

THE IMPACT OF BUFFER AND BANDWIDTH ON THE SCALING BEHAVIOR OF NETWORK TRAFFIC

Guoqiang Mao

School of Electrical and Information Engineering
The University of Sydney

Abstract—The effect of network mechanisms (i.e. buffer, bandwidth, statistical multiplexing and traffic control protocols) on traffic is on two aspects: first they contribute to network performance by limiting the timescale range of interest for performance analysis; second they shape the traffic, i.e. changing the scaling behavior of traffic. While some research has been done on the first aspect in terms of critical timescale, the second aspect has not been given enough attention. This paper investigates the effect of buffer and bandwidth on the scaling behavior of traffic. Our analysis shows that they have different effect on traffic at small timescale and large timescale. Both buffer and bandwidth can significantly affect the small timescale scaling exponent, but they seem ineffective on the large timescale scaling exponent. Moreover the use of smaller buffer size and less bandwidth will reduce the energy of traffic at small timescale. Therefore traffic may become less bursty.

I. INTRODUCTION

It is well known that some characteristics of Internet traffic fall beyond the conventional framework of Markov traffic modelling. Leland et al. discovered the existence of traffic self-similarity in a local area network (LAN) environment [1]. Beran et al. demonstrated self-similarity in variable-bit-rate (VBR) video traffic [2] and Crovella et al. showed self-similarity for WWW traffic [3]. Recent measurements further revealed that WAN traffic has complex multifractal characteristics on small timescales, and is self-similar on large timescales [4], [5].

There are a number of different, not equivalent, definitions of self-similarity. Refer to [6] for definitions of self-similarity. A measure of self-similarity is the Hurst parameter H .

Long-range dependence (LRD) is another widely used term in this area. Let the mean and the covariance function of a stationary sequence $X(t)$ be denoted by $\mu = E[X(t)]$ and $C_X(k) = E[(X(t+k) - \mu)(X(t) - \mu)]$. An LRD sequence can be defined via a slow, power-law decay of $C_X(k)$: $C_X(k) \sim C_\gamma k^{-\beta}$, $0 < \beta < 1$, where C_γ is a finite positive constant, and the symbol \sim means that the ratio of the two sides tends to one in the limit of large k . β is related to the Hurst parameter by $H = 1 - \beta/2$. Equivalently, LRD can also be defined via a power-law divergence at the origin of its spectrum:

$$f_x(v) \sim c_f |v|^{-\alpha}, \quad |v| \rightarrow 0 \quad (1)$$

where $f_x(v)$ satisfies, in the case of discrete time processes

$$\sigma_x^2 = \int_{-1/2}^{1/2} f_x(v) dv, \quad (2)$$

σ_x^2 being the variance (or power) of $X(t)$ [7]. Parameter c_f is the frequency domain equivalent of C_γ . Parameter α is related to the Hurst parameter by $\alpha = 1 - \beta = 2H - 1$.

Self-similarity and long-range dependence are different concepts. However in the context of network traffic analysis, they both refer to the fact that the cumulative effect of long-term correlations of a traffic process cannot be ignored. Therefore they are often used interchangeably.

Despite the well established presence of the scaling phenomenon, its impact on teletraffic issues and network performance is still the subject of some confusion and uncertainty. Specifically, although network traffic may exhibit the scaling behavior across a very wide range of timescales, it never exists alone. These scaling properties must exert their influence through a network of finite dimension, i.e. a network of finite size, finite capacity and finite queue. Moreover, traffic coming from different sources may mutually interact and this traffic will also be subject to the restrictions of traffic and congestion control protocols. All these network mechanisms (i.e. finite buffer, finite bandwidth, statistical multiplexing and traffic and congestion control protocols) cast a limit on the impact of traffic scaling. As a result, we only need to consider a finite timescale range when performing performance analysis [8]. Moreover, as traffic passes through the network, these network mechanisms may also shape the traffic, i.e. change the scaling behavior of traffic. Some earlier work exists in the area [9], [10], [11], [12]. However most of the work focuses on the contribution of network mechanisms to the finite timescale range of interest. The contribution of the network to shaping the traffic has not received much attention.

In this paper, we investigate the impact of some network mechanisms on shaping the traffic. Specifically, we investigate the impact of buffer and bandwidth on the scaling behavior of traffic. The rest of the paper is organized as follows: in section II, we present some qualitative analysis on the effect of buffer and bandwidth; analysis tool used in the paper is introduced in section III; section IV gives a brief introduction to the traffic trace used in our analysis; the effect of buffer and the bandwidth limiting effect are analyzed in section V and section VI respectively; and finally some conclusions are given in section VII.

II. QUALITATIVE ANALYSIS

In this section, we present some qualitative analysis on the effect of buffer and bandwidth on the scaling behavior of traffic.

A. Finite Buffer

Consider a buffer with capacity B and receives input at deterministic time. Let X_i be the number of arrivals at discrete time T_i . Let d be the number of traffic that is processed during

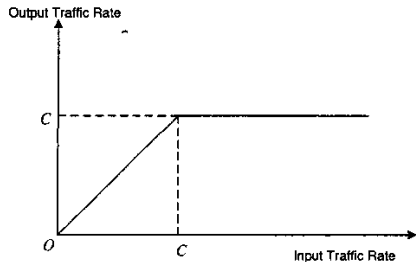


Fig. 1. Relationship between input and output traffic rate in a queue with a link capacity C

$[T_i, T_{i+1})$, referred to as the i^{th} interval, and let V_i be the buffer content at the end of the i^{th} interval. Then,

$$V_i = \min \left\{ (V_{i-1} + X_i - d)^+, B \right\}. \quad (3)$$

Analysis on (3) reveals that only traffic arriving in the same busy period T_{busy} interacts with each other and determines the buffer content in the i^{th} interval. Traffic arriving in a different busy period has no impact on buffer content in the i^{th} interval although it may be statistically correlated with traffic in the i^{th} interval. This is referred to as the *resetting effect* [13]. Resetting effect appears whenever buffer content reaches zero. It introduces a “break” in the impact of traffic correlations. Moreover, when buffer content becomes full another effect, referred to as the *truncating effect*, emerges. Truncating effect enhances resetting effect and they both diminish the performance impact of traffic correlations. Another effect of the finite buffer is that every busy period containing an overflow is shorter than the corresponding busy period in an infinite buffer version. This effect and its performance impact have been shown both graphically and analytically in our paper [8].

In addition to its performance impact, a finite buffer also contributes to shaping the traffic. A finite buffer has the effect of a low-pass filter in network. At small timescales, it may effectively smooth traffic and remove some small timescale traffic variations. This effect becomes more pronounced when traffic passes through a tandem of queues instead of a single queue.

B. Finite Bandwidth

Any traffic travelling in the network is subject to the restriction of a network link with a finite bandwidth. This effect is referred to as the bandwidth limiting effect. Fig. 1 shows the relationship between the input traffic rate and output traffic rate in a queue with a link capacity C . When the link utilization is low, the link works in a linear region close to the origin. Therefore traffic variations are almost fully preserved when passing through the link. The bandwidth limiting effect of a finite bandwidth link has little effect. However, when the link utilization is high, the link works in a region close to the link capacity C . The saturation point is easily reached, the limited remaining bandwidth will only allow a small variation of the traffic. Thus the bandwidth limiting effect may significantly change the traffic characteristics. Moreover, in normal network conditions, it is rare that the input traffic rate can exceed the link capacity for a long time period. Therefore it is expected that the bandwidth limiting effect will mainly affect the small timescale features of

traffic. The large timescale features will remain intact. Erramilli et al. observed that the performance impact of small timescale and large timescale traffic components are different at different link utilizations [5]. Small timescale features can affect performance substantially at low and intermediate utilizations, while the large timescale self-similarity is important at intermediate and high utilizations. Their observation actually witnessed the impact of the bandwidth limiting effect.

III. ANALYSIS METHOD

A lot of methods have been developed in the literature to analyze the scaling behavior of network traffic, to name but a few, variance-time plot, Higuchi’s Method, R/S method, Periodogram method, Whittle estimator. A comprehensive overview of these methods can be found in [14]. In this paper, we are going to perform our analysis using wavelet tools. Wavelets have many advantages when used in traffic analysis. Fundamentally, this is due to the non-trivial fact that the analyzing wavelet family itself possesses a scale invariant feature, a property not shared by other analysis methods. Quite different kinds of scaling features can be analyzed by the same technique and the same set of computations.

Wavelet analysis is based on the decomposition of a signal using orthogonal bases. Discrete wavelet transform (DWT) consists of the collection of coefficients

$$c_{jk} = \langle X, \varphi_{jk}(t) \rangle, d_{jk} = \langle X, \psi_{jk}(t) \rangle, k \in \mathbb{Z}, j \leq J \quad (4)$$

where $\langle *, * \rangle$ denotes inner product, $\{d_{jk}\}$ are the wavelet coefficients and $\{c_{jk}\}$ are the scaling coefficients. Equation (4) compares the signal X to be analyzed with a set of analysis functions

$$\psi_{jk}(t) = 2^{-j/2} \psi(2^{-j}t - k). \quad (5)$$

This set of analysis functions is constructed from a reference pattern $\psi(t)$ called the mother-wavelet by a time-shift operation and a dilation operation. $\psi(t)$ is a band pass function. Function $\varphi_{jk}(t)$ is a time shifted function of the scaling function $\varphi_J(t)$: $\varphi_{jk}(t) = \varphi_J(t - k)$. $\varphi_J(t)$ is a low-pass function which can separate the large timescale (low frequency) component of the signal. Thus wavelet transform decomposes a signal into a large timescale approximation (coarse approximation) and a collection of details at different smaller timescales (finer details). The original signal can be recovered from the wavelet coefficients and the scaling coefficients using

$$X(t) = \sum_k c_J(k) \varphi_{Jk}(t) + \sum_{j=0}^J \sum_k d_j(k) \psi_{jk}(t). \quad (6)$$

Theoretically the scale j can span from $-\infty$ to ∞ . For practical signals, i.e. network traffic, we limit the scale to $0 \sim J$, where scale J is the largest timescale and scale 0 is the smallest timescale.

The wavelet transform decomposes a signal into different frequency components and analyzes each component with a resolution matched to its scale. We can use coefficients of a wavelet decomposition to directly study the scale (or frequency) dependent properties of the signal. The coefficient $|d_{jk}|^2$ measures the amount of energy in a signal X about the time $t_0 = 2^j k$ and about the frequency $2^{-j} f_0$ or timescale $2^j / f_0$, where f_0 is

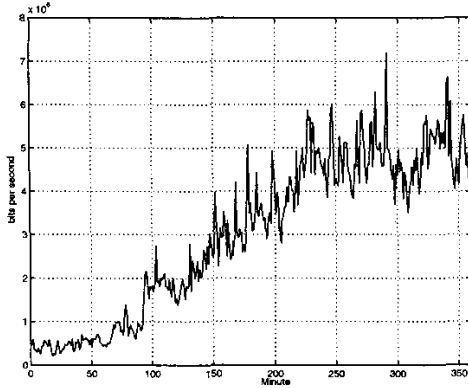


Fig. 2. Traffic rate measured on 1 minute interval

a reference frequency depending on the mother wavelet. Let E_j denotes the average of $|d_j(k)|^2$ at each scale:

$$E_j = \frac{1}{N_j} \sum_k |d_j(k)|^2, \quad (7)$$

where N_j is the number of wavelet coefficients at scale j , then E_j is a measure of the energy that lies within a given bandwidth 2^{-j} around frequency $2^{-j}f_0$. Using the logarithm of E_j , parameters of LRD c_f and α in (1) can be estimated:

$$\log_2(\mathbf{E}(E_j)) = j\alpha + \log_2(c_f C), \quad (8)$$

where C is a constant determined by the mother wavelet.

IV. NETWORK TRAFFIC

The network traffic used in our analysis was collected by WAND research group at the University of Waikato Computer Science Department. It is the LAN traffic at the University of Auckland on campus level. The traffic trace was collected on June 11, 2000 between 6am and 12pm on a 100Mbps Ethernet link. IP headers in the traffic trace are GPS synchronized and have an accuracy of $1\mu s$. More information on the traffic trace and the measurement infrastructure can be found on their webpage: <http://atm.cs.waikato.ac.nz/wand/wits/auck/6/>. Fig.2 shows the traffic rate of the traffic trace measured on 1 minute intervals. As shown in the figure, the traffic presents significant non-stationarity during the measurement period. In order to minimize the effect of non-stationarity, only a piece of the whole traffic trace during a two-hour time interval, i.e. 240 minute - 360 minute, is used for analysis in the rest of the paper. The average traffic rate during the two-hour time interval is $4.847 Mbps$.

Daubechies 5 wavelets is used and the choice of number of vanishing moment of 5 for Daubechies wavelet is used to further diminish the effect of polynomial trend in traffic due to non-stationarity. The analysis program is modified from the program of Veitch et al. [7]. The data being analyzed is the incoming traffic rate measured in the number of bytes per $10ms$. $10ms$ interval is chosen to avoid the situation that there may be no incoming packets during a time interval. Fig.3 shows the logscale diagram of $\mathbf{E}(E_j)$ together with the 95% confidence interval of the estimation. The bottom axis of the figure shows the scale j

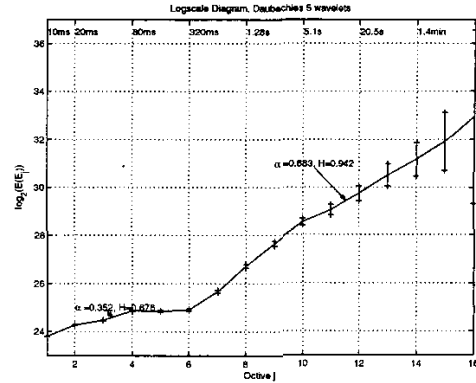


Fig. 3. Wavelet analysis of the traffic

and the top axis of the figure shows the corresponding timescale. As shown in the figure, the LAN traffic presents complex biscaling behavior. The traffic trace is consistent with asymptotic self-similarity or LRD at large timescale, but the large timescale and small timescale have distinct scaling behavior. The traffic fluctuation at small timescale is less correlated, which is indicated by a much less scaling exponent α (or equivalently H). A value of α equal zero indicates the traffic fluctuation at small timescale is independent. The transition from small timescale behavior to large timescale behavior occurs around timescale $80ms - 320ms$, which is the typical round-trip time of the network. Since the small timescale features of traffic play an important role in network performance [11], [15], the distinct scaling behavior of traffic in small timescale and large timescale indicates that an exact self-similar model which is characterized by a single scaling exponent, e.g. Fractional Gaussian Noise model, is inappropriate for modelling LAN traffic.

In the following sections, the small timescale and large timescale features of traffic are analyzed separately.

V. THE IMPACT OF BUFFER ON THE SCALING BEHAVIOR

In this section, we investigate the impact of buffer on the scaling behavior of traffic using simulations. The simulations are implemented using OPNET. We let the traffic analyzed in section IV pass through a queue with a link capacity of 1.5, where the link capacity is normalized by the average traffic rate, and vary the buffer size of the queue from $1ms$ to $100ms$ to observe the impact of buffer on traffic. Here the size of the buffer is expressed in time unit, which is obtained by dividing the buffer size by the link capacity. Fig.4 shows the variation of the small timescale scaling exponent with the buffer size together with the 95% confidence interval for the scaling exponent estimation. Please refer to [7] for the detailed procedure on calculating the confidence interval. Fig.5 shows the variation of the large timescale scaling exponent. Fig.6 shows the variation of $\mathbf{E}(E_j)$ with the buffer size for some typical buffer sizes. Table I shows the packet loss ratio (PLR).

Simulation results reveal that buffer size has different impact on the scaling behavior of traffic at large timescale and small timescale. At large timescale, buffer size has essentially no impact on the scaling exponent of traffic. However buffer

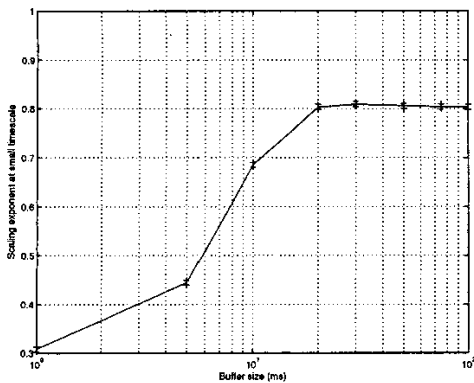


Fig. 4. Variation of scaling exponent α at small timescale with buffer size

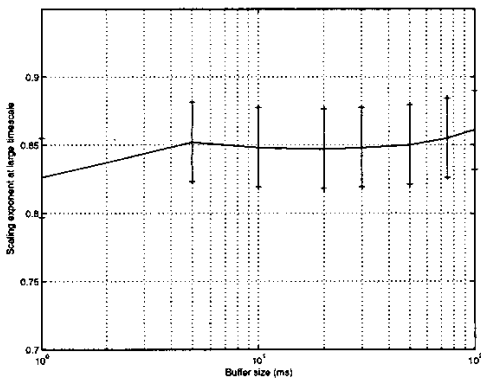


Fig. 5. Variation of scaling exponent α at large timescale with buffer size

size can dramatically change the scaling exponent of traffic at small timescale. When the buffer size is increased from $1ms$ to $100ms$, the scaling exponent at small timescale increases from 0.308 to 0.804. Fig.6 shows that generally with the decrease of buffer size, $E(E_j)$ at both large timescale and small timescales decreases. This indicates that energy of the traffic decreases both at small timescale and at large timescale with decreasing buffer size. Therefore traffic will be less bursty with decreasing buffer size. This effect is most likely due to traffic loss. Fig.6 also reveals the intriguing behavior that when buffer size varies between $10ms$ and $100ms$, while the large timescale energy increases with increasing buffer size, the small timescale energy decreases slightly with increasing buffer size. Table I indicates that the traffic has moderate loss when buffer size falls in this region. This fact seems to suggest that in this region, with increasing buffer size, the buffer can smooth the transient component of traffic and thus transfer a portion of the small timescale energy to large timescale. In real network, traffic loss is mostly caused by bursty traffic, which is the high frequency (small timescale) component of traffic. Therefore this delicate exchange of energy between small timescale and large timescale, although small in

TABLE I
VARIATION OF PACKET LOSS RATIO WITH BUFFER SIZE

buffer (ms)	1	5	10	20	30	50	75	100
PLR	.38	.081	.035	.016	.011	.007	.005	.005

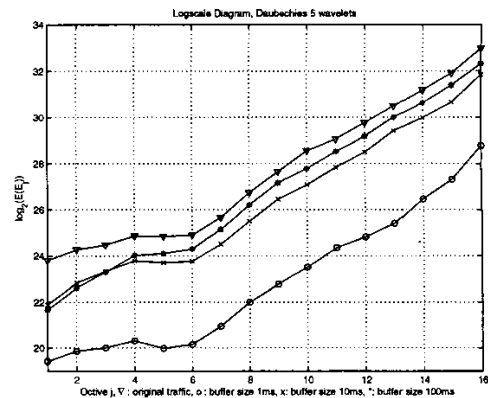


Fig. 6. Variation of $E(E_j)$ with buffer size

magnitude, may have significant impact of traffic loss. Further investigation is being performed to fully understand this effect.

VI. BANDWIDTH LIMITING EFFECT

In this section, we shall investigate the impact of the bandwidth limiting effect on the scaling behavior of traffic. We let the traffic pass through a queue with varying link capacity and fix the buffer size at $90.89kB$, which gives a normalized buffer size of $100ms$ at a link capacity of 1.5. The link capacity is normalized with respect to the average traffic rate. Fig.7 shows the variation of the small timescale scaling exponent with the link capacity. Fig.8 shows the variation of the large timescale scaling exponent. Fig.9 shows the variation of $E(E_j)$. The PLRs at link capacities 1.5 and 2 are 0.0047 and 0.0017 respectively. When the link capacity is increased beyond 3, the PLR becomes essentially zero.

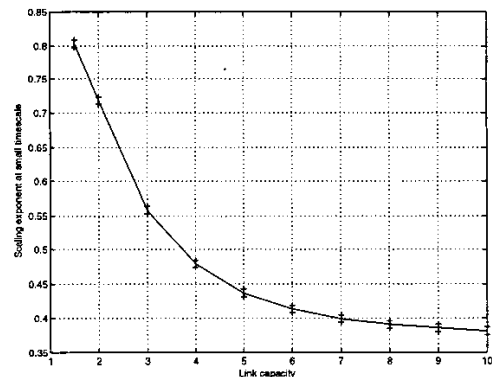


Fig. 7. Variation of scaling exponent α at small timescale with the link capacity

It is shown in the figures that the bandwidth limiting effect has different impact on the scaling behavior of traffic at large timescale and small timescale. At large timescale, the bandwidth limiting effect has almost no impact on the scaling exponent. At small timescale the scaling exponent increases dramatically with decreasing link capacity. As shown in Fig. 3, the small timescale scaling exponent of the traffic before passing the link is 0.352. Fig. 7 shows that when the traffic passes through a link with a large link capacity, the small timescale

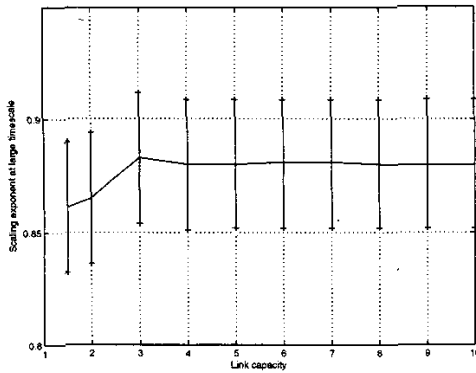


Fig. 8. Variation of scaling exponent α at large timescale with the link capacity

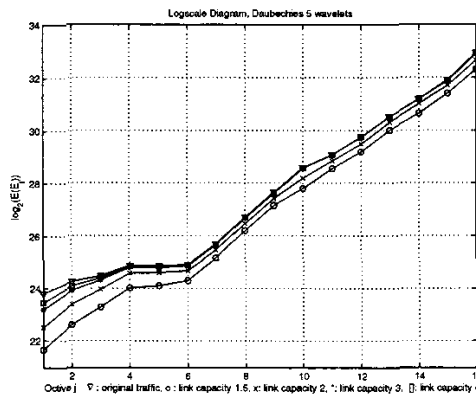


Fig. 9. Variation of $E(E_j)$ with the link capacity

scaling exponent remains almost constant. But when the traffic passes through a link with a much smaller link capacity, the small timescale scaling exponent increases dramatically.

Moreover, when the link capacity is increased beyond 3, there is no traffic loss. However both the small timescale scaling exponent and energy, as indicated by $E(E_j)$ in Fig.9, still change with the link capacity. Specifically, the small timescale scaling exponent increases with decreasing link capacity and the small timescale energy decreases with decreasing link capacity (increasing link utilization). In contrast the large timescale scaling exponent and energy is essentially constant in this region. This reveals that it is not necessary for traffic loss to occur in order for the network to modify the scaling behavior of traffic and smooth the traffic.

Finally, Fig.9 shows that the energy of the signal at both small timescale and large timescale decreases with increasing link utilization. The decrease at small timescale is more significant than that at large timescale. This shows that at heavy link utilization, the traffic will become smoother both at small timescale and at large timescale but small timescale variations of traffic is more heavily affected by the link utilization.

VII. CONCLUSION

In this paper, we analyzed the impact of network mechanisms, i.e. bandwidth and buffer, in shaping the traffic. Analysis results showed that both bandwidth and buffer can significantly

change the small timescale scaling exponent of traffic. However their effect on the large timescale scaling exponent is almost negligible. Moreover the use of smaller buffer and smaller link capacity can effectively reduce the energy of traffic at small timescale, thus smooth the traffic. Simulation results indicated that this effect can be achieved even when there is no traffic loss. A plausible application of this result is that the use of a small but sufficient link capacity such that no traffic loss occurs at the network edge can achieve better performance than using a very large link capacity. The reason is that traffic becomes less bursty with smaller link capacity. Therefore traffic entering into the core network will be smoother and less traffic loss can be expected at the core network, which is usually the bottleneck of the traffic congestion. Moreover, our preliminary analysis also revealed the intriguing behavior of buffer that it can transfer a portion of small timescale energy to large timescale, therefore the traffic becomes less bursty at small timescale. This is a direct evidence that buffer can effectively smooth small timescale (high frequency) traffic variations and reduce the burstiness of traffic. Further research is being performed to investigate this effect and its performance implication in real network.

REFERENCES

- [1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.
- [2] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1566–1579, 1995.
- [3] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.
- [4] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of internet wan traffic," 1998, Proceedings of SIGCOMM'98, pp. 42–55.
- [5] A. Erramilli, O. Narayan, A. Neidhardt, and I. Sanicic, "Performance impacts of multi-scaling in wide area tcp/ip traffic," in *INFOCOM 2000*, Tel Aviv, Israel, 2000, vol. 1, pp. 352–359.
- [6] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Process: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 1994.
- [7] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *Information Theory, IEEE Transactions on*, vol. 45, no. 0018-9448, pp. 878–897, 1999.
- [8] G. Mao, "Finite timescale range of interest for self-similar traffic measurements, modelling and performance analysis," in *IEEE International Conference on Networks*, 2003, pp. 7–12.
- [9] M. Grossglauser and J. C. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Transactions on Networking*, vol. 7, no. 5, pp. 629–640, 1999.
- [10] M. Grossglauser and D. N. C. Tse, "A time-scale decomposition approach to measurement-based admission control," in *IEEE INFOCOM 1999*, New York, NY, USA, 1999, pp. 1539–1547.
- [11] A. Erramilli, M. Roughan, D. Veitch, and W. Willinger, "Self-similar traffic and network dynamics," *Proceedings of the IEEE*, vol. 90, no. 5, pp. 800–819, 2002.
- [12] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "The effect of statistical multiplexing on the long-range dependence of internet packet traffic," Bell labs technical report, Bell Labs, 2002.
- [13] D. P. Heyman and T. V. Lakshman, "Long-range dependence and queueing effects for vbr video," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., pp. 285–318. John Wiley and Sons, Inc., 2000.
- [14] M. Taqqu, V. Teverovsky, and W. Willinger, "Estimators for long-range dependence: An empirical study," *Fractals*, vol. 3, no. 4, pp. 785–798, 1995.
- [15] R. H. Riedi and W. Willinger, "Toward an improved understanding of network traffic dynamics," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., pp. 507–530. John Wiley & Sons, Inc., 2000.