Evaluation of dial-up behaviour of Internet users

Dipl.-Ing. Johannes Färber[1], Dipl.-Ing. Stefan Bodamer[1], Dipl.-Ing. Joachim Charzinski[2]
[1] University of Stuttgart, Institute of Communication Networks and Computer Engineering (IND), Germany
[2] Siemens AG, Public Communication Networks, Germany

Abstract

In times of Internet access becoming a popular consumer application even for „normal“ residential users, some telephone exchanges are congested by customers using modem or ISDN dial-up connections to their Internet Service Providers. In order to estimate the number of additional lines and switching capacity required in an exchange, the Internet access traffic must be characterised in terms of holding time and call interarrival time distributions. In this paper, we analyse six months worth of log files tracing the usage of the central modem pool and ISDN access lines at University of Stuttgart. Mathematical distributions are fitted to the measured data and the fit quality is evaluated with respect to the blocking probability observed in a multiple server loss system loaded by the described traffic.

1 Introduction

Internet access, especially for WWW based services is becoming a popular application with residential as well as business users. While business users are often connected to the Internet via leased lines, most residentials use the local telephone network to establish a dial-up connection to their Internet Service Provider (ISP). Apart from seizing a modem in the ISP's modem pool, each dial-up connection occupies one telephone line in a local exchange. In order to evaluate the additional traffic load imposed on a telephone exchange by Internet access traffic, the traffic characteristics must be well known.
In this paper, we present an analysis of the Internet access traffic logged at a modem pool (56 modems shared by 3620 students) and an ISDN line access pool (30 B channels shared by 372 users) at the University of Stuttgart computing centre (RUS). Data were collected during six months from May through October 1997 with a resolution of one minute (modem pool) or one second (ISDN access). Section 2 describes the mean and distribution of holding times as well as access call interarrival times and the resulting daily traffic load. In Section 3, different mathematical distribution functions are fitted to the holding time and call interarrival time distributions. In Section 4, the fit quality is evaluated by comparing the blocking probability observed by loading different numbers of servers (i.e. phone lines) with random traffic generated according to the fitted distributions.

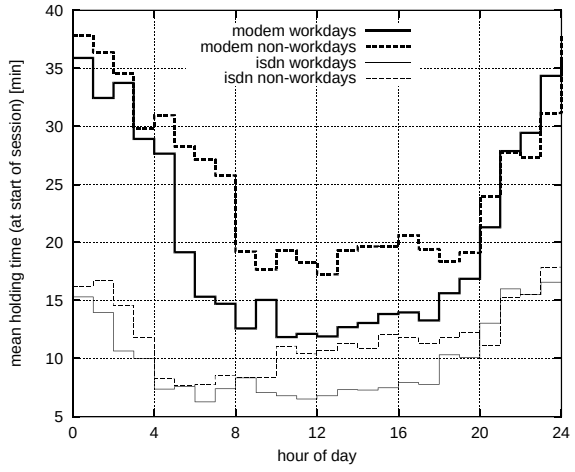

Fig. 1: Session holding time during course of the day

2 Session Behaviour

The automatic monitoring of user login times at the dial-up access service of the University of Stuttgart allows the evaluation of characteristic measures on session level. The data allow to distinguish between modem users and ISDN users. Note that there is no way to specify the type of traffic the user has started. While in most cases it can be expected to be a World Wide Web session, it may also be a telnet session, an ftp retrieval, , an email transfer or a mix of those traffic types.
In the following sections we describe the holding time of the sessions, the interarrival time between session starts and the mean daily traffic profile for traffic load. These measures allow to characterise the frequency and duration of a typical user session.

2.1 Holding Time

By the holding time of a session we mean the duration of the seizure of an access line. The mean holding time for modems was 21 minutes while ISDN users

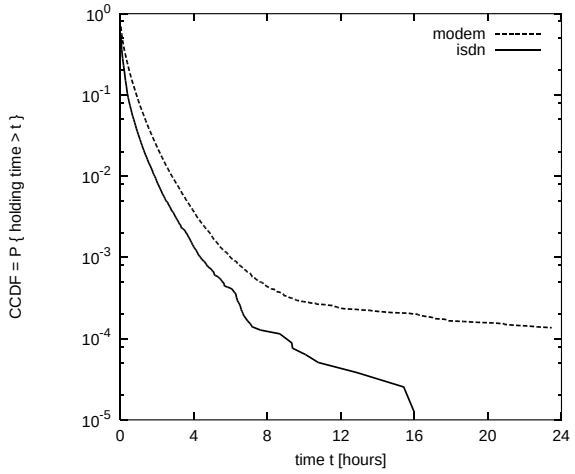

Fig. 2: Session holding time ccdf

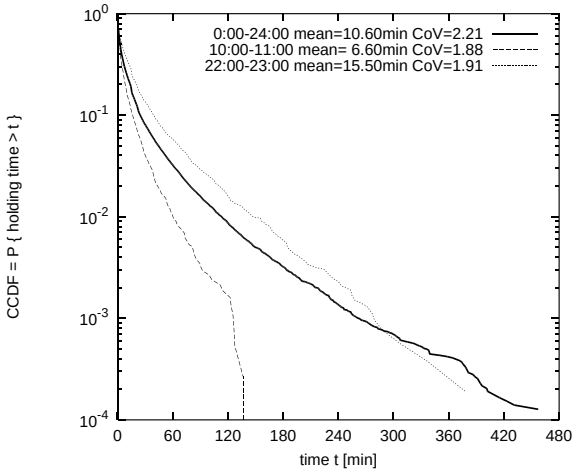

Fig. 3: Session holding time ccdf in different periods of the day (ISDN only)

were online for an average of 11 minutes. The holding time shows a high variability, i.e. it varies strongly and reaches a maximum of as much as 4 days.

Also, the average holding time varies during the course of the day. **Fig. 1** shows the average holding times of sessions of both user groups associated with the time of the session start for each hour of the day. Although this representation has to be regarded with caution (the average during the night is calculated from a relatively small number of sessions), it allows the conclusion that long sessions start mainly during the night and early morning hours. The mean session length at night is significantly larger than during day time. Sessions on non-workdays are longer in the average. Note that sessions of ISDN users seem to be much shorter than those of modem users. An explanation for this is given in Section 2.2.

In [8] Morgan reports a significant peak in holding time at 4 a.m. If the holding time is associated with the session endings, a similar peak is observed in our data at 4 a.m. This means that among sessions ending in the early morning have lasted for a long time.

The high variability of the holding time is visible in the complementary cumulative distribution function (ccdf) which is depicted in **Fig. 2**. The function shows the probability of the holding time being greater than the value on the horizontal axis. While there is a high probability for short sessions, the logarithmic presentation reveals that there is a small but not negligible probability for very long sessions of 20 hours and more. This so called „heavy tail“ is an indication for high variability of large values [4]. The coefficient of variation (CoV) of the holding time data is around 2.5 for modem and 2.2 for ISDN based access.

The shift of the average holding time during the course of the day shown in **Fig. 1** suggests that the instationarity of mean holding times contributes to the high variation of the overall holding time distribution. Therefore, the variability of the holding time might not be as high if a shorter period is regarded instead of the whole day. In **Fig. 3** we present the ccdfs of the holding times of only several specific hours of a day (the most interesting busy hours, see section 2.3). The coefficients of variation for those periods are in fact smaller, but not much, i.e. the „heavy tail“ is still visible.

2.2 Interarrival Time

In our context, the session interarrival time is the time between two consecutive session beginnings of the aggregate traffic as seen by the access provider. This means that blocked calls are not detected and that session arrivals are not to be confused with call attempts.

The mean session interarrival time was 44 seconds for modem sessions and 110 seconds for ISDN sessions. To allow a comparison, those absolute numbers have

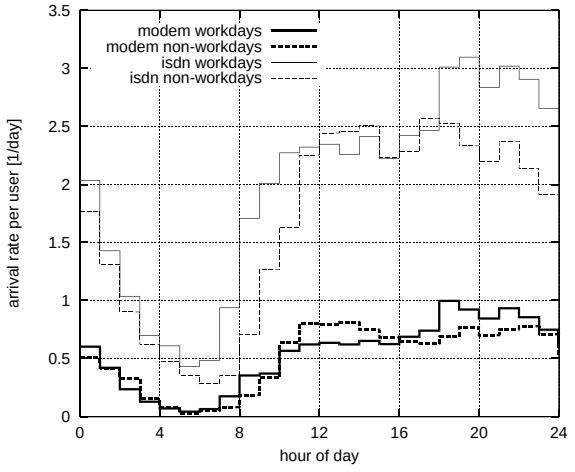

Fig. 4: Session arrival rate during course of the day

to be put into context to the number of users producing the summary traffic. Therefore the values in the following two figures are related to the corresponding numbers of users.

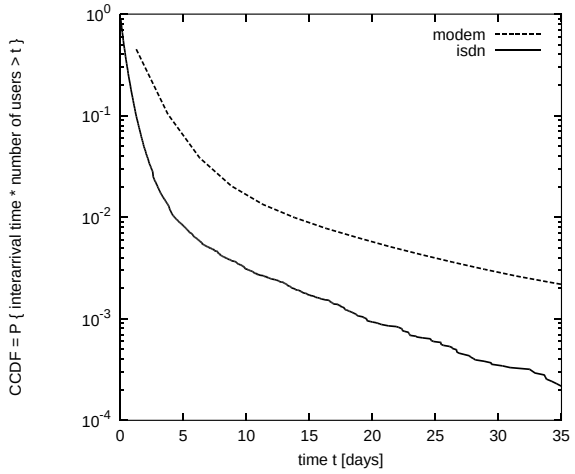

Fig. 5: Scaled ccdf of session interarrival time

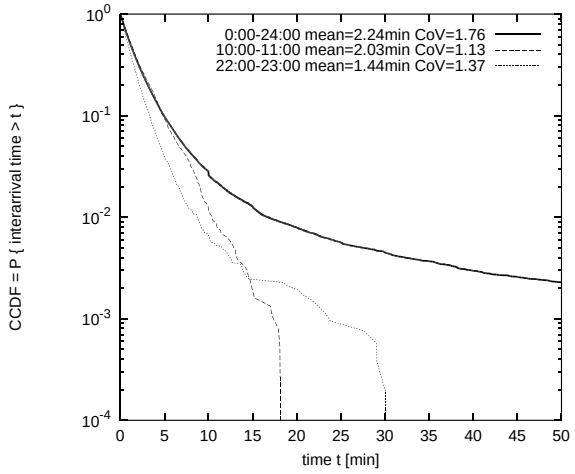

Fig. 6: Session interarrival time ccdfs in different periods of the day (ISDN only)

Fig. 4 depicts the session arrival rate per user (the reciprocal of the interarrival time) during the course of the day. It shows that only a few sessions start in the early morning hours and that most sessions take place during the day and evening hours. The significant step at 6 p.m. on workdays is due to the cheaper telephone tariff starting at that hour in Germany. There is a remarkably high number of sessions starting in the late night.

It can be seen that ISDN users cause a much higher call rate than modem users. When considering the arrival rates together with mean holding times (**Fig. 1**), it is clear that the overall session lengths per user and per day are roughly identical for modem and ISDN users. The fast and easy setup of connections via ISDN seems to lead to a different user behaviour, i.e. more but shorter sessions are generated.

Again we find a high variability for the interarrival time ranging from zero to several hours. The scaled ccdf in **Fig. 5** also shows the typical heavy tail. From the above, it is obvious that modem sessions have a higher probability for longer interarrival times than ISDN sessions.

Note that this presentation is only valid for the description of summary traffic of multiple users. As reported in [6], the ccdfs of the interarrival time of sessions of an individual user show significant periodic steps every 24 hours.

Similar to the preceding section, we present the ccdfs of interarrival times during several specific hours only in **Fig. 6**. The ccdf for the period between 10 a.m. and 11 a.m. is almost an exponential function (a straight line in the logarithmic presentation). The corresponding coefficient of variation is much smaller than e.g. between 10 p.m. and 11 p.m. where we still observe a high variability.

2.3 Traffic Load

To cope with the originated traffic, a telephone network must offer sufficient resources in terms of bandwidth (i.e. telephone lines) and connection setup capacity (i.e. processing power). Therefore we distinguish between traffic load related to user traffic (the seizure of the modem lines and ISDN channels) and signalling load, corresponding to the call arrival rate, to describe the actual load of the access network.

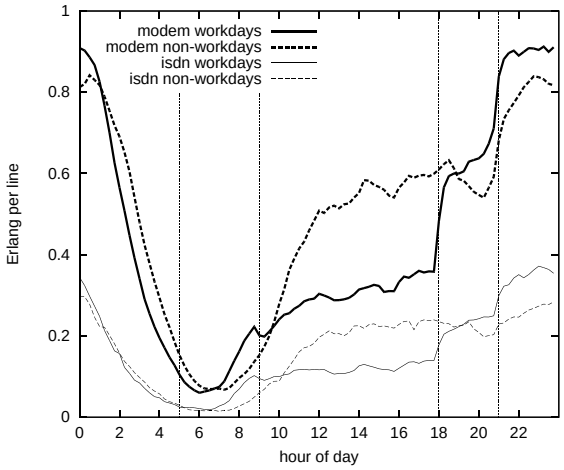

Fig. 7: Mean daily traffic profile

While the average signalling load can be seen in **Fig. 4**, the user traffic load is shown in **Fig. 7**. By depicting the traffic load per line, the utilisation of both services can be compared. While the 30 channels of the ISDN service are sufficient for 370 users, 56 modems are obviously not enough for the 3600 students. The flat

shape of the modem traffic load around midnight indicates that all modem lines were busy on most days and it can be assumed that many calls have been blocked.
The general shape of the profiles is almost complementary to the typical telephone traffic profile, which has its busy hour from 10 a.m. to 11 a.m. and only a small traffic load in the evening and during the night.
The most striking characteristics of the traffic profile are the two steps at 6 p.m. and 9 p.m. on workdays. As mentioned above, these times mark the beginning of cheaper telephone tariffs during the observed period. The boundaries of the tariff periods are drawn as vertical lines. A small peak is also visible right before 9 a.m., when the expensive day tariff starts. On weekends and holidays this day tariff between 9 a.m. and 6 p.m. is does not exist and only a sharp increase in traffic load at 9 p.m. can be observed. The user behaviour follows the telephone tariffing scheme amazingly accurately.
The time consistent busy hour for user traffic load is found at 10 p.m. but for the arrival rate it is found earlier at 6 p.m. In [3] Bolotin points out that these two busy hours are significantly shifted against each other for Internet traffic compared to telephone traffic. The phenomenon is caused by much longer holding times of around 20 minutes compared to 3 minutes for classical telephone calls.

3 Modelling Session Behaviour

For performance evaluation of communication systems by performance analysis or simulation, source traffic has to be modelled to assess the system behaviour. Complex traffic is best described with the help of empirical data. Either a logged traffic trace is replayed into the system model, or a mathematical description can be found to generate stochastic traffic of similar characteristics.
To describe traffic load on session level, it is important to know about the holding time and the session interarrival time. The complementary cumulative distribution functions of these measures capture their most important characteristics. If mathematical functions can be found that describe the ccdfs, they can be used for performance evaluation.
While the classical telephone traffic is appropriately described with negative-exponentially distributed holding times and call interarrival times, this is no longer true for Internet access session traffic any more. The high variability of the measures described above is not captured by this distribution. For a more accurate description of the new traffic other distributions have been proposed like Pareto, hyperexponential, Weibull or lognormal, e.g. in [1], [2], [5], [7].

In a previous attempt to describe the ccdfs, we used a least square fitting algorithm to fit some of the distributions mentioned above to the empirical ccdfs [6]. Although the results looked quite good, they were by far inferior to the exponential distribution if used for the simulation of a loss system.

Exp.	$f(t) = \frac{1}{m} \exp\left(-\frac{t}{m}\right)$
Hyperexp. (order k)	$f(t) = \sum_{i=1}^{k} \lambda_i \, p_i \exp\left(-\lambda_i \, t\right)$
Log-normal	$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma t} \exp\left(-\frac{\left(\ln t - \mu\right)^2}{2\sigma^2}\right)$
Weibull	$f(t) = \alpha \, \beta^{-\alpha} \, t^{\alpha-1} \exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)$

Tab. 1: Probability density functions for exponential, hyperexponential, lognormal and Weibull

In this paper we fit the above mentioned distributions by calculating their parameters from the mean and the coefficient of variation given by the empirical ccdfs instead. **Tab. 1** shows the mathematical formulae for probability density functions of the most promising distributions, which can easily be determined by mean and CoV. Also, we evaluate the ccdfs for the most critical periods of the day, i.e. the busy hours. **Tab. 2** shows the mean and CoV of the ccdfs for these cases, which are also depicted in **Fig. 3** and in **Fig. 6**.
For simplicity and because of a much better granularity of the raw data, the following sections deal only with the modelling of ISDN sessions.

	mean [min]	CoV
interarrival time:		
0 a.m. - 12 p.m.	2.24	1.76
10 a.m. - 11 a.m.	2.03	1.13
10 p.m. - 11 p.m.	1.44	1.37
holding time:		
0 a.m. - 12 p.m.	10.60	2.21
10 a.m. - 11 a.m.	6.60	1.88
10 p.m. - 11 p.m.	15.50	1.91

Tab. 2: Mean and coefficient of variation for interarrival time and holding time (ISDN)

Fig. 8 and **Fig. 9** shows the resulting ccdfs for the holding time and the interarrival time, respectively, between 10 p.m. and 11 p.m. in comparison to the empirical distribution. All functions have the same mean and variability (except the exponential distribution with a CoV of 1). The corresponding figures

for the whole day and for 10 a.m. - 11 a.m. are not depicted but look very similar.

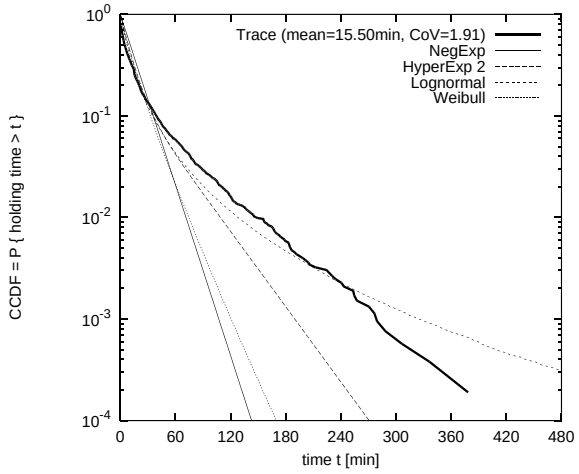

Fig. 8: Fitted functions to ccdf of the session holding time between 10 p.m. and 11 p.m.

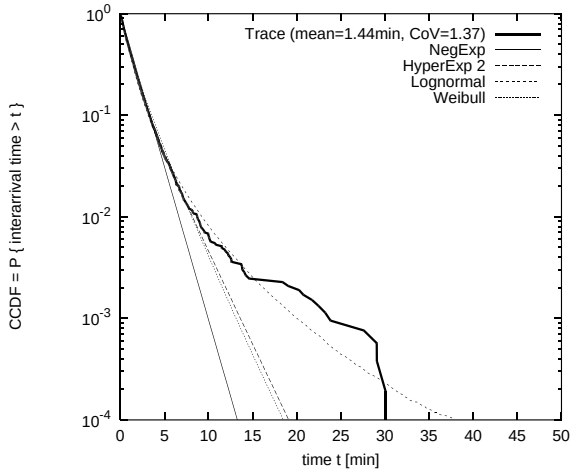

Fig. 9: Fitted functions to ccdf of the session interarrival time between 10 p.m. and 11 p.m.

All fitted ccdfs in both diagrams are quite close to the emprical ccdf in the range of small values for holding and interarrival time, respectively. The distribution tails, however, can only be approximated reasonably well by the lognormal distribution, while the other fitted ccdfs do not capture the „heavy tail“ effect.

4 Performance Evaluation

The fitting results are validated for the case of a simple G/G/n loss system as it is often used to model communication systems. In this model, *n* represents the number of servers which could be, e.g., modems, ISDN cards, or access lines.

In the case of an M/M/n system where interarrival and holding times are both assumed to be exponentially distributed, the loss probability *B* can be obtained analytically using the well-known Erlang loss formula:

$$B = \frac{\frac{A^n}{n!}}{\sum_{i=0}^{n} \frac{A^i}{i!}}$$

where *A* denotes the offered load defined as the ratio of mean holding time and mean interarrival time.
If interarrival and holding times are described by distributions different from the exponential distribution, simulations are used to obtain the loss probabilities. This is the case for the empirical distributions obtained from the trace evaluation (see Section 2) as well as for the approximating distributions of hyperexponential, lognormal and Weibull type found in Section 3.

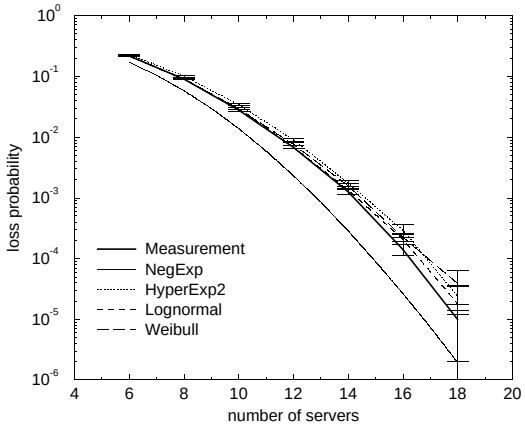

Fig. 10: Loss probability for 0 a.m. -12 p.m. traffic

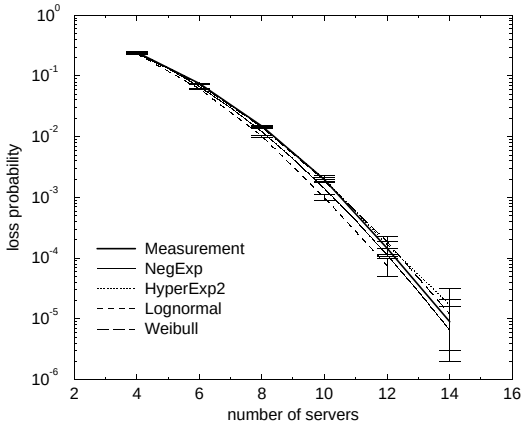

Fig. 11: Loss probability for 10 a.m. -11 a.m. traffic

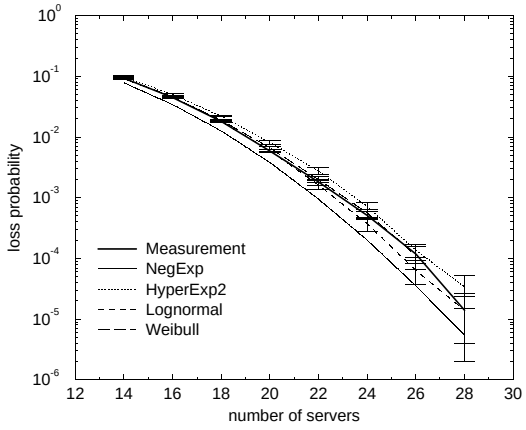

Fig. 12: Loss probability for 10 p.m. -11 p.m. traffic

The results of the performance analysis and simulation for the empirical and the fitted distributions for the periods of the day presented in Section 3 are depicted in **Fig. 10**, **Fig. 11**, and **Fig. 12**. The corresponding distribution type indicated for each curve in the diagrams is related to both interarrival and holding times.

For all investigated periods of the day it becomes obvious that the M/M/n approximation using negative-exponential distributions underestimates the actual loss probability. An underestimation of the loss probability means that a dimensioning based on the M/M/n loss system would give a value for the necessary number of servers which is too small. The effect is most significant in the case referring to the whole day traffic, but also for the period from 10 p.m. to 11 p.m.

The Weibull approximation evolves to give the most exact results in all three cases. The hyperexponential distribution yields a conservative approximation, i. e. the loss probability curve is quite close to that associated to the empirical distribution but always a little bit above. The results obtained for the lognormal distribution on the other hand are also quite exact, but too optimistic.

5 Conclusion

We have presented the dial-up behaviour of modem and ISDN users at the University of Stuttgart. Major results of the data evaluation are:

- the same long holding times as reported by other researchers have been observed,
- ISDN users generate more but shorter sessions (likely due to fast setup),
- the dial-up usage follows the telephone tariffing scheme with high accuracy (the Internet access itself was provided for free) and
- the session interarrival time and the session holding time show very high variability (heavy tailed distributions)

The presented results are based on empirical data of a special user group, i.e. the students and members of staff of the University of Stuttgart. Although this group might not represent the general Internet user, it does represent a large group of users going online after work hours.

The empirical ccdfs of interarrival time and holding time during the busy hours were approximated by fitting several distributions. The fit was based on the mean and the coefficient of variation instead of applying a least square fit as it was done in an earlier work.

The quality of the fitted distributions has been proven by the performance evaluation of a simple loss system. The models with Weibull and the hyperexponentially distributed interarrival and holding times gave the best approximations of the loss probability compared to the empirical distributions.

6 References

[1] Bolotin, V.A.: Telephone Circuit Holding Time Distributions, Proceedings of the ITC 14, pp. 125-134, Antibes, France, 1994.

[2] Bolotin, V.A.: Modelling Call Holding Time Distributions for CCS Network Design and Performance Analysis, IEEE Journal on Selected Areas in Communications, pp. 433-438, April 1994.

[3] Bolotin, V.A.: New Subscriber Traffic Variability Patterns for Network Traffic Engineering, Proceedings of the 15th International Teletraffic Congress ITC 15, Vol. 2, pp. 867-878, Washington, 1997.

[4] Crovella, M., Bestavros, A.: Performance Characteristics of World Wide Web Information Systems, Tutorial at the SIGMETRICS‘97, 1997.

[5] Crovella, M., Bestavros, A.: Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes“, Proceedings of the ACM SIGMETRICS`96, pp. 160-169, 1996.

[6] Färber, J., Bodamer, S., Charzinski, J.: Measurement and Modelling of Internet Traffic at Access Networks, EUNICE’98, Munich , 1998.

[7] Feldmann, A., Whitt, W.: Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models, Performance Evaluation 31, pp. 245-179, 1998.

[8] Morgan, S.: The Internet and the Local Telephone Network: Conflicts and Opportunities, IEEE Communications Magazine, pp. 42-48, January 1998.