A Unified Framework for Understanding Network Traffic Using Independent Wavelet Models

Xusheng Tian, Jie Wu and Chuanyi Ji

Abstract—**Properties of heterogeneous network traffic have been investigated from different aspects, resulting in different understanding. Specifically, one recent work discovers that the variance of network traffic exhibits a linear relationship with respect to the mean. Such a linear relation suggests that the traffic is "Poisson-like", and thus "smooth". On the other hand, prior work has shown that the heterogeneous traffic can be long-range dependent, and is thus bursty. The focus of this work is to investigate these seemingly contradictory issues, and to provide a unified understanding on the burstiness of heterogeneous traffic. In particular, we use a simple statistic, the variance of the traffic, for our investigation. We first study variance-mean relations at a single time scale. We then investigate the behavior of variances at multiple time scales, which determines the temporal correlation structure. Finally, we provide a unified view to include most important understanding of the network traffic.**

I. Introduction

Heterogeneous network traffic possesses complex statistical properties. Those properties have been investigated from different aspects, resulting in different understanding [1] [2]. Specifically, the prior work discovers that the heterogeneous traffic can be long-range dependent [3] [4] [5] [6] [7]. This implies that heterogeneous traffic possesses a strong temporal correlation across various time scales, and is thus "bursty". The multi-fractal property [8] [9] [10] of the network traffic at small time scales suggests that the traffic has even more variation than self-similarity (for comparison between self-similar and multi-fractal, please refer to [11]). On the other hand, a recent work [12] shows that the network traffic can be "smooth". This is based on a discovery that the variance of the network traffic is a linear function of the mean at a single time scale. Such a linear relationship suggests that the traffic is "*Poisson*-like" and is thus not "bursty"[1]. The above mentioned work on properties of Internet traffic has been mostly focused on the packet (or byte) counts in a fixed interval of time. Recently, an interesting study of the Internet traffic has been reported based on other traffic processes: packet size and inter-arrival time [13] [14]. Based on an empirical study of 3,026 packet traces collected from 6 monitors and mathematical theory of superposition of marked point processes, it is found that the long-range dependence of the inter-arrivals and sizes goes locally to independence as the active connection load on an Internet link increases.

Questions then arise as to whether these findings are contradictory, and whether the network traffic is bursty at all. Motivated by the prior work, we investigate these problems in order to gain a unified understanding of heterogeneous network traffic. Such a understanding is important not only to controlling and managing current networks but also to designing next generation networks.

[1]To be consistent, we use the same concept for "smooth" and "bursty" as suggested in [12] throughout the paper. In other words, if the variance and mean has linear relation, the traffic is "smooth"; if the variance and mean has quadratic relation, the traffic is "bursty".

In this work, we use a normalized byte counts process in order to study the behavior of the Internet traffic under different network loads. The normalized byte counts process is defined as the byte counts in a fixed interval of time (sampling interval) normalized by the maximum allowable byte counts in this interval. In other words, the normalized byte counts process is the average link utilization within the sampling interval.

The questions we would like to investigate are:

1. What is the relationship between the variance and the mean of the network traffic at a single time scale under different link utilization?
2. What is the relationship between the "smoothness/burstiness" and the "short-range/long-range" dependence?
3. How to combine the two relationships from the above questions to provide a unified understanding on the burstiness of the heterogeneous traffic?

We start obtaining experimental evidence to answer the first question. One issue of importance is to obtain a complete picture on the variance-mean relationship over all link utilization. We notice that a linear variance-mean relation [12] is discovered using the network measurements obtained at a non-bottleneck link and under low (about 1%) link utilization. We first extend the prior work to all link utilization to get a better understanding of the relationship. Since it is difficult to obtain network measurements under all link load conditions, we perform our studies using simulated data. The simulated data are traffic traces obtained at a link of a network simulated by network simulator *ns-2* . The network has a simplified server-client topology for web application. The workload models used to drive the simulation are similar to those in [15] [16]. Traffic data are then collected and normalized to obtain link utilization traces. The (sample) variance-mean relations are calculated using the traces.

Our results show that at a small time scale, when the sampling interval is equal to the transmission time of a single packet, the variance-mean has a perfect quadratic relation. At a single time scale, the variance-mean relation can be approximated as linear when the link utilization is relatively small or large. However, for moderate link utilization (around 50%), the non-linearity can not be ignored. Motivated by these experimental findings, we derive analytic results to provide insight. We first use a simple *Bernoulli* model to establish a quadratic variance-mean relation at a small time scale when the sampling interval is equal to the transmission time of a single packet. The maximum of the quadratic function occurs at the utilization 50%. This phenomenon, which we call limited bandwidth effect, can be understood intuitively as follows. Utilization of a link comes from two related quantities: the traffic load offered, and the bandwidth available in the network which allows/limits the traffic variation. A low utilization results in a low traffic load but full variation of the traffic. This variation, however, is intrinsically small since

the amount of traffic in the network is small. A high utilization results in a high traffic load but a small variation which is caused by the limited remaining bandwidth. A moderate utilization results from both a high enough traffic load as well as the extra bandwidth to allow its maximum variation, leading to the largest variance. We then show that the quadratic relation is indeed a upper bound of the variance-mean relation at a greater sampling interval. Our results show that the variance-mean curve is upper bounded by the quadratic function. To further demonstrate the limited bandwidth effect at a large time scale, we use a *Gaussian* model to analyze the variance-mean relation.

To answer the second question, we establish a theory among time scales using the independent wavelet model investigated in [17] [18] [19]. The key advantage of the independent wavelet model is its ability to provide a simple multiple time scale representation of heterogeneous traffic in the wavelet domain. Making use of this unique property, we derive a recursive relation between the variances of wavelet coefficients and that of cumulative traffic in the time domain. We verify this relationship analytically with DFGN process, and empirically with simulated data, and find that it provides a precise fit to the group of variance-mean relations.

Our results provide two different views to heterogeneous network traffic: the variance-mean relation within a time scale in the time domain, and the variance v.s. time scale relation at multiple time scales in the wavelet domain. We show that each view provides a unique characterization to the (bursty) heterogeneous traffic. Specifically, the first view ignores the temporal correlation [2], while the second one explicitly unveil the underlying temporal correlation. The traffic can have a short-range temporal correlation as specified through the variance-time scale relation but quadratic variance-mean relation within a given time scale. Meanwhile, the traffic can also have a long-range temporal correlation but a linear variance-mean relation within a given time scale. Therefore, the discoveries in the prior work are in fact complementary rather than contradictory. In other words, a unified understanding of heterogeneous traffic can be achieved by putting both views together. We then provide a unified view of the traffic by including most of the important traffic characteristics.

The rest of the paper is organized as follows. We first study the variance-mean relation at a single time scale based on the traffic traces obtained from the simulations in Section II. We extend our investigation from one time scale to multiple time scales to unfold the relations between the notion of burstiness and temporal correlation. We use the independent wavelet model to bridge the gap between these two concepts in Section III. In Section IV, we provide a unified view of network traffic. We conclude the paper in Section V.

II. The variance-mean relation at a single time scale

In this section, we use both simulation and analytic arguments to study the variance-mean relation at a single time scale. We start with the simulation setup in Section II-A. We present the simulation results in Section II-B. In Section II-C, we explain the simulation results analytically.

A. Simulation setup

We use the LBNL network simulator version 2 (*ns-2*) (see [20] [21] [15] and reference therein) to run our simulations. In the simulations, we use a typical dumbbell-type topology (Fig. 1) and four different kinds of workloads (Table I).

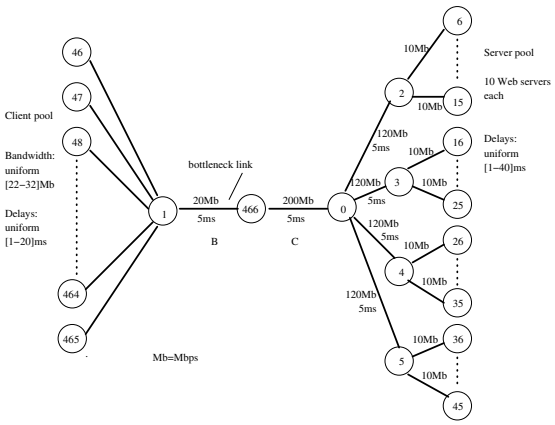

Fig. 1. Complex topology

The dumbbell-type topology is a simplified client-server topology. It is to simulate the scenario that a set of clients connected to an access network which in turn connects to a set of web servers. Similar topology has been used in [15] [16]. The network consists of a web server pool (node 6 — 45), a client pool (node 46 — 465), seven intermediate routers (or nodes) (node 0 — 5, node 466) and a number of links. The 40-node web servers consist of four groups with ten in each group. The servers in each group are connected to an access node (node 2 — 5) via 10Mbps links. The four access nodes in turn are connected to the network via 120Mbps links. 420 clients are connected to the network via 22—32Mbps links to node 1. Link B is the bottleneck of the network. The bandwidth and delay for each link are specified in the figure. The RTT[3] varies in the range of 34—150 msec, which is in consistent with the normal RTT for web applications. The average RTT is 92 msec. We use TCP Reno and HTTP 1.0 protocol stack in the simulations.

The web workload models we use are almost the same as those in [15] [16], which are similar to SURGE developed at Boston University [22]. The only difference is that we vary the number of web sessions (from 200 to 2400) to achieve different link utilization (from about 6% to 85%). A web session has a hierarchical structure. For the details of its structure, please refer to [15] [16]. Four types of workload models are used in the simulations, namely, Pareto1, Pareto2, Exp1 and Exp2, please see Table I for details. Since the simulation results for all four workload models are very similar, we only present the simulation results of Pareto1 in this paper.

B. Simulation results

B.1 Basic statistics

To explore the traffic characteristics for various average link utilization scenarios, we let the number of web sessions vary

[2]To be more accurate, the first view hide the temporal correlation.

[3]Possible queuing and processing delays are ignored.

TABLE I
WORKLOAD MODEL

Workload	Pareto1	Pareto2	Exp1	Exp2
inter-page	Pareto (mean=50, shape=2)	Pareto (mean=4, shape=1.2)	Pareto (mean=2.5, shape=2)	Exp[4]. (mean=10)
objs./page	Pareto (mean=4, shape=1.2)	Pareto (mean=3, shape=1.5)	Constant 1	Constant 1
inter-obj.	Pareto (mean=0.5, shape=1.5)	Pareto (mean=0.5, shape=1.5)	-	-
obj. size	Pareto (mean=12, shape=1.2)	Pareto (mean=12, shape=1.2)	Exp. (mean=12)	Exp. (mean=12)

from 200 to 2400. Every simulation runs 12 times with different random seeds and about 1.5 hours each time. We collect traffic traces at bottleneck link B with sample interval $T = 0.4$ msec. We then extract the stable part of the trace. The number of data points we actually used is 2^{21}, which is approximately 839 seconds or 14 minutes long. We then normalize the trace with respect to the bandwidth of link B to obtain the utilization data of each sample interval. The utilization trace enables us to compare among traffic traces obtained from different link bandwidth and/or different network topologies. Moreover, from the practical standpoint, an ISP (Internet service provider) cares more on utilization than the actual traffic.

Table II summaries the relationship between the number of web sessions and the average link utilization. Fig. 2 visualize the relationship. It is clear that their relation can roughly be considered as linear.

TABLE II
THE NUMBER OF SESSIONS V.S. LINK UTILIZATION

Sessions	Utilization (%)	90% CI
200	6.69	$\pm$1.13
400	13.71	$\pm$3.68
600	21.65	$\pm$2.32
800	29.64	$\pm$6.76
1000	36.01	$\pm$8.92
1200	42.15	$\pm$4.36
1400	51.65	$\pm$11.11
1600	57.22	$\pm$6.30
1800	60.98	$\pm$6.57
2000	70.62	$\pm$11.32
2200	78.94	$\pm$6.63
2400	82.71	$\pm$8.49

B.2 The variance-mean relation

To generate the variance-mean scatter plot, we first divide each trace into non-overlapping segments of length 512 points. We then estimate the (sample) mean and variance based on the 512-point segments to generate the scatter plot.

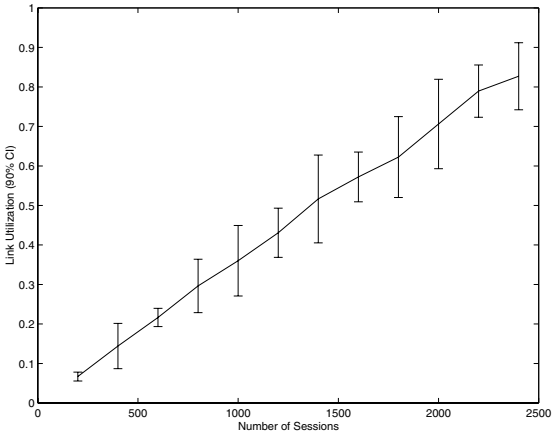

Fig. 2. The number of sessions v.s. the average link utilization (90% confidence interval)

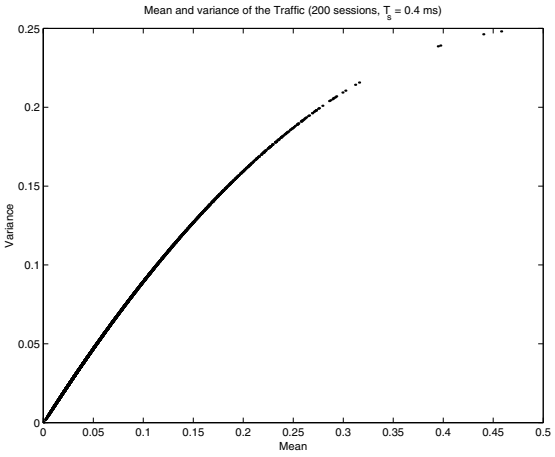

Fig. 3. Variance v.s. mean scatter plot ($T_s = 0.4$ msec, 200 sessions).

Fig. 3 illustrates the variance-mean relation for 200 web sessions. The average utilization of the trace is 6.69%, which represents a low utilization case. As we can see from the figure, it clearly depicts a near linear relation.

In Fig. 4, the variance-mean relation is shown for 1400 web sessions. This represents a moderate link utilization (51.65%). Obviously, the curve overlaps with the dashed-line, which is the perfect quadratic relation as expressed in Equ. (5).

Fig. 5 is the variance-mean relation for 2400 sessions, which represents a high utilization (82.72%) case. The curve is nonlinear, as well. Notice, however, for relatively high utilization (roughly 70%), the curve can well be approximated as linear.

From Fig. 3 to Fig. 5, we can have a sense that the variance-mean relation varies with respect to average utilization. The sample interval (T_s) for all of the three plots are $T(= 0.4$msec). The maximum value of the variances is 0.25. In fact, the three curves are all part of a quadratic relation described in Equ. 5, which we explain in detail in Section II-C.1.

B.3 The variance-mean relation at a different time scale

To see how the variance-mean relation varies with respect to time scale, we present the relations for the same set of simulations at a different sample interval ($T_s = 64T$). We use

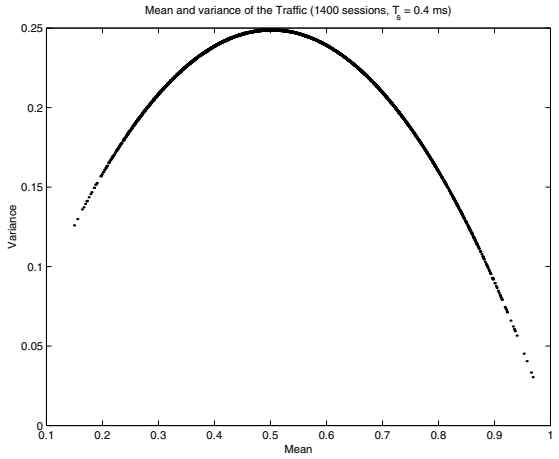

Fig. 4. Variance v.s. mean scatter plot ($T_s = 0.4$ msec, 1400 sessions).

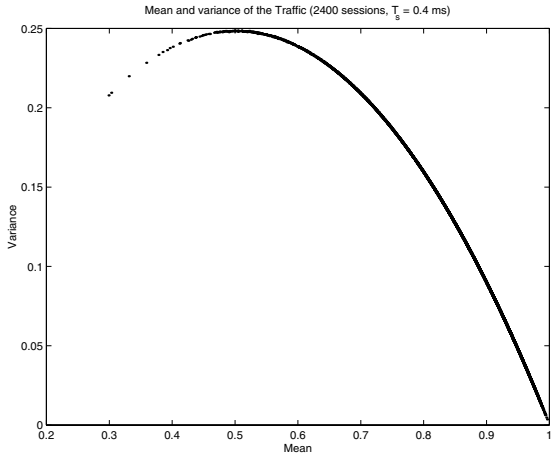

Fig. 5. Variance v.s. mean scatter plot ($T_s = 0.4$ msec, 2400 sessions).

Equ. (1) to generate traffic traces at different sampling intervals $T_s = mT$, for $m \in Z^+$.

Fig. 6, Fig. 7 and Fig. 8 are the variance-mean scatter plots for 200, 1400, and 2400 web sessions, respectively. As we can see, when the link utilization is relatively low (Fig. 6) or high (Fig. 8), the linear trend still exists. In the case of moderate link utilization (Fig. 7), a non-linear trend is shown. In all three cases, the previously observed trends still persist. But the trends are relatively vague as compared against Fig. 3 to Fig. 5. The maximum value of the variances is much smaller than 0.25.

C. Analytic explanation of variance-mean relation

In this section, we use two simple models to explain the simulation results analytically. To help with further presentation, we use the following notation to define traffic traces.

Definition 1: Let $\mathbf{X} = \{\mathbf{X}_k\}_{k=0}^{\infty}$ be the link utilization during a sample interval T, and $\mathbf{X}^{(m)} = \left\{\mathbf{X}_k^{(m)}\right\}_{k=0}^{\infty}$ be rescaled version of $\mathbf{X}$ with sampling interval mT ($m \in Z^+$), where

$$\mathbf{X}_k^{(m)} \triangleq \frac{1}{m}\sum_{i=0}^{m-1} \mathbf{X}_{km+i}. \qquad (1)$$

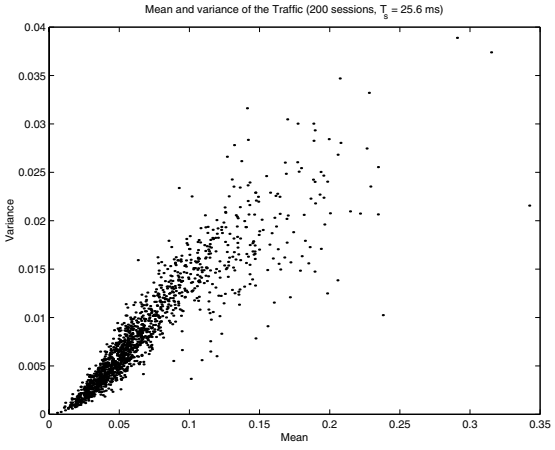

Fig. 6. Variance v.s. mean scatter plot ($T_s = 25.6$ msec, 200 sessions).

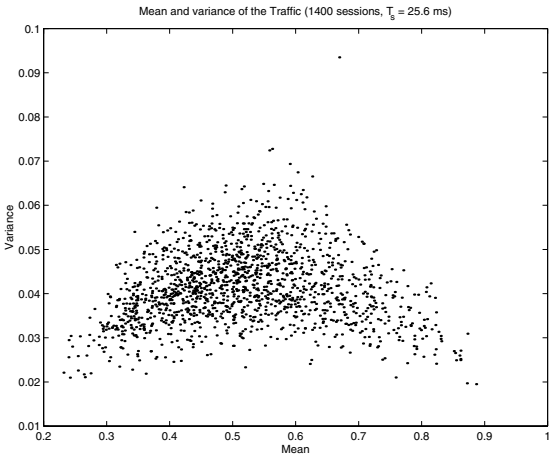

Fig. 7. Variance v.s. mean scatter plot ($T_s = 25.6$ msec, 1400 sessions).

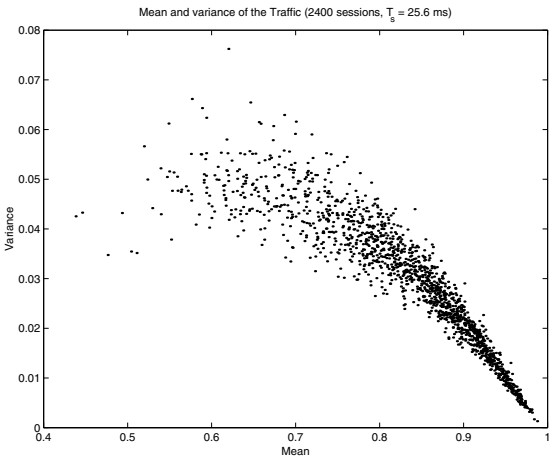

Fig. 8. Variance v.s. mean scatter plot ($T_s = 25.6$ msec, 2400 sessions).

Suppose $\mu_{\mathbf{X}} \triangleq \mathbf{E}[\mathbf{X}]$, $\sigma^2_{\mathbf{X}} \triangleq \mathbf{Var}[\mathbf{X}]$, and $R[i] \triangleq \mathbf{E}[(\mathbf{X}_k - \mu_{\mathbf{X}})(\mathbf{X}_{k+i} - \mu_{\mathbf{X}})]$ are the mean, variance, and autocorrelation of $\mathbf{X}$, respectively. Suppose $\mu^{(m)}_{\mathbf{X}} \triangleq \mathbf{E}\left[\mathbf{X}^{(m)}_k\right]$ and $\left(\sigma^{(m)}_{\mathbf{X}}\right)^2 \triangleq \mathbf{Var}\left[\mathbf{X}^{(m)}_k\right]$ are the mean and variance of $\mathbf{X}^{(m)}_k$, respectively. Based on the definition, one can easily verify the following properties.

$$\mu^{(m)}_{\mathbf{X}} = \mu_{\mathbf{X}}, \tag{2}$$

$$\left(\sigma^{(m)}_{\mathbf{X}}\right)^2 = \frac{1}{m^2}\sum_{i=0}^{m-1}\sum_{j=0}^{m-1} R[i-j], \tag{3}$$

and

$$\left(\sigma^{(m)}_{\mathbf{X}}\right)^2 \leq \sigma^2_{\mathbf{X}}. \tag{4}$$

Intuitively, Equ. (2) tells that when the sampling interval is mT, the traffic trace $\mathbf{X}^{(m)}$ is the link utilization, too. Equ. (3) depicts that the variance of $\mathbf{X}^{(m)}$ is the accumulation of autocorrelation of $\mathbf{X}$. In other words, the variance of $\mathbf{X}^{(m)}$ has included the autocorrelation of $\mathbf{X}$. If $\mathbf{X}$ is uncorrelated, i.e. $R[i] = \sigma^2_{\mathbf{X}}\delta[i]$, Equ. (3) can be further simplified as $\left(\sigma^{(m)}_{\mathbf{X}}\right)^2 = \frac{1}{m}\sigma^2_{\mathbf{X}}$, which means that under the above assumption the traffic variation becomes smaller as the sampling interval increases. Equ. (4) gives a loose bound (see Fig. 3 to Fig. 8 to verify) on the variance of rescaled traffic. Intuitively, it means that as the sampling interval increases, the variance of the underlying traffic decreases. In general, given more information about the autocorrelation of the traffic, a tighter bound can be obtained.

In Section II-C.1, we use a *Bernoulli* model to character the transmission of packet at the very small time scale when there can be either one or none packet transmitted within the sample interval. In Section II-C.2, we use a *Gaussian* model to examine the behavior of traffic at a large time scale when the central limit theorem becomes valid.

C.1 *Bernoulli* model

The motivation on using Bernoulli model is to capture the variance-mean relation within the smallest time scale[5].

Assume packets transmitted over a link have a fixed size, and the sample interval is exactly equal to the transmission time of a packet. Therefore, there is either one or none packet transmitted in any sample interval. The link utilization of a sample interval can be modeled by a random variable $\mathbf{X} = \{1, 0\}$ with *Bernoulli* distribution, where $\Pr\{\mathbf{X} = 1\} = p$ is the probability that there is a packet transmitted and $\Pr\{\mathbf{X} = 0\} = 1 - p$ is the probability there is no packet transmitted. We have the mean $\mu_{\mathbf{X}} = p$, and the variance $\sigma^2_X = p(1-p)$. Therefore,

$$\sigma^2_X = \mu_{\mathbf{X}} - \mu^2_{\mathbf{X}}. \tag{5}$$

Equ. (5) clearly explains why the variance-mean relation is quadratic instead of linear at small time scales. Another important finding from Equ. (5) is that when $\mu_{\mathbf{X}} = 0.5$, we have $\sigma^2_X = 0.25$ at its peak; when $\mu_{\mathbf{X}} < 0.5$, the variance increases with the mean; and, when $\mu_{\mathbf{X}} > 0.5$, the variance decreases with the mean. Intuitively, it suggests that the maximum burstiness is allowed when link utilization is at 50%. The above relation is plotted as dashed-line in Fig. 4. The dashed-line overlaps with the actual variance-mean scatter plot (dots) since in this case the sample interval is exactly equal to the time a packet is transmitted across the link [6].

In fact, based on the upper bound provided in Equ. (4), the quadratic relation shown in Equ. (5) is an upper bound to sampling interval mT. It is clearly true from our simulation experiments as shown from Fig. 3 to Fig. 8.

If independent arrival of packets is further assumed, $\mathbf{X}^{(m)}$ will have binomial distribution, with $\Pr\left\{\mathbf{X}^{(m)} = \frac{k}{m}\right\} = \binom{m}{k} p^k (1-p)^{m-k}$, and $\left(\sigma^{(m)}_{\mathbf{X}}\right)^2 = \frac{1}{m}\mu^{(m)}_{\mathbf{X}}\left(1 - \mu^{(m)}_{\mathbf{X}}\right)$. Clearly, the independence assumption is far from realistic, that is why maximum value of variances is greater than the upper bound ($0.25/64 \approx 3.9e-3$) predicted by the independent arrival as the sampling interval increases (shown in Fig. 6 to 8).

C.2 *Gaussian* model

Gaussian distribution is a feasible model for aggregated traffic obtained at a sufficiently large time scale, and thus used here to analyze the effect of limited link capacity on variance-mean relations.

Intuitively, finite link capacity limits traffic variation especially at the high utilization. Quantitatively, to illustrate that a limited link bandwidth itself can contribute to the change of the variance-mean relation, we use the following *Gaussian* model.

Suppose that the number of arrivals $\mathbf{X}$ (packet counts or byte counts) at a fixed time interval is *Gaussian* distributed with mean $\mu > 0$ and variance σ^2. Suppose

$$\mathbf{Y} = \begin{cases} \mathbf{X}/\Pr\{\mathbf{X} > 0\} & \text{if } \mathbf{X} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

is the positive part of $\mathbf{X}$. $\mathbf{Y}$ is then fed into a bufferless server with maximum service capacity of $C > \mu$ per time interval. We want to study the relationship of variance and mean on the output ($\mathbf{Z}$) side of the server.

$$\mathbf{Z} = \begin{cases} \mathbf{Y} & \text{if } \mathbf{Y} < C, \\ C & \text{otherwise.} \end{cases} \tag{7}$$

The analytic forms of the mean (μ_Z) and variance (σ^2_Z) have been derived, but are rather complex. We visualize their relationship in Fig. 9. The curves are normalized with respect to link capacity (C). $\alpha = \mu/C$ is the normalized input load. As we can see from the figure, the variance-mean depicts non-linear relation for all three different input loads. Another observation is that at small and large utilization, the non-linear relation can be approximated as linear.

D. Summary

To summarize, as a result of limited bandwidth, the variance-mean has a non-linear relation in general. The non-linear relation is upper bounded by a quadratic relation shown in Equ. (5).

[5] As will soon be discussed, temporal correlation only has no effect within a time scale.

[6] Please be aware that *ns-2* generates data packets with fixed size, namely 1000 bytes.

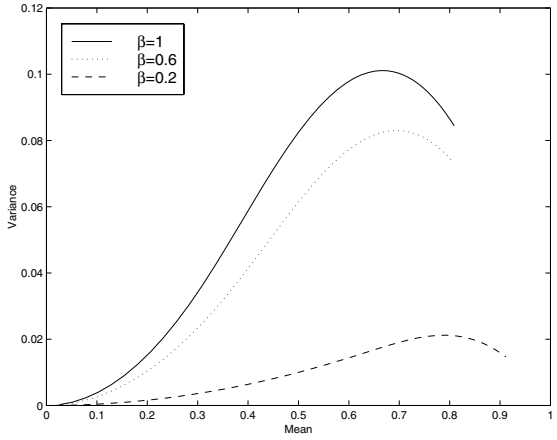

Fig. 9. Variance-mean relation: output. ($\beta = \frac{\sigma}{\mu}$)

For small or large link utilization, the non-linear relation can be approximated as linear. But, for moderate link utilization, the non-linearity can not be ignored. The variance-mean relation is non-linear at small time scales, while the relation becomes vague at large time scales. This implies that at a single fixed time scale (sampling interval) the traffic can be consider as either bursty or smooth depending on the range of link utilization and the sample interval rather than the temporal correlation. The linear relation discovered in previous work is an approximation of the non-linear relation for relatively small utilization.

III. Variances at multiple time scales

In the previous section, we have shown through both simulation and analytic arguments that at small time scales the variance-mean has non-linear relation and at large time scales the non-linear relation is vague. However, the variances information at one time scale is not enough to capture one of the key characteristics of the traffic, namely, the temporal correlation. One way to consider the temporal correlation of the traffic is to study the variances of the traffic at multiple time scales. In this section, we study the variances of the traffic at multiple time scales in both time and wavelet domain. It has been shown [23] that the variance-time relation is equivalent to the relation of variances of wavelet coefficients v.s. time scale. We use the independent wavelet model to investigate the relationship of different variances.

A. Wavelets and the independent wavelet models

We begin by introducing wavelets and the independent wavelet models [24]. The network traffic has complex temporal correlation structure: both short-range and long-range properties exist. But the complex correlation structure in the time domain is found to be simple in the wavelet domain. In fact, a simple independent model of the wavelet coefficients can provide sufficiently good performance. As a result, the synthesized traffic from the independent wavelet model has the same temporal correlation asymptotically, and generates queuing results close to original traffic.

Let $x(t)$ be a random process generated from an independent wavelet model for discrete time t ($t \geq 0$). That is, through the inverse wavelet transform, we have

$$x(t) = \sum_{j=1}^{\infty} \sum_{m=0}^{\infty} d_j\,[m]\,\psi_{j,m}\,(t) + \psi_0, \tag{8}$$

where $\psi_{j,m}\,(t)$ and $d_j\,[m]$ are the wavelet basis functions[7] (see [18] [25] for details) and the corresponding wavelet coefficients, respectively, at the time scale j ($j \geq 1$, and an integer) and shift m ($m \geq 0$, and an integer). The wavelet basis functions are obtained by dilating and translating a wavelet function $\psi(t)$, where $\psi_{j,m}\,(t) = 2^{-j/2} \cdot \psi(2^{-j} \cdot t - m)$, and in particular for *Haar* wavelet, $\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2, \\ -1 & \text{if } 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$ For the shape of the *Haar* wavelet, please refer to Fig. 10.

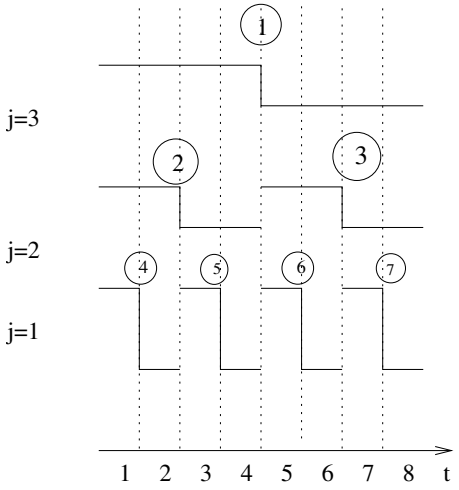

Fig. 10. The *Haar* wavelet functions

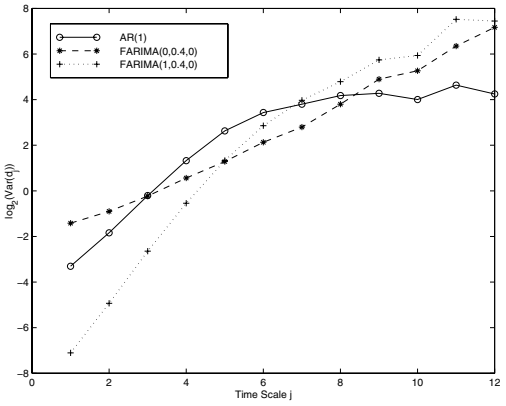

Fig. 11. $\log_2 \sigma_j^2$ v.s. j for traces with different temporal correlation structures. Solid-line: AR(1); dashed-line FARIMA(0,0,4,0); dotted-line: FARIMA(1,0.4,0).

The wavelet coefficients, $d_j\,[m]$'s, can be obtained through the wavelet transform

$$d_j\,[m] = \sum_{t=0}^{+\infty} x(t)\psi_{j,m}\,(t)\,. \tag{9}$$

[7] ψ_0 represents the mean of $x(t)$, for the *Haar* wavelet. In general, ψ_0 is the projection of $x(t)$ to the coarsest time scale.

The independent wavelet models are defined as below.

Definition 2: In the independent wavelet models (IWMs), the wavelet coefficients $d_j[k]$'s are assumed to be independent random variables. For a given j ($j \geq 1$), $d_j[k]$'s are independent and identically distributed (i.i.d.) random variables with a zero mean and a variance σ_j^2. That is,

$$\mathbf{E}\left[d_{j_1}[k_1]\, d_{j_2}[k_2]\right] = \begin{cases} \sigma_j^2 & k_1 = k_2 \text{ and } j_1 = j_2 = j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

To illustrate how the variances of the wavelet coefficients σ_j^2 characterize the temporal correlation of the traffic, Fig. 11 depicts $\log_2 \sigma_j^2$ v.s. j for three typical types of traffic: AR(1) represents short-range dependent traffic; FARIMA(0,0.4,0) represents pure long-range dependent traffic; while FARIMA(1, 0.4,0) represents traffic with both short and long range dependency.

B. The relationship among variances in time and wavelet domain

The relationship among variances in both time and wavelet domain is summarized as follows.

Theorem 3: Suppose $\left(\sigma_X^{(m)}\right)^2 = \mathbf{Var}\left[\mathbf{X}^{(m)}\right]$ is the variance of rescaled random process $\mathbf{X}^{(m)}$ as defined in Equ. (1), and $\left(\sigma_j^{(m)}\right)^2$ is the variance of *Haar* wavelet coefficient at time scale j when the sampling interval is mT. The following relation holds true

$$\left(\sigma_k^{(1)}\right)^2 = 2^k \cdot \left(\left(\sigma_X^{\left(2^{k-1}\right)}\right)^2 - \left(\sigma_X^{\left(2^k\right)}\right)^2\right). \tag{11}$$

Proof of the theorem is omitted due to page limit. The significance of the Theorem is that based on the independent (*Haar*) wavelet model, Equ. (11) provides a unified view of variances both in time and wavelet domain for different sampling intervals. It is noteworthy to mention that Equ. (11) is to be understood as that the difference of the (sample) variances in time domain is determined by the variances of wavelet coefficients, not the other way around. The variances of wavelet coefficients are determined by the nature of traffic itself. One can, of course, calculate the variances of the wavelet coefficients based on the (sample) variances in the time domain.

From the non-negativity property of Equ. (11), one immediate conclusion is that as the sampling interval increases, the variance of the measured traffic decreases. Using this property and σ_j^2, one can provide a tighter upper bound for variances in time domain.

Substitute Equ. (3) into Equ. (11), we have

$$\left(\sigma_k^{(1)}\right)^2 = \frac{1}{m}\left(\sum_{i=0}^{m-1}(2m-3i)\,R[i] - \sum_{i=m}^{2m-1}(2m-i)\,R[i]\right) - R[0],$$

where $m = 2^{k-1}$. It is clear that the variances of the wavelet coefficient have included the autocorrelation of $\mathbf{X}$.

A simple example to verify Equ. (11) is as following. If $\mathbf{X}$ is from discrete Fractional Gaussian Noise (DFGN), it is known that [26]

$$\left(\sigma_X^{(m)}\right)^2 = \sigma_X^2 m^{2H-2}, \tag{12}$$

and

$$\left(\sigma_j^{(1)}\right)^2 = \sigma_X^2\, 2^{j(2H-1)}\left(2^{2-2H}-1\right), \tag{13}$$

where H ($0.5 \leq H < 1$) is the *Hurst* parameter. Through some algebraic manipulation, it is not difficult to verify that Equ. (11) holds for the case of DFGN.

To show how accurate Equ. (11) is when used to estimate variances of the wavelet coefficients, we compared the estimate against the definition of variances of wavelet coefficients using all the simulation traces. The relative error of Equ. (11) is less than 2.5% for all traces.

In Fig. 13, we plot the variance-mean relation for different sampling intervals with 95% confidence interval. As we can see from the figure, when the sampling interval is fixed, the variance-mean exhibits quadratic relation at small time scales. When we look across different sampling intervals, it is obvious that as we increase the sampling interval, the variances decrease. For all the curves, the upper bound provided in Equ. (4)(5) is always true.

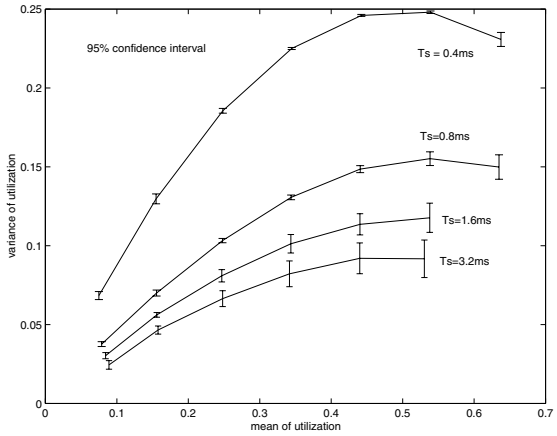

Fig. 12. The variance-mean relation at different time scales ($T_s = 0.4$ msec, 0.8 msec, 1.6 msec, and 3.2 msec from top down).

Fig. 13 depicts the variances of wavelet coefficients v.s. time scale at different sampling intervals, which demonstrates the correctness of Equ. (11). It is the counterpart of Fig. 12 in wavelet domain. More interestingly, it is clear from Fig. 13 that $j = 9$ separates the plot into two parts: to the left, $\log_2 \sigma_j^2$ varies a lot; to the right, the curves depict constant slope (i.e. the Hurst parameter). In terms of time, $j = 9$ is equivalent to 102.4 msec ($= 2^{9-1}T$). If we recall that the average RTT is 92 msec for our traces, $j = 9$ is actually the closest time scale to average RTT. The demarcation by average RTT is consistence with findings of the multifractal feature of network traffic.

IV. A UNIFIED VIEW

In previous two sections, we discussed variances at a single time scale and relationship of variances at multiple time scales in both time and wavelet domain. In this section, we put different views of the traffic together in a unified framework.

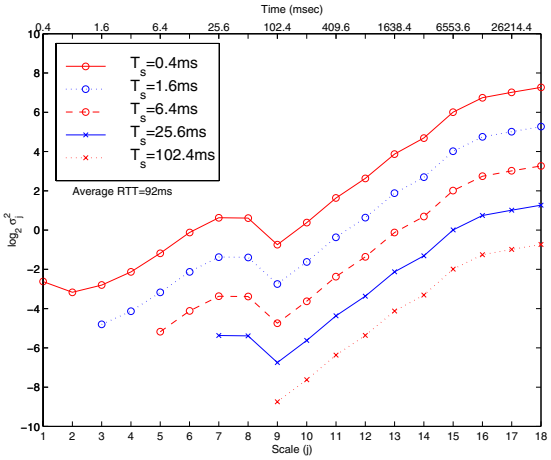

Fig. 13. The variances of wavelet coefficients v.s. time scales: average RTT=92 msec ($T_s = 0.4$ msec, 1.6 msec, 6.4 msec, 25.6 msec, and 102.4 msec from top down).

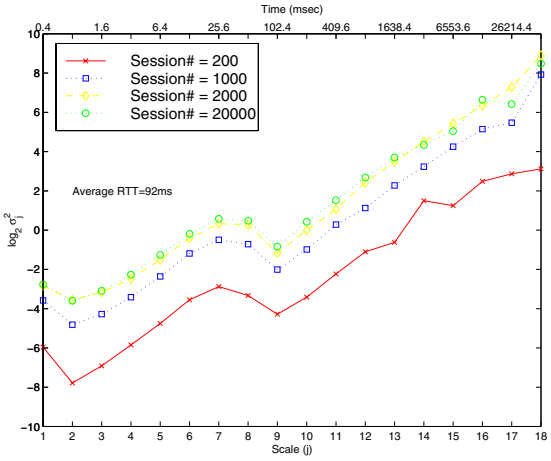

Fig. 14. The variances of wavelet coefficients v.s. time scales: average RTT=92 msec, $T_s = 0.4$ msec.

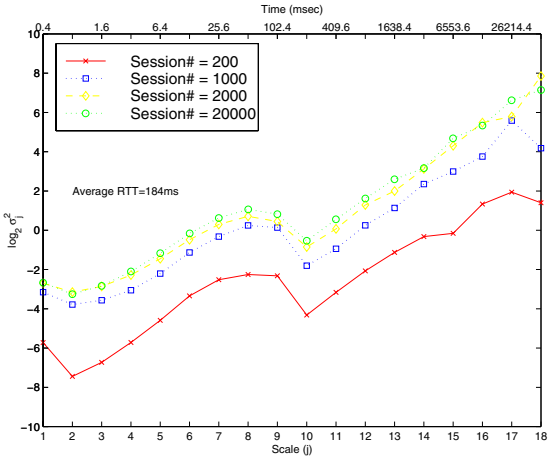

Fig. 15. The variances of wavelet coefficients v.s. time scales: average RTT=184 msec $T_s = 0.4$ msec.

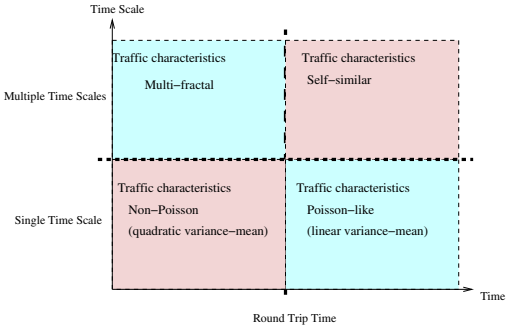

Fig. 16. The unified view of network traffic.

The unified view of network traffic is shown in Fig. 16. Traditional voice traffic characterized by *Poisson* models belongs to the lower-right corner of the framework. As the network carries more and more data traffic, it is insufficient to study traffic at a single time scale. Therefore, the *Poisson* model fails and the self-similar property of traffic becomes a common sense in multiple time scales perspective. This covers the upper-right part of the figure.

As we gets down to finer time scales, the variation of traffic has become even more severe than self-similar. This brings us to the upper-left corner of the figure, where multi-fractal and non-Gaussian are among the key characteristics of the traffic. In this paper, we fill the lower-left corner of the figure by unfolding the quadratic variance-mean relation of the traffic at a single small time scale.

We use the RTT to demarcate the left and right part of the figure. It is obvious from Fig. 13 that average RTT clearly separates the large time scales and small time scales. A. Feldmann, et al. [9] use port-to-port flows to separate the two different time scales. They suggest to use RTT in future work as well. It is noted that the average RTT may not be the only criteria to separate small and large time scales behavior of the traffic since the topology used to the conclusion is simple compared with real networks. Further investigation based on actual network measurement is needed to verify the result. The critical time scale may be a good candidate to be the criteria to separate small and large time scales as well. By any means, we call the time scale which demarcates the small and large times as the "best" time scale.

This work and [12] [13] [14] brought us back to where we started, the lower-right corner. However, this should not be understood as we go back to where we were and disregard the multiple time scales nature of the traffic. At the "best" time, use the traditional *Poisson* approach. For example, bufferless multiplexing of streaming flows has made the loss rate not depend on temporal correlation [27], which provide a chance to use traditional *Poisson* approach at flow level. In this paper, we bridge the single and multiple time scale(s) view with the independent wavelet models.

V. Conclusion

We have investigated in this work on how to provide a unified understanding of heterogeneous traffic. We have shown that the concept of "smooth" and "bursty" is about traffic at a single

time scale, while the long-range dependence is the characteristic of traffic at multiple time scales. In particular, we have shown through both simulation studies and analytic developments that the traffic at a small time scale has quadratic variance-mean relation, and therefore is "bursty". Nevertheless, the linear approximation of variance-mean relation is valid for low or high link utilization, or large time scales, where the traffic can be treated as if it were "smooth" within a given time scale.

We then obtain a relationship of the variances at multiple time scales using the independent wavelet model. The wavelet model essentially provides a framework to combine two different views: one is for variance-mean within a given time scale, and the other is for variance-time scale relation at multiple time scales. As a result, the two seemly different discoveries/concepts are actually complementary to each other: the concept of "smoothness/burstiness" at a single time scale does not take the temporal correlation of the traffic into consideration, whereas the concept of "short-range/long-range dependence" captures the temporal correlation through variances across all time scales.

A unified view of heterogeneous traffic is then a provided. At large time scales, traffic is self-similar, Gaussian and has *Poisson*-like variance-mean relation. While at small time scales, traffic has more variations with multi-fractal, non-Gaussian, and quadratic variance-mean relation. The average RTT is the key to demarcate small and large time scales. The network traffic itself has multiple time scales properties: self-similar or multi-fractal. It does not necessarily mean the end of the traditional traffic engineering.

As the importance of average RTT needs further confirmation by actual network measurement, we are actively working on the issue. Research efforts are expected to apply the current understanding to help network design and operation.

REFERENCES

[1] Bong K. Ryu and Anwar Elwalid, "The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities," in *Proc. ACM SIGCOMM'96*, San Francisco, CA, 1996, pp. 3–14.

[2] Matthias Grossglauser and Jean-Chrysostome Bolot, "On the relevance of long-range dependence in network traffic," in *Proc. ACM SIGCOMM'96*, San Francisco, CA, 1996.

[3] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, February 1994.

[4] Mark W. Garrett and Walter Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc. ACM SIGCOMM'94*, London, England, UK, 1994, pp. 269–280.

[5] Vern Paxson and Sally Floyd, "Wide-area traffic: The failure of Poisson modeling," in *Proc. ACM SIGCOMM'94*, London, England, UK, 1994, pp. 257–268.

[6] Diane E. Duffy, Allen A. Mcintosh, Mark Rosenstein, and Walter Willinger, "Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks," *IEEE J. Select. Areas Commun.*, vol. 12, no. 3, pp. 544–551, 1994.

[7] Mark E. Crovella and Azer Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, December 1997.

[8] Petteri Mannersalo and Ilkka Norros, "Multifractal analysis of real ATM traffic: a first look," Tech. Rep. COST257TD(97)19, VTT Information Technology, January 1997.

[9] A. Feldmann, A.C. Gilbert, and W. Willinger, "Data networks as cascades: investigating the multifractal nature of Internet WAN traffic," in *Proc. ACM SIGCOMM'98*, Vancouver, B.C., Canada, 1998, pp. 42–55.

[10] Rudolf H. Riedi, Mattew S. Crouse, Vinay J. Ribeiro, and Richard G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. Inform. Theory*, vol. 45, no. 3, pp. 992–1018, April 1999.

[11] R. H. Riedi and W. Willinger, "Toward an improved understanding of network traffic dynamics," in *Self-similar Network Traffic and Performance Evaluation*. Wiley, June 1999.

[12] Robert Morris and Dong Lin, "Variance of aggregated web traffic," in *Proc. IEEE INFOCOM'2000*, Tel Aviv, Israel, 2000, pp. 360–366.

[13] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun, "Internet traffic tends to Poisson and independent as the load increases," Tech. Rep., Bell Labs, 2001.

[14] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun, "The effect of statistical multiplexing on Internet packet traffic: theory and empirical study," Tech. Rep., Bell Labs, 2001.

[15] Anja Feldmann, Anna C. Gilbert, Polly Huang, and Walter Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," in *Proc. ACM SIGCOMM'99*, Cambridge, MA, 1999.

[16] Youngmi Joo, Vinay Ribeiro, Anja Feldmann, Anna C. Gilbert, , and Walter Willinger, "On the impact of variability on the buffer dynamics in IP networks," in *Proc. of 37th Annual Allerton Conf. on Comm., Control, and Computing*, 1999.

[17] Sheng Ma and Chuanyi Ji, "Modeling video traffic in the wavelet domain," in *Proc. IEEE INFOCOM'98*, San Francisco, CA, April 1998, vol. 1, pp. 201–208.

[18] Chunayi Ji, Sheng Ma, and Xusheng Tian, "Approximation capability of independent wavelet models to heterogeneous network traffic," in *Proc. IEEE INFOCOM'99*, New York, NY, March 1999, vol. 1, pp. 170–177.

[19] Sheng Ma and Chunayi Ji, "Modeling heterogeneous network traffic in wavelet domain," *IEEE/ACM Trans. Networking*, vol. 9, no. 5, pp. 634–649, October 2001.

[20] "http://www.isi.edu/nsnam/ns," .

[21] Kihong Park, Gi Tae Kim, and Mark E. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," in *Proc. of the Fourth International Conference on Network Protocals*, October 1996.

[22] P. Barford and M. E. Crovella, "Generating representative Web workloads for network and server performance evaluation," in *ACM SigMetrics'98*, Madison, WI, 1998, pp. 151–160.

[23] Darryl Veitch and Patrice Abry, "A wavelet based joint estimator of the parameters of long-range dependence," *IEEE Trans. Inform. Theory*, vol. 45, no. 3, pp. 878–897, April 1999.

[24] Sheng Ma and Chunayi Ji, "Modeling heterogeneous network traffic in wavelet domain," *IEEE/ACM Trans. Networking*, vol. 9, no. 5, pp. 634–649, October 2001.

[25] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia: SIAM, 1992.

[26] David L. Jagerman, Benjamin Melamed, and Walter Willinger, "Stochastic modeling of traffic processes," in *Frontiers in queueing*, chapter 10, pp. 271–320. 1997.

[27] Jim W. Roberts, "Traffic theory and the Internet," *IEEE Communications Magazine*, vol. 39, no. 1, pp. 94–99, January 2001.