Modelling of individual and aggregate Web traffic

E. Casilari J.M. Cano-García F.J. González-Cañete F. Sandoval

Dpto. Tecnología Electrónica, E.T.S.I. Telecomunicación, University of Málaga, Spain

Campus de Teatinos, 29071 Málaga, Spain

ecasilari@uma.es cano@dte.uma.es equinoxe@dte.uma.es sandoval@dte.uma.es

ABSTRACT

This paper describes the behaviour of individual Web users as well as aggregate HTTP traffic basing on actual traces taken from a transoceanic link between Japan and USA. The proposed model characterises the traffic at different levels, ranging from packets to Web pages. Thus, the study investigates the effects of aggregation on the heavy-tailed nature and the long range dependence present in the variables which govern Internet traffic.

Topics

Traffic models for high speed networks

Keywords

Web traffic model, long range dependence

1. INTRODUCTION

In spite of the efforts for integrating interactive multimedia services in IP networks, TCP connections and, in particular, those generated by Web transactions are still a major traffic source in Internet. The fact that the terms 'Web' and 'Internet' are commonly confused can be regarded as a sign of the prevalence of HTTP flows within the Net of Networks. Since the mid-90s, the growth and predominance of traffic conveyed through TCP port 80 has not been questioned by any other service. As it refers to the bandwidth consumption, just the apparition of Peer-to-peer (P2P) file transfer applications has introduced another 'killing' application in Internet. Thus, a proper characterisation of HTTP connections is a key aspect not only for dimensioning the Network at different levels but also to evaluate new improvements in the protocols which interact in the generation and management of Web traffic (e.g.: HTTP, cache policies in the browsers, flow management in servers and routers,…) and even in TCP/IP. So, many research works have been devoted to this issue during last six or seven years.

Due to the intrinsic asymmetry of Web traffic, based on a client-server nature, most studies base their results in traces taken at one of both end points. At the client side, the traffic is normally captured in a LAN or MAN environment (normally situated in academic premises) by utilising a traffic 'sniffer' which runs at the transport layer. At the server side, the traffic is analysed by investigating the logs that describe the client accesses to the Web servers.

On the other hand, the most efficient way to face the modelling of Web traffic is to adopt a structural or multilayer strategy [6]. According to this viewpoint, the underlying factors that, at diverse time scales, impact on the generation of Web traffic, are separately analysed and characterised. In this sense, at the highest scale, the models suppose the existence of traffic sessions consisting in a series of visits to different Web pages. The visualisation of the objects contained in a Web page may require in turn opening one or several TCP connections, which are definitively composed by a flow of IP packets. Consequently, a structural vision of Web traffic defines at least four levels which are normally independently modelled. In this work this modelling strategy, which has been incorporated to popular simulating tools such as *network simulator* [24], is applied to the Web traffic collected in a transoceanic link. The employed traces allow comparing the models obtained by the literature in LAN networks with the tendencies detected in long-distance connections. In the same way, the heavy multiplexing of heterogeneous users which is performed in the studied link permits to analyse in detail the effects of aggregation on the statistical characteristics of traffic at every aforementioned level.

This paper is organised as follows: section 2 briefly describes the traces utilised for the analyses, in section 3 we perform the modelling of individual users at the different levels while the traffic as an aggregate is characterised in section 4. Finally section 5 summarises the conclusions.

2. DESCRIPTION OF THE TRACES

The employed traces include the traffic served by a router which connects USA and the Japanese network of WIDE project through an 18 Mbps CAR (*Committed Access Rate*) link. The WIDE Internet forms a shared research platform connecting about 140 organizations. The traces were collected in a local Fast Ethernet segment situated in Japan one hop before the international link (see [9] for more details). The samples were captured by means of the widely-used software *tcpdump* [22]. In particular, the analysed traffic corresponds to the packets generated by the accesses of local Japanese clients to the Web. For our study we chose six day traffic samples from 29th January to 3rd February 2003. These traces are available and daily updated in [16].

To discriminate Web traffic from other services we selected the traffic through TCP port 80 (the traffic volume resulting from Web transactions flowing through ports different from 80 is normally negligible).

The main characteristics of the traces are described in Table 1. The table shows the importance of Web browsing with almost 50% of the traffic load (in bytes), followed by Napster and similar P2P applications which generated a 20%.

The HTTP packets in the traces were filtered with *tcpdump* and then post-processed with *tcptrace* [23], a program which reconstructs the packet flows and extracts basic properties of the TCP connections such as the initiation time, the size or the duration. In order to investigate both the individual behaviour of Web users and the effects of aggregation, for each connection, the client and the server are supposed to be identified by their IP addresses. Making so, we also consider as a single client the traffic generated by 'multiplexing elements' such as proxies or NAT (Network Address Translator) servers. As a consequence, the model of individual user will be performed from the perspective of the backbone network.

Table 1. Basic characteristics of the employed traces

Total number of packets	Relative weight of several services over the whole traffic in % bytes (in % packets)			No. of detected HTTP clients	No. of detected HTTP connections
	Web Servers (Flows from port 80)	Web Clients (Flows to port 80)	Napster & P2P		
54818015	44.33 % (23.01%)	4.14 %´ (20.60%)	19.16% (13.75%)	39278	1281256

3. MODELLING OF INDIVIDUAL USERS

After separating the HTTP flows of each detected IP address, we applied the hierarchical model commented in section 1. In this paper, due to the length of the traces (just three days) the model is focused on the lower levels (page, connection and packet). For modelling the session level and the user behaviour, the parameters offered by sociological studies are of more relevance than those compiled by papers on networking. For example, Nielsen Netratings informs that the sessions of Web surfing in America have a mean duration of 33 minutes while the mean number of sessions per user and month is about 31 (data corresponding to September 2003, see [17] for details)

3.1 Page level

The strong burstiness of Web traffic is mainly caused by the accesses of HTTP clients to the heterogeneous contents or 'objects' (text, images, PDF files,…) present in the Web. Objects are organised in units (Web pages) visited in a discontinuous way. The load of a Web page provokes sudden traffic bursts which are followed by inactivity periods during which the client reads the page or just simply minimises the browser.

This bursty nature of Web traffic can be approximated, in the simplest way, by an On-Off process [2].

If the traffic is not analysed at the application layer by interpreting the HTTP messages contained in the payload of the packets (as it is performed in [8]), the only way of inferring the existence of pages is to detect the presence of these inactivity periods between consecutive pages. This technique, which is employed in several works as [15] or [20], requires to set up a temporal threshold (T_{Th}) that, if exceeded or not reached, allows to decide if two consecutive connections (or packets) belong to the same page. Taking into account that the access to a page ordinarily implies to open at least one TCP connection, we operated this procedure on the time between the initiation instants (t_i) of the consecutive connections of the clients, as it is illustrated in Figure 1. In this way, from the point of view of teletraffic, a page is the collection of bytes transmitted by connections whose initiations are nearby in time.

To avoid a heuristic election of T_{Th}, we utilised different values for this variable to identify the bounds of the pages. Figure 2 seems to indicate that the number of detected pages starts to stabilise for a threshold above 60 s, which was the elected value for T_{Th}. What is more, this period coincides with the time-out or maximum time that a persistent HTTP 1.1 connection is kept open (without being used) by certain browsers (e.g.: MS Explorer).

If we compare with other studies, [20] utilises a threshold of 30 s while [25] chooses 3 s with a slightly different procedure. On the other hand, [15] fixes the page bounds by comparing the time between the end of a connection and the beginning of the next one with a limit of 1 s. With a similar strategy [14] also utilises a threshold of 1 s, while [16] uses this value to cluster into pages the internal connections between the users of a LAN and their proxy server.

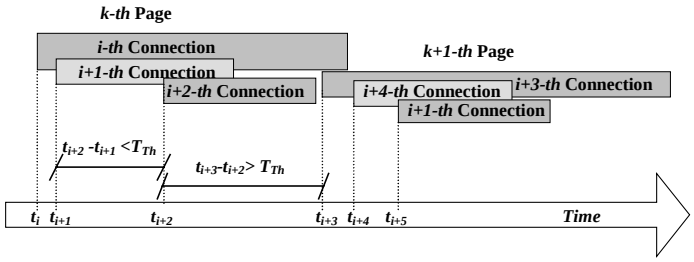

Figure 1. Grouping of the connections into pages

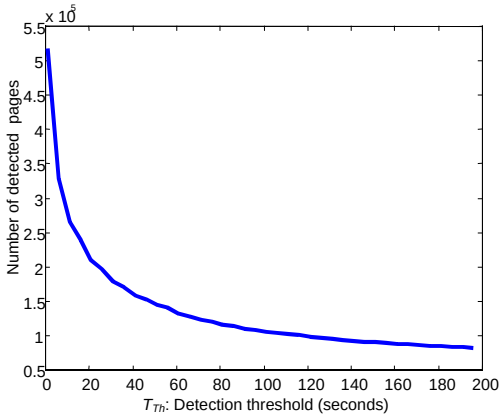

Figure 2. Number of detected pages as function of the chosen detection threshold

Anyhow, it must be considered that these detected 'pages' do not necessarily coincide with actual visits to Web pages. For example, in the case of loading several pages simultaneously (by opening more than one window with a browser) or visiting a number of pages situated in the same server through the same TCP connections (benefiting from the concept of persistent connections), the generated traffic will be considered as belonging to a single page. In the same way, the presence of a cache memory in the browsers and proxies increases this decoupling between the accesses to real pages and the 'pages' as units of traffic, since the cache permits to visualise Web contents avoiding traffic generation. Moreover, the employed traces, captured in a Wide Area Network (WAN) link, just include the traffic motivated by visits to remote (international) servers while the accesses to local or national servers are assumed to be routed through a different link. So, for the threshold of 60s, the measured number of visited pages per hour (17.91 pages/hour, see Table 2) is under the value reported by other sociological and market studies. The experience of Nielsen Netratings, which manipulates the browsing software of a wide population of users in order to track their Web activity, computes a mean of 50 visited pages per session [17] with a mean session duration about 30 minutes (this implies that a mean user consults about 100 pages/hour). Table 2, which also shows the statistics at the connection level, suggests characterising this variable with a Log-normal distribution. For this purpose, we

contrasted the adjustment of several standard distributions, which were employed for this and all the variables in this study. The distributions, which are tabulated in Table 3, were fitted to match the measured mean (μ_X) and the standard deviation (σ_X) of the actual variables, except for the case of Pareto, for which the tuned parameters were the mean and the hyperbolical decay rate (α) of the distribution tail. Similar results are obtained if a Maximum Likelihood Estimation (MLE) is applied to define these standard distributions. In order to compare the performance of the candidate distributions, we employed different methods, mainly the visual inspection of the fittings and the quantile-quantile plots. We also considered quantitative techniques such as Kolmogorov-Smirnov, χ^2 y λ^2 tests. However, although these methods offer an objective measurement of the quality of the adjustments, they are too sensitive to the number of intervals which are considered for their computation and they are strongly determined by the interval or range of values for which the approximation is the poorest.

The log-normal behaviour of the number of visited pages is also detected in the time between pages, which is heavily dependent on the reading time of the Web surfers. The log-normally distributed character of these two variables can be justified by the logarithmic characteristics of the human perception of time [4]. On the contrary, the variability of the number of connections per page stems from other non-psychological factors, mainly the number of objects in the page but also the performance of the cache memory and the degree of re-utilisation of persistent connections. Hence, this parameter exhibits a clear heavy-tailed nature. The presence of the 'Noah effect' or syndrome of the infinite variance is evidenced by a high ratio between the deviation and the mean value as well as between the mean and the median. Table 2 proposes to model this variable through a Pareto distribution for which the estimated value of α is below 2, which implies an infinite variance.

Table 2. Individual model of IP clients

Level	**SESSION**	**PAGE**		**CONNECTION**			
	No. of visited pages (visited pages/ hour)	**Time between pages of the same client (s)**	**No. of initiated connections/ page**	**Total Size of the connections in the sense Server-Client (Useful size) (No. of packets)**	**Connection Duration** (s)**	**Time between the beginning of consecutive connections (in seconds):**	
						Of the same client and page	**Of the same client**
Mean	3.43 (17.91 p/h)	315.5007	9.5132	17.9644 KByte (17.4891 KByte) (19.0297 pack.)	17.7396	4.2259	28.1338
Standard Deviation	3.80 (15.05 p/h)	366.3200	159.1550	342.2354 Kbyte (332.5738 KByte) (241.9442 pack.)	56.3309	9.8528	131.4023
Median	2.00 (13.19 p/h)	168.4999	1.0000	1.2960 (1.2590) (5 packets)	1.8524	0.3883	0.4773
Proposed model	LOGNORMAL (μ_Y=2.6181, σ_Y=0.7309)	LOGNORMAL (μ_Y=5.3274, σ_Y=0.9239)	PARETO (α=1.0439, β=0.41167)	PARETO (α=1.4890, β=8784.16)	PARETO (α=1.5149, β=9.1336)	WEIBULL (a=1.9978, b=0.4854)	PARETO (α=1.2711, β=7.7950)

Table 3. Standard distributions employed in the analysis (μ_X and σ_X are the mean and deviation of the series to approximate)

Distribution	**Marginal density function**	**Parameters**	**Adjustment**
Exponential	$f_E(x) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)$	μ	$\mu=\mu_X$
Gamma	$f_\Gamma(x) = \frac{a^b}{\Gamma(b)}\exp(-a\cdot x)\cdot x^{b-1} \quad si \quad a,b \geq 0; \quad 0 \leq x$ where $\Gamma(b) = \int_0^\infty y^{p-1}\exp(-y)dy$	a, b (Scale and shape)	$a = \frac{\mu_X}{\sigma_X^2}; \quad b = \frac{\mu_X^2}{\sigma_X^2}$
Weibull	$f_W(x) = \frac{b}{a^b}\cdot\exp\left[(-\frac{x}{a})^b\right]\cdot x^{b-1} \quad if \quad 0 \leq x \leq \infty$	a, b (Scale and shape)	Solving: $\mu_X = a\cdot\Gamma\left(\frac{b+1}{b}\right)$ $\sigma_X = a^2\cdot\left[\Gamma\left(\frac{b+2}{b}\right)-\Gamma\left(\frac{b+2}{b}\right)^2\right]$
Log-normal	$f_{LN}(x) = \frac{1}{x\sqrt{2\pi\sigma_Y^2}}\exp\left[\frac{-(\ln x-\mu_Y)^2}{2\sigma_Y^2}\right], \quad 0 < x$	μ_Y, σ_Y	$\mu_Y = \frac{1}{2}\log\left(\frac{\mu_X^4}{\mu_X^2+\sigma_X^2}\right)$ $\sigma_Y = \log\left(1+\frac{\sigma_x^4}{\mu_X^2}\right)$
Pareto	$f_P(x) = \frac{\alpha}{\beta^\alpha}\cdot(\beta+x)^{-\alpha-1} \quad si \quad 0 \leq x$	α, β	α calculated by means of a regression in the tail of the distribution to match $\beta = \mu_X\cdot(\alpha-1)$

3.2 Connection level

At this level it is necessary to characterise three parameters: size, duration and time between the arrivals of client connections. The size of the connections is directly connected to the size of the files (objects) distributed in the Web, even though they are modulated by the preferences of the users and the effects of the cache [11]. The analysis of the traces corroborates the heavy-tailed nature of this parameter which has been repeatedly reported in the literature as the main cause [19] of the self-similarity or Long Range Dependence (LRD) present in the Web traffic. Figure 3 proves that this heavy-tailed nature is exhibited not only by the total downlink size of the connections (bytes transferred from the server) but also by the effective or useful size of the connections (without considering the retransmitted bytes). Even though it has been shown that retransmissions can introduce by itself LRD properties in TCP traffic [21], from Figure 3 it could be deduced that TCP protocol has a minor impact on the distribution of the connection size, at least for the range of the losses measured in the traces (above 2-3% of the bytes, a typical value in most transfers of current Internet).

With reference to the duration of the connections, this parameter basically depends on the number of transported packets (which is in turn proportional to the connection size), the delay between the end systems (described by mean of the RTT or Round Trip Time) and the behaviour of TCP protocol. The linear relationship that could be initially presumed between the duration and the connection size is altered for the reactive control of TCP, especially in the presence of losses or small sized connections that do not surpass the initial phase of slow start. Figure 4 represents the mean estimated duration of the connections as a function of their downlink size. The figure shows that the correlation between duration and size is different depending on the transmitted bytes. For small connections (under 10 Kbytes, which corresponds with connections with less than eight 1500 byte packets), the slow start phase seems to set up a non linear association between the variables. Oppositely, for connections larger than 10 Kbytes the increase of the duration smoothes and tends to a linear dependence. Given the hyperbolic decay of the tail of the size distribution, this linearity between the duration and the large connection sizes could justify the heavy-tailed nature detected in the duration distribution, which is proposed to be modelled through a Pareto function.

On the other hand, an accurate model for the temporisation of the connections can be critical as long as the bursty nature of connections arrivals may influence the performance of routers' CPU as well as the policies of bandwidth provisioning in the network nodes [12]. As it has been commented for the page level, in contrast with other TCP services with longer and fewer connections per session (such as FTP, Telnet or file transfer applications), the opening and closing of a connection in a HTTP service rely on a wide set of factors. Among these factors we can include the human actions (the habits when browsing, the reading times, etc) but also the nature of the contents and the particularities of the implemented protocols (HTTP and TCP) in the extremes. This variability of the factors that impacts on the connection dynamics at different time scales gives an explanation for the heavy-tailed nature of the time between connections of each client (see Table 2). Nevertheless, within each page, the decision of opening (or reusing) a connection is essentially governed by the ability of the browser to parse an HTML document and send GET commands to import its embedded objects (the 'active' OFF periods, as they are called in [2]). So, the interarrival times of the connections within the same page present a lower variability which can be characterised by a Weibull distribution, as it is also proposed in [12].

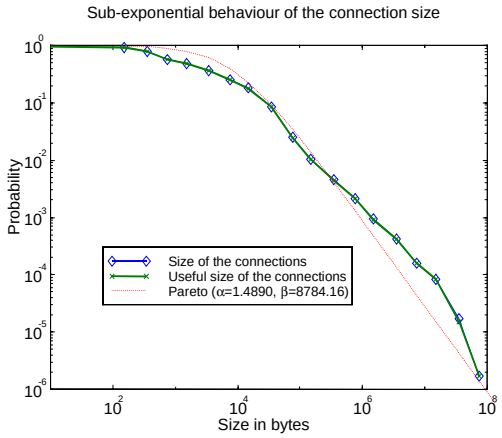

Figure 3. Complementary cumulative distribution of the total and useful size of the connections (considering the six traces)

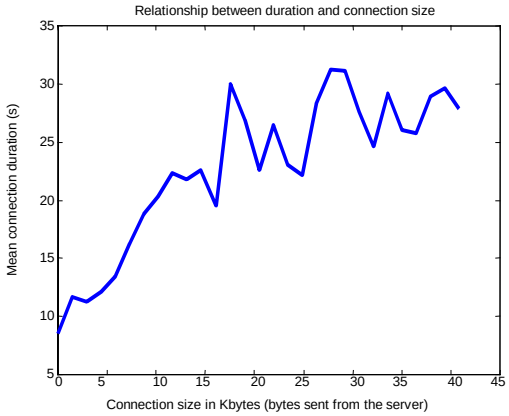

Figure 4. Relationship between the mean duration and the downlink connection size (Note: connections are grouped in intervals of 1500 bytes)

Moreover, as it is shown in Figure 5, these interarrival times between neighbouring connections exhibit a higher autocorrelation which may be necessary to model and which would be neglected if the page level is not considered.

Another way of observing the strong intermittency of Web activity is to study the degree of simultaneity of the connections. Table 4 presents the distribution of the client session duration according to the number of simultaneous active connections. This Table shows that the client is inactive about 75% of time while during 1% of time clients keep opened six or more connections, generating sudden and sporadic traffic peaks.

On the other hand, the presence of elements such as NAT servers or proxies is straightforwardly verified by inspecting the maximum number of parallel connections that each IP address is capable of keeping open. We must bear in mind that for commercial browsers this number is restricted to small values. For example, Netscape permits up to 4 simultaneous connections to the same server while the specifications of 1.1 version of HTTP recommend not opening more than two persistent connections to same server [13]. Under these considerations, the presence of Web clients with more than 200 simultaneous connections (see Table 4) confirms the existence of traffic aggregates through a single IP address.

These percentages indicate a higher intermittency than that reported in [7]. Basing on traces monitored in a LAN, authors in [7] register than users keep inactive about 40% of the time. In any

case, these long inactive periods confirm the On-Off pattern followed by individual Web traffic. If the On periods, which are related to the duration of the connections, have been shown to be heavy-tailed, a complete On-Off model would require to characterise the distribution of these idle periods. In this sense, [2] and [10] suggest a sub-exponential nature for this variable, which is also detected in our traces (see the approximations in Figure 6). The heavy-tailed characteristics of the Off state offer another base [20] for the LRD properties present in Web traffic, normally founded on the distribution of the connection sizes, which definitively determine the duration of the activity (On) periods.

3.3 Packet level

At this level the parameters to characterise are the packet size and the time between packets. Table 5 shows that the packet sizes flowing from the server practically follow a multimodal distribution which is determined by the existence of the following typical sizes:

- Void packets (0 bytes) provoked by TCP messages (mainly acknowledgment packets)

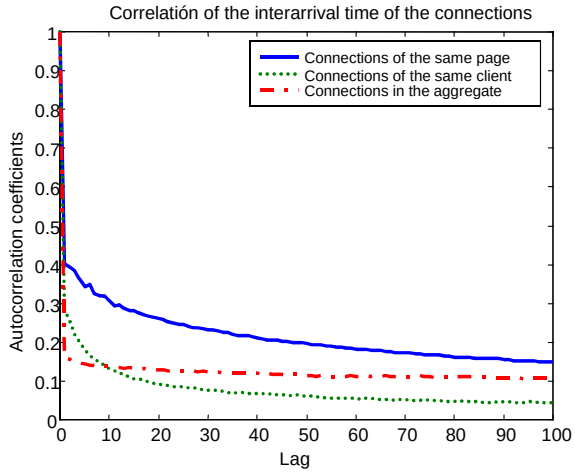

Figure 5. Autocorrelation coefficients of the interarrival time between connections

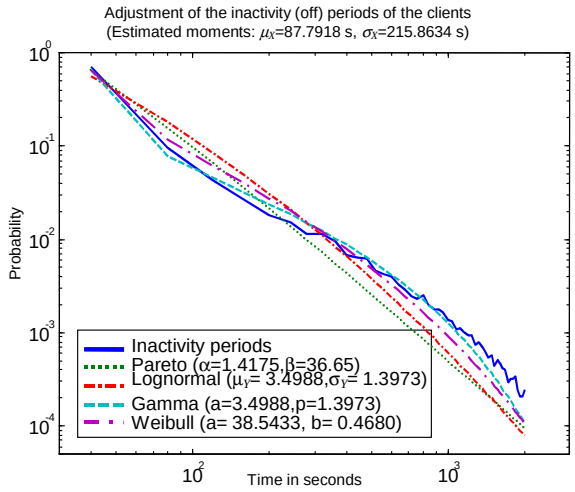

Figure 6. Subexponential Adjustment o of the density distribution of the inactivity periods of the clients

- Packets with 1460 bytes. This size, which accounts for more than 40% of packets, is the result of the limit imposed by the MTU (Maximum Transfer Unit) under Ethernet networks (1500 bytes: 1460 byte payload plus the 40 bytes TCP/IP header), which are definitely the most extended LAN technology.
- Packets with 1448 bytes, corresponding to those packets with the maximum Ethernet size and a TCP/IP header of 52 bytes containing an extra 12 bytes timestamp (in the Options field of the TCP header) for a better estimation of the RTT.
- Packets of 512 and 532 bytes, which follow the MSS recommended by the IETF [3]. This recommendation is implemented in certain operating systems such as in the SUN UNIX station when sending packets to non local IP addresses.

A previous study [20] proposes a similar distribution for this variable. Nevertheless Table 5 illustrates the decline of the packets of 512 bytes and the increasing presence in our traces of the TCP packets that incorporate the timestamp, which were practically inexistent some years ago.

As it refers to the time between packets of the same client, the packet interleaving performed at the routers makes the low values of the interarrival times be very dependent on the capturing point and on the effects of client aggregation. Conversely, the frequent inactivity periods introduce high peaks in this parameter which clearly affect the estimation of the first and second moments. For this reason Table 5 also shows the statistics if we neglect these values and only consider interarrival times under 60 s (the threshold that was chosen to separate the pages). In any case, just Pareto functions are able to approximate the strong variability that this parameter exhibits.

4. EFFECTS OF AGGREGATION

From the previous section we can conclude that Web traffic of individual clients experience a high variability at several time scales. This variability, which has been corroborated by different studies in the literature, is noticeable by the high values of the ratio between the two first central moments of different variables. However a still open issue is to calibrate the effects of multiplexing clients on the traffic properties and to determine the time scale and the degree of aggregation at which this variability begins to be mitigated.

The analysis of the aggregate in the employed traces (see Table 6 and Figure 7) proves that the aggregation of clients produces an exponential tendency in the interarrival times at all the levels (sessions, pages, connections, packets). At the session level (which can be assimilated to the call level in a phone network) this exponentiality was expectable because of the independence of client arrivals and has been reported in other works on Web traffic modelling such as [20]. In the case of the interarrival time of pages and connection, the results show that the variability that individual browsing imposes is smoothed, so that the distributions also tend to an exponential distribution even though this convergence is slower as the considered scale (session, page or connection) diminishes. Particularly, the correlation of the connection arrivals within each page could explain that the time between aggregated connections slightly differs from an exponential evolution.

As it refers to the aggregation at the packet level, [5] asserts that the time between packets tends to an exponential distribution as the aggregation increases. In particular, it analyses the ability of the Weibull distribution to match this parameter in different set of traces with diverse degree of multiplexing. The study concludes that when the packet rate is above 3000 packets/second the shape parameter (*b*) of the Weibull approximation is close to 1, indicating that it is an exponential variable. In our trace the measured rate is

around 900 packets/s, which justifies that the variable still presents some sub-exponentiality.

This fact evidences that at the packet level the aggregate has not completely smoothed the heavy intermittency of the individual sources (see Figure 8).

Table 4. Degree of simultaneity of the connections within client sessions

Maximum detected number of simultaneous connections	Distribution of the session time according to the number of simultaneous opened connections						
	0	**1**	**2**	**3**	**4**	**5**	**6 or more**
205	75.59 %	12.91 %	7.24 %	1.59 %	1.33 %	0.37 %	0.97 %

Table 5. Individual and aggregate model at the packet level

Distribution of the packet size (TCP payload) of the HTTP flows in the sense server-client Mean size=885.11 bytes			**Statistics of the time between packets (in seconds): (In parenthesis the results if only values under 60 s are considered)**			
0 bytes	**512 bytes**	**536 bytes**	**Statistic**	**Aggregate**	**Of the same client**	**Of the same connection**
20.32%	1.11%	10.71%	**Mean value**	1.0808 ms	0.4144 s (0.1734)	5.2854 s (0.4818 s)
			Deviation	1.7267 ms	12.4608 s (1.5565 s)	76.42 s (2.8312 s)
1448 bytes	**1460 bytes**	**Others**	**Median**	0.3660 ms	0.0016 s (0.0016 s)	0.0027 s (0.0024 s)
3.55%	42.58%	21.73%	**Proposed model**	WEIBULL ($a=7.8974\cdot10^{-4}$, $b=0.6490$)	PARETO ($\alpha=1.6818$, $\beta=0.3285$)	PARETO ($\alpha=1.0856$, $\beta=0.0148$)

Table 6. Modelling of aggregate traffic

Level	**SESSION**	**PAGE**	**CONNECTION**
	Time between the beginning of client sessions *	**Time between the beginning of the pages of all the clients**	**Time between the beginning of the connections of all the clients**
Mean value	0.5804 s	0.1241 s	0.0130 s
Deviation	0.5993 s	0.1303 s	0.0166 s
Median	0.3881 s	0.0834 s	0.0076 s
Proposed model	EXPONENTIAL (μ=0.5804)	EXPONENTIAL (μ=0.1241)	EXPONENTIAL (μ=0.0130)

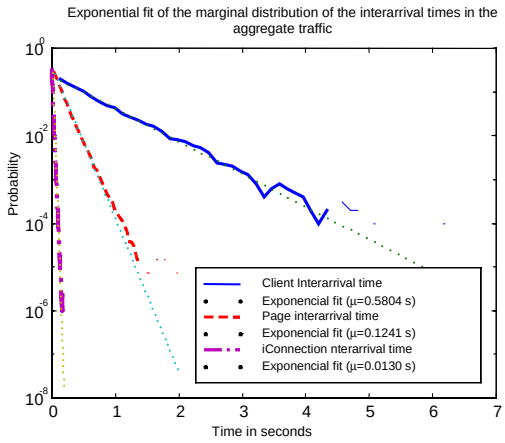

Figure 7. Exponential adjustment of the distribution of the interarrival time of clients, pages and connections in the aggregate traffic

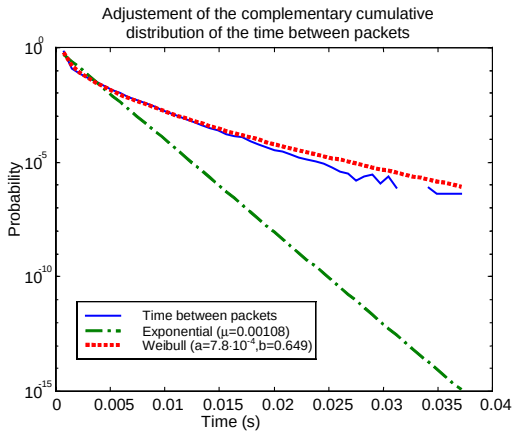

Figure 8. Adjustment of the density distribution of the time between packets (aggregate traffic)

In order to analyse the effect of the aggregation and the potential fractal nature of traffic, we define the variable $N_i^n(T)$ as the number of initiated connections by n clients during the *i-th* interval of duration T. Thus, the variation coefficient $\gamma^n(T)$, defined as the ratio between the standard deviation and the mean value of $N_i^n(T)$, offers an interesting measurement of the variability of this parameter:

$$\gamma^n(T) = \frac{\sqrt{Var(N_i^n(T))}}{E\left[N_i^n(T)\right]} \quad (1)$$

This coefficient must accomplish:

$$\gamma^n(T) \underset{T\to\infty}{\propto} T^{H-1} \quad (2)$$

where H is the so-called Hurst parameter, which describes the self-similarity in the series. If the aggregate provoked a Poissonian behaviour (with exponential interarrival times), H would tend to 0.5 (no self-similarity) as n increases, so that:

$$\gamma^n(T) \underset{\substack{T\to\infty \\ n\to\infty}}{\propto} \frac{1}{\sqrt{T}} \quad (3)$$

If we assume that the client arrivals are independent and identically distributed and that the two first central moments of the count process $N_i^1(T)$ for one client are finite, we can apply the Limit Central Theorem (LCT), resulting that:

$$\gamma^n(T) \underset{n\to\infty}{=} \frac{1}{\sqrt{n}}\gamma^1(T) \quad (4)$$

Aiming to evaluate the validity of the two previous equations, we separate the traffic of each client in the traces. Then we compare the variation coefficients when applied to the traffic resulting from multiplexing different number of clients. In particular, for the hour in which more connections were detected we arranged the clients according to their arrival time and performed a decimation in such a way that they were homogeneously distributed in new reduced traces (with less clients). Following this policy we considered combinations of 10, 100 and 1000 clients. These combinations offered a mean load of 3.3, 33 and 333 simultaneous clients, respectively.

The results of the variation coefficient for the decimated and global traces are depicted in Figure 9. In this figure it is observed that the traffic gradually adopts a Poissonian behaviour as the aggregation grows, which is manifested in the exponential decay with T (linear in a logarithmic scale) of $\gamma^n(T)$ for the trace containing all clients.

Figure 10 shows the results of the previous experiment when the variation coefficient is referred to $X_i^n(T)$, defined as the number of bytes received by n HTTP clients during the *n-th* time interval of duration T. In this case, the traffic volume is regulated not only by the connection arrivals but also by the duration and, especially, the connection size (two parameters which present an intense heavy-tailed distribution). Under these conditions, the Web traffic can be perceived as the aggregate of On-Off sources with heavy-tailed distributed On and Off periods. This structure of traffic generation intrinsically causes a self-similar nature on Web traffic which is not moderated by the increase of the number of multiplexed sources [21].

Anyhow it must be remarked that, following the LCT, figures 9 and 10 show that the aggregation always provokes an attenuation by $1/\sqrt{n}$ of the value of $\gamma^n(T)$ and, consequently, of the traffic variability itself.

So, for all the analysed time scales, the value of $\gamma^n(T)$ drops under the unity since a few hundreds of clients are concurrently multiplexed. This implies that the unpredictability of the bandwidth requirements is drastically reduced (with values for the deviation under the mean), what should be kept in mind when dimensioning a link designed to support a certain degree of client aggregation.

On the other hand, this work has neglected the effects of the non-stationarity of the traffic. This non-stationarity, especially the seasonality of the Web traffic, is evident for temporal scales longer than one hour. As a result, the time of the day significantly affects the central moments of the parameters previously studied. For example, the study in [1] compares the hourly evolution of the interarrival time of the connections in a MAN network, showing that the moments change while the distribution shape is not modified by the time. So, in any case, a global characterisation of Web traffic should incorporate the modulating effect of this seasonality on the moments of the proposed matching distributions.

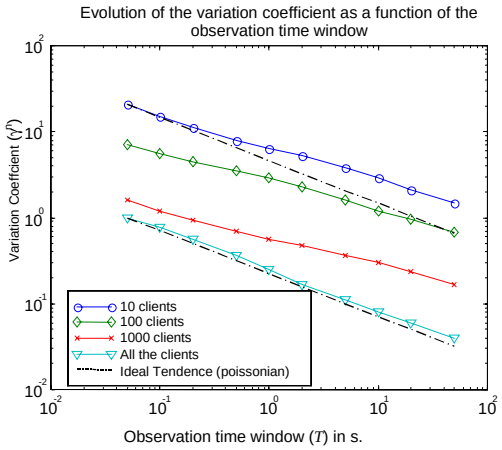

Figure 9. Effect of the client aggregation on the evolution of the variation coefficient (γ) of the number of initiated connections as a function of the observation time window.

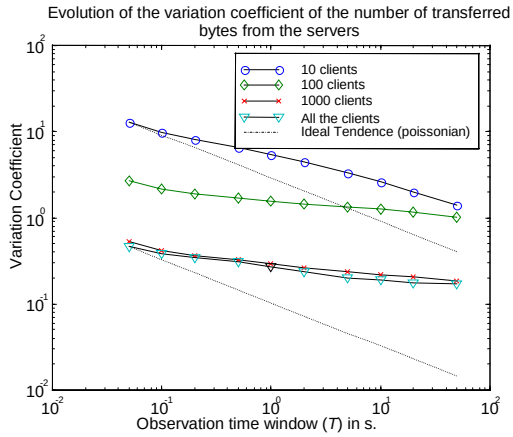

Figure 10. Effect of the client aggregation on the evolution of the variation coefficient (γ) of the number of transmitted bytes from the server as a function of the observation time window.

5. CONCLUSIONS

This work has developed a multi-scale model of the traffic induced by Web clients. The model of individual clients, which contemplate the levels of session, page, connection and packets, is parametrised basing on different traffic traces monitored in a transoceanic link between Japan and USA. From the parametrisation, for which a set of possible standard distributions is considered, it can be deduced that those variables which rely on the

human perception of time (time between pages or number of pages per session) present a log-normal distribution. On the contrary, those parameters depending on the distribution and typology of the Web contents as well as on the particular implementations of the software (connection size or duration, time between connections,…) exhibit the syndrome of the infinite variance or, at least, a high variability that requires to be modelled by distributions such as Pareto.

Similarly, it was shown that the client aggregation does not reduce the LRD nature of the Web traffic but limits its variability in absolute terms. Moreover, multiplexing of individual sources provokes a tendency to Poisson in the interarrival time of the considered traffic units (clients, pages, connection, packets). This tendency is revealed to be slower as the time scale decreases. So for certain scales (packets, for example), if the number of multiplexed clients is not enough, the interarrival time is better modelled through a Weibull distribution.

From a practical point of view, the drastic reduction of the sub-exponentiality that is achieved through multiplexing, could allow to employ Poissonian assumptions in many teletraffic issues such as the management of connections in core routers.

6. ACKNOWLEDGMENTS

This work was partially supported by the project No.TEL99-0755. We also wish to express our gratitude to Dr. Kenjiro Cho and MAWI (Measurement and Analysis on the WIDE Internet) Working Group (Japan) for releasing the Web traces

7. REFERENCES

[1] J. Aracil and D.Morató, “Characterizing Internet Load as a Non-regular Multiplex of TCP Streams”, *Proceedings del International Conference on Computer Communications and Networks (ICCCN 2000)*, Las Vegas, Nevada (USA), October, 2000, pp. 94-99.

[2] P. Barford, *Modeling, Measurement and Performance of World Wide Web Transactions*, Ph.D. Thesis, Boston University (USA), 2001.

[3] R.Braden, “Requirements for Internet Hosts – Communication Layers”, RFC 1122, IETF, October, 1989.

[4] 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.

[5] J. Cao, W. S. Cleveland, D.Lin and D. X. Sun, ”Internet Traffic Tends Toward Poisson and Independent as the Load Increases”, in *Nonlinear Estimation and Classification*, Springer, New York, 2002.

[6] E. Casilari, A. Reyes-Lecuona, F.J. González-Cañete, A. Díaz-Estrella and F. Sandoval, "Characterisation of Web Traffic", *Proceedings of IEEE GLOBECOM 2001*, San Antonio (Texas, USA), November, 2001,

[7] J. Charzinski, “Measured HTTP Performance and Fun Factors”, *Proceedings of the 17th International Teletraffic Congress (ITC'17)*, Salvador (Brasil), December, 2001, pp.1063-1074.

[8] H. Choi and J. Limb, “A Behavioral Model of Web Traffic”, *Proceedings of International Conference of Networking Protocol’99 (ICNP 99)*, Toronto (Canada), September, 1999.

[9] K. Cho, K. Mitsuya and A. Kato, ″Traffic Data Repository at the WIDE Project″, USENIX 2000 FREENIX Track, San Diego, CA, June, 2000. Available traces at MAWI group Web page in: http://tracer.csl.sony.co.jp/mawi/

[10] M. E. Crovella and A. Bestavros, “Explaining World Wide Web Traffic Self-Similarity”, Technical Report TR-95-015, Boston University (USA), August, 1995.

[11] 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

[12] A. Feldmann, "Characteristics of TCP Connection Arrivals", ATT Technical Report, December,1998.

[13] R. Fielding, J. Gettys, J. Mogul, H.Frystyk Nielsen and T. Bernes-Lee, “HTTP Version 1.1”, RFC 2068, IETF, January 1997.

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

[15] B.A. Mah, “An Empirical Model of HTTP Network Traffic”, *Proceedings of the IEEE INFOCOM’97*, Vol. 2, Kobe (Japan), April, 1997, pp. 592-600.

[16] M. Molina, P. Castelli, and G. Foddis, “Web Traffic Modeling: Exploiting TCP Connections’ Temporal Clustering through HTML-REDUCE”, *IEEE Network*, May/June, 2000,pp. 46-55.

[17] Nielsen//NetRatings Inc., “Global Internet Usage”, available data at http://www.nielsen-netratings.com/

[18] A. Reyes Lecuona, E. González, E. Casilari, J.C. Casasola and A. Díaz Estrella "A Page-oriented WWW Traffic Model for Wireless System Simulations", *Proceedings of the 16th International Teletraffic Congress (ITC'16)*, Edinburgh (United Kingdom), June, 1999, pp. 1271-1280

[19] K. Park, G. Kim, and M. Crovella, “On the relationship between file sizes, transport protocols, and self-similar network traffic”, *Proceedings of International Conference of Networking Protocol’96 (ICNP 96)*, Columbus (Ohio, USA), October, 1996, pp. 171-180.

[20] B. Ryu and S. Lowen, "Fractal Traffic Models for Internet Simulation", *Proceedings of the Fifth IEEE Symposium on Computers and Communications (ISCC 2000)*, IEEE Computer Society Press, Los Alamitos (CA,USA), July, 2000, pp. 200-206.

[21] B. Sikdar and K. S. Vastola, “The Effect of TCP on the Self-Similarity of Network Traffic”, *Proceedings of 35th Conference on Information Sciences and Systems*, Baltimore (MD, USA), March, 2001.

[22] Tcpdump, available software at http://www.tcpdump.org/

[23] Tcptrace, available software at http://irg.cs.ohiou.edu/software/tcptrace/tcptrace.html

[24] The Network Simulator, ns-2, available software at http://www.isi.edu/nsnam/ns/

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