What is The Required Number of Users for The Generation of Aggregated H-ss Traffic?

C. Sansores-Perez, L. Rizo-Dominguez and J. Ramirez-Pacheco.

***Abstract*—It is well known that network traffic can be well modeled by the use of *self-similar* processes with parameter H. The use of this kind of traffic is important for the design and performance evaluation of high performance computer networks. Simulation plays a very important role in the context of performance analysis. In the context of simulation, however, the impact of the number of sources has not been sufficiently emphasized for the generation of synthetic *self-similar* traffic. In this paper we describe a simulation scenario suitable for the testing of performance issues under *self-similar* traffic. Our analysis was centered on the effect of traffic aggregation over the *self-similarity* degree, determining the necessary number of sources to approach the verified relation $H = (3 - \min \alpha)/2$. Besides, we highlighted the performance of several *Hurst* parameters estimators for this type of simulation scenarios, identifying the most suited ones.**

***Index Terms*—Self-Similar, Heavy-Tail Distributions, Estimators.**

I. Introduction

Simulation plays an important role for the design, performance evaluation and dimensioning of computer networks. Diverse network features can be studied with the aid of simulation scenarios. The effect of traffic behaviour on the *QoS* metrics such as *delay, delay jitter, packet loss,* etc. is such an example[13][14][15]. In this context, packet network traffic has shown to be of *self-similar* nature[2][3]. Thus, current simulation scenarios must take this behaviour into account. A well known and amply cited simulation scenario is given in [1][4], where *self-similar* traffic was generated by the transmission of files of size Z by an ensemble of $i = 32$ users. A particular feature of this scenario is that the distribution of files, Z_i, transmitted by user i is *heavy-tailed* with parameter α_i, giving rise to highly variable file sizes. In the limit as the number of users $i \to \infty$, the traffic in the network node is *self-similar* with $H = (3 - \min \alpha_i)/2$. Unfortunately, the *Hurst* parameter obtained in that paper is highly variable and overestimates when $\alpha > 1.6$ and underestimates when $\alpha < 1.6$. From the above it is noted that the scenario described in [1][4] can not be used for simulation studies where accurate tuning of the *Hurst* parameter is required. In addition, estimators used to test the presence of *self-similar* behaviour are not the most robust. In this paper, we propose some changes to the simulation scenario and determine the neccesary(and finite) number of users to approach the limit $H = (3 - \min \alpha_i)/2$ with minimum variation. It is shown that our simulation scenario can be used to effectively generate a *self-similar* process with *Hurst* parameter H, where H shows little variation. Thus, we propose a simulation scenario which can be used to study the behaviour of network algorithms under *self-similar* traffic, and where H can be effectively and accurately tuned by the values of α. Also, we complement the study presented in [1][4] by including in our study several estimators of *Hurst*-index. In this context, the paper is organized as follows, section II reviews fundamentals concepts related to *heavy-tail* distributions, *self-similar* processes and the methods for generating *self-similar* processes from *heavy-tailed* distributions. It also reviews the main estimators for both of them. Section III provides description of the simulation scenario and points out the differences with the one described in [1][4]. Section IV shows the results of the simulation and finally section V presents the concluding remarks.

II. Internet Traffic Models

A. Heavy-Tail Distributions

Heavy-tailed distributions are distribution functions whose tails $P(X > x)$ and $P(X \leq -x)$, for positive x, decrease slower than exponential rate[34]. The latter, e.g., normal and exponential distributions, are said to be of light tails while Pareto distribution are said to exhibit *heavy tails*. We will concentrate on right *heavy-tails*, i.e., on distributions whose survival function $P(X > x),\ x \geq 0$ behaves as a power law. Let X be a random variable defined on the probability space $\{\Omega, F, P\}$, we said that X has a heavy right tail if the following asymptotic behavior holds

$$P[X > x] \sim x^{-\alpha} L(x),\ x \to \infty, \tag{1}$$

where $L(x)$ is a slowly varying function, i.e., $\lim_{i\to\infty} L(ix)/L(i) = 1$ and $\alpha \in (0, 2)$ is the tail-index. When $\alpha > 2$, the random variable X has finite mean and variance; when $\alpha \in [1, 2)$, X has infinite variance but finite mean; finally when $\alpha \in (0, 1)$, X has infinite variance and infinite mean. Qualitatively, typical features of sample paths of *heavy-tailed* distributions are; most observations take small values, intermediate values occur frequently and extreme values occur rarely but with non-negligible probability. The thicker the tail of a distribution, F, the more probable the appearance of an extreme value is. As above-mentioned tail-index indicates the existence of moments in a random variable X. Let $E(X^\beta)$ be the moment of order β of X, if $\beta < \alpha$ then $E(X^\beta) < \infty$, otherwise if $\beta > \alpha$ then

Manuscript received December 9, 2009. The associate editor coordinating the review of this letter and approving it for publication was Dr. XXXXXXX. This work was partially sponsored by CONACyT

Luis Rizo and Julio Ramirez are with Research Center and Advanced Studies, CINVESTAV, Guadalajara, Jal., Mxico. (email: lrizo, cramirez@gdl.cinvestav.mx). C. Sansorez-Perez is with University of Caribe, Complex Systems Simulation Laboratory, Cancun, Quintana Roo., 77528, Mexico. (email: csansores@ucaribe.edu.mx). Digital Object Identifier

$E(X^{\beta}) = \infty$. For more information on properties, estimators and methods of generation of *heavy-tailed* distributions please refer to [5][6][7][24][25][26]. Heavy-tailed distributions has for instance be used to model network delay[36].

B. Estimators of tail-index

Several estimators of the tail-index α have been proposed, next subsections review *Hill*-based and QQ plots used for estimating α.

1) Standard Hill Estimator: Let $X_1, X_2, ..., X_n$ be a discrete time series with distribution $F_X(x)$. Now let $X_{(1)} > X_{(2)} > ... > X_{(n)}$ denote the ordered statistics of time series $X_1, .., X_n$. The *Hill* estimator of $\gamma = \alpha^{-1}$ based on $k+1$-upper ordered statistics, $1 < k \leq n$, is defined according to the following formula:

$$H_{k,n} = k^{-1} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}}. \qquad (2)$$

The parameter α is estimated by plotting k versus $H_{k,n}$ for $1 < k \leq n$ and looking for a stable region in the plot. The stable region must sit at height α. Usually the *Hill* estimator works better when the underlying *heavy-tailed* distribution is Pareto. When the distribution is not of Pareto-type the *Hill* estimator shows volatility, i.e., irregular erratic behavior.

2) Smooth Hill Estimator: The smooth *Hill* estimator, $smooHill$, is obtained by applying a smoothing technique to the standard *Hill* estimator in order to reduce the volatility in the standard *Hill* plot. Let again $X_{(1)} > X_{(2)} > ... > X_{(n)}$ be the ordered statistics, the $smooHill$ estimator is defined as

$$smoo\ \hat{\alpha}_{k,n,u} = \frac{1}{\frac{1}{(u-1k} \sum_{j=k+1}^{uk} H_{j,n}}, \qquad (3)$$

where $u \in \{2, 3\}$. Again a plot of k versus $smoo\ \hat{\alpha}_{k,n,u}$ should stabilize at a region $\hat{\alpha}$.

3) Alternative Hill Estimator: Another variant of the standard *Hill* estimator is the alternative *Hill* estimator, *altHill*, which changes the scale of the *Hill* estimator. The *altHill* estimator can be applied to the standard *Hill* estimator and the *smooHill* estimator. When applied to the *smooHill* estimator, it results in the *altsmooHill* estimator of $\hat{\alpha}$. The *altHill* estimator is defined as

$$H_{\lceil n^{\theta} \rceil, n} = \lceil n^{\theta} \rceil^{-1} \sum_{i=1}^{\lceil n^{\theta} \rceil} \log(\frac{X_{(i)}}{X_{(\lceil n^{\theta} \rceil + 1)}}), \qquad (4)$$

where $\lceil y \rceil$, is the smallest integer greater of equal to $y \geq 0$. For the estimation of $\hat{\alpha}$ we plot θ versus $H_{\lceil n^{\theta} \rceil, n}$ for $0 \leq \theta \leq 1$. The stable region in the plot should be the estimated value of $\hat{\alpha}$.

4) QQ-Plot: Let $X = (X_1, X_2, ..., X_n)$ be $i.i.d$ observations with common distribution F. Now let $X_{(1)}, X_{(2)}, ...X_{(n)}$ be the upper order statistics of X, i.e., $X_{(i)} > X_{(j)}$ iff $i < j$. Pick k upper order statistics and neglect rest $k+1$. The distribution of the k exceedances, $X_{(1)}, .., X_{(k)}$ should be Pareto if F is *heavy-tailed*. Taking the logarithm of the k exceedances makes its distribution approximately exponential, thus the plot of the empirical quantiles of the exceedances against the theoretical quantiles of the exponential distribution should yield a straight line with slope α^{-1}. More formally the plot of

$$\{(-\log(1 - \frac{j}{k+1}), \log X_{(k-j+1)}, 1 \leq j \leq k\}, \qquad (5)$$

should yield approximately a α^{-1} slope straight line if the distribution of $X_1, X_2, ..., X_n$ satisfies the asymptotic behavior of (1). The slope of the line is computed by least squares regression through the points in (5) and is called the QQ estimator, i.e.,

$$\widehat{\alpha^{-1}}_{k,n} = \frac{\sum_{i=1}^{k}(\nu_{i,k})\xi_{i,k} - \sum_{i=1}^{k}(\nu_{i,k})H_{k,n}}{k(\frac{1}{k}\sum_{i=1}^{k}(\nu_{i,k})^2 - (\frac{1}{k}\sum_{i=1}^{k}\nu_{i,k})^2)}, \qquad (6)$$

where $\nu_{i,k} = -\log(\frac{i}{k+1})$ and $\xi_{i,k} = \log(\frac{X_{(i)}}{X_{(k+1)}})$. There are two different versions of the QQ-plot, namely the dynamic and static QQ-plot. The dynamic QQ-plot is similar to the *Hill* plot and is obtained by plotting $\{(k, 1/\widehat{\alpha^{-1}}_{k,n}), 1 \leq k \leq n\}$ and finding a stable region in the plot. The static plot is obtained by choosing an appropiate value of k, plotting the points in (5) and finding a region where the plots looks linear, then in the linear region apply (6) which should yield the value of α^{-1}.

C. Self-similarity and long-memory

Processes with some form of *scaling* behaviour can be defined as stochastic signals possesing invariance properties on all or a set of scales(i.e., no characteristic scale can be identified). Examples of such processes include *self-similar*[29], *long-memory*, *fractal* and *multifractal* processes[30][27][28][31]. The paper deals with *self-similar* and *long-memory* processes, the most known of them. Strict *self-similar* signals(H-ss), $X = \{X_t, t \in \mathbb{R}\}$, are defined as those for which appropiate changes of scale of time and space do not vary its statistical properties, i.e., processes for which $X_{at} = a^H X_t$, for any $t \in \mathbb{R}$, $a, H > 0$, where the equality is in terms of finite-dimensional distributions. Weak *self-similarity*, a more often used version, is defined as processes for which $\mathbb{E}X_{at}X_{as} = a^{2H}\mathbb{E}X_tX_s$, for any $t, s \in, \mathbb{R}$, $a, H > 0$. Note that strict self-similarity implies nonstationarity, *long-memory* processes on the other hand is often defined for stationary processes. Long-memory property of finite-variance stationary signals $Y = \{Y_t, t \in \mathbb{R}\}$ is possesed if $\mathbb{E}Y_tY_{t+\tau} \sim c_{\gamma} \mid \tau \mid^{\beta-1}$(equivalently as its PSD $f(\nu) \sim c_f \mid \nu \mid^{-\beta}$) as $\tau \to \infty$(as $\nu \to 0$). Indeed, a strong relationship between these two processes exists and a given *self-similar* process(H-ss) with stationary increments(Hsssi) possess *long-memory* in its first increment process, i.e., $\mathbb{E}Y_tY_{t+\tau} \sim c\tau^{\beta-1}$ provided $Y = \Delta^1 X(t;1) = X(t+1) - X(t)$ and X belongs to the space of finite variance H-sssi processes. The above for example holds true for the unique Gaussian *H-sssi* process, namely, fractional Brownian motion(fBm) with $0 < H < 1$. Many estimators of *Hurst*-index have been proposed[20][37][38], *R/S* statistic, variance based(aggregated, differenced, detrended), periodogram-based(GPH, cumulated, whittle), wavelet based estimators(abry, delbeke)[23][33], etc.

D. *Estimators of the* self-similarity *parameter*

1) R/S Statistic:: The *R/S Statistic*[6][8][9][10][38] developed by E. *Hurst* when studying Nile river is defined for a process $Y(t)$ in the interval $(\tau_i, \tau_i + n)$ as

$$\frac{R}{S}(\tau_i, n) := \frac{\max\big(W(\tau_i, n)\big) - \min\big(W(\tau_i, n)\big)}{S(\tau_i, n)}, \qquad (7)$$

where $W(\tau_i, n) = Y(\tau_i + u) - Y(\tau_i) - uE(\tau_i, n)$ and $E(\tau_i, n)$ and $S(\tau_i, n)$ denote the mean and standard deviation in the interval $(\tau_i, \tau_i + n)$. *Hurst* found that for long-memory records, (7) behaves as $E\{\frac{R}{S}(\tau_i, n)\} \sim n^H, H > 0.5$, in constrast, short-memory processes follow $E\{\frac{R}{S}(\tau_i, n)\} \sim n^{0.5}$. A log-log plot of the mean values of the *R/S* statistic values versus n is an estimator of H.

2) Block averaged methods: Variance and Absolute Moment: Consider the aggregated series $\Gamma_m(\{x_i\}) = X_i^{(m)}$ of a length N time series[11][16]. The sample variance of the block averaged process $\text{Var}\big(\Gamma_m(\{x_i\})\big)$ for *long-memory* series behaves asymptotically as $\text{Var}\big(\Gamma_m(\{x_i\})\big) \sim cm^{-\beta}$, where c is a constant and $\beta = 2 - 2H$. From this result, a log-log plot of $\text{Var}\big(\Gamma_m(\{x_i\})\big)$ versus m, for different values of m, and such that $m_{i+1}/m_i = C \in \mathbb{R}+$ is an estimator of H. The absolute moment of $X_i^{(m)}$, $AM^{(m)}$, behaves asymptotically as $AM^{(m)} \sim m^{-\beta/2}$, thus a log-log plot of $AM^{(m)}$ versus m results in a line with slope $-\beta/2 = H - 1$ from which H is inferred. First method is called the variance method and the latter the absolute moment one.

3) Periodogram based methods: Periodogram and Whittle: The periodogram, $I(\upsilon) = 1/(2\pi N) \mid \sum_{j=1}^{N} X_j e^{ij\upsilon} \mid^2$ for the series $\{X_j\}$ is also an estimator of H. The periodogram for a *long-memory* time series behaves as $I(\upsilon) \sim\mid \upsilon \mid^{1-2H}$ for $\upsilon \to 0$, therefore a log-log plot of $I(\upsilon)$ versus υ is used to obtain H. The *Whittle method* [37][32][21] [38][6] is a non-graphical MLE estimator strongly related to the periodogram defined by the following relation $Q(\eta) := \int_{-\pi}^{\pi} \big(I(\upsilon)/f(\upsilon;\eta)\big)d\upsilon + \int_{-\pi}^{\pi} \log(f(\upsilon;\eta))d\upsilon$, where η is a vector of unknown parameters and $f(\upsilon;\eta)$ is the spectral density at frequency υ of the studied function, the value of vector η that minimizes the function Q is considered the *Whittle Estimator*. A discretized version of $Q(\eta)$ is obtained as $Q^*(\eta) = \sum_{j=1}^{(N-1)/2} I(\upsilon)d\upsilon/f^*(\upsilon_j;\eta)$ where N is the series length. The *Whittle MLE* specifies the functional form of the spectral density at all frequencies and the *Local Whittle*[37][32] [6] assumes only the functional form when ν approaches zero, namely $f(\upsilon) \sim G(H)\,|\upsilon|^{1-2H}$ as $\upsilon \to 0$ and from $Q^*(\eta)$ the task is reduced to minimize the function

$$R(H) = \log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{I(\upsilon_j)}{\upsilon_j^{1-2H}}\right) - (2H-1)\frac{1}{M}\sum_{j=1}^{M}\log\upsilon_j \qquad (8)$$

Its computation involves the introduction of the parameter M which is an integer less than $\frac{N}{2}$, and satisfying$\frac{1}{M} + \frac{M}{N} \longrightarrow 0$ as $N \longrightarrow \infty$.

4) Wavelet based methods: Let $d_x(i,j)$ denote the wavelet coefficients of a particular finite length sequence $\{x_i\}$, it is known that for *long-memory* processes the variance at level i of the coefficients is given by $\text{Var}(d_x(i,.)) = \frac{\sigma^2}{2} V_\psi(H)(2^j)^{2H+1}$, where $V_\psi(H)$ depends on the particular wavelet and the *Hurst*-index and is defined by:

$$V_\psi(H) = -\int_{-\infty}^{\infty} \gamma_\psi(\tau) \mid \tau \mid^{2H} \, d\tau \qquad (9)$$

taking the logarithm at $\text{Var}(d_x(i,.))$ should result in $\log(\text{Var}(d_x(i,.))) = (2H+1)j + K$, where K is a constant. Abry and Veitch have suggested an *Hurst*-index estimator based on this behaviour using Daubechies wavelets[33][19][23]. First a time average μ_i of $d_x(i.j)$ is computed at a given scale, where μ_i is defined as $\mu_i = (n_i)^{-1}\sum_{j=1}^{n_i} d_x^2(i.j)$, where n_i is the wavelet coefficient number at scale i and n the time series points. The estimated *Hurst*-index is then obtained from the slope of a linear regression method for $\log_2(\mu_i) = \log_2(\frac{1}{n_i}\sum_{j=1}^{n_i} d_x^2(i,j))$, where $i = 1, 2, \ldots, [\log_2(n)]$.

5) Sources of inaccuracies: Algorithms' accuracy are often affected by some parameters such as cut-off selection, number of aggregation levels and the minimum number of points in block size in regression based methods. Also other parameters include number of frequencies for periodogram methods, begining and ending octave, etc. These parameters are sources of inaccuracies and bias the esttimates. They must be selected carefully.

E. *Self-similarity through high-variability*

Self-similar traffic can be generated using the Lamperti transformation based on a stationary stochastic process or can be generated by the superposition of an infinite number of users which are superposed in a node. In this paper we concentrate in the generation of *self-similar* traffic based on *heavy-tailed* distributions. Let X_i be a random variable with a *heavy-tailed* distribution. Suppose the random variable can represent the file size of traffic source i or the period of transmission between succesive packets. As the number of users $i \to \infty$, then, the traffic aggregated(or superposed at a node) is *self-similar* with *self-similarity* parameter $H = (3 - \min\alpha_i)/2$[17][18][35]. We used the high variability of interdepartures times for the generation of *self-similar* traffic.

III. Simulation scenario

This section presents the proposed simulation scenario which generates *self-similar* traffic with parameter H. This scenario turns to be an appropiate model for simulations where the degree of traffic *self-similarity* needs to be finely and precisely adjusted. The simulation scenario is shown by the network model of figure 1. In this network, *self-similar* traffic is generated by an ON/OFF model, where the ON and/or OFF are heavy-tailed [17][18]. Although the required number of independent users should be infinite along this model, in practice, this condition is not feasible giving rise to *self-similar* traffic generators using diverse number of sources. This diversity has an important impact over traffic *self-similarity* generation and measurement. For instance, in the works [1][4], oriented to study the relationship between file sizes and *self-similarity* phenomena, the numbers of sources was set up to $i = 32$ and its variation seemed not to be significant to their results. In contrast, in the experiments we performed, a significant relationship between this parameter and the generated *self-similar* traffic was found and thus, the network configuration of figure 1 was proposed in order to calibrate this feature. As shown in the figure 1, the network consists of i nodes(or sources) and l output links in a packet switched configuration. The parameter i is customizable and represents statistically independent UDP sources, i.e, $S_i, S_2, \ldots, S_i$ are i.i.d. N_1 and N_2 represents routers through which packets from sources S_i are processed and forwarded to the destination sources R_i. In our configuration N_1 represents the node over which traffic is superposed and thus represents our measurement point. Queue length of node N_1 was set up to 1000 packets with buffer size of 312.5kB and its output link bandwidth was set up to 32.768Mb and latency of 30ms. Each link from S_i to N_1 and from R_i to N_2 has a bandwidth of 8.2Mb and a latency of 20ms. In order to obtain *self-similar* traffic in N_1, the traffic sources S_i have a Pareto random variable generator for the inder-departure time t_i. Recall that if t_i is a Pareto random variable, its *CDF* is given by:

$$\mathrm{P}(t_i \leq t) = 1 - (\frac{t_{min}}{t})^{\alpha}, \qquad (10)$$

where the minimum value of t_i is t_{min} and α is the tail-index. The Pareto random variable has infinite variance when $1 < \alpha < 2$. In this case tha mean is finite. Then, in order to keep $E(t) < \infty$ for all sources, the simulation was performed for $\alpha \in \{1.1, \ldots, 1.9\}$. Eventhough the mean of the inter-departure time rely upon the value of α, in our configuration it is constant, i.e., $E(t) = 500\mu$s for all given values of α. Likewise to normalize the data mean rate for all sources, t_{min} was tuned to each values of α, with an initial value of 0.041ms and fixed packet size of 320 bytes. The network configuration just reviewed was used in all the experiments of the paper. We used the well-known network simulator *ns-2* in a 2x2.8GHz Quad-Core Intel Xeon Macintosh platform. All results were obtained from several hundreds of runs executed for 300 simulated seconds and varying number of sources.

IV. Results

A. Generation of Pareto series

In order to check the correctness of our simulation scenario, we first test the appropiate generation of *heavy-tailed* traces in `ns-2` from which users send the packets (recall that Pareto series simulate inter-departure times of packets). Figure 2 shows typical packet inter-departure time series trace in `ns-2` with $\alpha \in \{1.2, 1.8\}$. Top plot correspond to Pareto series with $\alpha = 1.2$ while bottom plot to Pareto with $\alpha = 1.8$. Note that traces behave qualitatively as *heavy-tailed* process, i.e extreme values occur frequently. As above mentioned, the number of extremes values(i.e., of silent periods) in `ns-2` generated Pareto series occur with a non-negligible probability. Note that the 'usual' values are below `100ms` and that the lower the value of α is the greater this value. A *Hill*-plot and a *CCDF* plot will confirm the appropiate generation of Pareto series in our simulation scenario. Figure 3 shows the *Hill* plots associated to traces of figure 2. Top plot corresponds to Pareto with $\alpha = 1.2$ while the bottom to $\alpha = 1.8$. Note that *Hill*-plots stabilize in a region and this region corresponds to the true value of α. CCDF plots are also helpful for testing if a given model follows a particular probability distribution $F_X(x)$. CCDF plots, therefore can be used to test if a series follows a Pareto distribution. Figure 4 shows the CCDF plots corresponding to traces presented in figure 2. Again as before top plot corresponds to `ns-2` generated Pareto time series with $\alpha = 1.2$ while bottom plot to $\alpha = 1.8$. Note from the figure that both time series follow accurately the reference line corresponding to an exact Pareto time series with known α. From the above it is seen that the generation of Pareto time series with `ns-2` is accurate since *Hill* and CCDF plots estimate correctly the given α. Similar results were obtained

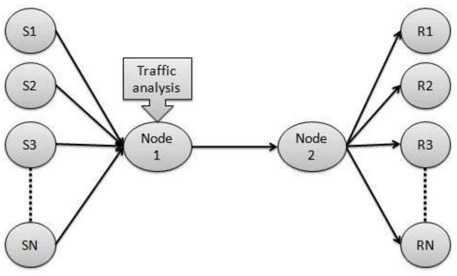

Fig. 1. Simulation scenario

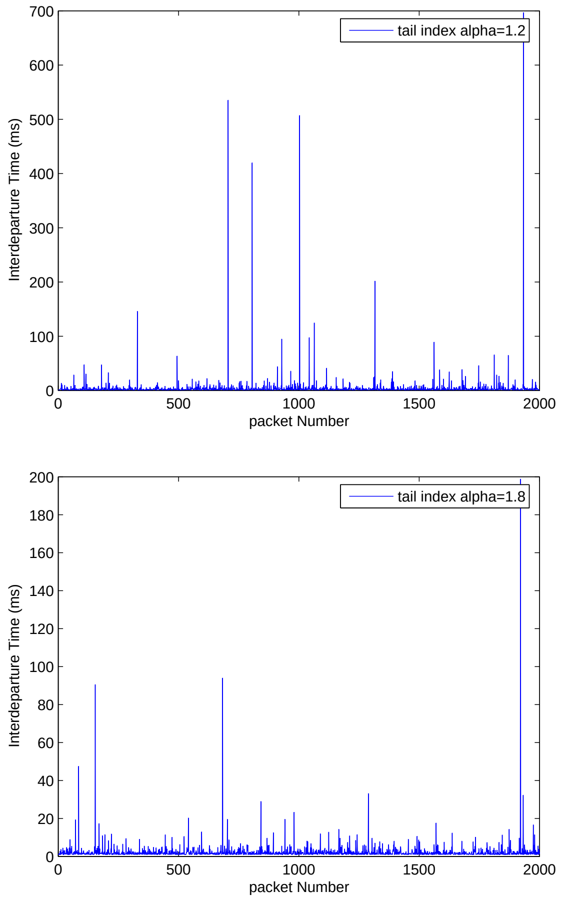

Fig. 2. Typical Pareto series generated with ns-2.

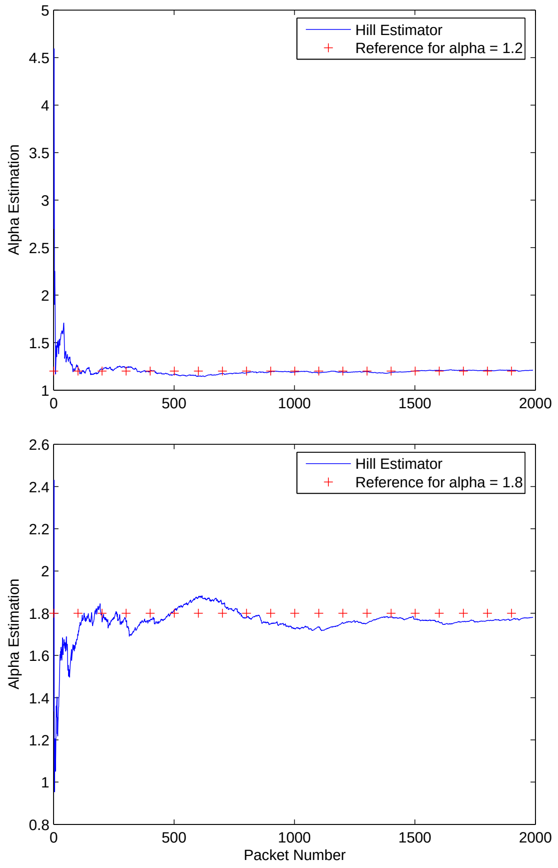

Fig. 3. Typical Hill plots for ns-2 generated Pareto series

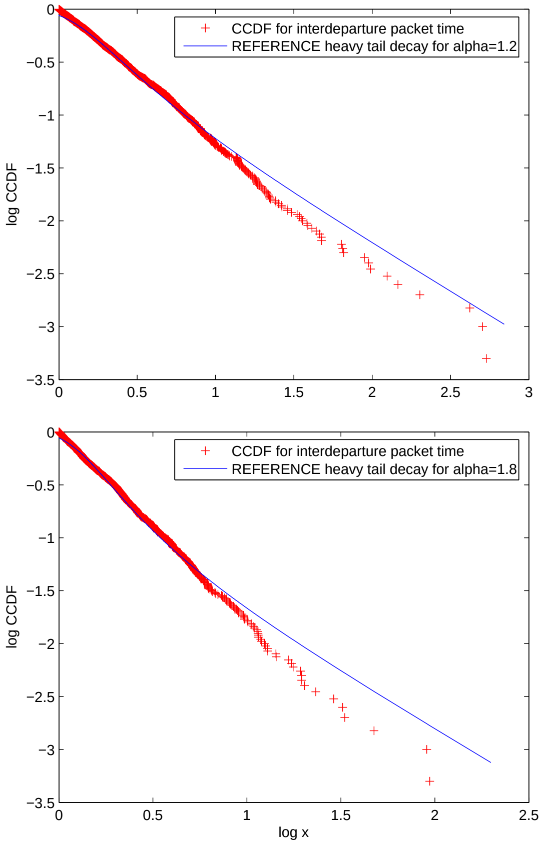

Fig. 4. Typical CCDF plots for ns-2 generated Pareto series

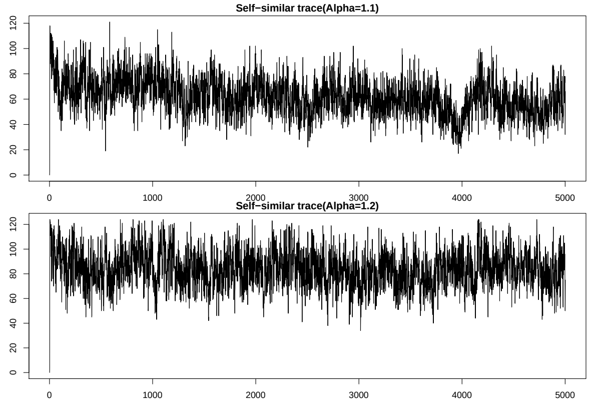

Fig. 5. Self-Similar Series obtained from Pareto distributions

when analyzing other time series corresponding to a given source or sources.

B. Generation of Self-Similar series

In this subsection we experimentally verify that aggregate traffic from N sources, where each source send packets with inter-departure according to a Pareto series, follows and can be modelled by a *self-similar* process of parameter H. Recall that as $N \to \infty$, $H = (3 - \alpha)/2$. This subsection only test that ggregate traffic is indeed a *self-similar* process. Figure 5 shows typical traffic traces obtained in a network node in our simulation scenario. Note that the series obtained behaves in accordance with a *self-similar* trace. The bottom trace corresponds to a trace with $H = 0.9$ and the top plot to a trace with $H = 0.95$.

C. Self-similarity and α relation

Figure 6 shows the simulation results when considering 30 traffic sources. The same number of traffic sources was used in [1][4]. As can be noted from the figure, the same kind of behaviour is obtained as those of [1][4]. Note that no estimator can follow the reference line $H = (3 - \min \alpha_i)/2$. In fact it is noted that the estimates $\hat{H} \sim H_{ref} + k$, where k is a constant. From this it can be said that no simulation scenario, neither [1][4] nor the proposed by us is capable of finely and accurately generating *self-similar* traffic for performance purposes when $N = 32$. Precisely generating *self-similar* traffic is important for testing the behaviour of algorithms or novel protocols and checking its behaviour under varying degrees of correlation or persistence. In fact, this degree can be accurately varied based on the tail-index of the traffic source. Figure 7 shows the variation of the *Hurst*-index when estimated with five different estimators. Note that variance-type method presents high variability. *R/S* statistic presents low variability but unfortunately its bias is high. From the two figures is concluded that when using

30 traffic sources in the simulation scenario shown above, *self-similar* traffic is effectively generated but the *Hurst*-index of this generated traffic presents high bias and variability. Figure 8 shows the simulation results when using 50 traffic sources. Note that significant improvements in the bias are obtained. In fact, Whittle, wavelet and *R/S* statistic behave reasonable well. Periodogram and variance method present high bias and variability. Recall that the most robust estimators are those based on wavelets and the MLE estimation ones(Whittle). Our paper takes these estimators into account and our conclusions area based on the results obtained from these. As above, figure 9 shows the standard deviation of the estimations of the *Hurst*-index but when using 50 traffic sources. Note that although *R/S* statistic shows low bias, the variance is high and thus is not suggested for deciding which the number of required traffic sources is. Whittle and wavelet are the most robust among all the estimators[37][38][12][23] studied and thus can be used for the task of deciding the required number of traffic sources for obtaining *self-similar* traffic with low-bias and variance. We also performed the same kind of analysis to 70 traffic sources obtaining similar results as those for 50, this behavior can be observed in Figure 10. From the above figures, we can conclude that the required number of users neccesary to generate accurate *self-similar* traffic is at least 50 traffic sources. Also, variance and *R/S* statistic methods can not be used for such a task. In our work, Whittle and wavelet based methods were used to decide the required number of users for the accurate generation of *self-similar* signals. References [1][4] showed the results for 30 traffic sources and the methods used to test the presence were variance and *R/S* statistic. We suggest that the relationship between *self-similarity* parameter and *QoS* parameter presented in that paper must be re-evaluated.

V. Concluding Remarks

This paper described background information on *self-similar* processes and *heavy-tailed* distributions. It reviewed the main estimators both for *self-similar* and *heavy-tailed* stochastic processes. It showed the appropiateness of the simulation scenario first in the generation of Pareto and then aggregated *self-similar* traces. The correct generation of Pareto was tested with Hill based estimators while the generation of correct aggregated

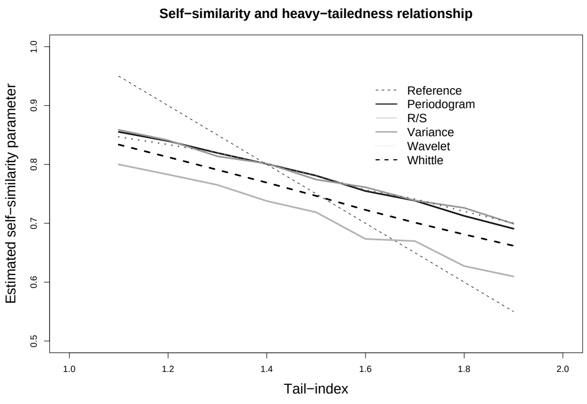

Fig. 6. Estimated *Hurst*-index for 32 users

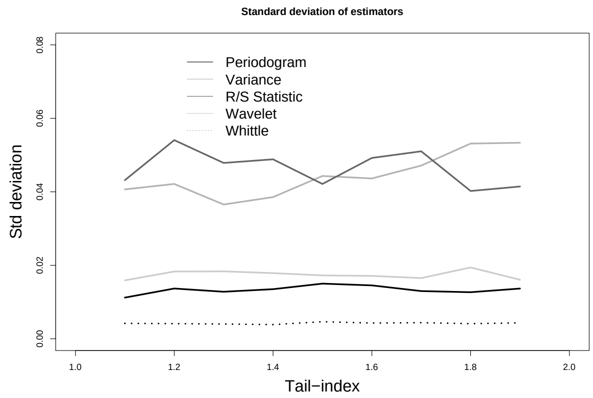

Fig. 7. Variation of *Hurst*-index estimation for 32 users

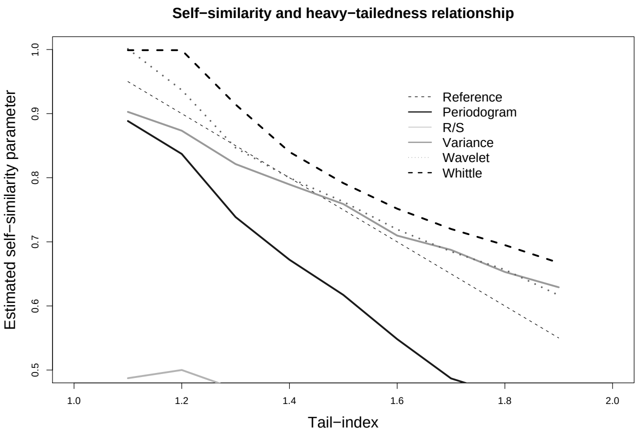

Fig. 8. Estimated *Hurst*-index for 50 users

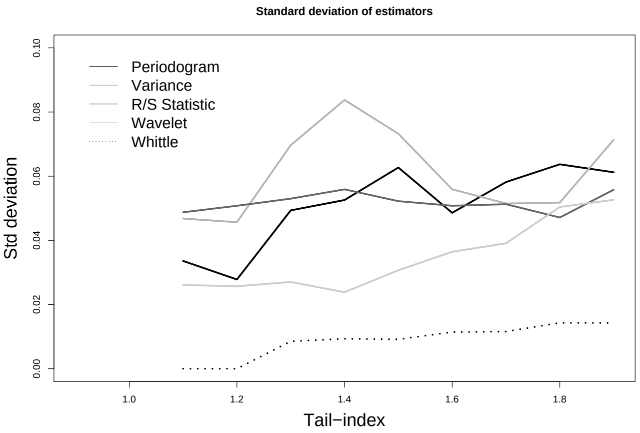

Fig. 9. Variation of *Hurst*-index estimation for 50 users

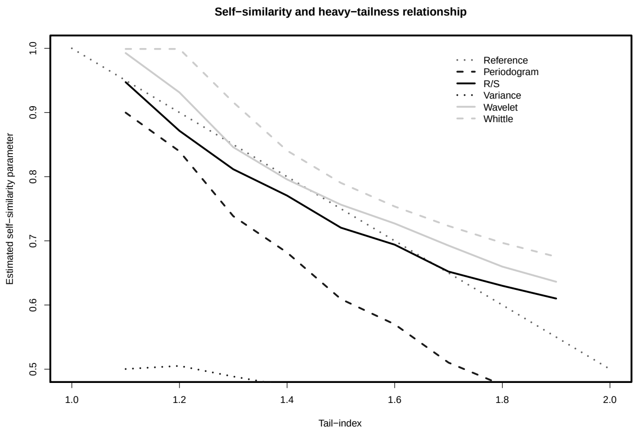

Fig. 10. Estimated *Hurst*-index for 70 users

self-similar traces was tested with time-domain, frequency-domain and time-scale estimators. It also detailed the simulation scenario for generating *self-similar* traffic from *heavy-tailed* sources. This scenario diverges from previous reported results in two aspects: $i)$ the number of sources and $ii)$ the increment in the number of *Hurst* parameter estimators evaluated. Based on extensive simulation results we found that the number of independent users impact the accuracy of the *Hurst*-index and conclude that the required number of independent traffic sources must be at least 50. By incrementing the number of sources we obtain higher accuracy but with higher computational cost. Also, according to variation of *Hurst*-index estimation, the Whittle and wavelet methods were the most suited for this type of simulation scenarios. As further work we propose the analysis of the relationship of *self-similarity* parameter and *QoS* performance under the scenario proposed. Also, it would be interesting to include several *self-similar* traffic flows in the topology to study its relation on adjacent nodes and also to establish the mathematical relationship among *Hurst*-indexes.

Acknowledgements: This work has been supported by Project QROO-2008-01-92231 funded by Consejo Nacional de Ciencia y Tecnolog?a (CONACYT) and Project PROMEP/103.5/08/2968 funded by PROMEP program of Mexico. Authors would like to thank Jos? Antonio Careaga for his help in this work.

REFERENCES

[1] K. Park, G. Kim and M. Crovella, On the Effect of Traffic Self-Similarity on Network Performance, *Proc. SPIE International Conference on Performance and Control of Network Systems*, 1997.

[2] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the Self-Similar Nature of Ethernet Traffic(Extended Version), *IEEE/ACM Transaction on Networking,* 2(1), 1994, pp. 1–15.

[3] W. Willinger, M.S. Taqqu, W.E. Leland and D.V. Wilson, Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements, *Statistical Science*, 10(1), 1995, pp. 67–85.

[4] K. Park, G. Kim and M. Crovella, On the Relationship Between File Sizes, Transport Protocols, and Self-Similar Network Traffic, *Proc. of the Fourth International Conference on Network Protocols (ICNP'96),* 1996, pp. 171–180.

[5] M. Crovella and M. Taqqu, Estimating the Heavy-Tailed Index from Scaling Properties, *Methodology and Computing in Applied Probability* 1(1), 1999, pp. 55–79.

[6] R. Adler, R. Feldman and M. Taqqu, *A Practical Guide to Heavy-Tails: Statistical Techniques and Applications*, Birkhauser, Boston, 1998.

[7] J. Berlaint, Y. Goegebeuer, J Segers and J. Teugels, *Statistics of Extremes: Theory and Applications*, John Wiley & Sons, 2004.

[8] B. Mandelbrot and J. Wallis, Computer Experiments with Fractional Gaussian Noises Parts I,II,III,*Water Resources Research* 5, 1969, pp. 228–267.

[9] L. Giraitis, L. Kokoszka, P Leipus and G. Teyssiere, On the Power of the R/S-type Test against Contiguous and Semi-Long-Memory Alternatives,*Actae Aplicandae Mathematicae* 78, 2003, pp. 285–299.

[10] V. Teverovsky, M. Taqqu and W. Willinger, A Critical Look at Lo's Modified R/S Statistic,*Journal of statistical Planning and Inference* 80(1,2), 1999, pp. 211–227.

[11] B. Tsybakov and N. Georganas, Self-similar Processes in Communication Networks,*IEEE Transactions on Information Theory* 44(5), 1998, pp. 1713–1725.

[12] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, New York, 1994.

[13] C. A. Fulton, and S Li, Delay jitter first-order and second-order statistical functions of general traffic on high-speed multimedia networks, *IEEE/ACM Transactions Networking*, Vol. 6, pp. 150-163, April 1998.

[14] L. Qiong, D. Mills, Jitter-based delay-boundary prediction of wide-area networks, *IEEE/ACM Transactions on Networking*, Vol. 9, pp. 578 590, October 2001.

[15] E. J. Daniel, C. M. White and K.A. Teague, An inter-arrival delay jitter model using multi-structure network delay characteristics for packet networks, *Conf. on Signals, Systems and Computers*, pp. 1738 1742, 2003

[16] K. Park and W. Willinger,*Self-Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, New York, 2000.

[17] W. Willinger, M. Taqqu, R. Sherman and D. Wilson Self-similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *Proc of ACM SIGCOMM Comp. Comm. Rev*, 1995, pp. 100–113.

[18] W. Willinger, M. Taqqu, R. Sherman and D. Wilson Self-similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *IEEE/ACM Transactions on Networking* 5(1), 1997, pp. 71–86.

[19] D. Veitch and P. Abry A Wavelet Based Joint Estimator of the Parameters of Long-Range Dependence, *IEEE Transactions on Information Theory* 45(3), 1999, pp. 878–897.

[20] J. Beran, R. Sherman, M. Taqqu and W. Willinger Long-Range Dependence in Variable-Bit-Rate Video Traffic, *IEEE Transactions on Communications* 43, 1995, pp. 1566–1579.

[21] J. Lopez-Ardao, J. Lopez-Garcia, C. Suarez, A. Fernandez and M. Rodriguez On the Use of Self-Similar Processes for Network Simulation, *ACM Transactions on Modelling and computer Simulation* 10(2), 2000, pp. 125–151.

[22] V. Paxson and S. Floyd Wide-Area Traffic: The Failure of Poisson Modelling, *IEEE/ACM Transactions on Networking* 3(3), 1995, pp. 226–244.

[23] P. Abry and D. Veitch Wavelet Analysis of Long-Range Dependent Traffic, *IEEE Transactions on Information Theory* 44(1), 1998, pp. 2–15.

[24] J. Berlaint, G. Dierckx, Y. Goegebeur and G. Matthys Tail-Index Estimation and an exponential regression Model, *Extremes* 2, 1999, pp. 177–200.

[25] J. Berlaint, P. Vynckier and J. Teugels Tail index estimation, Pareto quantile plots, and regression diagnostics, *Journal of the American statistical Association* 91, 1996, pp. 1659–1667.

[26] N. Bingham, N. Goldie and J. Teugels, *Regular Variation*, Cambridge University Press, 1987.

[27] R. Voss, *Fractals in nature: from characterization to simulation. In Peitgen HO. Saupe D (eds) the science of fractas images*, Springer, Berlin Heidelberg. New York, 1988.

[28] J. Bassingthwaighte and G. Raymond Evaluation rescaled range analysis for fractal time series, *Annals of Biomedical Engineering* 22, 1994, pp. 1659–1667.

[29] B. Tsybakov and N. Georganas On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probabilities Bound and Cell Delay Distribution, *IEEE/ACM Transactions on Networking* 5(3), 1997, pp. 397–409.

[30] M. Taqqu, V. Teverovsky and W. Willinger Is Network Traffic Self-Similar or Multifractal?, *Fractals* 5, 1997, pp. 63–74.

[31] J. Bassingthwaighte and G. Raymond Evaluation of the dispersional analysis method for fractal time series, *Annals of Biomedical Engineering* 23, 1996, pp. 491–505.

[32] M. Taqqu and V. Teverovsky, Robustness of Whittle-Type Estimators for Time-Series with Long-Range Dependence, *Stochastic Models* 13(4), 1997, pp. 723–757.

[33] P. Abry, D. Veitch and P. Flandrin Long-Range Dependence: Revisiting Aggregation with Wavelets, *Journal of Time Series Analysis* 19(3), 1998, pp. 253–266.

[34] G. Samorodnistsky and M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, Chapman and Hall/CRC, 1994.

[35] M. Crovella and A. Bestavros Explaining World-Wide Web Traffic Self-Similarity, *Tech Report Computer Science Department, Boston University)*, 1995.

[36] D. Munoz, S. Villareal, C. Vargas, M. Angulo, D. Torres and L. Rizo Heavy-Tailed Network Delay: An Alpha Stable, *Computación y Sistemas* 1, 2006, pp. 71–86.

[37] M. Taqqu and V. Teverovsky, Semiparametric Graphical Estimation Techniques for Long-Memory Data, *Athens Conference on Applied Probability and Time Series Analysis* 115(1), 1996, pp. 420–432.

[38] M. Taqqu, V. Teverovsky, and W. Willinger, Estimators for Long-Range Dependence: An Empirical Study, *Fractals* 3, 1995, pp. 785–798.