WEBCASTING AND WEB NETWORK TRAFFIC MEASUREMENTS AND MODELING WITH α-STABLE SELF-SIMILAR PROCESSES

Anestis Karasaridis and Dimitrios Hatzinakos

University of Toronto
Dept. of Electrical and Computer Engineering
Toronto, Ont. M5S 3G4, Canada
email: {anestis,dimitris}@comm.toronto.edu

ABSTRACT

In this paper, we present and discuss measurements of Webcasting and aggregated Web traffic in our research group's local area network. The Webcasting traffic is multiplexed to simulate the effect of having many clients running simultaneously webcasting software in the background. The multiplexed Webcasting and the aggregated Web traffic appear to be asymptotically self-similar. The α-stable self-similar stochastic process, originally proposed in [5] to model aggregated Ethernet LAN and WAN traffic, is applied to the new measurements and the results, implications and extensions are discussed.

1. INTRODUCTION

In the past three years, World Wide Web (WWW or simply web) applications have become ubiquitous. As a result, network traffic related to web applications has increased considerably to a conservative estimate of 30% of total network traffic. Therefore, accurate and robust models for this particular type of traffic would be essential for resource allocation and evaluation of network performance, and for future design of networks and protocols.

Among those applications of the web that are expected to attract considerable interest in the next few years is webcasting. Webcasting is the automatic transmission of web content from the server to the client without any intervention of the user. This type of traffic is inherently different than traffic created by the user "surfing" the net and downloading information by request. It is also expected to dominate the traffic, since it involves information broadcasting to a large number of users and also because it becomes the popular choice of people for frequently updated web content. Finally, webcasting will take advantage of the user's idle time to download information relevant to the user's interests, and therefore the related traffic is expected to be very bursty.

Recent network traffic measurements over Local Area Networks (LANs) [6], Wide Area Networks (WANs) [8], and of source level Variable Bit Rate Video (VBR) traffic [2], have revolutionized the field by showing that the actual traffic does not account for the simple widely accepted models such as the Poisson or Modulated Markov processes. There is also an indication that the traffic generated by the transmission of web content does not follow the traditional models at the peak times of utilization by the users [3]. In contrast, real network traffic exhibits burstiness over a very large range of time-scales and long-range dependence, which can be accounted by using statistical self-similar distributions.

Since the two main aspects of high speed network traffic are impulsiveness and self-similarity, there is a need for models that can capture both of these properties in a unified and parsimonious fashion. The problem is currently addressed from two different directions: The first one is the use of heavy-tailed distributions (e.g. Pareto) to account for the impulsiveness [6], and the second is the use of self-similar processes (e.g. fractional Gaussian noise) to account for the statistical self-similarity of the data [7]. While these approaches give better results than the simple Poisson or Compound Poisson models, they fail to unify the desired model properties.

On the other hand, α-stable self-similar processes can capture both the impulsiveness, since the underlying distribution is heavy-tailed, and the self-similarity. Furthermore, they provide a physical interpretation of how the observed data appear as the superposition of independent effects according to the Generalized Central Limit Theorem [4].

The rest of the paper can be summarized as follows: In section 2.1 we present measurements of traffic on the client site, using the Pointcast Network [9], as the webcasting source. Section 2.2 deals with the multiplexing of the traffic measured, to simulate the effect of having a number of clients in the local network running webcasting software in the background. In section 2.3, measurements of web traffic on our local network are provided and analyzed. The α-stable self-similar stochastic process which was originally proposed in [5] as a model for aggregated Ethernet LAN and WAN traffic, is presented in section 3. The methodology to estimate the parameters of the suggested model along with simulations are presented in section 4. By applying the Rescaled Adjusted Range Statistic (or R/S) we show that both the multiplexed webcasting and the aggregated web traffic are asymptotically self-similar. We provide results from simulations, where we synthesize traffic by using the proposed model. We summarize the conclusions of the paper and future research directions in section 5.

2. NETWORK TRAFFIC MEASUREMENTS

In the following three subsections, we provide measurements taken in the Communications Group's LAN during non-overlapping times of the day during a span of one week in June 1997. The first two subsections describe measurements and multiplexing of webcasting traffic, while the next subsection presents measurements of web traffic. Information about the size, date and time that the data were captured, is provided in tables 1 and 2.

2.1. Webcasting traffic

To measure webcasting traffic we used the following experimental setup: A networked station (Windows95-PC) was dedicated to run the Pointcast Network [9] on the background (we will refer to it as the client). Another station (Solaris-SparcStation) was used to measure the traffic addressed to the Pointcast client, coming from the www application port. The last condition ensures that we measure only the web traffic addressed to the client. No web browsing was allowed on the Pointcast client to isolate only the webcasting traffic.

The measurements consist of 10 binary files, each one containing information about 1000 Ethernet packets which arrived at the client during different times of the day. From those files, we extracted only the time-stamps and the lengths of the Ethernet packets. From the time-stamps we constructed packet-count plots, which are depicted in figure 1.

The installation of the Pointcast client was typical, with the default number channels and subchannels. Figure 1 illustrates that the updates are periodic and occur approximately every hour.

2.2. Generation of aggregate webcasting traffic

In this work, we are interested in modeling the aggregate webcasting traffic. As mentioned above, measurements presented in figure 1 correspond to traffic addressed to a single client. We are interested to observe how the network traffic patterns change as the number of stations receiving webcasting content increases.

We assume that the server does not support any sort of multicasting, and therefore that the information transmitted from the server to the client is point-to-point. This assumption agrees with the current implementation of most of the webcasting software available. This is also supported by the fact that the information distributed from the server to the client could be entirely personalized.

FILENAME	SIZE(pkts)	DAY	Tstart	Tstop
cap-1.bin	1000	Jun 19	9:32P	3:03A
cap-2.bin	1000	Jun 19	12:42P	4:25P
cap-3.bin	1000	Jun 20	1:22P	6:28P
cap-4.bin	1000	Jun 20	4:25P	9:31P
cap-5.bin	1000	Jun 20	3:25P	8:30P
cap-6.bin	1000	Jun 23	9:55A	1:58P
cap-7.bin	1000	Jun 23	3:49A	8:55A
cap-8.bin	1000	Jun 23	6:52A	10:56A
cap-9.bin	1000	Jun 24	9:09A	12:28P
cap-10.bin	1000	Jun 24	7:19P	11:23P

Table 1: Information about the webcasting traffic measurements for one client.

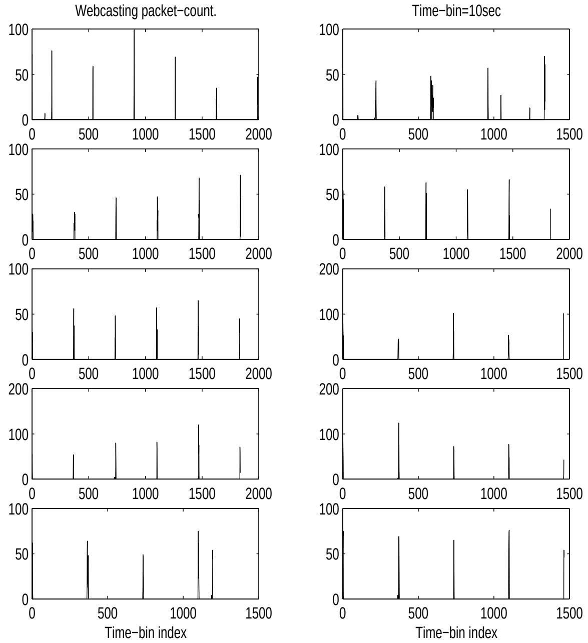
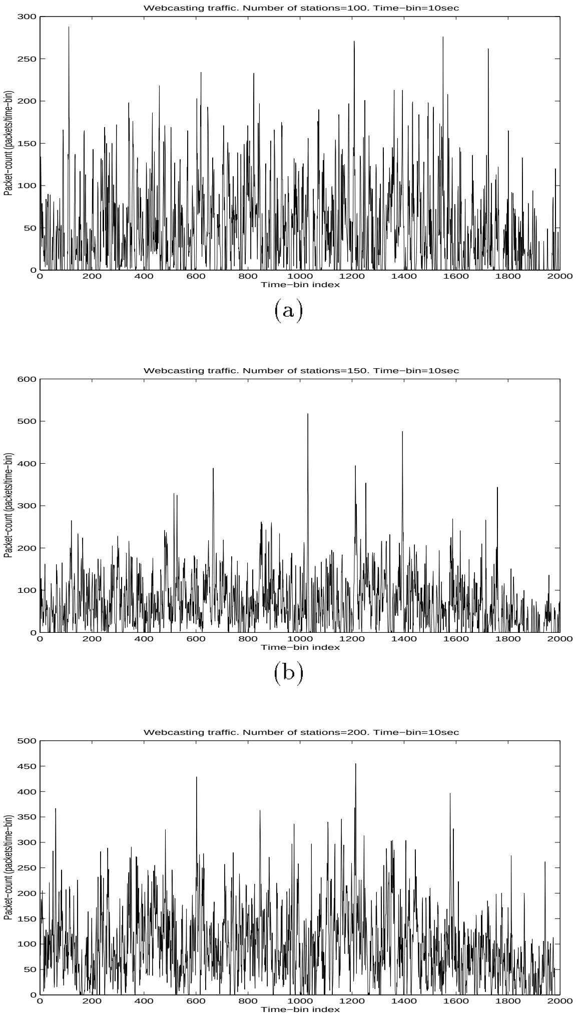


Figure 1: Webcasting measurements for one client over different times of the day. Traffic is represented by packet-counts. Each plot gives the packet-count of a record of 1000 Ethernet packets. Data were captured on the dates and times shown in table 1.

FILENAME	SIZE(pkts)	DAY	Tstart	Tstop
cap-weball1.bin	100000	Jun 19	4:30P	11.18P
cap-weball2.bin	72088	Jun 20	9:00A	12:41P
cap-weball3.bin	100000	Jun 20	1:00P	6:00P
cap-weball4.bin	100000	Jun 23	9:00A	1:16P
cap-weball5.bin	31962	Jun 23	3:00P	4:45P
cap-weball6.bin	100000	Jun 24	11:00A	3:52P

Table 2: Information about the aggregated web traffic measurements.

To simulate the effect of having a varying number of client stations in the local network requesting personalized information in the background, we used the set of actual measurements in the following way: for each station $k = 1, 2, \ldots, N$ in the network we pick randomly, in a uniform fashion, one of the actual measurements delayed by a random amount of time $t_o = Uniform[0, T)$, where T is the period of content updates for each client. The overall traffic is generated by multiplexing the N data records. Plots of packet-counts generated using the above method are shown in figure 2 for different numbers of client stations $N = 100, 150$ and 200.



(a)

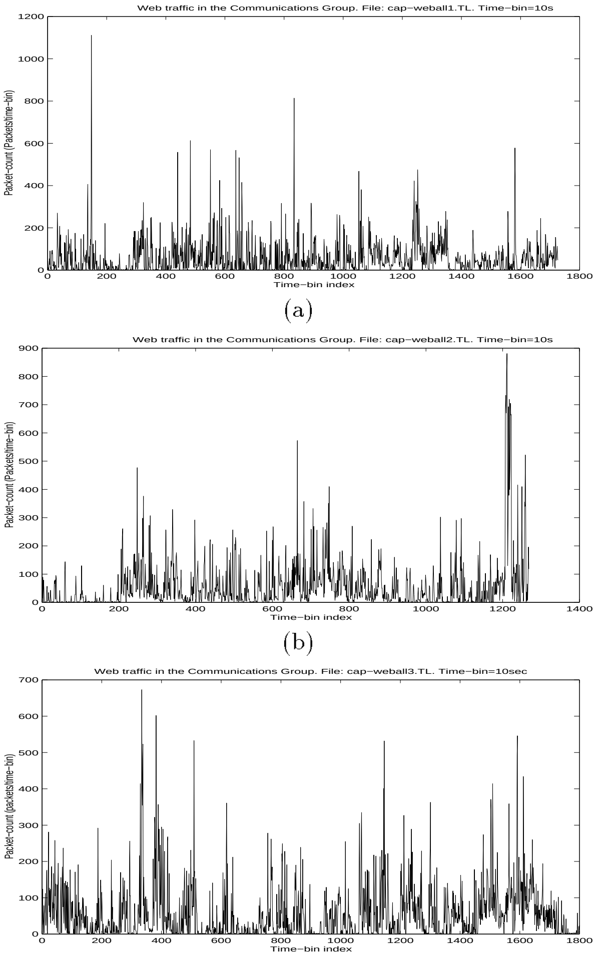
(b)

(c)

Figure 2: Aggregate webcasting traffic generated by multiplexing randomly delayed actual measurements. The number of stations is 100 in (a), 150 in (b) and 200 in (c).

2.3. Web traffic

A SparcStation was used to collect web traffic measurements over our research group's LAN. Web traffic was isolated from general Ethernet traffic by restricting the captured packets to have www destination or source port number. Our local network consists of approximately 120 workstations and one web server. We captured six segments of the traffic over different times of the day in a span of one week, each consisting of information for 100,000 packets. As in the case of webcasting measurements, we extracted the time-stamps and the lengths of the Ethernet packets and then calculated the corresponding packet-counts. Plots of the packet-counts of the 6 records are shown in figures 3 and 4.



(a)

(b)

(c)

Figure 3: Part I web traffic measurements over the Communications Group's LAN. Plots correspond to the first three files shown in table 2.

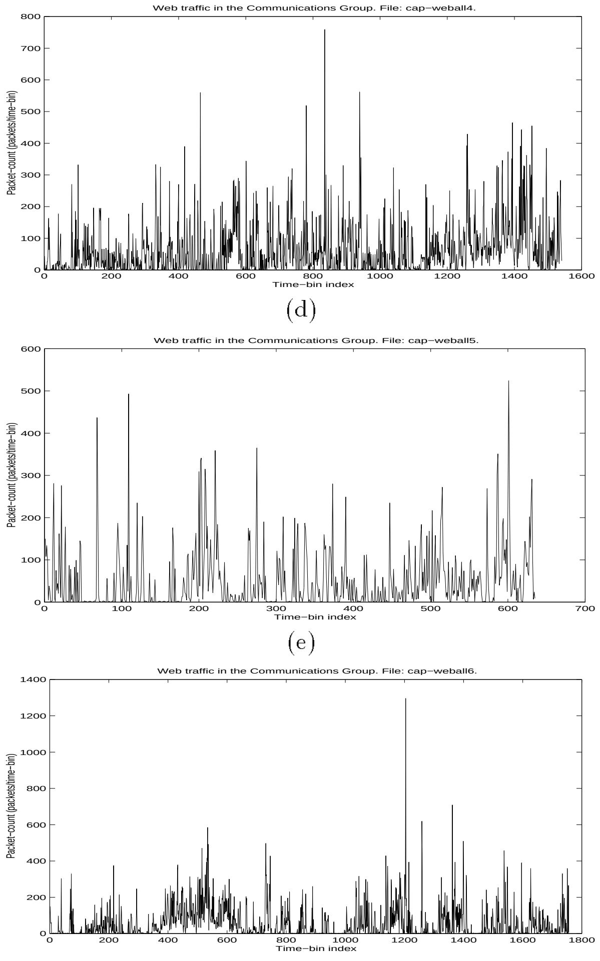
3. MODELING USING α-STABLE SELF-SIMILAR PROCESSES

Our proposed model M, is defined as follows:

$$M(i) = c_1 \cdot L_{\alpha,H}(i) + c_2, \quad c_1, c_2 \in \mathcal{R}^+, \quad i \in \mathcal{Z}^+, \qquad (1)$$

where c_1 and c_2 are positive real constants and $L_{\alpha,H}(i)$ is Linear Fractional Stable Noise (LFSN) [10], with $\beta = 1, \sigma = 1$, $\mu = 0$ and $H > 1/\alpha$ to ensure long-range dependence. This model was first proposed in [5] to capture the statistical behavior of aggregate WAN Ethernet traffic measured at Bellcore [6].

Since $\beta = 1$, the LFSN process is *totally skewed.* This does not imply that the density function has support only on the positive X axis for all α's. It is strictly positive only for $\alpha < 1$, but this condition is very restrictive for our modeling since we impose the inequality $\alpha > 1/H$, where $0 < H < 1$. Also the condition that α is greater than 1, ensures that the mean of the LFSN exists, according to the properties of α-stable distributions.



(d)

(e)

(f)

Figure 4: Part II web traffic measurements over the Communications Group's LAN. Plots correspond to the last three files shown in table 2.

The expected value of the model is

$$\mathbf{E}[M] = c_1 \cdot \mathbf{E}[L_{\alpha,H}] + c_2 = c_2, \qquad (2)$$

since $\mathbf{E}[L_{\alpha,H}] = \text{const} \cdot \mu = 0$ [5].

4. MODEL PARAMETER ESTIMATION AND SIMULATIONS

The proposed stochastic model in equation 1 depends only on a set of four parameters: (α, H, c_1, c_2). Our goal is to investigate whether this parsimonious model, can be accurately fitted to webcasting and web traffic as it was done to aggregate Ethernet traffic in [5].

The first step in the modeling is to estimate the self-similarity parameter H, which indicates whether the time-series representing the packet-count is asymptotically self-similar or not, and what is the degree of self-similarity. To estimate H we use the R/S statistic. The results are summarized in table 3 where we see that all of the webcasting and web data records exhibit self-similarity in various degrees.

Webcast records	Estimated H
100 stations	0.69
150 stations	0.66
200 stations	0.71
Web records	
1	0.59
2	0.85
3	0.82
4	0.75
5	0.62
6	0.79

Table 3: Estimated self-similarity parameter H for aggregated webcasting and web traffic records. The aggregated webcasting records were constructed with the method described in section 2.2 and the web records are constructed from the files given in table 2.

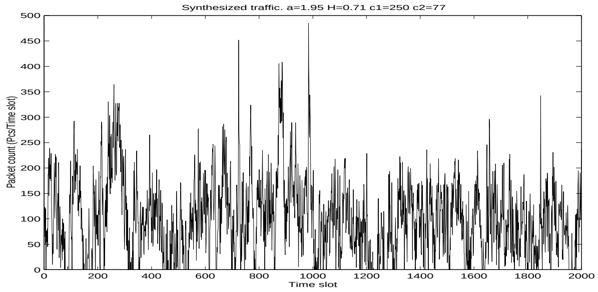


Figure 5: Traffic generated by the model extracted from the multiplexed webcasting traffic with 200 stations shown in figure 2(c).

Parameter c_2 is estimated simply by the mean of the packet-count process as equation 2 suggests. The other parameters α and c_1 can be estimated by minimizing the mean absolute error between the model and the real data:

$$\min_{1/H<\alpha\leq 2, c_1>0} \mathbf{E}|\mathbf{X} - c_1 \cdot \mathbf{L}_{a,H} - c_2|, \qquad (3)$$

where $\mathbf{X}$ is the vector of the real data corresponding to the packet-count. The existence of the mean absolute error is guaranteed since $\alpha > 1$ for the permissible range of H, as mentioned above.

We conducted simulations following the above parameter estimation method, and the results are shown in figures 5 and 6. Figure 5 presents synthesized traffic based on the model extracted from the multiplexed webcasting traffic with 200 stations, shown in figure 2(c), while figure 6 depicts traffic generated by the model extracted from the real web traffic record which is shown in figure 3(c). In the first case, the parameter set is $(\alpha = 1.95, H = 0.71, c_1 = 250, c_2 = 77)$, while in the second case, it is $(\alpha = 1.7, H = 0.85, c_1 = 100, c_2 = 55)$. In both cases the tail parameter α is smaller than 2, which justifies the choice of an underlying heavy-tailed distribution in the model. Both synthesized records look statistically very similar to the original time-series.

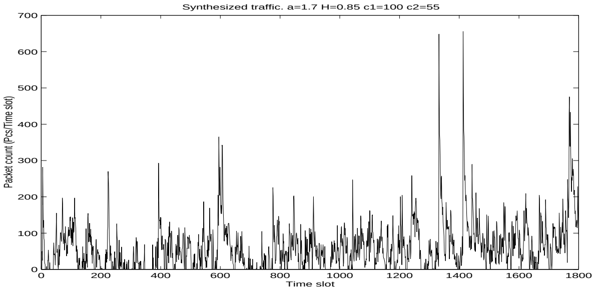


Figure 6: Traffic generated by the model extracted from the aggregated web traffic shown in figure 3(c).

5. CONCLUSION

We showed that multiplexed webcasting traffic and aggregate web traffic exhibit self-similarity asymptotically. Since web traffic already comprises a large portion of today's overall network traffic, and webcasting is expected to be the next resource-hungry application, network engineering will need statistically accurate, robust and parsimonious models to characterize them. We presented an α-stable self-similar stochastic process that was used as a model with the desirable properties. We are currently working on the investigation of more efficient and reliable algorithms for model parameter estimation and on the evaluation of our proposed model for characterization of network performance.

Acknowledgment

The authors would like to thank Prof. Irene Katzela at the University of Toronto for many helpful discussions and suggestions.

6. REFERENCES

[1] Bates S., McLaughlin S., "An investigation of the impulsive nature of Ethernet data using stable distributions", UK Perf. Eng. Workshop, Sept. 12, 1996.

[2] Beran J., Sherman R., Taqqu M.S., Willinger W., "Long-Range Dependence in Variable-Bit-Rate Video Traffic", IEEE Trans. Comm., vol. 43, no. 2/3/4, Feb/Mar/Apr 1995.

[3] Crovella M.E, Bestavros A., "Explaining World Wide Web Traffic Self-Similarity", Technical Report TR-95-015, Boston University, 1995.

[4] Feller W., *An Introduction to Probability Theory and Its Applications*, Volume II, John Willey & Sons, 1971.

[5] Karasaridis A., Hatzinakos D., "On the Modeling of Network Traffic and Fast Simulation of Rare Events using α-Stable Self-Similar Processes", Proc. IEEE Sign. Proc. Workshop on Higher-Order Statistics, Banff, Alberta, Canada, 1997.

[6] Leland W.E., Taqqu M.S., Willinger W., Wilson D.V., "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", IEEE Trans. Networking, vol. 2, no. 1, Febr. 1994.

[7] Norros I., "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks", IEEE Sel. Areas in Comm., Vol. 13, No. 6, Aug. 1995.

[8] Paxson V., Floyd S., "Wide-Area Traffic: The Failure of the Poisson Modeling", IEEE/ACM Trans. Networking, 3(3), pp. 226-244, June 1995.

[9] Pointcast Network, Web site at *www.pointcast.com.*

[10] Samorodnitsky G., Taqqu M.S., *Stable Non-Gaussian Random Processes*, Chapman & Hall, 1994.

[11] Willinger W., Taqqu M.S., Sherman R., Wilson D.V., "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Soruce Level", IEEE Trans. on Networking, 5(1), 1997, pp. 71-76.