A Simulation Study on the Relevant Time Scales of the Input Traffic for a Tandem Network

A. Nogueira, R. Valadas
Institute of Telecommunications - University of Aveiro
Campus Universitário Santiago, 3810-193 Aveiro, Portugal

Abstract - **Network traffic processes can exhibit properties of self-similarity and long-range dependence, i.e., correlations over a wide range of time scales. However, as already shown by several authors for the case of a single queue, the second-order behavior at time scales beyond the so-called correlation horizon or critical time scale does not significantly affect network performance. In this work, we extend previous studies to the case of a network with two queuing stages, using discrete event simulation. Results show that the introduction of the second stage provokes a decrease in the correlation horizon of the input traffic, meaning that the range of time scales that need to be considered for accurate network performance evaluation is lower than predicted by a single stage model. We also resorted to simulation to evaluate the single queue model. In this case, the estimated correlation horizon values are compared with those predicted by a formula derived by Grossglauser and Bolot, which presumes the approximation of the input data by a traffic model that enables to control the autocorrelation function independently of first-order statistics. Results indicate that although the correlation horizon increases linearly with the buffer size in both methods, the simulation ones predict a lower increase rate.**

Keywords - **traffic modeling, self-similarity, long-range dependence, wavelets, packet loss ratio, time scales, correlation horizon.***

I. INTRODUCTION

There is wide experimental evidence that network traffic processes exhibit properties of self-similarity and long-range dependence (LRD) [2][3][4]. Both phenomena are related to certain scale-independent statistical properties. A process is exactly self-similar if its aggregated processes have first and second-order properties that are indistinguishable from those of the original process itself. Long-range dependence means that the correlation decays slower than exponentially, i.e., it decays hyperbolically.

When studying the performance of a networking system, different types of time scales must be taken into account: time scales of the input traffic and of the system (which show up, for example, because of finite buffer queues). For the case of a single queue system, it was shown that second-order behavior at time scales beyond the correlation horizon (CH) or critical time scale (CTS) does not significantly affect network performance [1][5].

The time scale concept considered here is defined in the context of the autocovariance function. CH and CTS are two different terms representing essentially the same concept: the time scale or lag that separates relevant and irrelevant correlation with respect to the performance metric of interest. In this work, the metric is assumed to be the packet loss ratio (PLR). The CH depends on the correlation structure of the input traffic, on the system under study and on the PLR metric. In real networks, buffers have finite length. A finite length buffer "forgets" about the past as soon as it is either empty or full. Thus, while correlation on all time scales has an impact on the performance of an infinite queue, only the correlation up to the CH has an effect in a finite buffer queue. It was shown in [5] that the CH is finite, has a small value for small buffer sizes and is a non-decreasing function of the buffer size.

It is known that processes with the same correlation structure can generate vastly different queuing behavior. Performance evaluation depends not only on the time scales relevant to the system under study and on the correlation structure of the source traffic but also on other characteristics like the marginal distribution of the arrival rate process [1][6]. Our goal in this work is to concentrate on the joint influence of correlation structure of the input traffic and of the queuing system characteristics in the overall queuing performance, leaving other parameters such as the first-order statistics of the input traffic fixed.

In order to calculate the amount of correlation that needs to be taken into account for performance evaluation and to analyze its behavior for different queuing networks with variable number of stages, we consider a scenario (Fig. 1) consisting of buffers with adjustable parameters (queue length, link rate) and an input traffic stream exhibiting the LRD property. Several tasks are performed: (i) analysis of the LRD characteristics of the input traffic stream, using the autocorrelation function and the Wavelet based estimator [7]; (ii) estimation, via simulation, of the CH and analysis of its behavior; (iii) for the single buffer case, comparison of the CH values obtained using our simulation approach and the method proposed in [1]. The network that constitutes the basis for our simulation studies replicates a realistic scenario consisting of a network of switching elements, e.g., routers (Fig. 1).

* This work was part of project 34826/99 SCALE "Statistical Characterization of Telecommunications Traffic", funded by Fundação para a Ciência e Tecnologia, Portugal. We also acknowledge the support of Portugal Telecom Inovação.

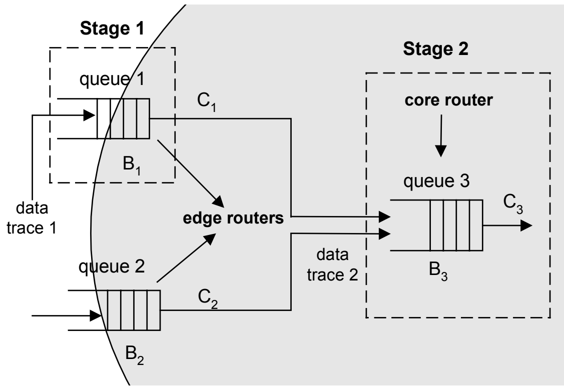

Fig. 1: Network of finite queues used in the study of the correlation horizon

The buffers of stage 1 belong to edge routers, located at the entry points of the network. For the sake of simplicity, we restrict our scenario to two input queues only. The buffer of stage 2 belongs to a core router, located at the core of the network and is typically fed by the superposition of several traffic streams, which are output flows of other queues.

The remaining part of this paper is structured as follows: section 2 analyzes the characteristics of the input trace, including the study of its LRD behavior using the wavelet based estimator tool; in section 3, we present the methodology used in this study; in section 4, the efficiency of the external shuffling procedure will be evaluated; in section 5, the main results are presented and discussed; finally, in section 6, we draw the main conclusions.

II. CHARACTERISTICS OF THE INPUT TRACE

Our study of the correlation horizon is based on discrete event simulation and uses the August Bellcore data trace [2], which will be designated in the remaining part of the paper by pAug. This trace has a total of 10^6 packet arrival instants and packet size values; the average packet size is 434.29 bytes and the average interarrival time value is 3.1 ms.

The analysis of the autocovariance function, represented in Fig. 2, lead us to suspect that the trace has a LRD behavior, due to the slow decay for large time lags.

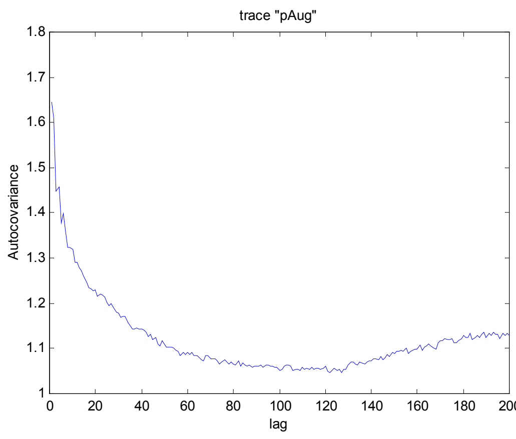

Fig. 2: Autocovariance function of the interarrival times for trace pAug

The presence of LRD behavior in the input trace is confirmed using the scaling analysis method described in [7]. This method resorts to the so-called Logscale Diagram which consists in the graph of y_j against j, together with confidence intervals about the y_j, where y_j is a function of the wavelet discrete transform coefficients at scale j. Traffic is said to be LRD if, within the limits of the confidence intervals, the y_j fall on a straight line, in a range of scales from some initial value j_1 up to the largest one present in data.

The results of the scaling analysis for trace pAug are presented in Fig. 3, where it can be seen that y_j values are aligned between a medium octave, 11, and octave 18, the highest octave present in data.

III. METHODOLOGY PROPOSED TO STUDY THE CORRELATION HORIZON

The study of the CH will be based on two queuing scenarios: a single-buffer system and a two-buffer tandem system, which will be designated in the remaining part of the paper by single-stage and two-stage queuing networks, respectively.

Applying the pAug data trace to queue 1, we control the “extent” of its correlation using a shuffling technique. Data trace 2 in Fig. 1 represents another traffic flow of the same service and is simply a shifted version of the output trace from queue 1. It will be denoted by Queue1Shift.

We use the external shuffling procedure, described in [8], to eliminate correlation in the input process beyond a certain lag. In this procedure, a time series representing a realization of a process is divided up into blocks and the blocks are shuffled. However, the structure of the time series inside a block remains unchanged. In this way, external shuffling removes correlation from the time series beyond a lag equal to the length of a block.

The main parameters of the queuing systems for both stages are selected according to the order of magnitude expected for the PLR values. For stage 1, we chose an output link capacity C_1 according to the average arrival rate λ_i of the input data trace. For the stage 2 buffer, the average arrival rate is $2\lambda_o < 2\lambda_i$. In order to have the same utilization ratio at both stages, the selected capacity C_3 must be slightly lower than $2C_1$.

Applying pAug data trace to the single-stage network, we determine through simulation the PLR values for varying block lengths and normalized buffer sizes, B/C. The normalized buffer size represents the time it takes to empty a buffer of length B packets at an output rate of C packets/s.

The CH, being the time scale that separates relevant and irrelevant correlation with respect to PLR, corresponds to the block size value beyond which PLR does not change significantly. According to this definition, we first interpolate linearly the *PLR vs block size* (defined according to the shuffling method) curve and try to minimize the error by

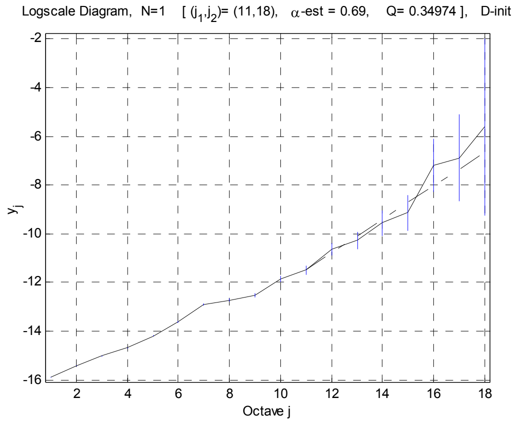

Fig. 3: Scaling analysis for trace pAug.

performing additional simulations around the threshold block size value (the one that corresponds to CH value). Considering a tolerance of *x*% for the PLR, the threshold block size value is given by

$$b_t = \frac{(1-x)P_f(B_2 - B_1) - P_1 B_2 + P_2 B_1}{P_2 - P_1} \tag{1}$$

where P_f is the PLR corresponding to the unshuffled trace, P_2 and P_1 are the PLR values for the higher and lower endpoints of the segment that includes $(1-x)P_f$, respectively, and B_2 and B_1 are the block size values corresponding to the higher and lower endpoints of the segment. In our case, we assumed that CH corresponds to the block size beyond which PLR has a value less than or equal 5% of the final value (corresponding to the unshuffled trace). This value represents the tolerance admitted in PLR values for the estimation of the CH.

For the two-stage queuing network, pAug data trace is applied to the first queue and the resulting output flow is applied, together with trace Queue1Shift, to the second queue. The new CH values are calculated using the same procedure described above. Note that in this scenario we have two queuing systems whose parameters can be varied independently, leading to a larger number of possible parameter combinations.

In the case of a single-buffer system, in order to evaluate the accuracy of the estimated CH values we approximate the input data trace by a traffic model based on the process proposed in [1]. In this model, the correlation structure of the traffic source can be controlled independently of other parameters (such as the marginal distribution) and the CH values can be calculated analytically. These features enable us to examine the impact on PLR of considering only the relevant time scales of the traffic source. For this scenario, the calculation of the CH is then performed in two different ways: analytically and using our simulation approach.

The traffic model proposed in [1] has *N* states and is characterized by the arrival rate matrix, Λ, a vector Π representing the marginal distribution of the arrival rates and the distribution of the states duration, $F_T(t)$. The duration of each state is modeled by a truncated Pareto distribution, with cumulative distribution function given by:

$$F_T(t) = \begin{cases} \left(\dfrac{t+\theta}{\theta}\right)^{-\alpha} & if\ t < T_C \\ 0 & otherwise \end{cases} \tag{2}$$

where α, $1 < \alpha < 2$ represents the exponential decay. The parameter T_C is called cutoff lag and, since the duration of a state cannot exceed T_C and since the rates in consecutive intervals are independent, there is no correlation in the rate process beyond lag T_C. Therefore, this parameter represents the CH.

The fitting algorithm that we used to estimate the parameters of this model from the empirical trace is based on the fitting approach presented in [1]. Vectors Π and Λ are estimated from a constant bin-size histogram of the number of arrivals. In the conversion step from the arrival instant to the number of arrivals formats, the sample interval was set to 10 ms. The construction of the histogram was based on 50 bins in all our experiments, as in [1].

In order to completely define the truncated Pareto distribution of (2), one must estimate the parameters T_C, θ and α. The correlation horizon, T_C, can be calculated analytically using the result presented in [1]:

$$T_{CH} = \frac{B}{2\sqrt{2}\sigma_T \sigma_\lambda erf^{-1}(p)}(\mu + \beta\sigma_T) \tag{3}$$

where μ and σ_T represent the average and variance of the interarrival times, respectively, σ_λ is the variance of the arrival rate distribution, B is the normalized buffer size and p is the probability of emptying or overflowing the buffer.

The exponential decay α is calculated from the Hurst parameter. Using a variance-time plot, a Hurst parameter of 0.8745 was found, which gave a α value of 1.2509 ($\alpha = 3 - 2H$).

In order to estimate θ, we first calculate the average number of consecutive samples in the trace that fall within the same histogram bin. Then, θ is calculated assuming that the mean interval duration, which is given by

$$E[T_n] = \frac{\theta}{\alpha - 1}\left[1 - \left(\frac{T_c}{\theta} + 1\right)^{1-\alpha}\right] \tag{4}$$

matches that empirical mean, for each bin. Equation (4) is solved numerically for θ.

IV. Results and Discussion

A. Single-stage network

Applying the pAug data trace to the single-stage queuing network, we simulated the *PLR vs block size* function, for

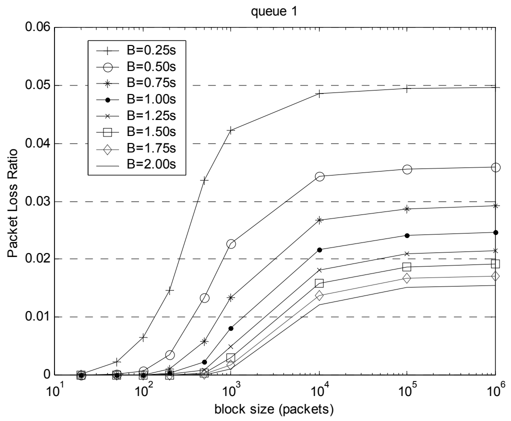

Fig. 4: Packet loss ratio vs block size in the single-stage network, for different normalized buffer sizes, B_1.

different normalized buffer sizes. In order to have reliable results, from a statistical point of view, ten simulations were performed for each parameter set and the results plotted in Fig. 4 correspond to the mean values of those runs. The corresponding confidence intervals were very small in all situations, so they are not represented in the plots.

We used normalized buffer sizes of up to 2 seconds, which are typical values for currently available switches. For example, an ATM switch with a buffer of 7000 cells will produce a maximum delay in the buffer approximately equal to $2\,s$, considering a T1 link at 1.5 Mbit/s. The selection of the output link capacity depends on the utilization ratio in each queue and must result in PLR values that span over several orders of magnitude. For the first stage, the capacity was set to 1.4 Mbit/s.

The correlation horizon values calculated for the single-stage network are represented in the lower curve of Fig. 5. We can see that the CH varies almost linearly with the normalized buffer size.

The work presented in [1] predicts that the slope of the *CH vs normalized buffer size* curve is one, as can be confirmed from (3) and from the top curve of Fig. 5. However our simulation results predict a lower slope.

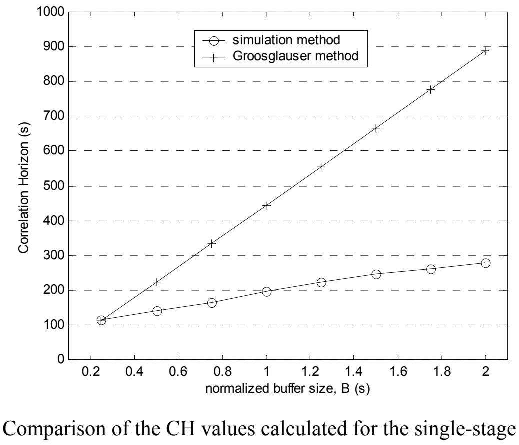

Fig. 5: Comparison of the CH values calculated for the single-stage network using the simulation method and the method described in [1]

Both methods lead to similar CH values only for small normalized buffer sizes, and the deviation increases with the normalized buffer size. For $B = 0.50$ s, for example, the difference between both values is about 48%. The effect, in terms of PLR, of these differences in the estimated CH values is illustrated in Fig. 6. Three *PLR vs normalized buffer size* curves are plotted, corresponding to queuing results obtained when we fed a buffer with (i) the original trace pAug; (ii) traffic generated according to Grossglauser traffic model with parameters estimated using the fitting procedure presented in [1] and (iii) traffic generated according to Grossglauser traffic model but whose parameters were estimated using our simulation method. We can conclude that (3) provides a pessimistic estimation of the relevant correlation. This equation was derived based on the assumption that the *buffer occupancy plus sum of the excess work in* n *consecutive intervals* is approximately normally distributed. Our simulation results show that real relevant correlation that needs to be taken into account for PLR prediction is much smaller than that, making our modeling tasks of the autocorrelation function easier to perform. This, in fact, enlarges the set of traffic models that can be used to match long-range dependence characteristics.

Note that in our simulation approach we use the real packet lengths of the Bellcore empirical traces, while the simulation of the queuing systems based on traffic generated according to the Grossglauser traffic model is performed using fixed length packets, whose size equals the average packet size of the Bellcore trace.

B. Two-stage network

This scenario includes both queuing stages of Fig. 1. The output flow of stage 1 is applied, with trace Queue1Shift, to the second-stage queue. The calculation of the CH is again performed by simulating the end-to-end *PLR vs block size* function, for different normalized first and second buffer sizes. With two queuing stages, there are several possibilities regarding the relative influence of each stage in the end-to-end PLR performance. We consider three cases, corresponding to different values of the second-stage output link capacity, C_3: (i) $C_3 = 2.1 Mbit/s$ - in this case, the relative influence of the second-stage PLR in the end-to-end PLR is much higher than the first-stage influence; (ii) $C_3 = 2.7 Mbit/s$ - this situation corresponds to approximately equal relative weights of both stages in the end-to-end PLR; (iii) $C_3 = 3.5 Mbit/s$ - in this case, there is a much higher relative influence of the first-stage PLR in the end-to-end PLR. Ten simulations were performed for each parameter set and the corresponding mean values and confidence intervals were calculated.

Considering case (i), we plot in Fig. 7 and Fig. 8 the end-to-end PLR curves for varying normalized second-buffer sizes, and for two extreme values of the normalized first-buffer size: $B_1 = 0.25s$ and $B_1 = 2.00s$, respectively.

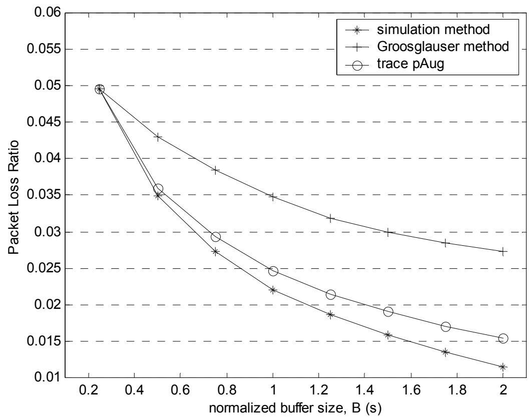

Fig. 6: Comparison of the effect, in terms of PLR, of the CH values calculated using the simulation method and the method described in [1]

As we can see, there is a slight increase in the end-to-end PLR values as the buffer size B_1 increases. In fact, we verified that first-stage PLR has decreased, while second-stage PLR has increased in a higher proportion, leading to the increase of the overall PLR values. However, the general behavior of the various *PLR vs block size* curves remains the same.

Regarding the CH, Fig. 9 plots the *CH vs normalized second-buffer size* curves for case (i), considering the extreme values of B_1 mentioned above. Comparing both curves, we see that as the size of the first buffer increases the CH value at the input of the overall network slightly increases (for the same size of buffer 2). However, the variation is not very significant because the influence of B_1 in the end-to-end PLR (and therefore, in CH) is small.

The corresponding plots for case (iii) are also presented in Fig. 9. Here, it is clear that first-buffer size has a strong influence in the end-to-end CH, as can be concluded from the large separation that exists between both curves and from their very low slopes.

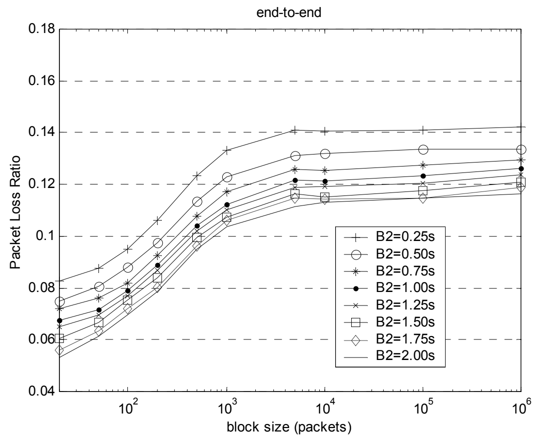

Fig. 7: End-to-end PLR vs block size, for different normalized second-buffer sizes and for a normalized first-buffer size of $B_1 = 0.25s$.

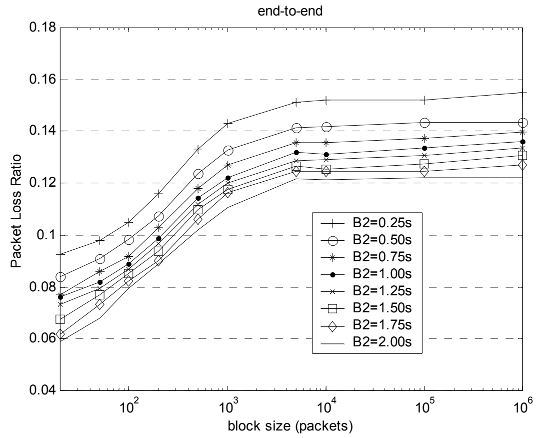

Fig. 8: End-to-end PLR vs block size, for different normalized second-buffer sizes and for a normalized first-buffer size of $B_1 = 2.00s$.

For case (ii), the results obtained demonstrate that the variation at the input CH is induced, in similar proportions, by the normalized buffer sizes of both stages.

From the analysis of these results it is clear that CH behavior at the input of the two-stage network is essentially determined by the most critical stage, from the PLR point of view, that is, the one that induces higher PLR values.

In the two-stage queuing scenario, CH also varies almost linearly with the normalized buffer size of each queue, as already seen in the single-buffer system. The slope of the curve is also less than one.

Comparing directly the CH values obtained for the single-stage network (Fig. 5) with those corresponding to the two-stage network, we see that the first ones are always higher, which means that in the second scenario the relevant correlation (in terms of PLR) ends at lower lags.

V. CONCLUSIONS

In this paper we evaluated the amount of correlation that needs to be taken into account for performance evaluation of queuing networks having different number of queuing stages. More precisely, two queuing networks were considered: one with a single buffer and another with two buffers in tandem. In order to perform this study, we (i) presented a discrete-event simulation methodology for the estimation of the correlation horizon values; (ii) calculated and compared the results obtained for both queuing networks, and (iii) compared the results obtained for the single-stage case with the ones presented by Grossglauser and Bolot. We note that results presented so far in the literature deal only with single queuing stages.

Our results show that CH values for the two-stage network are lower than the corresponding values for the single-stage case, which means that the relevant correlation (in terms of packet loss ratio) ends at lower lags.

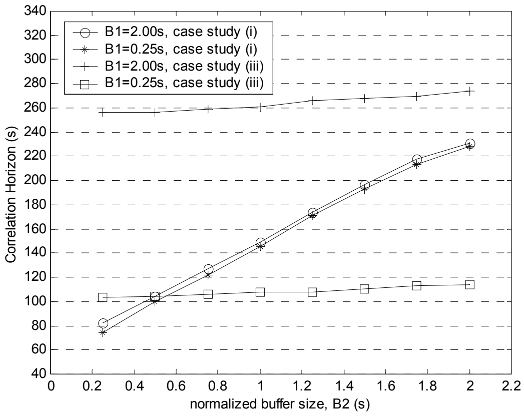

Fig. 9: Comparison of the *end-to-end CH vs normalized second-buffer size*, for two different values of B_1 and for case studies (i) and (iii)

For the single buffer case, Grossglauser and Bolot have concluded that when keeping the interarrival time and arrival rate distribution functions unchanged the CH varies linearly with the normalized buffer size with a slope equal to 1. However, our simulation results showed that the slope may be lower and that the deviation from unitary slope increases with the normalized buffer size.

REFERENCES

[1] M. Grossglauser and J. C. Bolot, "On the Relevance of Long-Range Dependence in Network Traffic", IEEE/ACM Trans. On Networking 7, October 1999.

[2] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic", IEEE/ACM Trans. On Networking 2, pp. 1-15, February 1994.

[3] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modelling", IEEE/ACM Trans. On Networking 3, pp. 226-244, June 1995.

[4] J. Beran, R. Sherman and W. Willinger, "Long Range Dependence in Variable Bit Rate Video Traffic", IEEE Trans. On Communications 43, pp. 1566-1579, February 1995.

[5] B. K. Ryu and A. Elwalid, "The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities", Proc. ACM SIGCOMM'96, Stanford, CA, August 1996.

[6] A. Anderson and B. Nielsen, " A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence", IEEE Journal on Selected Areas in Communications 16, pp. 719-732, June 1998.

[7] D. Veitch and P. Abry, "A wavelet based joint estimator for the parameters of LRD", Special issue on Multiscale Statistical Signal Analysis and its Applications - IEEE Trans. Inform. Theory 45, April 1999.

[8] A. Erramilli, O. Narayan and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic", IEEE/ACM Trans. On Networking 4, April 1996.

[9] S. Q. Li and C. L. Hwang, "Queue Response to Input Correlation Function: Continuous Spectral Analysis", IEEE/ACM Trans. On Networking 1, pp. 678-693, March 1993.