To appear IEEE Transactions on Networking, 2006

Multiscale Queuing Analysis

Vinay J. Ribeiro, *Member, IEEE,* Rudolf H. Riedi, *Member, IEEE,* and Richard G. Baraniuk, *Fellow, IEEE*

***Abstract*— This paper introduces a new multiscale framework for estimating the tail probability of a queue fed by an arbitrary traffic process. Using traffic statistics at a small number of time scales, our analysis extends the theoretical concept of the critical time scale and provides practical approximations for the tail queue probability. These approximations are non-asymptotic; that is they apply to any finite queue threshold. While our approach applies to any traffic process, it is particularly apt for long-range-dependent (LRD) traffic. For LRD fractional Brownian motion, we prove that a sparse exponential spacing of time scales yields optimal performance. Simulations with LRD traffic models and real Internet traces demonstrate the accuracy of the approach. Finally, simulations reveal that the marginals of traffic at multiple time scales have a strong influence on queuing that is not captured well by its global second-order correlation in non-Gaussian scenarios.**

***Index Terms*— Multiscale, queuing, long-range-dependence, fractional Brownian motion, multifractals, wavelets, critical time scale, network provisioning, admission control, marginals**

I. Introduction

We model a router queue as an infinite length queue with constant service rate [2] and study the probability that the queue size Q exceeds a threshold b, $\mathbb{P}\{Q > b\}$, also called the *tail queue probability*. The tail queue probability is a useful metric for various applications including techniques for maintaining low packet queuing delays and jitter at router queues such as admission control and network provisioning [3], [4]. Low network delays are critical for the viability of real-time streaming media applications for telephony, telemedicine, videoconferencing, economic transactions etc.

We can predict $\mathbb{P}\{Q > b\}$ in several ways. First, we can model network traffic using different processes (also called *traffic models*) and use any exact formula for $\mathbb{P}\{Q > b\}$ that is available. Second, in case exact results are unavailable for a particular process we can employ analytical results that only approximate $\mathbb{P}\{Q > b\}$, which we call *queuing approximations*. Third, if modeling traffic with a standard random process is cumbersome or inadequate then we can predict $\mathbb{P}\{Q > b\}$ directly from measured traffic statistics. In such a scenario it is desirable to use a small number of traffic statistics in order to reduce data acquisition and computational requirements.

V. J. Ribeiro and R. H. Riedi are with the Department of Statistics and R. G. Baraniuk is with the Department of Electrical and Computer Engineering at Rice University, 6100 Main Street, Houston, TX 77005, USA (email: {vinay, riedi, richb}@rice.edu; Web: spin.rice.edu).

Supported by NSF grants ANI–9979465, ANI–0099148, and ANI-0338856, DoE SciDAC grant DE-FC02-01ER25462, DARPA/AFRL grant F30602–00–2–0557, Texas ATP 003604-0036-2003, and the Texas Instruments Leadership University program.

A conference version of this paper with preliminary results was presented at INFOCOM 2000 [1].

In this paper we develop a new approach to queuing analysis called the *multiscale queuing analysis* that addresses the second and third scenarios mentioned above.

While our analysis is relevant to any traffic process we focus on processes with non-summable correlations (called long-range-dependence (LRD)) since LRD is a ubiquitous property of real-traffic [5].[1] Classical Poisson and Markov queuing techniques are unsuitable for LRD traffic which creates the need for new analytical tools. Up to now exact formulas for the queuing delay of LRD processes, other than for asymptotically large delays [6]–[8], have not been found and we are thus forced to use approximations.

To date, most approximations for the tail queue probability of queues fed with LRD processes have been based on the notion of the *critical time scale* [6]–[13]. Given a queue size threshold b, the critical time scale is the most likely amount of time it takes for the queue to fill up beyond b. While the critical time scale is a powerful theoretical tool, computing it directly from empirical measurements is impractical because this requires traffic statistics at *all* time scales.

By using traffic statistics at only a finite set of time scales, θ, our approach provides three practical approximations for $\mathbb{P}\{Q > b\}$: the *max* approximation, the *product* approximation, and the *sum* approximation. These have several important features:

- they apply to any finite queue threshold b, that is, they are non-asymptotic;
- they apply to any traffic model including non-stationary ones; and
- they are simple to employ because they require traffic statistics only at few time scales θ.

We prove numerous non-asymptotic error bounds, large-queue asymptotic results, and other bounds for the three approximations for different traffic models including fractional Brownian motion (fBm), fractional Gaussian noise (fGn), the wavelet-domain independent Gaussian model (WIG), and the multifractal wavelet model (MWM). We also compare the different approximations through numerical experiments.

Determining an appropriate candidate for θ is a key issue we address. The choice of θ involves a trade-off between the accuracy of the approximations and the requirements for computation and data acquisition. For example, a sparse θ decreases the accuracy of the max approximation but simultaneously requires the computation of traffic statistics and data acquisition at fewer time scales. We prove that the choice

[1]Note that standard LRD traffic processes typically become short-range dependent (i.e. non-LRD) with a special choice of parameters. Our results for LRD models also hold for these non-LRD cases.

of *exponential* time scales for θ is *optimal* with respect to this trade-off for a queue with fBm input traffic. A significant advantage of exponential time scales is their sparsity; just a few exponential time scales span a wide range. This result thus strongly recommends the use of traffic statistics at exponential time scales in queuing applications.

Traffic models such as fGn and the WIG suffice to capture the queuing behavior of traffic in Gaussian scenarios that can occur with traffic aggregation as on backbone links [11]. However they do not perform as well in non-Gaussian traffic scenarios. We term the distributions of traffic at different time scales as *marginals*. Through simulations with the WIG and the MWM which have different marginals we demonstrate the strong impact of marginals on queuing. This result supports similar findings in [12], [14].

Our main contributions are thus (i) a novel multiscale approach to queuing analysis that provides practical queuing approximations, (ii) optimality and error bounds related to the approximations for various traffic models, and (iii) the demonstration that marginals can strongly influence queuing behavior.

Paper Organization: Section II reviews previous work on the critical time scale. In Section III we present the multiscale queuing analysis of the paper and derive the various queuing approximations. Section IV describes the fBm, the fGn, the WIG, and the MWM traffic models. In Section V we prove the optimality of exponential time scales for fBm. Section VI proves that large buffer asymptotic results and Section VII proves bounding results for the different queuing approximations. Section VIII demonstrates the accuracy of the approximations through simulations with Internet and synthetic model data and also demonstrates the impact of marginals on queuing. We conclude in Section IX. The proofs of various results are placed in the Appendix.

II. Review of Critical Time Scale Analysis

In this section we review previous work on the critical time scale queuing analysis to set the stage for our multiscale queuing analysis in subsequent sections.

A. Queue size as a multiscale function

Consider a continuous-time fluid queue with constant service rate c with traffic process X_t, $t \in \mathbb{R}$ as input. We refer to

$$K_t[\tau] := \int_{t-\tau}^{t} X_\omega \mathrm{d}\omega \tag{1}$$

as the traffic process at *time scale* τ. To avoid notational ambiguity we occasionally add superscripts such as in $K_t^{\{X\}}[\tau]$ to identify the traffic process. For the ease of notation we drop the subscript t for all time-invariant quantities.

Assuming that the queue was empty at some time instant prior to t, the *queue size* Q_t equals the difference between the total traffic that arrived at the queue and the total traffic serviced since the time instant the queue was last empty. This is succinctly captured by Reich's formula

$$Q_t := \sup_{\tau>0} \left(K_t[\tau] - c\tau\right). \tag{2}$$

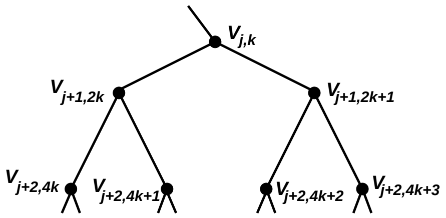


Fig. 1. Multiscale tree representation of a traffic trace. *Nodes at each horizontal level in the tree correspond to the sum (aggregates) of the process in non-overlapping blocks of sizes of powers of two, with lower levels corresponding to smaller block sizes. Each node is the sum of its two child nodes.*

We address the requirement of an empty queue prior to t with mathematical rigor in Section IV-E.

A key interpretation of (2) is that Q_t equals a function of $K_t[\tau]$, the traffic process at all time scales τ. The question arises as to whether or not we can accurately approximate $\mathbb{P}\{Q > b\}$ using the distribution of $K_t[\tau]$ at a *single* time scale τ.

B. Critical time scale queuing approximation

Most proposed approximations of $\mathbb{P}\{Q > b\}$ for queues fed by LRD traffic are indeed based on a *single* time scale called the *critical time scale* [6]–[13]

$$\lambda_t(b) := \arg\sup_{\tau>0} \mathbb{P}\{K_t[\tau] - c\tau > b\}. \tag{3}$$

We term the associated queue tail approximation the *critical time scale approximation*

$$\begin{aligned} C_t(b) &:= \sup\nolimits_{\tau>0} \mathbb{P}\{K_t[\tau] - c\tau > b\} \\ &= \mathbb{P}\{K_t[\lambda_t(b)] - c\lambda_t(b) > b\}. \end{aligned} \tag{4}$$

Clearly $C_t(b)$ is a lower bound of $\mathbb{P}\{Q_t > b\}$ since by (2) $K_t[\lambda_t(b)] - c\lambda_t(b) \le Q_t$; thus

$$C_t(b) \le \mathbb{P}\{Q_t > b\}. \tag{5}$$

Earlier work based on large deviation theory has shown that $C_t(b)$ has the same log-asymptotic decay as $\mathbb{P}\{Q_t > b\}$ when $b \to \infty$ for a large class of input traffic processes including fBm [6], [7]. As the simulations in Section VIII demonstrate, $C_t(b)$ is also a good approximation for $\mathbb{P}\{Q_t > b\}$ for any finite b for fBm-fed queues. The intuition for the accuracy of $C_t(b)$ is that *"rare events occur in the most likely way."* In other words given that $\{Q_t > b\}$ is a rare event, if the queue size is conditioned to fill up greater than b then it does so in time $\lambda_t(b)$ in which this is most likely. That is, conditioned on $\{Q_t > b\}$, we have that Q_t is approximately equal to $K[\lambda_t(b)] - c\lambda_t(b)$.

While the critical time scale is a powerful tool that has advanced the state-of-the-art in queuing theory, using it in practice is not straightforward. First, consider the problem of computing $C_t(b)$ for a queue fed with an arbitrary process, solely from empirical traffic measurements. From (4) we see that we require the distribution of $K_t[\tau]$ for all possible τ. This is infeasible to obtain empirically. Even if we replace purely empirical schemes by techniques that use both empirical statistics and analytical models, similar computational problems may persist. For example if we use traffic models for which analytical expressions for $C_t(b)$ are unknown then we may have to employ computationally intensive algorithms to determine $C_t(b)$.

Second, say that we wish to compute the critical time scale approximation when two independent processes X and Y are multiplexed and input to a queue. Such a scenario often arises in admission control and network provisioning [3], [4]. Obtaining $C_t^{\{X+Y\}}(b)$ directly from the statistics of X and Y is again fraught with similar problems.

III. Multiscale Queuing Approximations

In this section we develop three new queuing approximations that do not have the computational problems that are associated with using the critical time scale approximation. A key factor that simplifies their computation is that they use traffic statistics only at a *fixed finite* set of time scales $\theta \subset \mathbb{R}_+$. Note that while some of our theoretical results are for countably infinite sets θ, in practice we always employ a truncated, finite set θ when computing the queuing approximations. We typically choose the set θ to span the range of time scales in which we expect the critical time scale $\lambda(b)$ to lie, for values of b relevant to a particular application.

A. Max approximation

In analogy to the queue size formula and the critical time scale (see (2) and (3)) define

$$Q_t^{[\theta]} := \sup_{\tau\in\theta} \left(K_t[\tau] - c\tau\right) \tag{6}$$

and

$$\lambda_t^{[\theta]}(b) := \arg\sup_{\tau\in\theta} \mathbb{P}\{K_t[\tau] - c\tau > b\} \tag{7}$$

for $\theta \subset \mathbb{R}_+$. This leads to the *max* approximation

$$\boxed{\begin{array}{rcl} M_t^{[\theta]}(b) & := & \sup_{\tau\in\theta} \mathbb{P}\{K_t[\tau] - c\tau > b\} \\ & = & \mathbb{P}\left\{K_t\left[\lambda_t^{[\theta]}(b)\right] - c\lambda_t^{[\theta]}(b) > b\right\}. \end{array}} \tag{8}$$

Comparing (4) to (8) we see that the max approximation is similar to the critical time scale approximation with the difference that the supremum is taken over a finite set in (8) instead of over all time scales in (4). From (4), (5), and (8) we have the bounds

$$M_t^{[\theta]}(b) \le C_t(b) \le \mathbb{P}\{Q_t > b\}. \tag{9}$$

We note from (2) and (6) that

$$Q_t = Q_t^{[\mathbb{R}_+]} \ge Q_t^{[\theta]} \tag{10}$$

and from (6), (8), and (10) that

$$M_t^{[\theta]}(b) \le \mathbb{P}\left\{Q_t^{[\theta]} > b\right\} \le \mathbb{P}\{Q_t > b\}. \tag{11}$$

The max approximation is a practical replacement for $C_t(b)$. Since the max approximation requires estimates of $\mathbb{P}\{K_t[\tau] - c\tau > b\}$ only for $\tau \in \theta$, the difficulties associated with computing $C_t(b)$ as we described earlier do not arise. First, consider the problem of obtaining the max approximation from empirical traffic measurements. We simply compute histograms of the traffic at time scales $\tau \in \theta$ and then estimate $\mathbb{P}\{K_t[\tau] - c\tau > b\}$. Second, consider the problem of computing the max approximation when two independent processes X and Y are multiplexed and input to a queue. By simply convolving the distributions of $K_t^{\{X\}}[t]$ and $K_t^{\{Y\}}[\tau]$ for $\tau \in \theta$ we obtain the corresponding distributions of $K_t^{\{X+Y\}}[\tau]$, which immediately give the max approximation.

B. Product and sum approximations

Two additional approximations of $\mathbb{P}\{Q_t > b\}$ based on the set of time scales θ are the *product* approximation

$$\boxed{P_t^{[\theta]}(b) := 1 - \prod_{\tau\in\theta} \mathbb{P}\{K_t[\tau] - c\tau < b\}} \tag{12}$$

and the *sum* approximation

$$\boxed{S_t^{[\theta]}(b) := \sum_{\tau\in\theta} \mathbb{P}\{K_t[\tau] - c\tau > b\}.} \tag{13}$$

Note that the product approximation equals $\mathbb{P}\left\{Q_t^{[\theta]} > b\right\}$ if the events $\{K_t[\tau] - c\tau > b\},\ \tau \in \theta$, are independent,[2] and that the sum approximation equals $\mathbb{P}\left\{Q_t^{[\theta]} > b\right\}$ if the same events are mutually exclusive.

C. Intuition for the accuracy of the approximations

The max, product, and sum approximations inherit the accuracy of the critical-time scale approximation while being practical. If there exists an element of θ close enough to the critical time-scale then $M_t^{[\theta]}(b)$ will be close to $C_t(b)$ (see (4) and (8)). Moreover, if a single probability term dominates the summation in (13), then the product and sum approximations will closely approximate $M_t^{[\theta]}(b)$ and hence $C_t(b)$. Simulations below in Section VIII demonstrate that the product and sum approximations are often closer to $\mathbb{P}\{Q_t > b\}$ than the max approximation.

IV. Traffic Models

This section describes four traffic models that we focus on in this paper. While all have been shown to model the LRD in real Internet traffic well, they differ in their ability to model other properties of traffic.

A. Fractional Brownian motion

Fractional Brownian motion (fBm) is the unique Gaussian process with stationary increments and the following scaling property for all $a > 0$, $t \in \mathbb{R}$, and $0 < H < 1$

$$B_{at} \overset{d}{=} a^H B_t. \tag{14}$$

The symbols "$\overset{d}{=}$", "var", $\mathbb{E}$ and "cov" denote *equality in distribution*, *variance*, *expectation*, and *covariance* respectively.

[2] If events $E_i,\ i \in \mathbb{N}$, are independent then so are their complements.

B. Fractional Gaussian noise

Fractional Gaussian noise (fGn) is the increment process of fBm. While fGn is stationary, fBm is itself non-stationary by definition. Denote the stochastic differential of B_t as $\Delta_t B$. We denote fGn by

$$G_t[\tau] := K_t^{\{\Delta_t B\}}[\tau] = B_t - B_{t-\tau}. \tag{15}$$

While it is difficult to define $\Delta_t B$ rigorously, its aggregate $K_t^{\{\Delta_t B\}}[\tau]$ is well-defined. Often one is interested only in the time series $\{G_{i\tau'}[\tau']\}_{i\in\mathbb{Z}}$ with τ' a constant time lag. From (14) and (15) we have that

$$K_t^{\{\Delta_t B\}}[t] = B_t \stackrel{d}{=} t^H B_1 \tag{16}$$

and thus

$$\mathrm{var}(G_{i\tau'}[\tau']) = \mathrm{var}\Big(K^{\{\Delta_t B\}}[\tau']\Big) = \sigma^2(\tau')^{2H} \tag{17}$$

where $\sigma^2 = \mathrm{var}\,(B_1)$. When $\frac{1}{2} < H < 1$, fGn is LRD.

C. Wavelet-domain independent Gaussian (WIG) model

The WIG is a Gaussian traffic model that is able to approximate fBm and fGn as well as processes with more general scaling than (14) and (17). It uses a *multiscale tree* to model traffic over the time interval $[0, T]$ [15], [16]. The nodes $V_{j,k}$ on the tree correspond to the total traffic in the time interval $[k2^{-j}T, (k+1)2^{-j}T]$, $k = 0, 1, \ldots, 2^j - 1$ (see Fig. 1).

Starting at node $V_{j,k}$, the WIG models nodes $V_{j+1,2k}$ and $V_{j+1,2k+1}$ using independent *additive* random innovations $Z_{j,k}$ through

$$\begin{aligned} V_{j+1,2k} &= (V_{j,k} + Z_{j,k})/2, \\ V_{j+1,2k+1} &= (V_{j,k} - Z_{j,k})/2. \end{aligned} \tag{18}$$

In practice one uses a WIG tree of finite depth n to obtain a discrete-time process $V_{n,k}$. The $Z_{j,k}$ have the same variance within each scale j, thus guaranteeing that $V_{n,k}$ is a first-order stationary process. The root $V_{0,0}$ and all $Z_{j,k}$ are Gaussian which ensures that all tree nodes are Gaussian.

To *fit* a traffic model means to choose its parameters either to match key statistics of observed traffic or to ensure that the model has certain prespecified statistical properties. Fitting the WIG involves choosing its parameters to obtain a required variance progression of var$(V_{j,k})$. The WIG can provide a Gaussian approximation for *any* stationary discrete-time process X; that is the WIG can be fit to obtain

$$\mathrm{var}(V_{n-j,k}) = \mathrm{var}\left(K^{\{X\}}[2^j]\right). \tag{19}$$

We will refer to a WIG model for which (19) holds as a "WIG model of X" in the rest of the paper.

D. Multifractal wavelet model (MWM)

The MWM is a *non-Gaussian* model based on a multiscale tree that, like the WIG, allows a more general scaling behavior of the variance of tree nodes than fGn [17]. Unlike the WIG, it ensures positivity at all time scales, an intrinsic property of real data traffic that is often ill approximated by Gaussian models. Setting $V_{0,0} \geq 0$ the MWM uses independent *multiplicative* innovations $U_{j,k} \in [0, 1]$ to model the two children of node $V_{j,k}$ through

$$\begin{aligned} V_{j+1,2k} &:= V_{j,k}U_{j,k}, \\ V_{j+1,2k+1} &:= V_{j,k}(1 - U_{j,k}). \end{aligned} \tag{20}$$

Because the product of independent random variables converges to a lognormal distribution by the central limit theorem, the nodes $V_{j,k}$ become approximately lognormal with increasing j.

Following [17], we model the $U_{j,k}$'s and $V_{0,0}$ as symmetric beta random variables. The tree node $V_{j,k}$ is thus the product of several independent beta random variables. Using Fan's result [18], we approximate the distribution of $V_{j,k}$ as another beta distribution with known parameters in order to compute different queuing approximations for the MWM.

Fitting the MWM involves choosing its parameters to obtain a required variance progression of var$(V_{j,k})$. The MWM can model *any* stationary discrete-time process X with *positive* autocovariance in the sense of (19).

While the WIG and MWM models are first-order stationary, they are not second-order stationary. This is apparent from Fig. 1. Observe that $V_{j+2,4k}$ and $V_{j+2,4k+1}$ have the same parent node while $V_{j+2,4k+1}$ and $V_{j+2,4k+2}$ do not. Thus the correlation of $V_{j+2,4k+1}$ with its two neighbors, $V_{j+2,4k}$ and $V_{j+2,4k+2}$, are different. Both models however have a time-averaged correlation structure that is close to the stationary process X that they model (see [16], [17] for details).

E. Queuing analysis setup for fBm, fGn, WIG, and MWM

We now state precisely the queuing setup for the fBm, fGn, WIG, and MWM models that we analyze in subsequent sections. We set the initial queue size to be empty to satisfy the sufficient condition for (2) to hold (see Section II-A).

In this paper all queuing results for queues with fBm input correspond to a continuous-time queue with service rate c, initial value $Q_0 := 0$, and $K_t[\tau] = K_t^{\{\Delta_t B + m\}}[\tau] = B_t - B_{t-\tau} + m\tau$. We have

$$\begin{aligned} \mathbb{P}\{Q_t > b\} &= \mathbb{P}\left\{\sup_{0\leq\tau\leq t}(K_t[\tau] - c\tau) > b\right\} \\ &\xrightarrow{t\to\infty} \mathbb{P}\left\{\sup_{\tau\geq 0}(K_0[\tau] - c\tau) > b\right\} \\ &=: \mathbb{P}\{Q_\infty > b\} \end{aligned} \tag{21}$$

where the limit holds because of stationarity of fBm increments and Lemma 13 (in the Appendix). We assume that $\widehat{c} := c - m > 0$ and study the quantity $\mathbb{P}\{Q_\infty > b\}$ as defined in (21).

For fGn, WIG, and MWM traffic we consider discrete-time queues that are initialized to $Q_0 := 0$ and evolve according to

$$Q_{t+1} = \max\left(Q_t + X_t - \overline{c}, 0\right), \quad t \in \mathbb{Z}_+. \tag{22}$$

Defining $K_t[\tau] := \sum_{k=t-\tau}^{t-1} X_k$ for $\tau = 1, 2, \ldots, t$ and $t = 1, 2, \ldots, \infty$, and $K_t[0] := 0$ we have

$$Q_t := \max_{\tau=0,1,\ldots,t}\left(K_t[\tau] - \overline{c}\tau\right). \tag{23}$$

For fGn we set $\overline{c} = c\tau'$ and $X_t = G_{t\tau'}[\tau']$ for $t = 0, 1, \ldots, \infty$. We study the quantity $\mathbb{P}\{Q_\infty > b\}$ which is defined as in (21) with the difference that τ and t take integer values.

For the WIG and MWM we consider Q_t only for $t = 0, 1, \ldots, 2^n - 1$ with $\overline{c} = \widetilde{c}^{(n)} := cT2^{-n}$ where n is the depth of the multiscale tree. Here $X_t = V_{n,t}$. We assume that

$$\mathbb{E}(V_{n,k}) < \widetilde{c}^{(n)} \tag{24}$$

and study $\mathbb{P}\{Q_t > b\}$ which is a time-varying quantity.

For the fGn, WIG, and MWM models $K_t[\tau]$ is only defined for $\tau = 0, 1, \ldots, t$. For these models we define $M_t^{[\theta]}(b)$, $P_t^{[\theta]}(b)$, and $S_t^{[\theta]}(b)$ as in (8), (12), and (13) except that we replace θ by $\theta \cap \{0, 1, \ldots, t\}$.

V. Optimality of Exponential Time Scales for the Max Approximation of an fBm queue

Comparing (4) and (8) we see that the more dense θ is in $\mathbb{R}_+$, the closer the max approximation $M^{[\theta]}(b)$ is to the critical time scale approximation $C(b)$. However, we simultaneously have to acquire data at more time scales, and the computational cost of the max approximation increases (see (8)). In this section we prove that the sets of exponential time scales

$$\theta_\alpha := \{\alpha^k : k \in \mathbb{Z}\}, \quad \alpha > 1, \tag{25}$$

optimally balance this trade-off in accuracy versus computational cost.

First, for a queue with fBm input we first define a metric to characterize the accuracy of $M^{[\theta]}(b)$. Second, we prove that θ_α is the most sparse of all sets θ that satisfy a particular accuracy criterion for $M^{[\theta]}(b)$. Third, we obtain a non-asymptotic bound on the error of $M^{[\theta_\alpha]}(b)$ in approximating $C(b)$. This bound proves that $M^{[\theta_\alpha]}(b)$ accurately approximates $C(b)$ for a wide range of α.

A. Metric to characterize accuracy of $M^{[\theta]}(b)$

Consider a queue fed by fBm traffic as described in Section IV-E. Then for $\tau > 0$, using (14) it is easily shown that [9]

$$\mathbb{P}\{K[\tau] - c\tau > b\} \quad = \quad \Phi\left(g(b, \tau)\right) \tag{26}$$

where

$$g(b, \tau) := \frac{b + \widehat{c}\tau}{\sigma\tau^H} = \frac{b + (c - m)\tau}{\sigma\tau^H} \tag{27}$$

and Φ is the complementary cumulative distribution function of a zero mean unit variance Gaussian random variable. From (4) and (8) we have

$$C(b) = \sup_{\tau > 0} \Phi\left(g(b, \tau)\right) = \Phi\left(\inf_{\tau > 0} g(b, \tau)\right) \tag{28}$$

and

$$M^{[\theta]}(b) = \sup_{\tau \in \theta} \Phi\left(g(b, \tau)\right) = \Phi\left(\inf_{\tau \in \theta} g(b, \tau)\right). \tag{29}$$

Given a range of time scales $\mathcal{T}$, we characterize the accuracy of $M^{[\theta]}(b)$ in terms of the following metric

$$h_\theta(\mathcal{T}) := \sup_{b:\lambda(b) \in \mathcal{T}} \frac{\inf_{\tau \in \theta} g(b, \tau)}{\inf_{\tau > 0} g(b, \tau)} = \sup_{b:\lambda(b) \in \mathcal{T}} \frac{\inf_{\tau \in \theta} g(b, \tau)}{g(b, \lambda(b))}. \tag{30}$$

Intuitively, the closer $h_\theta(\mathcal{T})$ is to 1 the tighter we can bound the error of $M^{[\theta]}(b)$ in approximating $C(b)$ for all queue thresholds b whose corresponding critical time scale $\lambda(b)$ lies in $\mathcal{T}$. We refer to $h_\theta(\mathcal{T})$ as h_θ when $\mathcal{T} = (0, \infty)$. We use the following function to evaluate the accuracy metric

$$\zeta(s, H) := \frac{(s - 1)H^H(1 - H)^{1-H}}{(s - s^H)^{1-H}(s^H - 1)^H}. \tag{31}$$

Let $\theta = \{\tau_k^{[\theta]}\}_k$ be a set of time scales with elements arranged in increasing order. We begin by studying $h_\theta(\mathcal{T}_k^{[\theta]})$ where $\mathcal{T}_k^{[\theta]} := [\tau_{k-1}^{[\theta]}, \tau_k^{[\theta]}]$. Remarkably $h_\theta(\mathcal{T}_k^{[\theta]})$ is solely a function of the ratio of time scales $s_k^{[\theta]} := \tau_k^{[\theta]} / \tau_{k-1}^{[\theta]}$ and does not depend on any other property of θ. We denote the largest ratio of consecutive scales in θ by $d_\theta := \sup_k s_k^{[\theta]}$.

Theorem 1: *The accuracy metric over the range of scales $\mathcal{T}_k^{[\theta]}$ equals*

$$h_\theta\left(\mathcal{T}_k^{[\theta]}\right) = \zeta\left(s_k^{[\theta]}, H\right). \tag{32}$$

Corollary 2: *If θ extends from 0 to ∞, that is $\sup_k \tau_k^{[\theta]} = \infty$ and $\inf_k \tau_k^{[\theta]} = 0$, then the accuracy metric for $\mathcal{T} = (0, \infty)$ equals*

$$h_\theta = \zeta(d_\theta, H). \tag{33}$$

The proof of Theorem 1 relies on the fact that for a fixed threshold b, whose corresponding critical time scale $\lambda(b)$ lies in $\mathcal{T}_k^{[\theta]}$, the function $g(b, \tau)$ is minimized over $\tau \in \theta$ at either $\tau = \tau_{k-1}^{[\theta]}$ or $\tau = \tau_k^{[\theta]}$. Thus $h_\theta(\mathcal{T}_k^{[\theta]})$ depends only on $\tau_{k-1}^{[\theta]}$ and $\tau_k^{[\theta]}$ and not on other elements of θ. This fact and the elegant scaling properties of fBm result in (32).

B. Optimality of exponential time scales θ_α

Given a range of time scales $\mathcal{T} = (\underline{\tau}, \overline{\tau}),\ 0 < \underline{\tau} < \overline{\tau}$, we wish to find that time-scale set which is the most sparse (i.e., has the fewest elements) in $\mathcal{T}$ while guaranteeing a certain accuracy of $M^{[\theta]}(b)$.

The next theorem proves that there exists an exponential set of time scales that is most sparse among all sets θ that have accuracy metric $h_\theta(\mathcal{T})$ less than a specified threshold. Define

$$\Gamma(\alpha) := \{\theta : h_\theta(\mathcal{T}) \le \zeta(\alpha, H)\} \tag{34}$$

and let $\#\theta$ denote the number of elements of θ that lie in $\mathcal{T}$. Define the *generalized* exponential time scales as

$$\theta_{\alpha,\nu} := \{\nu\alpha^k : k \in \mathbb{Z}\} \tag{35}$$

where $\nu > 0$.

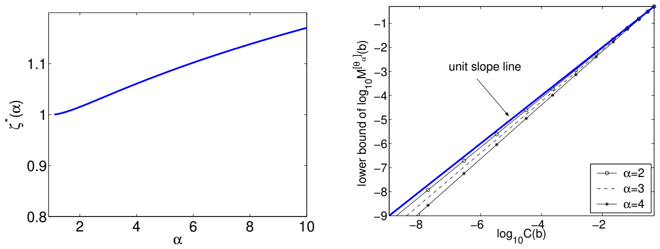
Theorem 3: *For all ν, we have $\theta_{\alpha,\nu} \in \Gamma(\alpha)$ and*

$$\#\theta_{\alpha,\nu} \le 1 + \min_{\theta \in \Gamma(\alpha)} \#\theta. \tag{36}$$

Moreover there exists $\xi > 0$ such that

$$\#\theta_{\alpha,\xi} = \min_{\theta \in \Gamma(\alpha)} \#\theta. \tag{37}$$

Theorem 3 follows from Theorem 1. Note that $h_\theta(\mathcal{T}_k^{[\theta]})$ increases with $s_k^{[\theta]}$ because $\zeta(s, H)$ is an increasing function of s (see (31) and (32)). Thus for any $\theta \in \Gamma(\alpha)$, if $\mathcal{T}_k^{[\theta]} \subset \mathcal{T}$ we must have $s_k^{[\theta]} \le \alpha$. In the exponential set θ_α, the ratio of



(a) $\zeta^*(\alpha)$ vs. α (b) $M^{[\theta_\alpha]}(b)$ vs. $C(b)$

Fig. 2. *(a) For a large range of* α, $\zeta^*(\alpha)$ *is close to* 1. *(b) Observe that the lower bound of* $M^{[\theta_2]}(b)$ *is almost identical to the upper bound* $C(b)$.

all consecutive time scale elements, $s_k^{[\theta_\alpha]}$, equals the maximum allowed value of α. Thus θ_α is the most sparse of all sets in $\Gamma(\alpha)$.

C. *Accuracy of* $M^{[\theta_\alpha]}(b)$

We use Theorem 1 to obtain the maximum error of $M^{[\theta_\alpha]}(b)$ in approximating $C(b)$ for all possible fBm traffic processes satisfying $\widehat{c} > 0$. Define $\zeta^*(\alpha) := \max_{H\in(0,1)} \zeta(\alpha, H)$.

Theorem 4: *For fBm input traffic with* $\widehat{c} > 0$

$$\Phi\left(\zeta^*(\alpha)\Phi^{-1}(C(b))\right) \leq M^{[\theta_\alpha]}(b) \leq C(b). \tag{38}$$

In Fig. 2(a) we see that the plot of $\zeta^*(\alpha)$ versus α, which we obtained numerically, is close to 1 for a large range of values of α. As a result the lower bound of $M^{[\theta_\alpha]}(b)$ from (38) for different values of α is close to $C(b)$ as depicted in Fig. 2(b). In fact $M^{[\theta_2]}(b)$ is almost identical to $C(b)$ when $C(b) > 10^{-8}$. Thus $M^{[\theta_2]}(b)$ is for all practical purposes as accurate as $C(b)$ in approximating $\mathbb{P}\{Q_\infty > b\}$.

VI. Asymptotic Analysis of fBm Queues

In this section, for a queue with fBm input, we study the accuracy of the max, product, and sum approximations of $\mathbb{P}\{Q_\infty > b\}$ for asymptotically large queue thresholds, that is as $b \to \infty$. While asymptotic queuing results are not always directly applicable to scenarios with finite queues, they often provide intuition for network design [6], [7].

We begin with some terminology. If

$$\lim_{b\to\infty} \frac{\Omega(b)}{\Upsilon(b)} = 1 \tag{39}$$

we say that Ω and Υ have the same *asymptotic decay* and denote it by $\Omega(b) \simeq \Upsilon(b)$. If $\log\Omega(b) \simeq \log\Upsilon(b)$ we say that Ω has the same *log-asymptotic decay* as Υ. Under the assumption that $\Omega(b) \to 0$ as $b \to \infty$, it is easily shown that an asymptotic decay implies a log-asymptotic decay, that is,

$$(\Omega(b) \simeq \Upsilon(b)) \;\Rightarrow\; (\log\Omega(b) \simeq \log\Upsilon(b)) \tag{40}$$

but not vice versa. We call Υ an *asymptotic upper bound* of Ω if

$$\lim_{b\to\infty} \frac{\Omega(b)}{\Upsilon(b)} = 0. \tag{41}$$

A. *Related work*

Large deviation principles reveal that $\mathbb{P}\{Q_\infty > b\}$ and $C(b)$ have the same log-asymptotic decay (see [6], [7])

$$\log\mathbb{P}\{Q_\infty > b\} \simeq \log C(b) \simeq -\frac{\eta b^{-(2-2H)}}{2} \tag{42}$$

where $\eta > 0$ is a constant depending on the traffic parameters and independent of b. However $\mathbb{P}\{Q_\infty > b\}$ and $C(b)$ do not have the same asymptotic decay: $\mathbb{P}\{Q_\infty > b\}$ is an asymptotic upper bound of $C(b)$. Interestingly under *transient* conditions, that is for a fixed t, $\mathbb{P}\{Q_t > b\}$ has the same asymptotic decay as $\sup_{0\leq\tau\leq t}\mathbb{P}\{K_t[\tau] - c\tau > b\}$ [13].

Recent results show that for fBm $\mathbb{P}\{Q_\infty > b\}$ has a Weibull asymptotic decay [8], [19], [20]

$$\mathbb{P}\{Q_\infty > b\} \simeq \vartheta b^{(1-H)(1-2H)/H} e^{-\eta b^{2-2H}/2}, \tag{43}$$

where $\vartheta > 0$ is a constant independent of b. When $1/2 < H < 1$, which implies that fBm's increment process is LRD, this Weibull decay is slower than the exponential decay for a queue fed with traffic that is not LRD, for example fBm with $H = 1/2$ [5].

From (43) we obtain that $e^{-\eta b^{2-2H}/2}$ is an asymptotic upper bound of $\mathbb{P}\{Q_\infty > b\}$ when $1/2 < H < 1$, since

$$\lim_{b\to\infty} \frac{b^{(1-H)(1-2H)/H} e^{-\eta b^{2-2H}/2}}{e^{-\eta b^{2-2H}/2}} = 0. \tag{44}$$

This asymptotic upper bound was derived as the *maximum variance* approximation in [11]. For a detailed discussion on large queue asymptotics of LRD traffic see Chs. 4 to 11 of [21] and the references therein.

B. *Asymptotic decay of approximations*

We now compare the log-asymptotic and asymptotic decay rates of the max, the product, and the sum approximations with that of $\mathbb{P}\{Q_\infty > b\}$. Define

$$b_k := \alpha^k \widehat{c}(1-H)/H, \quad k \in \mathbb{Z}, \tag{45}$$

where $\alpha > 1$ is arbitrary. We only consider the case $\theta = \theta_\alpha$. The next theorem summarizes our results.

Theorem 5: *The max, product, and sum approximations have the same log-asymptotic decay as* $\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}$ *and* $\mathbb{P}\{Q_\infty > b_k\}$*; that is as* $b_k \to \infty$ *we have*

$$\begin{aligned}\log M^{[\theta_\alpha]}(b_k) &\simeq \log P^{[\theta_\alpha]}(b_k) \simeq \log S^{[\theta_\alpha]}(b_k)\\ &\simeq \log\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\} \simeq \log\mathbb{P}\{Q_\infty > b_k\}.\end{aligned} \tag{46}$$

Moreover the max, product, and sum approximations all have the same asymptotic decay as $\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}$*; that is as* $b_k \to \infty$ *we have*

$$M^{[\theta_\alpha]}(b_k) \simeq P^{[\theta_\alpha]}(b_k) \simeq S^{[\theta_\alpha]}(b_k) \simeq \mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}. \tag{47}$$

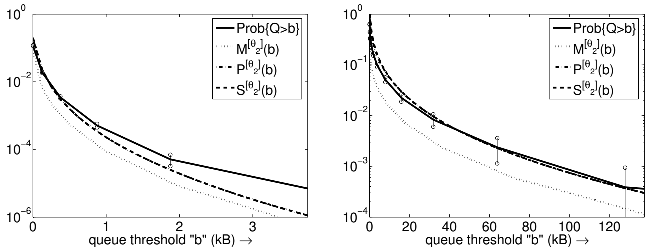
However

$$\lim_{k\to\infty} \frac{\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}}{\mathbb{P}\{Q_\infty > b_k\}} = 0. \tag{48}$$

Theorem 5 reveals the strengths and limitations of using traffic statistics only at exponential time scales θ_α to capture queuing behavior. Recall from (2) and (6) that $Q^{[\theta_\alpha]}$

approximates the queue size Q using traffic only at time scales $\tau \in \theta_\alpha$. From (46) we see that θ_α is dense enough in $\mathbb{R}_+$ to ensure that $\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}$ and $\mathbb{P}\{Q_\infty > b_k\}$ have the same log-asymptotic decays for a particular unbounded increasing sequence of queue sizes b_k. However, θ_α is not dense enough to ensure that $\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}$ and $\mathbb{P}\{Q_\infty > b_k\}$ have the same asymptotic decay.

We also observe from (47) that the max, product, and sum approximations have the same asymptotic decay as $\mathbb{P}\left\{Q^{[\theta_\alpha]} > b_k\right\}$. As a result they have the same log-asymptotic decay but different asymptotic decay as $\mathbb{P}\{Q_\infty > b_k\}$. We next present non-asymptotic results comparing the different queuing approximations to $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$.



(a) 30% utilization (b) 80% utilization

Fig. 3. *Comparison of the max the product and the sum approximations to $\mathbb{P}\{Q > b\}$ for fGn traffic with parameters $H = 0.8$, link capacity 10Mbps and $\sigma = 8 \times 10^5$ bits. For different utilizations in (a) and (b), the product and sum approximations are close to $\mathbb{P}\{Q > b\}$ for a wide range of queue thresholds b. The max approximation is a lower bound of $\mathbb{P}\{Q > b\}$ and is accurate to an order of magnitude.*

VII. Bounds for the Approximations

The knowledge of whether or not a queuing approximation is an upper or lower bound of $\mathbb{P}\{Q > b\}$ aids different applications. For example if we provision the queue service rate such that the critical time scale approximation $C(b)$ equals 10^{-6}, then we must expect the actual tail queue probability $\mathbb{P}\{Q > b\}$ to exceed 10^{-6} since $C(b)$ lower bounds $\mathbb{P}\{Q > b\}$ (see (5)). If $C(b)$ is an accurate approximation of $\mathbb{P}\{Q > b\}$ to an order of magnitude, as our simulations with fGn traffic in Section VIII affirm, then we would effectively be provisioning for $\mathbb{P}\{Q > b\} < 10^{-5}$. If we replace the lower bound $C(b)$ by an approximation that is an upper bound of $\mathbb{P}\{Q > b\}$, then $\mathbb{P}\{Q > b\}$ is guaranteed to be less than 10^{-6}.

In this section we prove bounding results for the max, product, and sum approximations, which we compare to $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$ rather than $\mathbb{P}\{Q > b\}$. Note from (10) that lower bounds of $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$ are also lower bounds of $\mathbb{P}\{Q > b\}$. However, the queuing approximations that are upper bounds of $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$ are not necessarily upper bounds of $\mathbb{P}\{Q > b\}$.

A. Bounds for general input traffic processes

We first state a general result that holds for a queue fed by *any* traffic random process and then present model-specific results.

Lemma 6: *For a discrete or continuous-time queue of infinite size, with an arbitrary input traffic process and constant service rate*

$$M_t^{[\theta]}(b) \leq \mathbb{P}\left\{Q_t^{[\theta]} > b\right\} \leq S_t^{[\theta]}(b) \tag{49}$$

and

$$M_t^{[\theta]}(b) \leq P_t^{[\theta]}(b) \leq S_t^{[\theta]}(b), \tag{50}$$

where θ is any countable subset of $\mathbb{R}_+$.

From Lemma 6 we see that max and sum approximations are always lower and upper bounds respectively of both $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$ and the product approximation. In the rest of this section we compare the product approximation to $\mathbb{P}\left\{Q^{[\theta]} > b\right\}$.

B. Product approximation bounds for Gaussian traffic

For queues fed with traffic from a large class of Gaussian processes, including fBm and fGn, $P_t^{[\theta]}(b)$ is an upper bound of $\mathbb{P}\left\{Q_t^{[\theta]} > b\right\}$.

Theorem 7: *Consider a Gaussian traffic process X_t input to an infinite buffer queue with constant service rate (discrete or continuous-time). If $\mathrm{cov}(K_t[\tau], K_t[r]) \geq 0$ for all $\tau, r \in \theta$ then*

$$\mathbb{P}\left\{Q_t^{[\theta]} > b\right\} \leq P_t^{[\theta]}(b), \tag{51}$$

where θ is any countable subset of $\mathbb{R}_+$.

Note that fBm satisfies the requirements of Theorem 7 since for all $\tau, r \geq 0$ and $0 < H < 1$

$$\begin{aligned} \mathrm{cov}\left(K_t^{\{\Delta_t B\}}[\tau], K_t^{\{\Delta_t B\}}[r]\right) &= \frac{\tau^{2H} + r^{2H} - |\tau - r|^{2H}}{2} \\ &\geq 0. \end{aligned} \tag{52}$$

Similarly fGn too satisfies the requirements of Theorem 7.

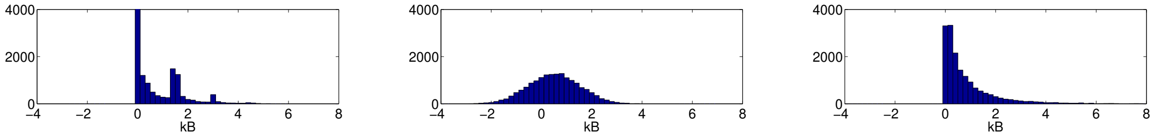
C. Product approximation bounds for WIG and MWM traffic

For the WIG and MWM we restrict our attention to the set of dyadic time scales, that is $\theta = \theta_\alpha$ with $\alpha = 2$. Recall from Section IV-C that the WIG and MWM are non-stationary traffic models. As a consequence $P_t^{[\theta_2]}(b)$ changes with time location t. We first compare $P_t^{[\theta_2]}(b)$ to $\mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\}$ for $t = 2^n$, that is at the final time instant of the tree process, and then at all other time instants t. We denote the final time instant 2^n by "end".

Theorem 8: *For the WIG and MWM with arbitrary model parameters*

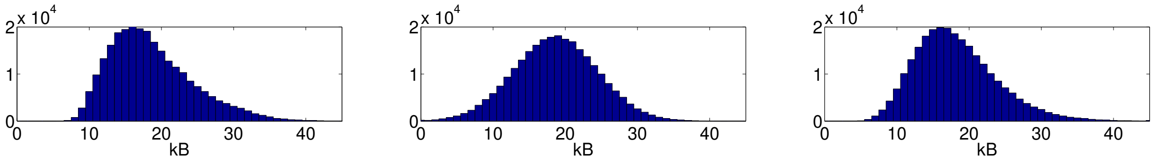
$$\mathbb{P}\left\{Q_{\mathsf{end}}^{[\theta_2]} > b\right\} \leq P_{\mathsf{end}}^{[\theta_2]}(b) \quad \forall b > 0. \tag{53}$$

Theorem 8 states that $P_t^{[\theta_2]}(b)$ is an upper bound of $\mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\}$ at the final time instant for the WIG and the MWM for arbitrary model parameters. The only ingredient of the proof of Theorem 8 is the fact that the quantities $K_{\mathsf{end}}[2^j],\ j = 1, 2, \ldots, n$ that determine $P_{\mathsf{end}}^{[\theta_2]}(b)$ are nodes along the right edge of the tree and hence are related through



(a) AUCK traffic (b) WIG synthesis (c) MWM synthesis

Fig. 4. *Histograms of the bytes-per-time processes at time-scale* 2ms *for (a) wide-area traffic at the University of Auckland (trace AUCK) [22], (b) one realization of the WIG model, and (c) one realization of the MWM.*



(a) VIDEO traffic (b) WIG synthesis (c) MWM synthesis

Fig. 5. *Histograms of the bytes-per-time processes at time-scale* 2.77ms *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.*

independent innovations (see Fig. 1). Since this fact is true for arbitrary model parameters, so is (53).

Generalizing the proof of Theorem 8 so that (53) holds for all time instants t is not straightforward because the quantities $K_t[2^j],\ j = 1, 2, \ldots, n$ are not always tree nodes for arbitrary t and are hence not related through independent innovations as the quantities $K_{\mathsf{end}}[2^j],\ j = 1, 2, \ldots, n$ are. However, for a WIG model of fGn we can extend (53) to all t as stated next.

Theorem 9: *For the WIG model of fGn*

$$\mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\} \le P_t^{[\theta_2]}(b) \le P_{\mathsf{end}}^{[\theta_2]}(b) \qquad \forall t. \tag{54}$$

As a consequence

$$\frac{1}{2^n}\sum_{t=1}^{2^n} \mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\} \le P_{\mathsf{end}}^{[\theta_2]}(b). \tag{55}$$

Theorem 9 reveals that for a WIG model of fGn $P_{\mathsf{end}}^{[\theta_2]}(b)$ is an upper bound of the time average of $\mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\}$.

Earlier work on the queuing behavior of the WIG model of fGn proved that the time average of the tail queue probability $\mathbb{P}\{Q_t > b\}$ has the same log-asymptotic behavior as that of fGn [16].

We demonstrate through simulations in Section VIII that $P_{\mathsf{end}}^{[\theta_2]}(b)$ approximates the time average of $\mathbb{P}\{Q_t > b\}$ well for a large range of queue sizes b for both the WIG and the MWM.

VIII. Simulations

In this section we demonstrate the accuracy of the max, product, and sum approximations of $\mathbb{P}\{Q > b\}$ through simulations with fGn, WIG, and MWM synthetic traces as well as with video and measured Internet traces. We also demonstrate that the tails of marginals of traffic at different time scales have a significant impact on queuing in certain scenarios by comparing the queuing behavior of the WIG and MWM models with that of measured Internet traffic. We restrict our attention to exponential time-scales with $\alpha = 2$ (that is $\theta = \theta_2$). All error bars in the plots correspond to 95% confidence intervals.

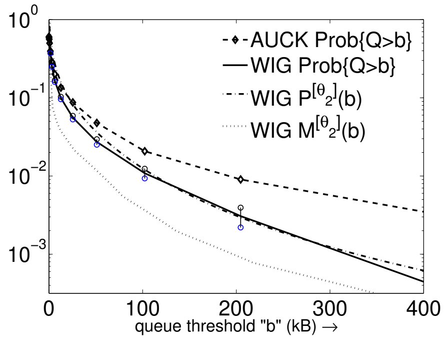
A. Comparison of queuing approximations for fGn traffic

We now compare the different approximations of $\mathbb{P}\{Q > b\}$ through simulations with fGn traffic. The simulations use fGn traces with Hurst parameter $H = 0.8$ and standard deviation at the 1s time-scale $\sigma = 8 \times 10^5$ bits that are generated using the method described in [23]. We set $\tau' = 10^{-4}$s and $c =$ 10Mbps and vary the mean rate of the traces to obtain different utilizations.

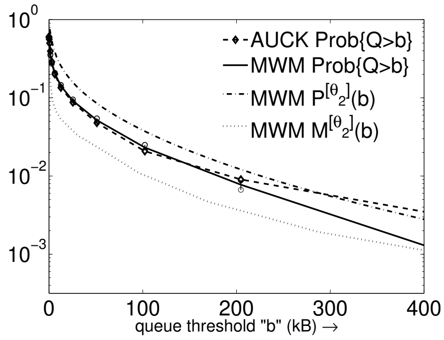
We estimate $\mathbb{P}\{Q_\infty > b\}$ for each simulation run as the fraction of time for which the queue size exceeds b. To eliminate transients we only make estimates using queue sizes during the second half of the simulation. The plots of tail queue probability correspond to the mean obtained from 300 simulation runs. Each run uses a trace of length 2^{19} data points corresponding to a 52s simulation time.

The simulation results for two different utilizations are depicted in Fig. 3. We obtain the various queuing approximations using (8), (12), and (13) by choosing $\theta = \{\tau', 2\tau', \ldots, 2^{20}\tau'\}$ which is equivalent to θ_2 truncated to lie within a fixed range of time-scales. Observe that in all cases $M^{[\theta_2]}(b)$ is a lower bound of $\mathbb{P}\{Q > b\}$ as predicted by (9). We also see that $M^{[\theta_2]}(b)$ is within an order of magnitude of $\mathbb{P}\{Q > b\}$ for a wide range of values of $\mathbb{P}\{Q > b\}$ ($\in [10^{-6}, 1]$). We conclude that $C(b)$, which lies between $M^{[\theta_2]}(b)$ and $\mathbb{P}\{Q > b\}$ (see (9)), is also within an order of magnitude of $\mathbb{P}\{Q > b\}$ for the same range of $\mathbb{P}\{Q > b\}$.

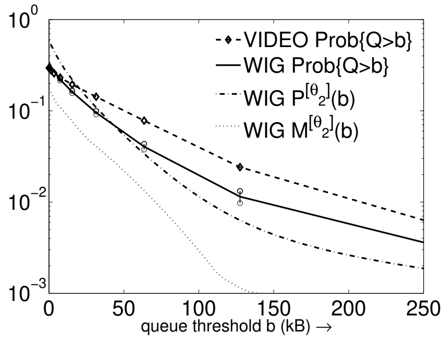
From Fig. 3 observe that the product and sum approximations are almost identical and accurately track $\mathbb{P}\{Q > b\}$ for a wide range of queue sizes b. Also observe that they are better approximations than the max approximation in general. However unlike the max approximation, which is a guaranteed lower bound of $\mathbb{P}\{Q > b\}$, these two approximations do not bound $\mathbb{P}\{Q > b\}$ from above or from below and in fact intersect it at some point. Call the queue threshold at which the product approximation and $\mathbb{P}\{Q > b\}$ intersect b'. We observe that in all cases the product approximation is greater than $\mathbb{P}\{Q > b\}$ at $b = 0$ and for $b > b'$ is always less than $\mathbb{P}\{Q > b\}$. Thus for $b > b'$ the product approximation



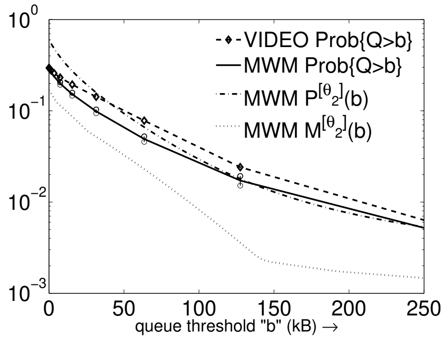
(a) AUCK vs. WIG



(b) AUCK vs. MWM



(c) VIDEO vs. WIG



(d) VIDEO vs. MWM

Fig. 6. Queuing performance of real data traces and synthetic WIG and MWM traces at high utilization. *In (b), we observe that the MWM synthesis matches the queuing behavior of the AUCK data closely, while in (a) the WIG synthesis is not as close. In (c) and (d), we observe that both the WIG and the MWM match the queuing behavior of VIDEO. We also observe that the product approximation ($P^{[\theta_2]}(b)$) is close to the empirical queuing behavior for both synthetic traffic loads (both WIG and MWM) and that it performs better than the max approximation, $M^{[\theta_2]}(b)$.*

lies between the max approximation and $\mathbb{P}\{Q > b\}$ which guarantees that it is a better approximation than the max approximation. The sum approximation has a similar behavior.

B. Impact of marginals on queuing

The impact of different traffic statistics on queuing has been extensively studied. Several studies have debated the importance of LRD for queuing [9], [24]–[27]. LRD is however only a function of the asymptotic second-order correlation structure of traffic (or equivalently the *variance* of traffic at different time scales).

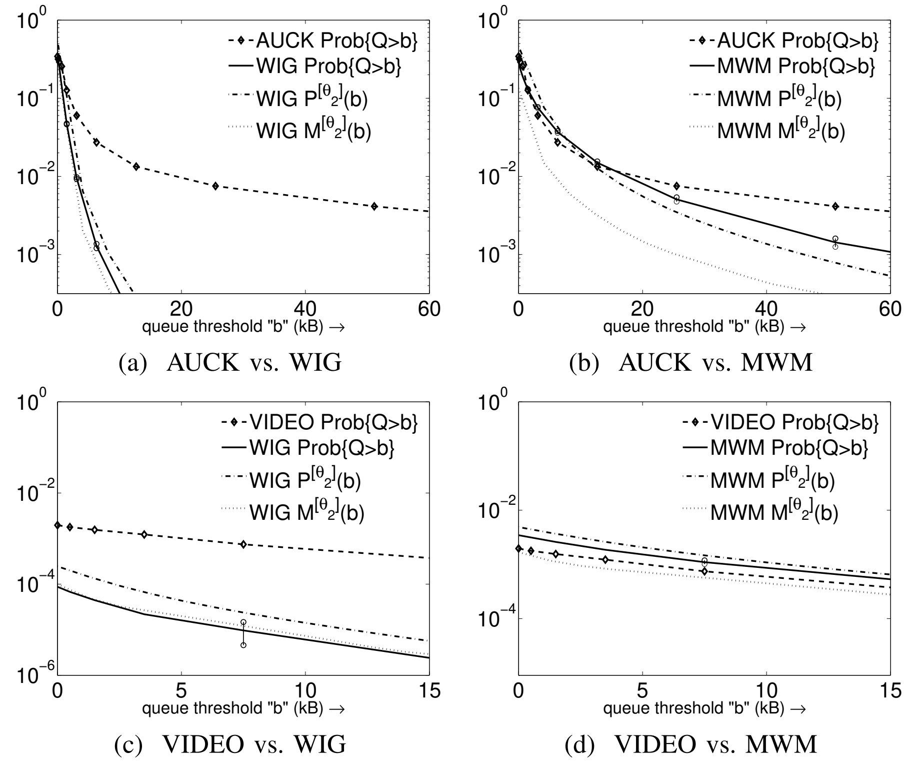
In this section we move beyond second-order statistics and demonstrate the impact of the *tails* of traffic marginals at different time scales on queuing. We do so by comparing the queuing behavior of the WIG and MWM processes with video and Internet WAN traces through simulations. Recall from Section IV that both the WIG and the MWM can capture a wide range of second-order correlation structures. They however differ in their marginal characteristics: the WIG process is Gaussian whereas the MWM process is non-Gaussian. We interpret our results using the product approximation and the conclusions of earlier work which studied the influence of link utilization on queuing [12].

Traces: The two traces we use are AUCK, which contains the number of bytes per 2ms of recorded WAN traffic [22] and VIDEO, which consists of 15 video clips multiplexed with random starting points [28]. 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. From Fig. 4 and Fig. 5 observe that AUCK has strongly non-Gaussian marginals while VIDEO's marginals resemble a Gaussian distribution.

Simulation results: We fit the WIG and MWM to the real data and then generated synthetic traces from the models. We then compared the queuing behavior of the synthesized WIG and MWM traces with that of the real data when they are input to a queue of infinite length with constant service rate. The plots of $\mathbb{P}\{Q > b\}$ correspond to the mean obtained from 1000 simulation runs.

We first present results for high link utilizations ($> 70\%$). Observe from Figs. 6(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 (43)). 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 Figs. 6(c) and (d)), which is much closer to a Gaussian process than AUCK,



(a) AUCK vs. WIG (b) AUCK vs. MWM

(c) VIDEO vs. WIG (d) VIDEO vs. MWM

Fig. 7. Queuing performance of real data traces and synthetic WIG and MWM traces at low utilization. *The MWM outperforms the WIG even more than at higher utilizations.*

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 Fig. 5), the approximately lognormal MWM is suitable for Gaussian traffic [17]. The link capacity we use is 69Mbps, which corresponds to a utilization of 77%.

In the case of lower link utilizations ($< 50\%$) from Fig. 7 we see that the MWM outperforms the WIG for both AUCK and VIDEO traces to a greater extent than in the high utilization case. For both the MWM and WIG we observe that the product approximation is close to $\mathbb{P}\{Q > b\}$ (see Figs. 6 and 7). The max approximation is within an order of magnitude of $\mathbb{P}\{Q > b\}$.

Interpretation using the product approximation: Accepting the product approximation $P^{[\theta_2]}(b)$ as a close approximation to the actual tail queue probabilities, a closer look at (12) unravels how the marginals affect queue sizes. For traffic with heavier tailed marginals, the terms $\mathbb{P}\left\{K[2^i] < b + c2^i\right\}$ are smaller and the product approximation is larger. Since the MWM marginals are more heavy tailed than the Gaussian WIG marginals, the MWM has a larger product approximation than the WIG.

In the case of VIDEO, which shows marginals much closer to Gaussian (see Fig. 5), both the WIG and MWM perform similarly in terms of capturing the tail queue probability at a high utilization, while at a low utilization the MWM outperforms the WIG. This result is easily explained using the finding in [12] that fine time-scale statistics influence queuing more than coarse time-scale statistics at low utilizations. Since fine time-scale marginals of VIDEO are more non-Gaussian than coarse time-scale marginals, obviously the MWM performs better than the WIG at low utilizations.

IX. Conclusions

We have developed a new approach to queuing analysis of network traffic that uses traffic statistics at a fixed finite set of time scales. The queuing analysis provides three approximations for the tail queue probability of an infinite buffer queue with constant service rate. Theoretical and simulation results strongly support their use for different applications.

We also proved that exponential time scales are optimal for fBm traffic with respect to a trade-off in accuracy vs. computational cost of the max approximation. Applications can thus obtain accurate approximations to the tail queue probability by employing traffic statistics only at a few sparse exponential time scales.

Our simulations demonstrated the impact of the tails of marginals at different time scales on queuing. We observed that in non-Gaussian traffic scenarios the correlation structure (short and long term) does not characterize the queuing behavior well.

There remain several open research problems that we have not addressed in this paper. First, we have developed a multiscale queuing analysis only for a single queue with constant service rate. Our intuition suggests that a multiscale queuing paradigm can help analyze more complex systems consisting of multiple queues with arbitrary service disciplines.

Second, we have ignored the case of finite length queues where packet drops occur. Our results thus are more useful in predicting packet queuing delays rather than packet losses.

Third, our analysis is for open-loop traffic models while real Internet traffic is mainly composed of closed-loop TCP traffic. Closed-loop traffic reacts to changes in network conditions unlike open-loop traffic [29]. For example, unlike open-loop traffic, closed-loop traffic will reduce its offered load if a bottleneck link speed is reduced. Thus one must use open-loop queuing results with caution in the Internet while ensuring that one does not affect network properties (delay and loss) which can influence the TCP traffic significantly. Possible applications of open-loop models are for Internet backbone provisioning (for very low delay/loss ISPs) [4], and obviously in networks dominated by open-loop traffic like certain UDP streaming applications.

APPENDIX

Proof of Theorem 1: We prove the theorem in three steps. For the ease of notation we drop the superscript from $\tau_k^{[\theta]}$, $\mathcal{T}_k^{[\theta]}$ and $s_k^{[\theta]}$.

Step 1: Determine $\inf_{\tau>0} g(b,\tau)$ for a fixed value $b>0$. From (27) we obtain:

$$\frac{\partial g(b,\tau)}{\partial \tau} = \frac{\widehat{c}\tau(1-H) - bH}{\sigma \tau^{1+H}}. \tag{56}$$

Thus $g(b,\tau)$ is minimized at $\tau = \lambda(b)$ where

$$\lambda(b) = \frac{bH}{\widehat{c}(1-H)}. \tag{57}$$

In addition $g(b,\tau)$ is non-decreasing with τ as we move away from $\tau=\lambda(b)$. Clearly

$$\inf_{\tau>0} g(b,\tau) = g(b,\lambda(b)) = \frac{b^{1-H}\widehat{c}^H}{\sigma H^H (1-H)^{1-H}}. \tag{58}$$

Step 2: Find $\varsigma(b) := \inf_{\tau\in\theta} g(b,\tau)/g(b,\lambda(b))$ for fixed $b \in \mathcal{B}_k$ where

$$\mathcal{B}_k := \left[\lambda^{-1}(\tau_{k-1}), \lambda^{-1}(\tau_k)\right], \tag{59}$$

and $\lambda^{-1}(\tau)$ is the inverse of $\lambda(b)$ given by

$$\lambda^{-1}(\tau) := \widehat{c}\tau(1-H)/H. \tag{60}$$

From (59) and (60) observe that $\mathcal{B}_k = \{b : \lambda(b) \in \mathcal{T}_k\}$. Consider

$$f(b,\tau) := \frac{g(b,\tau)}{g(b,\lambda(b))} = \frac{(b+\widehat{c}\tau)(\sigma H^H(1-H)^{1-H})}{\sigma\tau^H(b^{1-H}\widehat{c}^H)}. \tag{61}$$

Since $g(b,\tau)$ is non-decreasing as we move away from $\tau = \lambda(b)$, we must have that

$$\varsigma(b) = \min\{f(b,\tau_{k-1}), f(b,\tau_k)\}. \tag{62}$$

Step 3: Determine $\sup_{b\in\mathcal{B}_k} \varsigma(b)$.

By elementary calculus we obtain that $f(b,\tau_{k-1})$ monotonically increases with b when $b > \lambda^{-1}(\tau_{k-1})$. Also $f(b,\tau_k)$ monotonically decreases with increasing b when $b < \lambda^{-1}(\tau_k)$. If there exists $a_k \in \mathcal{B}_k$ such that $f(a_k,\tau_{k-1}) = f(a_k,\tau_k)$, then $\varsigma(b)$ must attain its supremum over $\mathcal{B}_k$ at this point (from (62)). Indeed such an a_k does exist. From (61) we obtain a_k as

$$a_k = \frac{\widehat{c}\tau_{k-1}\tau_k(\tau_{k-1}^{H-1} - \tau_k^{H-1})}{\tau_k^H - \tau_{k-1}^H} = \frac{\widehat{c}\tau_k}{s_k}\cdot\frac{s_k - s_k^H}{s_k^H - 1}. \tag{63}$$

As a result, after simplification

$$\begin{aligned}\sup_{b\in\mathcal{B}_k} \varsigma(b) = f(a_k,\tau_k) &= \frac{(s_k-1)H^H(1-H)^{1-H}}{(s_k^H-1)^H(s_k-s_k^H)^{1-H}} \\ &= \zeta(s_k,H). \end{aligned}\tag{64}$$

□

Proof of Corollary 2: The proof relies on the following claim.

Claim 10: $\zeta(s_k,H)$ increases with s_k for all $H\in(0,1)$.

Proof of Claim 10: Note from (64) that $\zeta(s_k,H)$ equals $f(a_k,\tau_k)$. It is thus sufficient to prove that $f(a_k,\tau_k)$ increases with s_k. Without loss of generality we study how $f(a_k,\tau_k)$ changes by varying τ_{k-1} keeping τ_k fixed. Note that this is equivalent to varying s_k. We have from (63)

$$\frac{1}{\widehat{c}\tau_k}\cdot\frac{\partial a_k}{\partial \tau_{k-1}} = \frac{\tau_{k-1}^{2H-1}s_k^{H-1}(Hs_k - s_k^H + (1-H))}{(\tau_k^H - \tau_{k-1}^H)^2} \tag{65}$$

It is easily shown that the function $Hs_k - s_k^H + (1-H)$ equals 0 at $s_k = 1$ and has a positive derivative for $s_k>1$. Thus $\frac{\partial a_k}{\partial \tau_{k-1}} > 0$ for all $s_k > 1$. Using this fact, the knowledge that $a_k < \lambda^{-1}(\tau_k)$, and the fact that $f(b,\tau_k)$ monotonically decreases for $b<\lambda^{-1}(\tau_k)$ we see that $f(a_k,\tau_k)$ decreases with increasing τ_{k-1}, or equivalently it increases with increasing s_k. Claim 10 is thus proved.

From the fact that $\cup_k \mathcal{T}_k = \mathbb{R}_+$, (59) and (60) we have that $\cup_k\mathcal{B}_k = \mathbb{R}_+$. Using this fact, (64), Claim 10, and exploiting the continuity of $\zeta(s_k,H)$ (see (64)) we have

$$h_\theta = \sup_{b\in\mathbb{R}_+}\varsigma(b) = \sup_k \zeta(s_k,H) = \zeta(d_\theta,H). \tag{66}$$

□

Proof of Theorem 3: From (30) $h_{\theta_{\alpha,\nu}}(\mathcal{T}) \le h_{\theta_{\alpha,\nu}} = \alpha$. Thus $\theta_{\alpha,\nu}\in\Gamma(\alpha)$.

Consider $\mathcal{D}\in\Gamma_\alpha$, such that $\#\mathcal{D} = \min_{\theta\in\Gamma(\alpha)}\#\theta$. We can write $\mathcal{T}$ as a union of the following $\#\mathcal{D}+1$ intervals: $(\underline{\tau}, \tau_i^{[\mathcal{D}]})$, $[\tau_i^{[\mathcal{D}]}, \tau_{i+1}^{[\mathcal{D}]})$, …, $[\tau_{i+\#\mathcal{D}-1}^{[\mathcal{D}]}, \overline{\tau})$, for some i. Because $\mathcal{D}\in\Gamma_\alpha$, from Theorem 1 the ratio of supremum to infimum of each of these intervals must be less than or equal to α. Consider $\theta_{\alpha,\nu}$ for arbitrary ν. Clearly by definition $\theta_{\alpha,\nu}$ can have at most one element in each of the these intervals. Thus (36) is proved. Consider $\theta_{\alpha,\xi}$ where $\xi = \tau_i^{[\mathcal{D}]}/\alpha^i$. Clearly $\theta_{\alpha,\xi}$ has no element in the first set of the union mentioned above and at most one element in the other sets of the union. □

Proof of Theorem 4: Note that by the construction of θ_α (see (25)), $d_{\theta_\alpha} = \alpha$. Thus from (9), (28)–(30), and (33) we have

$$\begin{aligned} C(b) &\ge M^{[\theta_\alpha]}(b) = \Phi\left(\inf_{\tau\in\theta_\alpha} g(b,\tau)\right) \\ &\ge \Phi\left(h_{\theta_\alpha}\inf_{\tau>0} g(b,\tau)\right) = \Phi\left(\zeta(\alpha,H)\Phi^{-1}(C(b))\right) \\ &\ge \Phi\left(\zeta^*(\alpha)\cdot\Phi^{-1}(C(b))\right). \end{aligned}$$

□

Proof of Theorem 5: The proof relies on the following two claims.

Claim 11: $M^{[\theta_\alpha]}(b_k) \simeq S^{[\theta_\alpha]}(b_k)$.

Claim 12: $\lim_{b\to\infty} \frac{C(b)}{\mathbb{P}\{Q_\infty > b\}} = 0$.

From (50), (49), and Claim 11 we have (47). From (45) and (57) note that

$$\lambda(b_k) = \lambda^{[\theta_\alpha]}(b_k) = \alpha^k. \tag{67}$$

Thus (4) and (8) yield

$$M^{[\theta_\alpha]}(b_k) = C(b_k) \quad \forall k. \tag{68}$$

From (42), (47), (68) and (40) we have (46). Finally Claim 12 combined with (47) gives (48).

We now prove the two claims. Recall the definition of $g(b, \tau)$ (see (27)). For the ease of notation we denote $g(b_k, \alpha^l)$ by $g_{k,l}$. From (58) and (67) we have

$$\inf_{\tau>0} g(b_k, \tau) = g(b_k, \lambda(b_k)) = g(b_k, \alpha^k) = g_{k,k}. \tag{69}$$

Proof of Claim 11: From (8), (13), (26), and (69) we have

$$S^{[\theta_\alpha]}(b_k) = \sum_{l\in\mathbb{Z}} \Phi(g_{k,l}) \tag{70}$$

and

$$M^{[\theta_\alpha]}(b_k) = \sup_{l\in\mathbb{Z}} \Phi(g_{k,l}) = \Phi(g_{k,k}). \tag{71}$$

We now prove that the maximum term, $\Phi(g_{k,k})$, dominates the summation of (70). We note two properties of $\frac{g_{k,l}}{g_{k,k}}$. First, we have from (45) and (27)

$$\begin{aligned} \frac{g_{k,l}}{g_{k,k}} &= \frac{b_k + \widehat{c}\alpha^l}{\sigma\alpha^{lH}} \cdot \frac{\sigma\alpha^{kH}}{b_k + \widehat{c}\alpha^k} \\ &= (1-H)\alpha^{(k-l)H} + H\alpha^{(l-k)(1-H)} \\ &\geq \epsilon_H \alpha^{|l-k|\epsilon_H}, \end{aligned} \tag{72}$$

where $\epsilon_H = \min(H, 1-H)$.

Second, from (56) and (67) observe that $\frac{g_{k,l}}{g_{k,k}}$ monotonically increases with increasing l when $l \geq k$ and also with decreasing l when $l \leq k$. We then have $\forall l \neq k$

$$\frac{g_{k,l}}{g_{k,k}} \geq \min\left(\frac{g_{k,k+1}}{g_{k,k}}, \frac{g_{k,k-1}}{g_{k,k}}\right) =: \mathcal{I}_H > 1 \tag{73}$$

Now $g_{k,k}$ is an increasing unbounded function of k. From (70), and using the following estimates of Φ [30],

$$(1 - 1/\delta^2)e^{-\delta^2/2}/(\delta\sqrt{2\pi}) \leq \Phi(\delta) \leq e^{-\delta^2/2}/(\delta\sqrt{2\pi}) \tag{74}$$

we have

$$\begin{aligned} \Phi(g_{k,k}) &\leq S^{[\theta_k]}(b_k) \\ &= \Phi(g_{k,k}) + \sum_{l>k} \Phi(g_{k,l}) + \sum_{l<k} \Phi(g_{k,l}) \qquad (75) \\ &\leq \Phi(g_{k,k}) \left(1 + \frac{2}{\epsilon_H} \cdot \frac{g_{k,k}^2 e^{-(\mathcal{I}_H^2 - 1)g_{k,k}^2/2} \alpha^{-\epsilon_H}}{(g_{k,k}^2 - 1)(1 - \alpha^{-\epsilon_H})}\right). \end{aligned}$$

From (73) and the fact that $g_{k,k} \overset{k\to\infty}{\longrightarrow} \infty$ we have

$$\lim_{k\to\infty} \frac{S^{[\theta_k]}(b_k)}{\Phi(g_{k,k})} = 1, \tag{76}$$

which proves Claim 11.

Proof of Claim 12: From (74) observe that $\Phi(\delta) \simeq \frac{e^{-\delta^2/2}}{\delta\sqrt{2\pi}}$.

Set $\eta := \left(\frac{\widehat{c}^H}{\sigma H^H (1-H)^{1-H}}\right)^2$. From (28) and (58) we then have

$$C(b) = \Phi(b^{1-H}\eta^{1/2}) \simeq \frac{b^{-(1-H)}}{\eta^{1/2}\sqrt{2\pi}} e^{-b^{2-2H}\eta/2}. \tag{77}$$

When $1/2 < H < 1$ we have $0 < \frac{2H-1}{H} < 1$ which implies that

$$\lim_{b\to\infty} \frac{b^{-(1-H)}}{b^{-(1-H)(2H-1)/H}} = 0. \tag{78}$$

Claim 12 follows from (43), (77) and (78). ☐

Proof of Lemma 6: Consider a queue with constant service rate c bits per unit time. Clearly

$$\sup_{\tau\in\theta} \mathbb{P}\{\mathcal{E}[\tau]\} \leq \mathbb{P}\{\cup_{\tau\in\theta}\mathcal{E}[\tau]\} \leq \sum_{\tau\in\theta} \mathbb{P}\{\mathcal{E}[\tau]\}, \tag{79}$$

where $\mathcal{E}[\tau] := \{K_t[\tau] - c\tau > b\}$. From (6) we see that $\mathbb{P}\{Q^{[\theta]} > b\}$ is identical to $\mathbb{P}\{\cup_{\tau\in\theta}\mathcal{E}[\tau]\}$. Then (8), (13) and (79) give (49). Note that (50) is equivalent to

$$\max_{k=1,\ldots,l} (1 - a_k) \leq 1 - \prod_{k=1}^{l} a_k \leq \sum_{k=1}^{l} (1 - a_k), \tag{80}$$

for $0 \leq a_k \leq 1$, $k = 1, \ldots, l$, which is elementary. ☐

Lemma 13: (from [31])*If events $E_1 \subset E_2 \cdots$ and $E = \cup_i E_i$ then $\lim_{i\to\infty} \mathbb{P}\{E_i\} = \mathbb{P}\{E\}$. If $E_1 \supset E_2 \cdots$ and $E = \cap_i E_i$ then $\lim_{i\to\infty} \mathbb{P}\{E_i\} = \mathbb{P}\{E\}$.*

The following lemma helps us prove Theorem 7.

Lemma 14: (page 6 in [32]) *Let $\Upsilon[t]$ and $\Psi[\tau]$, $\tau \in \theta$, be separable Gaussian random processes, where θ is a parameter set. If the following relations hold for their covariance functions:*

$$\begin{aligned} \text{var}(\Upsilon[\tau]) &= \text{var}(\Psi[\tau]), \quad \forall \tau \in \theta && (81) \\ \text{cov}(\Upsilon[\tau], \Upsilon[r])) &\leq \text{cov}(\Psi[\tau], \Psi[r])), \quad \forall \tau, r \in \theta && (82) \end{aligned}$$

plus their expected values are the same $\forall\tau$: then for any $x \in \mathbb{R}$

$$\mathbb{P}\left\{\sup_{\tau\in\theta} \Upsilon[\tau] < x\right\} \leq \mathbb{P}\left\{\sup_{\tau\in\theta} \Psi[\tau] < x\right\}. \tag{83}$$

Proof of Theorem 7: For $\tau \in \theta$ define independent Gaussian random variables $\Upsilon[\tau] \sim \mathcal{N}(\mathbb{E}(K_t[\tau] - \tau c), \text{var}(K_t[\tau]))$ and set $\Psi[\tau] := K_t[\tau] - \tau c$. Label the elements of θ as $\{\tau_k\}_{k\in\mathbb{Z}}$. Using Lemma 13 with the events $E_i := \cap_{k=-i}^{i}\{\Upsilon[\tau_k] < b\}$ and $E := \cap_i E_i = \{\sup_{\tau\in\theta} \Upsilon[\tau] < b\}$ we have

$$P^{[\theta]}(b) = 1 - \prod_{\tau\in\theta} \mathbb{P}\{\Upsilon[\tau] < b\} = 1 - \mathbb{P}\left\{\sup_{\tau\in\theta} \Upsilon[\tau] < b\right\}. \tag{84}$$

Then the fact that $\sup_{\tau\in\theta} \Psi[\tau] = Q_t^{[\theta]}$ along with (83) and (84) prove the theorem. ☐

Lemma 15: *Assume that the events W_i are of the form $W_i = \{I_i < \kappa_i\}$, where $I_i = R_0 + R_1 + \ldots + R_i$ 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*

$$\mathbb{P}\{W_i | W_{i-1}, \ldots, W_0\} \geq \mathbb{P}\{W_i\}. \tag{85}$$

Proof: By f_L and F_L denote the probability density function (p.d.f) and cumulative distribution function (c.d.f.), respectively, of a random variable L. Furthermore, we denote by $F_{L|E}(l)$ the c.d.f. of L conditioned on knowing the event E. For convenience, let us write $W_i := \{I_i < \kappa_i\}$ for short, and let us introduce the auxiliary random variables $Y_0 := L_0 := I_0 := R_0$,

$$Y_i := I_i | W_{i-1}, \ldots, W_0; \;\; L_i := I_i | W_i, \ldots, W_0, \;\; i \geq 1. \tag{86}$$

To prove the lemma, it is enough to show that

$$F_{Y_i}(r) \geq F_{I_i}(r) \tag{87}$$

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

We prove (87) by induction. First note that $F_{Y_0}(r) \geq F_{I_0}(r)$. Next, we assume that (87) holds for i and show that it holds also for $i+1$. Bayes' rule yields

$$F_{L_i}(r) \;\; = \;\; \left\{ \begin{array}{ll} \frac{F_{Y_i}(r)}{F_{Y_i}(\kappa_i)}, & \text{if } r \leq \kappa_i \\ 1, & \text{otherwise} \end{array} \right\} \geq F_{Y_i}(r). \tag{88}$$

Note that $Y_{i+1} = L_i + R_{i+1}$, where R_{i+1} is independent of I_j and hence of W_j for $j \leq i$. In short, R_{i+1} is independent of L_i. This fact, (87) and (88) allow us to write

$$\begin{aligned} F_{Y_{i+1}}(r) &= \mathbb{P}\{L_i + R_{i+1} < r\} \\ &= \int_{-\infty}^{\infty} F_{L_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1}) \, \mathrm{d}r_{i+1} \\ &\geq \int_{-\infty}^{\infty} F_{Y_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1}) \, \mathrm{d}r_{i+1} \\ &\geq \int_{-\infty}^{\infty} F_{I_i}(r - r_{i+1}) f_{R_{i+1}}(r_{i+1}) \, \mathrm{d}r_{i+1} \\ &= F_{I_{i+1}}(r). \end{aligned} \tag{89}$$

This proves the claim by induction. □

Proof of Theorem 8: Let us first show that Lemma 15 applies to the WIG and the MWM for the events $W_i = K_{\mathsf{end}}[2^{n-i}] < b$. To this end we need only show that these W_i can be written in the appropriate form. Recall that we have $K_{\mathsf{end}}[2^{n-i}] = V_{i,2^i-1}$.

WIG: The WIG uses additive innovations $Z_{j,k}$ arranged on a tree as in Fig. 1. It is immediate from (18) that $K_{\mathsf{end}}[2^{n-i}]$ becomes

$$K_{\mathsf{end}}[2^{n-i}] = V_{i,2^i-1} = 2^{-i} V_{0,0} - \sum_{j=0}^{i-1} 2^{j-i} Z_{j,2^j-1}. \tag{90}$$

It suffices, thus, to set $\kappa_i = 2^i b + 2^n \widetilde{c}^{(n)}$, $R_0 = V_{0,0}$ 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 $U_{j,k}$. Recalling (20), $K_{\mathsf{end}}[2^{n-i}]$ becomes

$$K_{\mathsf{end}}[2^{n-i}] = V_{i,2^i-1} = V_{0,0} \prod_{j=0}^{i-1} (1 - U_j). \tag{91}$$

Taking logarithms, we write the events W_i in the required form by setting $\kappa_i = \ln(b + 2^{n-i}\widetilde{c}^{(n)})$, $R_0 = \ln(V_{0,0})$, and $R_i = \ln(1 - U_{i-1})$. Using (85),

$$\begin{aligned} \mathbb{P}\left\{Q_{\mathsf{end}}^{[\theta_2]} > b\right\} &= 1 - \mathbb{P}\left\{Q_{\mathsf{end}}^{[\theta_2]} < b\right\} \;=\; 1 - \mathbb{P}\{\cap_{i=0}^{n} W_i\} \\ &= 1 - \mathbb{P}\{W_0\} \prod_{i=1}^{n} \mathbb{P}\{W_i | W_{i-1}, \ldots, W_0\} \\ &\leq 1 - \prod_{i=0}^{n} \mathbb{P}\{W_i\} = P_{\mathsf{end}}^{[\theta_2]}(b). \end{aligned}$$

□

The following lemma helps prove Theorem 9.

Lemma 16: (Theorem 5 in [16]) *For a WIG model of fGn with* $1/2 < H < 1$

$$\mathrm{var}(K_{\mathsf{end}}[\tau]) \geq \mathrm{var}(K_t[\tau]) \tag{92}$$

for $t = 1, 2, \ldots, \tau$ *and for* $t = 1, \ldots, 2^n$.

Proof of Theorem 9: Note that

$$\mathbb{P}\left\{K_t[\tau] - \widetilde{c}^{(n)}\tau < b\right\} = 1 - \Phi\left(\frac{b + \widetilde{c}^{(n)}\tau - \mathbb{E}(K_t[\tau])}{\sqrt{\mathrm{var}(K_t[\tau])}}\right). \tag{93}$$

From (24) we have that

$$b + \widetilde{c}^{(n)}\tau - \mathbb{E}(K_t[\tau]) > 0. \tag{94}$$

Since the process $V_{n,k}$ is first-order stationary, $\mathbb{E}(K_{\mathsf{end}}[\tau]) = \mathbb{E}(K_t[\tau])$ for all t and τ. This fact along with Lemma 16, (93), and (94) then give

$$\mathbb{P}\left\{K_t[2^i] - \widetilde{c}^{(n)} 2^i < b\right\} \leq \mathbb{P}\left\{K_{\mathsf{end}}[2^i] - \widetilde{c}^{(n)} 2^i < b\right\}, \tag{95}$$

$i = 0, \ldots, \lfloor \log_2 t \rfloor$, $t = 1, \ldots, 2^n$. We thus have

$$P_{\mathsf{end}}^{[\theta_2]}(b) \geq P_t^{[\theta_2]}(b), \;\; \forall t = 1, \ldots, 2^n. \tag{96}$$

To complete the proof we show the following claim.

Claim 17: $\mathrm{cov}(K_t[\tau], K_t[r]) \geq 0$ for $0 \leq \tau, r \leq t$.

It is easy to show that the covariance of any two arbitrary leaf nodes is positive for a WIG model of fGn with $1/2 < H < 1$, that is $\mathrm{cov}(V_{n,k}, V_{n,l}) > 0 \;\forall k, l$. Because $K_t[\tau] = \sum_{k=t-\tau}^{t-1} V_{n,k}$ it follows that $\mathrm{cov}(K_t[\tau], K_t[r])$ is a linear combination of covariances of leaf nodes with *positive* weights. This proves Claim 17. Claim 17 and Theorem 7 give

$$P_t^{[\theta_2]}(b) \geq \mathbb{P}\left\{Q_t^{[\theta_2]} > b\right\}, \;\; \forall t = 1, \ldots, 2^n. \tag{97}$$

Combining (96) and (97) proves the theorem. □

References

[1] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," *IEEE INFOCOM*, pp. 1026–1035, Mar. 2000.

[2] K. Papagiannaki, S. Moon, C. Fraleigh, P.Thiran, F. Tobagi, and C. Diot, "Analysis of measured single-hop delay from an operational backbone network," *IEEE INFOCOM*, Jun. 2002.

[3] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control," *IEEE INFOCOM*, pp. 1233–1242, Mar. 2000.

[4] C. Fraleigh, F. Tobagi, and C. Diot, "Provisioning IP backbone networks to support delay-based service level agreements," *IEEE INFOCOM*, Apr. 2003.

[5] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[6] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.

[7] 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.

[8] J. Hüsler and V. Piterbarg, "Extremes of a certain class of gaussian processes," *Stochastic Processes and their Applications*, vol. 83, pp. 257–271, 1999.

[9] A. L. Neidhardt and J. L. Wang, "The concept of relevant time scales and its application to queuing analysis of self-similar traffic," *ACM SIGMETRICS*, pp. 222–232, Mar. 1998.

[10] M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *Computer Communication Review*, vol. 26, no. 4, pp. 15–24, Oct. 1996.

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

[12] A. Erramilli, O. Narayan, A. Neidhardt, and I. Sanjee, "Performance impacts of multi-scaling in wide area TCP/IP traffic," *IEEE INFOCOM*, pp. 352–359, Mar. 2000.

[13] K. Debicki and T. Rolski, "A note on transient Gaussian fluid models," *Queueing Systems*, vol. 41, pp. 321–342, 2002.

[14] B. Hajek and L. He, "On variations of queue response for inputs with the same mean and autocorrelation function," *IEEE/ACM Trans. on Networking*, vol. 6, no. 5, pp. 588–598, 1998.

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

[16] S. Ma, "Network traffic modeling and analyisis," *Ph.D. Thesis, Rensselaer Polytechnic Institute*, May 1998.

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

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

[19] O. Narayan, "Exact asymptotic queue length distribution for fractional Brownian traffic," *Advances in Performance Analysis*, vol. 1, no. 1, pp. 39–63, 1998.

[20] L. Massoulié and A. Simonian, "Large buffer asymptitocs for the queue with fBm input," *Journal of Applied Probability*, vol. 36, no. 3, pp. 894–906, 1999.

[21] *Self-similar network traffic and performance evaluation, Wiley Interscience*, 2001, eds. K. Park and W. Willinger.

[22] 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.

[23] M. Crouse and R. G. Baraniuk, "Fast, exact synthesis of Gaussian and non-Gaussian long-range dependent processes," *submitted to IEEE Trans. on Information Theory*, 1999.

[24] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent traffic," *IEEE/ACM Trans. on Networking*, vol. 4, no. 2, pp. 209–223, Apr. 1996.

[25] 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, no. 3, pp. 301–317, Jun. 1996.

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

[27] B. K. Ryu and A. Elwalid, "The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities," *ACM SIGCOMM*, vol. 26, no. 4, pp. 3–14, 1996.

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

[29] Y. Joo, V. Ribeiro, A. Feldmann, A. C. Gilbert, and W. Willinger, "TCP/IP traffic dynamics and network performance: A lesson in workload modeling, flow control, and trace-driven simulations," *Computer Communication Review*, vol. 31, no. 2, pp. 25–37, Apr. 2001.

[30] R. J. Adler, *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. Institute of Mathmatical Statistics Lecture Notes - Monograph Series, 1990.

[31] P. Billingsley, *Probability and Measure*. Wiley Interscience, 1995.

[32] V. I. Piterbarg, *Asymptotic methods in the theory of Gaussian processes and fields*. Translations of mathematical monographs, vol. 148, American Mathematical Society, 1996.

PLACE PHOTO HERE

Vinay J. Ribeiro (M'05) received his B.Tech. degree from the Indian Institute of Technology Madras in 1997 and his M.S. and Ph.D. degrees from Rice University in 1999 and 2005 respectively, all in electrical engineering. He was an invited fellow at the Institut Mittag-Leffler in Sweden in 2004 and has held research internship positions at AT&T Labs and Sprint Advanced Technology Labs. He received the best student paper award at the Passive and Active Measurement Workshop 2003, a Texas Instruments Fellowship from Rice University, and is a member of Eta Kappa Nu. His email address is: vinay@rice.edu.

PLACE PHOTO HERE

Rudolf H. Riedi (M'98) received the M.Sc. degree in 1986 and the Ph.D. degree in 1993, both in Mathematics, both from the Federal Institute of Technology ETH Zurich (Switzerland). From 1993-1995 he was with B. Mandelbrot at the Mathematics Department of Yale University in New Haven, Connecticut. After spending 1995-1997 with the National Research Institute in Automation and Computing INRIA in Rocquencourt near Paris, France, he joined the Electrical and Computer Engineering Department at Rice University in Houston, Texas. From there he moved to the Department of Statistics at Rice University where he is currently an associate professor. His research interests lie in the theory and practice of multiscale stochastic modeling and statistical inference, especially for networks in communication and economics. Dr. Riedi won the ETHZ Polya prize in 1986 and received a postdoctoral fellowship from the National Science Foundation of Switzerland in 1993. At Rice he has won research grants from NSF (2001, co-PI), DoE (2001, co-PI), DARPA (2000, acting PI), NSF (2003, PI) and ATP (2003, co-PI), and consulted with AT&T Labs. He served as a guest-editor for the IEEE Transaction on Signal Processing in 2002 and is on the technical program committees of the IEEE Infocom conferences 2003 through 2006, as well as the Workshop on Mobile Peer-to-Peer Computing 2005 and 2006. His email address is: riedi@rice.edu.

PLACE PHOTO HERE

Richard G. Baraniuk (F'01) received the B.Sc. degree in 1987 from the University of Manitoba, the M.Sc. degree in 1988 from the University of Wisconsin-Madison, and the Ph.D. degree in 1992 from the University of Illinois at Urbana-Champaign, all in Electrical Engineering. After spending 1992–1993 at Ecole Normale Supérieure in Lyon, France, he joined Rice University in Houston, Texas, where he is currently the Victor E. Cameron Professor of Electrical and Computer Engineering and founder of the Connexions Project. He spent sabbaticals at Ecole Nationale Supérieure de Télécommunications in Paris in 2001 and Ecole Fédérale Polytechnique de Lausanne in Switzerland in 2002. His research interests include multiscale statistical signal and image processing and networking.

Dr. Baraniuk received a NATO postdoctoral fellowship from NSERC in 1992, the National Young Investigator award from NSF in 1994, a Young Investigator Award from ONR in 1995, the Rosenbaum Fellowship from the Isaac Newton Institute of Cambridge University in 1998, the C. Holmes MacDonald National Outstanding Teaching Award from Eta Kappa Nu in 1999, the Charles Duncan Junior Faculty Achievement Award from Rice in 2000, the ECE Young Alumni Achievement Award from the University of Illinois in 2000, and the George R. Brown Award for Superior Teaching at Rice in 2001 and 2003. He was elected a Fellow of the IEEE in 2001. His email address is: richb@rice.edu.