prove Lemmata related to the MSQ formula.

II. Classical Multiscale Models for LRD Processes

A. Long-range dependence

Consider a discrete-time, wide-sense stationary random process $\{X[t],\ t \in \mathbb{Z}\}$ 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 $X^{(m)}[t]$, which is obtained by summing $X[t]$ over non-overlapping blocks of length m

$$X^{(m)}[t] \ := \ X[tm - m + 1] + \cdots + X[tm]. \tag{2}$$

Denote the auto-covariance of $X^{(m)}$ by $r_X^{(m)}$. 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 [33]. Two additional equivalent definitions of LRD require that $r_X^{(m)}[0]/m \ \to\ \infty$ as $m \ \to\ \infty$ or that the power spectrum $S_X(f)$ is singular at $f\ =\ 0$ [33].

One example of an LRD process is *fractional Gaussian noise* G (fGn), the increment process of the fractional Brownian motion B (fBm): $G[k] := B\left[(k+1)\Delta\right] - B\left[k\Delta\right]$ with Δ a constant time lag. fGn's autocorrelation is given by

$$r_G[k] = \frac{\sigma^2}{2} |\Delta|^{2H} \left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H} \right) \tag{3}$$

with $\sigma^2 = \mathrm{var}\left(B[1]\right)$. The LRD of fGn is captured by the Hurst parameter H. For $H \ >\ 1/2$, the correlation is positive with increasing strength as $H \ \to\ 1$. Note that[1] $r_G[k] \ \simeq\ k^{2H-2}$

[1] In the sequel, "$\simeq$" denotes asymptotic decay.

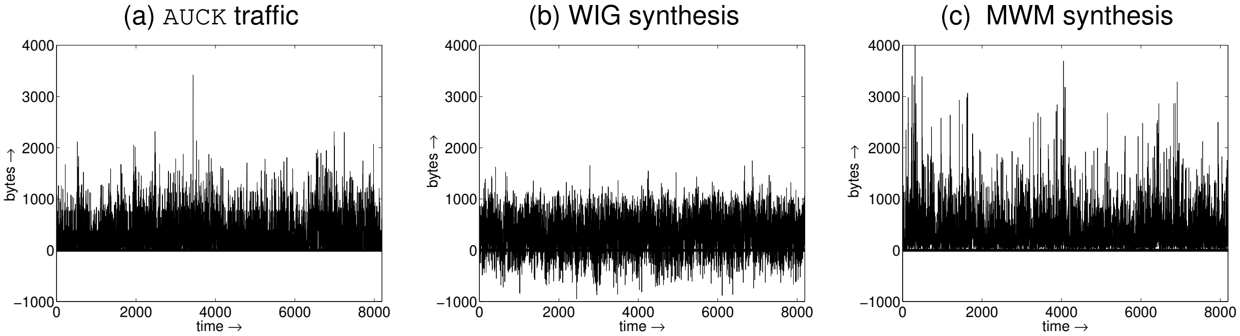


Fig. 1. Modeling bursty traffic data. *Arrival processes of bytes per* 8ms *for (a) wide-area traffic at the University of Auckland (*AUCK*) [25], (b) one realization of the state-of-the-art wavelet-domain independent Gaussian (WIG) model [21], and (c) one realization of the multifractal wavelet model (MWM) synthesis. The MWM trace closely resembles the real data, while the WIG trace does not. Note in particular the many negative values of the WIG.*

and thus $H \leq 1$. A simple estimation procedure for H, called the *variance-time plot*, makes use of the fact that $r_G^{(m)}[k] = r_G[k]m^{2H}$. Thus, a line fitted to the plot of an estimate of $\log \text{var}[G^{(m)}]$ against $\log m$ will have slope $2H$. The main advantage of this method lies in its simplicity. However, it is biased and often inaccurate. More reliable estimators of H have been devised [34], in particular a wavelet-based unbiased one [35].

B. Multiscale trees

Multiscale trees provide a simple framework for visualizing and exploiting the relationship between the aggregates $X^{(m)}$ of a process at multiple time scales. They are the basic ingredient of the multiscale models and queuing analysis we will discuss in this paper.

Consider a process X, for example the number of bytes per millisecond of traffic arriving at a router, and let us iteratively construct a particular multiscale representation. Due to the nature of this representation, we will consider a process of 2^n data points for some integer n; say for simplicity $X[k]$ for $k = 0, \ldots, 2^n - 1$. To start the iteration, let $V_{n,k} := X[k]$. The $V_{n,k}$ form the nodes at the lowest level n of the binary multiscale tree of Figure 2(a). Nodes at higher levels correspond to coarser resolutions of the process $X[k]$; i.e.,[2]

$$V_{j,k} := X^{(2^{n-j})}[k]. \quad (4)$$

Clearly, every parent node is the sum of its two children at the next finer scale. The coarsest resolution consists of a single node, the tree-root $V_{0,0}$, that equals the total traffic that arrives at the router in the 2^n time units.

To perform a *multiscale analysis* of a given traffic trace $X[k]$, we flow up the multiscale tree, forming nodes at coarser time scales by aggregating nodes at finer scales. Multiscale analysis techniques such as variance-time plot analysis (see Section II-A) and multifractal analysis (see Section III-C) use traffic statistics at different levels of the multiscale tree.

Apart from analysis, multiscale trees can also be used for *synthesis* of traffic. Starting from the tree-root, we can flow down the tree generating nodes at finer scales to synthesize a process

[2]This notation differs from that in our other papers on the MWM [1, 26, 36].

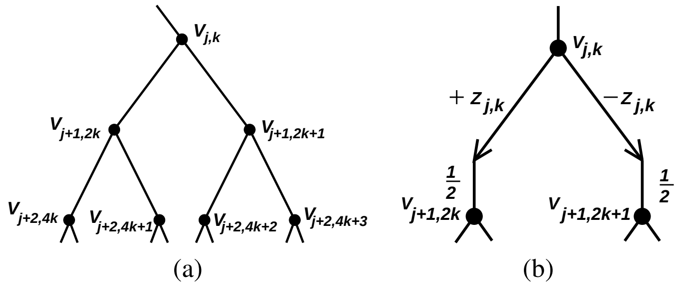


Fig. 2. *(a)* Multiscale tree representation of a traffic trace. *Nodes at each horizontal level in the tree correspond to the sum (aggregates) of the process in non-overlapping blocks of sizes of powers of two, with lower levels corresponding to smaller block sizes. Each node is the sum of its two child nodes. (b)* The WIG model. *The sibling nodes $V_{j+1,2k}$ and $V_{j+1,2k+1}$ are generated as the sum and difference of the parent node $V_{j,k}$ and a random Gaussian innovation $Z_{j,k}$. Within each scale (i.e., for fixed j) the $Z_{j,k}$'s are i.i.d. $\mathcal{N}(0, \sigma_j^2)$ random variables.*

$X[k]$ with certain statistics. We next describe two such synthesis techniques.

C. Multiscale trees for Gaussian signals: WIG model

The additive nature of computing nodes at coarser levels in the multiscale tree (see Figure 2(a)) suggests a simple additive multiscale model for LRD traffic [17–21] that is equivalent to the Haar wavelet system [37].

Starting at the node $V_{j,k}$, we model its two child nodes $V_{j+1,2k}$ and $V_{j+1,2k+1}$ using *independent additive random innovations* $Z_{j,k}$, through (see Figure 2(b))

$$\begin{aligned} V_{j+1,2k} &:= (V_{j,k} + Z_{j,k})/2, \\ V_{j+1,2k+1} &:= (V_{j,k} - Z_{j,k})/2. \end{aligned} \quad (5)$$

Since the $Z_{j,k}$'s are (up to normalization constants) exactly the Haar wavelet coefficients of the resulting process, we call this model the *wavelet-domain independent Gaussian* (WIG) model. The $Z_{j,k}$'s must be identically distributed within scale j to provide a first-order stationary process. The pyramid structure of (5) results in a rapid $O(N)$ algorithm for synthesizing an N-point traffic trace [21].

Due to the Central Limit Theorem (CLT), we expect the distributions of the nodes $V_{j,k}$ to become Gaussian as $j \to \infty$. It is natural, then, to use Gaussian variables for $V_{0,0}$ and $Z_{j,k}$.

In the sequel, when referring to moments of multiscale random variables that are identically distributed within scale, we drop their location coordinate for convenience. For example we write $\mathbb{E}[V_{j,k}]$ as $\mathbb{E}[V_j]$.

To synthesize an LRD process with a desired variance-time plot, we merely to choose the variances of the $Z_{j,k}$ appropriately, since

$$\mathrm{var}[V_{j+1}] = \left(\mathrm{var}[V_j] + \mathrm{var}[Z_j]\right)/4. \tag{6}$$

In particular, choosing[3]

$$\begin{aligned} V_{0,0} &\sim \mathcal{N}(m2^n, \sigma^2 2^{2nH}) \ \text{ and} \\ Z_{j,k} &\sim \mathcal{N}\Big(0, \sigma^2 2^{2(n-j)H}\left(2^{2-2H} - 1\right)\Big) \end{aligned} \tag{7}$$

provides a Gaussian process with $\mathrm{var}[V_j] = \sigma^2 2^{2(n-j)H}$, i.e., with the same variance-time plot as fGn with Hurst parameter H, mean m, and variance σ^2 at the finest time scale n.

III. Multifractal Wavelet Model

A. Multiscale trees for positive signals: MWM model

The WIG model is Gaussian by construction, and any additive model of the form (5) will be so at least approximately. This is in sharp contrast with the fact that network traffic signals (such as loads and interarrival times) can be highly "spiky" and non-Gaussian (recall Figure 1(a)). We seek a more accurate marginal characterization for these bursty, non-negative LRD processes yet wish to retain the simplicity of a tree-based model.

The *multifractal wavelet model* (MWM) we proposed in [26] achieves this with a simple modification to the WIG model. Unlike the additive WIG, the MWM uses independent *multiplicative innovations* $M_{j,k}$, modeling the two children $V_{j+1,2k}$ and $V_{j+1,2k+1}$ of node $V_{j,k}$ as (see Figure 3(a))

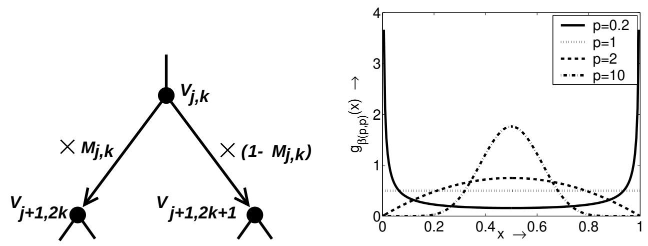
$$\begin{aligned} V_{j+1,2k} &:= V_{j,k} M_{j,k}, \\ V_{j+1,2k+1} &:= V_{j,k}(1 - M_{j,k}). \end{aligned} \tag{8}$$

By choosing $M_{j,k} \in [0, 1]$ and $V_{0,0} \geq 0$, we ensure positive values at all nodes on the tree. Requiring that the multipliers $M_{j,k}$ be symmetric about $1/2$ and identically distributed within scale ensures first-order stationarity of the process at the finest scale n [26]. A similar multiplicative model has been developed in [38], where it is applied to Bayesian estimation of the intensity of a Poisson process.

It remains to choose a distribution for the multipliers $M_{j,k}$ and the tree-root $V_{0,0}$. Due to the CLT and the multiplicative scheme, we expect an approximately lognormal process at finer time scales. Since a product of lognormal random variables is also lognormal, the most natural choice for the multipliers is a lognormal distribution. However, we require $M_{j,k} \in [0, 1]$ and so suggest the *symmetric beta distribution* [39], i.e.,

$$M_{j,k} \sim \beta(p_j, p_j) \tag{9}$$

[3]The symbol "$\sim$" denotes "distributed as."



(a) (b)

Fig. 3. The MWM model. *(a) At scale j, generate the multiplier $M_{j,k} \sim \beta(p_j, p_j)$, and then form the two nodes at scale $j+1$ by multiplying $V_{j,k}$ with $M_{j,k}$ and $1 - M_{j,k}$. (b) Probability density function (PDF) of a $\beta(p, p)$ random variable. For $p = 0.2$, $\beta(p, p)$ resembles a binomial distribution, while for $p = 1$ it reduces to the uniform density. For $p > 1$, the density is close to a truncated Gaussian with increasing resemblance as p increases.*

with variance

$$\mathrm{var}[\beta(p_j, p_j)] = 1/(4 + 8p_j). \tag{10}$$

This distribution is amenable to closed-form calculations and has a flexible shape (see Figure 3(b)). We use the beta parameters p_j to match the second-order statistics of traffic. By using multipliers with more parameters than the β-distribution, the MWM can exactly match higher-order moments of training data [26].

In general, any distribution with positive support can be used for $V_{0,0}$. However, in this paper we choose to model $V_{0,0}$ as a beta random variable, i.e.,

$$V_{0,0} \sim aM_{-1} \tag{11}$$

with $a \geq 0$ a constant and $M_{-1} \sim \beta(p_{-1}, q_{-1})$. Even though (11) bounds $V_{n,k}$ to a maximum value a, we choose a beta distribution to facilitate approximations in the queuing analysis of the MWM in Section IV below. Note that if $p_{-1} \neq q_{-1}$, M_{-1} is an asymmetric beta random variable. In case traffic is not Gaussian at the coarsest resolution, an asymmetric beta random variable is an appropriate choice for M_{-1}. However, often at coarse enough resolutions, traffic marginals appear Gaussian due to aggregation. Since symmetric beta random variables resemble truncated Gaussian random variables (see Figure 3(b)) we model M_{-1} as a symmetric beta random variable, i.e., we set $p_{-1} = q_{-1}$.

The pyramid scheme of (8) and Figure 3(a) constructs N samples of the MWM signal in $O(N)$ computations. In fact, synthesis of a trace of length 2^{18} data points takes just seconds of workstation CPU time.

B. MWM model training

Training the MWM requires setting the parameters a and p_j, $j = -1, 0, 1, \ldots, n-1$. We use these parameters to match the mean and second moment of the tree nodes of a given traffic trace. The parameters a and p_{-1} (see (11)) of $V_{0,0}$ are set as

$$a = 2\,\mathbb{E}[V_{0,0}] \ \text{ and } \ p_{-1} = \frac{1}{2}\left(\frac{(\mathbb{E}[V_{0,0}])^2}{\mathrm{var}[V_{0,0}]} - 1\right). \tag{12}$$

From (8), we obtain

$$\mathbb{E}[V_{j+1}^2] \;=\; \mathbb{E}[V_j^2]\mathbb{E}[M_j^2]$$

$$= \quad a^2 \prod_{i=-1}^{j} \mathbb{E}[M_i^2], \;\; j = -1, 0, \ldots, n-1. \quad (13)$$

Then (10) and (13) give

$$p_j = \frac{\mathbb{E}[V_j^2] - 2\mathbb{E}[V_{j+1}^2]}{4\mathbb{E}[V_{j+1}^2] - \mathbb{E}[V_j^2]}, \;\; 0 \leq j \leq n-1. \qquad (14)$$

Thus, by choosing the p_j's, we can match estimates of $\mathbb{E}[V_{j+1}^2]$. Since $\mathbb{E}[V_j] = 2^{-j}\mathbb{E}[V_{0,0}]$, matching the second moments of the V_j's is equivalent to matching their variance. The parameters p_j thus control the variances on all time scales and in particular the LRD parameter H for self-similar traffic. With one parameter per scale, the MWM has approximately $\log_2 N$ parameters for a trace of length N. To provide a more parsimonious model, this number could be drastically reduced to, say, match only the asymptotic behavior of the variance at multiple scales, i.e., to match only the LRD parameter H, by setting

$$p_j = 2^{2H-2}(p_{j-1}+1) - 1/2, \;\; j = 0, \ldots, n-1. \qquad (15)$$

Now the only parameters are H, p_{-1} and a.

C. Measuring burstiness on small scales

So far we have only addressed the non-Gaussianity of the MWM model. We now explain why its multiplicative structure produces a closer match to the burstiness present in traces of traffic loads.

Research on the bursty or fractal nature of traffic at *large* time scales has had a considerable impact on networking. The large time-scale burstiness of network traffic is now well understood in two respects, statistically in terms of second-order self-similarity (which is captured in the LRD parameter H) and conceptually through the heavy tailed on-off model [2, 9–11]. Also, fractal queuing analysis has provided networking-related insights [5, 6, 8], mostly regarding network design and management.

Performance control and QoS, however, also require *finer* time-scale information not captured by the large time-scale LRD characteristics. We thus need a different tool capable of addressing small time-scale variability and bursts. Multifractal analysis does exactly this.

Current analyses of flow control mechanisms consider time-scales of the order of round-trip times and either simplify or ignore finer time-scale dynamics to make the system amenable to analysis [40]. However, network loads vary considerably and "instantaneous arrival rates" change drastically at small time scales, which as we will show in Section IV affects queuing behavior. Since the notion of a constant rate at infinitely fine scales is meaningless, how can we measure the strength of traffic arrivals in a meaningful way?

One way of measuring burstiness is through *scaling exponents*. Consider a positive continuous-time process $Y(t)$, which should be thought of as the total traffic arriving in the time interval $[0, t]$. Consider the case where Y varies considerably and so is not differentiable. One way of saying that Y is not differentiable is that it cannot be approximated locally by a linear function. The strength of the burst of arrivals at time t and on scale δ is best measured by a scaling exponent $\alpha(t)$:

$$|Y(t+\delta) - Y(t)| \approx \delta^{\alpha(t)}. \qquad (16)$$

The smaller $\alpha(t)$, the larger the increments of Y around time t, and the "burstier" Y is at that time t.

The MWM construction (see (8) and Figure 3(a)), when continued to the limit of infinite resolution, is exactly that of a multifractal binomial cascade [26]. The term *multifractal* refers to the fact that the degree of burstiness (or scaling exponent) assumes different values and varies drastically as a function of time t.

At the heart of multifractal analysis lies the *multifractal formalism* [26], which relates the scaling behavior of sample moments of a trace to the frequency of occurrence of "bursts" of different strength $\alpha(t)$ in that trace. This formalism exploits the moments of *all* orders, unlike the concept of LRD which relies on second-order statistics only.

The multifractal formalism relates non-Gaussianity and burstiness explicitly and furnishes a solid formalism on which to explain the superiority of the multiplicative MWM over the additive WIG in modeling bursty network traffic loads. Moreover, as we elaborate in Appendix B, this formalism is instrumental in relating the burstiness of traffic to the range of queue sizes for which the analytic queuing formula of Section IV is valid.

D. Revealing structure in burstiness

Through the multifractal formalism, assessing the presence of bursts of different strength in an efficient, compact manner becomes numerically feasible. To this end, let $N_j(\alpha)$ denote the number of bursts of strength α at scale $\delta = 2^{-j}$ (see (16)).

In a nutshell, the multifractal formalism relates $N_j(\alpha)$, the frequency of bursts of given strength, to the decay rate $T(q)$ of the moments of Y across dyadic scales:

$$\mathbb{E}\left[\sum_{k_j=0}^{2^j-1} \left|Y((k_j+1)2^{-j}) - Y(k_j 2^{-j})\right|^q\right] \approx 2^{-jT(q)}. \quad (17)$$

The formalism states that under suitable conditions [41]

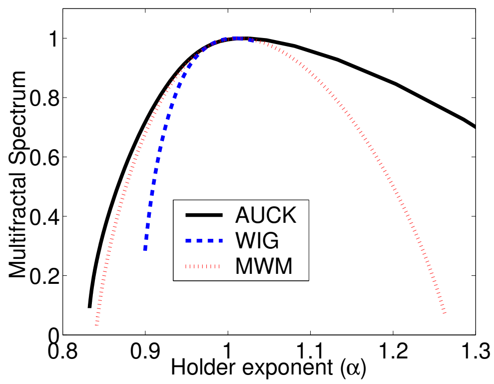
$$N_j(\alpha) \approx 2^{jT^*(\alpha)}, \qquad (18)$$

where T^* is the *Legendre transform* of T, i.e.,

$$T^*(\alpha) := \inf_q (q\alpha - T(q)). \qquad (19)$$

The Legendre transform $T^*(\alpha)$ gives the smallest distance between the straight line $q\alpha$ and the function $T(q)$. If $T^*(\alpha)$ is negative for some α, then (18) means that there is no exponent $\alpha(t) = \alpha$ in the signal. Since the time instances with equal $\alpha(t)$ form highly interwoven fractal sets, $T^*(\alpha)$ is called the *multifractal spectrum*. (For a more rigorous, in-depth presentation as well as a set of relevant references, see [26, 41].)

In practice, traffic traces possess a minimum time resolution and hence a finite range of dyadic time scales. In order to estimate $T(q)$ from such data sets, it is customary to extract the



(b)

Fig. 4. Multifractal spectra of the AUCK data and one realization of the WIG and MWM models. *The MWM spectrum matches that of the real data closely except for large values of α (corresponding to small values of the signal). In particular, it captures the large bursts in the traffic corresponding to small values of α better than the WIG model. Gaussian moments do not exist for $q < -1$, hence the WIG spectrum is not defined for $\alpha > 1$.*

scaling laws (17) from log-log plots. For a tree-based model, this reads as

$$T(q) = -\frac{1}{j}\log_2 \mathbb{E}\left[\sum_{k=0}^{2^j-1} |V_{j,k}|^q\right], \qquad (20)$$

which we use to obtain the multifractal spectra of Figure 4.

For the MWM, assuming that the multipliers $M_{j,k}$ converge in distribution to a limiting random variable $M \sim \beta(p,p)$, we find

$$\begin{aligned} T_{\text{MWM}}(q) &= -1 - \log_2 \mathbb{E}[M^q] \qquad (21)\\ &= \begin{cases} -1 - \log_2 \frac{\Gamma(p+q)\Gamma(2p)}{\Gamma(2p+q)\Gamma(p)} & \text{if } q > -p \\ -\infty & \text{if } q \le -p. \end{cases} \end{aligned}$$

For the WIG with fBm scaling, i.e., $V_j \sim \mathcal{N}(0, 2^{-2jH})$, we obtain

$$T_{\text{WIG}}(q) = \begin{cases} qH - 1 & \text{for } q > -1 \\ -\infty & \text{for } q \le -1. \end{cases} \qquad (22)$$

Due to its linear shape, the Legendre transform of T_{WIG} degenerates to a single point. This corresponds to the fact that fBm possesses only one degree of "burstiness" ($\alpha(t) = H$), which is omnipresent [41]. Consequently, fBm (or its increments process fGn) cannot capture the complicated multifractal behavior or burstiness of real WAN traffic.

Even without a strict fBm scaling, the WIG cannot possess *multifractal* properties similar to the MWM [26]. For an additive model to exhibit multifractal behavior, the variances of the $Z_{j,k}$'s must depend not only on scale j but also on the location k or the line of ancestors. The narrow spectrum of the WIG in Figure 4 demonstrates this fact, though it is not a single point due to difficulties inherent to any numerical estimation of the partition function $T(q)$ and its Legendre transform.[4]

E. Matching burstiness and LRD

We now describe the procedure for matching or fitting multiscale tree models to given traffic traces and then use two traces to compare the WIG with the MWM.

The fit: Given a trace of 2^n data points, we split the trace into $32 = 2^5$ subtraces of length 2^{n-5} and fit trees of depth $n-5$ to these subtraces. We do this to retain enough statistics for reliable estimation of var$[V_j]$ at the coarsest level.

Using the estimates for var$[V_j]$ as well as the overall mean, we obtain the tree parameters using (6) for the WIG and using (12) and (14) for the MWM. **The traces:** The two traces we use are AUCK, which contains the number of bytes per 2ms of recorded WAN traffic (mostly TCP packets) [25] and VIDEO, which consists of 15 video clips multiplexed with random starting points [42]. The finest time scale in VIDEO corresponds to 2.77ms, $1/15$ the duration of a single frame. The mean rates of AUCK and VIDEO are 1.456Mbps and 53.8Mbps, respectively. AUCK contains 1.8×10^6 data points and VIDEO 2^{18}. The Hurst parameter of AUCK obtained from the variance-time plot using time scales 512ms to 262.144s is $H = 0.86$. For VIDEO, we find $H = 0.84$ using time scales 354ms to 90.76s.

WIG vs. MWM: Figure 1 demonstrates that the MWM produces positive "spiky" data akin to the AUCK traffic, contrary to the WIG model. The marginals of the traces and models displayed in Figures 5 and 6 confirm this fact. This is remarkable, since the parameters of both the WIG and the MWM are used solely to match the correlation structure (or equivalently the variance-time plot) of the given trace (see Figures 5(d) and 6(d)). The superiority of the MWM indicates that both its multiplicative structure and the choice of β-distributions for the multipliers are natural for modeling this data set.

For a more sophisticated comparison that takes higher-order correlations into account, see the multifractal spectra $T^*(\alpha)$ of AUCK in Figure 4. As explained in Section III-C, these plots visualize the amount of burstiness in a signal. Small values of α correspond to strong bursts, large α to smooth parts in the signal. The larger the $T^*(\alpha)$, the more often bursts of this strength α are encountered in the signal. As should be expected from the above comparison, the MWM is superior, matching the spectrum of the trace particularly well in the bursty parts.This result will become useful when we discuss queuing performance in Section IV.

This comparison is summarized in Table I. It is notable that the multiplicative MWM is somewhat more accurate than the Gaussian WIG even in the case of VIDEO, a trace that comes close to a Gaussian process. We mention here also, that the MWM is flexible enough to replicate mono-fractal (self-similar) features at coarse scales and multifractal properties on fine scales [26].

[4] Let us mention two such difficulties. First, sample moments are only an approximation to the true moments, and so we cannot expect to numerically obtain a single point spectrum. Second, the moments can actually be infinite, as is the case with $T_{\text{WIG}}(q)$ for $q < -1$ (see (22)). Consequently, we have omitted these q-values from the computation of $T^*_{\text{WIG}}(\alpha)$ which limits the plot of T^*_{WIG} to small values of α in Figure 4.

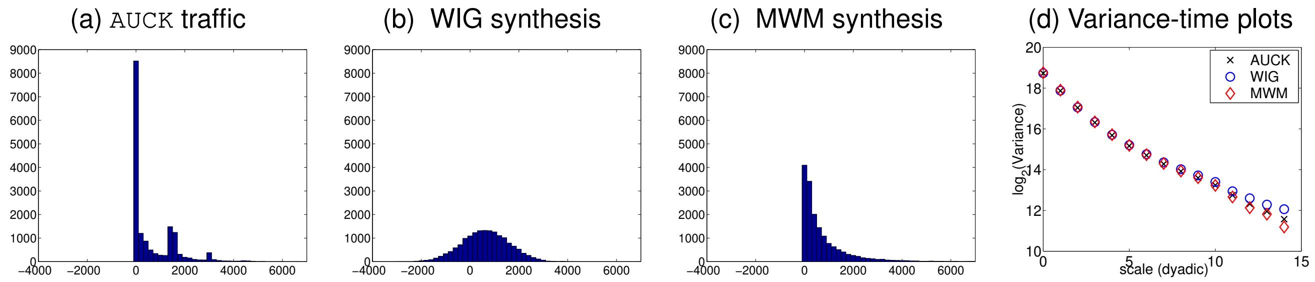


Fig. 5. *Histograms of the bytes-per-time processes for (a) wide-area traffic at the University of Auckland (trace* AUCK*) [25], (b) one realization of the WIG model, and (c) one realization of the MWM. Note the large probability mass over negative values for the WIG model. (d) Variance-time plots of* AUCK *and synthetic WIG and MWM data traces.*

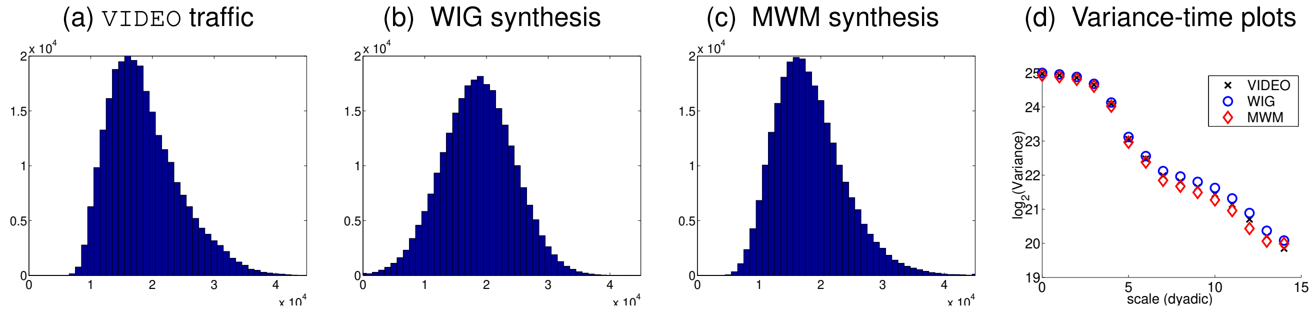


Fig. 6. *Histograms of the bytes-per-time processes for (a) video traffic formed by multiplexing* 15 *video traces (trace* VIDEO*), (b) one realization of the WIG model, and (c) one realization of the MWM. Note that the MWM matches the marginal of the video traffic better than the WIG; however, the video traffic is more Gaussian than the* AUCK *traffic. (d) Variance-time plots of* VIDEO *and synthetic WIG and MWM data traces.*

TABLE I

Comparison of the tree-based WIG and MWM models. For approximating a signal with a strict fGn covariance structure as in (3), both the WIG and MWM require only three parameters (mean, variance, and H).

	WIG	MWM
nature		
building blocks	independent wavelet coeffs.	independent multipliers
marginals	Gaussian	asymp. lognormal
bursts	monofractal	multifractal
parameters	$2+\log_2 N$	$2+\log_2 N$
complexity	$O(N)$	$O(N)$
AUCK		
marginals	not matched	close fit
LRD	matched	matched
bursts	not matched	close fit
VIDEO		
marginals	close fit	close fit
LRD	matched	matched
bursts	matched	close fit

IV. Multiscale Queuing Analysis

In this section, we develop a novel queuing analysis that is particularly adapted to *multiscale representations* of signals and processes. More precisely, exploiting the binary tree structure used in both the WIG and the MWM traffic models, we derive approximate formulas for their tail queue probability.

A. Analytic queuing for tree-based multiscale models

Consider a discrete-time random process $L[i]$, $i \in \mathbb{Z}$, the amount of traffic per time unit that enters an infinite buffer, single server queue with service rate of c bytes per time unit.

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; i.e.,

$$K[r] := \sum_{i=-r+1}^{0} L[i]. \tag{23}$$

In the sequel, we refer to $K[r]$ as representing the data at time scale r. Set $K[0]=0$. Recursively applying Lindley's equation [30],

$$Q[0] = \max\{Q[-1]+K[1]-c, 0\}, \tag{24}$$

it is easily shown that

$$Q[0] = \max\{Q[-r]+K[r]-rc, K[r-1]-(r-1)c, \cdots, K[0]\}. \tag{25}$$

Since $Q[-r] \geq 0$ for all r, we must have

$$Q[0] \geq \sup_{r\in\mathbb{N}} (K[r]-rc). \tag{26}$$

Denoting by $-t$ the last instant when the queue was empty before time instant 0 (we set $-t=0$ if $Q[0]=0$), we obtain

$$Q[0] = K[t]-tc \leq \sup_{r\in\mathbb{N}} (K[r]-rc). \tag{27}$$

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

$$Q[0] = \sup_{r\in\mathbb{N}} \left(K[r]-rc\right). \tag{28}$$

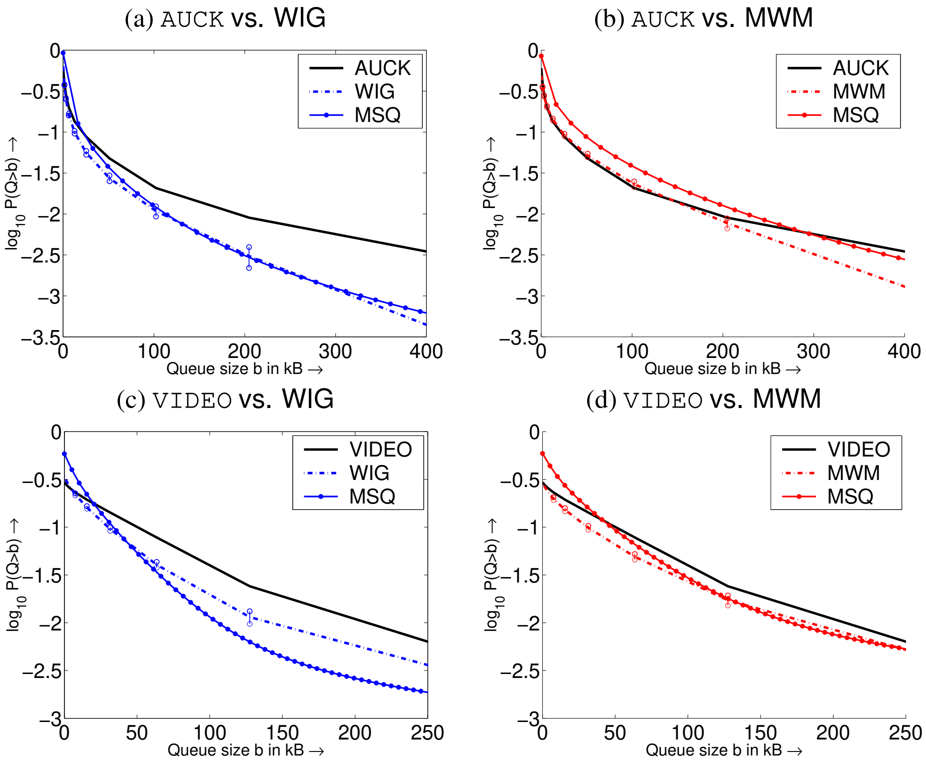


Fig. 7. Queuing performance of real data traces and synthetic WIG and MWM traces. *In (b), we observe that the MWM synthesis matches the queuing behavior of the* AUCK *data closely, while in (a) the WIG synthesis does not. In (c) and (d), we observe that both the WIG and the MWM match the queuing behavior of* VIDEO*. We also observe that the analytic multiscale queuing formula (MSQ) is a close approximation to the empirical queuing behavior for both synthetic traffic loads (both WIG and MWM). The link capacity we use for experiments with* AUCK *is* 2*Mbps (*72% *utilization) while that with* VIDEO *is* 69*Mbps (*77% *utilization). In all experiments in this paper, the confidence intervals correspond to a confidence level of* 95%.

In the sequel we will study exclusively $Q[i]$ at $i = 0$ and write $Q := Q[0]$ for ease of notation.

Note that (28) provides a direct link between the queue size Q and the aggregates 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.

We make the following three assumptions in our queuing analysis, which we justify in Sections IV-C to IV-E:

A1. Statistics at dyadic time scales accurately capture the effect of all time scales on the tail queue probability $\mathrm{P}[Q > b]$.

A2. When using a multiscale tree model, the joint distribution of the $K[2^j]$ (the traffic volume that has arrived in the past 2^j time instants) is well-modeled by that of the nodes $V_{j,2^j}$ on the right-edge of the tree (see Figure 2). In short, we can set $K[2^{n-j}] = V_{j,2^j-1}$.

A3. The events of large arrivals at the various dyadic time scales are independent.

Combining these assumptions, we claim that the following approximation is valid:

$$\begin{aligned}\mathrm{P}[Q < b] &\approx \mathrm{P}[\sup_{m\in\{0,\cdots,n\}} (K[2^m] - c2^m) < b] \\ &= \mathrm{P}[(K[2^m] - c2^m) < b,\, m \in \{0,\cdots,n\}] \\ &\approx \prod_{m=0}^{n} \mathrm{P}[K[2^m] < b + c2^m]. \end{aligned} \tag{29}$$

This leads us to propose the following *multiscale queuing formula* (MSQ) as an approximation to the true queue tail probability:

$$\boxed{\mathrm{MSQ}(b) := 1 - \prod_{i=0}^{n} \mathrm{P}\left[K[2^{n-i}] < b + c2^{n-i}\right].} \tag{30}$$

Note that *multiscale marginals*, i.e., the distributions of $K[2^i]$, $i = 0, \ldots, n$, enter (30) and not only $\mathrm{var}(K[2^i])$ (the correlation structure [33] of the process).

Before going into a more detailed argument supporting this approximation, we invite the reader to inspect Figure 7 for convincing numerical simulations that indicate that

$$\boxed{\mathrm{P}[Q > b] \approx \mathrm{MSQ}(b)} \tag{31}$$

for the bursty AUCK and VIDEO traces.

The only parameter not yet specified in the MSQ formula (30) is the depth n of the modeling tree, i.e., the difference between the coarsest and finest time scales. We will remark on the dependence on n in Section IV-F.

B. MSQ for the WIG and MWM

To derive explicit expressions for the MSQ, we need only compute the quantities $\mathrm{P}[E_i] = \mathrm{P}\left[K[2^{n-i}] < b + c2^{n-i}\right]$, which we obtain from the cumulative distribution functions (CDFs) of the marginals at multiple time scales. For tree-based models, these can be computed explicitly and simply from the tree-root $V_{0,0}$ and the multiscale innovations $Z_{j,k}$ or $M_{j,k}$ (see Figure 2(b) and Figure 3(a)).

WIG: The WIG model is a particularly simple case. Due to its additive character, $K[2^{n-i}] = V_{i,2^i-1}$ is the sum of independent Gaussian variables, namely the tree-root and the independent innovations (5). Thus, it is Gaussian itself with mean μ_i and variance σ_i^2. We set μ_i and σ_i^2 to equal the sample mean and variance of the nodes $V_{i,k}$ on the multiscale tree of the modeled traffic (see Section II-B). Alternatively, given the variances of $V_{0,0}$ and $Z_{j,k}$, we compute σ_i^2 using (6).

Let us denote the CDF of $V_{i,2^i-1}$ by Φ_{μ_i,σ_i}. Then,

$$\mathrm{MSQ}_{\mathrm{WIG}}(b) = 1 - \prod_{i=0}^{n} \Phi_{\mu_i,\sigma_i}(b + c2^{n-i}). \tag{32}$$

There are numerous approximations for Gaussian CDFs [39].

MWM: The multiplicative structure of the MWM model makes obtaining the multiscale marginals less straightforward, since $K[2^{n-i}]$ is the *product* of $i+1$ independent variables, i.e., the tree-root and the multiplicative multiscale innovations. If we impose lognormal distributions for the tree-root and innovations, then we obtain explicitly known lognormal distributions for the $K[2^{n-i}]$. However, this choice is not feasible, since we require the random innovations to be bounded between 0 and 1.

Recall from Section III that we settled for symmetrical beta distributions for the multipliers in the MWM model (see (9)). Using Fan's result [39, 43], we approximate the distribution of the product of independent beta random variables as another beta distribution with known parameters. For $K[2^{n-i}]$ (compare (8) and (46)) this yields

$$K[2^{n-i}] \;\sim\; a\,\beta(d_i, e_i), \tag{33}$$

with a a constant. The parameters d_i and e_i are given by

$$d_i = \zeta(\theta - \zeta^2)^{-1}(\zeta - \theta), \qquad e_i = d_i(1-\zeta)/\zeta, \tag{34}$$

where $\zeta = 2^{-i}$ and

$$\theta = \prod_{j=-1}^{i-1} \frac{(p_j+1)}{2(2p_j+1)}. \tag{35}$$

This approximation of the distribution of $K[2^{n-i}]$ matches the mean and variance exactly and closely approximates the first 10 moments [43]. The parameters p_j and a are obtained through the fitting procedure of Section III-B.

Denote the CDF of $\beta_{0,M}(d_i, e_i)$ by B_{M,d_i,e_i} (see [39] for numerical approximations). Then, we obtain

$$\mathrm{MSQ}_{\mathrm{MWM}}(b) = 1 - \prod_{i=0}^{n} \mathrm{B}_{M,d_i,e_i}(b + c2^{n-i}). \tag{36}$$

In the next three sections, we study the assumptions of our queuing analysis (see Section IV-A) in detail.

C. Restriction to dyadic time scales (A1)

First we restrict the supremum in (28) to time scales that appear naturally in a multiscale representation, i.e., the *dyadic time scales*:

$$Q_D := \sup_{m\in\{0,\cdots,n\}} (K[2^m] - c2^m). \tag{37}$$

The first approximation of our analysis is then

$$\textbf{A1}: \qquad \mathrm{P}\left[Q > b\right] \approx \mathrm{P}\left[Q_D > b\right]. \tag{38}$$

Clearly, $Q_D \le Q$ and $\mathrm{P}[Q > b] \ge \mathrm{P}[Q_D > b]$. To justify **A1** we must argue for using in (37)

1. only time scales smaller than 2^n, and
2. only dyadic time scales.

We assume that 2^n is larger than the longest busy period of the queue. This justifies using only time scales smaller than 2^n.

To justify using only dyadic time scales, we employ the notion of a *critical time scale* (CTS) [14–16]. This notion provides close approximations to both $\mathrm{P}[Q > b]$ and $\mathrm{P}[Q_D > b]$ that are easily computable. The CTS is defined as

$$r^* := \arg\sup_{r\in\mathbb{N}} \mathrm{P}\left[K[r] - cr > b\right], \tag{39}$$

and the *critical time scale queue* (CTSQ) is defined as

$$\mathrm{CTSQ}(b) := \mathrm{P}\left[K[r^*] - cr^* > b\right]. \tag{40}$$

It has been shown [14–16] that

$$\mathrm{CTSQ}(b) \approx \mathrm{P}\left[Q > b\right]. \tag{41}$$

Clearly, we have

$$\mathrm{CTSQ}(b) \le \mathrm{P}\left[Q > b\right]. \tag{42}$$

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

$$r_D^* := \arg\sup_{m\in\{0,\cdots,n\}} \mathrm{P}\left[K[2^m] - c2^m > b\right] \tag{43}$$

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

$$\mathrm{CDTSQ}(b) := \mathrm{P}\left[K[r_D^*] - cr_D^* > b\right]. \tag{44}$$

The CDTSQ is a computationally efficient substitute for the CTSQ, since it requires statistics at only a few dyadic time scales. Obviously,

$$\mathrm{CDTSQ}(b) \le \mathrm{P}\left[Q_D > b\right] \le \mathrm{P}\left[Q > b\right]. \tag{45}$$

We start by presenting experimental evidence to validate our first assumption **A1**. In Figure 8 we plot $\mathrm{P}[Q < b]$ against b for the WIG and the MWM with parameters corresponding to an fGn correlation structure (3) with $H = 0.8$. We observe that the CDTSQ is within an order of magnitude of $\mathrm{P}[Q < b]$ for both the WIG and the MWM. Clearly, from (45) this implies **A1**. When we vary H and the mean and variance of the models, the same result holds.

Finally, we give an intuitive explanation for **A1**. Dyadic time scales form only a small subset of $\mathbb{N}$, and so Q (28) and its approximation Q_D (37) could be very different. However dyadic time scales span the entire range of time scales of interest. Thus, r^* will be sandwiched between two dyadic time scales, one of which is likely to be r_D^*. Since the time scales r_D^* and r^* are "close", we can expect that CTSQ≈CDTSQ. From Figure 8, we observe that indeed CTSQ≈CDTSQ for the WIG.[5] From (41) and (45), this implies **A1**.

*D. Modeling joint distributions of traffic arrivals in the multi-scale tree (**A2**)*

To compute $\mathrm{P}[Q_D > b]$, it is necessary to know the joint distribution of the quantities $K[2^j]$ (see (37)), the total traffic that arrived in the past 2^j time instants. We claim here that the ideal time instant on a *tree model* at which to study the distribution of the queue size is the last time instant (or the right-most leaf of the tree). At this time instant, the quantities $K[2^j]$ and their joint distributions are explicitly available on the tree, which makes Q_D both more accurate and simple simultaneously.

For the sake of the argument, consider a queuing analysis at a time instant $t = 4k + 2$ at scale $j + 2$ (see Figure 2). Then $K[1] = V_{j+2,4k+2}$ and $K[2] = V_{j+2,4k+1} + V_{j+2,4k+2}$. Clearly, the dependence between $K[1]$ and $K[2]$ is not well represented by the tree model at this t, since $K[2]$ is not a node on the tree. Choosing $t = 4k + 1$, on the other hand, we have $K[1] = V_{j+2,4k+1}$, $K[2] = V_{j+2,4k} + V_{j+2,4k+1} = V_{j+1,2k}$, and $K[4] = V_{j+2,4k-2} + \ldots + V_{j+2,4k+1} = V_{j+1,2k-1} + V_{j+1,2k}$. Now, the dependence between $K[1]$ and $K[2]$ is explicitly modeled in the tree since $K[1]$ and $K[2]$ are tree nodes (see (5) and (8)). However, $K[4]$ is not a node on the tree.

In order to have all quantities $K[1], \ldots, K[2^m]$ as nodes on the tree, we must perform the queuing analysis at time instant $2^n - 1$ at scale n, where 2^{-n} is the time unit of interest. Since the largest scale captured in the tree is 2^n, we need $n \geq m$. This results in

$$K[2^i] = V_{n-i,2^{n-i}-1}, \quad \text{for } i = 0, \ldots, m. \tag{46}$$

Typically, we will work with $n = m$, since this allows us to exploit all levels of the tree.

*E. Approximate independence of large arrivals on dyadic time scales (**A3**)*

Obviously (as we just elaborated), the traffic volumes $K[1], K[2], K[3], \ldots$ are not independent. However, we argue here that very large arrivals over dyadic time intervals, i.e., the events[6] E_i^c, where

$$E_i := \{K[2^{n-i}] < b + c2^{n-i}\}, \tag{47}$$

can be assumed to be nearly independent of each other. We also show that this assumption is conservative in a certain sense.

The near independence of the E_i's is more intuitive than the independence of the E_i^c's, since it is based on the fact that the E_i's are highly probable events: Most of the numbers $\mathrm{P}[E_i]$ are nearly indistinguishable from 1 as our computations show. In the Appendix B we also show that $\mathrm{P}[E_i]$ converges exponentially fast to 1 as i increases. Thus, knowing that E_i has occurred (a highly probable event) does not tell us much about the occurrence of E_j ($j \neq i$). This implies that events E_i^c of large queue sizes are nearly independent as well, as we claimed.

The following Lemma, which we prove in Appendix A, helps us understand the implications of **A3** rigorously.

Lemma 1: Assume that the events E_i are of the form $E_i = \{S_i < b_i\}$, where $S_i = R_0 + \ldots + R_{i-1}$ for $1 \leq i \leq n$ and where $R_0, \ldots, R_n$ are independent, otherwise arbitrary random variables. Then, for $1 \leq i \leq n$, we have

$$\mathrm{P}[E_i | E_{i-1}, \ldots, E_0] \geq \mathrm{P}[E_i]. \tag{48}$$

For both the WIG and the MWM models, the events E_i (see (47)) can be written in the form required to apply Lemma 1 (see Appendix A). Using (48) we find

$$\begin{aligned}
\mathrm{P}[Q_D > b] &= 1 - \mathrm{P}[Q_D < b] \;=\; 1 - \mathrm{P}[\cap_{i=0}^n E_i] \quad (49)\\
&= 1 - \mathrm{P}[E_0] \prod_{i=1}^n \mathrm{P}[E_i | E_{i-1}, \ldots, E_0] \\
&\leq 1 - \prod_{i=0}^n \mathrm{P}[E_i] =: \mathrm{MSQ}(b). \quad (50)
\end{aligned}$$

If the E_i were indeed independent, then we would have equality: $\mathrm{MSQ}(b) = \mathrm{P}[Q_D > b]$. We conclude that the MSQ is a *conservative* approximation of the actual *dyadic* queue tail probability. In summary, we have

$$\mathrm{P}[Q > b] \geq \mathrm{P}[Q_D > b] \leq \mathrm{MSQ}(b). \tag{51}$$

F. Dependence on the tree depth

We have not as yet specified the coarsest and finest time scales to be used in the MWM. Clearly, the coarsest time scale must be chosen so that the model looks far enough into the past to ensure that the queue was empty in the covered time range with probability as close to 1 as desired or possible. We will assume this largest scale to be fixed for the remainder and discuss suitable choices for the finest time scale, i.e., the depth of the modeling tree n.

For clarity, let us index quantities computed on a tree of depth n with superscript (n), i.e., $\mathrm{MSQ}^{(n)}$, $K^{(n)}[2^j]$, etc. Also, for the ease of notation denote $\mathrm{MSQ}^{(n)}(b)$ by $\mathrm{MSQ}^{(n)}$.

As we increase the depth of the tree n, i.e., model the traffic at finer and finer time resolutions, the number of factors $\mathrm{P}[E_i^{(n)}]$ in the $\mathrm{MSQ}^{(n)}$ (see (30) or (50)) increases. Since $\mathrm{P}[E_i^{(n)}] \leq 1$, $\mathrm{MSQ}^{(n)}$ could potentially approach 1 as $n \to \infty$, especially if the $\mathrm{P}[E_i^{(n)}]$ were considerably smaller than 1. We claim this cannot happen.

First, note that $E_j^{(n)}$ is the event of large traffic arrivals at a level j steps below the tree-root (see Figure 2(a)). Having fixed

[5] Analytical formulas for the CTSQ of fGn make this comparison between CTSQ and CDTSQ possible for the WIG. For the MWM, explicit formulas for the marginal distributions at all time scales are not available.

[6] Here E_i^c denotes the complement of the event E_i.

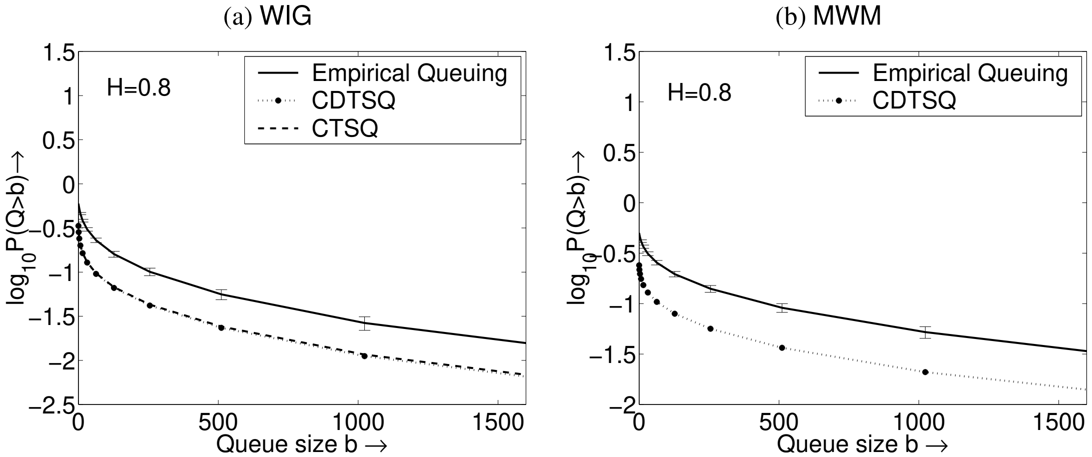


Fig. 8. Numerical justification that dyadic time scales provide sufficient information to describe queuing (**A1**). *In this experiment, synthetic WIG and MWM traces have an fGn correlation structure with Hurst parameter $H = 0.8$, mean $= 7$ units, and standard deviation $= 7$ units (see (7) and (15)). The link capacity was set to 10 normalized units. In (a) and (b) observe that for the WIG and MWM, the critical dyadic time scale queue (CDTSQ) is within an order of magnitude of the empirical tail queue probability of the traces. Also observe from (a) that the CDTSQ is almost identical to the critical time scale queue (CTSQ) for the WIG, which suggests its use as a computationally efficient substitute for the CTSQ. A CTSQ formula is not available for the MWM.*

the largest modeled time scale, these events are obviously the same for all n, i.e., $E_j^{(n)} = E_j^{(m)}$ for all $m, n \geq j$. In other words,

$$1 - \mathrm{MSQ}^{(n)} = \prod_{i=0}^{n} \mathrm{P}\left[E_i^{(n)}\right] = \prod_{i=0}^{n} \mathrm{P}\left[E_i^{(i)}\right]. \qquad (52)$$

The following Lemma, which we prove in Appendix B, says that $\mathrm{P}\left[E_n^{(n)}\right] \to 1$ fast enough that the MSQ does not trivially converge to 1. Moreover, the Lemma provides an error bound for neglecting finer time scales. To this end, let us introduce the ideal infinite-resolution MSQ:

$$\mathrm{MSQ}^{(\infty)} := \lim_{n\to\infty} \mathrm{MSQ}^{(n)} = 1 - \prod_{n=1}^{\infty} \mathrm{P}\left[E_n^{(n)}\right]. \qquad (53)$$

We also require the notion of a *threshold scale* N: Let N be such that

$$\mathrm{P}\left[E_j^{(j)}\right] \geq 1 - 2^{-j}, \ \forall\, j \geq N. \qquad (54)$$

A practical value for N is that time scale where

$$\max_k V_{N,k} < b \qquad (55)$$

Lemma 2: For the MWM model, a threshold scale N as in (54) always exists. Furthermore

$$\begin{aligned} \mathrm{MSQ}^{(N)} &\leq \mathrm{MSQ}^{(\infty)} \qquad (56)\\ &\leq \mathrm{MSQ}^{(N)} \cdot (1 - 2^{-N})^2 + 2^{-N+1}. \end{aligned}$$

V. Impact of multiscale marginals on queuing

Understanding the impact of traffic loads on queuing delay and loss statistics is fundamental to network engineering and control. Analytical queuing formulas [1, 5, 6] can significantly influence the design of networks and lead to novel algorithms for control and estimation [31].

LRD is one crucial property of traffic that impacts queuing. The tail queue distribution of fGn traffic fed into an infinite buffer with constant service rate [5, 6] decays *asymptotically* with queue size as

$$\mathrm{P}\left[Q > b\right] \simeq \exp(-\gamma b^{2-2H}), \qquad (57)$$

where γ is a positive constant that depends on the queue service rate. Clearly (57) reveals that $\mathrm{P}[Q > b]$ decays like a Weibullian law for fGn with $H > 1/2$, i.e., much slower than the exponential decay (corresponding to $H = 1/2$) predicted by short-range dependent (SRD) classical models [3].

For small queue sizes b, however, the single parameter H does not describe $\mathrm{P}[Q > b]$ accurately. The LRD parameter H captures only the asymptotic decay with time scale of the *variance* of traffic. More recent work that refines (57) indicates that traffic characteristics at the CTS (39) impacts queuing more than H [13–16, 44, 45]. This implies that the variance of traffic at one particular time scale impacts $\mathrm{P}[Q > b]$ more than H (see Section IV-B).

Here, we move beyond second-order statistics (i.e., variance at multiple time scales) and demonstrate the impact of the entire *marginal distribution* of traffic at different time scales on queuing through simulation experiments and the queuing analysis we developed in Section IV.

Our experiments simulate single-server FIFO queues of infinite buffer lengths fed with real as well as synthetic traffic traces. By comparing the queuing behavior of WIG and MWM traces that have the same second-order statistics but different marginals, we establish the influence of marginals on queuing. Observe from Figures 7(a) and (b), where we used the WAN traffic trace `AUCK`, that the real and synthetic traces exhibit asymptotic Weibullian tail queue probabilities, in agreement with the theoretical findings for LRD traffic (compare (57)). However, apart from this asymptotic match, the MWM is much closer to the queuing behavior of the real trace. The link capacity we use is 2Mbps, resulting in a utilization of 72%.

In the experiments with `VIDEO` (see Figures 7(c) and (d)), which is much closer to a Gaussian process than `AUCK`, we observe that both the WIG and MWM closely match the correct queuing behavior. This confirms the influence of marginals and also reassures us that the MWM is flexible enough to model Gaussian traffic. Gaussian-like traffic, which must be positive, necessarily has a mean at least comparable to its standard deviation. Since for a large mean to standard deviation ratio the lognormal and Gaussian distributions resemble each other closely (see Figure 6) , the approximately lognormal MWM is suitable for Gaussian traffic [26]. The link capacity we use is 69Mbps, which corresponds to a utilization of 77%.

Accepting the MSQ as a close approximation to the actual tail queue probabilities, a closer look at (30) unravels how the marginals affect queue sizes (recall Figure 7). For traffic with heavier tailed marginals, the terms $\mathrm{P}\left[K[2^i] < b + c2^i\right]$ are smaller and the MSQ larger. Since the approximately lognormal MWM marginals are more heavy tailed than the Gaussian WIG marginals, the MWM has a larger MSQ than the WIG. This is intuitively reasonable: larger bursts of traffic at different time scales lead to larger queue sizes. Recall from the histograms of traffic (see Figure 5) and the multifractal spectra (see Figure 4 that the MWM closely models the large bursts of real traffic well while the WIG does not. Consequently, the MWM captures the queuing behavior of the "spiky" `AUCK` trace (see Figure 1), while the WIG does not. However, in the case of `VIDEO`, which shows marginals much closer to Gaussian (see Figure 6), both the WIG and MWM perform similarly.

Finally, the MSQ also explains why Gaussian LRD traffic gives rise to longer queues than Gaussian SRD traffic (assuming the same mean and variance at the finest time scale). The LRD traffic has a higher variance than the SRD traffic at multiple time scales. The LRD traffic thus has heavier tailed marginals at multiple time scales and thus has a larger MSQ.

VI. Conclusions

The importance of multiscale models that capture the scaling properties of traffic loads has now been well recognized [2, 26, 27, 46]. In this paper, we used the multiscale Gaussian WIG and non-Gaussian MWM to demonstrate the impact of the marginals of traffic at multiple time scales on queuing.

Both the WIG and the MWM are built on binary trees, which allow fast $O(N)$ algorithms for synthesizing N-point data sets. 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. We can reduce the number of parameters further by developing a "linear" parametric characterization of the correlation decay reminiscent of the LRD parameter H.

The main contribution of this paper is our *multiscale queuing* (MSQ) approach, which provides a closed form queuing formula for tree-based models valid for any buffer size. Unlike earlier work on queuing of LRD traffic [5,6], our formula is non-asymptotic and takes into account the entire CDF 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, these models are now viable for numerous networking applications, including congestion control, admission control and cross-traffic estimation techniques [31] as well as non-networking related fields requiring LRD models.

Second and most importantly, the MSQ is to our knowledge the first analytical tool for assessing the impact of multiscale marginals on queuing. Earlier queuing experiments have suggested that their influence on the queue length distributions of LRD traffic should not be neglected [36]. 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 the tail queue probability.

Thirdly, the MSQ closely approximates the queuing behavior of training data using statistics from just the dyadic time scales. This confirms that dyadic time scales, though few in number, effectively capture the queuing behavior of traffic.

Our current research is aimed at making the MWM and MSQ practical for numerous applications. We have obtained encouraging results in cross-traffic estimation [31]. 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 and the MSQ in real-time network protocols and control algorithms seems very promising.

Appendix A: Proof of Lemma 1

We first spell out some notation. By f_U and F_U we denote the probability density function (PDF) and CDF, respectively, of a random variable U. Furthermore, we denote by $F_{U|E}(u)$ the CDF of U conditioned on knowing the event E. For convenience, let us introduce the auxiliary random variables $Y_0 := U_0 := S_0 := R_0$,

$$Y_i := S_i | E_{i-1}, \ldots, E_0 \text{ and } U_i := S_i | E_i, \ldots, E_0, \;\; i \geq 1. \tag{58}$$

To prove the Lemma, it is enough to show that

$$F_{Y_i}(r) \geq F_{S_i}(r) \tag{59}$$

$\forall\, r \in \mathbb{R}$ and $\forall\, i$ and then set $r = b_i$.

We prove (59) by induction. First note that $F_{Y_0}(r) \geq F_{S_0}(r)$. Next, we assume that (59) holds for i and show that it holds also for $i+1$. Bayes' rule yields

$$F_{U_i}(r) = \left\{ \begin{array}{ll} \frac{F_{Y_i}(r)}{F_{Y_i}(b_i)}, & \text{if } r \leq b_i \\ 1, & \text{otherwise} \end{array} \right\} \geq F_{Y_i}(r). \tag{60}$$

The key to the proof is to note that $Y_{i+1} = U_i + R_{i+1}$, where R_{i+1} is independent of S_j and hence of E_j for $j \leq i$. In short, R_{i+1} is independent of U_i. This fact and (59) and (60) allow us to write

$$\begin{aligned} F_{Y_{i+1}}(r) &= \mathrm{P}\left[U_i + R_{i+1} < r\right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{r - r_{i+1}} f_{U_i}(U_i) f_{R_{i+1}}(r_{i+1}) \,\mathrm{d}u_i \,\mathrm{d}r_{i+1} \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} F_{U_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1})\,\mathrm{d}r_{i+1} \\
&\geq \int_{-\infty}^{\infty} F_{Y_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1})\,\mathrm{d}r_{i+1} \\
&\geq \int_{-\infty}^{\infty} F_{S_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1})\,\mathrm{d}r_{i+1} \\
&= \mathrm{P}\left[S_i + R_{i+1} < r\right] \\
&= F_{S_{i+1}}(r). \qquad (61)
\end{aligned}
$$

This proves the claim by induction. $\diamondsuit$

Let us now show that Lemma 1 applies to the WIG and the MWM for the events E_i as given in (47). To this end we need only show that these E_i can be written in the appropriate form. Recall that according to (46) we have $K[2^{n-i}] = V_{i,2^i-1}$ for both models.

WIG: Recall that the WIG uses additive innovations $Z_{j,k}$ arranged on a tree as in Figure 2. It is immediate from (5) that $K[2^{n-i}]$ becomes

$$K[2^{n-i}] = V_{i,2^i-1} = 2^{-i}V_{0,0} - \sum_{j=0}^{i-1} 2^{j-i} Z_{j,2^j-1}. \qquad (62)$$

It suffices, thus, to set $b_i = 2^i b + 2^n c$, $R_0 = V_{0,0}$ and $R_i = -2^{i-1}Z_{i-1,2^{i-1}-1}$.

MWM: The MWM employs the same tree structure as the WIG, however, with multiplicative innovations $M_{j,k}$. Recalling (8), $K[2^{n-i}]$ becomes

$$K[2^{n-i}] = V_{i,2^i-1} = V_{0,0}\prod_{j=0}^{i-1}(1 - M_j). \qquad (63)$$

Taking logarithms, it is a simple task to write the events E_i in the required form, this time by setting $b_i = \ln(b + c2^{n-i})$, $R_0 = \ln(V_{0,0})$, and $R_i = \ln(1 - M_{i-1})$.

Appendix B: Proof of Lemma 2

We start by showing the existence of the threshold scale defined in (54).

The events $E_n^{(n)}$ depend on the buffer size $b^{(n)}$, the link capacity $c^{(n)}$ at time scale 2^{-n}, and the arriving workload $K^{(n)}[2^{n-i}]$. First of all, the buffer is always the same, and so $b^{(n)} = b$ independently of the depth n of the tree. Second, $c^{(n)} = \tilde{c}2^{-n}$, where $\tilde{c}$ is the total number of bytes that can be emptied from the queue in one time unit corresponding to the coarsest scale. Third and finally, according to (8) and (9), $K^{(n)}[2^{n-i}] = V^{(n)}_{i,2^i-1}$, which is in distribution equal to $aM_{-1}\prod_{j=0}^{i-1}(1 - M_j)$. Note, that we do not need to indicate the depth n of the tree, since the trees grow "downwards" as n increases. Indeed, the multipliers at scale i do not change with n, since new ones are added at the bottom of the tree. Together with the fact that the M_i are symmetrical random variables, this yields

$$\mathrm{P}\left[E_n^{(n)}\right] = \mathrm{P}\left[aM_{-1}\ldots M_{n-1} < b + \tilde{c}2^{-n}\right]. \qquad (64)$$

For short, let us write $D_i := \log_2(M_i)$ and

$$\alpha_n := -\frac{1}{n}\log_2(b/a + \tilde{c}2^{-n+1}/a). \qquad (65)$$

For any $q > 0$, a simple application of the Jensen inequality and independence yields the Chernoff bound:

$$
\begin{aligned}
1 - \mathrm{P}\left[E_n^{(n)}\right] &= \mathrm{P}\left[-(1/n)(D_{-1} + \ldots + D_{n-1}) < \alpha_n\right] \\
&= \mathrm{P}\left[2^{(q(D_{-1}+\ldots+D_{n-1}))} > 2^{-qn\alpha_n}\right] \\
&\leq \frac{\mathbb{E}[2^{(q(D_{-1}+\ldots+D_{n-1}))}]}{2^{-qn\alpha_n}} \\
&= 2^{n(-T^{(n)}(q) - 1 + q\alpha_n)}. \qquad (66)
\end{aligned}
$$

Here we set (analogous to (21))

$$T^{(n)}(q) = -1 - (1/n)\sum_{i=-1}^{n-1}\log_2\mathbb{E}[M_i^q]. \qquad (67)$$

Now, taking logarithms and minimizing over $q > 0$ yields

$$
\begin{aligned}
(1/n)\log_2\left(1 - \mathrm{P}\left[E_n^{(n)}\right]\right) &\leq \inf_{q>0}\left(q\alpha_n - T^{(n)}(q)\right) - 1 \\
&= -1 + \left(T^{(n)}\right)^*(\alpha_n) \leq -1 \qquad (68)
\end{aligned}
$$

provided that α_n is small enough that $\left(T^{(n)}\right)^*(\alpha_n) < 0$. Here $\left(T^{(n)}\right)^*$ is the Legendre transform of $T^{(n)}$ (see (19)). We refer here to Figure 4, which recalls the concave shape of T^*, with the smaller zero being positive but smaller than 1. Also, we point out that α_n decreases to 0 and that $T^{(n)}$ converges to T, implying that the zero of $\left(T^{(n)}\right)^*$ will not change greatly once n is large.

We can thus assume that $\left(T^{(n)}\right)^*(\alpha_n)$ is negative for all n larger than some critical N, which is the condition needed for completing the proof rigorously. For $n \geq N$, we have then $1 - \mathrm{P}[E_n^{(n)}] \leq 2^{-n} \leq 2^{-N}$ which proves (54). Choosing $h_0 = \log_2(1-2^{-N})/(-2^{-N})$, we guarantee that $\log_2 \mathrm{P}[E_n^{(n)}] \geq -h_0\left(1 - \mathrm{P}[E_n^{(n)}]\right)$ for all $n \geq N$. We conclude that

$$
\begin{aligned}
\log_2\prod_{n=N}^{N'}\mathrm{P}\left[E_n^{(n)}\right] &\geq -h_0\sum_{n=N}^{N'}\left(1 - \mathrm{P}\left[E_n^{(n)}\right]\right) \\
&\geq -h_0\sum_{n=N}^{N'}2^{-n} \geq -h_0\sum_{n=N}^{\infty}2^{-n} \\
&\geq -h_0 2^{-N+1}. \qquad (69)
\end{aligned}
$$

Thus we may estimate the "neglected terms" in $\mathrm{MSQ}^{(N)}$ by $1 \geq \prod_{n=N}^{\infty}\mathrm{P}[E_n^{(n)}] \geq 2^{-h_0 2^{-N+1}}$, which leads to

$$
\begin{aligned}
1 - \mathrm{MSQ}^{(N)} &\geq 1 - \mathrm{MSQ}^{(\infty)} \\
&\geq \left(1 - \mathrm{MSQ}^{(N)}\right)2^{-h_0 2^{-N+1}} \\
&= \left(1 - \mathrm{MSQ}^{(N)}\right)(1 - 2^{-N})^2. \qquad (70)
\end{aligned}
$$

Note that (70) and (56) are equivalent. $\diamondsuit$

A practical way of choosing a threshold scale N is to apply the multifractal formalism (18): the condition $\left(T^{(n)}\right)^*(\alpha_n) < 0$ means in this context that no exponent α_n is observed, since $N_n(\alpha_n) = 0$. More precisely, all observed coarse Hölder exponents at scale 2^{-n} (or n levels below the tree-root) $\alpha(t) = -(1/j)\log_2|V_{n,k}|$ are larger than α_n. With (65), this translates immediately to $b + \widetilde{c}2^{-n+1} \geq |V_{n,k}|$. A conservative condition, which is easy to check and which will ensure that $\left(T^{(n)}\right)^*(\alpha_n) < 0$ for $n \geq N$, is then to require that

$$b \geq |V_{N,k}|. \tag{71}$$

which agrees with (55).

Acknowledgements

Thanks to Edward Knightly (Rice University) for many informative discussions.

References

[1] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," *Proc. IEEE INFOCOM 2000*, March 2000.

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

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

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

[5] 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.

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

[7] N. Likhanov, B. Tsybakov, and N. Georganas, "Analysis of an ATM buffer with self-similar input traffic," *Proc. IEEE INFOCOM*, pp. 985–992, 1995.

[8] J. Choe and N. B. Shroff, "Queueing analysis of high-speed multiplexers including long-range dependent arrival processes," *Proc. IEEE INFOCOM*, pp. 617–624, March 1999.

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

[10] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic. Evidence and possible causes," *IEEE/ACM Trans. on Networking*, vol. 5, pp. 835–846, December 1997.

[11] 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. on Networking*, vol. 5, pp. 71–86, Feb. 1997.

[12] N. Duffield, "Economies of scale for long-range dependent traffic in short buffers," *Telecommunication Systems*, vol. 7, pp. 267–280, 1997.

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

[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] 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.

[16] 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.

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

[18] 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.

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

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

[21] S. Ma and C. Ji, "Modeling video traffic in the wavelet domain," *Proc. of INFOCOM*, pp. 201–208, Mar. 1998.

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

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

[24] S. Bates and S. McLaughlin, "The estimation of stable distribution parameters from teletraffic data," *IEEE Transactions on Signal Processing*, vol. 48, pp. 865–870, March 2000.

[25] NLANR, "Auckland-II trace archive." Available at http://moat.nlanr.net/Traces/Kiwitraces/. Trace 20000125-143640, corresponding to 3:11:28 hours of mostly TCP traffic.

[26] R. H. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. on Info. Theory*, vol. 45, pp. 992–1018, April 1999.

[27] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic," *Proc. ACM/SIGCOMM 98*, vol. 28, pp. 42–55, 1998.

[28] A. C. Gilbert, W. Willinger, and A. Feldmann, "Scaling analysis of random cascades, with applications to network traffic," *IEEE Trans. on Info. Theory*, April 1999.

[29] R. H. Riedi and W. Willinger, *Self-similar Network Traffic and Performance Evaluation*, ch. Toward an Improved Understanding of Network Traffic Dynamics, pp. 507–530. Wiley, 2000. K. Park and W. Willinger eds.

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

[31] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk, "Multifractal cross-traffic estimation," *Proc. of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Sept. 2000.

[32] 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.

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

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

[35] 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, 2000.

[36] 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.

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

[38] K. E. Timmerman and R. D. Nowak, "Multiscale modeling and estimation of Poisson processes with application to medical imaging," *IEEE Trans. on Info. Theory*, vol. 45, pp. 846–863, April 1999.

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

[40] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP throughput: A simple model and its empirical validation," in *Proc. ACM-SIGCOMM*, 1998.

[41] R. H. Riedi, "Multifractal processes," *Technical Report, ECE Dept. Rice Univ., TR 99-06*. "Long range dependence: Theory and applications," eds. Doukhan, Oppenheim and Taqqu (2000). Available at www.dsp.rice.edu.

[42] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," Tech. Rep. 101, University of Wuerzburg. Institute of Computer Science Research Report Series, February 1995.

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

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

[45] 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.

[46] J. Lévy Véhel and R. Riedi, "Fractional Brownian motion and data traffic modeling: The other end of the spectrum," *Fractals in Engineering*, pp. 185–202, Springer 1997.