COMPARISON OF INTERNET TRAFFIC MODELS: ON-OFF SELF SIMILAR MODEL VERSUS FRACTIONAL GAUSSIAN NOISE MODEL

K. Comu[1], L. Gavrilovska[2]

[1] Electrical-Mechanical State High School, Gorgi Naumov
Partizanska b.b. PO Box 33, 7000 Bitola, Macedonia, fax: ++389(047)258065
[2] Institute of Telecommunications, Fac. El. Engineering, Univ. Ss. Cyril & Methodius
Karpos 2, b.b. PO Box 574, 1000 Skopje, Macedonia
koco0404@yahoo.com, liljana@etf.ukim.edu.mk

Abstract: This paper focuses on investigation which model is propriety for generation of artificial Internet traffic streams depending on number of sources n. Two major models applicable to capture the specifics of the Internet traffic are: ON-OFF self-similar model and Fractional Gaussian Noise (FGN) model. ON-OFF series are generated using Pareto distribution and FGN series are generated using spatial renewal process (SRP). The relevant analysis parameters are mean, Hurst parameter, variance and the probability density function (PDF). The result shows that for small n ON-OFF is more appropriate, and for large n the better solution is FGN.

Keywords: Internet traffic model, self-similar, fractal, ON-OFF, Pareto, FGN, SRP, PDF

1 Introduction

Empirical measurements suggesting self-similarity in the traffic behavior have a significant influence on the field of traffic modeling and simulation. The phenomenon has been observed in both LAN (Leland et al. 1994) and WAN data traffic (Paxon and Floyd, 1995). Long range dependence (LRD), referring to self-similarity of the second-order statistics, has become particularly relevant to the characterization of Internet traffic. Traffic possessing long range dependence exhibits sub-exponential (hyperbolically) decay in the time dependence structure as measured by the autocorrelation function. Such traffic produces significantly different behavior in queueing systems as compared to that produced by non-LRD traffic models such as Poisson. The measure of self-similarity is Hurst parameter named after hydrologist Hurst who formally introduced these models for river flows in the 50's (Hurst, 1951). Hurst parameter defines decay of autocorrelation function.

Two major models are mentioned in the literature as applicable to capture the specifics of the Internet traffic:

- ON-OFF self-similar model
- Fractional Gaussian Noise (FGN) model

The goal of this paper is to light out the advantages and weakness in each of this models with assay on number of sources. For comparison purposes we simulate both traffic models in different load conditions and for different number of sources.

The paper is organized as follow:

Section 2 describes self-similarity (Willinger and Paxon, 1998) and variance time plot (Bulmer, 2000) as a method for calculating the Hurst parameter.

Section 3 describes generation of sample paths with ON-OFF self-similar using Pareto distribution for ON and OFF periods (Kramer, 2001). This model simulates aggregated data traffic from n different independent sources, each with load L_i. Input parameters for this model are Hurst parameter H, number of sources n, capacity of one source C, and load of a single source L_i expressed as percent of C. Capacity of one source C, is maximum number of packets per observed time segment.

Section 4 describes generation of sample paths according to Fractional Gaussian Noise model using spatial renewal process (Taralp et al. 1998). Input parameters for this model are Hurst parameter H, mean value and variance of traffic flow.

Section 5 describes results from comparison of these two models. Subsection 5.1 captures the ON-OFF simulation model while subsection 5.2 captures FGN simulation model. These results are in agreement with empirical measurements introduced in (Lucas et al. 1997).

Section 6 gives conclusions of comparison.

2 Self – Similarity

There is strong evidence that the Internet **session** arrival process is Poisson. That is, human Internet users seem to operate independently at random when initiating access to certain Internet resources. This observation has been noted for several network applications. For example, (Paxson and Floyd, 1995) study telnet traffic and find that the session arrival process is well-modeled with a Poisson process, though with a time-varying rate (e.g., hourly). Similarly, (Arlitt and Williamson, 1997) find that the user requests for individual Web pages on a Web server are often well-modelled by a Poisson process.

Although session arrival process is Poisson, detailed studies of Internet network traffic show that the **packet** arrival process is not Poisson. That is, the inter-arrival times between packets are not exponentially distributed, nor are they independent. Rather, the packet arrival process is bursty: packets arrive in “clumps" that make the traffic far more bursty than predicted by a Poisson process (Paxon and Floyd, 1995). As a result, the queueing behaviour can be much more variable than predicted by a Poisson model. This non-Poisson structure is due to the protocols used for data transmission.

Statistically, temporal high variability in traffic is captured by *self-similarity*. Formally, for a self-similar stationary process $X= \{ X_i \mid i = 1, 2, 3, 4, ….. \infty \}$ holds (Willinger and Paxon, 1998):

$$Var(X^{(m)}) = Var(X)m^{2H-2} \qquad (1)$$

for $m \geq 1$.

Var(X) is variance of process X, and Var($X^{(m)}$) is variance of process $X^{(m)}$=\{ $X_i^{(m)}$ | i = 1, 2, 3, 4, …, ∞ \} where:

$$X_i^{(m)} = \frac{1}{m} \times \sum_{k=im-m+1}^{k=im} X_k \qquad (2)$$

The *degree of self-similarity* is expressed by the *Hurst parameter H* in equation (1). H varies between 0.5 and 1, where a larger value indicates a higher degree of self-similarity.

The resulting linear representation of log(Var($X^{(m)}$)) against log(m) is called the *variance time plot*. For some sample path X=\{ X_i | i = 1, 2, 3, 4, …. ∞ \}, Hurst parameter is calculated from the slope β of the least-squares line on variance time plot (Bulmer, 2000):

$$H = 1 + \frac{\beta}{2} \qquad (3)$$

For a short-range dependent process, such as the Poisson-based models the H parameter will be approximately 0.5, while for the measured Internet traffic typical values for H parameter are about 0.7 to 0.8 (Lucas et al. 1997). Thus Internet traffic can not be modeled with a Poisson based process.

3 ON-OFF self-similar traffic using with Pareto distribution

The self-similarity can be captured by generation of series that represent the aggregate traffic of n independent ON-OFF sources, each described with *heavy-tailed distribution* of ON and OFF periods. Such heavy-tailed distribution can be found at well-known *Pareto* function, originally introduced from mathematician Pareto for modeling the income within a population. Pareto distribution has the following probability density function:

$$P(x) = \frac{ab^a}{x^{a+1}} \qquad (4)$$

where a is a shape parameter (tail index), and b is minimum value of x. When $a \leq 2$, the variance of the Pareto distribution is infinite. When $a \leq 1$, the mean value is infinite as well. For self-similar traffic, a should be between 1 and 2. The lower the value of a, the higher the probability of an extremely large x. Distribution of ON periods has parameters a_{ON} and b_{ON}, while distribution of OFF periods has parameters a_{OFF} and b_{OFF}. For generating sample path with desired H, values of these parameters are (Kramer, 2001):

The a_{ON} is:

$$a_{ON} = 3 - 2H \qquad (5)$$

Theoretical a_{OFF} is arbitrarily, but practical a_{OFF} should be smaller then a_{ON}, because it is reasonable to assume that in real traffic, probability of having extremely large OFF periods is higher then the probability of having extremely large ON periods.

The minimum value of ON period b_{ON} is 1, which represents a case of generating just one packet:

$$b_{ON} = 1 \tag{6}$$

When traffic stream is not expressed in [packet/s] but in [bit/s], then b_{ON} is size of one packet expressed in bits.

The b_{OFF} is:

$$b_{OFF} = b_{ON} \frac{T_{OFF}(1 - S^{T_{ON}})(1 - L_i)}{T_{ON}(1 - S^{T_{OFF}})L_i} \tag{7}$$

where:

$$T_{OFF} = \frac{a_{OFF} - 1}{a_{OFF}};\ T_{ON} = \frac{a_{ON} - 1}{a_{OFF}} \tag{8}$$

S is smallest random number that computer can generate.

L_i is load of a single source expressed as percent of C, and C is capacity of one source (maximum number of packets or bits per observed time segment).

4 FGN self-similar traffic using SRP

The other model for simulating self-similar traffic is Fractional Gaussian Noise (FGN) model. In this paper we use Spatial Renewal Process (SRP) to get self-similar FGN.

SRP consists of two background processes. The first background process is represented with inter-renewal time sequence $\{T_n\}$ defined by the cumulative distribution function $F_T(t)$. This process is responsible for the time dependent structure (autocorrelation). The second background process, independent of the first, is represented with sequence of values $\{X_n\}$ which are distributed according to the desired steady-state marginal distribution of the traffic (in this case Gaussian, to obtain FGN). The SRP process $\mathrm{Y_t}$ is composed of a chain of renewal periods where nth period is T_n in length and the sample path during this period takes on the value of X_n.

The choice of a marginal distribution and autocorrelation are fully decoupled from each another in the SRP model. It follows from the independence of the X_n and T_n sequences.

To generate self-similar traffic with desired H, CDF $F_T(t)$ must be (Taralp et al. 1998):

$$F_T(t) = 1 - (2^{2H-1} - 2)^{-1} \begin{cases} 2^{2H-1} - 2, & 0 \le t < 1 \\ H(t+1)^{2H-1} - 2Ht^{2H-1} + H(t-1)^{2H-1}, & 1 \le t \end{cases} \tag{9}$$

$F_\mathrm{T}(t)$ is used to generate T_n, while Gaussian function is used to generate X_n with desired mean and variance. The result is block-like sample path $Y(t)$. To repress unnatural block-like shape, it is first generated and then summed 10 independent $Y_1(t)$, $Y_2(t)$, …, $Y_{10}(t)$ each with desired H parameter, but with normalized mean and variance. The summing do not affect time dependence structure (autocorrelation). The final, aggregated sample path $Z(t) = Y_1(t)+Y_2(t)+\ldots+Y_{10}(t)$ is with desired Hurst parameter, mean and variance.

5 Comparison of two models

In order to conclude on applicability of these two simulation models under different conditions (number of sources and load) we completed extensive numbers of simulations on them.

5.1 ON-OFF simulation

First we generate series of $5*10^5$ samples of one source (n=1) with desired mean 150 [packet/sec.] and Hurst parameter 0.7, using ON-OFF Pareto model. This series is analyzed by calculating its mean, variance, PDF and Hurst parameter. In our experiment it was taken that maximum number of packet by one source in observed time segment is C=600[packet/sec.], load is L_i=25%, and the time segment is 1[s]. The next Fig. 1, to Fig. 4 represents the generated series:

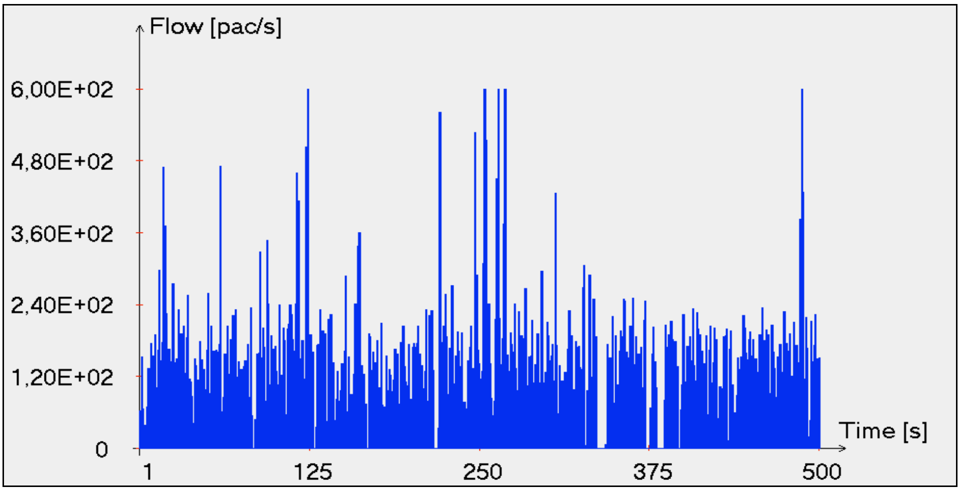

Fig. 1: First 500 samples of ON-OFF Pareto model with H=0.7, C=600[pac/s], L_i=25%, n=1

Fig. 1 represents the first 500 of $5*10^5$ samples of generated series. It is obvious that the traffic is bursty.

Fig. 2 represents the PDF of generated series. There are significant peaks for zero[pac/s] (due to OFF periods), and for 600[pac/s] (due to ON periods). PDF of generated series is not Gaussian at all. Common surface of generated PDF and Gaussian function with mean and variance as generated series is 76.94%. Mean X_{mean} of generated series is 149.77[pac/s] and it's as expected because $n*C*L_i$ = 1*600[pac/s]*25%

= 150[pac/s]. Variance of generated series is 7934[pac/s]^2, but we could not predict its value before the simulation, backing on input parameters.

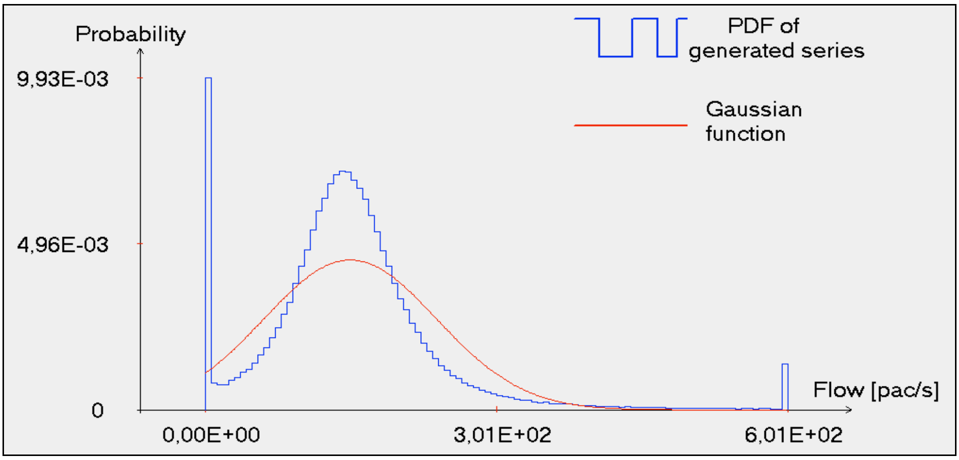

Fig. 2: PDF of ON-OFF Pareto model with *H*=0.7, *C*=600 [pac/s], L_i=25%, *n*=1, Mean=149.77 [pac/s], Variance = 7934 [pac/s] ^{2}, common surface is 76.94%

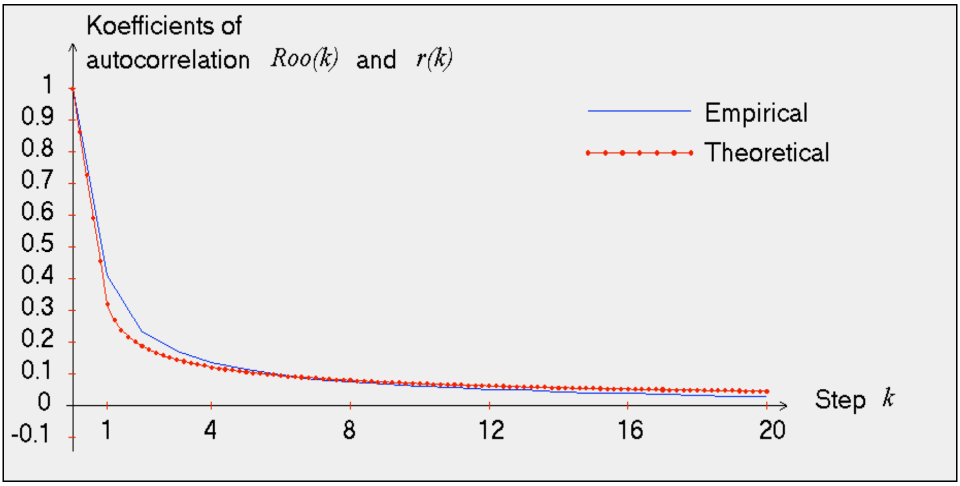

Fig. 3: Autocorrelation of ON-OFF Pareto model with *H*=0.7, *C*=600[pac/s], L_i=25%, *n*=1

Fig. 3 represents the autocorrelation function of generated (empirical) series compared to the desired (theoretical) autocorrelation function.

Formula for autocorrelation coefficients of generated (empirical) series is (Bulmer, 2000):

$$r(k)=\frac{\sum_{i=1}^{N-k}(X_i-X_{mean})(X_{i+k}-X_{mean})}{\sum_{i=1}^{N-k}(X_i-X_{mean})^2} \qquad (10\text{-}a)$$

where X_i is i-th sample of altogether N generated ($5*10^5$) samples, X_{mean} is mean value of generated samples, and k is the step.

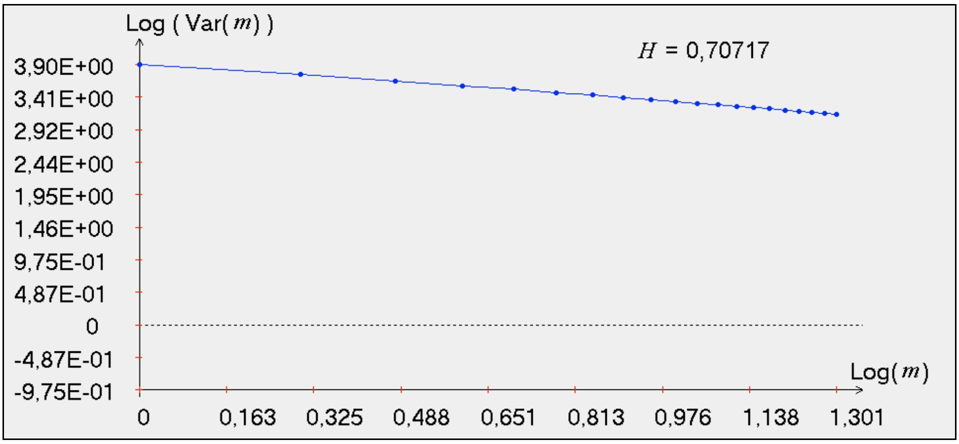

Fig. 4: Variance time plot of ON-OFF Pareto model with H=0.7, C=600[pac/s], L_i=25%, n=1

The desired (theoretical) autocorrelation coefficients for the self-similar process with Hurst parameter H is (Taralp et al. 1998):

$$R_{00}(k)=\begin{cases}1 & k=0\\ \frac{1}{2}[(k+1)^{2H}-k^{2H}+(k-1)^{2H}] & k=1,2,3,...\end{cases} \qquad (10\text{-}b)$$

The hyperbolical decay of both empirical and theoretical autocorrelation, characteristic for the long range dependence processes, is obvious form, Fig.3.

Fig. 4 represents the variance time plot for the generated series. Calculated Hurst parameter is 0.70717 and is very close to desired 0.7.

Our main interest is how the PDF will look like as n increases. Fig. 5 represents the PDF of generated ON-OFF Pareto series with H=0.7, C=600[pac/s], L_i=25% and n=2:

The mean of generated series is 299.86[pac/s] as expected because $n*C*L_i$=2*600[pac/s]*25%=300[pac/s]. Calculated Hurst parameter is 0.70791 as expected. Variance is 15907[pac/s]^2 and we could not predict it backing on input parameters. Common surface is 86.54% that is larger than in a previous case.

Fig. 6 represents PDF of generated ON-OFF Pareto series with H=0.7, C=600[pac/s], L_i=25% and n=6:

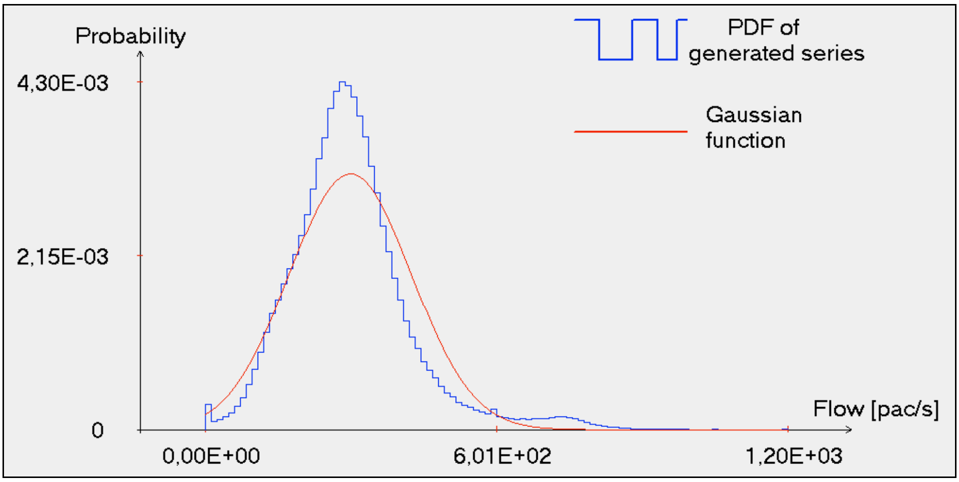

Fig. 5: PDF of ON-OFF Pareto model with H=0.7, C=600[pac/s], L_i=25%, n=2, Mean=299.86[pac/s], Variance = 15907[pac/s]^2, common surface is 86.54%

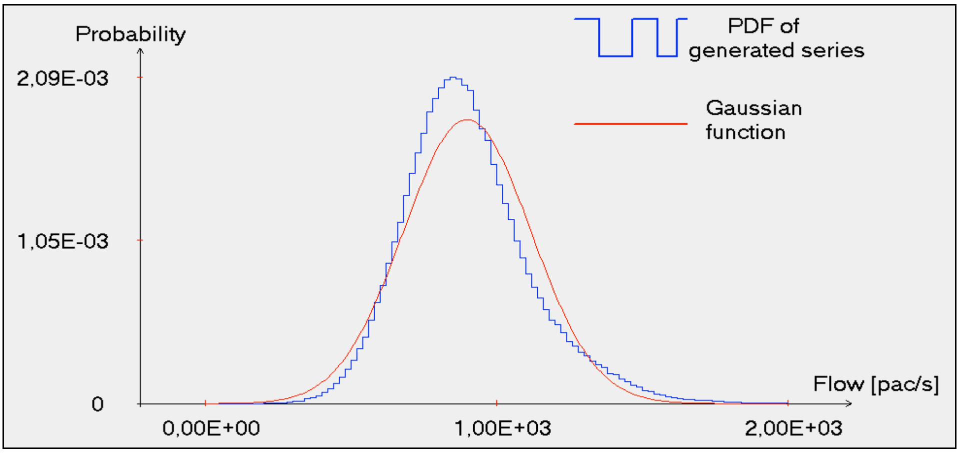

Fig. 6: PDF of ON-OFF Pareto model with H=0.7, C=600[pac/s], L_i=25%, n=6, Mean=899.96[pac/s], Variance = 47954[pac/s]^2, common surface is 90.83%

The mean of generated series is 899.96[pac/s] as expected because n*C*L_i=6*600[pac/s]*25%=900[pac/s]. Calculated Hurst parameter is 0.70991 as expected. Variance is 47594[pac/s]^2 and we could not predict it backing on input parameters. Common surface is 90.83% that is larger than in a previous case.

From Fig. 2, Fig. 5 and Fig. 6 can be seen that as n increases the PDF of generated series looks more and more like Gaussian function.

Fig. 7 representes the PDF changes dependening on load L_i. So, there is H=0.7, C=600[pac/s], n=6 as in the previous case, but L_i=10%:

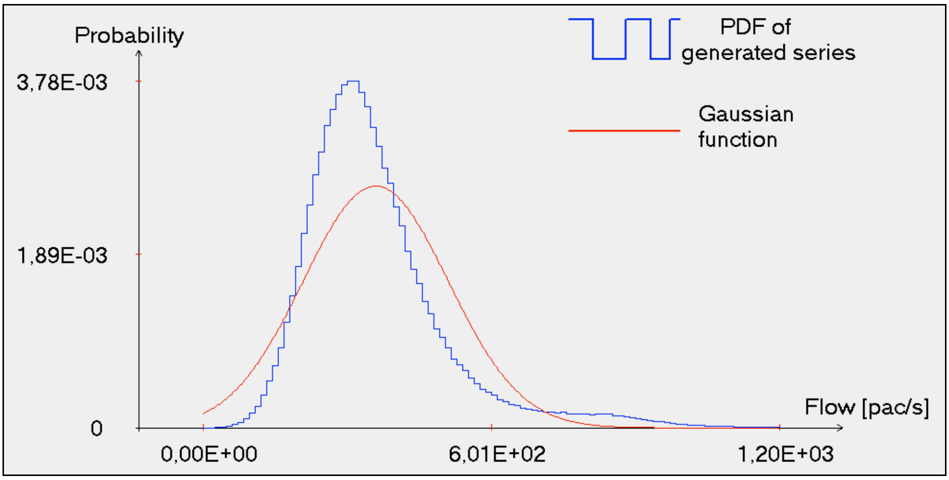

Fig. 7: PDF of ON-OFF Pareto model with H=0.7, C=600[pac/s], L_i=10%, n=6 Mean=359.99[pac/s], Variance = 22910[pac/s]^2, common surface is 79.72%

The mean of generated series is 359.99[pac/s] as expected because n*C*L_i=6*600[pac/s]*10%=360[pac/s]. Calculated Hurst parameter is 0.71218 as expected. Variance is 22910[pac/s]^2 and we could not predict it backing on input parameters. Common surface is 79.72% that is less than in a previous case.

From Fig. 6, and Fig. 7 it is obvious that as L_i increases the PDF of generated series looks more and more like Gaussian function.

5.2 FGN simulation

Now we generate 5*10^5 sample path with FGN SRP model. Generated path is with desired H=0.7, mean=900[pac/s], and variance 47954[pac/s]^2. These values are chosen according ON-OFF Pareto series shown on Fig. 6. The follow Fig.8 to Fig. 10 show the FGN SRP series:

Fig. 8 represents the sample path of generated FGN SRP series. The traffic also has bursty nature. Fig. 9 represents PDF, which is of course Gaussian. Fig. 10 represents hyperbolical decay of autocorrelation function. The mean of generated series is 899.3[pac/s], the variance is 47771[pac/s]^2 and calculated Hurst parameter is 0.69392 as expected.

6 Conclusions

This paper is analyzing two methods for simulation of Internet traffic. During simulations it is of interest to fit the larger number of process parameters (Hurst parameter, mean and variance). The PDF of real traffic is not Gaussian for low flow (Lucas et al. 1997), but looks more like PDF of ON-OFF models under low load. As flow increase, the PDF of real traffic is beginning to look like Gaussian more and more (Lucas et al. 1997), which is in agreement with CLT (central limit theorem). When load increase at ON-OFF models they begin to create sample paths with PDF that looks more and more like Gaussian. The ON-OFF model gives us the opportunity to fit the mean but not the variance. So, if we know the mean of real traffic at low load level, it is appropriate to simulate that traffic with ON-OFF models till PDF do not look like Gaussian. On the other hand, when we want to simulate traffic on high load level where PDF of real flow looks like Gaussian it is appropriate to use FGN model because this model gives us opportunity to fit both the mean and variance close to those at real traffic.

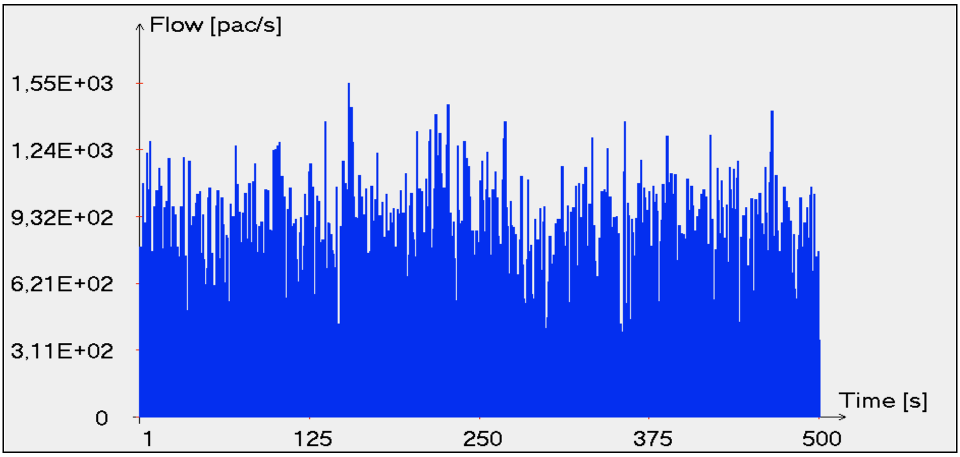

Fig. 8: Sample path of FGN SRP model with *H*=0.7, mean=900[pac/s] and variance=47954[pac/s]^2

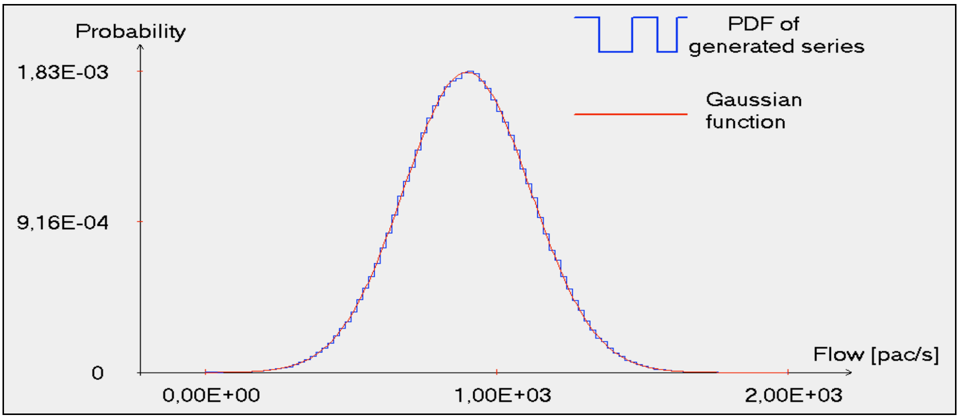

Fig. 9: PDF of FGN SRP model with *H*=0.7, mean=900[pac/s] and variance=47954[pac/s]^2

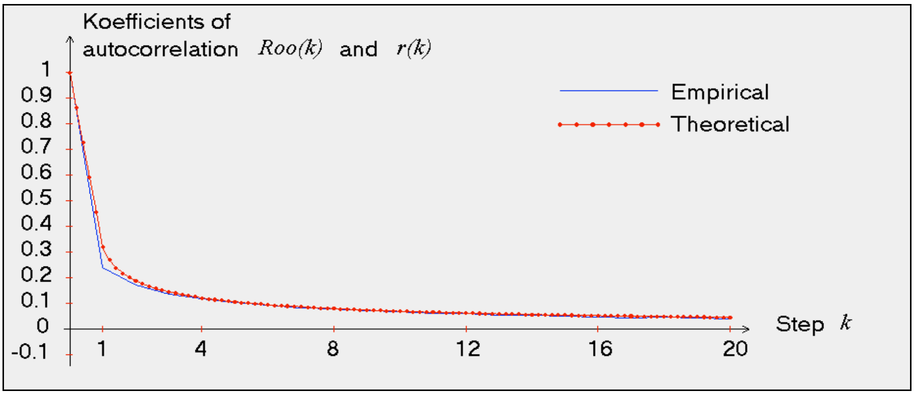

Fig. 10: Autocorrelation of FGN SRP model with H=0.7, mean=900[pac/s] and variance=47954[pac/s]^2

7 References

1. M. Arlitt and C. Williamson, “Internet Web Servers: Workload Characterization and Performance Implications", *IEEE/ACM Trans. Networking*, Vol. 5, No. 5, pp. 815-826, October 1997.
2. M. Bulmer, “Music from Fractal Noise”, *Proc. Mathematics 2000 Festival*, Melbourne, 10-13 January 2000
3. H. E. Hurst, “Long-term storage capacity of reservoirs,” *Transactions of the American Society of Civil Engineers*, pp. 770-808, 1951
4. G. Kramer “Self-similar Network Traffic”, *The notions and effects of self-similarity and long-range dependence*, 5.21.2001, Available at: http://www.csif.cs.ucdavis.edu/~kramer/papers/ss_trf_present2.pdf
5. W. Leland, M. Taqqu, W. Willinger, and D. Wilson, “On the Self-SimilarNature of Ethernet Traffic (Extended Version*)*,” *IEEE/ACM Trans. Networking, 2(1)*, pp.1-15, February 1994
6. M. T. Lucas, D. E. Wrege, B. J. Dempsey, A. C. Weaver, "Statistical Characterization of Wide-Area IP Traffic" *6th Int. Conf. Computer Communications and Networks (IC3N'97)* Las Vegas, NV, September 1997, Available at: www.ils.unc.edu/~bert/papers/ic3n-2col.pdf
7. V. Paxson and S. Floyd, “Wide-Area Traffic: The Failure of Poisson Modeling” *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226-244, June 1995.
8. T. Taralp, M. Devetsikiotis, I. Lambadaris, A. Bose, “Efficient Fractional Gaussian Noise Generation Using the Spatial Renewal Process”, *Proc. IEEE Int. Conference on Communications ‘98*, Atlanta, USA, June 7-11, 1998, Available at: www.research.att.com/~jyates/opticalnetworking/papers/serc/timeScales.ps
9. W. Willinger and V. Paxon, “Where mathematics meets the Internet”, *Notices of the American Mathematical Society*, Vol.45, No.8, pp.961- 970, September 1988