Modeling Network Traffic with Multifractal Behavior

António Nogueira (nogueira@av.it.pt)
University of Aveiro / Institute of Telecommunications, Aveiro, Portugal

Paulo Salvador (salvador@av.it.pt)
University of Aveiro / Institute of Telecommunications, Aveiro, Portugal

Rui Valadas (rv@det.ua.pt)
University of Aveiro / Institute of Telecommunications, Aveiro, Portugal

António Pacheco (apacheco@math.ist.utl.pt)
Instituto Superior Técnico - Technical University of Lisbon, Mathematics Department, CEMAT and CLC, Lisboa, Portugal.

Abstract. The traffic engineering of IP networks requires accurate characterization and modeling of network traffic, due to the growing diversity of multimedia applications and the need to efficiently support QoS differentiation in the network. In recent years several types of traffic behavior, that can have significant impact on network performance, were discovered: long-range dependence, self-similarity and, more recently, multifractality. The extent to which a traffic model needs to incorporate each of these characteristics is still the subject of much research. In this work, we address the modeling of network traffic multifractality by evaluating the performance of four models, which cover a wide range of traffic types, as mathematical descriptors of measured traffic traces showing multifractal behavior. We resort to traffic traces measured both at the University of Aveiro and at a Portuguese ISP. For the traffic models, we selected a Markov modulated Poisson process as an example of a Markovian model, the well known fractional Gaussian noise model as an example of a self-similar process and two examples of models that are able to capture multifractal behavior: the conservative cascade and the L-system. All models are evaluated comparing the density function, the autocovariance and the loss ratio queuing behavior of the measured traces and of traces synthesized from the fitted models. Our results show that the fractional Gaussian noise model is not able to perform a good fitting of the first and second order statistics as well as the loss rate queuing behavior, whereas the Markovian, the conservative cascade and the L-system models give similar and very good results. The cascade and the L-system models are intrinsically multifractal in the sense that they are able to capture and synthesize traffic multifractality, thus the obtained results are not surprising. The good performance of the Markovian model can be attributed to the parameter fitting procedure, that aggregates distinct sub-processes operating in different time scales, and matches closely both the first and second order statistics of the traffic. The poor performance of the self-similar model can be explained mainly by its lack of parameters.

Keywords: Traffic modeling, self-similar, multiscaling, multifractal, L-system, fractional Gaussian noise, Markov modulated Poisson process, conservative cascade.

1. Introduction

Efficient design and control of telecommunications networks need to take into account the main characteristics of the supported traffic. Therefore, it is important to measure packet flows and to describe them through appropriate

traffic models. Traffic modeling comprises three steps: (i) selection of one or more models that may provide a good description of the traffic type, (ii) estimation of parameters for the selected models, and (iii) statistical testing for election of one of the considered models and analysis of its suitability to describe the traffic type under analysis. Parameter estimation is based on a set of statistics (e.g. mean, variance, density function or autocovariance function, multifractal characteristics) that are measured or calculated from observed data. The set of statistics used in the inference process depends on the impact they may have in the main performance metrics of interest.

An effective traffic model has, at least, to reproduce the first and second order statistics of the original traffic trace. The density function defines the first order statistics whereas the second order statistics can be accounted for by the autocovariance function. The second order statistics play an important role in traffic modeling, because traffic correlation is an important factor in packet losses due to buffer and bandwidth limitations. However, the first two order statistics may not be sufficient to characterize real data traces, that are known to be bursty and spiky in nature. In these cases, higher order statistics must also be included in the models and fitting procedures in order to get a real picture of network data.

It has been shown through experimental evidence that network traffic may exhibit properties of self-similarity and/or long range dependence (LRD), which have significant impact on network performance. Matching LRD is only required within the time scales specific to the system under study and one of the consequences of this result is that, in principle, more traditional traffic models, such as Markovian models, can still be used to model traffic exhibiting LRD. The use of Markovian models also benefits from the existence of several mathematical tools for assessing queuing behavior, such as average delay and packet loss ratio.

However, more recent analysis of measured Internet WAN traffic has revealed that multifractal structures, such as random cascades, can help explaining the scaling behavior typically associated to networking mechanisms operating on small time scales (e.g. TCP flow control). A cascade (or multiplicative process) is a process that fragments a set (typically an interval) into smaller and smaller components according to a fixed rule, and at the same time fragments the measure of the components by another (possibly random) rule. Random cascades were introduced by Mandelbrot as a physical model for turbulence [1] and in this paper we consider a conservative cascade model, which is a special case of a random cascade. The multifractal nature of network traffic was first noticed by Riedi and Véhel [2]. Subsequently various studies have addressed the characterization and modeling of multifractal traffic, essentially within the framework of random cascades [3] [4] [5] [6] [7] [8] [9] [10].

In this study we evaluate the need for including different descriptors (e.g. the first and second order statistics, the Hurst parameter, the multiscaling or multifractal characteristics) of real multifractal network traffic in order to obtain an accurate fitting of its main statistics and queueing performance. This is accomplished by comparing the first and second order statistics and the queuing behavior of (i) original measured data traces and (ii) traces generated via discrete event simulation of the traffic models whose parameters are inferred from the measured data. Different types of models are considered as candidates to characterize the measured traffic traces, ranging from Markovian to multifractal models: specifically, the selected models are the Markov modulated Poisson process (MMPP) as an example of a Markovian model, the fractional Gaussian noise model (fGn) as an example of a self-similar process and the conservative cascade and L-system models as examples of multiscaling/multifractal models. These models cover a wide range of traffic characteristics, enabling us to assess their relevance for an accurate fitting of multifractal traffic. The traffic traces used were measured both at the University of Aveiro and in the premises of a Portuguese ISP.

The fitted traffic model can be used to enable Quality of Service (QoS) deployment in the network. For example, dimensioning Differentiated Services (DiffServ) networks requires the calculation of the traffic parameters of each service class, which can only be done accurately through appropriate traffic models.

The results obtained in this paper show that the fractional Gaussian noise model is not able to perform a good fitting of the first and second order statistics as well as the loss rate queuing behavior of the original traffic traces, because it is a too parsimonious model having only two adjustable parameters: the mean and the Hurst parameter. The fitting procedure selected for the MMPP model performs an accurate fitting of the first and second order statistics, and this accuracy reveals itself sufficient in the prediction of the empirical queuing behavior. Finally, the conservative cascade and the L-system models also give very good results (actually, they are close and even better to the ones predicted by the MMPP model) in predicting the empirical queuing behavior and fitting the first and second order statistics. This result is not surprising, since these models have the ability to include time-varying scaling characteristics in its framework, which are present in the real traces considered in this paper.

The paper is organized as follows. Section 2 presents some basic definitions regarding important aspects of network traffic characterization, like LRD, self-similarity and multiscaling. Section 3 presents the traffic models considered in this study, as well as their fitting procedures, identifying also their main characteristics, advantages and drawbacks. Section 4 briefly presents the data traces used in the simulations and in Section 5 we discuss the results of applying the proposed models and fitting procedures to

the measured and synthesized traces. Finally, Section 6 presents the main conclusions.

2. Some essential definitions

Consider the continuous-time process $Y(t)$ representing the traffic volume (e.g. in bytes) from time 0 up to time t and let $X(t) = Y(t) - Y(t-1)$ be the corresponding increment process (e.g. in bytes/second). Consider also the sequence $X^{(m)}(k)$ which is obtained by averaging $X(t)$ over non-overlapping blocks of length m, that is

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

A process is exactly self-similar (H-SS) if it has a statistical scale invariance property that holds for all of the distributions and moments of the process, not just the second order ones. More precisely, $Y(t)$ is H-SS when it is equivalent, in the sense of finite-dimensional distributions, to $a^{-H}Y(at)$, for all $t > 0$ and $a > 0$, where H $(0 < H < 1)$ denotes the Hurst parameter which represents the scaling parameter of self-similarity. Clearly, the process $Y(t)$ can not be stationary. However, if $Y(t)$ has stationary increments then again $X(k) = X^{(1)}(k)$ is equivalent, in the sense of finite-dimensional distributions, to $m^{1-H}X^{(m)}(k)$.

Long-range dependence is a property associated with stationary processes. Consider now that $X(k)$ is second-order stationary with variance σ^2 and autocorrelation function $r(k)$. Note that, in this case, $X^{(m)}(k)$ is also second-order stationary. A stationary stochastic process with finite second order moments is long range dependent if its autocorrelation function $r\,(k)$ is non-summable, that is $\sum_k r\,(k) = \infty$. Thus, in a rigorous way the definition of LRD applies only to infinite time series, but generally this definition is also used for finite series as well. Using an equivalent definition, one can say that a stationary stochastic process is LRD if its spectrum diverges at the origin, that is $f(v) \sim c_f|v|^{-\alpha}, v \to 0$, where α is the dimensionless scaling exponent, takes values in $[0, 1)$ and is the most important parameter describing the qualitative nature of the scaling. The parameter c_f takes positive real values, it has the dimensions of variance and describes the quantitative aspect or "size" of the LRD. The value of c_f has a large impact in the overall correlation: for example, confidence intervals around mean estimates of LRD data are essentially proportional to the square root of c_f. A short range dependent (SRD) process is simply a stationary process which is not LRD. Such a process has $\alpha = 0$ at large scales, corresponding to white noise at scales beyond the so called characteristic scale or correlation horizon.

There are several estimators of LRD: the estimator used in this study was proposed in [6] and is semi-parametric. In this case, prior to the estimation an analysis phase is necessary to determine the lower cutoff scale at which the LRD "begins" and to see if LRD is present at all. To do this one looks for alignment in the Logscale Diagram (LD), which is essentially a log-log plot of variance estimates of the wavelet details, against scale, complete with confidence intervals about these estimates at each scale. It can be thought of as a spectral estimator where large scale corresponds to low frequency. More precisely, the LD consists in a 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.

There is a close relationship between long-range dependent and self-similar processes. In fact, if $Y(t)$ is self-similar with stationary increments and finite variance, then $X(k)$ is long-range dependent, as long as $\frac{1}{2} < H < 1$. The process $X(k)$ is said to be exactly second-order self-similar ($\frac{1}{2} < H < 1$) if

$$r\left(n\right) = 1/2\left[\left(n+1\right)^{2H} - 2n^{2H} + \left(n-1\right)^{2H}\right] \tag{2}$$

for all $n \geq 1$, or is asymptotically self-similar if

$$r\left(n\right) \sim n^{-(2-2H)} L(n) \tag{3}$$

as $n \rightarrow \infty$, where $L(n)$ is a slowly varying function at infinity. In both cases the autocovariance decays hyperbolically, which indicates LRD. Any asymptotically second-order self-similar process is LRD, and vice-versa. Using the LRD estimator described above, the Hurst parameter can be obtained knowing that the slope α is related with H by $H = (\alpha + 1)/2$. Note that strictly speaking the Hurst parameter characterizes H-SS processes, although it is often also used to describe the LRD of "derivatives" of such processes, which can lead to some misunderstanding in some situations.

Examining the presence of LRD (using the LD, for instance) means to analyze the rate of decay of the correlations for large timescales, that is, to look at the traffic behavior at low frequencies. However, in real traffic traces it is possible to observe a wide range of irregular behaviors at all small timescales. The meaning and properties of this high-frequency behavior, which is closely related to multifractality, must be analyzed because, like LRD, multifractality may also have serious consequences on network performance [11].

If a process has scaling in some second order statistic, for example if the variance $S_2(j)$ of the wavelet coefficients at octave j obeys $S_2(j) \sim Cj^{\alpha}$, then it will very often have scaling for all moments, $S_q(j) = E[|d(j)|^q] \sim C_q j^{\alpha_q}$, where $d(j)$ represents the wavelet coefficient of the sequence at scale j. For simple scaling processes such as H-SS processes, the function α_q is

simply given by $\alpha_q = Hq + q/2$, a simple linear relationship. The Hurst parameter H controls each exponent, and this constitutes an example of a monofractal process. If on the other hand α_q is not linear, then we say that it exhibits multiscaling. Multifractals fall into this category. The $S_q(j)$ are known as the structure or partition functions. They are often based on the increments of a process, rather than the statistically more advantageous wavelet coefficients. It is common to define the exponents in a slightly different way:

$$\zeta_q = \alpha_q - q/2 \tag{4}$$

which gives the elegant relation $\zeta_q = Hq$ for H-SS processes. ζ_q is then the slope read directly in the so called Zeta Diagram (ZD) or Multiscale Diagram (MD), ζ_q as a function of q, in an analogous way to α_q which was the slope observed in the LD. Usually, a Linear Multiscale Diagram (LMD) is constructed, representing $h_q = \zeta_q/q$ as a function of q. Non-trivial multifractal scaling behavior is detected when there is no horizontal alignment in this diagram (within the limits of confidence intervals), reflecting the non-linear nature of ζ_q in the MD [12]. Note that the range of q is typically semi-infinite, from some value up to positive infinity. If the partition functions display the power-law scaling at small scales, then ζ_q can be estimated and thought of as a function of a continuous moment parameter q. The Legendre transform of this function is the Legendre Multifractal spectrum, another experimentally accessible form of the Multifractal Spectrum describing and defining Multifractals. The estimation of the ζ_q is semi-parametric based, and is analogous to that used at second order in the LD.

3. Selected models and fitting procedures

3.1. The fractional Gaussian noise model

Self-similar processes have gained much attention since it was empirically observed that LAN traffic was positively correlated and long range dependent. These facts definitely challenged Poissonian and Markovian traffic modeling, leading to other modeling approaches that could account for these new features. One of the most widely studied self-similar processes is fractional Gaussian noise (fGn), a stochastic process that is the formal derivative of fractional Brownian motion (fBm). Fractional Brownian motions are a family of Gaussian processes that are indexed by the Hurst parameter, H, in the interval $(0, 1)$. These processes are exactly self-similar, that is, the distribution of $Z(\alpha t)$ is identical to that of $\alpha^H Z(t)$, where $Z(t)$, $t \in (-\infty, +\infty)$, represents the fBm process. For H in $(1/2, 1)$, these processes have a LRD property that is characterized by the relatively slow decay of the correlation

(or covariance) function:

$$\langle Z(t_1) Z(t_2) \rangle = \frac{1}{2} \left(t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H} \right). \quad (5)$$

Several analytical closed-form results are available for these models, which is another important advantage supporting their use. Norros [13], for example, has derived several results for the queuing behavior obtained by driving a deterministic service time queue with a fBm process, including asymptotic lower bounds for the probability $P(V > x)$ that the queue length V exceeds *x*.

In this paper we use the method proposed in [14] for synthesizing fGn traffic. This method is based on the discrete time Fourier transform and only requires an estimation of the Hurst parameter of the empirical data. The mean and variance of the generated trace are then adjusted to the empirical mean and variance (using simple linear transformations), and the negative values of the generated trace are eliminated because they have no physical meaning. Since we are looking for a traffic process, the generated data values can be interpreted as the number of arrivals, packets or bytes within a time interval.

3.2. The Markov modulated Poisson process

Since matching LRD is only required within certain time scales, Markovian models can still be used to model traffic exhibiting this property. These models have been widely used in traffic modeling, specially MMPPs, which are the most popular. Several fitting procedures have been proposed in the literature for estimating their parameters from empirical data [15], [16], [17], [18], [19] and [20]. In this paper we use a fitting procedure for MMPPs, proposed in [21] and [22], that matches both the autocovariance and marginal distribution of the counting process.

Matching simultaneously the autocovariance and the marginal distribution is a difficult task since every MMPP parameter has an influence on both characteristics. With the purpose of achieving some degree of decoupling when matching these two statistics, the MMPP is constructed as a superposition of two MMPPs, where one MMPP (with 2^L states) is used to adjust the autocovariance and the other (with M states) is used to adjust the marginal distribution, taking into account the contribution of the first MMPP (figure 1). We will denote the resulting process as M2^L-MMPP.

The 2^L-MMPP matching the autocovariance is a superposition of L independent 2-MMPPs. Given that the autocovariance of a 2-MMPP is a single exponential, this approach allows matching the empirical autocovariance based on prior approximation by a weighted sum of exponentials, which results in a simple and accurate procedure. The M-MMPP matching the marginal distribution is forced to have null autocovariance at positive lags, to assure

that the autocovariance of the M2^L-MMPP equals that of the superposition of the L 2-MMPPs; the marginal distribution of the M-MMPP is obtained through deconvolution of the L 2-MMPPs and M2^L-MMPP marginal distributions, thus ensuring that the contribution of the L 2-MMPPs is taken into account. The autocovariance modeling is such that each 2-MMPP (in the set of L 2-MMPPs) models a specific time-scale. Here, the concept of time-scale is defined in the context of second-order statistics: each time-scale is associated with a characteristic time constant of the autocovariance function. A major feature of the procedure is that the number of states is not fixed a priori. It is an output of the fitting process, thus allowing the number of states to be adapted to the particular trace being modeled.

We note that, in order to boost the computational efficiency of the fitting procedure, the 2-MMPPs used to fit the autocovariance function are interrupted Poisson processes, and that the value of M of the M-MMPP is chosen as the smallest value that provides a given degree of matching between the data and the marginal probability function of the fitted model. Moreover, the assumed independence between the M-MMPP and the L 2-MMPPs introduces further constraints on the parametric form of the fitted $M2^L$-MMPP.

The fitting procedure that is used starts by approximating the autocovariance by a weighted sum of exponential functions. As part of this step, the relevant time-scales of the data are identified. After this, the procedure fits the M-dMMPP parameters in order to match the probability function, within the constraints imposed by the autocovariance matching. The final M2^L-dMMPP is obtained by superposing the L 2-dMMPPs and the M-dMMPP.

The proposed MMPP, with its associated fitting procedure, matches the first and second-order statistics of the empirical data, but it does not enable to incorporate directly time-dependent scaling behavior in its mathematical framework.

We have considered discrete time MMPPs (dMMPPs) instead of continuous time MMPPs, since they are more natural models for data corresponding to the number of arrivals in a sampling interval. Note that discrete time and continuous time MMPPs are basically interchangeable (through a simple parameter rescaling) as models for arrival processes, whenever the sampling interval used for the discrete time version is small compared with the average sojourn times in the states of the modulating Markov chain.

3.3. The Conservative Cascade model

The third traffic model considered in this study is a conservative cascade model, a special case of a random cascade able to fit the multiscaling characteristics of the empirical data. A random cascade can be used to construct traffic processes in an iterative way. It is characterized by an initial mass,

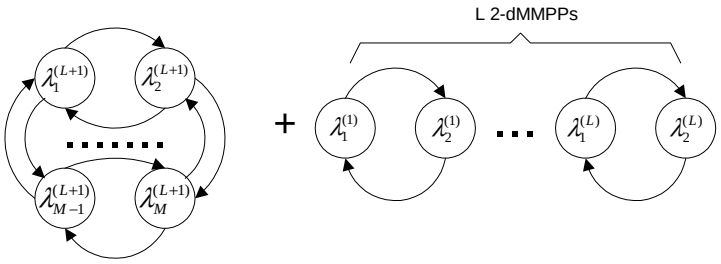

Figure 1. Superposition of M-dMMPP and L 2-dMMPP models.

uniformly distributed over a single interval, which is subdivided along the different stages of the cascade construction. In the case of a traffic process, the mass can be interpreted as the number of arrivals or bytes within a time interval. Each interval is divided in two (or more) identical subintervals and the mass is randomly assigned to each subinterval, according to a random variable W called the generator. Let $W_{i,j}, i = 1, ..., N_j, j = S-1, ..., 0$, denote an (independent) random variable, having the same distribution as W, that redistributes mass from time scale j into time interval i (belonging to subsequent time scale $j-1$).

The construction of the traffic process, starts at the coarsest time scale $S-1$ with an initial mass M distributed uniformly over a unit time interval. We will restrict our discussion to the case where a (parent) interval is subdivided in only two (child) intervals. In the first iteration, a finer time scale is produced, by dividing the time interval in two new subintervals of length $1/2$, and assigning mass $MW_{1,S-1}$ to the left subinterval and $MW_{2,S-1}$ to the right subinterval. At the second iteration, producing time scale $S-2$, each of these intervals generates new subintervals, one at the left and another at the right, giving rise to four subintervals of length $1/4$ with masses $MW_{1,S-1}W_{1,S-2}$, $MW_{1,S-1}W_{2,S-2}$, $MW_{2,S-1}W_{3,S-2}$ and $MW_{2,S-1}W_{4,S-2}$, respectively. Iterating this construction process, the mass of parent interval i from scale j is redistributed into child subintervals $2i-1$ and $2i$ of scale $j-1$ with probabilities $W_{2i-1,j}$ and $W_{2i,j}$, respectively. Note that the mass at each time scale is only preserved in expectation. Note also that the way mass gets redistributed is independent from mass itself, since the $W_{i,j}$ are independent of mass.

Feldmann et al. [3] proposed to use conservative cascades, a special case of random cascades, as a model for IP traffic. In conservative cascades, the generator W takes on values in $(0, 1)$, has mean $1/2$ and is symmetric about its mean. Furthermore, mass is redistributed such that the total mass assigned to left and right child subintervals remains equal to that of the parent interval. Thus, the mass is preserved throughout the splitting process: if the mass of the i^{th} parent interval is Q, the mass of the left child interval will be $QW_{2i-1,j}$

and that of the corresponding right child interval will be $Q(1 - W_{2i-1,j})$. As a result, the mass at all stages will be (exactly) M, if the initial mass is M. This is the mass preservation property. Feldmann et al. supported the adoption of conservative cascade models in the networking context by observing that the transmitted traffic is constructed through fragmentation at successive network layers, and that the total number of bytes is roughly preserved during this fragmentation process. A typical example is the dynamics of a Web session, where user clicks results in requests, requests give rise to connections, connections are made up of flows, and flows consist of individual packets.

In this study, we use the conservative cascade model proposed in [3] as an example of a traffic model that is able to characterize multiscaling behavior. The model parameters are inferred using the procedure presented in [3], where each time scale is fitted to a truncated normal distribution with a mean that is always equal to $1/2$ but with a variance that is adjusted individually at each time scale.

3.4. THE L-SYSTEM MODEL

The basic idea behind L-Systems is to define complex objects by successively replacing parts of a simple object using a set rules. The L-System is a feedback machine that operates on strings of symbols. The set of symbols is called the alphabet. Starting from an initial state (called axiom), an L-System operates, at each iteration, by applying the set of production (or rewriting) rules simultaneously to all symbols of an input string to give an output string. The production rules can be stochastic. In stochastic L-Systems there may be several production rules for one symbol, and the specific rule is selected according to a probability distribution. Stochastic L-Systems are a method to construct recursively random sequences with multifractal behavior [23]. For a comprehensive introduction to L-Systems see [23].

In the proposed model in [10], we work on a closed alphabet of ordered arrival rates defined by

$$\vec{\lambda} = \{\lambda_1, \lambda_2, ..., \lambda_L\}, \lambda_i \in I\!R_0^+, i = 1, ..., L. \tag{6}$$

and with production rules that randomly generate two arrival rates from a previous one. The traffic process is constructed progressively, governed by an L-System machine, where each iteration produces a new time scale. Starting with the coarsest time scale, where traffic is characterized by a single arrival rate over a single time interval, each iteration generates a finer time scale by (i) division of each (parent) time interval in two new equal length (child) subintervals and (ii) association of arrival rates to each new subinterval according to the production rules of the stochastic L-System. We allow the grouping of time scales in time scale ranges and the definition of different sets of production rules for each time scale range. This is motivated by the

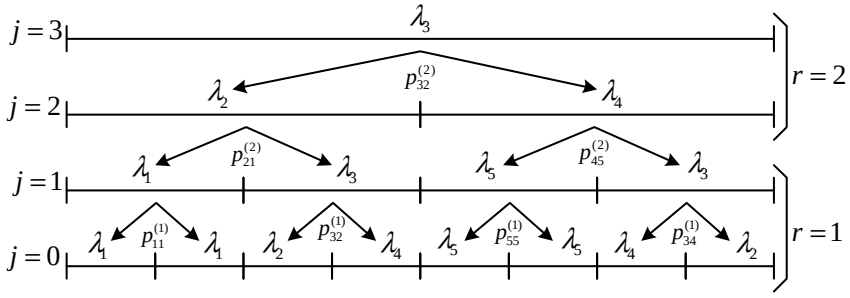

Figure 2. Construction of an L-System based traffic model.

fact that each set of production rules maps into a distinct scaling behavior [24]. The traffic process construction is illustrated in Figure 2.

To characterize the traffic process we define $X^{(i)}_{(j,r)}\epsilon\vec{\lambda}$ as the arrival rate at time interval i of time scale j and time scale range r. Let the number of scales be S and the number of ranges of scales be R. For convenience, we let j decrease from $j = S-1$ (at the coarsest time scale) to $j = 0$ (at the finest time scale). Also, we let r decrease from $r = R$ (the range of coarsest time scales) to $r = 1$ (the range of finest time scales). Thus, the number of time intervals at time scale j, which we will denote by N_j, is 2^{S-j-1}. Moreover, assuming a unitary width for the intervals of the finest time scale, $j = 0$, the width at scale j will be 2^j. To relate time scales and time scale ranges we define j_r as the coarsest scale j in range r. Thus, in Figure 2, $S = 4$, $R = 2$, $j_2 = 3$ and $j_1 = 1$.

The average arrival rate must be the same in all time scales, i.e., the mapping of arrival rates is such that the arrival rate averaged over the left and right child subintervals will be equal to the parent arrival rate. The traffic process generation can be described by axiom $X^{(1)}_{(S-1,R)}$, the arrival rate at the coarsest time scale, and production rules defined by

$$X^{(i)}_{(j,r)} = \lambda_l \xrightarrow{p^{(r)}_{lq}} \begin{cases} X^{(2i-1)}_{(j-1,r')} = \lambda_q \\ X^{(2i)}_{(j-1,r')} = 2\lambda_l - \lambda_q \end{cases} \tag{7}$$

where $\sum_{q=1}^{L} p^{(r)}_{lq} = 1, \forall l$. Thus, an arrival rate λ_l in interval i, scale j and range r produces, with probability $p^{(r)}_{lq}$, arrival rate λ_q at the left subinterval $2i-1$ and arrival rate $2\lambda_l - \lambda_q$ at the right subinterval $2i$, of next scale $j-1$ and range r'. The production rules can be totally described by R $L \times L$ matrices

$$\mathbf{P}^{(r)} = \left(p^{(r)}_{lq}\right), \quad l, q = 1, ..., L, \quad r = 1, ..., R. \tag{8}$$

The L-System construction defines, at scale j and range r, the sequence

$$Y_{(j,r)} = \{X^{(i)}_{(j,r)}, i = 1, ..., N_j\}. \quad (9)$$

3.4.1. *Fitting Procedure*

The fitting procedure determines the L-System parameters from real data observations. It starts by fixing a sampling interval Δ and considering the time series, $\{A_k, k = 1, 2, ..., K\}$, representing the total number of packet arrivals in each non-overlapping sampling interval. For convenience, the length of the time series K is considered a power of 2.

The first step of the inference procedure is the determination of the L-System alphabet and axiom. The alphabet of the L-System will consist in L equidistant arrival rate values, ranging from the minimum to the maximum values present in data. The axiom is inferred as the average arrival rate of $\{A_k\}$, rounded to the closest alphabet element, i.e.,

$$X^{(1)}_{(S-1,R)} = \Lambda\left((1/K\Delta)\sum\nolimits_{k=1}^{K} A_k\right) \quad (10)$$

where $\Lambda(x)$ represents a function that rounds x towards the nearest element of $\vec{\lambda}$.

The second step is the identification of time scale ranges, which is based on wavelet scaling analysis, more precisely on the (second-order) logscale diagram.

The final step is the inference of the L-System production rules, which are fully characterized by the $\mathbf{P}^{(s)}$ matrices. First, data is rounded in order to define sequence $Y_{(j,r)}$ at each time scale. This comprises obtaining the arrival rates $X^{(i)}_{(j,r)}$ from $\{A_k\}$ through

$$X^{(i)}_{(j,r)} = \Lambda\left((N_j/K\Delta)\sum\nolimits_{k=K(i-1)/N_j+1}^{Ki/N_j} A_k\right) \quad (11)$$

with $i = 1, ..., N_j$, for each j. Letting $c^{(r)}_{lq}$ represent the number of times that, at scale j and range r, the parent $X^{(i)}_{(j,r)} = \lambda_l$ produced the left child $X^{(2i-1)}_{(j-1,r')} = \lambda_q$, the production rule probabilities can be inferred as

$$p^{(r)}_{lq} = c^{(r)}_{lq} / \sum_{u=1}^{L} c^{(r)}_{lu}, \quad l = 1, ..., L, \quad r = 1, ..., R. \quad (12)$$

3.4.2. *Relation between L-Systems and Random Cascades*

In addition to the mass preservation property, also present in conservative cascades, L-Systems include an important feature that is not present in conservative cascades (nor in random cascades) and has a meaningful physical

explanation. In L-Systems the way mass is redistributed to left and right child subintervals can be made dependent on the mass of the parent interval, whereas in the conservative cascade construction this dependence is not allowed [10]. Lets take the example of [5] and consider the way Web requests are scheduled over time. The number of requests per time interval that a Web server can handle is limited, due to resource availability constraints. Therefore, the way user clicks produce requests depends on the overall number of clicks. If a Web server has more requests to process it will distribute them more sparsely over time. Thus, the mass itself (the number of clicks, in this case) influences the way distribution takes place. At a lower level, consider the way flows produce packets. If the network is more congested, the feedback control exercised by TCP imposes that packets will be more sparsely distributed over time. Thus, once more, the mass (the number of flows, in this case) influences the way distribution takes place. To conclude, the effect of resource availability limitations is an important factor with strong impact in the traffic generation process, which can be captured through an L-System based construction (but not through a conservative cascade).

4. Overview of the traffic traces

The selected traffic models and their associated fitting procedures were applied to three traces of IP traffic, one that was measured at the University of Aveiro (trace UA) and the others at the premises of a Portuguese ISP (traces ISP1 and ISP2). For all our measurements, the traffic analyzer was a 1.2 GHz AMD Athlon PC, with 1.5 Gbytes of RAM and running WinDump, and recorded the arrival instant and the IP header of each packet. The main characteristics of all selected traces are described in Table I.

Table I. Main characteristics of measured traces.

Trace name	Capture period	Trace size (pkts)	Mean rate (pkts/s)	Mean pkt size (byte)
UA	10.15am to 3.08pm, July 10^{th}2001	1 million	1138	557
ISP1	10.00am to 10.09am, July 27^{th}2001	7.9 million	24000	496
ISP2	09.00pm to 09.09pm, July 27^{th}2001	12 million	26000	470

The UA trace is representative of Internet access traffic produced within a University campus environment. The University of Aveiro is connected to the Internet through a 10 Mb/s ATM link and the measurements were carried out in a 100 Mb/s Ethernet link connecting the border router to the firewall, which

only transports Internet access traffic. The UA trace consists of 1 million packets captured on July 10^{th} 2001 from 10.15 AM to 3.08 PM. The mean arrival rate is 1138 pkts/sec and the sampling interval was 0.1 seconds.

The ISP1 and ISP2 traces are representative of IP traffic generated by a group of corporate clients. Trace ISP1 consists of 7.9 million packets captured on July 27^{th} 2001 from 10.00 AM to 10.09 AM and trace ISP2 consists of 12 million packets captured on July 27^{th} 2001 from 09.00 PM to 09.09 PM. The mean arrival rate is 24000 pkts/sec for trace ISP1 and 26000 pkts/sec for trace ISP2. The sampling interval was 0.01 seconds for both traces.

Both ISP traces comprise only 9 minutes of traffic due to buffer space limitations, since we were measuring at a high rate link carrying aggregated traffic from several users. However, we have analyzed several traces belonging to successive time periods, concluding that the general statistics were very similar within the same daily period. The sampling interval used in the ISP traces is 10 times lower than the one used for the UA trace, in order to have average values for the number of packets per interval with similar orders of magnitude in both cases.

The analysis of the autocovariance function (Figure 3) of trace UA lead us to suspect that it exhibits LRD behavior, due to the slow decay for large time lags. This is confirmed by the scaling analysis, since the y_j values in the logscale diagram are aligned between a medium octave (8) and octave 14, the highest one present in data. We will group the time scales of the original trace in time scale ranges having distinct scaling behaviors [24]. Classically, there will be scaling if the log-log plot of the q^{th} order energies (usual energy is $q = 2$) as a function of scale behaves linearly; if the plot is (globally) non-linear, different time scale ranges can be detected where linearity is observed (see [12], for example). The second-order logscale diagram for trace UA (Figure 4) identified 3 time scale ranges (within a total of 14 time scales) defined by $(1, 3), (3, 8)$ and $(8, 14)$.

Multifractality is assessed using the LMD referred in previous section, where non-trivial multifractal scaling behavior is detected when there is no horizontal alignment (within the limits of confidence intervals). Figures 5 and Figure 7 reveal trivial multiscaling behavior over small and large time scales whereas Figure 6 shows non-trivial multiscaling behavior over medium time scales. The behavior observed at small and large time scales suggests a trivial multiscaling model, such as a self-similar one, as a good approach at these time scale ranges, but at the medium time scales a good fitting will theoretically require a model having non-trivial multifractal scaling characteristics.

A similar analysis was made for traces ISP1 and ISP2, also revealing non-trivial multiscaling behavior in some time scale ranges. This evidence suggests the use of multifractal traffic models for an accurate fitting of their main statistics and queuing behavior. Among the selected models for this

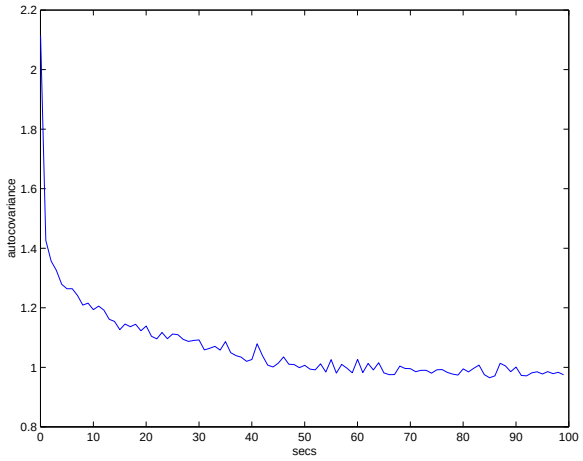

Figure 3. Autocovariance of packet counts, trace UA.

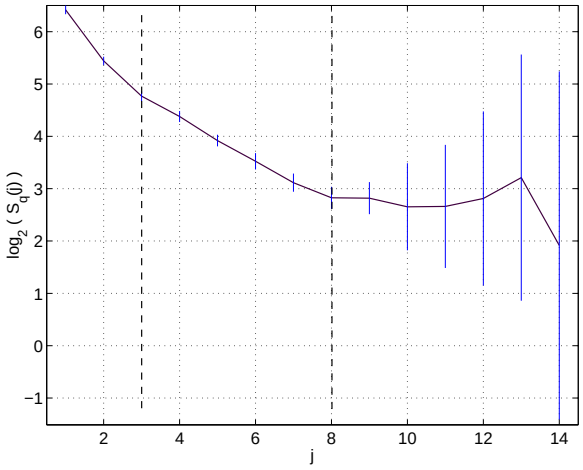

Figure 4. Second order Logscale Diagram, trace UA.

study, conservative cascades and L-systems possess intrinsically these characteristics, so we expect these models to give the best fitting results.

5. Numerical Results

We assess the suitability of the different traffic models and the accuracy of their fitting procedures using several criteria. Firstly, comparing both the probability and autocovariance functions of the packet counts (number of packet arrivals in the sampling interval) obtained with two traces: (i) the original data trace and (ii) a trace synthesized according to the fitted model.

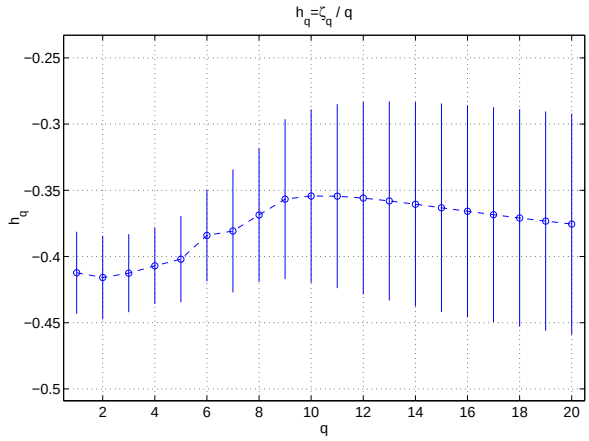

Figure 5. Linear Multiscale Diagram for time scale range $\{(j_1, j_2) = (1, 3)\}$, trace UA.

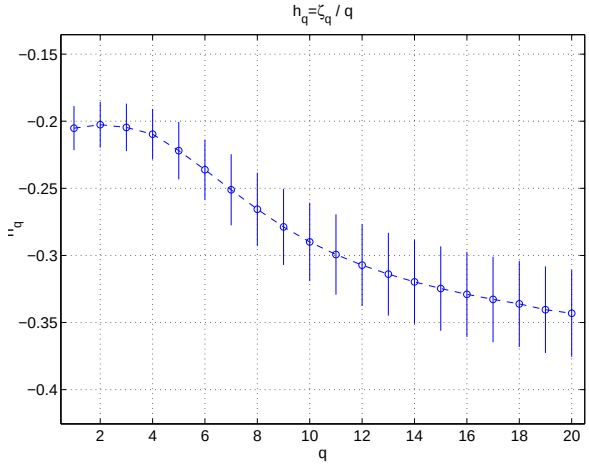

Figure 6. Linear Multiscale Diagram for time scale range $\{(j_1, j_2) = (3, 8)\}$, trace UA.

Secondly, analyzing the queuing behavior by comparing the packet loss ratio obtained, through trace-driven simulation, using those two traces. The simulations were carried out using a fixed packet length (equal to the mean packet length of the trace) because the selected models only deal with packet arrival instants.

Considering trace UA, we see that it was possible to achieve a good fitting of the first and second order statistics using a 12-MMPP: Figure 8 reveals a relative good agreement between the probability functions of the packet counts corresponding to the original and the fitted MMPP trace, and from Figure 9 we can see that there is also a good fitting of the autocovariance functions.

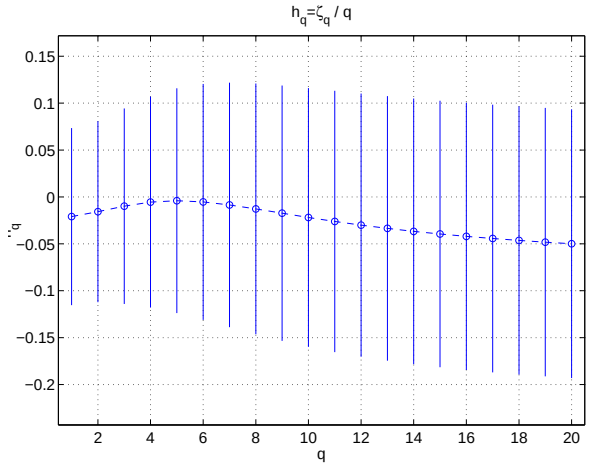

Figure 7. Linear Multiscale Diagram for time scale range $\{(j_1, j_2) = (8, 14)\}$, trace UA.

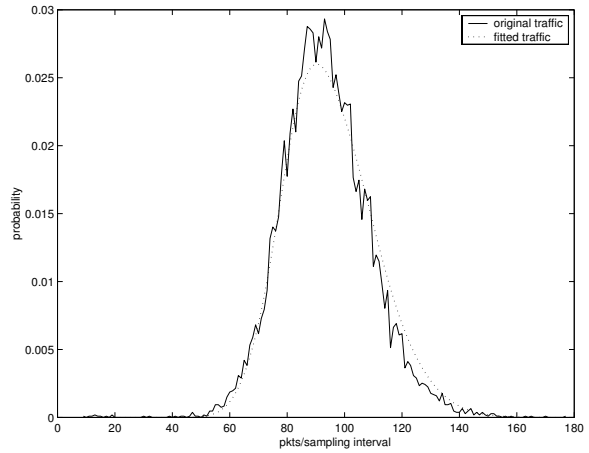

Figure 8. Probability function of packet counts, traces UA and MMPP fitted.

For the selected self-similar traffic model (fGn), only three parameters of the original trace were adjusted: the Hurst parameter, estimated from the second-order logscale diagram (the estimated value was 0.971), the variance and the average packet count. We can see that there is a reasonable fitting of the first order statistics (Figure 10), but the fitting of the autocovariance functions is much poorer (Figure 11).

Taking the conservative cascade model, we can see that there is a quite good fitting of the first (Figure 12) and second order statistics (Figure 13) of the packet counts. Note that the generator W in conservative cascades is assumed to be symmetric, which makes it difficult for conservative cascades

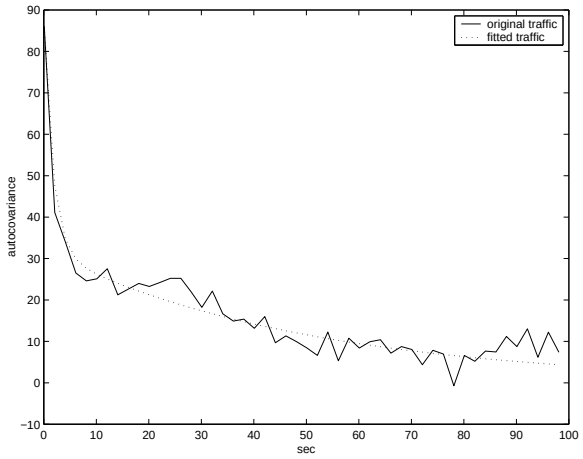

Figure 9. Autocovariance of packet counts, traces UA and MMPP fitted.

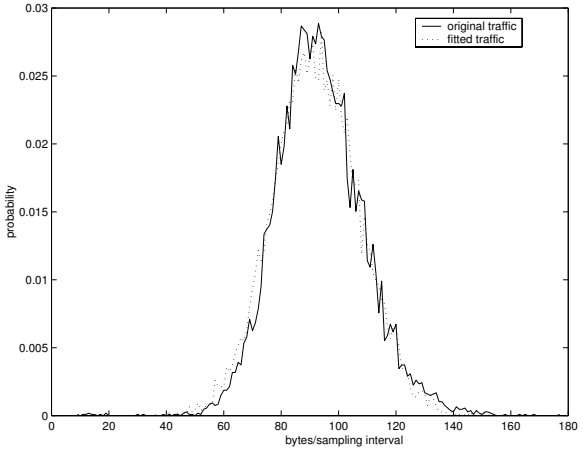

Figure 10. Probability function of packet counts, traces UA and fGn fitted.

to fit asymmetric probability distributions. However, in our case the empirical probability distribution is symmetric, so this drawback has no impact.

Finally, using the L-system model the UA trace was fitted to a stochastic L-System with an alphabet of $L = 169$ arrival rates, from the minimum to the maximum present in data, in steps of 10 pkts/sec (the minimum and maximum values were 90 pkts/sec and 1770 pkts/sec, respectively). The logscale diagram identified 3 time scale ranges (within a total of 14 time scales) defined by $j_1 = 3, j_2 = 8$ and $j_3 = 14$. The parameter estimation took less than 30 seconds, using a MATLAB implementation running in the PC described above. This shows that the fitting procedure is computationally very efficient. We can see from figures 14 and 15 that there is also a quite good fitting of

the first and second order statistics of the packet counts, so a good queuing behavior can be expected.

To assess queuing behavior, the buffer size was varied from 10 KBytes to 3.8 MBytes. The service rate was 547 KBytes/s (corresponding to an utilization of 0.95). The results for all considered traffic models are presented in Figure 16. We can observe that the fitting of the queuing behavior was very good for the MMPP, the conservative cascade and the L-system traffic models but significant differences occurred with the self-similar model. For this model, the result is not surprising since the fitting of the original autocovariance function was poor. This also shows that second-order statistics have a significant impact on network queuing performance, and that the fGn traffic model is too parsimonious, being unable to approximate complex traffic streams.

For the MMPP traffic model, there was a very good fitting of the probability and autocovariance functions of the packet counts. At least for this trace, the accuracy of this fitting seems to be sufficient to predict packet loss ratio performance. However, this can be attributed to the particular MMPP and the fitting procedure being considered, which results from the aggregation of distinct sub-processes that operate in different time scales. This in fact reproduces the way multifractal traffic processes are constructed and can help explaining the good performance in terms of queuing behavior.

For the conservative cascade model, the fitting of the autocovariance function was not perfect for all lag values, although the tail of the function was conveniently captured. Despite these differences in the fitting of the second order statistics, the queuing performance was accurately predicted, because the time-varying scaling characteristics of the input trace can be accounted for by the intrinsic multiscaling characteristics of the cascade model.

For the L-system model, the fitting of the autocovariance function at various lag values was even better than the other cases, so it does not surprise that the queuing performance is more accurately predicted, although the difference is not significant.

For the original traces ISP1 and ISP2 all four models were inferred using the inference procedures presented in Section 2 and we could see that the fitting of the first and second order statistics lead to similar conclusions.

In order to assess queuing behavior for the original trace ISP1, the buffer size was varied from 10 KBytes to 7.7 MBytes. The service rate was 12.30 MBytes/s (corresponding to an utilization of 0.98). The results for all selected traffic models are presented in Figure 17. In this case, we can observe that the fitting of the queuing behavior was good for both the MMPP and the conservative cascade traffic models, even slightly better for the L-system model, but significant differences occurred for the self-similar model. Note that, near the origin one can see a slight deviation between the PLR curves of the original and the fitted traffics, since for these low buffer size values even

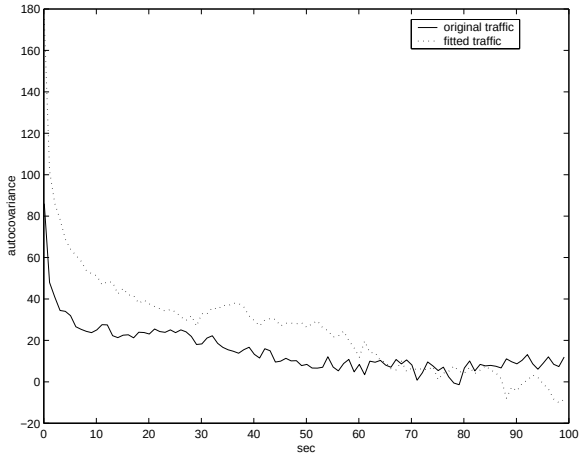

Figure 11. Autocovariance of packet counts, traces UA and fGn fitted.

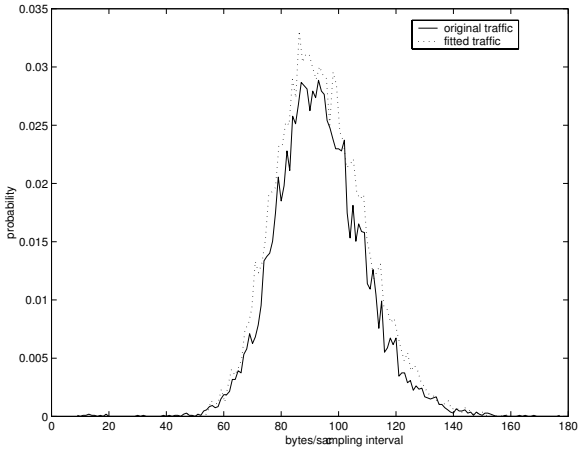

Figure 12. Probability function of packet counts, traces UA and conservative cascade fitted.

small differences in some peaks of the traces can produce significant changes in PLR values. So, these discrepancies are not too significant.

For the original trace ISP2, the queuing behavior was evaluated considering buffer sizes varying from 10 KBytes to 0.7 MBytes. The service rate was 12.99 MBytes/s (corresponding to an utilization of 0.95) and the results for all considered traffic models are presented in Figure 18. In this case the differences between the various models are not so pronounced. However, the self-similar model leads to a significant deviation in the knee of the queuing behavior curve.

Our results show that, in general, L-Systems achieves better performance than conservative cascades. Being two different approaches that are appropriate to capture the multiscaling/multifractal traffic characteristics, it is im-

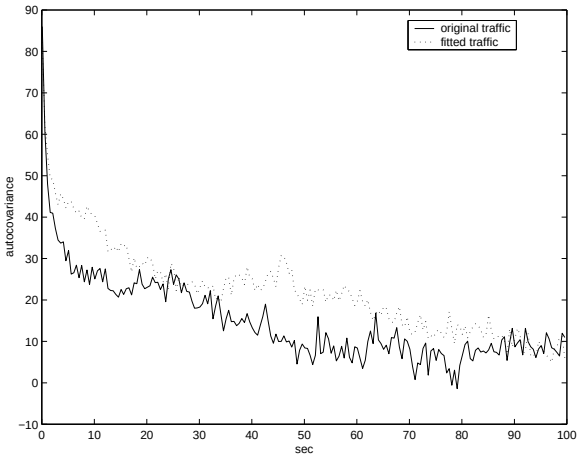

Figure 13. Autocovariance of packet counts, traces UA and conservative cascade fitted.

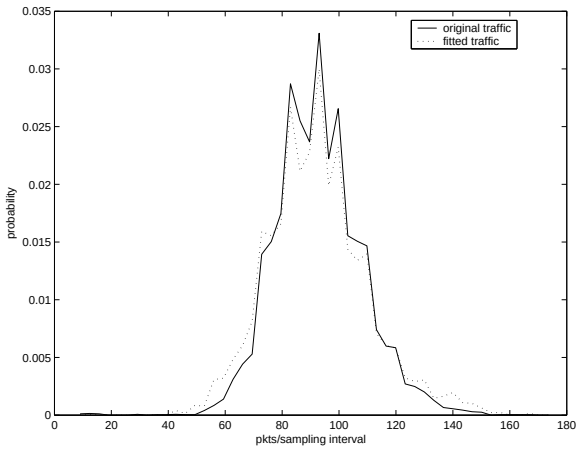

Figure 14. Probability function of packet counts, traces UA and L-system fitted.

portant to have a meaningful explanation. First, L-Systems allow the mass redistribution to depend on the mass itself, a feature that is clearly present in real observed data. Second, L-Systems provide a higher number of parameters, which are meaningful from the point of view of physical reality. Third, the generator in conservative cascades is assumed to be symmetric which restricts the fitting of the probability function.

In summary, the traffic multifractal behavior, clearly illustrated in figures 2, 3, 4 and 5, is captured by the Markovian, the conservative cascade and the L-system models. These traffic models are able to reproduce the different scaling behaviors observed, which is not the case with the fGn model.

We have analyzed several other traces measured at UA and at the portuguese ISP, obtaining similar conclusions. Thus, our main conclusion is that

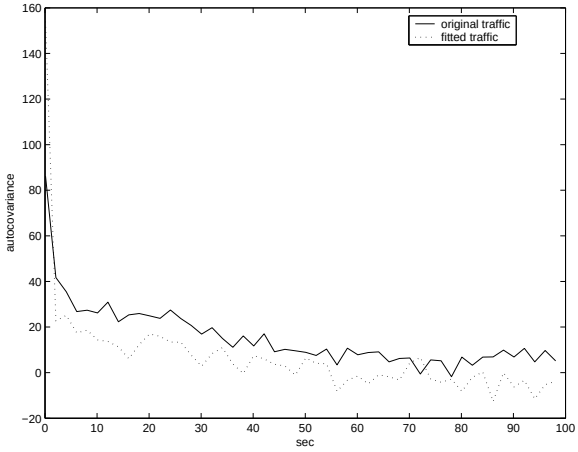

Figure 15. Autocovariance of packet counts, traces UA and L-system fitted.

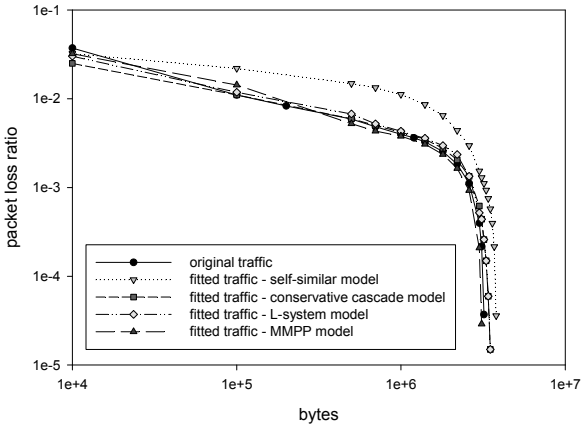

Figure 16. Packet loss ratio versus buffer size, trace UA.

MMPPs (with an appropriate fitting procedure), in addition to intrinsically multifractal models like conservative cascades and L-systems, can be used to model traffic exhibiting multifractal characteristics.

6. Conclusions

In this paper, we addressed the modeling of network traffic multifractality by evaluating the performance of four models, covering a wide range of traffic types, as mathematical descriptors of measured traffic traces showing multifractal behavior. We resorted to traffic traces measured both at University of Aveiro and at a Portuguese ISP. For the traffic models, we selected a Markov modulated Poisson process as an example of a Markovian model, the

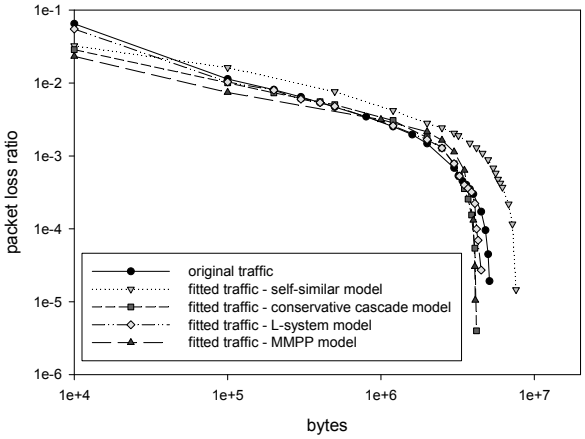

Figure 17. Packet loss ratio versus buffer size, trace ISP1.

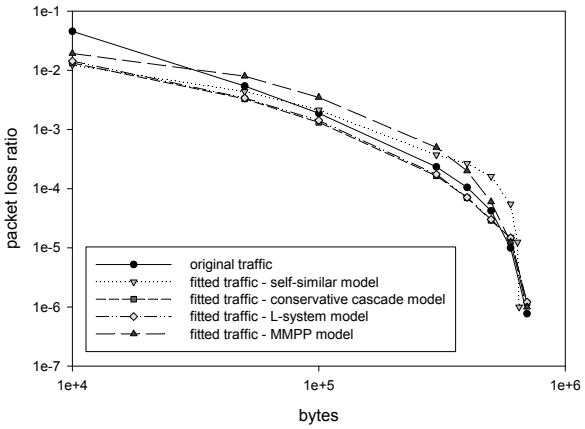

Figure 18. Packet loss ratio versus buffer size, trace ISP2.

well known fractional Gaussian noise model as an example of a self-similar process and the conservative cascade and L-system models as examples of multifractal models. These models were evaluated comparing the density function, the autocovariance and the loss rate queuing behavior of the measured traces and traces synthesized from the fitted models. Our results showed that the fractional Gaussian noise model was not able to perform a good fitting of the first and second order statistics as well as the loss rate queuing behavior, whereas the Markovian, the conservative cascade and the L-system models gave similar and very good results. The cascade and the L-system models are intrinsically multifractal, thus the obtained results were not surprising. The good performance of the Markovian model can be attributed to the fitting procedure that aggregates distinct sub-processes operating in different time scales, reproducing the way multifractal traffic processes are constructed, and

matches closely both the first and second order statistics of the traffic. The poor performance of the self-similar model can be explained mainly by its lack of parameters.

Acknowledgements

This research was supported in part by Fundação para a Ciência e a Tecnologia, the project POSI/42069/CPS/2001, and the grants BD/19781/99 and SFRH/BSAB/251/01.

References

1. B. Mandelbrot, "Intermittant turbulence in self-similar cascades: Divergence of high moments and dimensions of the carrier," *Journal of Fluid Mechanics*, vol. 62, pp. 331–358, 1974.
2. R. Riedi and J. Véhel, "Multifractal properties of TCP traffic: a numerical study," *Technical Report No 3129, INRIA Rocquencourt, France*, Feb 1997, Available at www.dsp.rice.edu/∼riedi.
3. A. Feldmann, A. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of internet WAN traffic," in *Proceedings of SIGCOMM*, 1998, pp. 42–55.
4. R. Riedi, M. Crouse, V. Ribeiro, and R. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 992–1018, April 1999.
5. A. Feldmann, A.C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," in *SIGCOMM*, 1999, pp. 301–313.
6. D. Veitch and P. Abry, "A wavelet based joint estimator for the parameters of LRD," *IEEE Transactions on Information Theory*, vol. 45, no. 3, Apr. 1999.
7. J. Gao and I. Rubin, "Multifractal analysis and modeling of long-range-dependent traffic," in *Proceedings ICC'99*, June 1999, pp. 382–386.
8. A. Gilbert, W. Willinger, and A. Feldmann, "Scaling analysis of conservative cascades, with applications to network traffic," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 971–992, April 1999.
9. A. Erramilli, O. Narayan, A. Neidhardt, and I. Saniee, "Performance impacts of multi-scaling in wide area TCP/IP traffic," in *Proceedings of INFOCOM'2000*, 2000.
10. P. Salvador, A. Nogueira, and R. Valadas, "Modeling multifractal traffic with stochastic L-Systems," in *Proceedings of GLOBECOM 2002*, November 2002.
11. J. Véhel, "Fractal and multifractal internet traffic," *Business Briefing: Global Optical Communications*, 2002.
12. P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation and synthesis of scaling data," *in Self-Similar Network Traffic Analysis and Performance Evaluation, K. Park and W. Willinger Eds*, 1999.
13. I. Norros, "A storage model with self-similar input," *Queueing Systems*, , no. 16, pp. 387–396, 1994.
14. Vern Paxson, "Fast approximation of self-similar network traffic," *Tech. Rep., Lawrence Berkeley Laboratory and EECS Division, University of California*, April 1995.

15. A. Andersen and B. 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, June 1998.
16. S. Kang and D. Sung, "Two-state MMPP modelling of ATM superposed traffic streams based on the characterisation of correlated interarrival times," *IEEE GLOBECOM'95*, pp. 1422–1426, Nov. 1995.
17. S. Li and C. Hwang, "On the convergence of traffic measurement and queuing analysis: A statistical-match and queuing (SMAQ) tool," *IEEE/ACM Transactions on Networking*, pp. 95–110, Feb. 1997.
18. K. Meier-Hellstern, "A fitting algorithm for Markov-modulated Poisson process having two arrival rates," *European Journal of Operational Research*, vol. 29, 1987.
19. C. Nunes and A. Pacheco, "Parametric estimation in MMPP(2) using time discretization," *Proceedings of the 2nd Internation Symposium on Semi-Markov Models: Theory and Applications*, Dec. 1998.
20. P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behaviour in an ATM multiplexer," *IEEE/ACM Transactions on Networking*, pp. 446–458, Aug. 1993.
21. P. Salvador, A. Pacheco, and R. Valadas, "Multiscale fitting procedure using Markov modulated Poisson processes," *Telecommunications Systems*, vol. 23, no. 1-2, pp. 123–148, June 2003.
22. P. Salvador and R. Valadas, "A fitting procedure for Markov modulated Poisson processes with an adaptive number of states," in *Proceedings of the 9th IFIP Working Conference on Performance Modelling and Evaluation of ATM & IP Networks*, June 2001.
23. H. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag, 1992.
24. P. Salvador, A. Nogueira, and R.Valadas, "Scaling behavior analysis of multifractal traffic models based on stochastic L-systems," Tech. Rep., University of Aveiro, 2002.