Finite Timescale Range of Interest for Self-Similar Traffic Measurements, Modelling and Performance Analysis

Guoqiang Mao
School of Electrical and Information Engineering
The University of Sydney

Abstract—**Although the existence of self-similarity in network traffic has been widely recognized, there is considerable debate on the impact of self-similarity on traffic engineering and network performance, and whether or not self-similar model should be used for traffic modelling. This paper reviews some major research in the area and summarizes limitations in current theoretical performance analysis with self-similar input. It is pointed out that these performance analyses are insufficient to make a definitive conclusion on the long-range dependence effects. Furthermore, it is pointed out that in a real system, the timescale range of interest for traffic measurements, modelling and performance analysis is limited, which can be characterized by an engineering timescale range (ETR). Only traffic correlations within the ETR will affect performance and are important for traffic measurement, modelling and performance analysis. Factors contributing to the ETR are identified. Further research is proposed on quantitatively identifying the ETR, and traffic measurements and self-similar traffic modelling using Markov models.**

I. Introduction

Bandwidth hungry computer and communications applications are on the rise with a variety of services, as a few examples, video conference, video on demand, voice over IP and high-definition television. Different from traditional data communication applications, most of these applications are real-time applications and they have stringent requirements on quality of service (QoS), i.e. traffic delay, jitter and loss. Traditional data communication technique such as retransmission does not apply to real-time traffic. Because even a packet is delivered to the destination eventually, if it does not arrive in time, it will be considered as lost. One of the major challenges in designing modern communication networks is providing QoS support to the individual applications.

It is well known that some of the characteristics of network traffic fall beyond the conventional framework of Markov traffic modelling. Leland et al. demonstrated self-similarity in a LAN environment (Ethernet) [1]. Paxson et al. showed self-similar burstiness manifesting itself in pre-World Wide Web WAN IP traffic [2]. Beran et al. [3] and Garrett et al. [4] demonstrated self-similarity in variable-bit-rate (VBR) video traffic, and Crovella et al. showed self-similarity for WWW traffic [5]. Collectively, these measurement works constitute strong evidence that scale-invariant burstiness is not an isolated, spurious phenomenon but rather a persistent trait existing across a range of network environments [6].

For a stationary sequence $X = \{X(i),\ i \geq 1\}$, let

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i), \quad k = 1, 2, ..., \qquad (1)$$

be the corresponding aggregated sequence with level of aggregation m, obtained by dividing the original series X into non-overlapping blocks of size m and averaging over each block. The index k, labels the block. A stationary sequence $X = \{X(i),\ i \geq 1\}$ is called exactly self-similar if it satisfies (2) for all aggregation levels m.

$$X \triangleq m^{1-H} X^{(m)}, \quad 0 < H < 1. \qquad (2)$$

It is said to be asymptotically self-similar if (2) only holds as $m \to \infty$. Parameter H is called the Hurst parameter [7] and is a measure of self-similarity. For self-similar processes its value is between 0.5 and 1 and the degree of self-similarity increases as the Hurst parameter approaches 1. Self-similarity in network traffic usually refers to asymptotic self-similarity. There are a number of other different, not equivalent, definitions of self-similarity. Refer to [8], [9], [10], [11], [12] for details.

Long-range dependence (LRD) is another widely used term in this area. Let the mean and covariance function of a stationary sequence $X(t)$ be denoted by $\mu = E[X(t)]$ and $C_X(k) = E[(X(t+k)-\mu)(X(t)-\mu)]$. A long-range dependent sequence can be defined via a slow, power-law decay of $C_X(k)$:

$$C_X(k) \sim C_\gamma k^{-\beta}, \quad 0 < \beta < 1 \qquad (3)$$

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 $H = 1 - \beta/2$.

Strictly speaking, self-similarity and long-range dependence are different concepts. However in the context of network traffic modelling and performance analysis, they are used to refer to the same phenomenon that the cumulative effects of long-term correlations of a traffic process cannot be ignored. Therefore they are often used interchangeably.

Although long-range dependence in network traffic has been widely recognized, QoS impact of long-range dependence is still an open issue [3], [5], [13], [4], [1], [14], [15], [16], [17], [18]. In this paper, some of the major research in the area are reviewed and it is pointed out that most performance analyses with self-similar input are asymptotic in nature and they fail to consider the impact of finite-length queue and network dynamics (i.e. call level dynamics, network control protocols, etc.) in shaping the traffic. As a result, although they provide useful insight into the performance impact of some aspects of self-similarity, they cannot be used to make a definitive conclusion on it. It is proposed that the interactions between traffic process and queue

0-7803-7788-5/03/$17.00

level dynamics, as well as network dynamics are incorporated into a united framework of performance analysis through the concept of engineering timescale range (ETR). Only traffic correlations within the ETR will affect performance and are important for traffic measurement, modelling and performance analysis. A traffic model which fits the network traffic characteristics within the ETR is sufficient for performance analysis. Therefore Markov traffic model can be used for modelling self-similar traffic.

The rest of the paper is organized as follows. In section II performance analyses on the LRD effects are reviewed. Limitations of most performance analysis with self-similar input are summarized in section III. The existence of engineering timescale range is justified in section IV. Further research on traffic measurements, modelling and engineering timescale range is proposed in section V.

II. Relevance of Self-Similarity - *The Pros and Cons*

Some researchers performed performance analysis on queueing systems with self-similar input. They demonstrated that queue length distribution decays polynomially instead of exponentially, which is typical of Markov traffic. They considered that the widespread existence of self-similarity challenges the basis of the traditional Markov-based traffic modelling, performance analysis and traffic engineering. Thus a major overhaul on the current Markov-based traffic engineering techniques is necessary.

Norros [19] showed that the marginal distribution $Q(x)$ of a stationary fluid queue in the fractional Brownian noise traffic model (FBM), which is self-similar, is asymptotically of Weibull type [20], [21], [22], that is:

$$logQ(x) \sim -\kappa x^{2(1-H)} \tag{4}$$

where κ is a positive constant. Krishnan [23] showed that the implication of Norro's result is that when a sufficiently large number of sources are multiplexed, high-H sources require more bandwidth than low-H sources. Erramilli et al. [24] demonstrated empirically that LRD has considerable impact on queueing performance and traffic engineering problems. Likhanov et al. [14] showed that the overflow probability of the multiplexing of a large number of on-off sources with Pareto distributed (heavy tailed) on periods, which is self-similar [25], [26], [27], has an asymptotic relationship with buffer size:

$$Q(x) \sim \alpha x^{-\gamma} \tag{5}$$

where α and γ are positive constants. Parulekar et al. [20] obtained the same asymptotic relationship using large deviations theory and the $M/G/\infty$ model [28] for self-similar process. There are many other research which demonstrated that the marginal distribution of a fluid queue under self-similar traffic decays slower than that under Markov traffic. A summary of these research can be found in [11].

However, there is considerable debate on the impact of LRD on traffic engineering and network performance, and whether or not self-similar model should be used for traffic modelling [29], [30], [17], [18], [15], [31], [32], [33]. They recognized the existence of self-similarity in network traffic, however, they took a more practical view at the problems. They pointed out that long-range dependence is not a crucial property in determining the behavior of real buffers with *finite buffer size*. Since the objective of traffic modelling is to enable performance analysis, and Markov traffic model is accurate enough to predict the performance of real buffers fed with real traffic sources, Markov traffic model should be used instead of self-similar traffic model.

Elwalid et al. found that buffer overflow probability decreases exponentially with buffer size [29], i.e.

$$Q(x) \approx \alpha e^{-\delta x}, \tag{6}$$

where α and δ are positive constants. They validated their result using simulations, where real video conference sequence coded by different algorithms are used as traffic sources. Their results show that (6) accurately captures the relationship between buffer overflow probability and buffer size, and $DAR(1)$ (discrete autoregressive process with order 1) traffic source model, which takes into account only short-range dependence, is accurate enough for admission control and bandwidth allocation of video conferences. In particular, the video conference sequences used for their simulations exhibit long-range dependence [3]. Therefore, their results effectively counter the assertion in [1] that when traffic is long-range dependent "overall packet loss decreases very slowly with increasing buffer capacity".

Heyman et al. [30] employed a generic buffer model to investigate the effects of long-range dependence. The buffer has capacity B, and receives input at deterministic times. Let X_i be the number of arrivals at discrete time T_i. Let d be the number of traffic that is processed during $[T_i,\ T_{i+1})$, referred to as the i^{th} interval, and let V_i be the buffer content at the end of the i^{th} interval. Then,

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

They showed that the *resetting effect* of the buffer when buffer becomes empty and the *truncating effect of finite buffers* when buffer become full diminish the long-range dependence effects. Since VBR video traffic has stringent QoS requirements, the traffic intensity for these services will not be large. Thus the resetting effect and truncating effect should be strong in practical regions. Their numerical examples confirmed that Markov model can accurately predict the QoS of real VBR video conference and entertainment video [31], which are long-range dependent. Based on it, they concluded that long-range dependence is not a crucial property in determining the buffer behavior of VBR video sources.

Ryu et al. investigated the practical implications of long-range dependence by studying the behavior of buffers with VBR video input over a range of desirable cell loss rates and buffer sizes [15]. Based on large deviations theory, they introduced the notion of *Critical Time Scale* (CTS). For a given buffer size, link capacity, and the marginal distribution of frame size, the CTS of a VBR video source is defined as the number of frames whose correlations contribute to the cell loss rate. They show that whether the model is Markov or long-range dependent, its CTS is finite. CTS assumes a small value for a small buffer, and is a non-decreasing function of the buffer size. In other words, under realistic scenarios of buffer dimensioning, the number of

frame correlations which affect buffer overflow probability is finite and small even in the presence of the long-range dependence property. Simulations were used to validate their result. They used the superposition of FBNDP (Fractional-Binomial-Noise-Driven Poisson Process) and $DAR(1)$ (discrete autoregressive process of order 1) as their long-range dependent VBR video traffic model, in which the long-term correlations and the short-term correlations can be effectively controlled. Their simulations showed that the buffer overflow probability of the long-range dependent traffic can be accurately captured by $DAR(p)$ process (discrete autoregressive process of order p), which is short-range dependent, *under the practical ranges of buffer size and cell loss ratio*. Their numerical results showed that:

- even in the presence of long-range dependence, long-term correlations do not have significant impact on cell loss rate; and
- short-term correlations have dominant effect on cell loss rate,

and therefore, well-designed Markov traffic models are effective for predicting QoS of long-range dependent VBR video traffic. They concluded that it is unnecessary to capture the long-term correlations of a real-time VBR video source under realistic buffer dimensioning scenarios as far as the cell loss rate and maximum buffer delays are concerned.

Grossglauser et al. found the existence of the CTS independently [17]. They argued that most of recent modelling works have failed to consider the impact of two important parameters, namely the finite range of time scales of interest in performance evaluation and prediction problems, and the first-order statistics such as the marginal distribution of the process. They introduced a modulated fluid traffic model in which the correlation functions of the fluid rate matches that of an asymptotically second-order self-similar process with given Hurst parameter up to an arbitrary cutoff time lag, then drops to zero. Numerical experiments are performed to evaluate the performance of a single server queue fed with the above fluid input process. They found that the amount of correlations that needs to be taken into account for performance evaluation depends not only on the correlation structure of the source traffic, but also on time scales specific to the system under study. For example, the time scale associated with a queueing system is a function of the maximum buffer size. For finite buffer queues, they found that the impact of correlations in the arrival process on traffic loss becomes nil beyond a time scale referred to as the correlation horizon. This means, in particular, for performance-modelling purposes, any model among the panoply of available models can be chosen as long as the chosen model captures the correlation structure of the traffic source up to the correlation horizon.

In [18], Grossglauser et al. studied a robust measurement-based admission control with emphasis on the impact of estimation errors, measurement memory, call-level dynamics and separation of time scales. Their work [17], [18] identifies a *critical time-scale* $\widetilde{T_h}$ such that aggregate traffic fluctuation slower than $\widetilde{T_h}$ can be tracked by the admission controller and compensated for by connection admissions and departures. Fluctuations faster than $\widetilde{T_h}$ have to be absorbed by reserving spare bandwidth on the link. Using Gaussian aggregate traffic model and heavy traffic approximations, the critical time scale is shown to scale as $T_h/\sqrt{n}$, where T_h is the average flow duration and n is the size of the link in terms of the number of flows it can carry. The

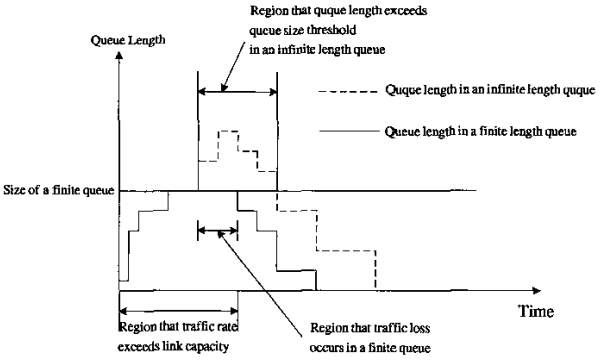

Fig. 1. Comparison between traffic loss in a finite length queue and an infinite length queue

major insight that can be gained from their work is that call level dynamics, i.e. connection admissions and departures can diminish the impact of long-range dependence on the performance of a MBAC.

III. Limitations of Current Research

The aforementioned research on the LRD effects seems to conflict with each other. This is due to the complexity of the problem. There are too many factors to be considered, which include, traffic characteristics, statistical multiplexing, call level dynamics, resetting and truncating effects of finite size buffers, network buffer size, network utilization level, etc.

At this stage, no definitive conclusions can be made on the LRD effects. However it is noticed that there are some limitations in most theoretical performance analyses with self-similar input.

They are asymptotic in nature where either buffer capacity is assumed infinite and marginal distribution $Q(x)$ of queue length is estimated as $x \to \infty$; or buffer capacity b is assumed finite but buffer overflow probability is computed as b becomes unbounded [34]. However it is not clear that how large a buffer should be in order that the marginal distribution conforms to the asymptotic relationship. If that buffer size is too large that it exceeds practical range of real buffers, this asymptotic relationship is not meaningful in real applications. Moreover, many asymptotic results are given in the form of the probability that the queue size exceeds a certain threshold in an infinite length queue. In a real system, the quantity of interest is the loss probability in a finite length queue. This is also the quantity we can measure in reality. These two values may not necessarily agree with each other. Fig. 1 shows that the traffic loss probability in a real buffer with size x is actually given by:

$$\text{loss probability} = Pr\{\text{traffic rate} > \text{link capacity}\} \times Q(x) \tag{7}$$

where $Q(x)$ denotes the probability that queue length exceeds the threshold value x in an infinite length queue. The first term in the equation $Pr\{\text{traffic rate} > \text{link capacity}\}$ is a small value in low to medium traffic intensity. Some connection admission control schemes based on bufferless fluid flow model achieve a link utilization of 60% to 80% while still controls the loss ratio below 10^{-4} [35]. Therefore it can be expected that using $Q(x)$

to approximate loss probability will produce a large error in non-heavy-traffic scenarios.

They tend to emphasize the impact of one aspect of statistical characteristics of traffic process while neglecting the other. Current research on traffic self-similarity focuses on Hurst parameter H because it defines the existence of self-similarity. Hurst parameter is the major concern in traffic analysis, measurements and modelling. As a result, most performance analyses are based on traffic models which match Hurst parameter only. The impact of the second parameter C_γ in (3) is often neglected. This is unfortunate because C_γ defines the magnitude of correlation variation and it plays a major role in fixing the absolute size of the LRD-generated effects [36], [37]. Concentrating on the impact of Hurst parameter only may lead to the over exaggeration of the LRD effects because if the value of C_γ in a traffic process is small, the LRD effects will be small and are possibly negligible in performance analysis. Moreover, the analysis of measured network traffic and resulting understanding of some of its underlying structure have led to the realization that while network traffic is consistent with asymptotic self-similarity behavior, its small time scaling features are very different from those observed over large timescales. Those small time scaling features play an important role in performance analysis [30], [38], [39], [40], [41]. Performance analysis considering the scaling features of traffic in both small timescale and large timescale is required to gain a complete understanding on the LRD effects.

The amount of correlation that needs to be taken into account for performance evaluation depends not only on the correlation structure of the traffic process, but also on the range of timescales specific to the system under study [17]. In a real system, the range of timescale of interest for traffic measurements, modelling and performance analysis is limited. Some performance analyses emphasize one aspect of the LRD effects, that is, the cumulative effects of long-term correlations can not be ignored in the performance analysis. Their results implicitly consider an infinite timescale for performance analysis. This may produce a large error in a real system and may lead qualitatively different conclusions

Most performance analyses with self-similar input are based on the "open loop" data traffic models [38]. Here "open loop" refers to the fact that while the traffic characteristics impact queueing behavior, the impact of finite-length queue and network dynamics (i.e. call level dynamics, network control protocols, etc.) in shaping the incident traffic is not modelled. It has been pointed out that a finite-length queue will interact with the traffic process to diminish the LRD effects [30], [31]. It is also observed that network traffic control protocols (e.g. TCP, UDP) can modify the self-similar scaling behavior of network traffic [38] and call level dynamics (i.e. connection arrivals and departures) will affect the LRD effects [18]. In [36], it was found that long-term correlations in the traffic affect performance at higher utilizations. Short-term correlations are important in complementary regimes and both are important at intermediate utilizations. These research together reveals a fact that the impact of finite length queue and network dynamics cannot be ignored in investigating the LRD effects.

In summary, although these theoretical analyses with self-similar input provide useful insight into the performance impact of some aspects of self-similar traffic, they cannot be used to make a definitive conclusion on it.

IV. The Existence of Engineering Timescale Range

In this section we try to heuristically establish that in a real system, the timescale range of interest for traffic measurements, modelling and performance analysis is limited. This finite timescale range of interest for practical traffic engineering is referred to as the engineering timescale range (ETR). Traffic correlations beyond the ETR will have no impact on performance. Several factors that may affect the engineering timescale range are identified.

While theoretically self-similarity may extend to an infinitely large timescale, the timescale range of interest is limited in a real system. For example, traffic congestion and performance degradation at 3pm is unlikely caused by some traffic sources at 3am although statistically strong large-timescale correlation may be measured between them [42]. It has been found that self-similarity in the aggregate traffic is caused by the high variability in the active/silence period distributions of individual connections [5], [26], [16]. This is not unexpected because any statistical characteristics of the aggregate traffic should be able to be traced back to the statistical characteristics of individual connections and network control mechanisms (e.g. traffic control protocols) shaping the traffic. The impact of the aggregate traffic at a past instant t on the performance of current system will eventually be realized through the individual connections that consist in the aggregate traffic at time t. However the impact of an individual connection can not go beyond its lifetime. Moreover the lifetime of a connection must be limited in a real system, that is, any connection cannot exist in the system indefinitely. Therefore it can be asserted that the state of the network system at time t will have no impact on the performance of the current system when all connections active at time t have left network. Furthermore, the impact of the system state at time t on the performance of the current system will be negligible when most active connections at t has left the network. Therefore a timescale, referred to as critical timescale, should exist such that traffic correlations beyond the critical timescale have no impact on performance. This timescale should be upper-bounded by a value which can be expressed as a function of the distribution of connection durations.

There are other factors which are important in determining the value of the critical timescale. It was pointed out that the resetting effect when a buffer becomes zero will diminish the LRD effects. The effects of LRD are significant only if LRD causes the busy periods to be long enough for the long lags to come into play. Similarly, the truncating effect of a finite buffer will also diminish the LRD effects [30], [31]. Actually in [15] and [18], critical timescale is considered to be a function of maximum buffer size. However maximum buffer size alone cannot determine critical timescale because both the truncating effect and the resetting effect are a function of utilization and they are stronger at low to medium utilization [31]. This conforms to the findings in [36], where it was found that long-term correlations in the traffic affect performance only at higher utilizations. Therefore network utilization is also an important factor

in determining critical time scale. This fact may affect traffic engineering techniques for different applications. Real-time applications such as video on demand, video conference usually have stringent requirements on QoS. Thus they usually run at low to medium utilization levels where LRD effects are less significant. Some non-realtime applications multiplexed with them have lower priority and will not affect the QoS of those realtime applications. Non-realtime applications such as data traffic do not have stringent QoS requirements and they may be run at high utilization to fill the capacity gap left by realtime applications. LRD effects are more significant for these non-realtime applications.

The effects of network traffic and congestion control protocols also need to be considered in determining critical timescale. For non-realtime traffic, it was found that TCP can modify the self-similar scaling behavior of network traffic [38]. For realtime traffic, which is not subject to TCP control because their stringent delay requirement, it was found that a measurement-based admission control scheme can investigate call level dynamics to diminish the LRD effects [18]. Moreover LRD will improve the predictability of network traffic due to larger correlations at large timescales. It is therefore expected that traffic prediction and performance prediction algorithms can be embedded in traffic control protocols to improve performance under self-similar traffic. The impact of these traffic and congestion control protocols on the LRD effects can be represented by their effects on determine the timescale of interest for performance analysis, i.e. critical timescale.

The critical timescale forms the upper boundary of the engineering timescale, which defines the range of long-term correlations that needs to be considered in performance analysis. The ETR should also have a lower boundary which defines the range of short-term correlations that needs to be considered. This is due to the fact that small timescale correlations, which present as rapid traffic rate fluctuations, below a certain limit can be effectively absorbed by a small buffer, thus are of no importance in performance analysis [43], [44].

The resulting ETR will possibly span both small timescale and large timescale. Only traffic correlations within the ETR will affect performance and are of importance for traffic measurements, modelling and resource provisioning.

V. Further Research

A. Determining the Critical Timescale

It is extremely difficult to consider the aforementioned contributing factors to critical timescale altogether to obtain the value of CTS. Therefore we suggest modelling the diminishing effects of these queue level dynamics and network dynamics on the LRD effects by identifying a critical timescale for each kind of diminishing effect. This includes:

- a CTS for finite queue size. The impact of network utilization on CTS can be considered when evaluating the resetting and truncating effects.
- a CTS for call level dynamics.
- a CTS for network control protocols.

Realtime applications and non-realtime applications should be treated differently when determining CTS because of different utilizations they are usually working at and different control protocols. The cumulative effects can be represented by a CTS value not larger than these individual CTS values.

B. Traffic Measurements

Traffic measurement is the first step toward accurate traffic modelling and performance analysis. Over the years, lots of measurement works have been done, to name but a few, [1], [4], [39], [40], [45]. They contribute significantly to our understanding of the network traffic characteristics. However, due to the progressive nature of our understanding about the impact of various statistical parameters, most of the earlier works concentrate only on certain limited aspects of the traffic process, e.g. marginal distribution, Hurst parameter, etc. In order to provide an adequate and complete description of actual network traffic, other parameters as the magnitude of correlation variation, short-term scaling parameters, also need to be measured.

There are a variety of tools available to measuring the scaling features of traffic sources, for example, variance-time plot and its variants, Higuchi's method [46], R/S method, Periodogram method and its variants, Whittle estimator and wavelet-based estimator [47], [42]. A comparison is made in [48]. Wavelet-based estimator appears to be a very promising technique because of its robustness, negligible bias, low variance and a key advantage that quite different kinds of scaling can be analyzed by the same technique.

C. Traffic Modelling and Performance Analysis

Some widely used self-similar traffic models such as FBN, heavy-tailed on-off sources are either insufficient to capture statistical behavior of real traffic or intractable to performance analysis. Since only traffic correlations within the ETR will affect performance and are important for traffic measurement, modelling and performance analysis, a traffic model which fits the network traffic characteristics within the ETR is sufficient for performance analysis. Therefore, Markov traffic models are considered to be strong candidates for traffic modelling. Markov traffic models have been successfully applied in the past to model the first-order statistical characteristics of network traffic. The benefit of using Markov models is obvious - a whole array of tools for estimating performance measures is already available. The major obstacle in the application of Markov models is their incapability in modelling LRD. The definition of ETR will remove this obstacle. Traffic model is only required to model the scaling behavior over a finite timescale range. Andersen et al. [32] demonstrated the capability of superposition of two-state MMPP (Markov-Modulated Poisson Process) in modelling LRD over several timescales. Further research in this area will be finding a suitable Markov model which is able to model both the first-order statical characteristic of network traffic and the scaling behavior of the traffic process across the whole range of ETR.

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] V. Paxson and S. Floyd, "Wide area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[3] 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.

[4] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar vbr video traffic," in *ACM SIGCOMM 1994*, 1994, pp. 269–280.

[5] 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.

[6] K. Park and W. Willinger, "Self-similar network traffic: An overview," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., pp. 1–38. John Willey & Sons, Inc., 2000.

[7] V. Klemes, "The hurst phenonmenon: A puzzle?," *Water Resources Research*, vol. 10, no. 4, pp. 675–688, 1974.

[8] B. Tsybakov and N.D. Georganas, "On self-similar traffic in atm queues: Definitions, overflow probability bound, and cell delay distribution," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 397–409, 1997.

[9] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Process: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 1994.

[10] J. Beran, *Statistics for Long-Memory Processes*, London: Chapman and Hall, 1994.

[11] K. Park and W. Willinger, Eds., *Self-Similar Network Traffic and Performance Evaluation*, John Wiley and Sons, Inc., 2000.

[12] W. Willinger, V. Paxon, R. Reidi, and M. Taqqu, "Long-range dependence and data network traffic," in *Long-range Dependence: Theory and Applications*, P. Doukhan, G. Oppenheim, and M. S. Taqqu, Eds., pp. 243–254. 2001.

[13] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, and F. Toomey, "Predicting quality of service for traffic with long-range fluctuations," in *IEEE International Conference on Communications 1995*, Seattle, WA , USA, 1995, vol. 1, pp. 473–477.

[14] N. Likhanov, B. Tsybakov, and N.D. Georganas, "Analysis of an atm buffer with self-similar ("fractal") input traffic," in *IEEE INFOCOM 1995*, Boston, MA , USA, 1995, vol. 3, pp. 985–992.

[15] B. K. Ryu and A. Elwalid, "The importance of long-range dependence of vbr video traffic in atm traffic engineering: Myths and realities," *Computer Communication Review*, vol. 26, no. 4, pp. 3–14, 1996.

[16] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.

[17] 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.

[18] 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.

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

[20] M. Parulekar and A. M. Makowski, "Tail probabilities for a multiplexer with self-similar traffic," in *IEEE INFOCOM 1996*, San Francisco, CA , USA, 1996, pp. 1452–1459.

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

[22] L. Massoulie and A. Simonian, "Large buffer asymptotics for the queue with fbm input," *Journal of Applied Probability*, vol. 36, no. 3, pp. 894–906, 1999.

[23] K. R. Krishnan, "A new class of performance results for fractional brownian traffic model," *Queueing Systems*, vol. 22, no. 3-4, pp. 277–285, 1996.

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

[25] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communication Review*, vol. 27, pp. 5–23, 1997.

[26] D. Heath, S. Resnick, and G. Samorodnitsky, "Heavy tails and long range dependence in on/off processes and associated fluid models," *Mathematics of Operations Research*, vol. 23, no. 1, pp. 145–165, 1998.

[27] O. J. Boxma and V. Dumas, "Fluid queues with long-tailed activity period distributions," Probability, Networks and Algorithms (PNA) PNA-R9705, April 30 1997.

[28] D. R. Cox, "Long-range dependence: A review," in *Statistics: An Appraisal*, H. A. David and H. T. David, Eds., pp. 55–74. The Iowa State University, Ames, Iowa, 1984.

[29] A. Elwalid, D. Heyman, T. V. Lakshman, and D. Mitra, "Fundamental bounds and approximations for atm multiplexers with applications to video teleconferencing," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1004–1016, 1995.

[30] D. P. Heyman and T. V. Lakshman, "What are the implications of long-range dependence for vbr-video traffic engineering?," *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 301–317, 1996.

[31] 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.

[32] A. T. Andersen and B. F. Nielsen, "A markovian approach for modeling packet traffic with long-range dependence," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 719–732, 1998.

[33] P.R. Jelenkovic, A. A. Lazar, and N. Semret, "The effect of multiple time scales and subexponentiality in mpeg video streams on queueing behavior," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 0733-8716, pp. 1052–1071, 1997.

[34] K. Park, "Future directions and open problems in performance evaluation and control of self-similar network traffic," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., pp. 531–554. John Wiley & Sons, Inc., 2000.

[35] G. Mao and D. Habibi, "Loss performance analysis for heterogeneous on-off sources with application to connection admission control," *IEEE/ACM Transactions on Networking*, vol. 10, no. 1, pp. 125–138, 2002.

[36] A. Erramilli, O. Narayan, A. Neidhardt, and I. Saniee, "Performance impacts of multi-scaling in wide area tcp/ip traffic," in *INFOCOM 2000*, Tel Aviv , Israel, 2000, vol. 1, pp. 352–359.

[37] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, no. 0018-9448, pp. 878–897, 1999.

[38] 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.

[39] A. Feldmann, A. Gilbert, W. Willinger, and T. Kurtz, "The changing nature of network traffic: Scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.

[40] M. S. Taqqu, V. Teverovsky, and W. Willinger, "Is network traffic self-similar or multifractal?," *Fractals*, vol. 5, pp. 63–73, 1997.

[41] 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 Willey & Sons, Inc., 2000.

[42] R. Roughan, D. Veitch, and P. Abry, "Real-time estimation of the parameters of long-range dependence," *IEEE/ACM Transactions on Networking*, vol. 8, no. 1063-6692, pp. 467–478, 2000.

[43] S. Q. Li and C. L. Hwang, "Queue response to input correlation functions: Continuous spectral analysis," *IEEE/ACM Transactions on Networking*, vol. 1, no. 6, pp. 678–692, 1993.

[44] Y. Kim and S. Q. Li, "Timescale of interest in traffic measurement for link bandwidth allocation design," in *IEEE INFOCOM 1996*, San Francisco, CA , USA, 1996, vol. 2, pp. 738–748.

[45] L. A. Kulkarni and S. Q. Li, "Measurement-based traffic modeling: Capturing important statistics," *Journal of Stochastic Model*, vol. 14, no. 5, 1998.

[46] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D*, vol. 31, pp. 277–283, 1988.

[47] M. Roughan and D. Veitch, "Measuring long-range dependence under changing traffic conditions," in *IEEE INFOCOM 1999*, New York, 1999, vol. 3, pp. 1513–1521.

[48] M. Taqqu, V. Teverovsky, and W. Willinger, "Estimators for long-range dependence: An empirical study," *Fractals*, vol. 3, no. 4, pp. 785–798, 1995.