Characterisation of Web Traffic

E. Casilari, A. Reyes-Lecuona, F.J. González, A. Díaz-Estrella and F. Sandoval
Dpto. Tecnología Electrónica, E.T.S.I. Telecomunicación, University of Málaga
Campus de Teatinos, 29071 Málaga (Spain),

***Abstract*-In this work the authors show how the behaviour of Web users strongly affects the dynamics of TCP connections in Internet. Analysing actual and systematically generated HTTP traces, it is proved that the time between the download of two pages is critical to determine the re-utilisation of TCP connections. On the other hand, the study also shows that the utilisation of 1.1 version of the HTTP standard does not significantly affect the traffic generated by HTTP 1.0. In this sense, the heavy-tailed nature of the size of HTTP connections can be considered as an invariant property.**

I. Introduction

Due to the dramatic growing of IP based networks, an accurate knowledge of the nature of Internet traffic is crucial to properly design network controls or even to dimension wide area or access networks. HyperText Transfer Protocol (HTTP) is by far the dominant traffic source in IP networks, as it is the protocol implemented by the browsers to surf Internet. Depending on the browser, the utilised version of the protocol is 1.0 or 1.1.

The first versions of HTTP 1.0 [1] opened and closed a new TCP connection for each object embedded in a Web page. This increased the user-perceived latency as well as the protocol and the processing overloads since the establishment of each TCP connections requires a 3-way handshake and augments the time for TCP to achieve its appropriate transmission speed. To cope with this problem, version 1.1 [2] proposed the utilisation of the so-called persistent connections. According to this new version, connections are kept open (until a time-out expires) once an object has been delivered to the browser. If the user emits a request of more objects, opened connections are re-utilised avoiding the need of opening new ones. As the first analysis of Web traces were performed when HTTP 1.1 was not very common, it was thought that the new version would imply a significant change in HTTP traffic. But in parallel, the "Keep-alive" extension had been included in HTTP 1.0 implementations. With some limitations (especially in the presence of proxies), this extension can imitate the technique of persistent connections in HTTP 1.1.

In this work the actual differences between HTTP 1.0 and 1.1 traffic are systematically studied proving that the utilisation of HTTP 1.1 does not impose notable differences on HTTP traffic. However, it is shown that the existence of a time-out for both Keep-alive technique and persistent connections establishes a heavy correlation between the user behaviour and the traffic generated at TCP level.

II. Modelling of Web Traffic

According to a "black-box" modelling strategy, Internet traffic traces can be characterised by means of "behaviourist" models, that is to say, abstract processes which try to capture the statistical traffic properties with independence of the subjacent mechanisms of traffic generation. For example, in [3] it is proposed a non-linear AutoRegressive Moving Average (ARMA) filter to model the IP traffic flowing from a campus LAN to the vBNS backbone. Similarly, Basu [4] defines several Fractional Integrated ARMA (FARIMA) filters to approximate the traffic of different IP services (SMTP, FTP, HTTP,...).

In opposition to this family of models, structural o hierarchical modelling offers a more comprehensive and meaningful way to fully characterise the complex nature of traffic. At the cost of defining a more elaborate (and usually mathematically intractable) statistical process, the main advantage of these models is that their parameters possess a physical meaning so they could be more easily accommodated to changes in the conditions of actual networks. Consequently, a proper parametrisation of the models can make them suitable to follow the evolution of the traffic load, the user behaviour, the number and size of multimedia objects in Web pages, etc.

In the case of HTTP flows, this structural modelling policy presents multilayer or multilevel processes which imitate the different components or levels existing behind Web browsing [5]. Thus, these models normally consist of the following levels:

- The session level, which must describe the user behaviour in terms of the number of Web sessions per day (week, month or year) and the distribution of the sessions along the day or, otherwise, the time between two consecutive sessions.
- The page level: as Internet navigation implies to visit a series of Web pages, this level focuses on determining the number of pages per session and the statistical distribution of the time between pages.
- The connection level: A Web page in turn consists of a bunch of objects (text, images, sound files, etc.), which are conveyed through one or more TCP connections. Consequently, for this level it is necessary to model the number of connections for each page, the time between two consecutive connections as well as the distribution of the connection sizes (in bytes). This last parameter is directly related to the dimension of the multimedia objects that are transported by each connection.
- The packet level: If a low granularity is required, the packet level must be incorporated to the model. Thus, the total amount of bytes for each connection has to be split in TCP/IP packets. For this purpose, this level must characterise the distribution of the packet sizes and the interarrival times of the packets. Obviously, these parameters are greatly affected by the physical and link layers of the path between the user terminal and the different Web servers (mainly by the maximum transfer unit along the path and the access rate of the user link to Internet).

0-7803-7208-5/01/$17.00 (C) 2001 IEEE

III. A REVIEW OF HTTP TRAFFIC MODELS

There exist many works in the literature which follow a structural or "page oriented" approach to model Web traffic. They normally consist in the characterisation of long HTTP traces collected in Campus or Local Area Networks. The traces are usually analysed at the transport layer (TCP), although there are studies such as [6], which examine the samples through the HTTP headers. So, in order to distinguish the pages and sessions (two concepts of the application layer), the set of packets generated by the access to each page (or session) is separated by searching an inactivity interval in the traffic flow (or between the beginning of two consecutive connections) longer than a certain threshold. Tables I to VII summarises the standard distributions and their mean values (m) which have been proposed to model the different levels of the models. As a general rule, the autocorrelation of these variables is measured to be very weak [6] so they are modelled by means of uncorrelated processes.

A. Session Level

At the session level, the Poisson assumption performs well, so call inter-arrival time can be reasonably approximated by an exponential function. If the model has to describe the daily or weekly behaviour of Web traffic the process should describe the seasonal (non stationary) nature of the access of Internet users, which extremely depend on the scenario (access rate, type of user, etc.). The scenario is also a key aspect in determining the number of pages consulted during each session, which is sometimes supposed to follow a geometric distribution for simplicity (Table III).

Table I. Characterisation of the time between the beginnings of two consecutive sessions

Ref.	Proposed distribution	Parameters
[7]	Exponential	m is considered to be an invariant
[8]	Non stationary process	m depends of the time of the day

Table II. Characterisation of the number of sessions per day

Ref.	Frequency (sessions/day)	Scenario
[9]	0.5	Modem connected users
[9]	1.5	ISDN connected users
[10]	4	Residential environment (USA)
[11]	0.85	Residential environment (USA)
[11]	1.43	Corporate environment (USA)

Table III. Characterisation of the number of pages per sessions

Ref.	Proposed distribution	Measured Mean Value (pages/session)
[12]	Geometric	m=50
[13]	Not modelled	m=19.6
[14]	Not modelled	m=40.8
[15]	Geometric	m=5
[11]	Not modelled	m=30
[5]	Lognormal	m=22-25

B. Page Level

The time between two consecutive pages is related to the reading time of the users. As the human perception of time follows a normal distribution in the logarithmic scale [17], the time between pages result to be sub-exponentially distributed. So sub-exponential distribution such as Pareto or Weibull have been suggested to characterise this parameter, although there are examples in the literature where it is proposed the exponential distribution (see Table IV).

Table IV. Characterisation of the time between two consecutive pages within the same session

Ref.	Proposed distribution	Measured Mean Value (s)
[17]	Weibull	m=21
[12]	Exponential	m=33
[15]	Geometrical	m=12
[13]	Pareto	m=81
[14]	Pareto	m=43.5
[18]	Pareto	m=3
[6]	Weibull	m=39.5
[5]	Gamma	m=25-35

On the other hand, the number of connections per page has been commonly considered to be settled by the number of embedded (in-line) objects within the pages. However, this parameter may be also strongly affected by mechanisms such as the persistent connection (or the "keep alive" extension of HTTP 1.0), the effect of the local cache or the maximum number of parallel connections permitted by the browsers (about 4 or 6 connections). As it can be observed from Table V, the literature has employed diverse standard distributions to model this variable but for all cases the estimated mean value was very low (below 6 connections per page).

Table V. Characterisation of the number of connections per page

Ref.	Proposed distribution	Measured Mean Value
[17]	Gamma	m=1.9
[12]	Geometric	m=2.5
[18]	Pareto	m=2.7
[13]	Not modelled	m=4
[14]	Not modelled	m=3.5
[19]	Not modelled	m=2.8-3.2
[6]	Gamma	m=5.5 (In-line objects)

C. Connection Level

Once the HTML document which describes a Web page is analysed, the browser may open one or several TCP/IP connections (up to the maximum allowed number) to download the in-line objects. In case of re-utilisation of existing connection a page could be completely loaded without this opening process. As a consequence, time between connections belonging to the same page are usually quite small (about tens of milliseconds, see Table VI). Only if the browser requires to exceed this maximum number, the opening of new connections will have to wait until the download of previous objects.

Table VI. Characterisation of the time between two consecutive connections within the same page

Ref.	Proposed distribution	Measured Mean Value
[17]	Gamma	m=0.148 s (between the first connection of the page and the rest)
[17]	Deterministic (a fixed time)	m=0 ms (between the second and consecutive connections of the page)
[12]	Exponential	m=500 ms
[8]	Lognormal	Not offered (Approximation for the connections of aggregate Web traffic during a loaded hour)
[8]	Exponential	Not offered (Approximation for the connections of aggregate Web traffic during a not loaded hour)
[6]	Gamma	m=0.860 s (mean time between the opening of an in-line object and the next)

As it refers to the connection sizes, in [20] it was proved that the size of Web objects follow a heavy-tailed distribution. This fact, which offers a physical explanation for the existence of self-similarity in Internet traffic, is reflected in the sub-exponential nature of the connection sizes in HTPP transfers (see Table VII). Thus, lognormal and Pareto distributions are normally employed to approximate this parameter whose mean value evolves in parallel with the increasing complexity of the contents in Web pages

Table VII. Characterisation of the connection sizes

Ref.	Proposed distribution	Measured Mean Value
[8]	Lognormal	Not offered
[18]	Composition of Pareto and Lognormal	m=7.2-14.8 KBytes
[17]	Lognormal	8.3 packets
[12]	Erlang	Not offered
[22]	Pareto	Not offered
[19]	Heavy-tailed	m=8-10 KBytes
[5]	Lognormal	m=7.7-10.7 KBytes
[20] [7]	Heavy-tailed	Not offered

IV. Effects of the Version of HTTP and the Time Between Pages

As it can be deduced, the previous models define each level independently from the others, presuming that they are influenced by different factors (the nature of the contents in the Web for the connection level, and the short and long term behaviour of the user for the cases of the page and session levels, respectively). However, HTTP can impose strong cross-correlations between the levels, which should not be neglected. In particular, it must be pointed out that each Web object does not have to be necessarily conveyed through a single and exclusive TCP connection. In fact, most Web pages include much more than 5 or 6 objects (the maximum mean number of connections per page which have been measured in the aforementioned studies). This implies that the relationship that is usually assumed between objects and connections should be revised. Moreover, it can be shown that the time-out of persistent connections establishes a dependence of the number of connections per page and the connection size with respect to the time between pages [23].

Aiming to analyse the impact of the utilised version of HTTP as well as the dependence between the page and the connection levels, we investigate the traffic generated by Web browsers (in particular we used MS Internet Explorer) when alternatively utilising the version 1.0 and 1.1 of the HTTP standard. For the analysis (performed in December 2000), we consider two scenarios which could represent two limit cases. In both scenarios, the browsers are programmed to consecutively visit a list of Web pages, waiting a fixed interval T between the visits (loads) of two successive pages.

- In the first scenario the list consists of more than 200 heterogeneous Web pages, all situated in different national, European and American Web servers. To make the sample representative of present Internet, the list includes the main Web pages of the 100 most visited sites in Internet. On the other hand, the results do not indicate any significant difference between enabling HTTP 1.0 or 1.1 in the browser. The number of connections per page remains practically unchanged for both versions as well as the connection size, as it can be concluded from Fig. 1. The figure also proves the heavy-tailed nature of the marginal distribution of this parameter, which is reflected in a hyperbolic decay (linear in the logarithmic scale) of the estimated normalised histogram of the connection sizes from the traces. As it has been pointed out before, this heavy-tailed property (or subexponentiality) in the HTTP connection sizes has been justified [20] by the heavy-tailed distribution of Web objects existing in Internet. Furthermore, the figure also proves that the connection size can be reasonably modelled by a Pareto distribution:

$$F(x)=\begin{cases}1-\left(\dfrac{b+x}{b}\right)^{-a} & for \quad x\geq 0\\ 0 & other \quad case\end{cases} \qquad (1)$$

where x represents the size.

For our traces, the parameters of this approximation (a,b) were designed to match both the estimated mean and the decay rate of the histogram. By using a regression, a was found to be between 1.6 and 1.7 for the two HTTP traces. A value of a lower than 2 indicates the presence of the 'syndrome' of the infinite variance (also called the Noah effect) which has also been reported in the literature. Fig. 1 illustrates the accuracy of the adjustment for two orders of magnitude.

- In the second scenario the navigation is entirely performed trough a single Web site. In particular, the browsers consecutively visit more than 150 and 110 pages situated in the Web Servers of Microsoft and Lycos, respectively, which are two of the 3 most popular Web sites in Internet, although similar results have been obtained with other Web sites. In contrast with the previous scenario, in this case the navigation increases the utilisation of persistent connections (or Keep-alive extension of HTTP 1.0) as the

HTTP server, which is the endpoint of TCP connections, is always the same. For this reason, we also investigate the influence of the interval *T* between the load of two pages.

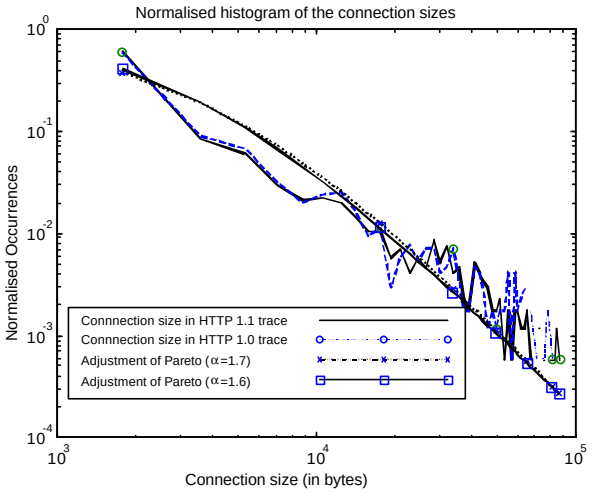

Fig. 1. Normalised Histogram of the connection sizes (first scenario: heterogeneous Web pages).

The results (in terms of the mean size of the connections and the mean number of connections per page) are depicted in figures 2 (for the case of Microsoft) and 3 (for the server of Lycos). Figures show the strong influence of the time between pages on the connection level, proving the existence of three clearly defined zones:

1. An initial zone of very low times between pages (less than 15 seconds), which could correspond to the situations in which the user does not wait until the page is completely loaded to click an hyperlink to another site. In this case, as the load of objects in the present page has not finished, before resetting present connections, the browser requires to open new connections to request and convey the objects of the new page. Consequently, the utilisation of persistent connections is poor, and the number of connections per page increases.
2. A second zone of low times between pages (between 15 and 60 seconds), which could represent quick visits to the pages (e.g.: the user just gets a glance at the embedded pictures or reads some lines before quitting the page). In this zone, the idea behind persistent connections is proved to be effective. So, the load of each page re-utilises the idle connections of the previous page, which are still open. As a consequence, HTTP traffic consists in fewer and longer TCP connections.
3. A third zone of high times between pages (more than 60 seconds), which may be related to longer reading or thinking times of the user before accessing another page. This time of inactivity can be equally motivated if the user iconises the browser while working with other application. For this zone, the time-out (in this case about 60 seconds) of the connections obliges to close them before a new page is solicited, so they cannot be re-utilised. By surpassing this threshold determined by the time-out, the number of connections per page notably increases while the mean load of each connection decreases.

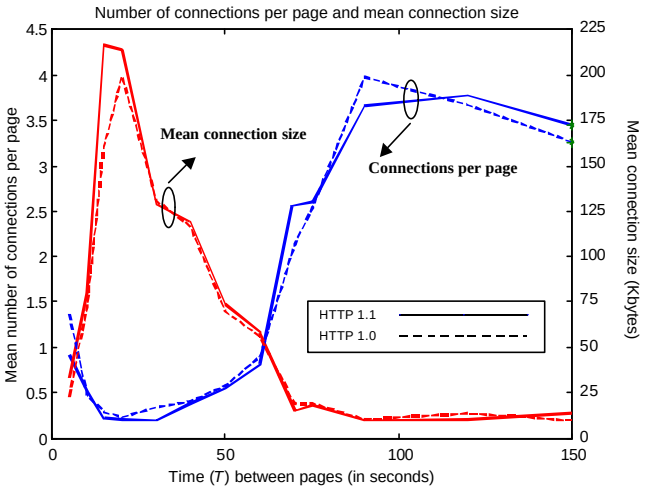

Fig. 2. Mean number of connections per page and mean connection size as a function of the interval *T* between pages (second scenario: Web pages in the server of Microsoft).

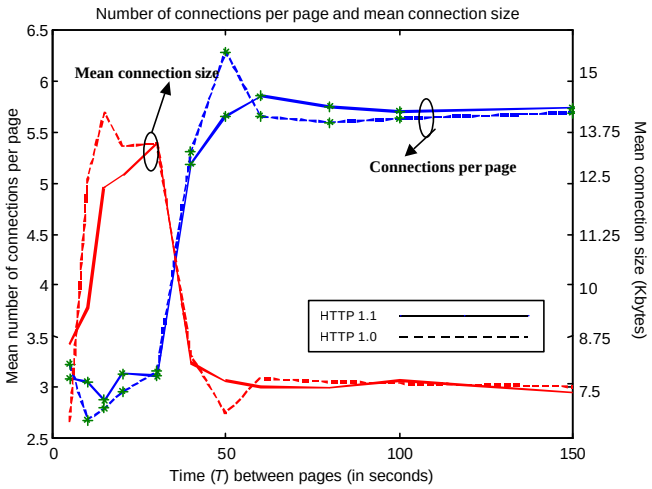

Fig. 3. Mean number of connections per page and mean connection size as a function of the interval *T* between pages (second scenario: Web pages in the server of Lycos).

Obviously, the particular values of the thresholds between the zones depend on the Internet latency suffered by the user as well as on the specific implementation of HTTP. Nevertheless, the three zones will be present in all HTTP transactions. In any case, the results for HTTP 1.0 and 1.1 present no significant variation while the distribution of connection sizes (see Fig. 4) for this second scenario are also demonstrated to be heavy-tailed.

IV. Conclusions

In this work it is proved that the behaviour of Web users strongly affects the nature of TCP connections in Internet. In

particular, it is shown that the time between two pages is critical to determine if existing connections can be re-utilised or if new ones have to be opened (increasing the traffic overload due to the typical TCP handshake of connection set-up). An accurate model for re-utilised connections is crucial as they present several traffic peaks provoked by the load of different pages. Consequently, this correlation between the page level and the connection level, which is neglected by the literature, should be incorporated to the strategies for modelling HTTP traffic. On the other hand, by means of alternatively enabling the two versions of HTTP protocol, the study also shows that for the same conditions HTTP 1.1 traffic exhibits the same properties than HTTP 1.0. In this sense, the heavy-tailed nature of the size of HTTP connections is proved to be an invariant.

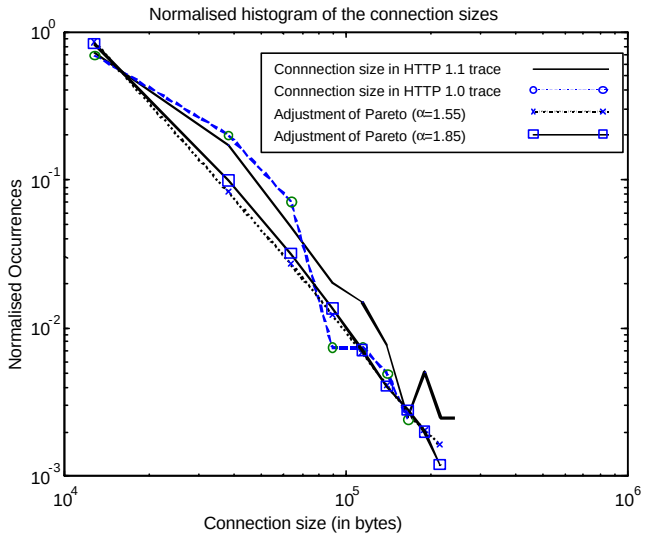

Fig. 4. Normalised Histogram of the connection sizes (second scenario: heterogeneous Web pages, Microsoft Server with a fixed time between pages of 75 s).

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Projects No. TEL99-0755 and by Funds of the European Union (Project FEDER-CICYT No. 1FD97-0918)

REFERENCES

[1] T. Berners-Lee, R. Fielding and R. Frystyk, "Hypertext Transfer Protocol – HTTP/1.0", Request for Comments (RFC) No. 1945, HTTP Working Group, May, 1996.

[2] R. Fielding, J. Gettys, J. C. Mogull, H. Frystyk and T. Berners-Lee, "Hypertext Transfer Protocol – HTTP/1.1", Request for Comments (RFC) No. 2068, HTTP Working Group, January, 1997.

[3] C. You, K.Chandra, "Time Series Models for Internet Data Traffic", *Proc. 24th Conf on Local Computer Networks*, October, 1999, Lowell (Massachusetts).

[4] S. Basu, A. Mukherjee and S. Klivansky, "Time Series Models for Internet Traffic", Proc. of INFOCOM'96, San Francisco, 1996.

[5] A. Reyes-Lecuona, E. González, E. Casilari, J. C. Casasola and A. Díaz-Estrella, "A Page-Oriented WWW Traffic Model for Wireless System Simulations", *Proc. of International Teletraffic Congress* (ITC-16), Vol. 3.b, pp. 817-826, Edinburgh (UK), June 1999.

[6] H. Choi and J. Limb, "A Behavioral Model of Web Traffic", International Conference of Networking Protocol'99 (ICNP 99), September 1999.

[7] W. Willinger and V. Paxson, "Where mathematics meet Internet", Notices of the American Mathematical Society, vol. 45, pp. 961-970, September 1998.

[8] N. Nabe, M. Murata and H. Miyahara, "Analysis and modelling of World Wide Web traffic for capacity dimensioning of Internet access lines", in *SPIE'97* (W.S. Lai and H. Kobayashi, Eds.), Dallas, November 1997.

[9] J. Farber, S. Bodamer and J. Charzinski, "Evaluation of dial-up behaviour of Internet users", in ITG-Fachbericht (K.D. Schenkel and J. Speidel, eds.), pp.73- 78, October 1998.

[10] S. Morgan and M. Delaney,"The Internet and the local telephone network: conflicts and opportunities", *IEEE Communications Magazine*, vol. 36, pp. 42-48, January 1998.

[11] Nielsen//NetRatings Inc., "Internet Audience Statistics", Technical Report, Available at http://www.nielsen-netratings.com/

[12] M.F. Arlitt and C.L. Williamson, "A Synthetic Workload Model for Internet Mosaic Traffic", *Proc. of the 1995 Summer Computer Simulation Conference (SCSC'95)*, Ottawa, pp. 24-26, July 1995.

[13] N. Vicari, "Measurement and Modeling of WWW-Sessions", Report No. 184, Research Report Series, Institute of Computer Science, University of Wurzburg (Germany), September 1997.

[14] N. Vicari, "Models of WWW traffic: A comparison of Pareto and Logarithmic histogram models", Report No. 198, Research Report Series, Institute of Computer Science, University of Wurzburg (Germany), 1998.

[15] "Selection procedures for the choice of radio transmission technologies of the UMTS", Technical Report 101 112 v3.1.0, ETSI, 1997.

[16] V. Bolotin, "Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis", IEEE Journal on Selected Areas in Communications, Vol. 12, No. 3, pp 433-438, April 1994.

[17] S. Khaunte and J. O. Limb, "Statistical Characterization of a WWW Browsing Session", Technical report GIT-CC-97-17, Georgia Tech. College of Computing, June 1997.

[18] P. Barford and M. Crovella, "Generating representative web workloads for network and server performance evaluation", Technical Report BU-CS-97-006, Computer Science Department, Boston University, 1997.

[19] B.A. Mah, "An Empirical Model of HTTP Network Traffic", *Proceedings of the IEEE INFOCOM'97*, vol. 2, Kobe (Japan), pp. 592-600, April 1997.

[20] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web. Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp. 835-846, December 1997.

[21] P. Barford, A. Bestavros, A. Bradley and Mark Crovella, "Changes in Web Client Access Patterns: Characteristics and Caching Implications", *World Wide Web*, vol. 2, pp. 15-28, January 1999.

[22] C. R. Cunha, A. Bestavros and M. E: Crovella, "Characteristics of WWW Client-Based Traces", Technical Report BU-CS-95-010, Computer Science Department, Boston University, July 1995.

[23] E. Casilari, F. González and F. Sandoval, "Modelling of HTTP Traffic", *IEEE Communications Letters*, in press.