# A page-oriented WWW traffic model for wireless system simulations

A. Reyes-Lecuona, E. González-Parada, E. Casilari, J. C. Casasola and A. Díaz-Estrella

Dpto. Tecnología Electrónica, E.T.S.I. Telecomunicación,
Universidad de Málaga, Campus de Teatinos, 29071 Málaga (Spain)*
Telephone no.: 34-952132755; FAX 34-952131447; E-mail: arcadio@dte.uma.es

This paper presents a new, page-oriented model of WWW traffic, based on real HTTP traces taken from corporate and educational environments. It was designed to generate synthetic workload traffic for simulations of wireless systems. Taking into account the characteristics of this kind of simulation, we designed a structural, multilevel model based on the behaviour of WWW users. A three-level structure was created, considering session, page and packet levels. A detailed traffic characterization is presented for both corporate and educational cases in order to fit the model parameters. We also present a detailed comparison between both environments within all the considered time scales. Finally, we report simulations comparing our proposed model empirically with the WWW traffic model proposed by ETSI, and show that the ETSI model underestimates packet losses and delay.

## 1. INTRODUCTION

Internet users are increasingly demanding that current Web services be available in non-fixed locations (out of the office, travelling, etc.) or with a wireless access network (at home, at work, etc.). The benefits of mobility are obvious and the advantages of avoiding wires in a building network or residential zone are also clear. Moreover, wireless connection gives users access in places where installation of cables is not feasible or profitable (remote places, old parts of a city, etc.). The current bandwidth limitations of wireless access are only important for a few Internet services, and will be overcome by innovation.

For performance evaluation of the different implementation alternatives of previous systems, it is necessary to carry out tests and measurements that are difficult or not feasible in real scenarios. Hence, the definition and validation of resource management techniques are usually based on simulations of the access system. In this context, the World Wide Web (WWW) is the major source of traffic in the Internet.

In order to acquire useful data from simulations, an accurate model of Internet client traffic is required. Recent studies [1, 3] show that Poisson-based modelling is inappropriate, for either individual connections or aggregated traffic, as Internet aggregated traffic exhibits long range dependence (LRD).

Several models have been proposed in the literature [1, 4, 8], but some are very complex, or are not well supported by measurements, or only model an educational environment. The

* This work was partially supported by the Spanish Comisión Interministerial de Ciencia y Tecnología (CICYT), Project No. TIC98-1040. We also wish to express our gratitude to the Radio Department of Alcatel ALSTHOM (Madrid, Spain) for enabling the capture of HTTP traces there.

European Telecommunication Standardisation Institute (ETSI) also proposed a WWW traffic model in its recommendation [5]. In this paper we propose an alternative structural model and compare it with the ETSI version.

This paper is organized as follows: section 2 describes the structure of WWW traffic, yielding a model structure appropriate for simulations of wireless systems; section 3 characterizes the traffic in both corporate and educational environments; section 4 presents implementation details and simulation results, together with comparisons with the ETSI model. Finally, in section 5 we present some conclusions about this work.

## 2. WWW TRAFFIC STRUCTURE

There are basically two approaches to data traffic modelling: the behaviourist or *black-box* approach, in which the statistical characteristics are modelled without taking into account the causes that lead to them; and the *structural* approach, in which the model design is based on the internal structure of the traffic generating system [13].

WWW traffic has very complex characteristics with LRD, and some authors have therefore postulated the existence of fractal or self-similar behaviour [7].

Fractal modelling of network traffic can be performed from various points of view, but all of them require fitting self-similar parameters, such as the Hurst (H) parameter, to a fractal process. Among the most important fractal processes are Fractional Gaussian Noise (FGN), Fractional Auto-Regressive Integrated Moving-Average (FARIMA) processes, and Fractal Renewal Processes (FRP). They all need the adjustment of parameters, such as H, with very obscure physical meanings. An advantage is that FRP and FGN processes are very parsimonious, needing very few parameters to model LRD. FARIMA processes need more parameters, but are capable of modelling both LRD and short-range dependence.

Internet traffic is evolving rapidly. It is therefore very important that modelling parameters should have physical meaning, because they can then be used to forecast the future evolution of traffic. An additional complication is that in the design of wireless access systems, it is necessary to perform simulations at different time scales.

For these reasons, we chose a structural multilevel approach to model WWW traffic, based on user behaviour when browsing the Web. This multilevel structure can fit LRD with its longer time scales and can fit short-range dependence with its shorter time scales, and all the parameters will have clear physical meanings because of the structuralist philosophy. In this way, it is possible to estimate models for future WWW traffic to the extent that we can forecast the evolution of its parameters.

The first step in constructing our model was to study the way in which WWW users generate traffic by browsing the Web. A user begins a WWW session by running a web browser. We considered a session to be the period of time between that starting moment and the exit from the browser. Within a session, the user visits several Web pages and reads them. Therefore, a session is really a set of page transfers separated by reading time.

Transfer of a Web page consists of several TCP connections that run in parallel. In turn, TCP connections are composed of IP packets. As the TCP connections for a page are simultaneous in time, we considered a page as a set of IP packets, including control packets, thus avoiding the complexity of managing TCP connections.

Once we had considered all these issues, we designed a three-level structural model, based on the behavior of a WWW user. We describe the different levels taken into consideration in table 1. In this way, a pair of parameters, as table 2 shows, can define each level.

Table 1
Level description

| LEVEL | DESCRIPTION |
|---|---|
| Session level | Working session of a user with a WWW browser: from the time of starting the browser to the end of the navigation. This level consists of pages, considered to be non-overlapped. |
| Page level | Visit to a Web page, considering a page as the set of files (HTML, images, sounds, videos, etc.) composing a Web page. |
| Packet level | Transmission of an IP packet. This is the lowest level. IP packet arrivals are assumed to be a point process, fully defined by the interarrival time and the packet size. |

Notice that uplink and downlink traffic are considered separately at the page and packet levels. We can therefore build two models: one for uplink traffic, and one for downlink traffic. Downlink traffic is different from uplink traffic; in fact, the largest part of the transferred information is in downlink direction. Thus, if the communication system is symmetric, the most critical direction is the downlink one. Our simulations were made with a downlink model. However, in asymmetric systems, and especially in dynamic bandwidth allocation systems, both directions need to be analysed separately.

Table 2
Parameter definition

| LEVEL | PARAMETER | DEFINITION |
|---|---|---|
| Session level | Session interarrival time | Time between commencements of two consecutive sessions |
| | Pages per session | Number of Web pages that the user visits in a session |
| Page level | Time between pages | Time between the end of one page and the beginning of the next in the same session. |
| | Page size | Total amount of information transferred per page. In this case, we can distinguish between uplink and downlink information. |
| Packet level | Packet interarrival time | Time between two consecutive packets inside the same page, and in the same direction. That is to say, we treat uplink and downlink packet interarrival times separately. |
| | Packet size | Number of bytes contained in each packet. Again, we consider uplink and downlink packets separately. |

## 3. WEB TRAFFIC CHARACTERIZATION

Once the model structure is defined, we need to fit probability density functions (pdf) to each parameter. This characterization was based on real traces captured from two different environments. We collected data from the University of Málaga, as an educational environment, and from Alcatel Alsthom in Madrid, as a corporate one. The main characteristics of these traces are shown in table 3. We considered all TCP/IP packets containing HTTP transactions (port 80). We obtained a record for each IP packet using a public domain sniffer. In both working environments, these protocols resided on an Ethernet network.

Table 3
Trace characteristics

| Environment | Study period | No. of users | No. of sessions | Time resolution |
|---|---|---|---|---|
| Alcatel | March and April, 1997 | About 2000 | 10198 | 52 ms |
| Univ. of Málaga | February, 1998 | About 150 | 593 | 52 ms |

In order to determine when a user started a session we defined a time threshold between two consecutive packets, belonging to the same user. If the time between two consecutive packets, belonging to the same user, exceeded this threshold, we considered that a new session had begun. We fixed that threshold to 1800 s.

We defined a page as the set of packets that belong to TCP connections starting close together in time. Two TCP connections that began less than 30 s apart were considered as belonging to the same page. Using this value in the captured traces, we estimated that, after the start of a new page, the number of received packets from a previous page of the same user was less than 1% of total. We therefore considered pages to be non-overlapping, and for modelling purposes these overlapped packets were construed as included in the current page. The end of a page was set at the last packet received before the beginning of the next page. This is illustrated in figure 1.

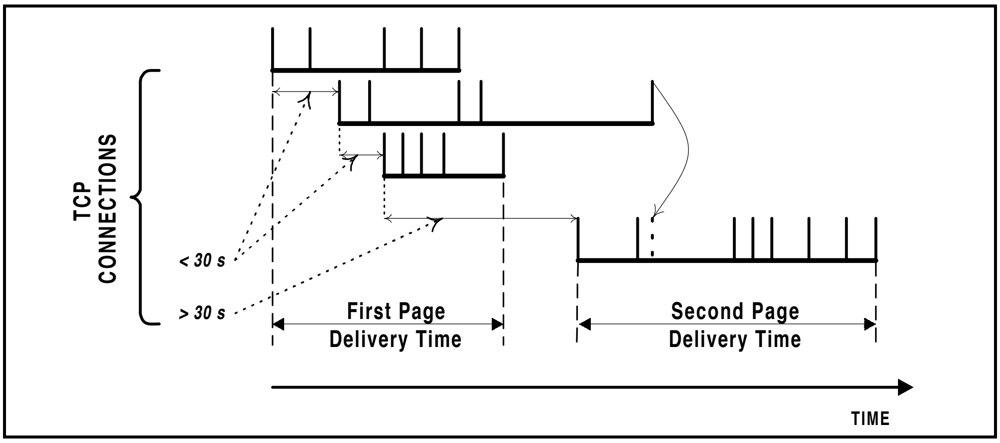


Fig. 1 Page detection strategy.

### 3.1. Session interarrival time

The session arrival process was modelled as a Poisson process, with interarrival time modelled as an exponentially distributed random variable. This is a reasonable assumption since different users initiate different sessions, and its validity was confirmed in our experiments with the collected traces. Thus, interarrival times are statistically independent.

In our model, the mean interarrival time will depend on the desired traffic load.

### 3.2. Pages per session

Table 4 shows the main statistics for this parameter, in both corporate and educational environments. Figures 2 and 3 plot the pdfs of this parameter for corporate and educational cases, respectively. Notice the similarity between them. We can conclude that they could be fitted to the same pdf. The quantil-quantil (Q-Q) plot of figure 4 compares the fits of several standard distributions.

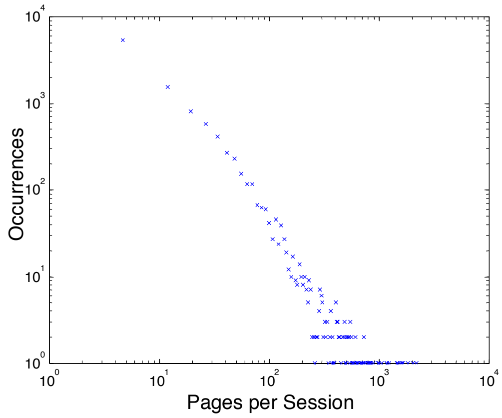


Fig. 2 Pages per session in corp. env.

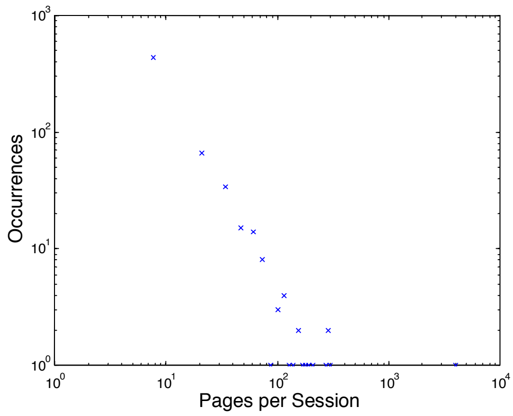


Fig. 3 Pages per session in educational env.

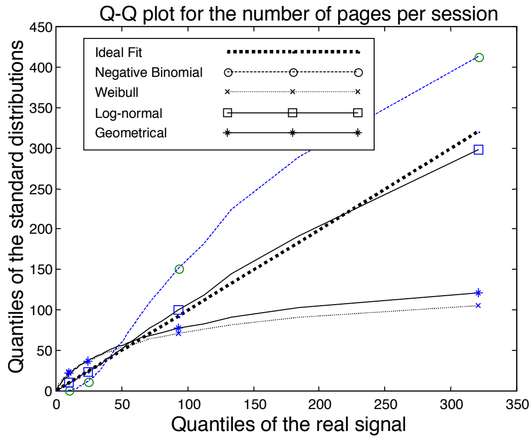


Fig. 4 Q-Q plot for number of pages per session.

Table 4
Pages per session (p.p.s.) main statistics

| | Corporate environment | Educational environment |
|---|---|---|
| Mean | 25.807 p.p.s. | 22.975 p.p.s. |
| Std. Deviation | 78.752 p.p.s. | 166.16 p.p.s. |
| Median | 8 p.p.s. | 5 p.p.s. |

In light of these results, we modelled the number of pages per session as a log normal-distributed random variable, with means and standard deviations as shown in table 4.

### 3.3. Time between pages

Table 5 shows the main statistics of this parameter, in both environments. The similarity of the two environments is evident. Figures 5 and 6 plot the pdf for each environment, and figure 7 shows the Q-Q plot of this parameter in relation to some standard pdf's.

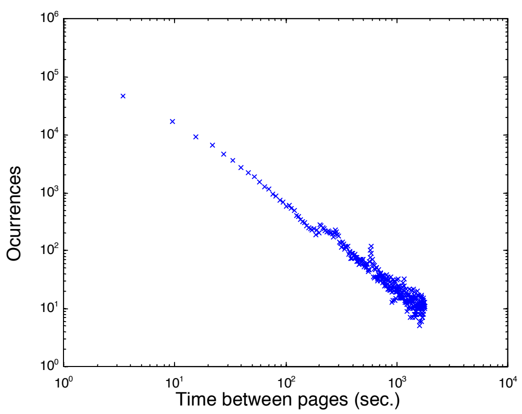


Fig. 5 Time between pages in corp. env.

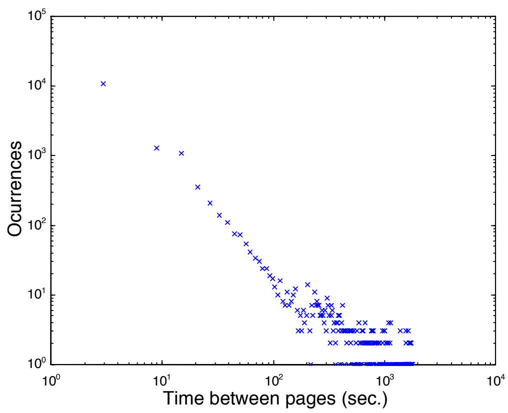


Fig.6 Time between pages in educat. env.

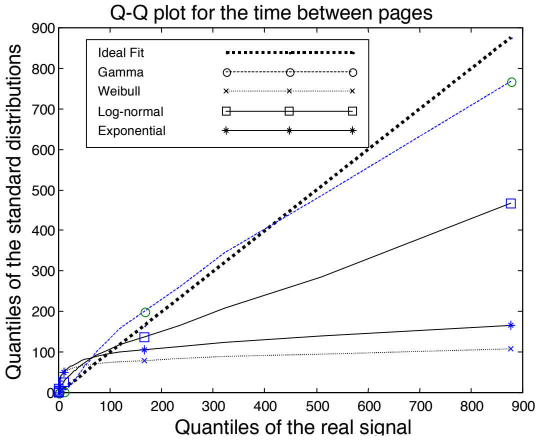


Fig. 7 Q-Q plot for the time between pages.

Table 5.
Time between pages mean statistics

| | Corporate environment | Educational environment |
|---|---|---|
| Mean | 35.286 s | 24.694 s |
| Std. Deviation | 147.39 s | 133.9 s |
| Median | 0.06 s | 0.25 s |

In light of these results, we modelled the time between pages as a gamma-distributed random variable, with means and standard deviations presented in table 5.

### 3.4. Page size

Table 6 shows the main statistics of this parameter (measured in bytes). Notice the huge variability, represented by a large standard deviation and a small median. Figures 8 and 9 show the pdf of this parameter. As we suspected an infinite variance syndrome (or "Noah" effect), we included the Pareto distribution in the Q-Q plot shown in figure 10.

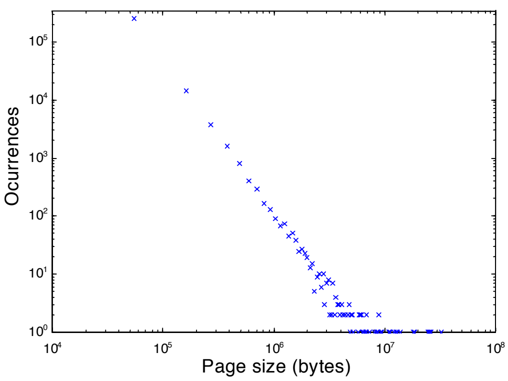


Fig. 8 Page size histogram for corp. env.

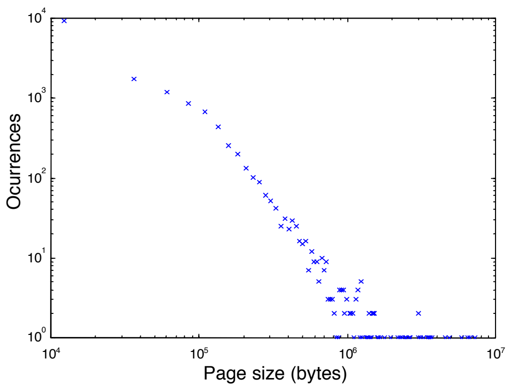


Fig.9 Page size hist. for educational env.

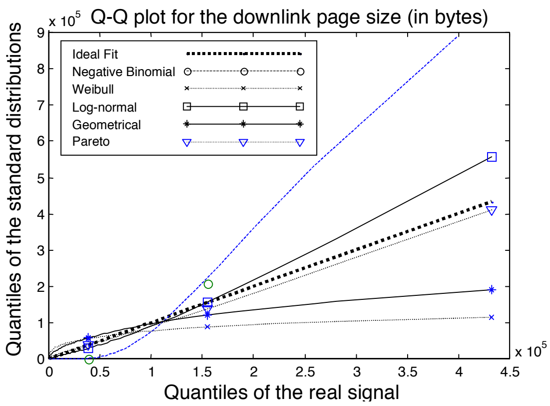


Fig. 10 Q-Q plot for the page size.

Table 6
Page size main statistics

| | Corporate environment | Educational environment |
|---|---|---|
| Mean | 40160 bytes | 56481 bytes |
| Std. Deviation | 1.964 E5 bytes | 1.927 E5 bytes |
| Median | 10269 bytes | 11635 bytes |

These results confirmed our hypothesis and so the page size was modelled with a Pareto-distributed random variable, whose Probability Distribution Function is as follows:

$$F_P(x) = \begin{cases} 1 - \left( \dfrac{\beta + x}{\beta} \right)^{-\alpha} & \text{if } 0 \le x \\ 0 & \text{other case} \end{cases} \tag{1}$$

The $\alpha$ parameter was obtained using the Hill estimator [7], resulting in $\alpha$ and $\beta$ as shown in table 8.

### 3.5. Packet size

This parameter is strongly related to the underlying link layer protocol, being independent of the type of user (corporate or educational). Therefore we studied the global packet size distribution.

Figure 11 shows the packet size histogram of downlink packets, indicating four important packet sizes. Table 7 shows the rate of occurrence of each packet size computed according firstly to all the observed packets (left column), and secondly to the four important packet sizes (right column).

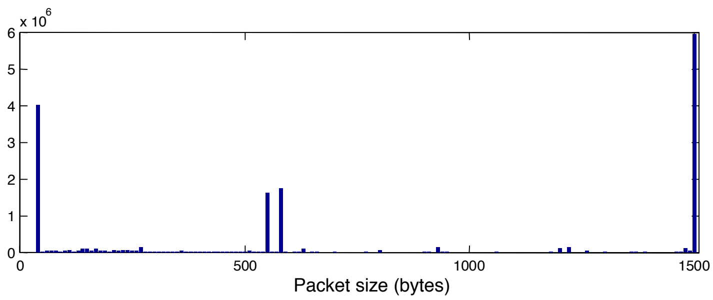


Fig. 11. Packet size histogram.

Table 7
Packet size distribution

| Packet size | % over observed packets | % over selected packets |
|---|---|---|
| 40 bytes | 23.62 % | 30.17 % |
| 552 bytes | 9.61 % | 12.27 % |
| 576 bytes | 10.24 % | 13.08 % |
| 1500 bytes | 34.83 % | 44.48 % |
| Other | 21.7 % | – |

Packets of 40-byte size result from the TCP/IP protocols. They correspond to empty packets, including just the TCP and IP headers.

Packets of 1500-byte size result from the link layer protocol, Ethernet in our case. This is because the maximum Ethernet packet size is 1500 bytes. IP packets are fragmented to avoid Ethernet segmentation.

Packets of 525- and 576-byte sizes result from the IETF recommendation [2].

We modelled this parameter over the four selected sizes with a multimodal distribution that fits the packet size rate, as shown in table 7. In this way, the mean packet size is 822.31 bytes.

Ninety percent of uplink packets are 40 bytes long because they are mostly acknowledgement packets. In a simplified modelling strategy we could consider all uplink packets as 40-byte responses to downlink packets.

### 3.6. Packet interarrival time

It is not appropriate to model this interarrival time in order to fit the measured distribution in the traces under study, because it strongly depends on the link layer protocol and the background traffic in the network.

In addition, as the simulated traffic will feed a queueing system, it is not essential to fit this pdf, since queues cause a low-pass filtering effect, removing high frequency components of the traffic. In any case, it is important that the mean simulated interarrival time fits the measured one.

Therefore, we modelled this parameter with an exponentially-distributed random variable, which provides reasonable variability without adding convergence problems in simulations.

In order to compute the mean interarrival time, it is necessary to measure the mean page delivery time. This time was 35.703 s in the corporate environment, and 73.519 s in the educational one, revealing a difference in the available bandwidth.

The mean packet interarrival time (PIT) can be obtained from this expression:

$$\overline{PIT} = \frac{\overline{PDT}}{\overline{BPP}/\overline{PS}} \tag{2}$$

where PDT is the Page Delivery Time; BPP the Bytes per Page; and PS the simulated Packet Size. The values obtained were 0.73 s and 1.07 s in the corporate and educational environments, respectively.

### 3.7. Parameter distribution summary

Finally, table 8 summarizes all characterization parameters.

Table 8
Parameter distribution summary

| Parameter | Distribution | Main Values In Corporate Env. | Main Values In Educational Env. |
|---|---|---|---|
| Session interarrival time | Exponential | – | – |
| Pages per session | Lognormal | μ=25.807 p.p.s.<br>σ=78.752 p.p.s. | μ=22.975 p.p.s.<br>σ=166.16 p.p.s. |
| Time between pages | Gamma | μ=35.286 sec.<br>σ=147.39 sec. | μ=25.694 sec.<br>σ=133.9 sec. |
| Page size | Pareto | α=1.7584<br>β=30458 bytes | α= 1.5549<br>β= 31341 bytes |
| Packet size | Multimodal | See § 3.5 | See § 3.5 |
| Packet interarrival time | Exponential | μ=0.73 sec. | μ=1.07 sec. |

## 4. SIMULATION AND RESULTS

The proposed WWW traffic model was implemented using a commercial event-driven simulation tool for communication systems. The ETSI model [5] was also implemented for comparison.

The model proposed by ETSI has a structure similar to ours, and models session arrival as a Poisson process as do we. However, the probability density functions used are different for every parameter. The number of pages (called *packet calls* in [5]) per session is geometrically distributed with a mean of 5 p.p.s. The time between two consecutive pages (called *reading time* in [5]) is geometrically distributed with a mean of 412 s. The number of packets in a page is geometrically distributed with a mean of 25 packets. The packet interarrival time is also geometrically distributed with a mean that depends on the channel bandwidth (we considered a 32 Kbps link, having an average packet interarrival time of 0.125 s). Finally, the packet size is modelled with a non-null minimum Pareto distribution with cut-off ($\alpha$=1.1; minimum k=81.5 bytes; cut-off maximum m=66666 bytes).

We designed a test system consisting of a traffic generator (based on these models) and a queue with a constant service rate of 2 KBps. Because of the model structure, each simulated session requires a separate process to manage it and generate its pages. This process is started when a session begins, but its termination is not scheduled until all its pages are concluded.

The test conditions were defined as follows: 2000 s of average session interarrival time and a queue size of 4 MBytes. With these test conditions we carried out simulations with the proposed model adapted to the corporate environment, and with the ETSI model, having a server utilization rate of 68% and 3% respectively. In addition, we developed an extra simulation with the ETSI model and a session interarrival time of 82.75 s to get a server utilization rate of 68%. Thus, we could compare both models in the same load conditions.

Table 9 shows the packet loss probability obtained, and the delay and jitter in the buffer. Notice that the ETSI model underestimates these parameters, even with the same server utilization rate. This is not a surprising result, since the traffic load generated by the ETSI model is lower than the measured one. Furthermore, the ETSI model has less burstiness than the measured one.

Table 9
Model comparison

| Parameter | Corporate Environment | ETSI Model | Adjusted Utilization ETSI Model |
|---|---|---|---|
| Measured packet loss probability | 0.14 | 0 | $1.9 \cdot 10^{-5}$ |
| Mean delay | 699.21 s | 13.74 s | 124.34 s |
| Jitter | 785.12 s | 31.93 s | 181.86 s |

Figure 12 shows buffer occupancy histograms for the three tests described above. These results confirm the previous conclusion. The ETSI model underestimates the buffer occupancy, giving totally different results.

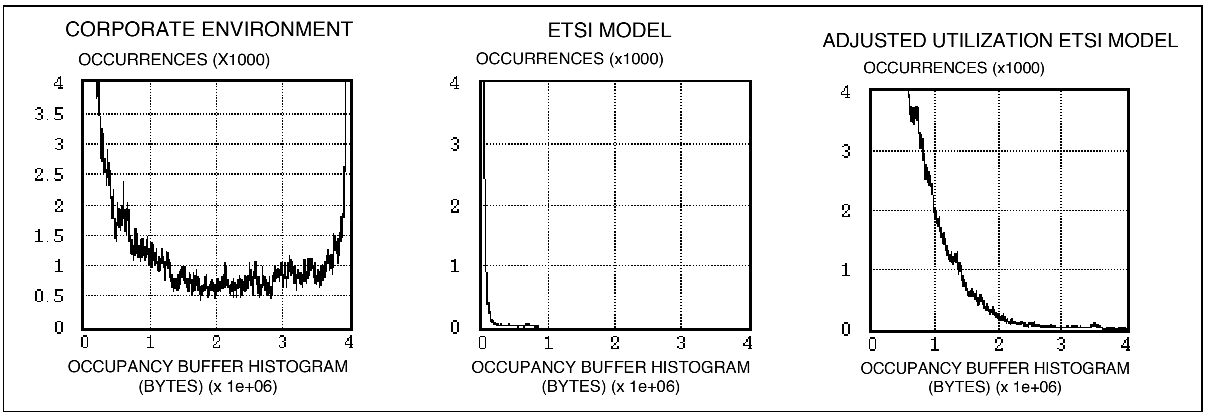


Fig. 12. Buffer occupancy comparison.

## 5. CONCLUSIONS

A new WWW traffic model for wireless systems simulations has been presented, with detailed characterization in corporate and educational environments. The proposed model is defined by a set of parameters that have clear physical meanings.

Both corporate and educational environments present very similar traffic characteristics. However, educational traffic is more dispersed and its users more heterogeneous than corporate users.

Interestingly, the measured page size was found to be Pareto-distributed, with infinite variance. This will cause self-similar behaviour, which has been reported by other authors [3, 10, 11, 12].

We have also compared our model with that proposed by ETSI [5], and demonstrated that the ETSI model underestimates packet losses and delay in a queue. We have confirmed that these results are due to the low load offered by the ETSI model, while our proposed model generates a traffic load similar to the measured one.

In addition, our proposed model generates traffic with much more burstiness than the ETSI one. Thus, even with the same server utilization rate, the ETSI model underestimates packet loss probability and delay.

## REFERENCES


1. Arlitt, M., and Williamson, C., "A Synthetic Workload Model for Internet Mosaic Traffic", SCSC '95, Ottawa, July, 1995.
2. Braden, R., "Requirements for Internet Hosts – Communication Layers", RFC 1122, IETF, October, 1989.
3. Crovella, M.E., Taqqu, M.S., and Bestavros, A., "Heavy-Tailed Probability Distributions at the World Wide Web: Traffic Self-Similarity", Technical Report, Univ. of Boston, 1996.
4. Danzig, P., and Jamin, S., "Tcplib: A Library of TCP Internetwork Traffic Characteristics", Report CS-SYS-91-01, Computer Science Department, South California University, 1991.
5. "Universal Mobile Telecommunication System (UMTS); selection procedures for the choice of radio transmission technologies of the UMTS", Technical Report TR 101 112 v3.1.0, ETSI, November, 1997
6. Fielding R. T, et al., "Hypertext Transfer Protocol – HTTP/1.1", Internet draft, IETF, 1996.
7. Greiner, M., "How to model Telecommunications (and Other) Systems Where 'Self-Similar' Traffic is Observed", Tech. Report from Technische Universität München, Germany, Sept., 1997
8. Mah, B.A., "An Empirical Model of HTTP Network Traffic", *Proceedings of the IEEE INFOCOM '97,* Kobe, April, 1997, Vol. 2, pp. 592-600.
9. Paxson, V., "Empirically-Derived Analytic Models of Wide-Area TCP Connections", *IEEE/ACM Transactions on Networking*, Vol. 2, 1994, pp. 316-336.
10. Park, K., Kim, G., and Crovella, M., "On the relationship between file sizes, transport protocols and self-similar network traffic".
11. Pitkow, J. E., "Summary of WWW Traffic Characterizations", Computer Networks and ISDN Systems, Vol. 30, 1998, pp. 551-558.
12. Vicari, N., "Measurement and Modeling of WWW-Sessions", Report Nº 184, Research Report Series, Institute of Computer Science, University of Wurzburg, September, 1997.
13. Willinger, W., and Paxson, V., "Discussion of 'Heavy Tail Modeling and Teletraffic Data' by S.R. Resnick".