TO APPEAR IN ICASSP 2001

WAVELETS AND MULTIFRACTALS FOR NETWORK TRAFFIC MODELING AND INFERENCE

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

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

ABSTRACT

This paper reviews the multifractal wavelet model (MWM) and its applications to network traffic modeling and inference. The discovery of the fractal nature of traffic has made new models and analysis tools for traffic essential, since classical Poisson and Markov models do not capture important fractal properties like multiscale variability and burstiness that deleteriously affect performance. Set in the framework of multiplicative cascades, the MWM provides a link to multifractal analysis, a natural tool to characterize burstiness. The simple structure of the MWM enables fast $O(N)$ synthesis of traffic for simulations and a tractable queuing analysis, thus rendering it suitable for real networking applications including end-to-end path modeling.

1. INTRODUCTION

Fractal models have made a major impact in communications, particularly in the arena of queuing analysis of data networks (such as local-area networks (LANs), wide-area networks (WANs), and the Internet). It has been convincingly demonstrated and confirmed by many studies that network traffic signals, such as the time series of number of bytes or packets arriving at a router, exhibit fractal properties such as self-similarity, burstiness, and long-range dependence (LRD) [1]. These properties are inadequately described by classical traffic models such as Poisson, Markov, and ARMA models [1], with the result that these models are far too optimistic in their predictions of performance.

Fractals are geometric objects that exhibit an irregular structure at all resolutions. Most fractals are *self-similar*; if we use a magnifying glass to "zoom" (in or out) of the fractal, we obtain a picture similar to the original. Deterministic fractals usually are constructed by predetermined iterative refinement steps and, thus, exhibit strong patterns that repeat at all scales. Real-world phenomena can rarely be described using such simple models. Nevertheless, "similarity on all scales" can hold in a statistical sense, leading to the notion of random fractals. For example the bytes per time traffic observed on a WAN when viewed on different time-scales displays a similar bursty structure (see Figure 1).

This work was supported by NSF grants CCR–9973188 and ANI–9979465, ONR grant N00014–99–10813, DARPA/AFRL grant F30602–00–2–0557, and by Texas Instruments. Email:{vinay, riedi, richb}@rice.edu. URL: www.dsp.rice.edu.

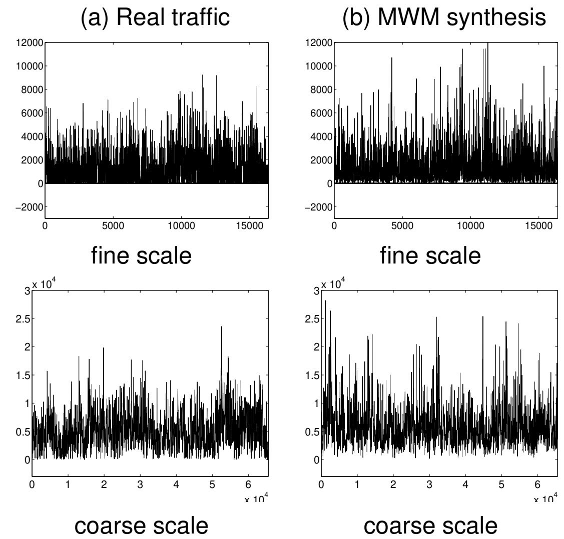

Figure 1: Modeling bursty traffic data. *Arrival processes of bytes per* $8ms$ *(top) and* $32ms$ *(bottom) for (a) real wide-area traffic [3] and (b) one realization of a multifractal wavelet model (MWM) synthesis. The MWM trace has the same bursty nature as the real data at different scales.*

As the pre-eminent random fractal model, fractional Brownian motion (fBm) has played a central rôle in many fields [1, 2]. fBm is the unique Gaussian process with stationary increments and the following scaling property for all $a > 0$

$$B(at) \stackrel{fd}{=} a^H B(t), \tag{1}$$

with the equality in (finite-dimensional) distribution. In other words, when "zoomed" into, fBm appears statistically the same up to a rescaling factor. The constant H, $0 < H < 1$, is known as the *Hurst parameter*. For $1/2 < H < 1$, fBm's increments process, fractional Gaussian noise (fGn), has an autocorrelation function that decays so slowly that it is non-summable, a property known as *long-range dependence* (LRD).

Wavelets are a powerful tool for the analysis and synthesis of LRD signals. Though LRD signals are highly correlated in the time domain, they become nearly decorrelated in the wavelet domain. Exploiting this fact, several authors have proposed wavelet-based generalizations of fGn [2]. Using efficient multiscale tree structures, these models provide fast $O(N)$ synthesis algorithms to synthesize N-point data sets. As a consequence of their Gaussian nature, however, these 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 marginal cannot capture the burstiness on small time scales that greatly affects queuing [4].

In [5], we proposed a simple multiplicative traffic model called the *multifractal wavelet model* (MWM). Set in the framework of multifractal cascades, the non-Gaussian MWM outperforms Gaussian LRD traffic models in capturing the "spiky" bursts [5] and queuing behavior of measured traffic [4]. The MWM's attractive features include linear time synthesis of traffic traces, a tractable queuing analysis, and strong multifractal properties that closely match those of real traffic. These make it viable for numerous networking applications including a novel cross-traffic estimation algorithm [6]. In this paper, we review the MWM and several of its applications.

2. WAVELETS AND LRD

The discrete wavelet transform is a multiscale signal representation of the form [7]

$$\begin{aligned} x(t) \quad = \quad & \sum_k u_k \, 2^{-J_0/2} \, \phi\big(2^{-J_0}t - k\big) \; + \\ & \sum_{j=-\infty}^{J_0} \sum_k w_{j,k} \, 2^{-j/2} \, \psi\big(2^{-j}t - k\big), \quad j,k \in \mathbb{Z} \end{aligned} \tag{2}$$

with J_0 the coarsest scale and u_k and $w_{j,k}$ the scaling and wavelet coefficients. The scaling coefficients may be viewed as providing a coarse approximation of the signal, with the wavelet coefficients providing higher-frequency "detail" information.

Wavelets serve as an approximate Karhunen-Loève transform for fBm [2], fGn, and more general LRD signals. In other words highly-correlated signals become nearly uncorrelated in the wavelet domain. In addition, the energy of the wavelet coefficients of continuous-time fBm decays with scale according to a power law [2]. While for *sampled* fBm the power-law decay is not exact [2], the *Haar* wavelet transform of fGn exhibits power-law scaling of the form[1]

$$\mathrm{var}(W_{j,k}) = \sigma^2 \, 2^{(2H-1)(j-1)} \, (2 - 2^{2H-1}). \tag{3}$$

Thus, by generating independent wavelet coefficients $W_{j,k}$ with appropriate decay of energy with scale and inverting the wavelet transform, we can synthesize Gaussian LRD processes. Gaussian processes, however, possess negative values that are unrealistic for real traffic and cannot capture the burstiness of traffic at finer scales [4].

[1] We use capital letters when we consider the underlying signal X (and, hence, its wavelet and scaling coefficients) to be random.

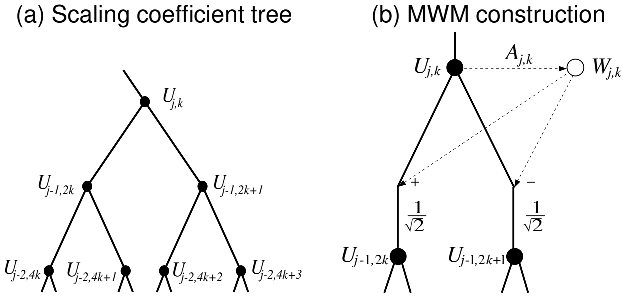

Figure 2: *(a) Binary tree of Haar scaling coefficients. (b) MWM construction: At scale j, we form the wavelet coefficient as the product $W_{j,k} = A_{j,k} U_{j,k}$. Then, at scale $j-1$, we form the scaling coefficients $U_{j-1,2k}$ and $U_{j-1,2k+1}$ as sums and differences of $U_{j,k}$ and $W_{j,k}$ (normalized by $1/\sqrt{2}$).*

3. MULTIFRACTAL WAVELET MODEL

The basic idea behind the MWM is simple. To preserve non-negativity, we use the Haar wavelet transform with special wavelet-domain constraints. To capture LRD, we mimic the wavelet energy decay as a function of scale.

3.1. Haar wavelets and non-negative Data

Before we can model non-negative signals using the wavelet transform, we must develop conditions on the scaling and wavelet coefficient values for x in (3) to be non-negative. While cumbersome for a general wavelet system, these conditions are simple for the Haar system. In a Haar transform, the scaling coefficients can be recursively computed using

$$\begin{aligned} u_{j,2k} &= 2^{-1/2}(u_{j+1,k} + w_{j+1,k}) \\ u_{j,2k+1} &= 2^{-1/2}(u_{j+1,k} - w_{j+1,k}). \end{aligned} \tag{4}$$

For non-negative signals, $u_{j,k} \geq 0, \; \forall \; j,k$, which with (4) implies that

$$|w_{j,k}| \leq u_{j,k}, \qquad \forall \; j,k. \tag{5}$$

3.2. Multiplicative model

The positivity constraints (5) on the Haar wavelet coefficients lead us to a very simple multiscale, multiplicative signal model for positive processes. Let $A_{j,k}$ be a random variable supported on the interval $[-1, 1]$ and define the wavelet coefficients recursively by

$$W_{j,k} = A_{j,k} \, U_{j,k}. \tag{6}$$

Together with (4), we obtain (see Figure 2(b))

$$\begin{aligned} U_{j,2k} &= 2^{-1/2}(1 + A_{j+1,k}) \, U_{j+1,k} \\ U_{j,2k+1} &= 2^{-1/2}(1 - A_{j+1,k}) \, U_{j+1,k}. \end{aligned} \tag{7}$$

We use beta distributions for $A_{j,k}$.

3.3. Multifractal analysis

The MWM is a multiplicative cascade [5]. Cascades are associated with a powerful tool called *multifractal analysis*, which provides a statistical language and calculus to characterize burstiness.

At the heart of multifractal analysis lies the *multifractal formalism* [5], which relates the scaling behavior of sample moments of a trace to the frequency of occurrence of "bursts" of different strength in that trace. This formalism exploits the moments of *all* orders, unlike the concept of LRD that 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 MWM over Gaussian models in modeling bursty network traffic loads. Moreover, this formalism is instrumental in relating the burstiness of traffic to the range of buffer sizes for which the multiscale queuing formula of Section 4.2 is valid [8].

4. NETWORKING APPLICATIONS

The MWM has proven to be a versatile model, mainly due to its simple structure. This section overviews some of its applications to networking.

4.1. Traffic synthesis

The MWM provides fast synthesis of traffic for simulation purposes. Starting from the top node, $U_{J_0,0}$, of the Haar scaling coefficient tree (see Figure 2(a)), the scaling coefficients at a finer scales are computed iteratively by applying (6) and (7) thus obtaining a realization of the process. In essence the algorithm simultaneously synthesizes the wavelet coefficients and inverts the wavelet transform, requiring only $O(N)$ operations to create a length-N signal.

By specifying the variances for the $A_{j,k}$, we can model the time-domain LRD or covariance structure of a signal through the energy decay of its wavelet coefficients with scale j [5]. The MWM construction guarantees decorrelated wavelet coefficients. Typically, the residual correlation between the wavelet coefficients of LRD processes is small, and therefore we can approximate the time-domain behavior of such LRD processes quite accurately.

4.2. Multiscale queuing

Data packets are multiplexed and queued at Internet routers, where they are delayed and often discarded (or dropped) due to overflow. The queuing behavior of traffic is thus crucial to network performance. Traffic models with tractable queuing analyses can help understand and ameliorate network congestion. Though the queuing analysis of classical models is well developed, most queuing results for LRD models are valid only for asymptotically large queue sizes, thus limiting their utility for real networks with finite router queue sizes.

The MWM, however, possesses a non-asymptotic queuing formula [4]. Exploiting the binary tree structure of the MWM (see Figure 2), we have developed a queuing formula

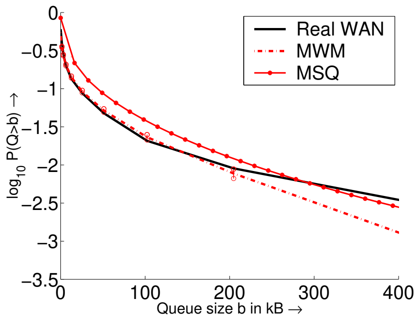

Figure 3: Accuracy of MWM queuing behavior and the multiscale queuing formula (MSQ). *The MWM matches the queuing behavior of real WAN traffic [3] accurately. The MSQ formula is an accurate approximation to the MWM queuing for all queue sizes.*

applicable to tree-based models for any finite queue size (see Figure 3).

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

$$Q = \sup_r (K[r] - rc). \tag{8}$$

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 over time scales of r time units. In the multiscale representation of the MWM model, such aggregates appear explicitly at dyadic time scales ($r = 2^m$) as the Haar scaling coefficients (up to normalization constants). We exploit the fact that the scaling coefficients are related to each other by independent random multipliers $2^{-1/2}(1 \pm A_{j,k})$ to derive an approximation to the tail queue probability $P(Q > b)$ called the *multiscale queuing formula* (MSQ) [4]. The MSQ can be directly computed from the distributions of $A_{j,k}$.

From Figure 3 observe that the MWM has tail queue probability very close to that of the real traffic and that the MSQ accurately tracks the MWM's queuing behavior. These make the MWM useful for numerous practical applications like end-to-end path modeling.

4.3. End-to-end path modeling

Packets from network connections, while traveling from one end of the Internet to the other, pass through several routers where they are multiplexed with traffic from other connections (end-to-end paths). A better understanding of the traffic dynamics on an end-to-end path will greatly benefit the design and development of future network control algorithms and protocols. Two facts make this task difficult. First, it is unrealistic to expect internal routers to determine and report these properties, because this would require their maintaining overwhelming amounts of state information. Thus it becomes necessary to infer the properties by sampling the current network state through probe packets (end-based measurements), which are relatively easy and inexpensive to

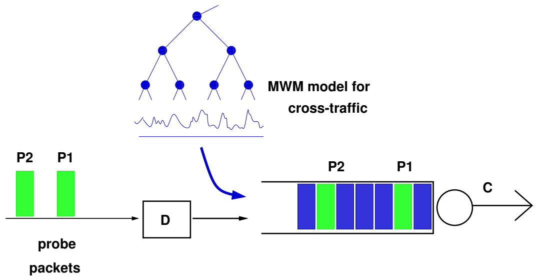

Figure 4: Cross-traffic estimation via end-to-end path modeling. *We model an end-to-end path as a single bottleneck queue fed by cross-traffic from the MWM and estimate cross-traffic volumes via efficient exponentially spaced "chirp" packet trains.*

make. Second, modeling every aspect of the several routers that comprise an end-to-end path has proved intractable, thus necessitating reduced-complexity models for end-to-end paths.

The MWM inspires a simple edge-based algorithm, *Delphi* for estimating instantaneous volumes of competing cross-traffic from delays experienced by probe packets [6]. Delphi uses a simple reduced-complexity model for an entire end-to-end path: a single bottleneck queue where probe packets are multiplexed with competing cross-traffic modeled by the MWM (see Figure 4).

Inherent in any probing scheme is an uncertainty principle or "accuracy-sparsity" tradeoff. The volume of cross-traffic entering a queue between two probes can be computed exactly from their delay spread at the receiver *provided* the queue does not empty in between. Unfortunately, this situation is guaranteed only if the probes are very closely spaced. However, sending long trains of narrowly spaced probes will congest the network and affect the very cross-traffic we are trying to measure. If probes are spaced far apart, then the queue can empty in between, which results in uncertainty in the cross-traffic volume.

Delphi balances this "accuracy-sparsity" tradeoff through a "chirp" probing packet train that matches the binary tree structure underlying the MWM (see Figure 5). The first three probes of the chirp are spaced close enough to provide exact estimates of the cross-traffic at the bottom of the binary tree, i.e., $U_{j,0}$ and $U_{j,1}$, and thus $U_{j+1,0}$. The succeeding probes are exponentially spaced to reduce the probe traffic load, thus reducing their impact on the cross-traffic. Using an approximate maximum likelihood estimator based on the MSQ, we can recursively estimate the scaling coefficients $U_{j+i,0}$ from $U_{j+i-1,0}$ and the queuing delay experienced by probe P_{i+2}. The recursion halts when a required coarsest scale is reached. Delphi does not require a priori statistics of cross-traffic and uses an adaptive algorithm to estimate the model parameters. It has performed well in simulation experiments (see Figure 6) and is being deployed in the Internet.

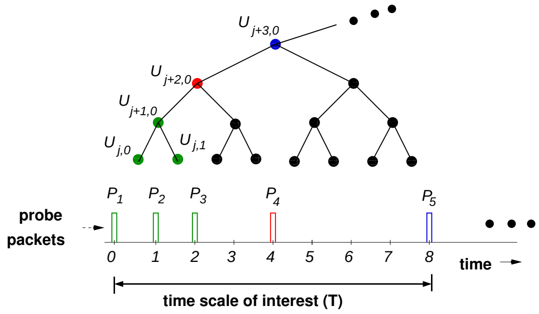

Figure 5: Chirp probe packet trains. *Probe packets are exponentially spaced to match the binary tree structure of the MWM.*

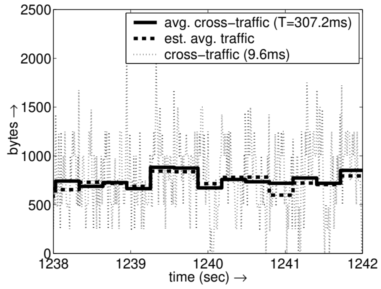

Figure 6: Cross-traffic estimates in simulation experiments. *Plotted is the cross-traffic at a fine time-scale (9.6ms) and the actual and estimated cross-traffic at a coarse time-scale (307.2ms). Observe that the estimates are accurate.*

5. CONCLUSIONS

There is a great need for new analytical tools to help understand and improve current networks. We have presented one such tool, the MWM, which though simple has proved useful for real networking applications, including traffic synthesis, queuing analysis, and end-to-end path modeling.

6. REFERENCES

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

[2] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 910–916, Mar. 1992.

[3] 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.

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

[5] R. H. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. Info. Theory*, vol. 45, no. 3, pp. 992–1018, April 1999.

[6] 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*, Sept. 2000.

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

[8] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," *IEEE Trans. Networking*, submitted.