Accurate and Fast Replication on the Generation of Fractal Network Traffic Using Alternative Probability Models

Stenio Fernandes, Carlos Kamienski & Djamel Sadok
Computer Science Center, Federal University of Pernambuco
CP 7851, Cidade Universitaria, Recife-PE, 50732-970
Brazil

ABSTRACT

Synthetic self-similar traffic in computer networks simulation is of imperative significance for the capturing and reproducing of actual Internet data traffic behavior. A universally used procedure for generating self-similar traffic is achieved by aggregating On/Off sources where the active (On) and idle (Off) periods exhibit heavy tailed distributions. This work analyzes the balance between accuracy and computational efficiency in generating self-similar traffic and presents important results that can be useful to parameterize existing heavy tailed distributions such as Pareto, Weibull and Lognormal in a simulation analysis. Our results were obtained through the simulation of various scenarios and were evaluated by estimating the Hurst (H) parameter, which measures the self-similarity level, using several methods.

1. INTRODUCTION

Recently, researchers identified some evidences of self-similar (or fractal) behavior in computer network traffic, as well as its severe implications in network performance [3, 10, 17, 19, 20]. Delving into the scaling phenomena, the fractal behavior could be labeled into three classification, that is when time scales go to infinity (self-similarity), go to zero (multi-fractality) and occurring over a limited range of timescales (pseudo self-similarity)[5]. Under such conditions and relevant time-scales, router's queue could work at a high level of occupancy, mainly due to presence of burst traffic in several time-scales leading to a higher end-to-end delay and packet losses. As consequence, this phenomenon could yield a low level utilization of the communication links. Therefore, an in-depth understanding of the self-similar nature in network traffic and the identification of its characteristics or implications in different scenarios and network topologies are vital for carrying out network management activities, keeping QoS assurances in suitable levels, making traffic engineering work and designing networks efficiently.

Considering a simulation environment, the performance evaluation of network protocols and mechanisms under proper conditions is significant for obtaining reliable results[8]. The selection of representative scenarios in computer networks must include the exploitation of fractal traffic.

There are some well-known analytical methods for the generation of synthetic self-similar traffic. However, due to the complexity of a physical interpretation, an alternative construction, closer to real traffic models in computer networks, is based on the aggregation and superposition of On/Off [1] sources, which activity and/or inactivity periods follow a heavy tailed probability distribution function (PDF).

There are related studies concerning the aggregation of heavy tailed sources and self-similar traffic. Some of them present results and analysis of traffic measurement in real networks[20], whereas others focus on purely statistic perspectives [7, 11]. Currently, a limitation in using such techniques is that there is no wide and effective self-similar traffic evaluation that could be confidently used to customize and parameterize simulation scenarios. For that reason, his paper analyses the trade-off between accuracy and computational efficiency on the generation of fractal traffic. The precision is determined by evaluating the error between a target Hurst parameter (usually used to measure the self-similarity level) and its actually estimated value from the traffic sample collected during a simulation experiment. The computational efficiency is related to the processing time needed to obtain a previous chosen precision. We present some results obtained from a simulation-based study where several distinct scenarios were evaluated in a number of experiments. The evaluation includes Lognormal, Weibull and Pareto PDF including several form parameters used to characterize the heavy tail behavior of the On/Off sources. In spite of Pareto PDF being frequently used to generate heavy tailed traffic, our results show that Weibull and Lognormal PDF require a smaller quantity of simultaneous sources in order to obtain the same precision as Pareto. In general, the results are significant in view of the fact that there is no need to inflate the aggregation of such sources since this decision could lead to unacceptable higher simulation processing times.

An important feature on the generation of self-similar traffic is associated to the relation between the form (shape) parameter of the heavy tailed PDF and the Hurst parameter H. For instance, for the Pareto PDF, analytical and empirical procedures show the relation $a = 3 - 2H$ [3], where a is its form parameter.

This paper is organized as follows. Section 2 presents fundamental theory about the self-similar phenomenon and its implications in network performance and it also describes some techniques to estimate its level. Some methods for generating synthetic self-similar traffic are presented in section 3. Section 4 describes the simulation environment and scenarios and section 5 presents the simulation results and comments. Finally, in section 6 we have concluding remarks and further discussion for future works.

2. SELF-SIMILARITY

The main concept related to self-similarity or general fractal behavior consists of the phenomenon of preserving the major characteristics of an entity in nature when observed in distinct time or space scales[2]. Particularly, in the case of stochastic objects such as the time series (e.g., computers network traffic), the self-similar behavior exhibits the same structural properties in several time scales. Without a suitable strong statistical approach, one should now assume if a realization of a stochastic process is aggregated in distinct time scales and keep its most important statistical properties (e.g., first- and second order moments), it is considered a fractal process.

2.1. Self-Similar Process

Let $X(t)$ be a strict-sense stationary time series, with mean m, variance s^2 and autocorrelation function $r(t)$. Additionally, let $X^m(t)$ be a new time series obtained from $X(t)$, through averaging it in non-overlapping blocks of size m. In other words, the aggregated series has the form $X^m(t) = (m^{-1})(X_{tm-m+1} + X_{tm-m+2} + \cdots + X_{tm})$ and $r^m(t)$ is the autocorrelation function. The process $X(t)$ is considered self-similar if $r^m(t) = r(t)$ for any $m = 1,2,3,\ldots$. In particular, if the autocorrelation function has the form $r(t) \rightarrow t^b L(t), t \rightarrow \infty$, where $L(t)$ is slowly varying at infinity, one could say that it is a self-similar process with a Hurst parameter H. The relation between the Hurst parameter and the decaying rate of autocorrelation function b is $H = 1 - b/2$. This kind of process exhibits Long-Range Dependence (LRD), which implies the autocorrelation function is not limited, that is $\sum_t r(t) \rightarrow \infty$. Another important property is related to the variance of the aggregated series that has a slow decrease as the aggregation level increases. Such characteristic could be used to estimate the self-similar level of a stochastic process. There are evidences that the LRD feature is firmly associated to heavy tailed behavior of the generating process. Additionally, the superposition of several independent heavy tailed sources yields self-similarity [20]. We give more details in section 3.

2.2. Self-Similar Traffic and Network Performance

Several empirical and analytical studies show evidences related to the phenomenon of self-similar in computer network traffic [3, 14, 17,20]. Some approaches show that aspects such as file sizes in Web servers and file transfer times under HTTP, cause unfavorable impact in network performance. Such characteristics yield traffic bursts in several time scales, which make difficult the determination of efficient algorithms of congestion control, admission control and traffic prediction[18]. For instance, in the presence of LRD traffic, increasing queue lengths do not produce fewer packets loss rates [6], as would be expected for traffic with short-range dependence. Besides, performance is seriously affected due to the high concentration of congestion periods and significant increase in queue delays[14]. Therefore, the traditional traffic source models, such as Poisson and Exponential PDF, which superposition does not exhibits self-similarity, must be replaced for more accurate models in order to obtain reliable simulation results[3]. For this reason, usual performance metrics, such as throughput, delay, jitter, packets loss and queue lengths, must be evaluated taking into account these evidences as a support for obtaining coherent results.

2.3. Estimation Techniques of Self-Similar Processes

As we have shown before, the Hurst parameter determines the self-similarity level of a time series. If H is in the [0.5, 1] range, there is a clear indication of the presence of self-similar behavior. In addition, H values closer to the unity point out a high self-similarity level. There are a number of methods to estimate the H parameter, which could be classified in heuristic and by inference ones. Heuristic methods are mainly useful as simple diagnostic tools and the best-known one is the analysis of the rescaled range R/S statistic. Other techniques include the log-log correlogram, the log-log plot of the variance of the aggregated processes versus the aggregation level, least squares regression in the spectral domain and inference by maximum likelihood estimation in the time and spectral domain (Whittle's estimator)[2].

In order to exemplify some self-similarity level estimation methods, we briefly describe the R/S statistic and the variance techniques. The R/S statistic is related to the H parameter by $E[R(n)/S(n)] \approx cn^H$, when $n \to \infty$ and c is constant and independent of n. It is easy to notice that $\log(E[R(n)/S(n)]) \cong H\log(n) + \log(c)$. This equation has the form $y = a + bx$ and consequently H could be estimated by linear regression, where $\hat{H} = \hat{b}$. Using the variance approach, the relation between the logarithm of the variance of the aggregated process $X^{(m)}$ and the block size m has the form $Var(X^{(m)}) \approx am^{-b}, m \to \infty$. As a result, $\log[Var(X^{(m)})] \approx -b\log(m) + \log(a)$ and H could also be estimated by linear regression that determines the negative slope b with $\hat{b} = 2 - 2\hat{H}$.

3 SELF-SIMILAR TRAFFIC GENERATION

Due to the importance of the fractal behavior in a number of areas (e.g., economy, telecommunications), several formal analytical models have been proposed which most of them are useful for generating such sequences. Some of them rely on Fractional Autoregressive Integrated Moving Average (FARIMA) processes[11], Fractional Gaussian Noise (FGN) and Wavelets[7]. We are also aware that the use of Fractional Brownian Motion (FBM) models[2] is a powerful tool to generate fractal traces with a high level of accuracy. However, using these approaches lead to difficulties to get some sense for network engineers and computer scientists.

In order to address this issue, an alternative proposal that has an authentic meaning to real networks is based on the aggregation and superposition of Renewal Rewards Process (On/Off)[16], which activity (On) and inactivity (Off) periods follow a heavy tailed PDF. We also know that there are some drawbacks associated to the deployment of this technique in network simulation procedures. Roughan, Yates and Veitch[9] alert about pitfalls in choosing time-scales of interest and number of samples. Despite these issues, this approach could allow an immediate use of widespread network simulation tools, such as Network Simulator 2 – ns2 or the software family from OPNET[13] since there is no need to extend their libraries to support such analytical models.

The M/Pareto process, also known as Poisson Pareto Burst Process – PPBP[1], is an excellent model that could be used for precise self-similar traffic generation. At the same time, it maintains the understanding of the physical process existing in local or wide area networks. The M/Pareto is a process composed of a number of overlapping bursts. Bursts arrive following a Poisson Process with rate l and have a Pareto distributed duration. Increasing l may be considered an increase in the level of activity of individual sources or in the number of sources. Each burst has a constant rate r and its length has the form $P_r(X > x) = 1 - F(x) = x^{-a}/d, x \geq d$, with $1 < a < 2, d > 0$, where d is the scale parameter. It's easy to verify that mean amount of work arriving in the PPBP model is $m = (lrda)/(a-1)$. It also is asymptotically self-similar with H parameter $H = (3 - a)/2$, where a is the form parameter of the Pareto PDF.

4. SIMULATION CONFIGURATIONS

The analysis of self-similarity undertaken in this paper involves the simulation of aggregation of On/Off traffic sources, traffic measurement and estimation of the Hurst parameter. As the simulation platform, we used the Network Simulator 2 (ns-2)[12] Different scenarios were simulated, varying the number of sources into the aggregation, the heavy tailed distribution, and its shape parameter, that defines the tail size.

Figure 1 depicts the topology used in our simulations. It consists of a variable number of traffic sources (S1, S2… Sn), connected to a router that in turn is connected to a destination node D1. All links were configured with a fixed capacity of 10 Mbps. In spite of its simplicity, this topology is able to yield self-similar traffic at the destination node D1, thus making it possible to significantly decrease the time required for running the simulation (compared to a more complex scenario).

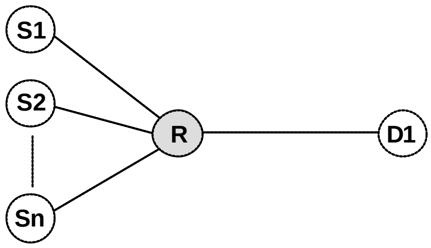

Figure 1 – Simulation topology

The numbers of traffic sources used in different scenarios were 1, 2, 5, 10, 20, 50, 100 and 1000, since one goal of this paper is to quantify the number required for an accurate self-similar traffic generation. Source average aggregate rate was set to 2 Mbps. During the activity periods (On), each source sends data at a rate of $4/n$ Mbps, where n is the number of simultaneous sources. On and Off average duration times were both set to 50ms, according to the Pareto, Weibull and Lognormal distributions, thus yielding an average rate of $r = 2/n$ for each source.

This model is frequently used for generating self-similar traffic, and it is comparable to the M/Pareto model. In our On/Off model, the number of sources is fixed for each scenario, while each source sends several bursts with random duration. On the other hand, the M/Pareto model uses a random number of sources (a Poisson process), but each source generates only a single burst with random duration. As an illustration, for the Pareto distribution (also used in the M/Pareto model) the average traffic level generated by this process is $m = nr$. In this case, the a parameter is implicitly used in the computation of r.

For each simulated scenario, 1 million traffic samples were collected at the destination node D1. Each sample corresponds to the average throughput during the sampling period of 100ms. For each set of 1 million traffic samples, the Hurst parameter was estimated through the R/S and variance methods (section 2.3). Some results (precisely identified) presented in section 5 were obtained through the execution of 30 replications for each scenario (Monte Carlo's simulation). In such cases, results represent the mean of the 30 replications and also 99% asymptotic confidence intervals were calculated. The estimation methods were implemented in the R system (version 1.6.1)[15] that is a free-software famous for its accuracy and efficiency.

5. SIMULATION RESULTS

This section presents results obtained from simulation experiments and Hurst parameter estimation performed in this study. One goal was to quantify the exact number of superposed sources required for yielding self-similar traffic with controlled accuracy and variation, for the Pareto, Weibull and Lognormal distributions. Another goal was to perform an exploratory study with the Weibull and Lognormal distributions concerning the generation of self-similar traffic, so that to characterize a relation of its shape parameter with the Hurst parameter estimated from the traffic samples. These distributions are frequently pointed out as having heavy tails, but so far we failed to find a reference with a scenario configuration and experiment results.

In general, it is basically believed that the aggregation of traffic generated by a large number of superposed sources produces self-similar traffic. An important question here is to determine the required number of sources wherefrom trustworthy results can be obtained. The larger the number of sources, the higher the accuracy is, since this behavior is asymptotic. However, it also implies a higher processing time for running the simulation. Therefore, knowing which number of sources is sufficient is a significant information for researchers, because complex simulations (with many replications) can easily consume hours, days, weeks or months of the processing time of machines with very reasonable capacity.

5.1 Pareto Distribution

Figure 2 shows the experiment results for the Pareto distribution. Pareto has a heavy tail when its shape parameter is between 1 and 2. For this study, we used the values 1.1, 1.2, 1.3, 1.4, 1.5, 1.6 and 1.7 corresponding to the target (intended) Hurst parameters of 0.95, 0.9, 0.85, 0.8, 0.75, 0.7 e 0.65, respectively. It is commonly accepted that from 0.7, traffic could cause some harm to the performance of a network[17]. As a general result, data show that the known relation

$a = 3 - 2H$ holds for the experiments we performed. For all values of the target H, the estimated H stabilizes from 10 sources, incurring in a maximum error of 7.5 %, which is coherent with[7].

It can be observed that from 50 sources on, for all shape parameters, there is no significant increase in the accuracy of the estimated Hurst parameter. It could be expected that from this number the processing time will start to become noticeable. A reasonable conclusion is that the computational cost associated to increasing the number of sources beyond 20 does not compensate the additional benefit in accuracy. Additionally, when the target Hurst parameter is enough high, e.g. starting from 0.75, the simulation results suggest there is no need to parameterize the model with more than 5 sources.

The largest interval was obtained for 10 sources. As an illustration, the mean for the 30 replications is 0.8, whereas the values of the estimated H are spread between 0.72 and 0.88. These arguments corroborate the conclusion that 20 sources is an acceptable trade-off between accuracy and computational cost for the Pareto distribution.

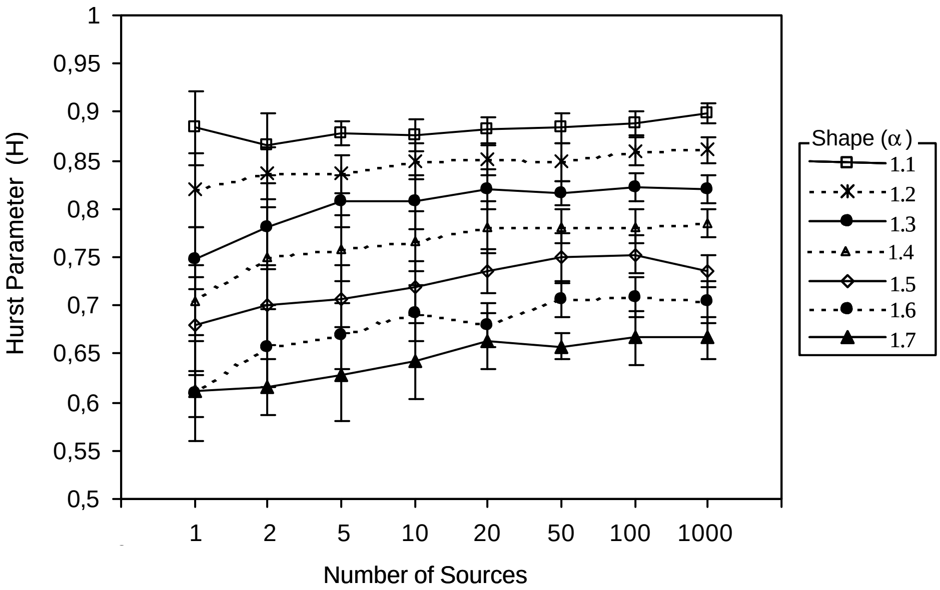

Figure 2 – Pareto with several shape (α) parameters

5.2 Weibull Distributi on

Figure 3 presents the results obtained in our evaluation for the Weibull distribution, with the shape parameter α varying between 0.05 and 0.2. For α = 0.2, it can be observed that the estimated H parameter is around 0.62, indicating that the traffic presents a low self-similarity level. We also simulated scenarios with values of α between 0.2 and 1, but no significant variations in the estimated Hurst parameter were obtained. In other words, the Weibull distribution has a known L shape for α values between 0 and 1, though only between 0 and 0.2 it is able to produce self-similar traffic, according to Figure 3.

Results obtained for the Weibull distribution differ from those obtained for Pareto, mainly related to stability and variability issues. Taking a visual comparison between figure 2 and 3, it could be presumed that using Weibull will present better reliability, represented by the narrower size of the confidence intervals, to obtain the target Hurst with a small number of sources. For the Weibull distribution both means of the Hurst parameter and confidence intervals presented a steady behavior. This fact points toward that from a few superposed Weibull traffic On sources the generation of self-similar traffic has a substantial statistical reliability.

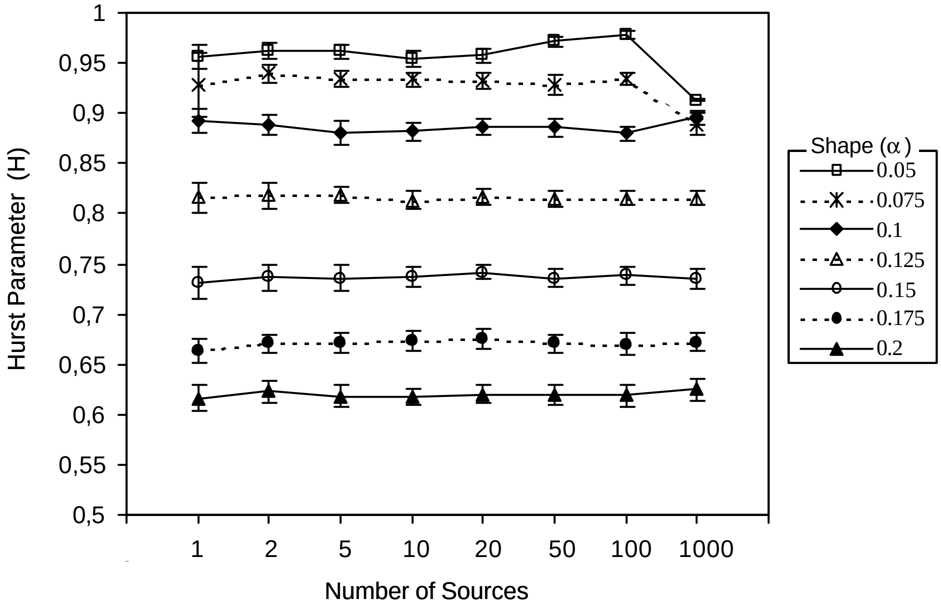

Figure 3 – Weibull with several shape (α) parameters

5.3 Lognormal Distribution

The results obtained for the Lognormal distribution are shown in Figure 4. The experiments with the Lognormal distribution were performed varying the s parameter (the standard deviation) between 2 and 5. Sometimes the standard deviation is called the shape parameter, because it determines the size of the distribution tail. When s = 5, the Lognormal distribution presents a heavy tail, and consequently generating a high level of self-similarity in the aggregated traffic. The results suggest an intermediate behavior in stability and variability as compared to Pareto and Weibull dis tribution. Similarly to Weibull distribution, the estimates present low variability, which can be observed by the narrow confidence intervals in Figure 4.

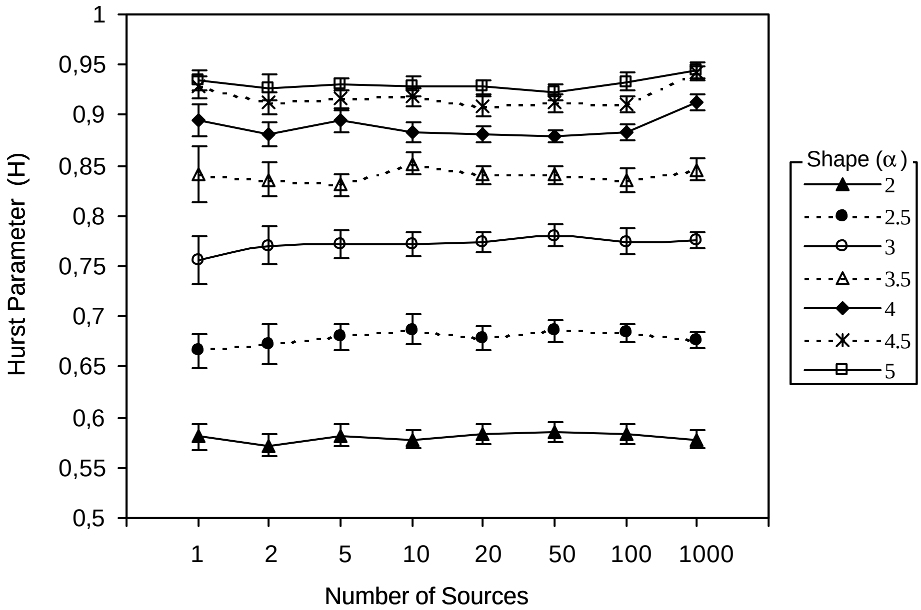

Figure 4 – Lognormal with several shape (s) parameters

5.4. Simulation Time

Figure 5 presents the simulation processing times considering Pareto, Weibull and Lognormal distributions associated to the number of sources. It could be noticed a significant increase in processing times as the sources aggregation augments from 100 to 1000.

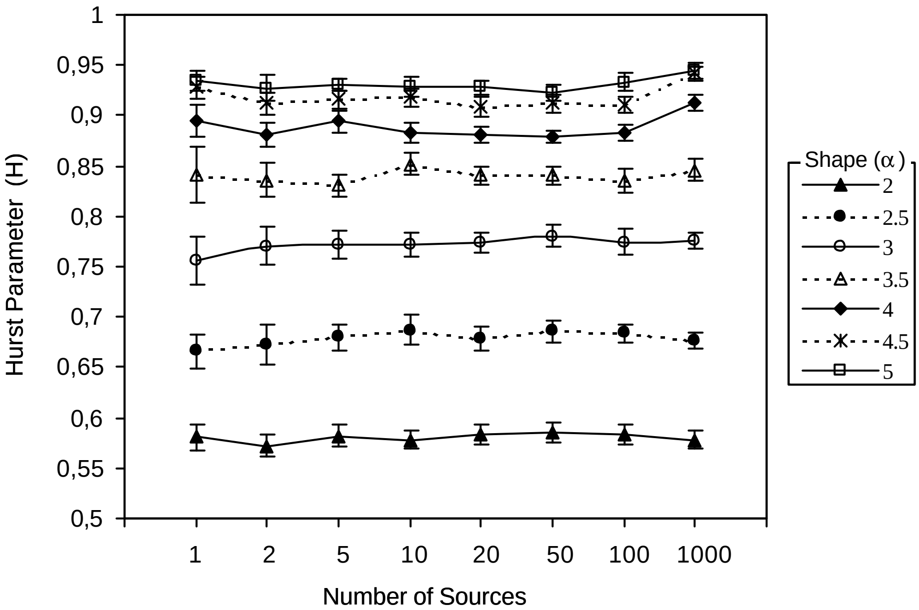

Figure 5 - Simulation Times

In addition, the simulation times variation associated to the number of sources and the probability distribution functions is better described in the Table 1 This table shows the processing times as an increment related to only one source and the previous number of sources. This could be used to quantify the computational effort linked to an increase in the source aggregation.

Table 1 - Simulation Time Variation

Increase (in %) related to:		**Number of sources**							
		1	**2**	**5**	**10**	**20**	**50**	**100**	**1000**
Pareto	1 source	-	31	59	63	69	67	90	707
	Last	-	31	21	3	4	-1	14	325
Weibull	1 source	-	32	62	70	81	94	121	296
	Last	-	32	23	5	7	7	14	79
Log normal	1 source	-	29	61	69	79	87	111	349
	Last	-	29	24	5	6	4	13	112

5.5. Aggregated Throughput

Along with the Hurst parameter, the measured throughput is another parameter that can tell us how accurate is the traffic generated for each simulated scenario. As described in section 4, the traffic aggregate was generated with a target rate of 2 Mbps. Figure 6 show the results collected for the Pareto distribution. For the number of simultaneous sources varying from 1 through 50, the maximum error was 4.5%, except when the shape parameter is 1.1 (that was 9.7%). As the graph show us, the measured throughput achieved higher values for the scenarios with 100 and 1000 simultaneous sources, yielding errors of up to 24,4% and 82,3%, respectively. This observation corroborates with previous conclusions that no more than 50 sources are desirable.

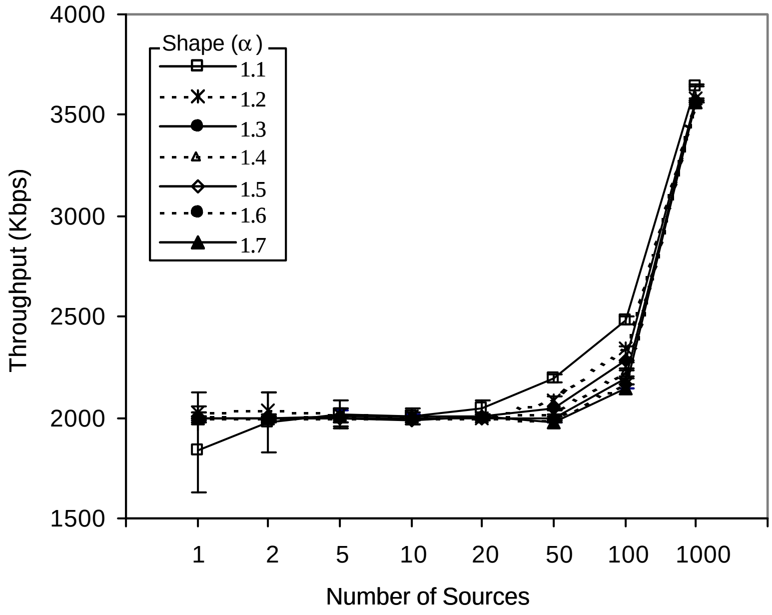

Figure 6 – Throughput for the Pareto Distribution

Results for the Weibull distribution are shown in Figure 7. It can be observed that when the shape parameter is 0.05, the measured throughput presented higher variability, and this explains the large confidence intervals. Furthermore, the error was also unacceptable for this scenario, achieving up to 94%. Unlike the Pareto distribution, the error was low only for the scenarios that used less than 50 sources.

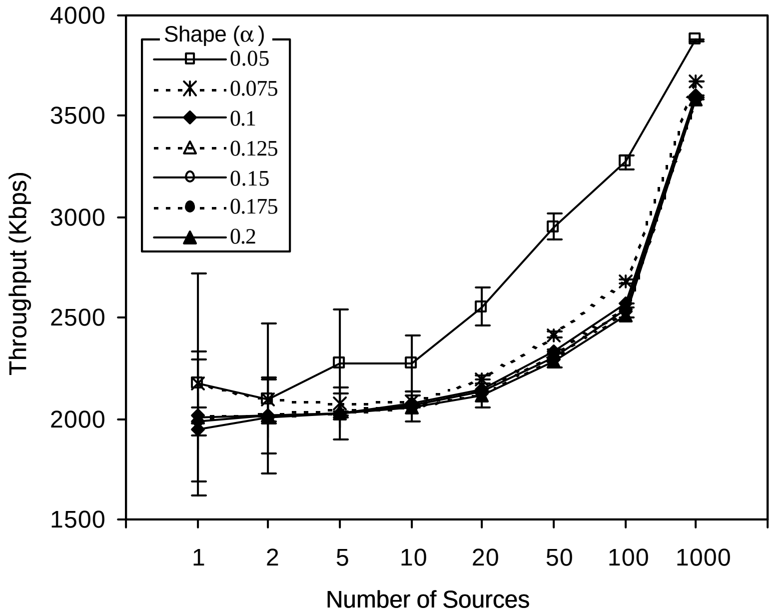

Figure 7 – Throughput for the Weibull Distribution

Finally, Figure 8 shows results of the measured throughput for the Lognormal distribution. This distribution presented the lowest variability among all scenarios simulated. In the same way as for the Weibull distribution, the error was low only for the scenarios that used less than 50 sources (less than 10%).

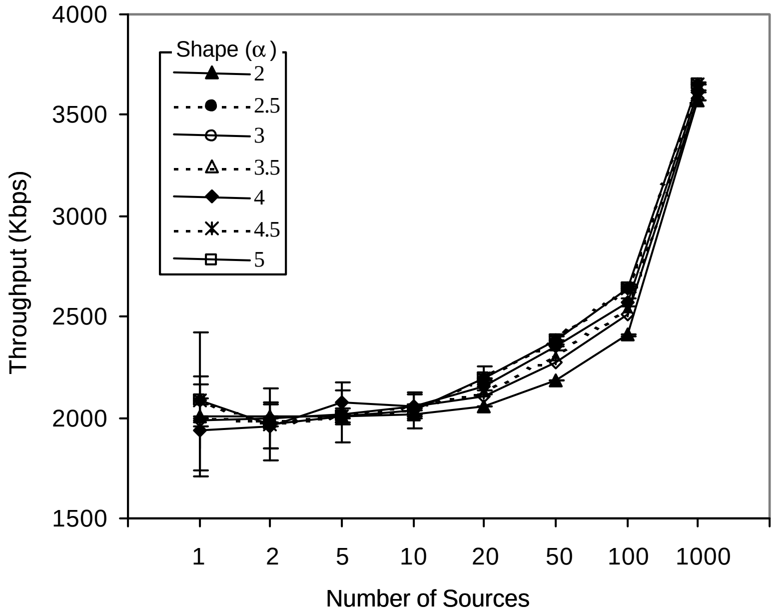

Figure 8 – Throughput for the Lognormal Distribution

6. CONCLUSION

Generating synthetic self-similar traffic through the aggregation of superposed sources is of great importance for undertaking simulation in the computer network area. Unlike other known analytical methods, the superposition of On/Off traffic source has an explicit explanation, since sources represent entities from the real world. The enlightenment for this method is the well-known relation between heavy-tailed burst times, long-range dependence and self-similarity. Another advantage of using superposed sources is the simplicity of generating self-similar traffic in network simulators that do not offer any built-in analytical generator, e.g. the ns-2 simulator that we used in our evaluations.

This paper showed a broad analysis of the parameterization employed for generating self-similar traffic with the Pareto, Weibull and Lognormal distributions. These three distributions were selected due to their ability to be parameterized in such a way to present heavy tails. They are frequently used to characterize the existence of heavy tailed distributions in the Internet. Pareto is more common, but Weibull and Lognormal also can be used for generating self-similar traffic, with the additional benefit of requiring a smaller number of superposed sources. Our results revealed that, for the evaluated scenarios, the best trade-off between accurate estimation of the Hurst parameter and efficient utilization of computational resources is around 20 Pareto and 10 Weibull (and Lognormal) On/Off sources.

Furthermore, this paper also provides values for the shape parameters whereto the Weibull and Lognormal distributions yield self-similar traffic. Consequently, they can be used in other related work.

As future work, we intend to find out analytical relations between the shape parameters of the Weibull and Lognormal and the Hurst parameter (such relation there exists for Pareto). In addition, we are working on a tractable and flexible heavy-tailed distribution able to model a variety of aspects of the self-similar traffic.

We would like to thank CNPq for the financial support.

REFERENCES

1. Addie, R.G., Neame, T.D. & Zukerman, M., "Performance Evaluation of a Queue Fed by a Poisson Pareto Burst Process", *Computer Networks*, **40** (3), p. 377-397, Oct. 2002.
2. Beran, J., *Statistics for Long-Memory Processes*, Chapman & Hall/CRC, 1ª ed., New York, 1994

3. Crovella, M. E. & Bestravos, A., “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes”, *IEEE Trans. Networking*, **5** (6), Dec. 1997.
4. Downey, A. B., “Evidence for Long-Tailed Distributions in the Internet”, ACM SIGCOMM IMW 2001, Nov. 2001.
5. B. Sikdar, K. Chandrayana, K. Vastola and S. Kalyanaraman, “On Reducing the Degree of Second-Order Scaling in Network Traffic”, Proceedings of IEEE GLOBECOM, Taiwan, November, 2002.
6. Grossglauser, M., Bolot, J.C., “On the Relevance of Long-Range Dependence in Network Traffic”, *IEEE Trans. Networking*, Vol. 7, p.629-640, Oct. 1999
7. Jeong, H.-D. J., McNickle, D. & Pawlikowski., K. "Fast Self-Similar Teletraffic Generation Based on FGN and Wavelets". IEEE ICON'99, Sept. 1999
8. Kamienski, C. A., et al., “Simulating the Internet: Applications in Research and Education”, *XXI JAI’2002* (in portuguese), Jul. 2002.
9. M. Roughan, J. Yates and D. Veitch, "The Mystery of the Missing Scales: Pitfalls in the Use of Fractal Renewal Processes to Simulate LRD Processes," Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics, 1999.
10. Leland, W., Taqqu, M., Willinger, W., & Wilson, D., “On the Self-Similar Nature of Ethernet Traffic”, *IEEE Trans. Networking*, **2** (1), Feb. 1994.
11. Liu, J., Shu, Y., Zhang, L., Xue, F. & Yang, O. “Traffic Modeling Based on FARIMA Models”, IEEE 1999 Canadian Conference on Electrical and Computer Engineering, May 1999.
12. Network Simulator (version 2.1b9a), http://www.isi.edu/nsnam/ns/, 2002.
13. OPNET Technologies Inc., http://www.opnet.com.
14. Park, K. Kim, G. & Crovella, M.E. "On the Effect of Traffic Self-similarity on Network Performance," SPIE International Conference on Performance and Control of Network Systems, Nov. 1997.
15. R System (v. 1.6.1), http://www.r-project.org, 2002.
16. Ross, S. M., *Introduction to Probability Models*, Academic Press, Inc., 5ª ed., London, 1993.
17. Sahinoglu, Z. & Tekinay, Sirin, “On Multimedia Networks: Self-Similar Traffic and Network Performance”, *IEEE Comm. Mag.*, **37**(1), Jan. 1999.
18. T. K. Lee, M. Zukerman and R. G. Addie, "Admission control schemes for bursty multimedia traffic", IEEE INFOCOM 2001, Apr. 2001.
19. Veres, A., Kenesi, Zs., Molnar, S. & Vattay, G., “On the Propagation of Long-Range Dependence in the Internet”, SIGCOMM´2000, August 2000.
20. Willinger, W., Taqqu, M.S., Sherman, R. & Wilson, D.V., “Self-Similarity Through High-Variability: Statistical Analisys of Ethernet LAN Traffic at the Source Level”, *IEEE Trans. Networking*, **5**(1), p. 71-86, 1997