Statistical evaluation and modelling of Internet dial-up traffic

Johannes Färber [a], Stefan Bodamer [a], Joachim Charzinski [b] *

[a] University of Stuttgart, Institute of Communication Networks and Computer Engineering (IND)

[b] Siemens AG, Information and Communication Networks Group

ABSTRACT

In times of Internet access being 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 (ISPs). In order to estimate the number of additional lines and switching capacity required in an exchange or a trunk group, Internet access traffic must be characterized in terms of holding time and call interarrival time distributions. In this paper, we analyse log files tracing the usage of the central ISDN access line pool at University of Stuttgart for a period of six months. Mathematical distributions are fitted to the measured data and the fit quality is evaluated with respect to the blocking probability caused by the synthetic traffic in a multiple server loss system. We show how the synthetic traffic model scales with the number of subscribers and how the model could be applied to compute economy of scale results for Internet access trunks or access servers.

Keywords: Internet traffic modelling, ISDN access, trace evaluation

1. INTRODUCTION

Internet access, especially for WWW based services, has become popular with residential as well as business users. Most of the residentials and some small businesses connect to the Internet using the local telephone network to establish dial-up connections to their Internet Service Providers. Apart from seizing a modem in the ISP's modem pool, each dial-up connection occupies one telephone line at least in the respective local exchanges, sometimes also on trunk groups leading to the local exchange where an ISP's modem pool is located.

In this paper, we present an analysis of the Internet access traffic logged at an ISDN line access pool at the University of Stuttgart computing centre (RUS). Section 2 describes characteristics of holding times as well as access call interarrival times and the resulting daily traffic load. In Section 3, we suggest a simple traffic model based on different mathematical distribution functions which are fitted to the holding time and call interarrival time distributions. The accuracy of the model is evaluated through a comparison of the blocking probability in a loss simulation for trace data and the data generated according to the fitted distributions. Furthermore we discuss how the aggregate arrival process scales with the number of users. We motivate the scaling behaviour with a rate modulated process, and finally we indicate how the results can be applied to dimensioning problems.

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 on which this study is based, were collected during six months from May through October 1997. Low monthly fluctuations of total online time, number of sessions or the number of active users allow to consider the dataset as stationary. The ISDN line access pool comprises 30 B channels which were shared by 372 users.

Note that there is no way to specify the type of session each 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 these. However, we have observed similar session duration characteristics for the WWW service in other instances.

In the following sections we describe the observed 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 durations of typical user sessions.

* Correspondence: Email: {faerber|bodamer}@ind.uni-stuttgart.de, Joachim.Charzinski@icn.siemens.de

2.1. Holding time

The holding time of a session is defined as the duration of the seizure of an access line in our context. The holding time varies strongly around a mean of 11 minutes† and reaches an observed maximum of as much as 4 days.

Also when considering the distribution of holding times observed during a specific hour of day, strong variations are observed. Figure 1 shows the average holding times of sessions (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 hours. The mean session length at night is significantly larger than during day time. Sessions on non-workdays are longer in the average. Morgan reports a significant peak in holding time at 4 a.m [9]. 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 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 Figure 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 half-logarithmic presentation reveals that there is a small but not negligible probability for very long sessions of 15 hours and more. The shape looks slightly „heavy tailed“ and is an indication for high variability of large values [4]. The coefficient of variation (CoV) of the holding time data is around 2.2.

The shift of the average holding time during the course of the day shown in Figure 1 suggests that the instationarity of the mean holding times contributes to the high variation of the overall holding time distribution. Therefore, the variability of the holding time should not be as high if a shorter period is regarded instead of the whole day. The ccdfs of the holding times of only several specific hours of a day (the most interesting busy hours, see Section 2.3) are depicted in Figure 2 as well. The coefficients of variation for those periods are in fact smaller, but not much.

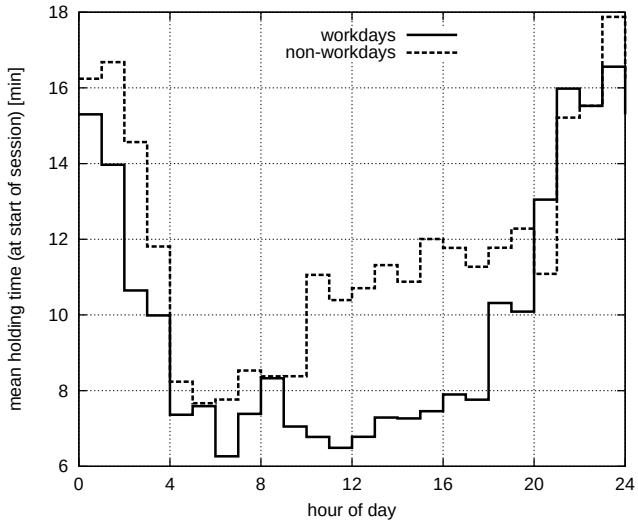

Figure 1: Session holding time during course of the day

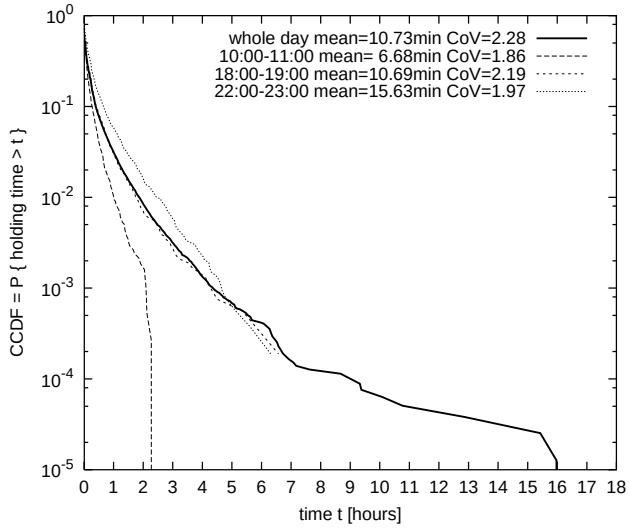

Figure 2: Session holding time ccdfs for different periods of the day

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 therefore session arrivals should not be confused with call attempts. However, since all 30 lines of the ISDN pool were never seized simultaneously we can say that no blocking due to limited access port availability has occurred in the observed period.

† Often, an average holding time of around 20 minutes is reported. This is true for the analog modem pool at the University of Stuttgart as well. ISDN connections however, seem to be shorter in general [7].

The mean session interarrival time of the aggregated traffic was 110 seconds. To allow a comparison with other data, this absolute number has to be put into context to the number of users producing the summary traffic. With a total of 372 users this leads to a session interarrival time of approximately 11 hours per user.

Figure 3 depicts the mean session arrival rate per user for each hour of the day. Note, that the rate is given in sessions per day for easier association with user behaviour. A rate of 2 from 9 a.m. to 10 a.m. means that a user would generate 2 sessions per day if this rate was maintained for the whole day. The profile 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.‡

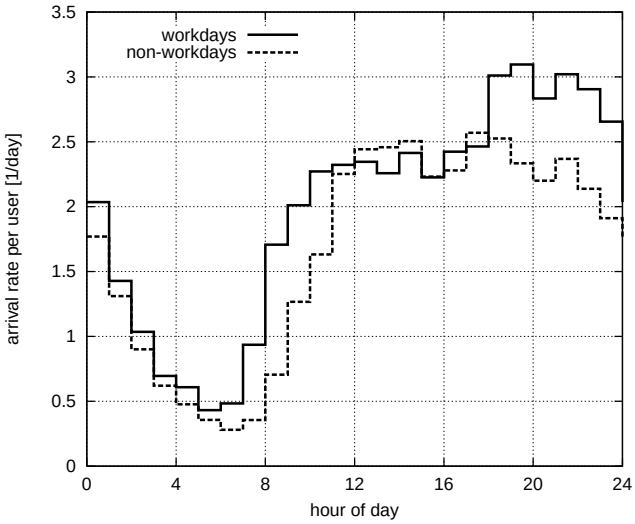

Figure 3: Session arrival rate per user during course of the day

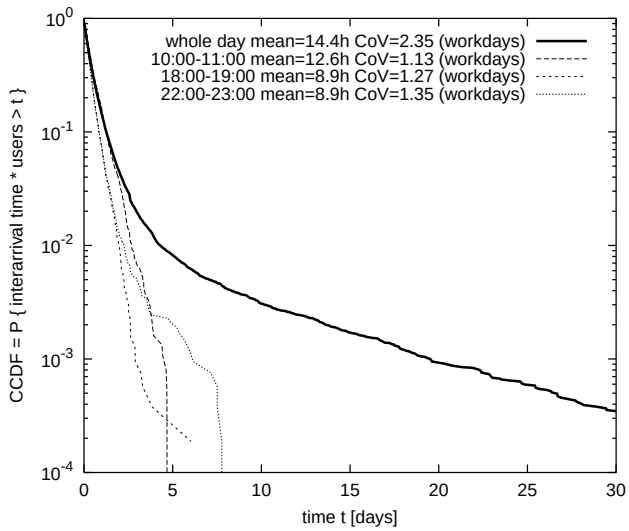

Figure 4: Session interarrival time ccdfs for different periods of the day

Again we find a high variability for the interarrival time ranging from zero to several hours. The scaled ccdf in Figure 4 shows the typical shape of a hyperexponential distribution. Similar to the preceding section, we also show the ccdfs of interarrival times during several specific hours. 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. Note that this presentation is only valid for the description of normalised summary traffic of multiple users. The interarrival time of sessions of an individual user shows significant periodic behaviour every 24 hours [6].

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 load related to connections (the seizure of the ISDN channels) and signalling load (corresponding to the call arrival rate) to describe the actual load of the access network. While the signalling load profile can be derived from Figure 3, the user traffic load profile is shown in Figure 5.

The general shape of the profiles is almost complementary to the typical business 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. Above, we also presented traffic characteristics for this busy hour to allow comparisons.

---

‡ We have observed that the call rate of ISDN users at the University of Stuttgart is more than twice as high as for modem users [7]. The fast and easy setup of connections via ISDN seems to lead to a different user behavior than with modem access, i.e. more but also shorter sessions are generated. In total, this leads to a longer user online time per day than was observed for modem access.

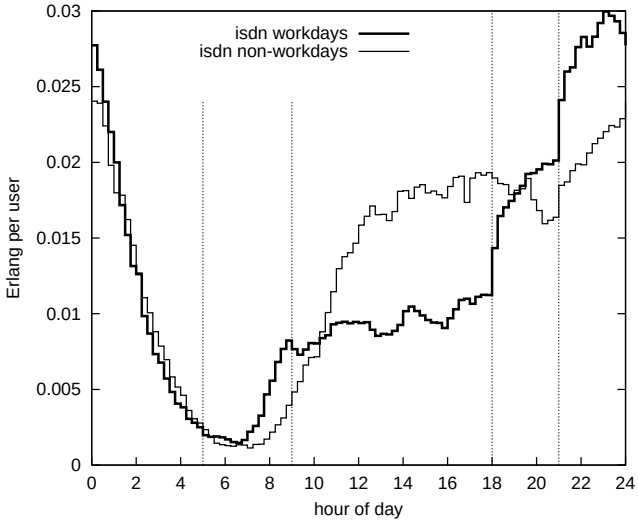

Figure 5: Mean daily traffic profile

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. does not exist and only a sharp increase in traffic load at 9 p.m. can be observed. During the day, the user behaviour follows the telephone tariffing scheme amazingly accurately.

The time consistent busy hour for user traffic load starts at 10 p.m. whereas for the arrival rate it starts earlier at 6 p.m. Bolotin points out that the shift between these two busy hours, which is observed for Internet traffic, is negligible for telephone traffic [3].

3. MODELLING SESSION BEHAVIOUR

An important issue in performance evaluation of communication systems is source traffic modelling. Complex traffic is best described with the help of empirical data. The simple approach of replaying a logged trace into the system in a simulation, however, is limited with regard to generality. Therefore the most promising way of traffic modelling is to find a mathematical description that allows the generation of stochastic traffic with similar characteristics.

We suggest a simple model for the generation of Internet access-like aggregated traffic that is based on our evaluation presented above. The purpose of the model is to produce the same state process (i.e. number of access lines occupied) in the access network. We validate it by loss simulations and compare the resulting blocking probabilities to those of corresponding loss simulations based on the trace data.

Although the trace does not capture blocking events, we can be sure that the logged information is complete and that therefore the trace is well suited for typical traffic description: the maximum observed number of seized access lines was 27 out of 30 available so that blocking due to limited access port availability can be excluded. However, we do not know if trunk line blocking has occurred somewhere else in the telephone network, but we assume that this blocking probability is low enough not to influence our data too much.

3.1. Modelling aggregated traffic with Renewal Processes

Description of traffic at session level requires knowledge about the holding time and the session interarrival time. The complementary cumulative distribution functions (ccdfs) of these measures capture their most important characteristics except for correlation effects. In a first step, we use renewal processes governed by the corresponding ccdfs to model the interarrival and holding times.

Whereas the classical telephone traffic is appropriately described by negative-exponentially distributed holding times and call interarrival times, this is no longer true for Internet access session traffic. The high variability of the measures described above is not captured by a negative-exponential distribution. A more accurate description of the traffic can be obtained from other distributions like Pareto, hyperexponential, Weibull or lognormal [1,2,5,8].

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 resulting plots of the ccdfs looked quite similar to the original, they were by far inferior to the exponential distribution if used to generate input traffic for the simulation of a loss system. 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. For the hyperexponential distribution we only regard the order of two and fit two of its three parameters by mean and variance. The third parameter is fixed by the symmetry condition $p_1 \cdot m_1 = p_2 \cdot m_2$ where p_1 and p_2 represent the branch probabilities and m_1 and m_2 the means, respectively, in the corresponding phase model. Furthermore, we concentrate on the ccdfs for the most critical periods of the day, i.e. the busy hours.

Figure 6 and Figure 7 show the fitted 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 tails are not well fitted by the chosen method but one has to consider that the tail effects are emphasised by the half-logarithmic presentation. The corresponding figures for the whole day and for 10 a.m. - 11 a.m. are not depicted but look very similar.

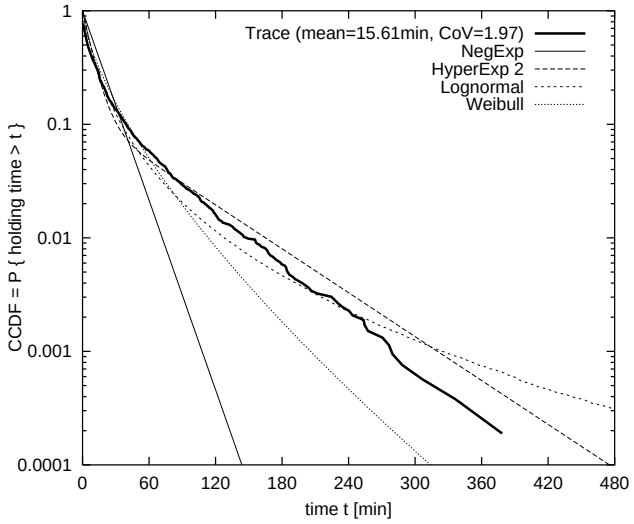

Figure 6: Functions fitted to ccdf of the session holding time between 10 p.m. and 11 p.m.

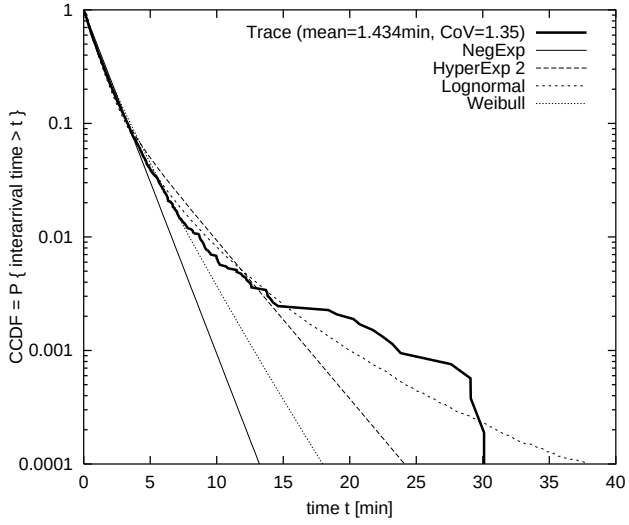

Figure 7: Functions fitted to ccdf of the session interarrival time between 10 p.m. and 11 p.m.

All fitted ccdfs in both diagrams are quite close to the empirical ccdf in the range of small values for holding as well as for interarrival time. The distribution tails, however, can only be approximated reasonably well by the lognormal and hyperexponential distribution.

The renewal process model based on these fitting results is 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 or telephone access lines.

In the case of an M/G/n system where the interarrival times are assumed to be exponentially distributed, the loss probability can be obtained analytically using the well-known Erlang loss formula. 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 approximating distributions of hyperexponential (H2), lognormal (LogN) and Weibull (W) type. Simulations were also used to calculate the reference loss probabilities by replaying the original trace into the loss system. Repeated call attempts for blocked calls were neither considered in our renewal process models, nor in the trace driven simulation because here the main emphasis was on finding processes which model the state process of the considered loss system correctly.

We have simulated all combinations for the holding time and interarrival time distributions and found that using the hyperexponential distribution for both measures provides the best approximation. The results for both the trace and the renewal process model with H2/H2/n are depicted in Figure 8. The M/G/n approximation using the negative-exponential distribution for the arrivals underestimates the loss probability resulting from the original traffic significantly. All other combinations lead to better

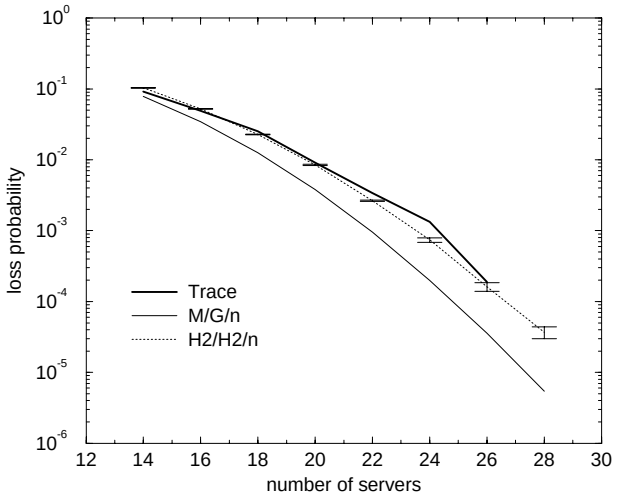

Figure 8: Loss probability for 10 p.m. - 11 p.m. traffic

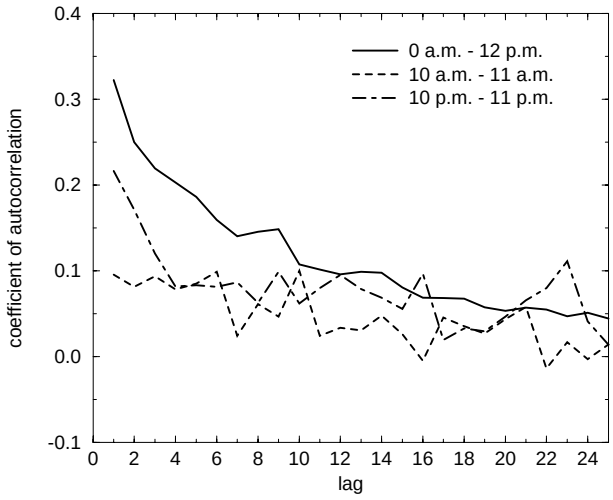

Figure 9: Autocorrelation of interarrival times

approximations than M/G/n (not depicted), but not better than the hyperexponential case. Thus for the busy hours evaluated here, renewal processes with hyperexponential distributions for holding time and interarrival time give the best description of the observed aggregated traffic.

3.2. Scalability

The model from Section 3.1 only describes aggregated traffic caused by the 372 users observed. In order to be of practical use, the model of the interarrival time should be adaptable to an arbitrary number of subscribers.

When asking whether the assumption of a renewal process for the interarrival time of the aggregate traffic was justified, we found a significant autocorrelation in the trace data especially for the interarrival times of the whole day traffic (see Figure 9). Also the curves for traffic between 10 a.m. and 11 a.m. and between 10 p.m. and 11 p.m. show positive coefficients of autocorrelation but they are smaller than in the case of the whole day traffic.**

Although the autocorrelation is not captured by the renewal process, we still have received good estimations when we used it for the approximation of the *aggregated* traffic. In order to use this approximation for a scalable model we need a different method to build the aggregation renewal process other than a superposition of renewal processes, which would yield an aggregated process with $\mathrm{CoV} \approx 1$ already for few aggregate processes.

To find such a method, we evaluate the multiplexing behaviour of real traffic by splitting our user group into non-overlapping subsets and evaluating the mean and the coefficient of variation (CoV) for the traces of each subset only. Figure 10 and Figure 11 show the resulting values for the mean and the CoV for each subset. The mean of those points resulting from subsets of the same size are connected with a line. The mean arrival rate (reciprocal of the interarrival time) is proportional to the number of evaluated users as one would expect. The CoV however, is almost independent of the number of users.

According to these results, the renewal process can be easily parameterised with the mean arrival rate which depends linearly on the number of participants and the coefficient of variation which can be assumed constant after Figure 11. This approach is based on empirical observations only and the question arises if it also can be motivated in theory.

As we have shown above, the aggregated traffic shows significant correlations which may stem from two effects:

- Single arrivals produced by different users are mutually dependent.
- Users show similar behaviour in the same period of time.

** For the estimation of the autocorrelation function of interarrival times of single hours of the day, the number of samples is much smaller as only correlations between values referring to the same hour of the same day are taken into account. Therefore, the statistical evidence of the curves in Figure 9 related to selected hours is relatively poor.

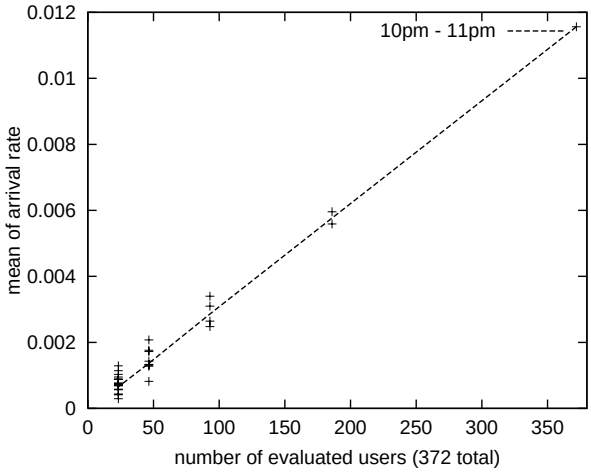

Figure 10: Mean of arrival rate for different subsets of 372 users (10 p.m.- 11 p.m.)

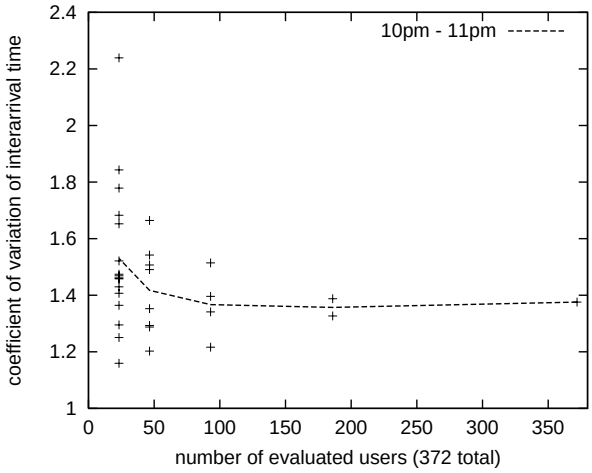

Figure 11: CoV of arrival rate for different subsets of 372 users (10 p.m.- 11 p.m.)

The first effect might be visible in a heavily underdimensioned system, which is not the case for the ISDN pool considered here, or in a system where single users dominate. In the regarded system, however, no motivation can be found for this kind of effect.

The second type of correlation can be observed when looking at variations in the daily traffic profile (peak hours, night hours, dependence on tariff boundaries). This is the reason for only taking the traffic observed during the busy hour (10 p.m. - 11 p.m.) into account. However, a similar effect is also visible when regarding traffic intensity on different days. Mean user activity may be influenced in a correlated way by common external events (weather, TV program).

The latter kind of correlation effects can be described by a CS-DMPP (continuous state deterministically modulated Poisson process). The process is a Poisson process with a certain arrival rate λ which is constant while the process remains in the same state. After a deterministic time period d the process changes into a different state with a different arrival rate for the underlying Poisson process. Continuous state means that there is a continuous state space, i.e. the arrival rate is drawn from a continuous distribution.

For our adaptation of the general model to Internet access traffic the state duration d is set to one hour (reflecting the busy hour). The distribution of λ is obtained by measuring the number of arrivals within the busy hour of each day. As shown in Figure 12, this empirical rate distribution can be well approximated by a normal distribution with mean of 0.7 and a CoV of 0.29 (which will be referred to as rate CoV below).

A CS-DMPP with parameters chosen as described above is able to produce characteristics very similar to the real traffic. Figure 13 shows that the interarrival time ccdf obtained for the CS-DMPP fits the empirical ccdf quite well. Also, the empirical interarrival time correlation function depicted in Figure 14 is reasonably well matched by the autocorrelation function of the CS-DMPP.

Again, this CS-DMPP describes only the aggregated process for 372 users. To estimate the behaviour of an arbitrary number $s \cdot 372$ of users (s is denoted as the "scaling factor"), the mean of the arrival rate λ has to be changed to $s \cdot \lambda$. The question is now how the CoV of λ depends on the scaling factor so that the resulting interarrival time CoV remains constant as been observed. The answer to that question can be obtained from Figure 15. The resulting curve is located somewhere in between of two extreme curves. The first one (denoted as "reduced CoV") represents the case without any correlation between arrivals. According to the central limit theorem the variance of λ increases with s, i.e. rate CoV ~ $1/\sqrt{s}$. If on the other hand the CoV is kept constant the second limiting curve is obtained. The rate CoV producing an interarrival time CoV independent of s is rather close the second curve, at least for $s > 1$.

One has to remark here that the purpose of the CS-DMPP was to give an explanation for the observed scaling behaviour. As a model for dimensioning we will further use renewal processes which have been shown to yield quite accurate results in the loss simulation.

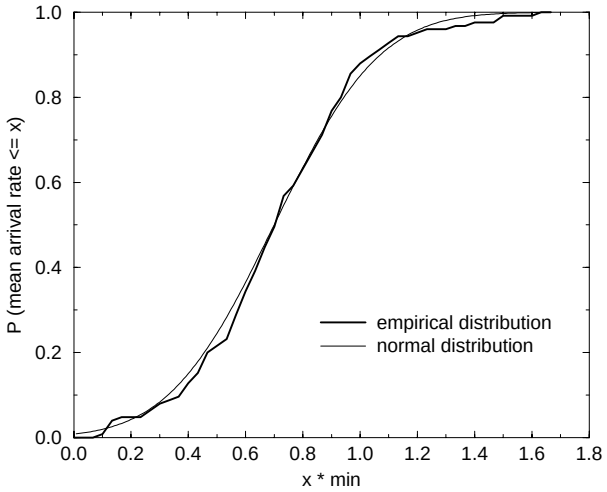

Figure 12: Approximation of empirical ccdf of 10 p.m. - 11 p.m. arrival rate averages

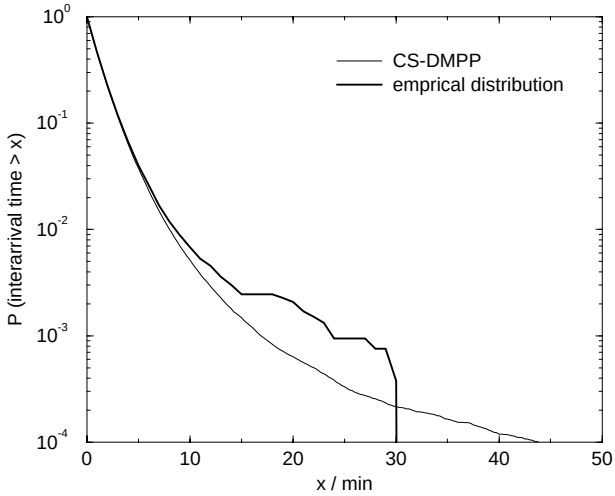

Figure 13: CS-DMPP interarrival time ccdf compared with empirical ccdf (10 p.m.- 11 p.m.)

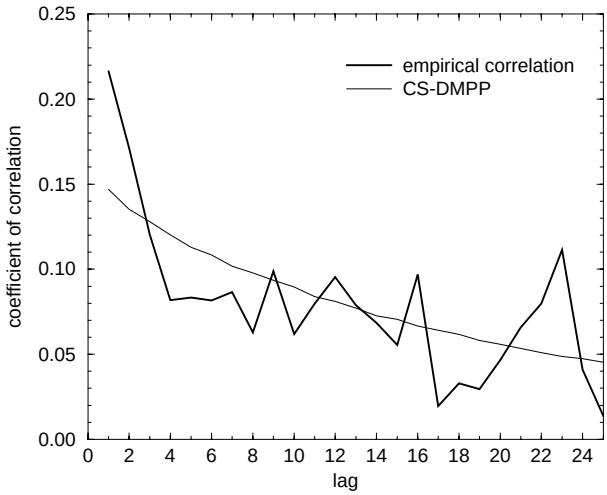

Figure 14: CS-DMPP interarrival time correlation compared with empirical correlation (10 p.m.- 11 p.m.)

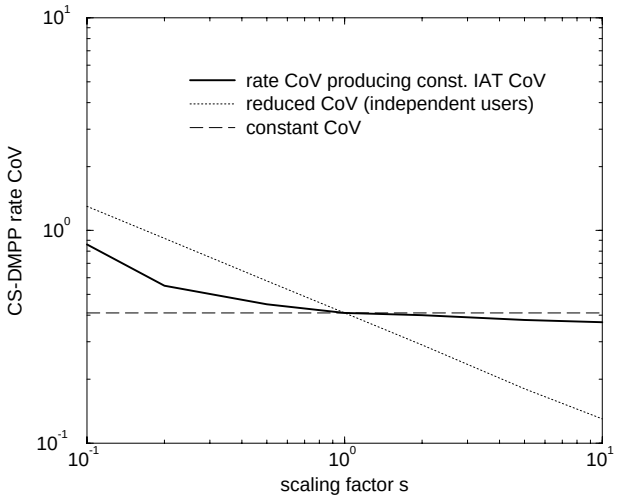

Figure 15: CS-DMPP rate CoV resulting in a constant interarrival time CoV

3.3. Economy of scale

If we use the results of Section 3.2 to scale our model of aggregated traffic, we are able to model network load for a different number of users. We assume that the characteristics for the holding time stay the same and adapt the distribution for the inter-arrival time by setting the mean reciprocal to the number of users and keeping the CoV constant.

Similarly to Figure 8 the loss simulations lead to the blocking probabilities for given numbers of servers and users. From this set of results we are able to conclude the required number of servers for a given number of users and a desired blocking probability B as depicted in Figure 16. The economy of scale, i.e. the gain in efficiency if the trunc size is increased for more users, is clearly visible in the figure. The required number of lines per user is slightly higher in the case where hyperexponential distributions are used for modelling than in the M/G/n case. The achievable system utilization, which is depicted in Figure 17, is generally smaller when the more exact kind of modelling is applied than if the system is dimensioned according to a Poisson model.

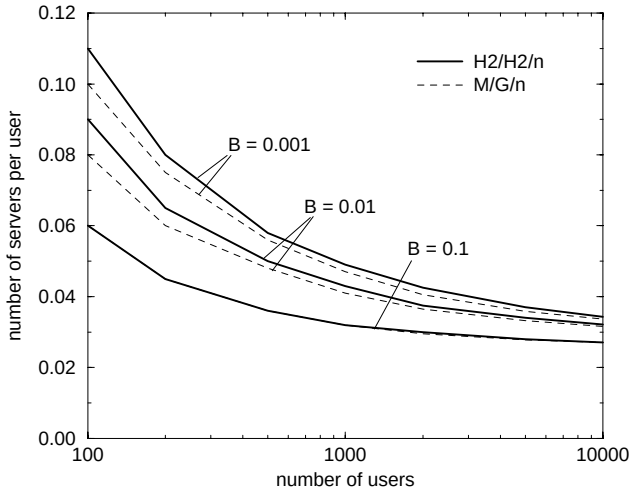

Figure 16: Required number of servers per user

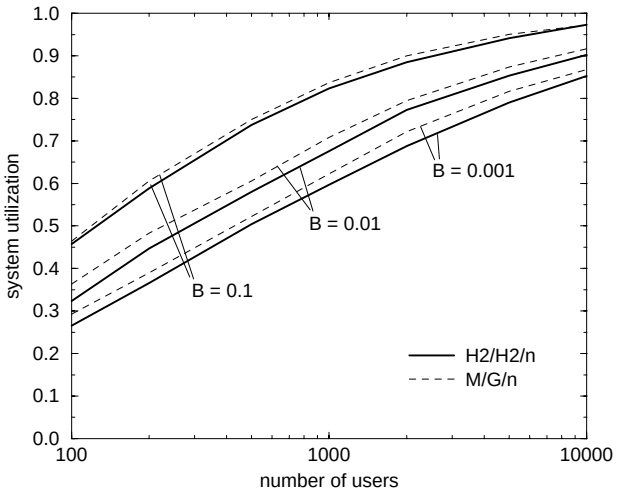

Figure 17: System utilization

4. CONCLUSIONS

In this paper we have described the dial-up behaviour of ISDN users at the University of Stuttgart. We presented a model for the aggregated traffic and evaluated it with respect to the performance in a loss simulation. We also suggested a method for scaling of our model.

Major results of the data evaluation are:

- dial-up sessions have much longer holding times than classical telephone calls,
- ISDN users generate more but shorter sessions (likely due to fast setup),
- dial-up usage follows the telephone tariffing scheme with high accuracy (Internet access itself was provided for free),
- session interarrival time and the session holding time show very high variability if regarded for the whole day and
- for the interarrival time and session holding time during busy hours, hyperexponential distributions give quite accurate models of the observed traffic with respect to loss behaviour.

The presented results are based on empirical data. Be aware that the described behaviour reflects the one of a special user group, i.e. the students and members of staff of the University of Stuttgart. Also, the constraints in terms of telephone tariffing and free Internet access should be taken into account when assessing our results. Although our description might not represent the general Internet user, we think that it does provide valuable insight in the behaviour of a large group of users.

We have shown that renewal processes based on second order hyperexponential distributions provide good estimations for the blocking probability during busy hours only by fitting the distributions to the first two moments of the trace data. Through the evaluation of the coefficient of variation for subsets of our user group, we were able to find a rule for the scaling behaviour of our model. Although we have found a theoretical explanation for this rule, evaluations of larger data sets are necessary to verify the exact behaviour of the coefficient of variation.

Before using our model in the range of high blocking probabilities, the effect of repeated call attempts has to be taken into account. This effect is expected to be much higher than with traditional telephony traffic due to computers being able to achieve a very high call repetition rate, especially if connected directly to an ISDN network.

ACKNOWLEDGEMENTS

We would like to express our thanks to our colleagues at the RUS, the computing centre of the University of Stuttgart, for their fruitful support without which this work could not have been carried out.

REFERENCES

1. Bolotin, V.A., “Telephone Circuit Holding Time Distributions”, *Proceedings of the 14th International Teletraffic Congress (ITC 14)*, Antibes, France, June 1994, pp. 125-134.
2. Bolotin, V.A., “Modelling Call Holding Time Distributions for CCS Network Design and Performance Analysis”, *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 3, April 1994, pp. 433-438.
3. Bolotin, V.A., “New Subscriber Traffic Variability Patterns for Network Traffic Engineering”, *Proceedings of the 15th International Teletraffic Congress (ITC 15)*, Washington D.C. , June 1997, pp. 867-878.
4. Crovella, M., Bestavros, A., “Performance Characteristics of World Wide Web Information Systems”, *Tutorial at the ACM SIGMETRICS ’97*, Seattle, June 1997.
5. Crovella, M., Bestavros, A., “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes”, *Proceedings of the ACM SIGMETRICS ’96*, Philadelphia, May 1996, pp. 160-169.
6. Färber, J., Bodamer, S., Charzinski, J., “Measurement and Modelling of Internet Traffic at Access Networks”, *EUNICE ’98*, Munich, August 1998, pp. 196-203.
7. Färber, J., Bodamer, S., Charzinski, J., “Evaluation of dial-up behaviour of Internet users”, *ITG-Fachtagung*, Stuttgart, October 1998, pp. 73-78.
8. Feldmann, A., Whitt, W., “Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models”, *Performance Evaluation*, Vol. 31, No. 3-4, January 1998, pp. 245-279.
9. Morgan, S., “The Internet and the Local Telephone Network: Conflicts and Opportunities”, *IEEE Communications Magazine*, Vol. 36, No. 1, January 1998, pp. 42-48.