Modeling of HTTP Traffic

Eduardo Casilari, Francisco J. González and Francisco Sandoval

***Abstract*-- In this work the authors show that 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 downloads of two pages is critical to determine the re-utilisation of connections. On the other hand, the use of 1.1. version of the HTTP standard does not significantly affect the traffic generated by HTTP 1.0 browsers. In this sense, the heavy-tailed nature of the size of HTTP connections can be considered an invariant property.**

***Index Terms*-- HTTP 1.0, HTTP 1.1, Web traffic, modelling of internet traffic**

I. INTRODUCTION

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 navigator, the utilised version of the protocol is 1.0 or 1.1. The first versions of HTTP 1.0 [1] opened a new TCP connection for each object embedded in a Web page. This increased the latency and the protocol overload, as the establishment of a TCP connection requires a 3-way handshake. To cope with this problem, version 1.1 [2] proposed the so-called persistent connections, which are kept open (until a time-out expires) once an object is delivered to the browser. If a new request is emitted, open connections are re-utilised avoiding the need of opening new ones. But in parallel, the "Keep-alive" extension had been included in HTTP 1.0 implementations. This extension was designed to imitate the technique of persistent connections in HTTP 1.1.

In this work it is shown that HTTP 1.0 and 1.1 traffics do not present notable differences. However, it is proved 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

To fully characterise the complex nature of HTTP flow, it is required a hierarchical model which imitates the underlying mechanism of Web browsing [3]. This model should consist of the following levels:

- The session level, which describes 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.
- The page level: as Internet navigation implies to visit a series of Web pages, this level determines the number of pages per session and the distribution of the time between pages.
- The connection level: A Web page in turn consists of a bunch of objects (text, images, sound files,...) which are conveyed through one or more TCP connections. Hence, for this level it is modelled the number of connections for each page, the time between two consecutive connections as well as the sizes of the connections.
- The packet level: Each connection has to be split in TCP/IP packets. Thus, this level must characterise the sizes and interarrival times of the packets, mainly determined by the physical and link layers of the path between the user terminal and the Web servers.

The models in the literature [3] [4] define each level independently from the others, assuming that they are influenced by different factors (the Web contents for the connection level, and the short and long term behaviour of the user for the page and session levels, respectively). However, HTTP imposes a strong correlation between these levels. In particular, in the next section we show how the time-out of persistent connections establishes a dependence between the pages and the connection levels.

III. ANALYSIS OF HTTP TRACES

We analyse the traffic generated by Web browsers when alternatively utilising the version 1.0 and 1.1 of the HTTP standard. In particular, we employed MS Internet Explorer 5.0, which allows to set the utilised version of the HTTP client. For the analysis, we consider two scenarios representing two limit cases. In both scenarios, the browser reads a local page containing an HTML frame which consecutively visits (by means of a script) a list of external Web pages. After requesting a page, the frame is programmed to wait a fixed time interval T before loading the following page in the list. The downstream traffic received by the browser is collected (through the monitoring software *Tcpdump*) in another terminal situated in the same LAN (Ethernet) segment. The experiments were performed in December 2000.

- In the first scenario the list consists of more than 200 heterogeneous Web pages, all situated in different national, European and American Web servers, including the main Web pages of the 100 most visited sites in Internet. The results for both versions 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 as well as the connection size, as it can be deduced from Figure 1. The figure also proves the heavy-tailed nature

Manuscript received February 12, 2001. This work was partially supported by the Spanish Comisión Inteministerial de Ciencia y Tecnología (CICYT) Project No. TEL99-0755 and by Funds of the European Union (Project FEDER-CICYT No. 1FD97-0918)

The authors are with the Department of Electronics Technology, University of Málaga (Spain) (e-mail: casilari@dte.uma.es).

of the marginal distribution of this parameter, which is reflected in a hyperbolic decay (linear in the logarithmic scale) of the normalised histogram of the connection sizes from the traces. This heavy-tailed property (or subexponentiality) in the HTTP connection sizes has been justified [5] 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 while α and β are the two parameters of the distribution.

By using a regression to match the decay rate of the histogram, 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, which has also been reported in the literature.

- In the second scenario the navigation is entirely performed through a single Web site. In particular, the browsers consecutively visit more than 150 pages situated in the Web Server of Microsoft, which is one of the 3 most popular Web sites in Internet, although similar results have been obtained with other sites. In contrast with the previous scenario, in this case the navigation increases the utilisation of persistent connections (or Keep-alive extension), as the endpoints of the TCP connections coincide for most transfers (only ten different external IP addresses were detected in the traces, which indicates that the embedded objects are concentrated in the same HTTP servers). For this reason, we also investigate the influence of the interval T between the pages.

The results are depicted in figure 2. The figure shows 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 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, the browser requires to open new connections to request the objects of the new page. Therefore, the utilisation of persistent connections is poor, and the number of connections per page increases.
2) A second zone of medium 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 before quitting the page). Now, as the transmissions of the objects of the previous page have finished, the load of the new page re-utilises those idle connections, 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 describe longer reading times of the user before accessing another page. 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. Thus, the number of connections notably increases while the mean load of each connection decreases.

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 (e.g.: the maximum number of connections to one server is a browser-configurable parameter whose default value differs from Netscape to MS Internet Explorer). 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 (not depicted here) is again demonstrated to be heavy-tailed.

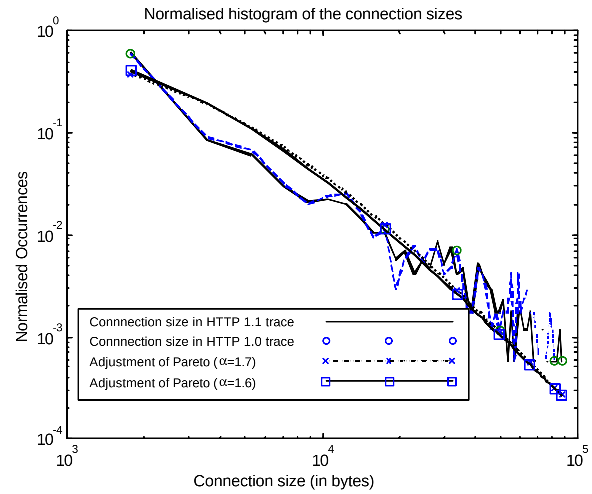

Fig. 1. Normalised Histogram of the connection sizes (first scenario: heterogeneous Web pages).

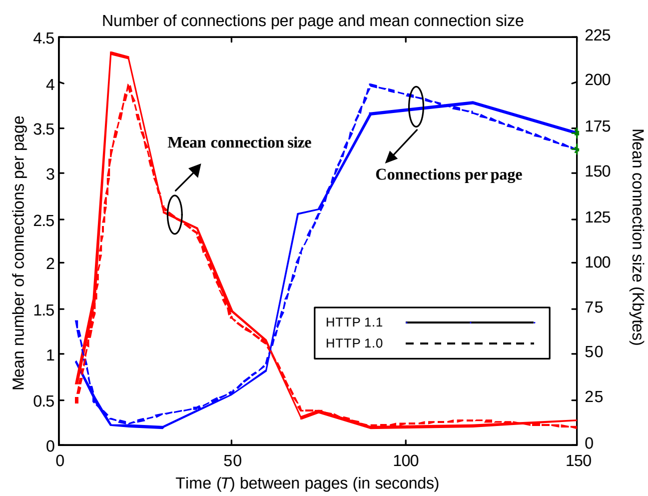

Fig. 2. Mean number of connections per page and mean connection size as a function of the interval T between pages (second scenario: Web pages in the server of Microsoft).

IV. CONCLUSIONS

In this work it is shown that the time between two Web pages is critical to determine if existing TCP 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 and is imposed by the dynamics of Web Users, 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, as it refers to the number of connections per page, 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. Obviously, longer traces should be collected to analyse the impact of other improvements introduced by the version 1.1 of the standard. For example, HTTP 1.1 saves bandwidth by means of the range requests [6], which permits a partial download of a Web object (e.g.: a file whose previous transfer was truncated in mid-stream). Similarly, the version 1.1 supports new mechanisms for negotiating the data compression as well as a new status code which avoids the transport of the bodies of rejected requests [6]. Depending on the application of these new capacities, the sizes of TCP connections could be affected by HTTP 1.1.

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] 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

[4] 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.

[5] 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.

[6] B. Krishnamurthy, J. C. Mogul and D. M. Kristol, "Key Differences between HTTP/1.0 and HTTP/1.1", *Proc. of the WWW-8 Conference,* Toronto, May 1999.