Synthesizing of Markovian and Self-similar LAN/WAN Traf c on Data Networks

Anibal D. A. Miranda Alessandro Anzaloni
Instituto Tecnológico de Aeronaútica (ITA)
Electronic Engineering/Telecommunications Department
São José dos Campos 12228-900, SP.
e-mail:{anibal,anzaloni}@ele.ita.br

Abstract— **Lately, the actual statistical behavior of LAN/WAN traf c is being confronted against classical Markovian-based models. Many theoretical and simulation work have been done to date. Almost all such disagreements are based on software simulations. In this article we show a simple yet reliable tool that can emulate in an ef cient manner many LAN/WAN traf c workloads. Spanning from Markovian-type traf c to fractal models. With this tool you can reproduce the statistical behavior you want your IP traf c ows would have over a real network environment. It can reproduce either real aggregated traf c (previously measured traces) or simply one workstation traf c behavior. This is a UDP-based tool, averting the complex TCP ow control mechanism (this was done so just for practical purposes only), nevertheless TCP ows can be reproduced though. All tests using this tool were supported with well-funded statistical analysis methods and was veri ed that statistical properties, overall concerning the Hurst (H) parameter, from such theoretical traces was maintained successfully, i.e., $H_{real} \rightarrow H_{theo}$. All these features make this tool very reliable to work with.**

Index Terms —**Workload characterization, Traf c engineering, Real Web Traf c Measurements.**

I. INTRODUCTION

LATELY , new paradigms about network traf c characterization are under discussion, but no one can argue against that there exist a very strong evidence that LAN/WAN traf c is best tted with fractal models rather than Markovian-based ones, this conclusion was coined based on the seminal work [1]. Therefore, the need of synthetical Local Area (LAN) and Wide Area (WAN) Network traf c following as close as possible the statistical behavior of real LAN/WAN traf c must have a high priority in the traf c engineering and performance measurement world. Short-range dependent (SRD) processes have been confronted with long-range (LRD) dependent processes, being the latter more successful in that those processes can emulate ef ciently real LAN/WAN traf c. For practical research purposes it is necessary a tool that can reproduce such network traf c dynamic, both SRD and LRD, just to study and verify whether network operating systems (*firmware*) can support such workloads given some quality of service (QoS) for end users.

To this end, a simple tool is being developed and its results are promising. This tool can emulate many data traces spanning from mathematical models, namely fractals [1] or Markovian-based [2] ones, through real IP traf c traces (previously measured) that show aggregated (Gaussian or non-Gaussian) or *single ow* (ON −OFF exponential or paretian) behavior.

As explained below, this tool can be used whether on connection oriented high-speed network border switches or Internet routers with QoS, where call admission control (CAC) algorithms for traf c management and policing are necessaries.

This work will not discuss about Markovian or self-similar theory, but shows how those theoretical models can be used to produce synthetical IP traf c ows instead.

Therefore, Section 2 brie y explains how traf c is generated, Section 3 shows two examples that argue in favor of the reliability of this tool, Section 4 presents the statistical analysis applied on those theoretical models and real IP traf c so obtained, Section 5 gives future research trends. Finally, Section 6 outlines the major conclusions of this work.

II. SYNTHETICAL IP TRAFFIC GENERATION

In order to obtain those synthetical IP ows, the following steps have to be followed:

- Install both sink (server) and source (client) programs on Linux based machines (`gcc` is enough to compile them), this client/server software has been tested on Linux, SunOS and AIX systems.
- The trace le, to be emulated, has to be stored in the source (client) machine. Integer values reading from this le will be the number of packets to be sent in that interval of time.
- This traf c generator runs as a command line interface [1] program.

Therefore, to reproduce any LAN/WAN traf c behavior, the procedure above is mandatory. Fig. 1 shows the test bed environment layout where tests took place. Beside the client/server computers, a third machine appears in this gure, this computer is the Network Control Center (NCC), it is intended for querying the router's network interface. This task was done via *GetRequests*/*GetResponses* to/from the router, the data so collected was stored to be processed later. Basically, NCC asks for the number of IP packets that are owing through it directly to the Management Information Base (MIB) of the router [4].

This is an UDP-based tool, so that, the TCP complex control loop is avoided (no window control), even though TCP traf c could be emulated as well. UDP is useful when single

[1] A GUI has not been developed yet

packets have to be sent with no previous positive acknowledgements [3]. Moreover, UDP is good for sending messages from one system to another when the order is not important and all of the messages to get to the other machine is not needed.

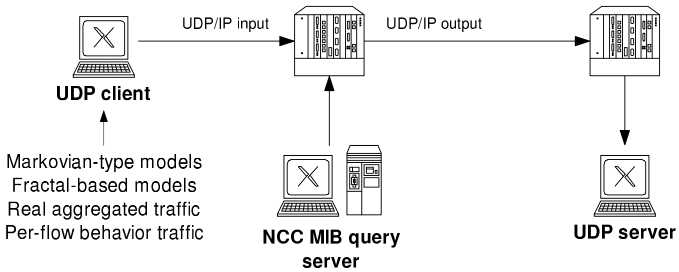

Fig. 1. Network con guration test bed environment

Owing the limited number of pages, just two[2] representative examples showing the power of this easy tool are depicted.

A. Traffic Characterization

Network traf c monitoring implies, in someway, a constant inquire task to the router interface, as the traf c generator produce packets with the same length, our main attention is focused in the number of packets per unit time i (see Fig. 2). Sampling process has a rate of 20 msec. per sample approximately. Therefore, unit time is $\Delta T \approx 20$ msec. Then

$$X_i = \frac{\#\text{ of packets per unit time } i}{\Delta T}$$

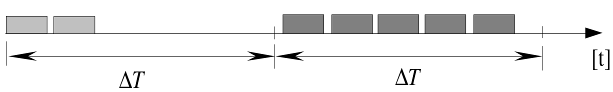

Fig. 2. Traf c characterization

It is assumed that the traf c so produced $X = \{X_i\}_1^N$ is *wide sense* stationary (WSS) and $N\Delta T$ = *measurement duration*. The process $X_i^{(m)}$ is known as the *aggregate process* and it is de ned as an average over m non-overlapped blocks, i.e., the original time series given by $X_i = (X_1, X_2, \ldots)$ and $X_i^{(m)} = (X_1^m, X_2^m, \ldots)$ lead us to

$$X_i^{(m)} = \frac{1}{m}\,(X_{im-m+1}, \ldots + X_{im})$$

this time series is also, for $m \geq 1$, a WSS stochastic process with m the *time scale*. For example, for $m = 3$ the time series under study will have non-overlapped blocks size of 3. X_i is the number of IP packets arriving in the i-*th* time interval. Basically, aggregation and averaging tend to smooth the structure of the original time series, X_i. Furthermore, $X_i^{(m)}$ is useful to estimate H, mathematically is given by

$$X_i^{(m)} = \frac{1}{m}\Big[\sum_{j=(i-1)m+1}^{im} X_j\Big] \tag{1}$$

[2]Many tests and trials were performed in our laboratory.

Therefore, X_i is *exactly self-similar* with parameter H if, for all aggregation levels m, the following sequences have the same distribution for $i \geq 1$, then

$$m^H X_i \overset{\text{def}}{=} m\,X_i^{(m)} \tag{2}$$

Where the equality, $\overset{\text{def}}{=}$, is in the sense of nite-dimensional distribution, i.e., their statistical properties are the same for the aggregated sum as for the original time series, a canonical example is the fractional Brownian motion (fBm) [8]. On the other hand, X_i is called *asymptotically self-similar* if eq.(2) holds as $m \to \infty$.

The concepts above were used to implement the synthetical workload traf c generator.

III. Real IP Traffic Examples

Recall that X_i is the process to be emulated, then for Markovian-type model the classical Markov Modulated Poisson Process (MMPP) was chosen and for self-similar model the fractional Gaussian noise (fGn) was selected.

A. Example 1: Markovian-type traffic $(H = 0.5)$

Markovian-type models were and still being useful for performance measurements purposes as stated in [2] and many other papers. This is due, such processes have a well-developed mathematical theory. It is also knew that those models obeys the non-memory property and therefore can be considerate as SRD processes [1][5][6]. Fig. 3 shows one trace of a Poisson process with a mean packet rate given by $\lambda = 40$. Fig. 4 shows the result obtained when this mathematical model was the input process to the synthetical traf c generator, the IP traf c so obtained follows closely the theoretical model behavior.

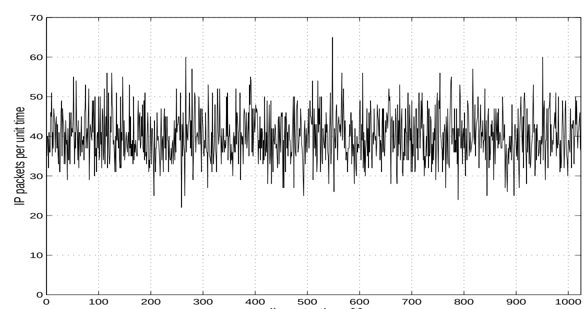

Fig. 3. Theoretical Markovian-based trace

B. Example 2: Self-similar traffic $(H = 0.84)$

One of the most outstanding result that has been reached in the past few years was the discovery of the *self-similar*, or in common terms, the fractal characteristic in statistical behavior of LAN/WAN traf c, as stated in [1]. Fractals can be considerate as long-range dependence (LRD) processes because their statistical properties vanishing slowly. Ever since, many technical papers, based on this seminal work [1], have been written trying to explain how LAN/WAN traf c has such behavior. Despite this brand new traf c paradigm in modern data networks,

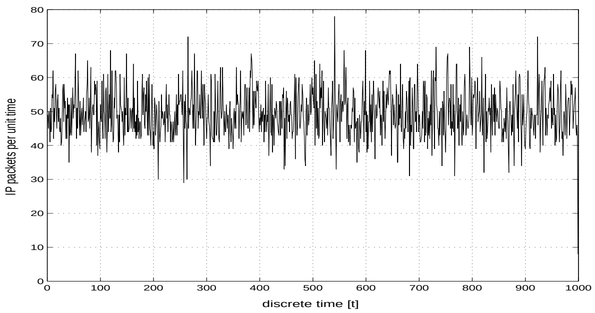

Fig. 4. Real IP traf c obeying that Markovian model

enormous theoretical breakthroughs have been reached on this subject.

This work uses the fGn as parsimonious model to emulate the cumulating arrival process, i.e., the aggregated workload traf c offered to a given network. Fig. 5 shows one fGn trace obtained via the *Random Midpoint Displacement* (RMD) algorithm [7] and Fig. 6 depicts the IP traf c obtained using the synthetical workload generator. The theoretical model was produced with the following parameters: mean packet rate $\mu = 25$, a variance given by $\sigma^2 t^{2H}$ and a Hurst parameter $H = 0.84$. Keep in mind that the degree of self-similarity is strictly characterized by Hurst parameter.

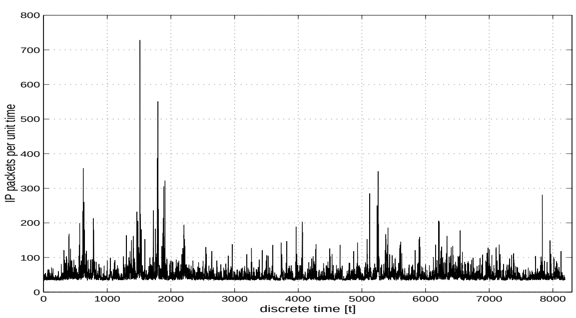

Fig. 5. Theoretical self-similar fGn trace

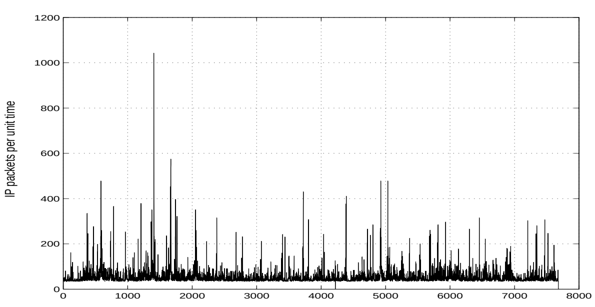

Fig. 6. Real IP traf c obeying the fGn theoretical model

IV. Statistical Analysis

In order to warrant the quality of the statistical properties of those synthetical IP traf c traces so generated, three well-known tests were performed. This helped us to verify how accurate Hurst parameter was maintained during the trials. The statistical methods were:

- The Maximum Likelihood Estimator of Whittle
- The Periodogram-based method
- The Wavelet-based method

A. The Maximum Likelihood Estimator of Whittle

This estimator provides asymptotically consistent and normally distributed estimators of the unknown parameters of both Gaussian and non-Gaussian time series. Here, we are not intended to show the involved theory of this estimator more about this method can be obtained in [8]. Fig. 7 shows this method applied over both theoretical theoretical processes and synthetical IP traf c ows. Fig. 8 shows the same method applied over the fGn process. From these gures is clear to verify that Hurst parameter was maintained for both mathematical models, i.e., for Markovian-type process it was $H \approx 0.5$ and for fGn model the estimation procedure minimized at $H \approx 0.84$, as expected.

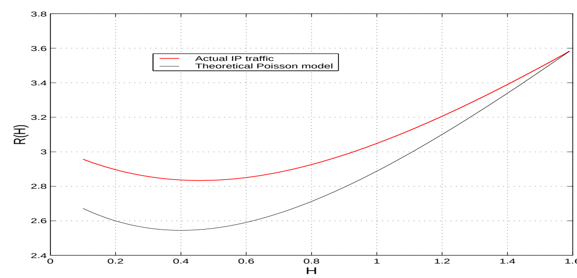

Fig. 7. Whittle estimator: Markovian-based model and IP traf c

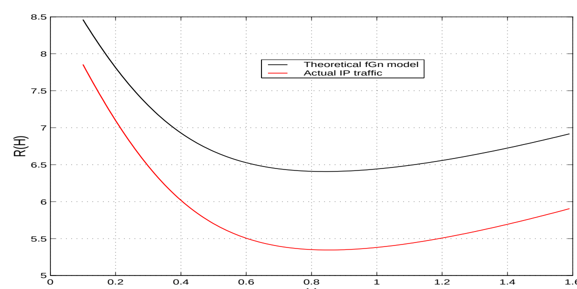

Fig. 8. Whittle estimator: self-similar fGn model and IP traf c

B. The Periodogram-based method

This method estimates the periodogram (based on Fourier technics) and reveals its behavior near the origin [1][6].

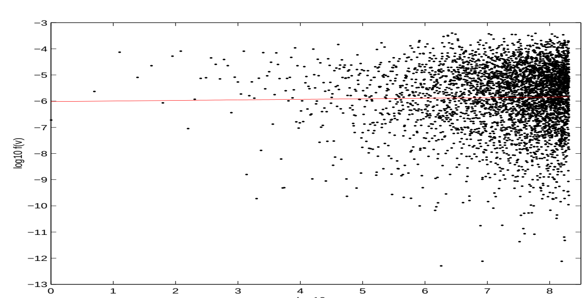

Fig. 9. Periodogram method: Markovian-based process

If the underlying process has non-LRD properties, the process will show a at spectrum over its entire range resembling a second order pure white Guassian noise process. Fig. 9 depicts such behavior for Markovian-type process. This process produced the synthetical workload IP traf c which in turn had its spectral behavior as in Fig. 10.

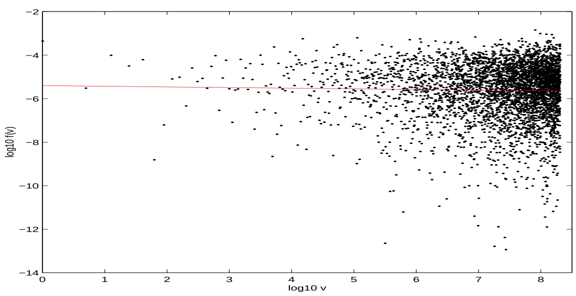

Fig. 10. Periodogram method: Markovian-based IP traf c

On the other hand, if the time series under study has intrinsically a LRD characteristic its spectral behavior should be proportional to $|\nu|^{1-2H}$ (see Table I), as $\nu \to 0$ [1][5]. Fig. 11 (corresponding to fGn theoretical model) and Fig. 12 (IP traf c obeying that model) shows clearly that behavior, i.e., the power law ($1/f^{\alpha}$) topology as $\nu \to 0$.

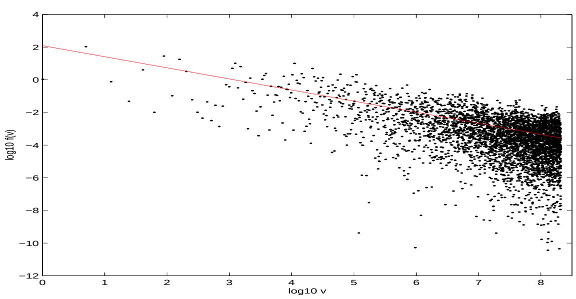

Fig. 11. Periodogram method: self-similar fGn process

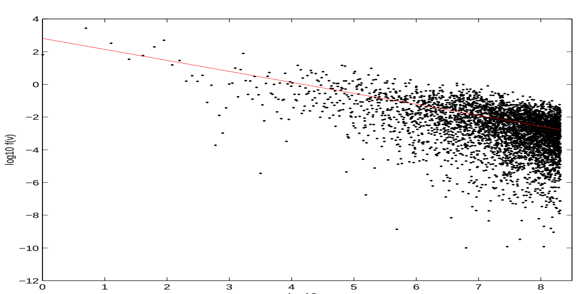

Fig. 12. Periodogram method: self-similar IP traf c

The singularity (a *pole*) around the origin, for fGn process, is clearly visible in Fig. 11, same characteristic was obtained when this model was used to produce the synthetical traf c. Fig. 12 shows the spectral behavior of the IP traf c so obtained. Table I summarizes these results.

C. The Wavelet-based method

This method is also known as the Abry-Veitch method and has proven to be unbiased under very general conditions and ef-

TABLE I

HURST PARAMETER ESTIMATION RESULTS

Periodogram-based Analysis	$\mathbf{1-2H}$	$\boldsymbol{H}$
Theoretical Poisson process	0.016735	0.49163
Actual IP Poisson traf c	-0.035106	0.51755
Theoretical fGn process	-0.679732	0.83986
Actual IP fGn traf c	-0.671513	0.83575

 cient under Gaussian assumptions. In the wavelet framework is necessary to study differences of aggregated series (eq.(1)). Regarding the simplest case one computes the difference between points in non-overlapping blocks of size 2 as de ned by the Haar wavelet. Consider now Y^{j+1} as being the series made by difference Y^j, note that Y^0 is the data series at highest time resolution.

$$Y_k^{j+1} = 2^{-\frac{1}{2}} \cdot (Y_{2k}^j - Y_{2k-1}^j) \tag{3}$$

Eq.(3) has $k = 1, 2, \ldots, N/2j$ and $j = 1, 2, \ldots$. The variance of Y^j decay according to a similar power-law as above

$$Var[Y^j] \approx 2^{j(2H-1)} \tag{4}$$

It turns out that the variance is equal to the second order moment. Since the expectation of Y is zero. In the frequency domain, the variance, $E[(Y^j)^2] = Var[Y^j]$, is equivalent to the signal energy in a frequency band depending on j (cf. [9][10]).

In resume, this method is based on the Multiresolution Analysis (MRA) and computes the Discrete Wavelet Transform (DWT), averages the squares of the coef cients (details) of the transform, and then performs a linear regression on the logarithm of the average, versus the $\log$ of j, the scale parameter of the transform. The result should be directly proportional to H, such linear relationship between $log_2[Var(\text{detail}_j)]$ and j implies LRD. Therefore, taking logarithms in eq.(4) one obtains

$$\log_2[Var(\text{detail}_j)] = (2H-1)j + c \tag{5}$$

where c is a nite constant. Eq.(5) is useful to estimate Hurst parameter with a time scale $= 2^j$. Large j implies the coarse time scale [9][10].

Markovian process: Fig.13 depicts this wavelet method performed on Markovian-based process. Fig. 14 is the actual IP traf c obtained when the synthetical workload generator was fed with such a process.

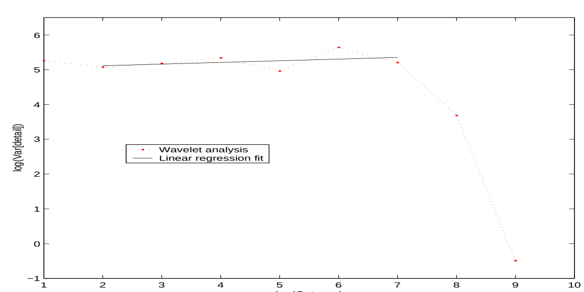

Fig. 13. Wavelet-based estimator: Markovian-based process

Results for theoretical Poisson model were $H = 0.524$ and for IP traf c that used this model, $H = 0.451$. Although slightly different from other methods, both results reveal an $H \rightarrow 0.5$ indicating SRD. Perhaps the time series length consisted of approximately 2^{10} samples would have produced this slightly biased result for $j = 5$ time scales.

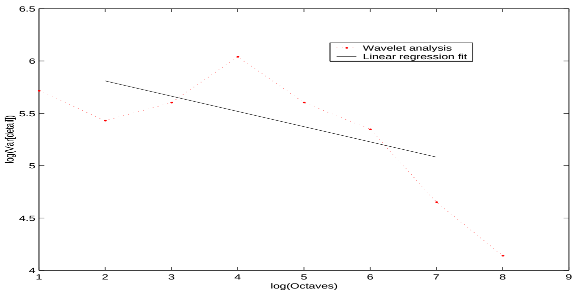

Fig. 14. Wavelet-based estimator: Markovian-based IP traf c

Self-similar process: Fig. 15 shows the analysis performed over the theoretical fGn process and Fig. 16 depicts the analysis over IP traf c obtained when such a fGn process was used to synthesize that workload traf c. Results for theoretical fGn process were $H = 0.798$ and for IP traf c using this model, $H = 0.756$. The linear relation in $log_2[Var(\text{detail}_j)]$ and j is clear and therefore this implies intrinsically the LRD property. In this case, estimation differed from that theoretical value $H = 0.84$ though.

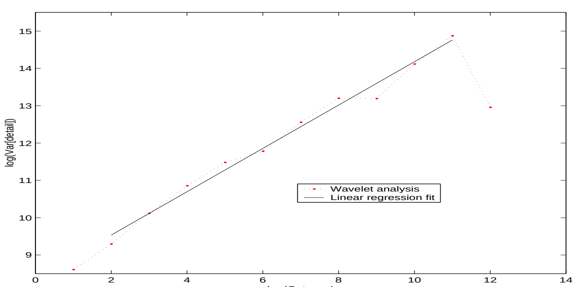

Fig. 15. Wavelet-based estimator: self-similar fGn process

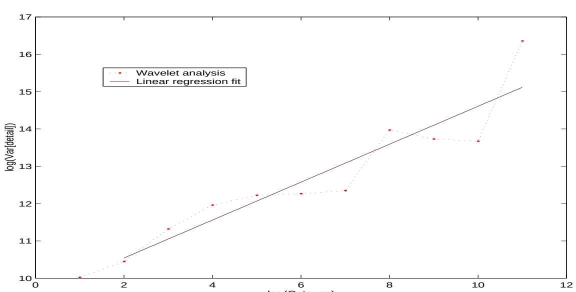

Fig. 16. Wavelet-based estimator: self-similar IP traf c

V. Two Simple Experiments

Once veri ed that our synthetical IP traf c generator produce streams with a well-de ned H parameter lying in $\frac{1}{2} \leq H < 1$, two simple experiments were bearing.

A. Experiment 1: IP packets transmission performance

Fig. 17 shows the IP packet transmission performance ratio, $\eta = \frac{R_x}{T_x}$, as a Hurst parameter function. Where R_x is the UDP/IP packets received and T_x is the whole UDP/IP packets that have been sent. As this gure shows the transmission performance is a decreasing function of H, i.e., $\eta \rightarrow 0$ as $H \rightarrow 1$.

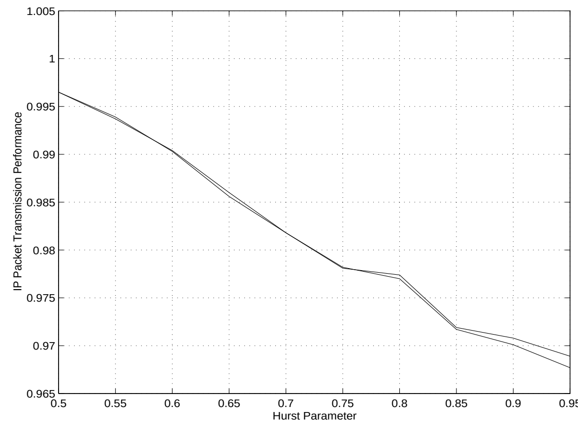

Fig. 17. UDP/IP transmission performance vs. H parameter

B. Experiment 2: IP packets loss

Fig. 18 shows the UDP/IP packets loss that this link has experienced. According to this gure the packet loss increases as $H \rightarrow 1$ as predicted by theory and software simulations. It is worth to note that our routers have worked with no DiffServ (DS) characteristics in its queueing mechanisms policing and just a default Class (*best-effort*) service was used, this is due we have not yet implemented the program code to write each UDP/IP packet header's DS eld with those bits intended for some type of service when passing through those routers. This feature will be part of future research directions.

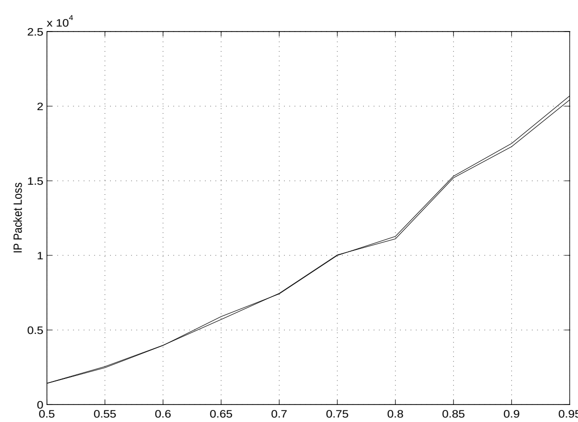

Fig. 18. UDP/IP packets loss vs. H parameter

These two simple experiments depict a real LAN/WAN networking performance problem that constitutes a great challenge for network device developers and system performance researchers all over the world.

VI. Future Research Directions

Future research work based on this tool can be addressed as follow:

- The TCP-based version for throughput and measurement purposes and comparisons with UDP.
- To include the **DiffServ** property to such IP streams data. These streams will flow through router devices supporting this feature, seeking whether its performance are according with what their developers have planned.
- Another important work being developed is depicted in Fig. 19, here we can see the scenario for a call admission control (CAC) project.

Basically, the CAC algorithm must be able to admit or reject a new incoming traffic maintaining its QoS demands whilst still maintaining all existing traffic's QoS demands. The CAC algorithm will take into account the so-called *Effective Bandwidth* concept as the LAN/WAN traffic management policing. At a first stage of CAC setup, the following parameters (among others) will be important to be negotiated for elastic traffic:

- the mean bit rate m,
- the maximum bit rate a,
- the Hurst parameter H, and
- the cell loss probability $P(X > x)$.

Based on the parameters above, it is possible to calculate the effective bandwidth C_{eff} for IP traffic input streams [12][13]. Perhaps one main tradeoff of this procedure will be the on-line estimation of such parameters, being H the most sensitive.

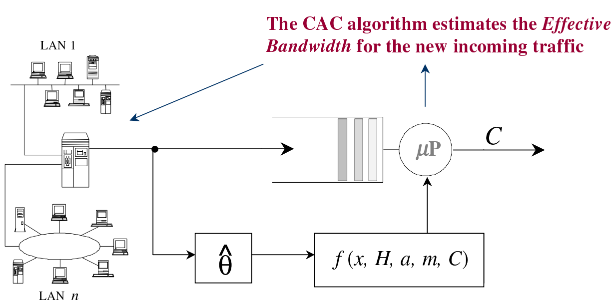

Fig. 19. High-speed network environment supporting QoS via CAC algorithm

VII. Conclusion

Definitively, the statistical properties of traffic streams in data packet transfers between client and servers machines have changed drastically. The statistical regularity in data traffic, as predicted by ubiquitous Markovian laws are highly questioned. Studies have arisen new models to characterize actual traffic patterns. The self-similar paradigm [1] has brought to researches a kind of *simplicity of expression* but a *depth of though* in traffic modeling. Nobody cannot argue against the idea that self-similarity can emulate very well the data traffic dynamics with a handful parameters (this is quite difficult with Markovian models) but the lack of a well-established fractal queueing theory turns this paradigm a hard thing to deal with. Despite this, great breakthroughs have been made to date. Although, self-similar and Markovian-based models are quite different from each other, Markovian models are still useful for research purposes, thereof to emulate both behaviors is imperative.

To this end, we have developed and showed that produce self-similar and Markovian-based models with high quality is possible. Therefore, from the results above, the Markovian-based traffic could be used when Voice over IP (VoIP) tests are being performed, this is because the ON$-$OFF model with exponential sojourn times on both states, still well suited for this type of traffic. This model fails when pure data traffic is assumed, in this case the ON$-$OFF model with heavy tails rather than exponential sojourn times either on one or both states can be used. It has been mathematically proven that many sources with such properties produce an accumulated process, $A(t)$ that, for practical purposes, could be enveloped as a fractional Gaussian process. Therefore, $A(t) = mt + \sqrt{am}\,Z(t)$ is the workload received in interval $(0, t]$, being $Z(t)$ bearing by H.

Beside this, the two simple experiments performed have shown clearly that as $H \rightarrow 1$, a finite buffer can be filled extremely fast, it suggests that is probably for self-similar traffic with a well-defined Hurst parameter between $\frac{1}{2} < H < 1$, even with low system utilization, to fill up a buffer with ease [12]. Hence LRD phenomenon might has a pervasive effect on queuing performance, as predicted by theory and software simulations.

Acknowledgments

The authors would like to acknowledge the support of this work by the International Civil Aviation Organization (ICAO), under grant project ITA-BRA006/95/802/805.

References

[1] Leland, W. E.; M. S. Taqqu; W. Willinger; D. V. Wilson. 1994. "On the Self-Similar Nature of Ethernet Traffic." *IEEE/ACM Trans. on Networking*, No 1, Feb.:1-15.
[2] Heffes, H., D. M. Lucantoni. 1986. "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance." *IEEE JSAC*, Sept. 1986.
[3] Pabrai O. Uday. 1996. *UNIX Internetworking*. Artech House, Inc., 685 Canton Street, Norwood, MA.
[4] Stallings W. 1993. *SNMP, SNMPv2, and CMIP The practical Guide to Network-Management Standards.* Addison-Wesley, Reading, Massachusetts.
[5] Samorodnitsky G. and M. S. Taqqu. 1994. *Stable Non-Gaussian Random Processes*. Chapman & Hall, One Penn Plaza, New York, NY.
[6] Beran J. 1994. *Statistics for Long-Memory Processes*. Chapman & Hall, One Penn Plaza, New York, NY.
[7] Mannersalo P. and I. Norros. 1997. "RMD-mn simulator: a fast and simple way to simulate Gaussian processes." Technical Report COST257TD(97). Finland: VTT Information Technology.
[8] Fox, R. and M. S. Taqqu. 1986. "Large-Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series." *Annals of Statistics*, Vol. 14, p. 517-532.
[9] Abry, P. and D. Veitch. 1998. "Wavelet Analysis of Long Range Dependent Traffic." *IEEE Transaction on Information Theory*, Vol. 44, No 1, p. 2-15.
[10] Popescu, A. 1999. "Traffic Self-Similarity." Technical Report, University of Karlskrona/Ronneby, Sweden.
[11] Willinger, W. and V. Paxson, V. 1998. "Where Mathematics Meets the Internet." *Notice of the AMS*, Vol. 45, No 8, Sep.: 961-971.
[12] Norros I. 1995. "On the Use of Fractional Brownian Motion in the Theory of Connectionless Traffic." *IEEE Journal on Selected Areas of Communications*, Vol. 13, p. 953-962.
[13] Norros I. 1994. "A Storage Model with Self-Similar Input." *Queuing Systems And Their Applications*, Vol. 16, p. 387-396.