COST 257 MID-TERM SEMINAR
INTERIM REPORT ON
SOURCE CHARACTERIZATION

Source Characterization in Broadband Networks

Edited by

Sándor Molnár, István Maricza

High Speed Networks Laboratory
Dept. of Telecommunications and Telematics
Technical University of Budapest
H–1117, Pázmány Péter sétány 1/D, Budapest, Hungary
E-mail: molnar@bme-tel.ttt.bme.hu
maricza@ttt-atm.ttt.bme.hu

Preface

This interim report provides an overview of results on source characterization obtained in the COST 257 Action on "Impacts of New Services on the Architecture and Performance of Broadband Networks".

First, the results related to fractal traffic characterization is presented. Multifractal analysis, the issue of non-stationarity and long-range dependent (LRD) traffic models are discussed. Three models have been investigated which can exhibit LRD features: quasi-Markovian models, superposed heavy-tailed on/off models and shifting level models. Finally, the impact of LRD on queueing performance is discussed in this section.

Second, the issue of spatial traffic characterization is addressed and the estimation and characterization of expected teletraffic demand is investigated.

Next, studies on the second order traffic descriptors to characterize MAPs and investigations on the application of generalized peakedness are outlined.

Finally, research results concerning Internet traffic characterization are presented.

The report is based on the COST 257 technical documents. The editors are thankful to P. Mannersalo for his contribution to the report. Contributors of the following authors have been used: A. T. Andersen, J. Arnold, C. Blondia, S. Bodamer, J. Charzinski, I. Cselényi, T. Daniels, J. Färber, M. Frater, M. Grasse, A. Koski, K. Laevens, T. Leskien, R. Macfadyen, P. Mannersalo, Gy. Miklós, S. Molnár, B. F. Nielsen, I. Norros, P. Tran-Gia, K. Tutschku, A. Vidács, J. Virtamo.

Contents

1 Introduction

Broadband networks are expected to support various applications. These applications (ranging from the traditional telephony to the multimedia services) can generate a heterogeneous mixture of traffic to the network. The nature of traffic can be very different considering the timeliness requirements of user applications (interactive or retrieval services) and the traffic profile at different time scales from macro levels (e.g. calls) to microscopic levels (e.g. cells or packets). Understanding the nature of this traffic, identifying its characteristics and developing appropriate traffic models are of crucial importance to the teletraffic engineering and the performance evaluation of our broadband networks being or to be developed.

One of the most important lessons learned from traffic measurements studies during the last two decades is that data traffic is highly variable and very bursty over many time scales [120]. Finding a good teletraffic framework dealing with this bursty data traffic led researchers to investigate fractal models [121]. Fractal traffic characterization is an important research topic of the COST 257 Action and activities on this field are summarized in Section 2 of this report.

Developments in the mobile world and especially the design of the third generation mobile communication networks resulted in a number of challenges in traffic modeling besides many other fields. The analysis of the distribution of the expected teletraffic demand in the complete service area is one of the important topics. The estimation and characterization of this traffic based on a geographical traffic model is discussed in Section 3 of the report.

What statistical descriptors of the arrival process can give an accurate prediction of the queueing behaviour? This important question is addressed in Section 4. It is well documented in the literature that in general queueing behaviour cannot be accurately predicted on the basis of first and second order properties of the counts of the arrival process. The topic is investigated in the framework of Markovian Arrival Processes (MAP) and results illustrated with simple examples. A study on the general framework of peakedness for traffic engineering is also presented in this section. We provide the computation of peakedness for a number of important discrete time models including the Markov modulated batch Bernoulli process and the batch renewal process. The analysis of generalized peakedness for measured data including MPEG video, ATM and Ethernet traffic is given.

Internet becomes more and more popular and there are many unsolved issues in the field of Internet traffic characterization. For example, we can observe that more and more users access Internet from their homes via public switched telephone network. To efficiently develop our telephone networks to this growing demand a good understanding of traffic characteristics is needed. This topic is addressed in Section 5.

2 Fractal traffic characterization

Recent traffic analysis studies based on measuraments taken from different LAN and WAN reported *high variability and burstiness* of network traffic over a wide range of time scales [120]. Statistically, this high traffic variability can be well captured by *long range dependence* (LRD), i.e. autocorrelations that exhibit a power-law decay. More precisely, a covariance-stationarity process $X = (X_k : k \geq 0)$ with autocorrelation function $r(k), k \geq 0$ is said to exhibit LRD if $r(k) \sim k^{2H-2}L(k)$ as $k \to \infty, 1/2 < H < 1$, where L is slowly varying at infinity, i.e. $\lim_{k\to\infty} \frac{L(tk)}{L(k)} = 1, \quad t > 0$ and $a(x) \sim b(x)$ means $a(x)/b(x) \to 1$ as $x \to \infty$. The parameter H is called the *Hurst parameter* and used as a measure of the degree of LRD. The LRD is also refered as *Joseph effect* or *persistence phenomenon*.

This hyperbolically decaying autocorrelation is an important property of *self-similar* processes. Self-similarity is a mathematical concept which is related to *fractals* and offer a promising framework of modeling highly bursty network traffic that exhibit LRD. Formally, a process X is *exactly self-similar* if $X_k \stackrel{d}{=} X_k^{(m)}$, where $X_k^{(m)} = \frac{1}{m^H} \sum_{i=(k-1)m+1}^{km} X_i$ and the equality is in the sense of finite-dimensional distributions. We can also define *second-order self-similarity*, i.e. $r^{(m)}(k) = r(k)$, and *asymptotically second-order self-similarity*, i.e. $\lim_{m\to\infty} r^{(m)}(k) = r(k)$ where $r^{(m)}(k)$ denotes the autocorrelation of process $X_k^{(m)}$.

Models based on the concept of self-similarity have been developed and applied in several research studies [120]. The *Fractional Brownian Motion* which can be found the limit traffic in many traffic aggregation is investigated in [105] and related results obtained in the COST 257 activity is found partly in this report, and partly in the interim report on "Queueing Systems". Other models, like the *fractional ARIMA processes* are also under investigations [105].

Several studies have been carried out to investigate the physical explanations of the observed fractal properties in network traffic. For example, it was shown by Willinger et al. that the self-similarity properties found in Ethernet LAN can be well explained by a structural model where each source is modeled by an ON/OFF model having heavy-tailed distributions with infinite variance of the ON and/or OFF periods. The aggregation of traffic generated by these sources produce self-similar traffic [119]. Other limit results for aggregated WAN traffic based on the $M/G/\infty$ *queueing model* due to Cox [18] and a more refined model due to Kurtz [59] can also provide models to explain self-similar traffic dynamics. An important finding of these models is that *heavy-tailed* distributions play an important rule in the appeared fractal properties [119].

An important generalization of self-similar processes is *multifractal* processes. Multifractals provide more flexible scaling properties which seem to be needed to capture local irregularities of some types of network traffic. In this section the COST 257 results related to multifractal analysis is presented.

Non-stationarity of network traffic can also produce properties detected by many statistical methods which are similar to fractal properties. Moreover, in some cases non-stationarity models can offer an alternative approach to capture these properties. This section performs the analysis of non-stationarity and introduces a video model based on a level shifting process. We also address the impact of LRD on queueing. Investigations based on both measured and synthetic LRD traffic are discussed.

2.1 Multifractal analysis

In this section, we consider quite a new approach in telecommunication engineering — multifractal analysis. Multifractals themselves are a relatively old topic first introduced by Mandelbrot in the context of turbulence [75, 76] in early 70's. In telecommunications, multifractals were introduced only recently: Appleby combined multifractal analysis of population distributions with network planning [5, 6], Riedi and Lévy-Véhel applied multifractal analysis to data traces [104], Taqqu, Teverovsky and Willinger considered whether network traffic is self-similar or multifractal [112], Feldmann, Gilbert and Willinger motivated the multifractal nature of data traffic using a cascade based construction [32], and Riedi et al. developed a multiscale modeling framework suitable for network traffic characterization [103].

Also in COST257 project, some work related to multifractal analysis has been done. Inspired by Riedi and Levy-Vehel's work we have considered multifractal traffic characterization with real ATM trace and multifractal spectra of some standard source models [79, 78], and on the other hand, following the Appleby's ideas, we have studied effects of the multifractal nature of a population distribution on network planning [77].

In this report, we present some aspect of multifractal analysis related to data traces.

2.1.1 Some basics of multifractal analysis

Multifractal analysis has clearly an advantage compared with standard statistical approaches because it gives information about both local and global properties of the observed data. The local, possibly singular, behavior is measured by the Hölder exponent at a point, and the global behavior is characterized by the statistical distribution of the occurring Hölder exponents. More strictly speaking, multifractal analysis provides an approach for characterizing a singular measure according to the distribution of the asymptotics of its local finite densities.

The reference [104] includes a nice introduction to the basic ideas of multifractal analysis and an extensive bibliography. Here we present only what is necessary for understanding the rest of this section.

Let us first consider the abstract setup in a simplified form. Let μ be a probability measure on $[0,1]$, and denote the Lebesgue measure by λ. Let I_n denote the partition into 2^n equal subintervals

$$I_n = \{[(k-1)2^{-n}, k2^{-n}),\ k = 1, \dots, 2^n\}$$

and let

$$C_n(x) = \text{ the element of } I_n \text{ containing } x.$$

We are interested in the distribution of the random variables $\mu(C_n(x))$ with respect to λ (in our application: the distribution in time of "cell arrival rates" defined at resolution

2^{-n}). Now, the theory of large deviations enters the game. For each n, define the "free energy function"

$$c_n(q) = \log E_\lambda \exp(q \log \mu(C_n(x))).$$

The Gärtner-Ellis theorem states that **if** the limit function

$$c(q) = \lim_{n\to\infty} \frac{c_n(q)}{n}$$

exists *and is differentiable* in its domain, assumed to contain 0 as an interior point, and if $|c'(q)| \to \infty$ as q approaches the boundary of $\{c < \infty\}$, then we have for every α

$$\lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log P_\lambda(|\frac{1}{n}\log\mu(C_n(x)) - \alpha| < \epsilon) = -c^*(\alpha),$$

where the *entropy function* c^* is defined as the convex conjugate

$$c^*(\alpha) = \sup_q (q\alpha - c(q)).$$

Note that $\frac{1}{n}\log\mu(C_n(x)) = \alpha$ means

$$\mu(C_n(x)) = \lambda(C_n(x))^{-\alpha/\log 2}.$$

We then call $h = -\alpha/\log 2$ the *coarse Hölder exponent* of the set $C_n(x)$. If the assumptions of the Gärtner-Ellis theorem hold, the distribution of coarse Hölder exponents is approximately given by the concave function $-c^*(-\alpha/\log 2)$.

Quite often in multifractal analysis, slightly different functions are used. Instead of the free energy function we use the *partition function*

$$\tau(q) = \lim_{\delta\to 0} \frac{\log S_\delta(q)}{\log \delta},$$

where the *partition sum* S_δ is defined by

$$S_\delta(q) = \sum_{C\in\tilde{I}_\delta} \mu(C)^q,$$

and the summation is done over the set $\tilde{I}_\delta = \{[k\delta,(k+1)\delta) \mid k \in \mathbb{N}, \mu([k\delta,(k+1)\delta)) \neq 0\}$. Correspondingly, instead of the entropy function, we have the *Legendre spectrum*

$$f_L(h) = \inf_q \left(hq - \tau(q)\right). \tag{1}$$

These functions are equal to c and c^* up to trivial transformations. Note that we write $\tilde{I}_\delta = \tilde{I}_n$ if $\delta = \delta_n$, and in this report, we consider only the case $\delta_n = 2^{-n}$.

However, the limit function $c(q)$, and thus $\tau(q)$, does not necessarily exist, and even if it exists, it need not be differentiable. In the latter case, c^* (and f_L) can be defined but it does not give the right information. In this case another, more sensitive notion of multifractal spectrum, the *coarse grained spectrum,* becomes important.

Using the coarse Hölder exponent, at the level of sampling $\delta_n = 2^{-n}$,

$$h(C_n) = \frac{\log \mu(C_n)}{\log \delta_n}$$

one can calculate the coarse grained spectrum

$$f_G(h) = \lim_{\epsilon \downarrow 0} \lim_{n \to \infty} \sup \frac{\log N_{\delta_n}(h, \epsilon)}{-\log \delta_n}, \tag{2}$$

where $N_{\delta_n}(h, \epsilon) = \# \left\{ C_n \in \tilde{I}_n \ : \ h(C_n) \in (h - \epsilon, h + \epsilon] \right\}$ (the number of intervals C_n of size δ_n with coarse Hölder exponent near h).

The coarse grained spectrum measures the exponential speed that the probability of observing a coarse Hölder exponent different from the expected value approaches zero as the resolution tends to infinity. If the conditions for the large deviation principle are satisfied then f_G can be calculated as the Legendre transformation of the partition function τ, i.e., $f_G = f_L$. Unfortunately, those conditions are rarely satisfied in traffic analysis. However,

$$\tau(q) = \inf_{\alpha \in \mathbb{R}} (\alpha q - f_G(\alpha))$$

holds in every case. This means that f_L is the concave hull of f_G. In spite of the fact that f_L does not give us any extra information about spectra, it is worth of computing. Especially the behavior of the partition sum $S_n(q)$ with different resolutions includes essential information about multifractal scaling.

2.1.2 Multifractal analysis of data

Suppose that we have a sampling of a measure μ at the resolution N. In order to check whether there exist multifractal scaling we calculate the partition sum with several values of q and over several resolutions:

$$S_m(q) = \sum_{C \in \tilde{I}_{\log N/m}} \mu(C)^q,$$

where $\tilde{I}_l = \{ [k2^{-l}, (k+1)2^{-l}) \ : \ k \in \mathbb{N}, \ \mu([k2^{-l}, (k+1)2^{-l})) \neq 0 \}$, and for example, $N = 2^n$ and $m = 1, 2, 2^2, \ldots, 2^n$. If $S_m(q)$ is a linear function of m in some region in the log–log scale we say that the region in question is the scaling region and the measure exhibits multifractal scaling there. Furthermore, we can approximate the partition function $\tau(q)$ by solving the equation

$$\log S_m(q) \approx \tau(q) \log m + \text{const.}$$

for $\tau(q)$ in the least square sense in the scaling region. After finding τ, an approximation for the Legendre spectrum is found numerically by applying (1) at the calculated values of q.

Approximating the coarse grained spectrum f_G is a more complicated task. In (2) we must take limits over two variables, namely, ϵ and n. In practice, given a sampling at the resolution 2^n one must determine a suitable value for ϵ, i.e., ϵ as a function of n, such that the corresponding approximation is near the right one. This can be done by some density estimation method, for example, by using the double kernel method [24].

2.1.3 Multifractal analysis of recorded ATM traffic

Let us apply multifractal approach into the analysis of real ATM traffic. Twenty traces of daytime ATM traffic were recorded at Tampere University of Technology. These recordings consist of arrival times of cells into a ATM link during the observation interval, each recording lasting about two seconds. The traffic is the output of a router with an ATM interface, transmitting Internet packets from TUT to the Finnish University Network. Four such traces are shown in Figure 1.

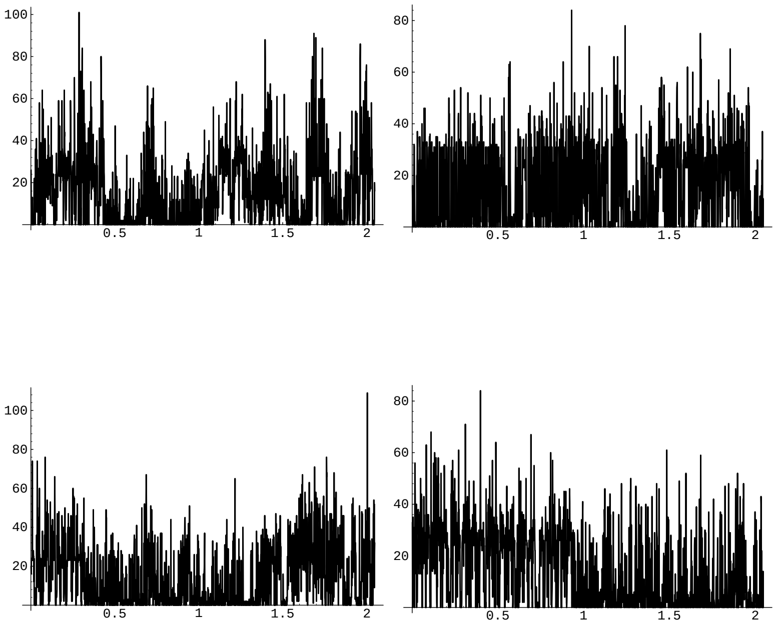


Figure 1: Four traces of ATM traffic observed at Tampere University of Technology. The number of cells per millisecond are plotted.

We consider the atomic measure counting arrivals. The corresponding measure is extremely well scalable confirming that calculating multifractal spectra is reasonable (see Figure 2). The Legendre spectra shows that there is a quite wide range of different intensities in the observed traffic, and that it is a bit more probable that traffic is light (because of the slight non–symmetry of the spectrum).

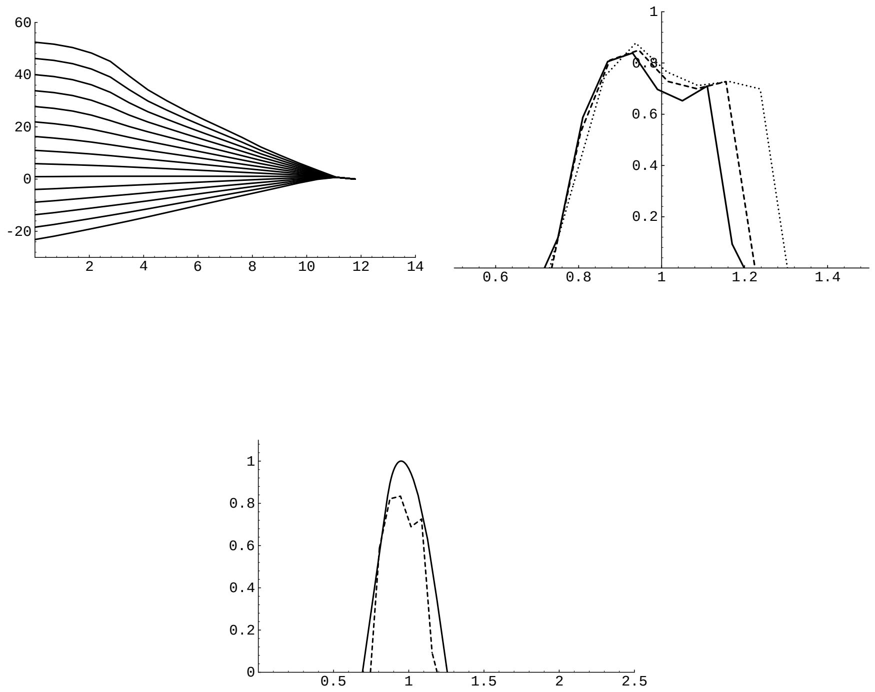


Figure 2: Multifractal analysis of the atomic measure of the arrival times originated form the recorded ATM traffic at TUT. The log–log plot of the partition sum is shown on the left ($q = -3.5, -3.0, \ldots, 3.0, 3.5$), the evolution of the approximation of the coarse grained spectra over resolutions 2^{17}, 2^{16} is on the left and 2^{15}, and the Legendre spectrum and the "best" approximate coarse grained spectrum are plotted below.

The really noteworthy result is seen in the coarse grained spectra. The two peaks seen suggests that the recorded traffic might include two different phases. In a matter of fact, the ATM traffic concerned includes two possible speeds that cells can be sent. It is quite probable that the peaks correspond to them.

2.1.4 Multifractal models

The standard source models, e.g., Poisson processes or heavy tailed renewal processes, do not have true multifractal nature and their scaling region is not wide enough (see [78]). Unfortunately, nor the multifractal properties of fractional Brownian motion are consistent with those of measured traffic traces (see [65]).

Binomial measures and their randomized versions are often introduced as the very first multifractal models, (see e.g., [104, 32]). Even these simplest models can reproduce some interesting statistical properties observed in real traffic. Here we consider a single example. Let the sequence of partitions I_n be defined as in section 2.1.1, and define the measure μ by the following randomized procedure. Let

$$M_{nk}, \quad n = 1, 2, \ldots, \quad k = 0, \ldots, 2^{n-1} - 1,$$

be i.i.d. random variables with values in $(0, 1)$ such that the distribution of $M = M_{nk}$ is symmetric around $1/2$, i.e., M and $1 - M$ have the same distribution.

The measure of the dyadic intervals is defined recursively by

$$\begin{aligned} \mu([2k2^{-n}, (2k+1)2^{-n})) &= M_{nk}\mu([(k2^{-n+1}, (k+1)2^{-n+1})) \\ \mu([(2k+1)2^{-n}, (2k+2)2^{-n})) &= (1 - M_{nk})\mu([k2^{-n+1}, (k+1)2^{-n+1})). \end{aligned}$$

The measure of a small dyadic interval $[k2^{-n}, (k+1)2^{-n})$ has approximately the distribution Lognormal$(nm, n\sigma^2)$, where

$$m = \mathrm{E} \log M, \quad \sigma^2 = \mathrm{Var}(\log M).$$

Lognormal marginal distributions have often been observed in real traffic (see, e.g., [105] p. 364). Here they arise as a consequence of a multiplicative structure of the measure. Such a structure can be thought of as arising from the hierarchical (multilayer) nature of telecommunication: the often sited "burstiness at several time scales". Of course, the distribution of M_{nk} would be in a more accurate model depend on n, but for qualitative understanding we can study the simplified scheme in a similar way as self-similar processes have be used in traffic modeling.

Let us consider the variance of the measure of a dyadic interval. Then we have

$$v(2^{-n}) \doteq \mathrm{Var}(\mu([k2^{-n}, (k+1)2^{-n}))) = (\mathrm{E}M^2)^n - 2^{-2n}.$$

Substituting $t = 2^{-n}$, this reads

$$v(t) = t^{-\log_2 \mathrm{E}M^2} - t^2.$$

The second term is similar to the error term in a sample variance (cf. [105] (13.6.1)). In any case, the first term dominates for small t. Thus the variance growth for small t is approximately of the self-similar form $v(t) = t^{2H}$ with

$$H = -\frac{1}{2} \log_2 \mathrm{E} M^2.$$

Note that $\mathrm{E}M^2$ lies in the interval $(\frac{1}{4}, \frac{1}{2})$ which allows for H all values in $(\frac{1}{2}, 1)$.

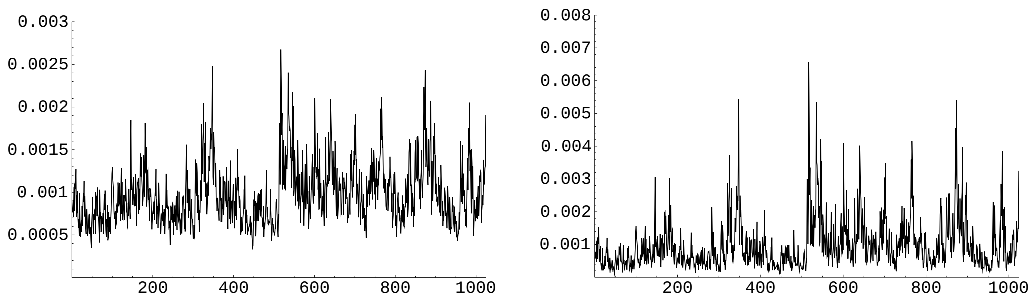


Figure 3: Binomial random measures at resolution 2^{-10}. $M = p + (1 - 2p)U$, U is Uniform(0,1) and $p = 0.4$ (left) and $p = 0.3$ (right).

Let us have a look at some simulated realizations where $M = p + (1 - 2p)U$, U is Uniform(0,1) and $p \in (0, \frac{1}{2})$ is a parameter. Figure 3 shows the cases $p = 0.4$ and $p = 0.3$. Both resemble visually real traffic traces with long range dependence. The first one could be sold to a non-statistician as fractional Brownian motion, whereas the second looks very different having strongly non-Gaussian character.

In Figure 4, scaling of the partition sums and the Legendre spectra are shown. It is quite evident that even with this very simple model by choosing properly the parameter p and the distribution of U, we could get traces whose multifractal properties resembles those of real traffic. For a more complicated models see, e.g., [103, 32].

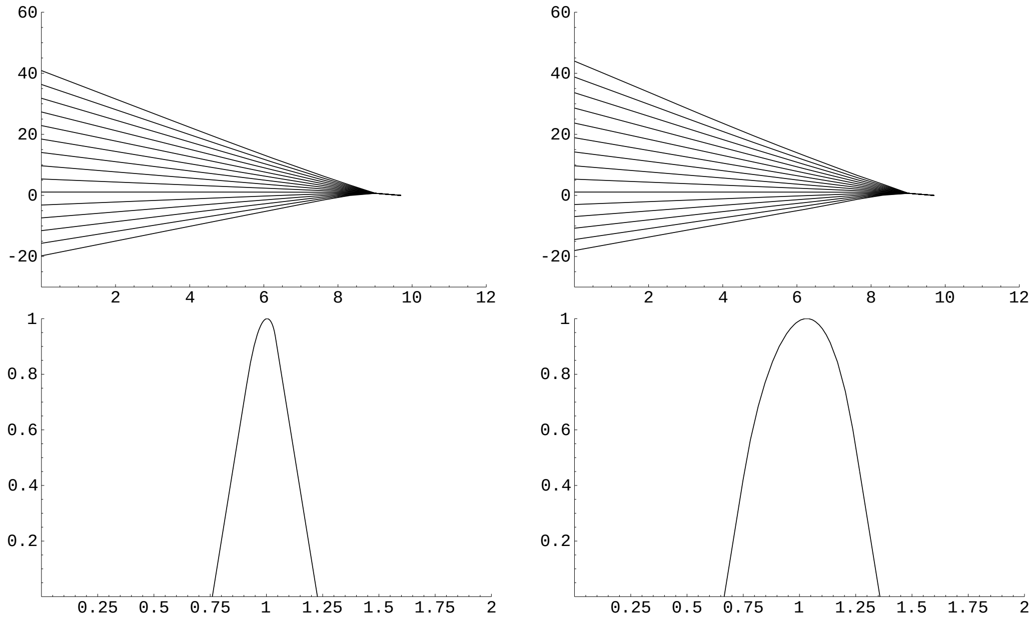


Figure 4: Multifractal scaling and Legendre spectra of binomial random measures. $M = p + (1-2p)U$, U is Uniform(0,1) and $p = 0.4$ (left) and $p = 0.3$ (right)

2.2 The problem of non-stationarity

This section addresses the issue of stationarity which is an important assumption of a number of traffic models. This issue becomes more important in the context of fractal models which tempts researchers to use this assumption on a global time-scale.

The analysis of testing for self-similarity and the estimation of the Hurst parameter are not easy in practice. The problem is that we are of course always dealing with finite data sets so it is not possible to check whether by definition a traffic trace is self-similar or not. We are therefore forced to look for different features of self-similarity and long range dependence in our actual measured traffic. However, the detection of long range dependence only by identified properties could be misleading. Several non-stationary processes, e.g. level shifting processes [25] which can be observed in the superposed effects of different protocol levels [54] (ATM interface card based bursts, IP frames, window mechanisms, session procedures, etc.) can produce such properties. *It means that if we found the traffic to be Hursty*[1] *it is due to long range dependence or non-stationarity.* Without any proof by rigorous statistical tests of stationarity in many cases it is only reasonable to discuss about *Hursty behaviour over a given time-scale for a given data set* [90, 92].

An analysis based on measured ATM WAN traffic is presented in this section. It is shown that the presence of non-stationarities can deceive our simple self-similarity tests and Hurst parameter estimation methods [90, 92]. As an example, based on a formal statistical test the assumption of weak stationarity of variable bit rate (VBR) video traffic is questioned [41].

2.2.1 ATM traffic measurements

The measurements were made on the FUNET ATM WAN network. 'FUNET' stands for 'Finnish University and Research Network', which provides primarily Internet services to its members based on TCP/IP-protocol. All these services are provided by CSC—Center for Scientific Computing which is a national service center that specializes in scientific computing and data communications providing modeling, computing and information services for universities, research institutes and industry. The FUNET long-distance network is built on Telecom Finland's ATM network. All the Nordic national networks (FUNET, DENnet, ISnet, SUNET and UNINETT) are connected to the Nordic Backbone Network (NORDUnet) which has a connection point in Stockholm, Sweden. NORDUnet has connections to the US backbones, the European backbones and to networks in Central and Eastern Europe [68].

The measurement was made at the CSC in Espoo, Otaniemi. This location is in the logical center of the whole network. All the international links start from here, including the main crosslink to Stockholm. Our measurement equipment was inserted between the network and the high-capacity ATM switch situated in Espoo (see Figure 5). From that point all the ATM traffic from the FUNET network transported through

[1] Hursty traffic means that the classical Hurst parameter estimation methods, e.g. R/S tests, estimate a Hurst parameter which is bigger than 0.5.

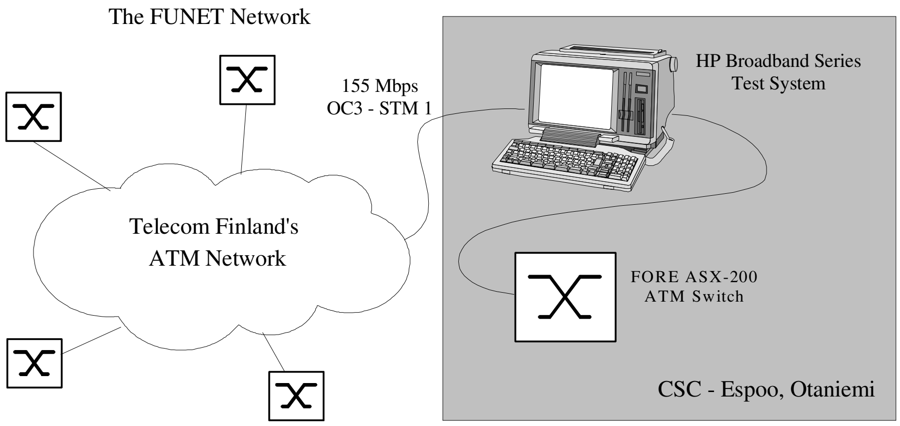


Figure 5: The FUNET measuring configuration.

the switch and the traffic generated at the CSC and transmitted to the rest of the world could be monitored. The measurements were made by an HP Broadband Series Test System equipment. During our work only the cell capture capability of the measurement unit was used: 131 072 ATM cells mapped into a 155 Mbps SONET/SDH signal can be recorded into the 8MB of capture memory of this equipment. All cells are timestamped with the calendar time with resolution 0.01 μs.

The aggregated traffic at the most heavily loaded point of the FUNET network was measured, including Internet traffic, data transfer and supercomputer usage. During the measurements, two types of data collections were made. In the first scenario the measured data was the time stamp of the arrival time instant for every single cell on the link. Because of the upper limit for the number of captured cells each measured data file contains 131 072 time stamps only, which corresponded to about 3–5 seconds according to the network load. For the long-term analysis longer measurement periods were needed, so in the second measurement scenario the recorded data was the number of cells received in a one second interval. In this case the time interval of the observation could take several minutes long. A summary of these data sets is given in Table 1. The files FUNET1, FUNET2 and FUNET3 contain traffic data captured from the incoming traffic from the whole country to the CSC, and the FUNET4 measurement was made on the outgoing link. In the case of the last two measurements in Table 1, the registered data was the number of cells received in every second on the incoming link. The average traffic load was about 14 Mbps for the first three measurements, and about 8 Mbps in the case of the FUNET4 data. In the following, we refer to the data above listed as the ‘FUNET measurements’.

As for the first four data sets, the measurement unit was able to register the VPI and VCI fields from the cell headers, too. Using this extra information we can reveal the structure of the aggregated traffic stream. Comparing the VPI/VCI fields the aggregated cell stream can be divided into independent connections. (Note that connection means

Table 1: Qualitative description of the measured data sets and the values of Hurst-parameter H calculated from different statistical methods

Filename	#packets	Time (sec)	$\hat{H}_{idc}$	$\hat{H}_{var}$	$\hat{H}_{rs}$	$\hat{H}_{per}$
FUNET1	131,072	3.9	0.7	0.7	0.68	0.68
FUNET2	131,072	5.1	0.67	0.67	0.67	0.73
FUNET3	131,072	4.4	0.66	0.66	0.68	0.68
FUNET4	131,072	6.4	0.72	0.72	0.74	0.78
FUNETSTA.T3	14,807,546	425	0.70	0.70	0.82	0.94
FUNETSTA.T4	43,768,430	1964	0.67	0.67	0.79	0.90

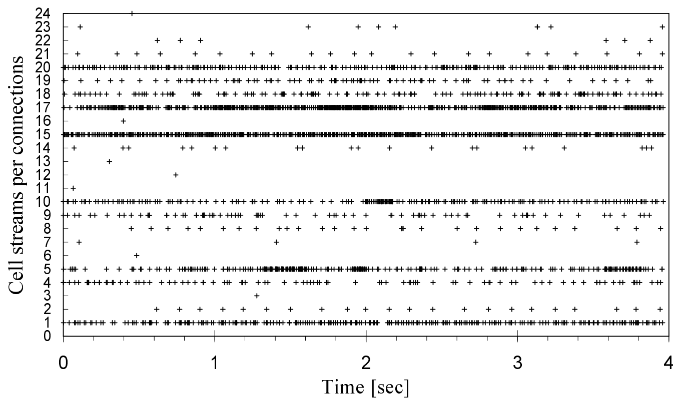


Figure 6: The structure of the FUNET1 traffic trace.

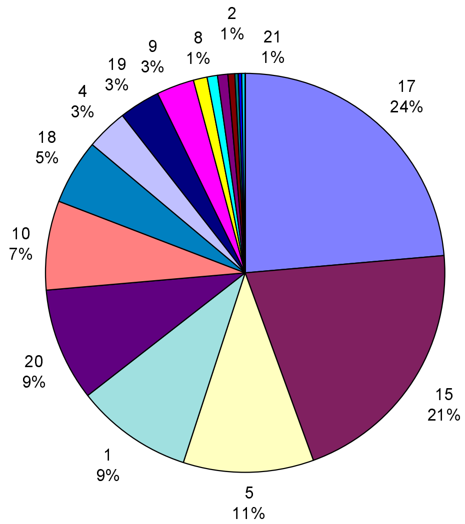


Figure 7: The bandwidths of connections in the FUNET1 data.

a cell stream with common VPI and VCI fields in the headers. We do not have any information about the type of traffic carried by these cell streams.) The most important piece of information for us is the number of connections and their relative cell rate compared to each other. A detailed analysis was made for the FUNET1 data set. Figure 6 shows the separated cell streams schematically. (Because of the huge number of cells each hair-cross represents every 50th cell arrival in a connection.) As can be seen from the figure, during the 4 second measured time period 24 connections were in progress. The connection with highest rate contains about 30 000 cells which is about 24 percent of the whole aggregated traffic as well as the first dozen with highest intensity contain 99% of all the cells. Figure 7 shows the pie-chart of bandwidths of connections.

In our investigation the question of stationarity is fundamental. As far as it can be concluded from Figure 6 without a comprehensive stationary analysis, the cell streams are homogeneous enough in time apart from the bursty nature of ATM traffic. There is no connection turned on or off in the middle of the measurement time and the rates apart from the burstiness are not changing considerably.

2.2.2 Hurst parameter estimation

Index of Dispersion for Counts is a commonly used measure for capturing the variability of traffic over different time scales [19]. For a given time interval of length t, the index of dispersion for counts (IDC) is given by the variance of the number of arrivals A_t during the interval of length t divided by the expected value of the same quantity:

$$IDC(t) = Var\{A_t\}/E\{A_t\}. \tag{3}$$

For a finite data set, the variance of A_t can be calculated by dividing the whole series into nonoverlapping blocks of length t and treat them as different instances of A_t.

Self-similar processes produce a monotonically increasing IDC of the form $m^{-1}t^{2H-1}$. Plotting $\log IDC(t)$ against $\log t$, this property results in an asymptotic straight line with slope $2H-1$ [62].

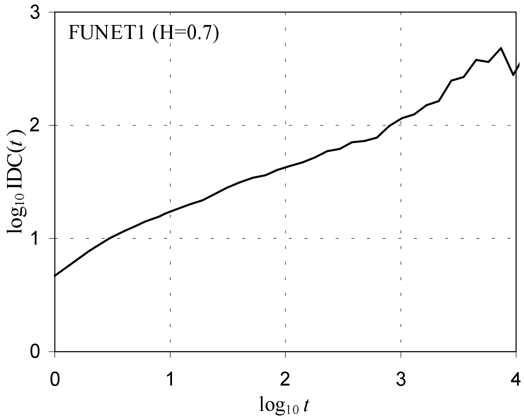
Figure 8(a) depicts the IDC curve corresponding to the trace FUNET1. The sequence of cell counts in every 100μs interval was analyzed. The IDC curve for the FUNET1 file increases monotonically throughout a time span that covers 3–4 orders of magnitude and shows an asymptotic slope that is strictly different from the horizontal line and is estimated to be about 0.4, resulting in an estimate $\hat{H}$ of the Hurst-parameter H of 0.7.

The same analysis was made for all the data sets. Table 1 shows the results: the values of the estimated Hurst-parameter $\hat{H}$. As can be seen from the table, the values of $\hat{H}$ are pretty much the same for all the data sets. It is remarkable that in the case of the last two data sets the analyzed process was the sequence of cell counts in each second instead of 100μs as in the case of the first four sets. In spite of the fact that the time scale was four orders of magnitude higher the Hurst-parameter remained the same.

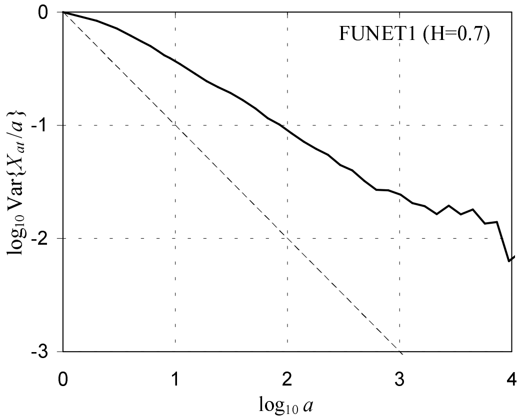
Variance-time analysis is a method based on the property that a self-similar process has slowly decaying variances. The so-called variance-time plot is obtained by plotting $\log Var\{X_{at}/a\}$ against $\log(a)$ and by fitting a simple least squares line through the resulting points in the plane, ignoring the small values of a. Values of the estimated asymptotic slope $\hat{\beta}$ between -1 and 0 suggest self-similarity, and the estimate for degree of self-similarity is given by $\hat{H} = 1 + \hat{\beta}/2$ [62].

The corresponding plot for the FUNET1 data set can be seen in Figure 8(b). The estimated values of $\hat{H}$ are listed in Table 1. Since the variance-time plots and the IDC diagrams are closely related statistical methods, the results obtained from this method are the same as in the previous subsection.

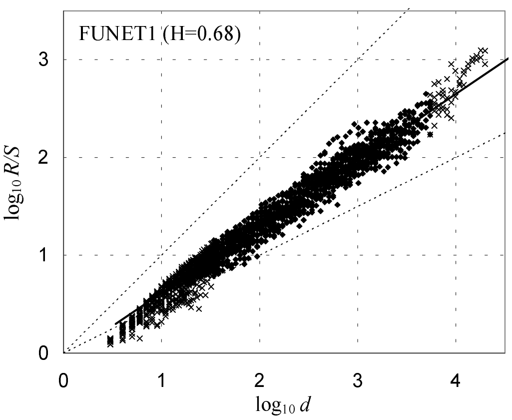
The *R/S analysis* tries to capture the Hurst parameter based on the rescaled adjusted range statistics. Given an empirical time series of length N ($X_k : k = 1, \ldots, N$), the whole series is subdivided into K non-overlapping blocks. Now, we compute the rescaled adjusted range $R(t_i, d)/S(t_i, d)$ for a number of values d, where $t_i = \lfloor N/K \rfloor (i-1) + 1$



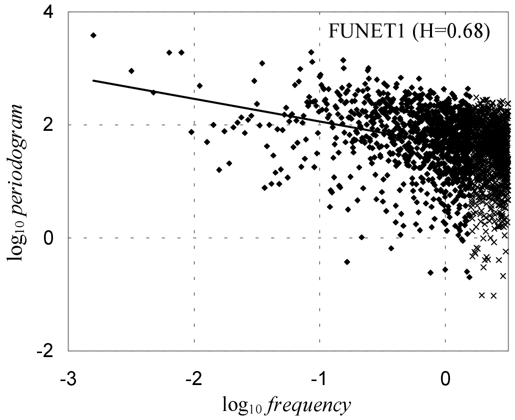
(a) IDC plot



(b) variance-time plot



(c) R/S diagram



(d) periodogram plot

Figure 8: Diagnostic plots for FUNET1 data

are the starting points of the blocks which satisfy $(t_i - 1) + d \leq N$.

$$R(t_i, d) = \max\{0, W(t_i, 1), \ldots, W(t_i, d)\} - \min\{0, W(t_i, 1), \ldots, W(t_i, d)\}, \qquad (4)$$

where

$$W(t_i, k) = \sum_{j=1}^{k} X_{t_i+j-1} - k \cdot \Big(\frac{1}{d}\sum_{j=1}^{d} X_{t_i+j-1}\Big), \quad k = 1, \ldots, d. \qquad (5)$$

$S^2(t_i, d)$ denotes the sample variance of $X_{t_i}, \ldots, X_{t_i+d-1}$. For each value of d one obtains a number of R/S samples, which decreases from K for larger values of d. One computes these samples for logarithmically spaced values of d, i.e., $d_{l+1} = m \cdot d_l$ with $m > 1$, starting with d_0 of about 10. Plotting $\log R(t_i, d)/S(t_i, d)$ vs. $\log d$ results in the R/S plot, also known as *pox diagram*.

Next, a least squares line is fitted to the points of the R/S plot, where both the R/S samples of the smallest and largest values of d are omitted. The slope of the regression line is an estimate for H [105].

Figure 8(c) shows the R/S plot for the FUNET1 data. The analyzed process was the sequence of cell counts in every 100μs. The estimated value of H for this data set is 0.68, which is nearly the same as the values calculated by the two previous methods. The same analysis was made for all the FUNET measurement data sets (see Table 1).

Periodogram-based analysis is used to identify the manifestation of self-similarity by frequency domain analysis of the measured data. Let $I(\cdot)$ denote the sample periodogram (i.e., power spectrum as estimated using a Fourier transform) defined by

$$I(\lambda) = \frac{1}{2\pi N}\left|\sum_{j=1}^{N} X_j e^{ij\lambda}\right|^2, \qquad \lambda \in [0, \pi). \qquad (6)$$

The spectral density of self-similar processes obeys a power law near the origin. Thus, the first idea to determine the Hurst parameter H is simply to plot the periodogram in a log-log grid, and to compute the slope of a regression line which is fitted to a number of low frequencies. This should be an estimate of $1 - 2H$. In most of the cases this will lead to a wrong estimate of H since the periodogram estimation method is unbiased and inconsistent. However, this method can reveal the power spectrum near the origin. The periodogram plot is obtained by plotting $\log(I(\lambda))$ against $\log \lambda$.

Figure 8(d) presents the periodogram plot for the FUNET1 data set, where the analyzed time series was the number of cells in every 1 msec. The slope of the low frequency part—in the present context, the regression line was fitted to the lowest 50% of all frequencies—is clearly different from zero, the slope estimate is about -0.36 which yields $H = 0.68$. This result corresponds to the previously calculated values of H.

The analysis was made for all the data sets, and the results are listed in Table 1.

Summary To summarize the results listed in Table 1, we conclude that:

- The estimated values of the parameter H are definitely greater than 0.5 for all cases.
- The values of H are nearly the same for all of the four analysis methods and for all the data sets. The common value for it is about 0.7. (Apart from the last two values for the FUNETSTA data sets.)

In spite of this, it would be too early to say that it follows from the results above that the measured traffic is self-similar with self-similarity parameter 0.7. To establish such a statement, we should carefully examine the applied analysis methods with their preliminary conditions and confidence intervals as well as the structure of the analysed data sets in more details. The applications of more robust and thorough methods (e.g. Whittle methods, wavelet methods) are also recommended.

In the next section, we investigate the problems arising during the calculation of the parameter H and determine those effects which can influence the results considerably.

2.2.3 Problems of testing

In practice, using measured data sets the estimated values of H obtained from different analysis methods are influenced by the dependences on estimating technique, sample sizes, time scales and data structure. These problems are discussed in [90, 92, 72] and some examples are presented in this section.

The choice of Hurst parameter for a given scenario is not straightforward. Indeed the Hurst varies considerably from customer to customer and over time. This ties in with comments made by Paxson and Floyd about the predictability, or lack of it, in Internet traffic [100]. In particular, the rate of change in the quantity and form of Internet traffic over recent years has shown how difficult it is to predict the traffic of the future.

Several universities were monitored to examine the Hurst parameter of a number of similar customers. 41 sites were monitored, each having a high-speed SMDS connection. These were monitored simultaneously for 1 hour at a 10 second resolution. The Hurst parameter was calculated using the rescaled adjusted range plot and variance time plot.

The Hurst parameter varied considerably between one access and another, despite the supposed similarity in their environment. The distribution of Hurst parameter values for these connections is shown below.

As expected from the notion of self-similarity, the aggregated traffic from all of the sites mentioned above was still self-similar. (The average Hurst parameter of individual sources weighted according to their mean was H=0.78; the aggregated traffic had a parameter of H=0.8.)

To further illustrate the difficulty in assigning a fixed value to the Hurst parameter for a particular customer's traffic, the following two traffic profiles represent two different busy periods for a single Frame Relay customer peaking around 1-2Mb/s. Traffic measurements were taken at a 2 second resolution.

The first 1-hour profile has a Hurst parameter of approximately 0.54. The second has a Hurst parameter of approximately 0.86. The graphs clearly illustrate the extreme variation in the Hurst parameter of even a single customer. In this example clearly

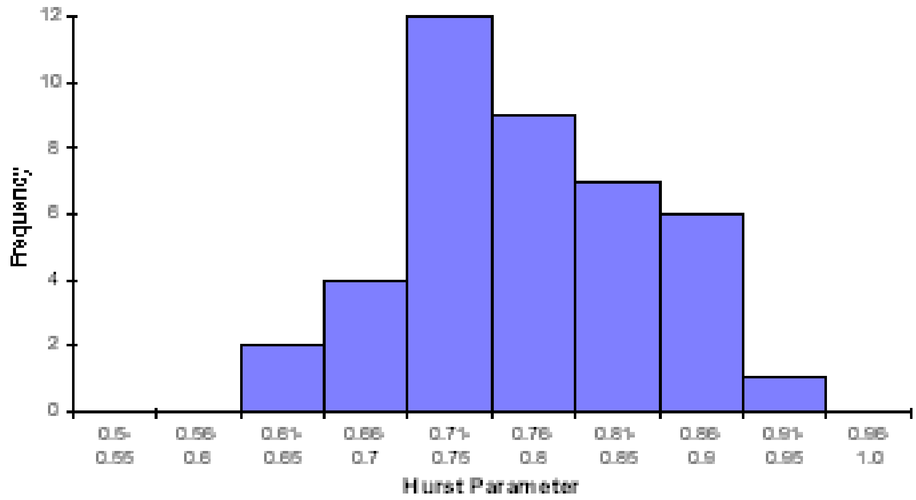


Figure 9: Distribution of Hurst Parameter for university traffic

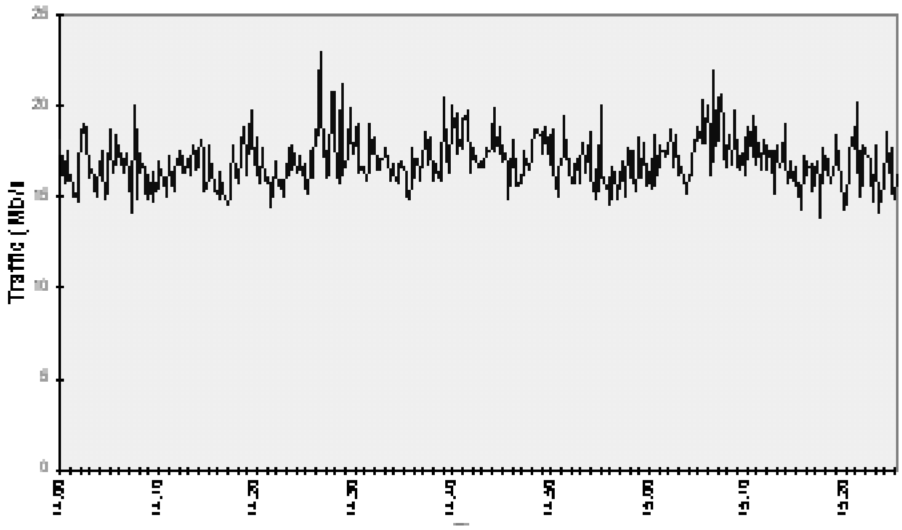


Figure 10: Aggregated SMDS traffic

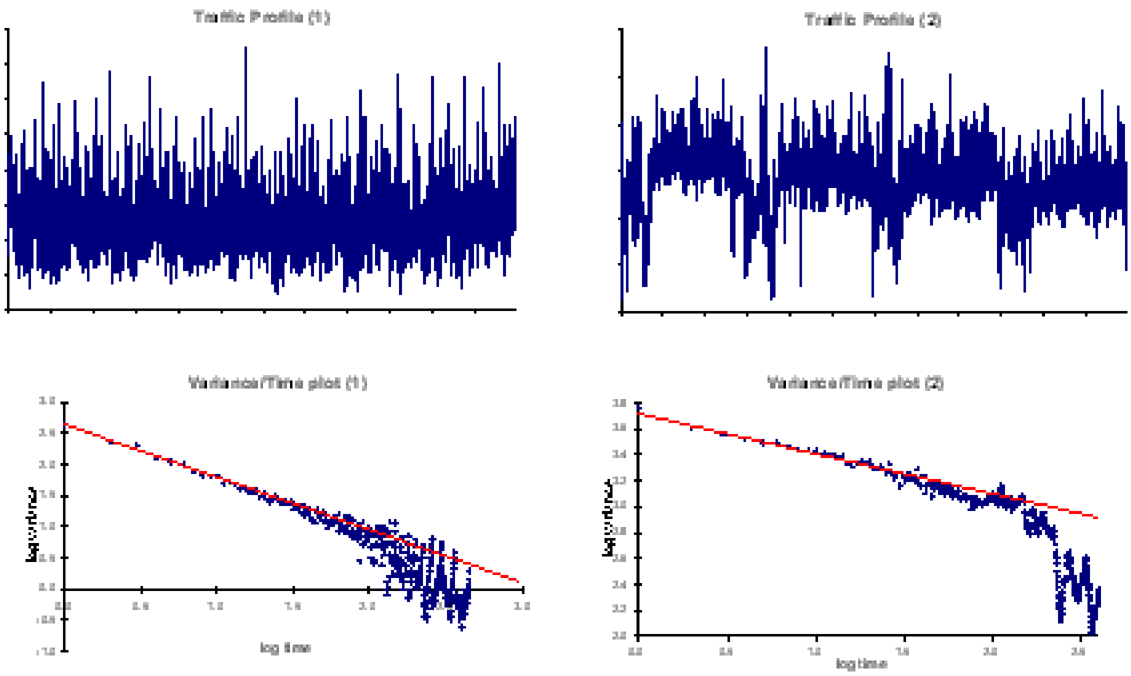


Figure 11: Traffic at different times of day for single customer

a major change in the usage had occurred. It may be argued that the self-similarity should be calculated based on a longer period, averaging out some of these anomalies, or alternatively that a two-second resolution is inadequate. However if the concept of self-similarity is to be useful, then it must be valid across the timescales of interest, to be able to relate the burst-scale characteristics to those found at the typical measurement resolution.

During our analysis we usually assume that the statistical properties of the measured cell stream are independent from time. This assumption is questionable, but—since the greatest part of statistical analysis methods require stationarity as a basic preliminary condition—it cannot be avoided. Strictly speaking, we assume that the measured process is stationary in the wide sense which means that its mean is finite and independent of time, and its autocorrelation function is finite and is invariant of time shift. (If we decided to treat our measured data sets as nonstationary sequences it would be almost impossible to make a comprehensive analytical study with meaningful results general enough to use elsewhere. Furthermore, in the case of finite data sets it is not possible to discriminate a stationary long-range dependent sequence from a nonstationary one.)

In this section we investigate the case when the assumption of stationarity does not hold (i.e., there is a level shift present in the measured traffic traces.) We examine how robust our statistical test is in case of a nonstationary cell sequence with a change in the mean as a function of time.

Example 1 In this first simple model nonstationarity is introduced by adding a CBR traffic to the second half of the measured data set. (Note, that this example represents not just a theoretical problem but a possible event in practice: while measuring the network traffic suddenly a new source may start to emit cells with constant cell rate.) Figure 12 shows the calculated IDC plots for these new multiplexed data sets. 2.8Mbps (CBR20—20% of the load of FUNET1) and 7Mbps (CBR50—50% of the load of FUNET1) CBR rates were applied.

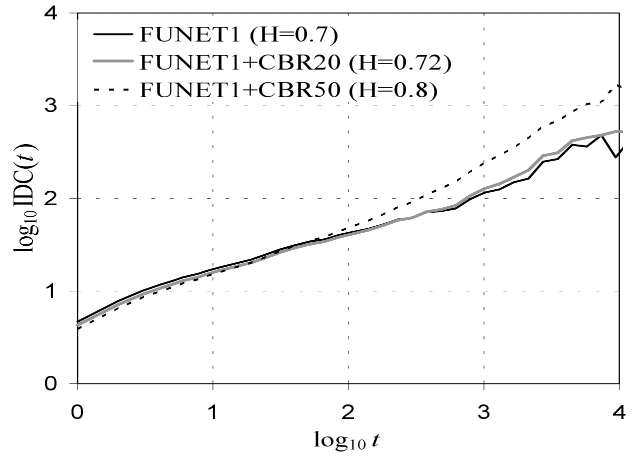


Figure 12: IDC plot for FUNET1 multiplexed with nonstationary CBR traffic.

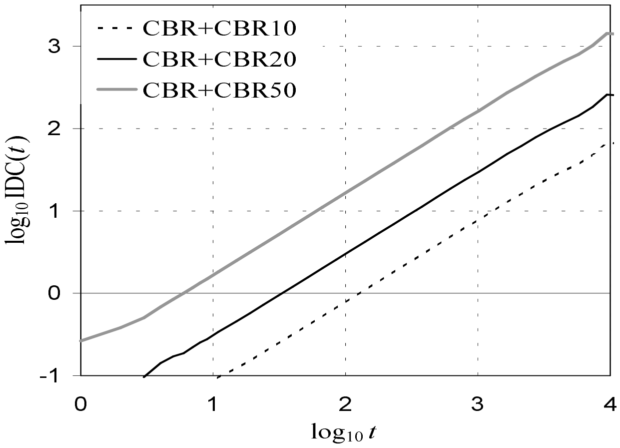


Figure 13: IDC plot for a CBR cell stream with level shift.

Discussion 1 The effect on the IDC plot is clearly visible. For the FUNET+CBR20 plot the upper part of the curve is moved up a bit as well as the lower segment shifted down slightly. As a result, the calculated Hurst parameter is greater, about 0.72. For the FUNET+CBR50 case the effect is sharper, the calculated value of $\hat{H}$ being 0.8.

Example 2 To understand the effects of level shift on the IDC plot more deeply, we investigated a simple CBR model in this example. As a starting point we chose a CBR traffic trace with the same rate as the mean rate of the FUNET1 traffic. The nonstationarity was introduced by increasing the CBR rate by 10, 20 and 50 percent abruptly at half time of the investigated time period. The IDC(t)value for an ideal CBR source without jitter is zero for all t which cannot be plotted on a logarithmical scale. The calculated IDC plots for the CBR traces with level shift can be seen on Figure 13.

Discussion 2 All the IDC curves are straight lines with slope 1. The only difference is that the IDC values are higher when the level shift is stronger.

The simplicity of the examined CBR model makes it possible to calculate the IDC(t) values analytically. (In practice, calculating the IDC plot for a finite data set means evaluating a double sum to estimate the mean and the variance. The following results are derived from this IDC estimator.) The IDC(t) for the above data sets is of the form:

$$IDC(t) = \frac{(a_1 - a_2)^2}{2(a_1 + a_2)} t,$$

where a_1 and a_2 are the cell rates for the first and second half of the data respectively. For $a_1 \neq a_2$ the IDC plot is given by:

$$\log IDC(t) \simeq const + \log t, \quad \text{where} \quad const = \log \frac{(a_1 - a_2)^2}{2(a_1 + a_2)}$$

which gives us a straight line with slope 1.

The main result here is the fact that although the CBR data with level shift has nothing to do with self-similarity, the estimated IDC is a monotonically increasing straight line with slope 1.

Example 3 The first example is generalized here by replacing the CBR traffic with a Poisson process. Again, the FUNET1 data was modified by adding a Poisson traffic to the second half of the measured data to increase the mean rate by 20 and 50 percent. The calculated IDC plots can be seen in Figure 14.

Discussion 3 The effect of nonstationarity in the plots is the same as in Example 1. The upper part of the curves moved up and the lower-left segments are shifted down simultaneously, resulting in higher Hurst parameter estimates.

Example 4 To make the effect of level shifts on the IDC plot clearer, in this example a simple but inhomogeneous Poisson process is examined which changes its intensity in time. Here we consider the case when the Poisson source emits cells with rate λ_1 and suddenly changes its intensity to λ_2. Figure 15 presents the analysis result for these data sets. (For every process λ_1 was set to 1 and λ_2 changes as noted in the figure.)

Discussion 4 For such simple inhomogeneous Poisson processes the IDC estimate can be derived analytically. Let λ_1 and λ_2 denote the intensity parameters of the process

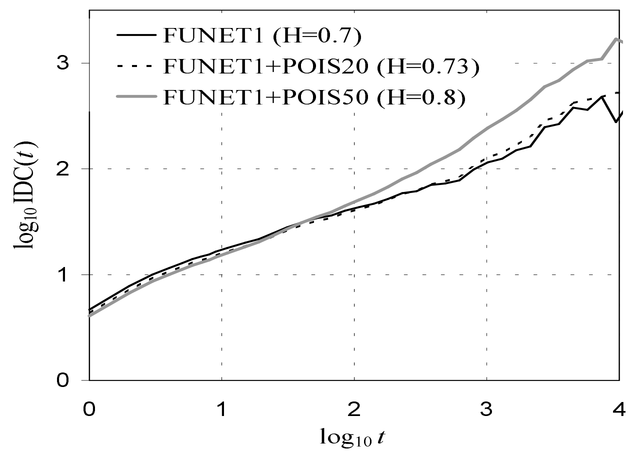


Figure 14: IDC plot for FUNET1 multiplexed with a nonstationary Poisson traffic.

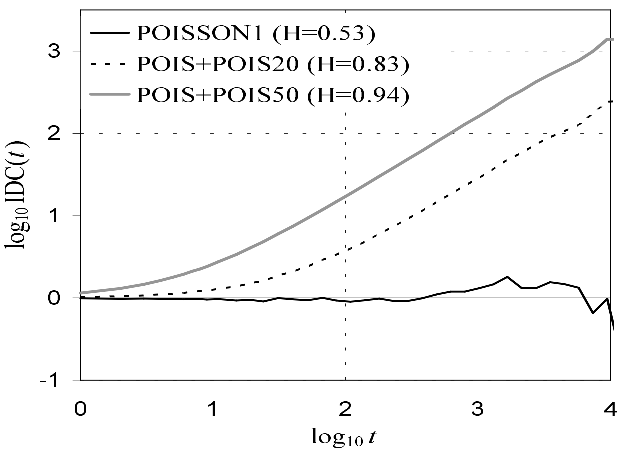


Figure 15: IDC plot for inhomogeneous Poisson processes.

for the two halves. Then, the IDC(t) value can be calculated as follows:

$$IDC(t) = 1 + \frac{(\lambda_1 - \lambda_2)^2}{2(\lambda_1 + \lambda_2)} t.$$

For the appropriate IDC plot for $\lambda_1 \neq \lambda_2$ and $t \to \infty$ we get:

$$\log IDC(t) \simeq const + \log t, \quad \text{where} \quad const = \log \frac{(\lambda_1 - \lambda_2)^2}{2(\lambda_1 + \lambda_2)}.$$

This equation gives a straight line with slope 1 as an asymptote.

The main message from this example is again that a monotonically increasing IDC does not necessarily come from the self-similar nature of the analysed data. Instead, it comes from the nonstationarity present in the sample trace. We mention that a linearly growing IDC curve over many time scales can also be created even with a simple stationary Markovian model (e.g. with an Interrupted Poisson Process). In this case the increasing IDC curve again nothing has to do with self-similarity.

Investigations have showed [90, 92] that the estimation of Hurst parameter can depend on many characteristics and require the stationarity assumption to be hold. Therefore the problem of 'deceiving self-similar tests' is a critical issue. These highlight the problem of 'how can we get the correct value for the Hurst parameter in practice?'.

Misinterpretation of results of statistical tests can lead to estaiblish wrong conclusions and we may fail to give a useful characterization for real traffic.

2.2.4 Non-Stationarity of MPEG2 Video Traffic

It has been reported that VBR traffic belongs to the class of long-range dependent processes (see, e.g. [10, 39]). This implies the weak stationarity of this traffic type [10]. While our investigations confirm that VBR traffic displays signs of long-range dependence, they strongly call into question the assumption of weak stationarity [41]. Furthermore, we will examine the character of non-stationarity in the following sections, which provides a basis for a new traffic model described in section 2.3.3.

The analysis is based on three traffic streams generated using the MPEG2 video compression algorithm [42]. Each traffic stream contains approximately 154,000 data points where each data point represents the total bit rate of one video frame (i.e. luminance and chrominance information, motion vectors and overhead information). In a video made from a movie original, both fields in each frame are identical. In our work, these two fields were averaged (thus improving SNR) and the resulting data (720×288 pixels, 25 frames/s) was coded with an MPEG2 software coder. However, in order not to overload the following discussion with side aspects of other MPEG2 coding schemes we focuss our attention on the most simple one: III traffic, where all frames are coded in intrafield mode.

For this investigation, the action movie 'Star Wars' was chosen. Among all existing genres of movies, action movies might be the most demanding ones in terms of network management, because of rapid scene changes, fast changing lighting conditions and the like.

Recently it became evident that variable bit rate video traffic displays signs of long-range dependence [10, 39], such as

- The autocorrelation ρ_k obeys a hyperbolic decay for large lags k: $\rho_k \stackrel{k\to\infty}{\longrightarrow} c_0 k^{-\beta}$
- The power spectral density $s(\omega)$ follows the law $\Gamma(\omega) \stackrel{\omega\to 0}{\longrightarrow} c_1\omega^{\beta-1}$ for small frequencies ω.
- The variance σ_n^2 of the sample mean decreases more slowly than the reciprocal of the sample size n: $\sigma_n^2 = \mathrm{Var}\,\bar{X}_n \stackrel{n\to\infty}{\longrightarrow} c_2 n^{-\beta}$, where $\bar{X}_n = \sum_{i=1}^n X_i/n$,

for some constants c_0, c_1, c_2. The constant $\beta \in [0,2]$ indicates the type of dependence: $0 \le \beta < 1$ indicates long-range dependence and $1 < \beta \le 2$ indicates short-range dependence. (The degree of persistence is expressed most often by the Hurst parameter $H = 1-\beta/2$). However, long-range dependence is defined within the framework of weak stationarity [9, 10].

Definition 2.1 *A stochastic process $X(t)$ is said to be weakly stationary if the moments up to order 2 are finite and constant over time and if its covariance*

$$\mathbf{E}\left\{(X(t_0)-\mu)(X(t_0+\Delta t)-\mu)\right\}$$

depends only on Δt, where μ is the mean.

Stationarity in conjunction with ergodicity allows one to infer statistical estimates such as mean and variance or model parameters from a single realization of data, or in our case a single time series. If this convenient assumption is violated, some measures, such as mean and variance, may become meaningless. Indeed, it has been reported that the mean of a VBR video time series converges only slowly [39], which may be caused by non-stationarity and not necessarily by long-range dependence. Inspecting the bit rate profile of *III* traffic (see Figure 16) suggests that non-stationarity is a more likely explanation of the observed long-range dependence [43].

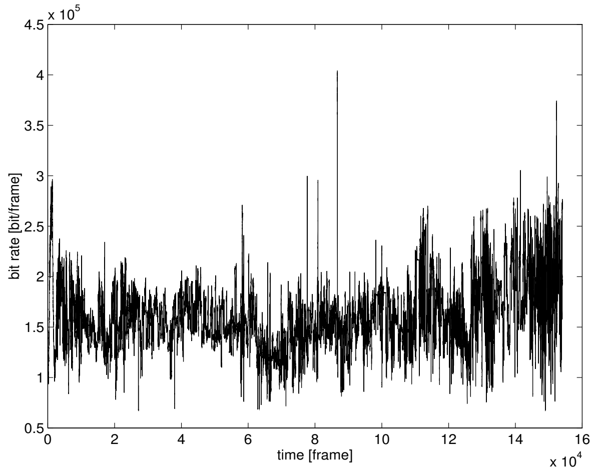


Figure 16: Bit rate profile of *III* traffic

2.2.5 Testing for Stationarity

Let $X(n)$, $n = 0, 1, 2 \ldots$, be a stochastic process with power spectral density $\Gamma(\omega)$. The periodogram

$$\hat{\Gamma}(\omega) = \frac{1}{2\pi N} \left| \sum_{n=0}^{N-1} (X(n) - \bar{X}) e^{-j\omega n} \right|^2, \tag{7}$$

where $\bar{X}$ is the sample mean, converges in distribution to $\frac{1}{2}\Gamma(\omega)\chi_2^2$ (see, e.g., [101]) for $\omega \neq 0, \pm\pi, \pm 2\pi, \ldots$. This implies that $\hat{\Gamma}(\omega)$ is for large N an unbiased estimate, but not a consistent one, since $\lim_{N\to\infty} \operatorname{Var}\hat{\Gamma}(\omega) = \Gamma^2(\omega)$. However, it holds that the periodogram ordinates $\hat{\Gamma}(\omega_1)$ and $\hat{\Gamma}(\omega_2)$ are approximately uncorrelated for two fixed frequencies ω_1 and ω_2. These properties hold as well for long-range dependent processes [9]. Applying a *spectral window* $\Lambda(\omega)$ gives a consistent estimate [101]

$$\bar{\Gamma}(\omega) = \int_{-\pi}^{+\pi} \hat{\Gamma}(\omega) \Lambda(\Theta - \omega) d\omega, \tag{8}$$

Choosing the *Bartlett-Priestley* spectral window [101] one obtains for the variance $\operatorname{Var}\bar{\Gamma}(\omega) \approx \frac{6M}{5N}\Gamma^2(\omega)$. Still the variance depends on the power spectral density itself. To overcome this functional dependence, a logarithmic *variance-stabilizing transformation* is used [55].

To first order accuracy one obtains

$$\mathbf{E}\left\{\log(\bar{\Gamma})\right\} \approx \log(\Gamma), \tag{9}$$

$$\operatorname{Var}\log(\bar{\Gamma}) \approx \frac{2\pi}{N}\int_{-\pi}^{+\pi}\Lambda^2(\Theta)d\Theta, \tag{10}$$

where $\omega \neq 0, \pm\pi, \ldots$. As a by-product the estimate $\log(\bar{\Gamma})$ is closer to normallity than the untransformed one [55]. To verify (or reject) the assumption of weak stationarity the process X is split into I segments centered at times t_i each of it has length N. For each segment i the sample power spectral density $\bar{\Gamma}_i(\omega)$ according to (8) is computed. Sampling the smoothed periodogram (8) at frequencies $\omega_j = \pi j/N$ ($j = j_0 + k\Delta j, k = 0, 1, \ldots J$) and taking the logarithm gives the two-dimensional random variable $Y_{ij} = \log(\bar{\Gamma}_i(\omega_j))$. The variate Y_{ij} is approximately normally distributed and uncorrelated if the frequencies ω_j as well as the times t_i are sufficiently wide apart [102]. Assuming approximate normallity and the Y_{ij} being uncorrelated in both dimensions implies approximate independence of Y_{ij}. Therefore, we can use the *analysis of variance* (ANOVA) technique (see, e.g., [55, 102]) to infer the underlying structure of random process Y_{ij}

$$Y_{ij} = \mu + a(t_i) + b(\omega_j) + c(t_i, \omega_j) + \eta_{ij} \tag{11}$$

where the η_{ij} are independent and identically normally distributed with zero mean and variance σ^2 defined by (10). The presence of $c(t_i, \omega_j)$ resp. $a(t_i)$ is tested using the quantities

$$S_{I+R} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - Y_{\cdot j} - Y_{i\cdot} + Y_{\cdot\cdot}\right)^2, \tag{12}$$

$$S_T = J\sum_{i=1}^{I}\left(Y_{i\cdot} - Y_{\cdot\cdot}\right)^2, \tag{13}$$

where the dot indicates the mean over the index it replaces, e.g. $Y_{\cdot j} = \sum_{i=1}^{I} Y_{ij}/I$. For a stationary process we expect $c(t_i, \omega_j)$ and $a(t_i)$ to vanish. In this case S_{I+R}/σ^2 resp. S_T/σ^2 are χ^2-distributed with $(I-1)(J-1)$ resp. $(I-1)$ degrees of freedom. The hypothesis of stationarity is rejected if one of the test statistics exceeds the upper 1% quantile of the corresponding chi-squared distribution.

Table 2 contains the results for *III* traffic, fractionally differenced white noise (FDWN) [52], AR(1) with autocorrelation $\rho_k = 0.90^k$ and the scenic model [38]. FDWN has a Hurst parameter of $H = 0.8$. AR(1) and FDWN are stationary series, whereas the scenic model is non-stationary in the mean.

It can be seen that stationarity has to be rejected on the 1% level for *III* traffic and the scenic model. FDWN and AR(1) are classified as stationary series. In fact stationarity has to be rejected as well for *IPP* and *IBBP* traffic [43]. Figures 17 and 18 demonstrate

Table 2: Test for Stationarity

	N	M	j_0	Δj	$\frac{S_{I+R}}{\sigma^2}$	$\chi^2_{\nu,1\%}$	ν_{I+R}	$\frac{S_T}{\sigma^2}$	$\chi^2_{\nu,1\%}$	ν_T
III	1024	512	200	16	6810.0	7888.0	7599	13376.8	191.3	149
SM	512	512	100	20	3694.3	5500.8	5260	7880.5	318.5	263
FDWN	1024	512	200	8	13743.0	15709.0	15300	130.8	192.4	150
AR(1), $\Phi = 0.90$	1024	512	50	16	7629.3	9009.1	8700	137.1	186.8	145

SM: scenic model FDWN: fractionally differenced white noise ($H = 0.8$)

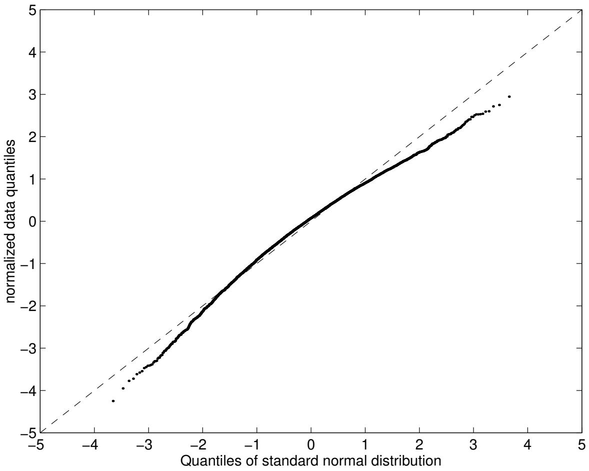


Figure 17: Q-Q plot of noise terms η_{ij}

that the noise terms are normally distributed and uncorrelated for *III* traffic. It has been argued that in the presence of long-range dependence this test might fail, because the noise is not normally distributed and not independent. As a result the test quantities were no longer chi-square distributed and the ratio $F = \frac{S_T/(I-1)}{S_{I+R}/((I-1)(J-1))}$ would not be F-distributed. The F-ratio was computed from 52 series of FDWN of length 155,000 frames and $H = 0.8$. Using the same parameters N, M, j_0 and Δj as for *III* traffic the F-ratio follows well an F-distribution. Hence, all conditions the test is based on are met.

2.2.6 The Type of Non-Stationarity of VBR Video Traffic

The test of stationarity suggests that the scenic model and *III* have the same structure $\mathbf{E}\{Y_{ij}\} = \mu + a(t_i) + b(\omega_j)$. In addition, the scenic model is non-stationary in the mean, i.e. it shows jumps in the data rate. Indeed, inspecting *III* traffic on a shorter time scale reveals the same behaviour (see Figure 19). This indicates that the convenient assumption of weak stationarity has to be given up and *long-range dependence must be seen as an artifact of non-stationarity.* This is further strongly supported by a publication by *Klemeš* [56]. The process he investigated is called a *shifting level process* (SLP).

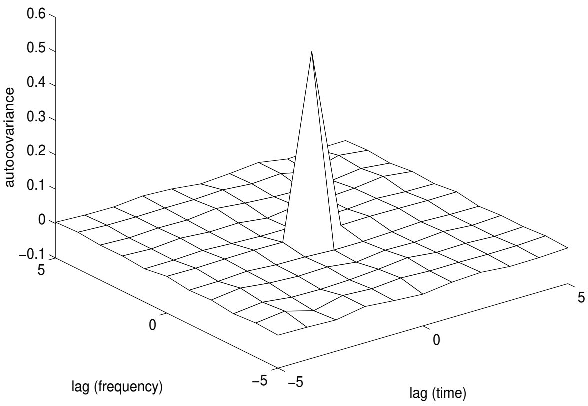


Figure 18: 2D autocorrelation of noise terms η_{ij}

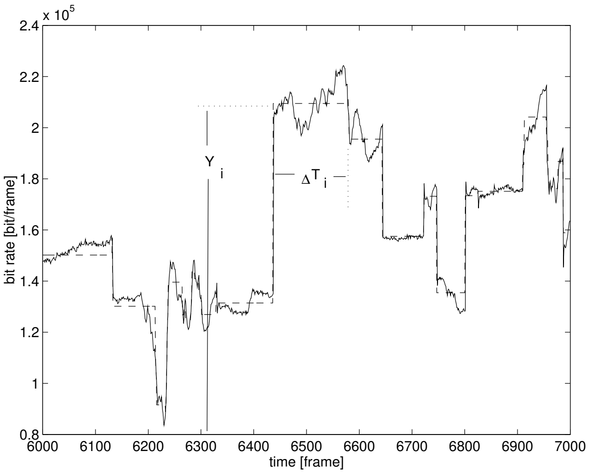


Figure 19: Section of *III* traffic

Definition 2.2 *Let Y_i be independent and identically distributed random variables with mean μ_Y and variance σ_Y^2 and possibly existing higher order raw moments $\mu'_{Y,r}$. Let $\Delta T_i := t_{i+1} - t_i$ (the epochs) be independent and identically distributed random variables with density f_t and mean μ_t. The stochastic process $X(t) = Y_i$ for $t_i \leq t < t_{i+1}$ $(i = 1, 2, \dots)$ is then called a shifting level process.*

Since these processes were introduced as an alternative explanation of the observed long-range dependence of hydrological time series, their connection to the Hurst phenomenon has been studied by several authors (see, e.g., [13]). The behaviour of moment estimators for these processes can be found in [17].

Admittedly, the concept of non-statonarity has at first glance not much appeal and seems to be a rather cumbersome one. However, for a process defined by (2.2) it can be shown that the empirical raw moments

$$M'_{X,r} = \frac{1}{T} \int_{t_1}^{t_{n+1}} X^r(t)dt = \sum_{i=1}^{n} Y_i^r \frac{\Delta T_i}{T}, \tag{14}$$

where $T := t_{n+1} - t_1$ and $r = 0, 1, \ldots,$ do exist with expectation

$$\lim_{T\to\infty} \mathbf{E}\left\{M'_{X,r}\right\} = \mu'_{Y,r}. \tag{15}$$

Even the autocovariance

$$\gamma(\tau) = \frac{1}{T} \int_{t_1}^{t_{n+1}} (X(t) - \mu_Y)(X(t+\tau) - \mu_Y)dt \tag{16}$$

exists for an infinitely long process

$$\lim_{T\to\infty} \mathbf{E}\left\{\frac{\gamma(\tau)}{\sigma_Y^2}\right\} = 1 - \frac{\tau}{\mu_t} + \frac{1}{\mu_t} \int_{\kappa}^{\tau} (\tau - t) f_t(t)dt, \tag{17}$$

for positive τ; κ is the shortest epoch length. The foregoing result is a generalization of a result given by *Mandelbrot* for discrete distributions f_t [74]. Hence, these processes are *asymptotically weakly stationary*, which provides a reasonable basis for pratical work. Assume now, that the epochs have a distribution with a Pareto shaped tail

$$f_t(t) \;=\; \begin{cases} f(t) & \text{for } 0 < \kappa \le t < t_0, \\ \dfrac{\Theta t_1^{\Theta}}{t^{\Theta+1}} & \text{for } t \ge t_0, \end{cases} \tag{18}$$

where $f(t)$ is some positive function, such that f_t is a proper density, and t_0, t_1 are some positive constants. It can be shown that the following theorem holds.

Theorem 2.1 *A shifting level process has epoch distribution (18) if and only if the corresponding autocorrelation function shows long-range dependence with Hurst parameter* $H = \frac{3-\Theta}{2}$.

Moreover, *Mandelbrot* demonstrates [74] that the power spectral density near the origin of an SLP with distribution (18) follows the same power law as stated in section 2.2.4. This demonstrates clearly that non-stationarity in the mean can cause long-range dependence. Moreover, (18) gives enough flexibility to model short term behaviour as well by specification of $f(t)$.

2.2.7 Summary

As a conclusion we emphasize that characterizing the network traffic by fractal models we have to be very careful not to mistake actual non-stationarities with stationary fractal behaviour. These effects can produce the same results to a lot of statistical tests. We should note that there are promising methods which try to distinguish between non-stationarity and long range dependence [9] or estimating the Hurst parameter in the presence of some types of non-stationarities [107].

A number of cases in practice we can talk about local stationarity only and it is important to specify the relevant time-scales of the stationer fractal behaviour. We note that in some cases beside statistical evidence of fractal behaviour the physical explanation of the traffic generation mechanisms can also support the idea of choosing fractal models.

2.3 Long-range dependent traffic models

2.3.1 Quasi-Markovian models

A discrete-time ATM traffic model which exhibits a long range dependence character is presented in this section [23]. The process results from the superposition of an infinite number of on/off sources which have an increasing mean on and off period duration. The condition under which the process has the long range dependence property is a simple function of the parameters of the on/off sources.

The traffic model The traffic model that is envisaged is defined in the framework of Markovian Arrival Processes. For completeness reasons we recall the definition of a Discrete-Time Batch Markovian Arrival Process (D-BMAP), the discrete-time version of the BMAP defined in [96] and [70]. (For more details, we refer the reader to [11]).
Consider a discrete-time Markov chain with transition matrix $\mathbf{D}$. Suppose that at time k this chain is in some state i, $1 \leq i \leq m$. At the next time instant $k+1$, there occurs a transition to another or possible the same state and a batch arrival may or may not occur. With probability $(d_0)_{i,j}$, $1 \leq i \leq m$, there is a transition to state j without an arrival, and with probability $(d_n)_{i,j}$, $1 \leq i \leq m$, $n \geq 1$, there is a transition to state j with a batch arrival of size n. We have that

$$\sum_{n=0}^{\infty} \sum_{j=1}^{m} (d_n)_{i,j} = 1.$$

Clearly the matrix $\mathbf{D}_0$ with elements $(d_0)_{i,j}$ governs transitions that correspond to no arrivals, while the matrices $\mathbf{D}_n$ with elements $(d_n)_{i,j}$, $n \geq 1$, govern transitions that correspond to arrivals of batches of size n.
The matrix $\mathbf{D} = \sum_{n=0}^{\infty} \mathbf{D}_n$ is the transition matrix of the underlying Markov chain. Let $\boldsymbol{\pi}$ be stationary probability vector of this Markov process, i.e.

$$\boldsymbol{\pi}\mathbf{D} = \boldsymbol{\pi}, \quad \boldsymbol{\pi}\mathbf{e} = 1,$$

where $\mathbf{e}$ is a column vector of 1's.
The fundamental arrival rate λ of this process is given by

$$\lambda = \boldsymbol{\pi}(\sum_{k=1}^{\infty} k\mathbf{D}_k)\mathbf{e}.$$

A D-MAP is a special case of a D-BMAP, where arrivals have a batch of size 1 (for examples we refer to [11]).
Now we define the processes which are used to obtain the long range dependent process. Consider a sequence $(X^{(i)})_{i\in N}$ of independent on/off sources with the following characteristics. Let $1 < b < a$. Assume that both the on and off period of the process $X^{(i)}$ are geometrically distributed with mean duration $(\frac{a}{b})^i$, resp. a^i. While on, the source

generates a cell in a slot with probability p, with $0 < p < 1$. Using matrix analytic notations, $X^{(i)}$ is a D-MAP with parameter matrices

$$\mathbf{D}_0^{(i)} = \begin{pmatrix} 1-(1/a)^i & (1/a)^i \\ (1-p)(b/a)^i & (1-p)\left(1-(b/a)^i\right) \end{pmatrix}$$

and

$$\mathbf{D}_1^{(i)} = \begin{pmatrix} 0 & 0 \\ p(b/a)^i & p\left(1-(b/a)^i\right) \end{pmatrix}.$$

The matrix $\mathbf{D}^{(i)} = \mathbf{D}_0^{(i)} + \mathbf{D}_1^{(i)}$ is the transition matrix of the underlying Markov chain of state transitions. The stationary distribution of $\mathbf{D}^{(i)}$ is given by

$$\boldsymbol{\pi}^{(i)} = \begin{bmatrix} \frac{b^i}{(1+b^i)} & \frac{1}{(1+b^i)} \end{bmatrix}.$$

The fundamental arrival rate $\lambda^{(i)}$ associated with $X^{(i)}$ is

$$\lambda^{(i)} = \boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e} = p/(1+b^i).$$

From the definition of $X^{(i)}$ we see that for increasing i, both the on and off periods become longer. This property of the process $X^{(i)}$ will be responsible for the long range dependence of the envisaged process.
Let us now characterize the correlation structure of the process $X^{(i)}$. From [11], we know that

$$\operatorname{Cov}\left(X_1^{(i)}, X_{1+k}^{(i)}\right) = \boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\left(\left(\mathbf{D}^{(i)}\right)^{k-1} - \mathbf{e}\boldsymbol{\pi}^{(i)}\right)\mathbf{D}_1^{(i)}\mathbf{e}.$$

Hence one can easily verify that

$$\operatorname{Cov}\left(X_1^{(i)}, X_{1+k}^{(i)}\right) = \left(1 - \left(\frac{1}{a}\right)^i - \left(\frac{b}{a}\right)^i\right)^k \frac{p^2 b^i}{(1+b^i)^2}.$$

In view of [11], p. 8, we know that a finite superposition $Y^{(M)} = \sum_{i=1}^{M} X^{(i)}$ of D-MAPs is a D-BMAP determined by the matrices

$$\begin{aligned}
\mathbf{C}_0^{(M)} &= \mathbf{D}_0^{(M)} \otimes \mathbf{D}_0^{(M-1)} \otimes \cdots \otimes \mathbf{D}_0^{(1)}, \\
&\vdots \\
\mathbf{C}_i^{(M)} &= \sum_{k_M+\ldots+k_1=i} \bigotimes_{j=M}^{1} \mathbf{D}_{k_j}^{(j)}, \\
&\vdots \\
\mathbf{C}_M^{(M)} &= \mathbf{D}_1^{(M)} \otimes \mathbf{D}_1^{(M-1)} \otimes \cdots \otimes \mathbf{D}_1^{(1)}.
\end{aligned}$$

The superposition $Y^{(\infty)}$ is not a D-BMAP any longer, but since the $X^{(i)}$ are independent, the expressions for the fundamental arrival rate $\lambda^{(\infty)}$ and the covariance structure are given by:

$$\lambda^{(\infty)} = \sum_{i=1}^{\infty} \frac{p}{1+b^i}$$

and

$$\operatorname{Cov}\left(Y_1^{(\infty)}, Y_{1+k}^{(\infty)}\right) = \sum_{i=1}^{\infty} \left(1 - \left(\frac{1}{a}\right)^i - \left(\frac{b}{a}\right)^i\right)^k \frac{p^2 b^i}{(1+b^i)^2}. \tag{19}$$

Properties of the Process $Y^{(\infty)}$ In this section the influence of the parameters a and b on the correlation structure of the arrival process $Y^{(\infty)}$ is examined.

Property 2.1 *The arrival process $Y^{(\infty)}$ is long range dependent if and only if $b^2 \leqslant a$.*

Proof. Following Definition 13.4.1 in [105, page 326] we have long range dependence if and only if the series is

$$\sum_{k=1}^{\infty} \operatorname{Cov}\left(Y_1^{(\infty)}, Y_{1+k}^{(\infty)}\right) \tag{20}$$

diverges. Using (19) and changing the order of summation we deduce that the series (20) diverges if and only if

$$\sum_{i=1}^{\infty} \frac{a^i b^i - b^i - b^{2i}}{\left(1+b^i\right)^3} = \infty. \tag{21}$$

This series is similar to a geometric one and hence it diverges if and only if $b^2 \leqslant a$. ■

Property 2.2 *There exist $0 < C_1 < C_2 < \infty$ such that*

$$C_1 k^{-\beta} < \operatorname{Cov}\left(Y_1^{(\infty)}, Y_{1+k}^{(\infty)}\right) < C_2 k^{-\beta} \tag{22}$$

with

$$\beta = \frac{\log b}{\log a - \log b}. \tag{23}$$

Proof. See [23]. ■
Now we state the main result of this section, namely an explicit expression for the Hurst parameter of the process $Y^{(\infty)}$.

Property 2.3 *The Hurst parameter H of the discrete-time arrival process $Y^{(\infty)}$ is given by*

$$H = \frac{1}{2}\left(2 - \frac{\log b}{\log a - \log b}\right). \tag{24}$$

Proof. Based on (22), we see that $\mathrm{Cov}\left(Y_1^{(\infty)}, Y_{1+k}^{(\infty)}\right)$ decreases as $k^{-\beta}$, with $\beta = \dfrac{\log b}{\log a - \log b}$. Hence, from [105, page 327], we immediately obtain Equation 24. ■

Clearly, if $b^2 \leqslant a$, then the Hurst parameter satisfies $\frac{1}{2} \leq H < 1$, a criterion for long range dependence of the process $Y^{(\infty)}$.
We will illustrate the above properties through numerical examples.

The Index of Dispersion for Counts of the Traffic Model In this section we investigate the correlation structure of the process $Y^{(\infty)}$ by means of the Index of Dispersion for Counts (IDC).
Denote N_k the number of arrivals in an interval of length k. The *Index of Dispersion for Counts* (IDC) at time k is defined to be the variance of the number of arrivals in an interval of length k divided by the the mean number of arrivals in this interval, i.e.

$$I(k) = \frac{\mathrm{Var}(N_k)}{\mathrm{E}(N_k)}.$$

It is well known that for a renewal process $I(k) = c_1^2$, for all $k \geq 1$, where c_1^2 is the squared coefficient of variation of the number of arrivals in a slot. In particular for a Bernoulli process, $I(k) = 1$, for all $k \geq 1$.
Denote $I^{(i)}(k)$ the IDC of the process $X^{(i)}$ with $\lim_{k\to\infty} I^{(i)}(k) = J^{(i)}$ and $I^{(\infty)}(k)$ the IDC of the process $Y^{(\infty)}$, with $\lim_{k\to\infty} I^{(\infty)}(k) = J^{(\infty)}$.
From [12], we know that

$$J^{(i)} = \frac{\boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e} - 3[\boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e}]^2 + 2\boldsymbol{\pi}\mathbf{D}_1^{(i)}\mathbf{Z}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e}}{\boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e}}, \tag{25}$$

with $\mathbf{Z}^{(i)}$ the fundamental matrix of the Markov chain $\mathbf{D}^{(i)} = \mathbf{D}_0^{(i)} + \mathbf{D}_1^{(i)}$, given by

$$\mathbf{Z}^{(i)} = [\mathbf{I} - (\mathbf{D}^{(i)} - \mathbf{e}\boldsymbol{\pi}^{(i)})]^{-1}.$$

From the expressions for $\mathbf{D}_0^{(i)}$ and $\mathbf{D}_1^{(i)}$ given in Section 2, it is easy to show that

$$\mathbf{Z}^{(i)} = \frac{1}{(1+b^i)^2}\begin{pmatrix} a^i + b^i(1+b^i) & 1 - a^i + b^i \\ b^i(1 - a^i + b^i) & 1 + b^i + a^i b^i \end{pmatrix}.$$

Hence,

$$\boldsymbol{\pi}^{(i)}\mathbf{D}_1^{(i)}\mathbf{Z}^{(i)}\mathbf{D}_1^{(i)}\mathbf{e} = \frac{p^2}{(1+b^i)^3}\,[1 + b^i(a^i - b^i)].$$

Using this expression in (25), we obtain that

$$J^{(i)} = 1 - 3\frac{p}{1+b^i} + 2\frac{p}{(1+b^i)^2}[1 + b^i(a^i - b^i)]. \tag{26}$$

Now we compute $J^{(\infty)}$, i.e. the limit of the IDC of the process $Y^{(\infty)}$.
Since $Y^{(\infty)} = \sum_{i=1}^{\infty} X^{(i)}$, we have that $I^{(\infty)}(k) = \sum_{i=1}^{\infty} I^{(i)}(k)$. Hence,

$$I^{(\infty)}(k) = \frac{\sum_{i=1}^{\infty} \mathrm{cov}(X_1^{(i)}, X_1^{(i)}) + \sum_{i=1}^{\infty} 2 \sum_{j=1}^{k-1} \frac{k-j}{k} \mathrm{cov}(X_1^{(i)}, X_{1+j}^{(i)})}{\sum_{i=1}^{\infty} \mathrm{E}[X_1^{(i)}]}.$$

Taking the limit for $k \longrightarrow \infty$, we obtain

$$J^{(\infty)} = \frac{\lambda^{(\infty)} - 3\sum_{i=1}^{\infty}(\lambda^{(i)})^2 + 2p^2 \sum_{i=1}^{\infty} \dfrac{1 + b^i(a^i - b^i)}{(1 + b^i)^3}}{\lambda^{(\infty)}}. \tag{27}$$

From equation (27) it follows that the limit of the IDC of the process $Y^{(\infty)}$ is infinite if $b^2 \leqslant a$, which is exactly the condition under which the process has the long range dependence property (see Property 2.1). This is in agreement with the criterion that a process is long range dependent if its IDC is diverging.

Queueing Behaviour We consider a queue of the G/D/1-type which has the arrival process $Y^{(\infty)}$ as input. It turns out that the mean queue length is ∞. This result is obtained by studying the sequence of queues with arrival processes $\sum_{i=1}^{M} X^{(i)}$. These queues are of the D-BMAP/D/1-type. Consider the D-BMAP/D/1-queue with arrival process $Y^{(M)} = \sum_{i=1}^{M} X^{(i)}$. From now on we will drop the index M to keep the notation simple. The stationary queue distribution $\mathbf{x}$ of the D-BMAP/D/1 queue satisfies the following steady state equations

$$\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \dots) = (\mathbf{x}_0, \mathbf{x}_1, \dots) \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \dots \\ \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \dots \\ \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{28}$$

together with

$$\mathbf{x}\,\mathbf{e} = 1. \tag{29}$$

This leads to

$$\mathbf{X}(z)\,(z\mathbf{I} - \mathbf{D}(z)) = (z-1)\mathbf{x}_0\mathbf{D}(z) \tag{30}$$

with generating functions $\mathbf{X}(z) = \sum_{n=0}^{\infty} \mathbf{x}_n z^n$ and $\mathbf{D}(z) = \sum_{n=0}^{\infty} \mathbf{D}_n z^n$. It is important to notice that $\mathbf{D}(z) = \mathbf{D}^{(M)}(z) \cdots \mathbf{D}^{(1)}(z)$. The mean queue length is given by the expression $\mathbf{X}'\mathbf{e}$ with $\mathbf{X}' = \frac{d}{dz}\mathbf{X}(z)|_{z=1}$. The computations made in [70] result in the following expression for the mean queue length

$$\mathbf{X}'\mathbf{e} = \frac{(\boldsymbol{\pi}\mathbf{D}''\mathbf{e} + 2\mathbf{x}_0\mathbf{D}'\mathbf{e} - 2\rho + (2\mathbf{x}_0\mathbf{D} + 2\boldsymbol{\pi}\mathbf{D}')(\mathbf{I} - \mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{D}'\mathbf{e})}{2(1-\rho)}, \tag{31}$$

with $\mathbf{D} = \mathbf{D}(1)$, $\mathbf{D}' = \frac{d}{dz}\mathbf{D}(z)|_{z=1}$, $\mathbf{D}'' = \frac{d^2}{dz^2}\mathbf{D}(z)|_{z=1}$ and $\boldsymbol{\pi} = \boldsymbol{\pi}^{(M)} \otimes \cdots \otimes \boldsymbol{\pi}^{(1)}$ and $\rho = \sum_{i=1}^{M} \lambda^{(i)}$. We show that the right hand side of (31) is diverging for $M \longrightarrow \infty$. Since $\boldsymbol{\pi}\mathbf{D}''\mathbf{e} > 0$ and $2\mathbf{x}_0\mathbf{D}'\mathbf{e} > 0$, and as the load ρ is bounded by some number, independent from M, it is sufficient to investigate the behaviour of the factor

$$2\mathbf{x}_0\mathbf{D}(\mathbf{I} - \mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{D}'\mathbf{e} + 2\boldsymbol{\pi}\mathbf{D}'(\mathbf{I} - \mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{D}'\mathbf{e} \tag{32}$$

for $M \longrightarrow \infty$. First we notice that

$$(\mathbf{I} - (\mathbf{D} - \mathbf{e}\boldsymbol{\pi}))^{-1} = \mathbf{I} + \sum_{k=1}^{\infty} \left(\mathbf{D}^k - \mathbf{e}\boldsymbol{\pi}\right). \tag{33}$$

Furthermore, since

$$\mathrm{Cov}\,(X_1, X_{1+k}) = \boldsymbol{\pi}\mathbf{D}' \left(\mathbf{D}^{k-1} - \mathbf{e}\boldsymbol{\pi}\right) \mathbf{D}'\mathbf{e} \tag{34}$$

we have

$$\boldsymbol{\pi}\mathbf{D}' \left(\sum_{k=1}^{\infty} \left(\mathbf{D}^k - \mathbf{e}\boldsymbol{\pi}\right) \right) \mathbf{D}'\mathbf{e} = \sum_{k=2}^{\infty} \sum_{i=1}^{M} \left(1 - \left(\frac{1}{a}\right)^i - \left(\frac{b}{a}\right)^i \right)^k \frac{b^i}{(1+b^i)^2} \tag{35}$$

The behaviour of $\mathbf{x}_0\mathbf{D}\left(\sum_{k=1}^{\infty}\left(\mathbf{D}^k - \mathbf{e}\boldsymbol{\pi}\right)\right)\mathbf{D}'\mathbf{e}$ is a bit more elaborated. From [70], we know that $\mathbf{x}_0 = (1-\rho)\mathbf{g}$, with $\mathbf{g}$ the steady state vector of the matrix $\mathbf{G}$, describing the first passage times from one level to another. Hence,

$$\mathbf{x}_0 = (1-\rho)\mathbf{u}^{(M)} \otimes \mathbf{u}^{(M-1)} \otimes \cdots \otimes \mathbf{u}^{(1)}$$

with $\mathbf{u}^{(i)}$ the first row of the matrix $\mathbf{D}^{(i)}$. Using the elementary properties of the Kronecker product $\otimes$, one obtains

$$\mathbf{x}_0\mathbf{D} \left(\sum_{k=1}^{\infty} \left(\mathbf{D}^k - \mathbf{e}\boldsymbol{\pi}\right) \right) \mathbf{D}'\mathbf{e} = -(1-\rho) \sum_{k=1}^{\infty} \sum_{i=1}^{M} \left(1 - \left(\frac{1}{a}\right)^i - \left(\frac{b}{a}\right)^i \right)^k \frac{1}{1+b^i}. \tag{36}$$

Hence, the mean queue lenght is diverging iff

$$\lim_{M\to\infty} \sum_{i=1}^{M} \sum_{k=0}^{\infty} \left(1 - \left(\frac{1}{a}\right)^i - \left(\frac{b}{a}\right)^i \right)^k \left(\frac{b^i}{(1+b^i)^2} - (1-\rho)\frac{1}{1+b^i} \right) = \infty. \tag{37}$$

One can check this is the case iff $b^2 \leqslant a$, in other words, iff the arrival process $Y^{(\infty)}$ is long range dependent. This result is in accordance with the one obtained in [67].

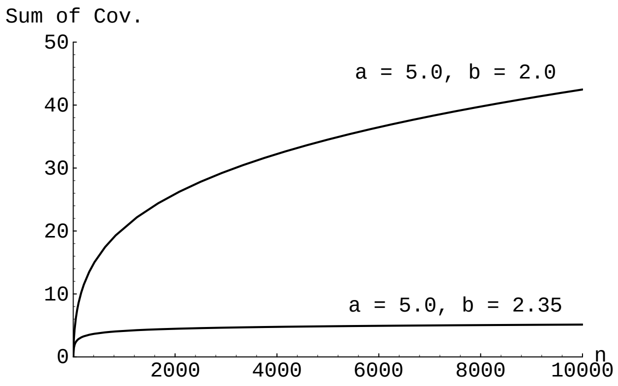


Figure 20: Influence of a and b on sum of covariances.

Numerical Examples

Example 1 In this example we illustrate Property 2.1. Consider two superpositions of on/off sources, the first with parameters $a_1 = 5$ and $b_1 = 2$ and the second with parameters $a_2 = 5$ and $b_2 = 2.35$. Application of Property 2.1 immediately shows that contrary to the second superposition, the first superposition is long range dependent (as $b_1^2 \leqslant a_1$). This is illustrated in Figure 20, where the sum of covariances of the first superposition clearly does not converge, while the second superposition does.

Example 2 In this example we consider the processes $Y^{(M)} = \sum_{i=1}^{M} X^{(i)}$. We illustrate the influence of the value M on the behaviour of the sum of covariances $\sum_{k=1}^{n} \mathrm{Cov}(Y_1^{(M)}, Y_{k+1}^{(M)})$ of increasing n.
Let $M = 6, 9, 14$. In Figure 21, we see that for higher values of M, the convergence of $\sum_{k=1}^{n} \mathrm{Cov}(Y_1^{(M)}, Y_{k+1}^{(M)})$ is slower than for smaller values. This result is in accordance with Property 2.2, which states that the sum is divergent for $M = \infty$.

Discussion In this section we have introduced a discrete-time traffic model resulting from the superposition of a sequence of on/off sources with increasing on and off period duration. Under a simple condition, the traffic model exhibits a long range dependence character. Moreover the Hurst parameter can be computed explicitly. Queueing problems in which this process is involved can be easily handled by considering a matrix-analytic approach. The correlation structure of the process is investigated by means of the IDC and its limit. Here again closed form formulas are obtained. The proposed

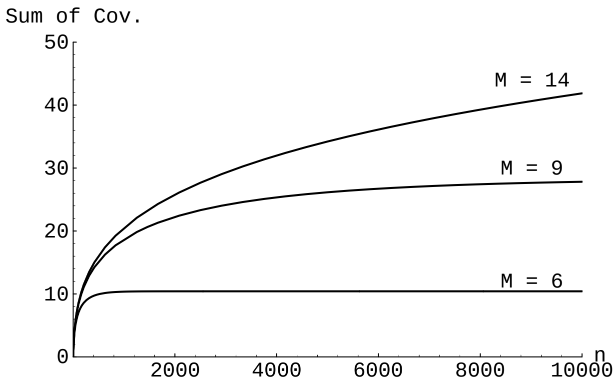


Figure 21: Influence of M on sum of covariances.

process will be used in future research to investigate the influence of long range dependent traffic on delays and loss probabilities in queueing systems when merging with Markovian traffic. Furthermore we will investigate how to choose the parameters when matching this process with data obtained from measurements.

2.3.2 Superposed heavy-tailed ON/OFF models

In this section a discrete-time on-off source model is considered. We focus mainly on an infinite-server queue model that arises when considering the superposition of a large (i.e., infinite) number of such sources. Various numerical examples are presented to illustrate the distinction between short- and long-range-dependent traffic [60, 61].

The source model On-off sources alternate between two states: the on-state (one cell generated per slot) and the off-state (no cells generated). The durations of the on- or off-periods - generically denoted by τ_A or τ_B - are iid random variables (rv's) characterized by the probability density functions (pdfs) $a(n) = \Pr[\tau_A = n]$ and $b(n) = \Pr[\tau_B = n]$ $(n = 1, 2, \dots)$ or the associated probability generating functions (pgfs)

$$A(z) = \mathrm{E}[z^{\tau_A}] = \sum_{n=1}^{+\infty} a(n) z^n \quad \text{and} \quad B(z) = \mathrm{E}[z^{\tau_B}] = \sum_{n=1}^{+\infty} b(n) z^n$$

respectively. Durations of on- and off-periods are mutually independent, their mean values equal $\mathrm{E}[\tau_A] = A'(1)$ and $\mathrm{E}[\tau_B] = B'(1)$. Variances are given by $\sigma_A^2 = \mathrm{Var}[\tau_A] = A''(1) + A'(1) - A'(1)^2$ and $\sigma_B^2 = \mathrm{Var}[\tau_B] = B''(1) + B'(1) - B'(1)^2$.

Traffic characteristics The number of cells generated by a single source during slot k, either 0 or 1, will be denoted by q_k. The average of q_k can be expressed as

$$\lambda = \mathrm{E}[q_k] = \frac{\mathrm{E}[\tau_A]}{\mathrm{E}[\tau_A] + \mathrm{E}[\tau_B]}$$

and its variance as $\sigma^2 = \mathrm{Var}[q_k] = \lambda(1-\lambda)$. An important second-order traffic characteristic is the so-called power spectral density (psd) $S(f)$, which is the Fourier-transform of the autocovariance function $C(m) = \mathrm{E}[(q_0 - \lambda)(q_m - \lambda)]$. Here, it is given by

$$S(f) = \sigma^2 \left(1 + Q(e^{j2\pi f}) + Q(e^{-j2\pi f})\right)$$

with

$$\sigma^2 Q(z) = \sum_{m=1}^{+\infty} C(m) z^m = \sigma^2 z \frac{P(z) - 1}{z - 1}$$

and

$$P(z) = \frac{A(z) - 1}{A'(1)(z-1)} \cdot \frac{B(z) - 1}{B'(1)(z-1)} \cdot \frac{[A'(1) + B'(1)](z-1)}{A(z)B(z) - 1}$$

By definition [62], long-range dependence is present when

$$\sum_{m=-\infty}^{+\infty} C(m) = S(0) = \sigma^2 \left\{ \frac{\sigma_A^2}{\mathrm{E}[\tau_A]} + \frac{\sigma_B^2}{\mathrm{E}[\tau_B]} - \frac{\sigma_A^2 + \sigma_B^2}{\mathrm{E}[\tau_A] + \mathrm{E}[\tau_B]} \right\} = +\infty$$

This will be the case when, for instance, $\sigma_A^2 = \mathrm{Var}[\tau_A]$ is infinite, given $\mathrm{E}[\tau_A]$ is finite. The main characteristic of long-rang-dependent traffic is its strong correlation structure, which is exactly what the above formula expresses.

Sample distributions Throughout this document, three different distributions for the on-periods will be used for illustrative purposes, as summarized in Table 3: a light-tailed

	variance	tail behavior
A	$< +\infty$	$\sim z_0^{-n}$
B	$= +\infty$	$\sim n^{-2.5}$
C	$< +\infty$	$\sim n^{-3.5}$

Table 3: Three different distributions

geometric distribution A, a heavy-tailed distribution B with infinite variance and a third distribution C, also heavy-tailed but with finite variance. Including distribution C will allow us to distinguish between 'long-range dependent' features and features originating from a 'heavy tail'. Mean values were set (arbitrarily) to $\mathrm{E}[\tau_A] = 100.0$ slots.

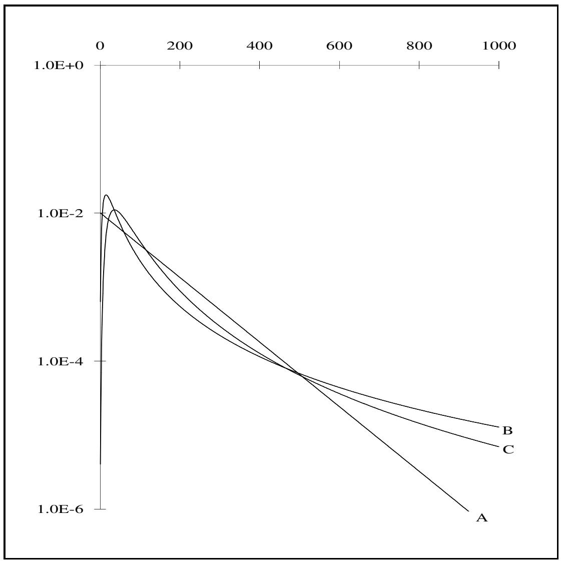


Figure 22: Tail behavior, $\log \Pr[\tau_A = n]$ versus n, for various types of distributions.

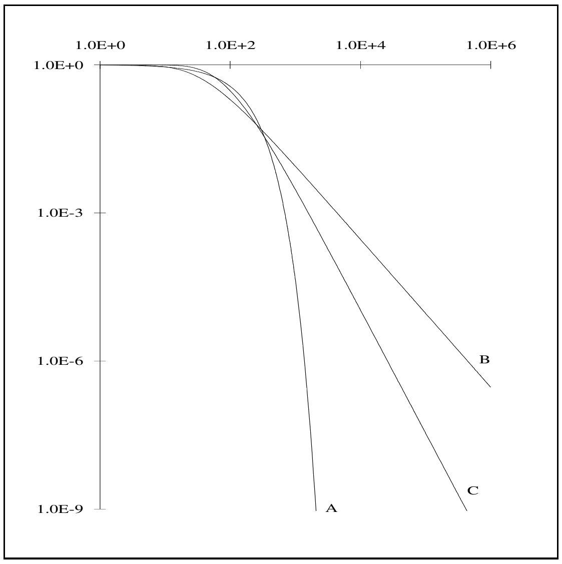


Figure 23: Tail behavior, $\log \Pr[\tau_A > n]$ versus $\log n$, for various types of distributions.

Both distribution B and C are based on the hypergeometric function

$$F(\alpha, \beta; \gamma; z) = \sum_{n=0}^{+\infty} \frac{\Gamma(\alpha+n)\Gamma(\beta+n)\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma+n)n!} z^n$$

More precisely, we used a pgf of the form

$$A(\alpha, \beta; \gamma; z) = z\frac{F(\alpha, \beta; \gamma; z)}{F(\alpha, \beta; \gamma; 1)}$$

The associated distribution has a heavy tail, in the sense that

$$\Pr[\tau_A = n] \approx \mathrm{Const} \cdot n^{-(\gamma-\alpha-\beta+1)}$$

(This particular choice seems promising, since the pgf used is based on a well-studied function for which numerical procedures are available [118] and three (real-valued) parameters are involved, which can easily be fitted to yield e.g. a given mean and tail decay. Note that the pgf involved has a branch point at $z = 1$.)

In Figure 22, where the sample distributions were plotted, the slow decay of the tails of distributions B and C clearly shows. In Figure 23, a log-log plot of the complementary cummulative distribution, this is even more apparent.

The psds of sources with on-periods as above, are shown in Figure 24. A geometrically distributed off-period was assumed for all cases, with mean 25.0 slots, yielding a traffic intensity of 0.8 Erlang. It is known [62] that, for long-range dependent sources, $S(f) \sim f^{-(1-\nu)}$ or $\log S(f) \sim -(1-\nu)\log f$, when $f \to 0$, while for short-range dependent sources, $\log S(f) \sim \log S(0)$. Both types of behavior are clearly distinguishable in

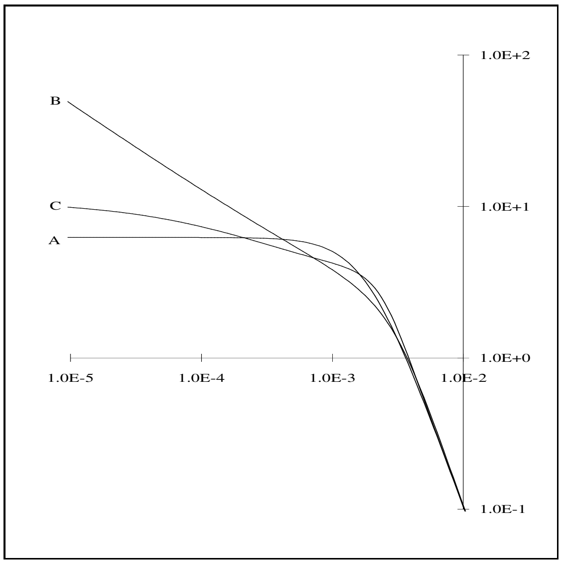


Figure 24: psd, $\log S(f)$ versus $\log f$, for various types of sources.

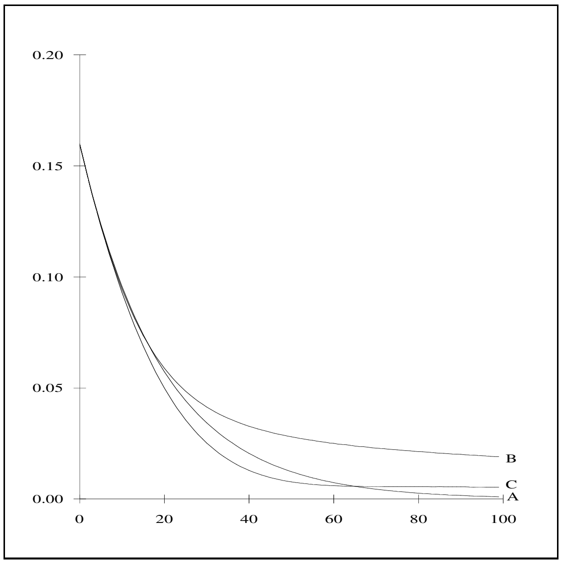


Figure 25: Autocovariance function, $C(m)$ versus m, for various types of sources.

Figure 24. Note that while the tail of distribution C is also hyperbolic and thus 'heavy', it decays too fast to yield long-range dependence in the strict sense. Corresponding autocovariance functions, obtained by numerical transform inversion, are given in Figure 25.

Superposition

N sources Assume N iid sources generate the aggregated traffic stream p_k with mean total arrival rate $\lambda_T = \mathrm{E}[p_k] = N\lambda$. It is easily shown that the psd of the aggregated process is given by $S_{(N)}(f) = NS(f)$, with $S(f)$ the psd of a single source. Figures 26 and 27 show $S_{(N)}(f)$ for a superposition of $1, 2, 5$ and an infinite number of sources. (The latter case is treated in more detail below.) Figure 26 is for a short-range dependent source of type A, Figure 27 for a long-range dependent source of type B. The total arrival rate was kept constant at 0.8 Erlang by varying the mean duration of the (geometrically distributed) off-periods.

The distinction between short- and long-range dependence is also clearly visible in the sample traces presented in Figure 28. The figure was obtained by aggregating the traffic over various timescales (1,10,100,... 10^6 slots respectively). The traffic was generated by a superposition of 5 sources, again with total traffic intensity 0.8 Erlang. In long-range-dependent traffic of type B, large fluctuations occur over large time-scales, while in short-range dependent traffic of types A and C, fluctuations die out quickly as the time scale increases.

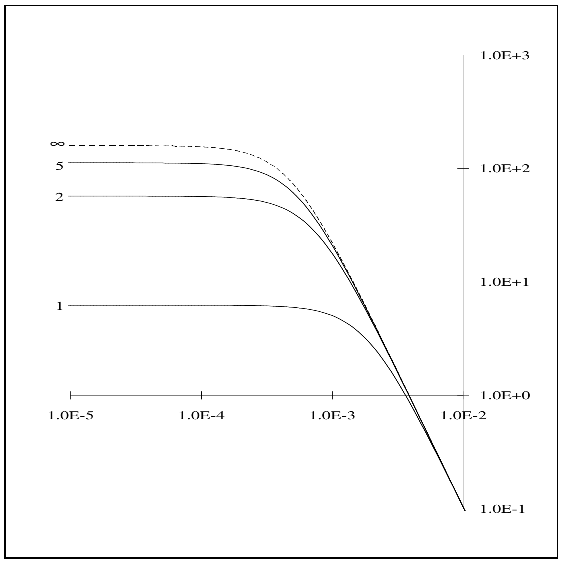


Figure 26: psd, $\log S_{(N)}(f)$ versus $\log f$, for a superposition of sources of type A.

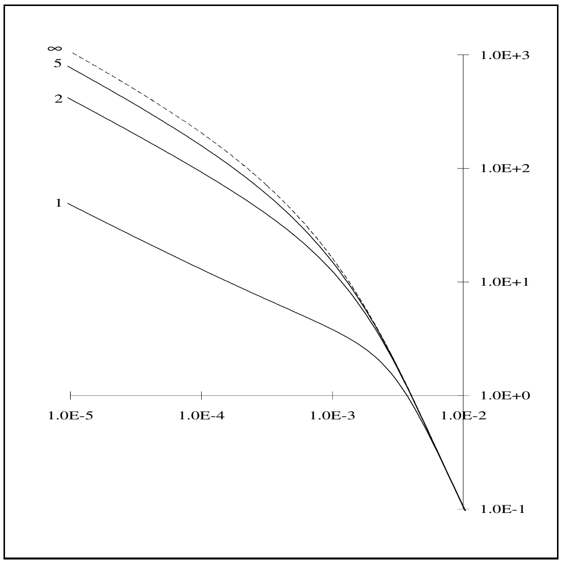


Figure 27: psd, $\log S_{(N)}(f)$ versus $\log f$, for a superposition of sources of type B.

The $N \to +\infty$ case An interesting case - from a mathematical point of view - is that whereby the number of sources grows infinitely. As illustrated by Figures 26 and 27, traffic characteristics quickly approach their limiting values as the number of sources increases. For the psd, we find by taking a limit

$$S_{(\infty)}(f) = \lim_{N \to +\infty} S_{(N)}(f) = \lambda_T \left(1 + Q_{(\infty)}(e^{j2\pi f}) + Q_{(\infty)}(e^{-j2\pi f})\right)$$

whereby

$$Q_{(\infty)}(z) = z \frac{A^*(z) - 1}{z - 1}$$

The pgf $A^*(z)$ is that of the residual duration of an on-period and is given by

$$A^*(z) = \frac{1}{\mathrm{E}[\tau_A]} \sum_{n=0}^{+\infty} \Pr[\tau_A > n] z^n$$

Observe that the number of arrivals in a slot is now equivalent with the number of customers in a discrete-time GI-G-∞ queue, the equivalent of the continuous time $M/G/\infty$ queue. One can show that the numbers of newly arriving 'customers' in each slot, i.e., the number of sources becoming active, are iid rv's with a Poisson distribution with mean $\lambda^* = \lambda_T / \mathrm{E}[\tau_A]$ and pgf $\exp\{\lambda^*(z-1)\}$. The service times of the customers are also iid rv's with pgf $A(z)$, i.e., the pgf of the on-time distribution. By analyzing this equivalent queue model on a slot-to-slot basis, it is quite straightforward to derive e.g. that

$$C(m) = \lambda^* \sum_{k=m}^{+\infty} \Pr[\tau_A > k]$$

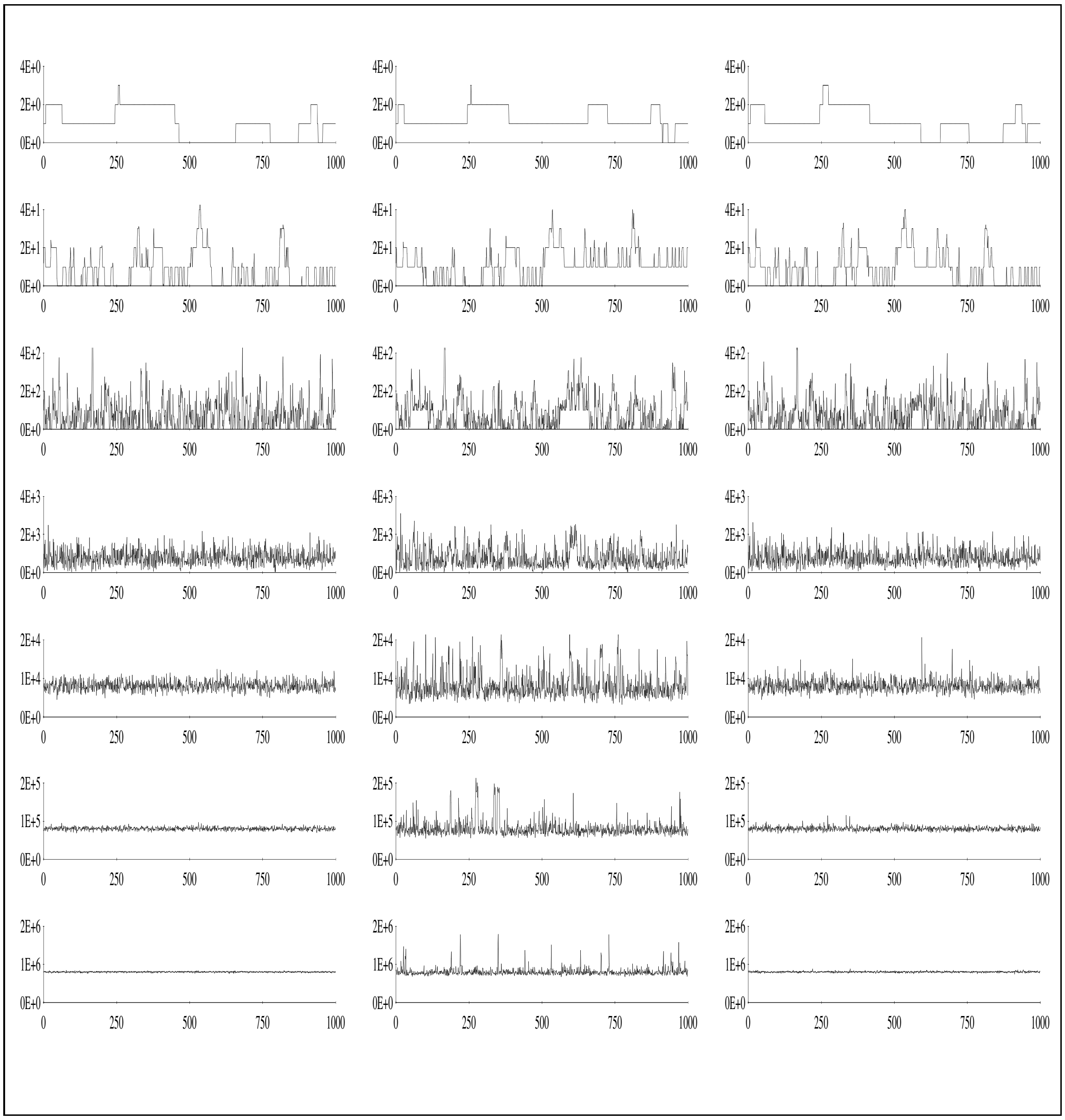


Figure 28: Aggregated traffic traces for source types A, B and C (left to right).

which is in full agreement with the expression for $Q_{(\infty)}(z)$ derived above. It illustrates once more that light-tailed on-periods lead to short-range dependence, since

$$\Pr[\tau_A = m] \sim z_0^{-m} \Rightarrow \Pr[\tau_A > m] \sim z_0^{-m} \Rightarrow C(m) \sim z_0^{-m} \Rightarrow \sum_{m=-\infty}^{+\infty} C(m) < +\infty$$

On the other hand, for heavy-tailed on-periods one has

$$\Pr[\tau_A = m] \sim m^{-q} \Rightarrow \Pr[\tau_A > m] \sim m^{-(q-1)} \Rightarrow C(m) \sim m^{-(q-2)} \Rightarrow \sum_{m=-\infty}^{+\infty} C(m) = +\infty$$

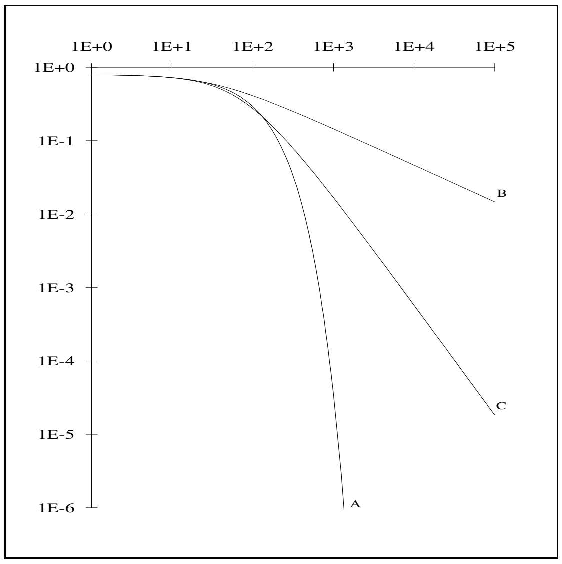


Figure 29: Autocovariance, $\log C(m)$ versus $\log m$, for the *GI-G-*∞ model and various types of on-time distributions.

when $2 < q \leq 3$ (the lower bound being required for $\mathrm{E}[\tau_A]$ to be finite). Note again, however, that also for $q > 3$, as for traffic of type C, the autocovariance function may decay quite slowly, i.e., correlation may extend over long time periods, notwithstanding it does not lead to long-range dependence in the strict sense. Figure 29 shows a log-log plot of the autocovariance function for the three different sample on-time distributions and $\lambda_T = 0.8$.

One can also derive an expression for the pgf of the total number of cells generated during m consecutive slots, namely

$$\begin{aligned} \mathrm{E}[z^{p_1+\ldots+p_m}] &= \exp\left\{\lambda^* \sum_{k=m}^{+\infty} \Pr[\tau_A > k](z^m - 1) + \lambda^* \sum_{k=0}^{m-1} \Pr[\tau_A > k](z^k - 1)\right. \\ &\qquad \left. + \lambda^* \sum_{k=0}^{m-1} \Pr[\tau_A > k](m-k) z^k (z-1)\right\} \end{aligned}$$

The first two sums in the RHS represent the contribution of 'old' sources, i.e., sources that were already active prior to slot 1. The last sum represents the contribution of sources that started generating cells during slot 1 or later. Taking derivatives and performing some algebra, one finds

$$\mathrm{Var}[p_1 + \ldots + p_m] = m^2 \lambda_T - \lambda^* \sum_{k=0}^{m-1} \Pr[\tau_A > k](m-k)(m-k-1)$$

which is in agreement with results obtained through a limiting procedure (omitted here). From this, one can easily calculate (numerically) e.g. the index-of-dispersion for counts.

For $m = 1$, one obtains

$$\mathrm{E}[z^{p_1}] = \exp\{\lambda^*(z-1)\mathrm{E}[\tau_A]\} = \exp\{\lambda_T(z-1)\}$$

The distribution of the number of active sources or, equivalently, the total number of cells generated in a random slot, is thus Poisson and function of the load λ_T only. This marginal distribution is rather smooth and independent of the exact form of the distribution of the on-periods. The latter does, however, strongly affect the correlation structure of the process.

An appealing property of the GI-G-∞ arrival process is that the aggregation of two or more such processes is again of that type. This is a consequence of the fact that the arrival process of 'new sources' is Poisson. The parameters of the aggregated GI-G-∞ arrival process are given by

$$\lambda^* = \lambda_1^* + \cdots + \lambda_N^*$$

and

$$A(z) = \frac{\lambda_1^* A_1(z) + \cdots + \lambda_N^* A_N(z)}{\lambda_1^* + \cdots + \lambda_N^*}$$

From this, it is easily seen that the tail of the aggregated on-period distribution will be dominated by the heaviest tale of the constituent distributions. In other words, long-range dependent sources will 'dominate' over short-range dependent ones.

Queueing In [60], two promising approaches to analyze the queueing behavior of traffic of the type described above were briefly discussed: the Beneš approach and a slot-to-slot approach. In this section, we present some further results on this matter, but have to refer to future work for conclusive results.

Simulation results, shown in Figures 30 and 31, for a GI-G-∞ arrival process (with intensity 0.8 Erlang) of type B and C respectively, give an indication of the magnitude of the queues - denoted by the variable u - that can build up. For instance, from Figure 30, we learn that for the long-range dependent case, the queue exceeds the order of 10^5 cells during 10% of the time. For the other case, the magnitude of the queue is about a hundred times smaller, but still very large. Although the simulations are too crude to draw detailed conclusions, the figures already point towards a hyperbolic decay of the queue contents (a straight line in a log-log plot).

The Beneš approach The Beneš approach yielded the following expression for the complementary cummulative distribution of the system contents.

$$Pr[u_{k+1} > m] = \sum_{l=0}^{+\infty} Pr[p_k + p_{k-1} + \ldots + p_{k-l} > m + l | u_{k-l} = 0] Pr[u_{k-l} = 0]$$

Intricate questions are what the link is between this general result and the general observation made in e.g. [46, 66] concerning the impact of the psd of the traffic at low

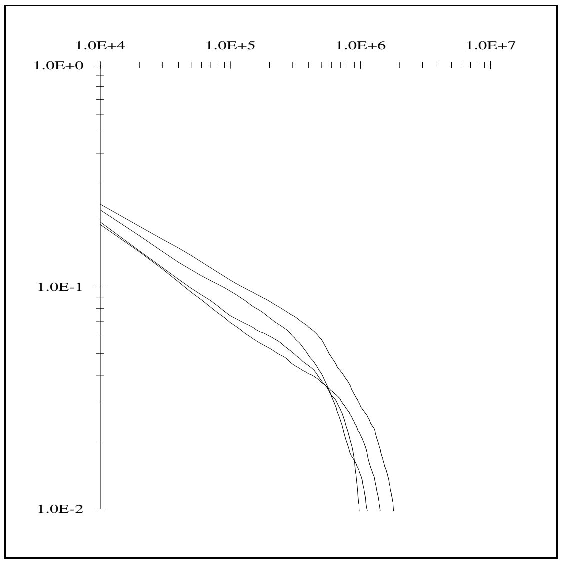


Figure 30: $\log \Pr[u > n]$ versus $\log n$, simulations for GI-G-∞ traffic of type B.

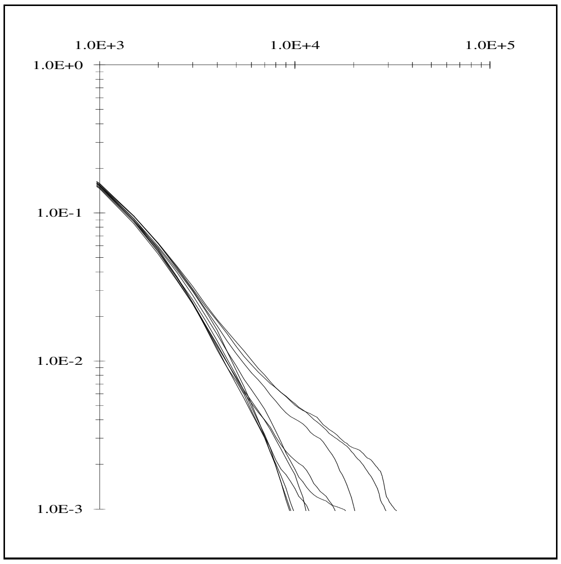


Figure 31: $\log \Pr[u > n]$ versus $\log n$, simulations for GI-G-∞ traffic of type C.

frequencies on its queueing behavior, and how, for instance, the index of dispersion for counts of the traffic relates to the probabilities in the RHS of the above formula. A (normal ?) approximation of the involved distribution, based on only first- and second-order characteristics of the traffic process, is an option that seems worthwhile considering.

One can show that, for the GI-G-∞ model,

$$\mathrm{E}[z^{p_k+p_{k-1}+\ldots+p_{k-l}}|u_{k-l}=0] = \exp\left\{\lambda^* \sum_{n=0}^{l} \Pr[\tau_A > n](l+1-n)z^n(z-1)\right\}$$

(Note that the condition $u_{k-l}=0$ determines the evolution of the traffic process in slots $k-l,\ldots,k$. It implies that no sources were active just prior to slot $k-l$.) After some further manipulations, one obtains the following result

$$\Pr[u_{k+1} > m] = (1-\lambda_T)\sum_{l=1}^{+\infty}\sum_{k=m+l}^{+\infty} \mathrm{Res}\left[\frac{1}{z^{k+1}}\exp\left\{\lambda^*\sum_{n=0}^{l-1}\Pr[\tau_A > n](l-n)z^n(z-1)\right\}\right]_{z=0}$$

It still remains to be determined if replacing residues around $z=0$ by residues around the other singularities (poles or branches) of the function involved, will lead to 'practical' results, or if an accurate numerical transform inversion is feasible.

A slot-to-slot approach The queueing of discrete-time on-off sources was studied by a slot-to-slot approach in e.g. [122] for a finite number of sources (with geometric off-times), and in e.g. [123] for a infinite number of sources. The model in the latter paper is more general than the model of Section 2.3.2, in that the number of new sources becoming active during a slot can have an arbitrary distribution. The special case of a Poisson distribution then leads to the GI-G-∞ arrival process considered here. It is worth noting that the Poisson distribution has a number of properties which simplify the analysis and results to some extent.

The analysis can proceed as follows. Consider $P_k(z, x_1, x_2, \ldots)$, the joint pgf of u_k, the number of cells in the system, and of $v_{i,k}$, the numbers of active sources which will still generate a single cell per slot during the i slots to come, all observed at the beginning of slot k. One can then establish the following relation.

$$\begin{aligned} &P_{k+1}(z, x_1, x_2, x_3, \ldots) \\ &= \quad z^{-1} \exp\left\{ \lambda^* \sum_{k=1}^{+\infty} a(k)(x_k - 1) \right\} \Big(P_k(z, z, zx_1, zx_2, \ldots) + (z-1) P_k(0, z, zx_1, zx_2, \ldots) \Big) \end{aligned}$$

The pgf $P_k(0, x_1, x_2, \ldots)$ can easily be obtained by observing that the queue being empty at the beginning of a slot (argument $z = 0$) implies no sources generated cells during the previous slot. (Note the connection with a similar observation made for the Beneš approach.) This straightforwardly leads to

$$P_k(0, x_1, x_2, \ldots) = \Pr[u_k = 0] \exp\left\{ \lambda^* \sum_{k=1}^{+\infty} a(k)(x_k - 1) \right\}$$

We will not go further into the details of the analysis here, but it is possible to derive from the above two formulas an expression for the mean buffer contents in regime. It is given by

$$\lim_{k \to \infty} \mathrm{E}[u_k] = \lambda_T + \frac{\lambda_T^2}{2(1-\lambda_T)} \left(\mathrm{E}[\tau_A] + \frac{\sigma_A^2}{\mathrm{E}[\tau_A]} \right)$$

This expression can also be found from that in [122] by a limiting procedure, or from that in [123] by assuming a Poisson arrival process for new sources. The formula contains the variance of the durations of the on-periods and is infinite for on-time distributions having infinite variance. This infinite mean points towards a (very) slowly decaying tail for the queue size distribution, as indicated by the simulation results.

We believe - but couldn't prove yet - that, in general, the following goes

$$\Pr[\tau_A = m] \sim m^{-q} \Rightarrow \Pr[u = m] \sim m^{-(q-1)} \Rightarrow \Pr[u > m] \sim m^{-(q-2)}$$

while

$$\Pr[\tau_A = m] \sim z_0^{-m} \Rightarrow \Pr[u = m] \sim z_0^{*-m} \Rightarrow \Pr[u > m] \sim z_0^{*-m}$$

Similar observations have been made for fluid-flow models, see e.g. [15]. Of course, in order for this result to be of 'practical' value, one should also be able to derive the constant of proportionality, i.e., the 'intercept' of the curve $m^{-(q-2)}$ or z_0^{*-m}. (In e.g. [123] it was assumed that the dominating singularity of the pgf of the system contents is an isolated pole z_0^*, somewhere in the interval $(1, +\infty)$ of the real line, which leads to geometric tail decay. However, this assumption is no longer valid when heavy-tailed on-time distributions are involved, since the corresponding pgf's have a branchpoint at $z = 1$. It remains to be studied how the approach has to be modified to deal with this case.)

Discussion The analysis of the discrete-time on-off source model and its use in studying the phenomenon of long-range dependence is presented. Various numerical examples clearly illustrated the distinction between short- and long-range-dependent traffic.

A limiting model for the aggregated traffic, the GI-G-∞ queue, was introduced. It was shown to have a number of nice mathematical properties, which might considerably simplify a future analysis of the queueing behavior of that traffic.

Two approaches towards this analysis, the Beneš approach and a slot-to-slot approach, were outlined. Naturally, both approaches are related, since they pertain to the same model.

2.3.3 Shifting level models

Some cases in practice long range dependence can be seen as an artifact of non-stationarity. For example, the observed long-range dependence of VBR traffic in section 2.2.4 can be well explained by shifting level processes. These processes are non-stationary on any finite time scale but converge to weak stationarity in the long run. Based on shifting level processes a new traffic model is introduced [41] and its implications are addressed in this section.

Modeling VBR Traffic The shifting level process introduced in section 2.2.4 is based on the assumption that Y_i and ΔT_i are independent. The validity of this assumption for *III* traffic can readily be checked by resampling without replacement one of the processes and keeping the sampling order of the other one unchanged. From Figure 32 it can be seen that much of the dynamical behaviour of *III* traffic is retained, if the epochs are resampled without replacement and the sequence of bit rates Y_i during an epoch remains unchanged. In contrast, Figure 33 shows that the dependence structure is destroyed, if the Y_i are resampled without replacement and the sequence of epochs remains unchanged.

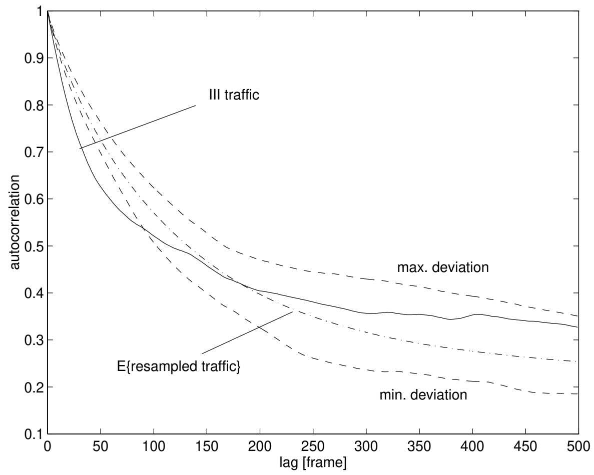


Figure 32: Resampling of ΔT_i in *III* traffic

Both figures prove that Y_i and ΔT_i are dependent processes. A further analysis of the individual processes shows that both of them can best be modeled by fractionally differenced white noise (with $H = 0.8$ for epochs and $H = 0.95$ for the mean bit rates) passed through an ARMA(1,1) filter. Since the original processes are not normally distributed (see Figure 34), one has to employ the probability integral transformation [93] to match the sampled distributions.

Figure 35 shows the autocorrelation function of *III* traffic along with the expected autocorrelation of the model and its minimum and maximum deviation from the expection. Figure 35 and Figure 36 prove that the new model captures well the behaviour of VBR traffic.

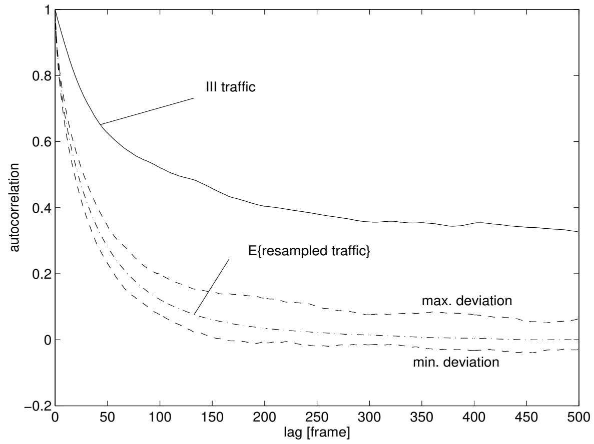


Figure 33: Resampling of Y_i in *III* traffic

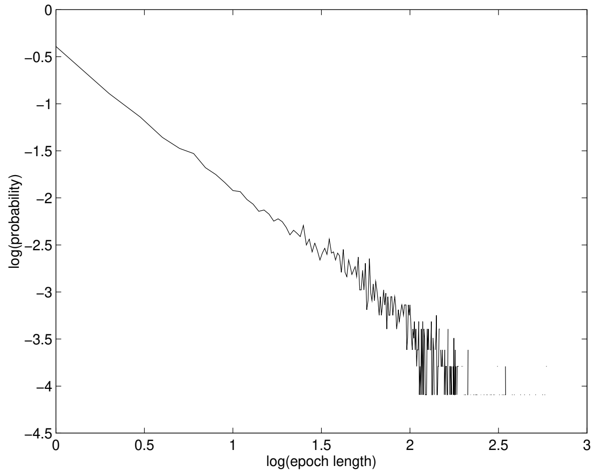


Figure 34: Distribution of ΔT_i for *III* traffic

Some Implications In earlier work it was assumed that VBR traffic can be modeled by ARMA or Markov processes. Many results for dimensioning network queues are based on this assumption. *Beran* et al [10] point out *" ... that the performance of queueing systems with long-range dependent input streams can be drastically different from the performance predicted by traditional short-range dependent models"*.

Simulations of VBR traffic with the new model have confirmed the slow convergence of the mean observed in [39]. Moreover, the confidence intervals obtained from these simulations are rather large for mean and standard deviation. This calls the meaning of simple traffic descriptors, such as mean and variance, into question at least in the case of live broadcasting. For pre-recorded events the situation is friendlier, since all information for the network can be extracted prior to transmission.

However, the broadcaster has to deal with a non-stationary process, but not necessarily the network. Simulations have shown that a superposition of VBR sources gives

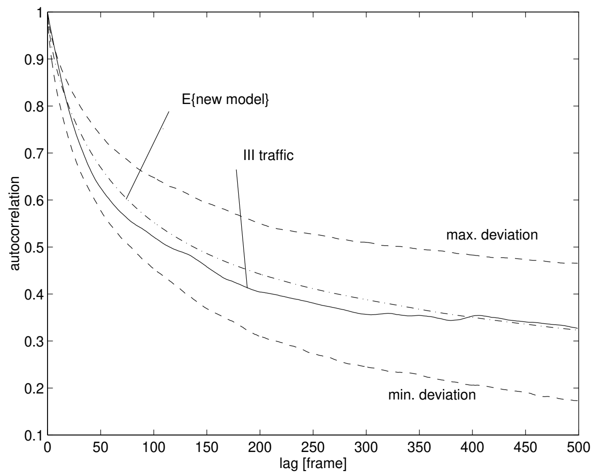


Figure 35: Autocorrelation of modeled *III* traffic

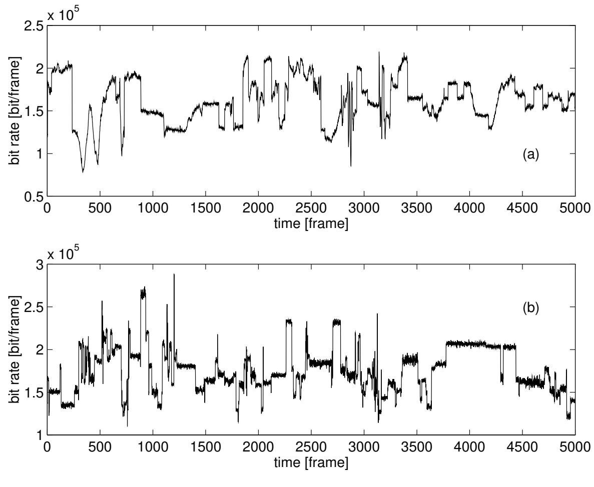


Figure 36: (a): *III* traffic, (b) New model

a weakly stationary and long-range dependent traffic stream. This implies whenever a large number of VBR traffic streams are mixed the Hurst parameter might be an appropriate measure. This situation changes if the number of mixed traffic streams is low. Then the non-stationarity dominates and the mixed stream can stay for a long time on a high bit rate level; a scenario which cannot be captured by conventional long-range dependent models, such as fractionally differenced white noise.

Discussion Based on a formal statistical test and by comparison with other time series it was shown that *III* traffic is non-stationary in the mean (see Section 2.2.4). This type of non-stationarity is best explained by a shifting level process, which is *asymptotically* weakly stationary. A simple resampling experiment proved that the individual processes, which form the shifting level process, are correlated. The correlation structure can be modeled by fractionally differenced white noise. It was shown that this model matches accurately the dependence structure of *III* traffic. It was pointed out that conventional long-range dependent models are accurate if the number of aggregated traffic streams is sufficiently large. However, a single stream is non-stationary and simple traffic descriptors may be insufficient.

2.4 Queueing performance of long-range dependent ATM traffic

There are different concerns about whether LRD is an important traffic characteristics for cell loss estimation or not [29, 45]. We investigate this question and try to identify the relevant correlation time scales of actual measured traffic. We have used a large amount of long traffic traces taken during a trial on the Swedish ATM wide area network and performed LRD, queueing and shuffling analysis [91, 117] in order to investigate

- the effect of LRD on cell loss,
- the relevance of different time scales on cell loss,
- the effect of different buffer sizes on relevant time scales, and
- the effect of different loads on relevant time scales.

2.4.1 ATM measurements

LAN interconnection is one of the most popular services provided by Telia, the Swedish network operator, on its ATM wide area network. Apart from business customers, different parts of the Swedish University Network (SUNET) are also attached to the Swedish ATM WAN. The aggregated traffic on the SUNET were analyzed during summer

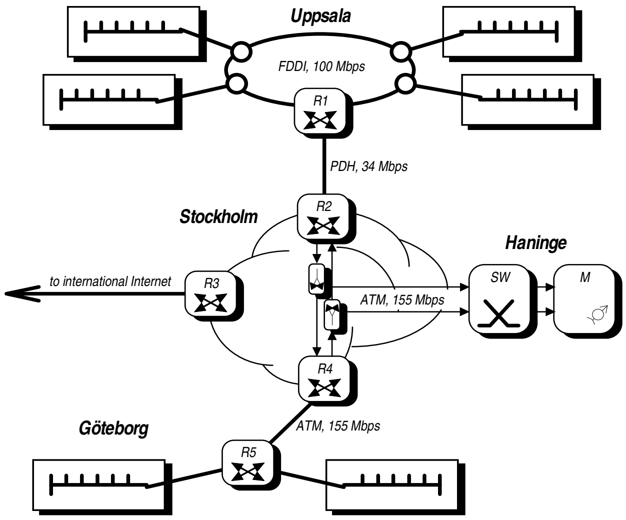


Figure 37: The configuration of measurements on the SUNET.

1996 in the framework of a common trial between the SUNET community and Telia Research. The LAN traffic of universities in the northern region, around Uppsala are connected to an FDDI backbone which is connected via R1, R2 routers and a 34 Mbps PDH link to the ATM backbone in Stockholm (Figure 37). This network joins the northern LANs of SUNET to the international Internet backbone and to the southern university networks around Göteborg. A CBR connection with 90 000 cps (38.16 Mbps) cell rate was established on the SDH link between the routers R4 and R5 for the trial. The measurements reported here were performed on the connections between Uppsala and Göteborg. ATM traffic streams were duplicated by means of optical splitters avoiding impacts on original traffic flows. The duplicated traffic streams were routed on dedicated links to Telia Research in Haninge, where almost one hundred traffic traces were collected with more than 8 million cell arrivals in each trace using a non-commercial custom built measurement instrument developed in the RACE Parasol project [82].

These connections used Telia's Guaranteed Traffic Class thus the influence from other traffic in the network was negligible. A good assumption is that the traffic was an ordinary mix of common Internet traffic types such as HTTP, FTP, telnet, chat, IPphone etc.

To estimate the Hurst-parameter H, the R/S and variance-time analysis [9] were performed for 45 data sets. The obtained values of H are plotted on Figure 38. To get more information about the Hurst-parameter of the measured traffic, H is plotted against the average load of the traces. As can be seen on the figure, H varies within the range $(0.8, 1)$ and does not depend significantly on the load. Figure 39 reveals an

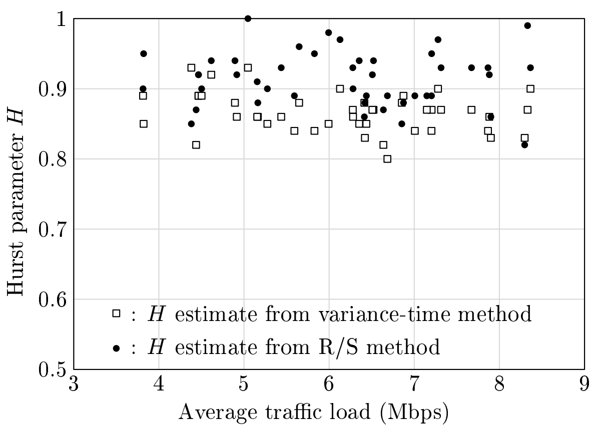


Figure 38: Estimated values of H as a function of traffic load.

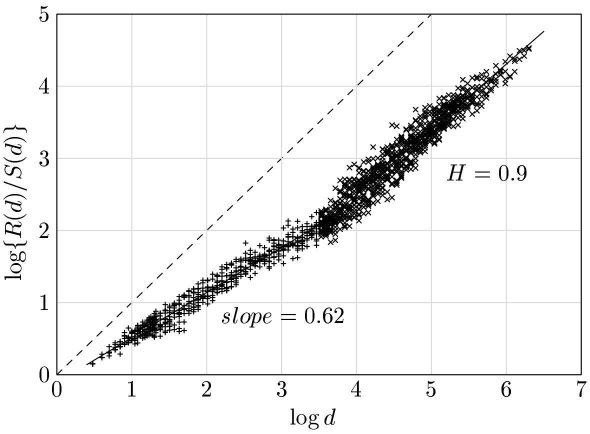


Figure 39: R/S values plotted against the logarithm of block size.

interesting phenomenon, namely, that there is a knee-point in the R/S diagram that separates two linear regions of sample points. Since LRD is an asymptotic property, the linear region to the right (sample points marked with an 'x') was used to determine the estimate of H. The same knee-point was found in all the data sets. The origin of this behaviour could be revealed by examining the burstiness structure of the data more deeply.

2.4.2 Relevance of time scales in queueing

An important question is what is the impact of LRD on queueing. Several engineering issues, such as buffer dimensioning and traffic control, are related to this question which makes it extremely important. There are two opposing viewpoints on this problem. One claim is that the queueing performance is determined by the time scale of busy periods of the queues and there is no practical impact of correlations above this time scale [45, 50, 57, 108]. The contradicting claim is based on several studies [29, 62] and states that LRD is one of the main characteristics of the traffic with significant effect on queueing behaviour. In this section we present some experimental results in order to clarify this question.

Queueing set-up In the performance analysis we used the queueing set-up shown in Figure 40. The pre-recorded traces were taken from our SUNET database as described in section 2. The correlation structure of the original data was artificially modified by a 'shuffler' (see next section). The functionality of this shuffler was investigated by computer simulation in our experiment. The QoS measures under investigation were the complementary queue length distribution and the cell loss ratio (CLR).

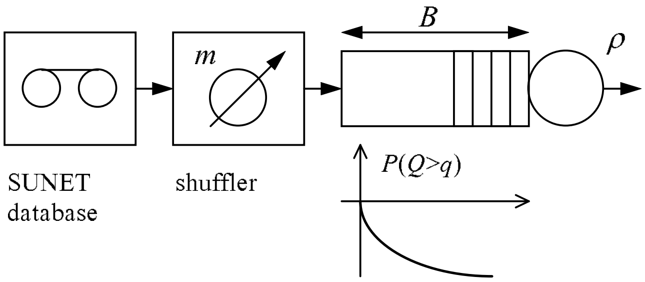


Figure 40: Queueing set-up.

External shuffling In our analysis, external shuffling was used as a tool to modify the correlation structure of the measured traffic traces. Using the terminology of [4, 29], we call 'external shuffling' the following method. First, divide the sequence of interarrival times[2] into blocks of size m. For a measured traffic trace containing N cell arrivals, there are N/m such blocks. Then the order of the blocks is shuffled, while preserving the cell sequence inside each block. Thus, for different values of m we preserve the short-range correlations (up to lag m) while eliminating the long-range correlations (beyond lag m).

Therefore by plotting the complementary queue length distributions for the original and the shuffled traces with different block sizes we have a simple tool to investigate the effect of short-term and long-term correlations in queueing (see Figure 41).

Queueing properties of LRD input In case of LRD traffic input, the tail of the complementary queue length distribution decays slower than exponentially. The uppermost solid line related to the original traffic trace in Figure 41 shows this phenomenon. (The

[2] Besides the permutation of a sequence of interarrival times, one could perform the same shuffling on the sequence of the number of arrivals in consecutive time slots [4].

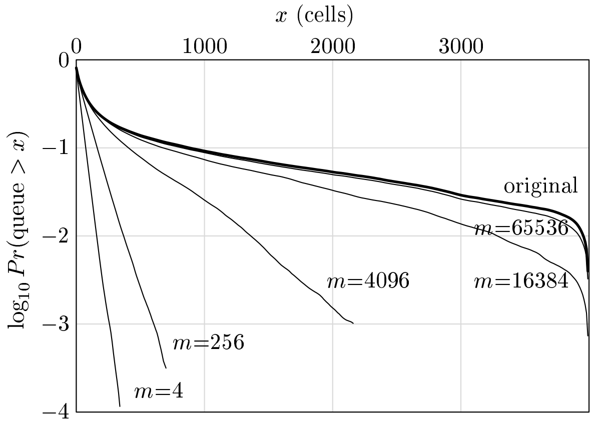


Figure 41: Complementary queue length distributions.

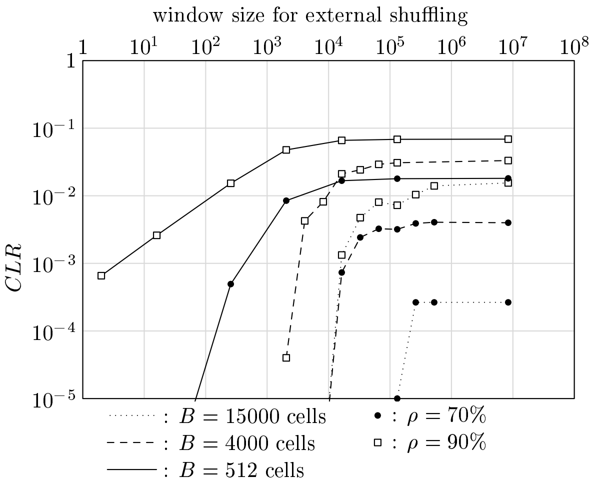


Figure 42: Cell loss ratio with external shuffling.

buffer size was set to 4000 cells and the service rate was chosen so that the utilization ρ was 0.7. The shuffling block size m is shown on the figure for each simulated curve.) By shuffling the input traffic, we can observe that if we increase the block size the curve approaches the distribution curve of the original trace. It means that we can find the time scale of the relevant correlations, i.e., the block size of shuffling where the distribution is approximately the same as for the original case, which are important for queueing. It can also be concluded that beyond that time scale there is no significant effect of correlations. In our case this time scale is in the range of 4–5 seconds (block size of 10^5) for load 0.7. This finding is in accordance with the results reported in [45].

Impacts of LRD on cell loss The queue length distribution in the infinite buffer case gives (almost) always an upper bound on the cell loss ratio (CLR) for the finite buffer case. In practice, the latter is of interest. To investigate the effects of LRD on CLR, we performed simulation studies with the measured traces as input for three different buffer sizes (512, 4000 and 15000 cells). The rate of service was set to obtain utilizations 0.7 and 0.9. The results are plotted on Figure 42.

We can observe in Figure 42 that *there is an upper time scale determined by the buffer size and the load where there is no effect of correlations on cell loss if we go beyond that time scale*. For example, in our experiments above cell lag 10^5 (approximately 5 seconds) the cell loss curves are practically constant even for the large (15000 cells) buffer case. It supports our finding derived from Figure 41, namely, that there is a time scale which determines the biggest lag of correlations which has effect on the cell loss. However, this upper time scale seems to be dependent on many parameters. First, it depends on the buffer size, i.e. the bigger the buffer the bigger the upper time scale. For moderate and large buffers there is a sharp cut-off in cell loss as a function of the shuffling block size but for small buffers the appearance of this upper time scale is not very pronounced. Secondly, this upper time scale is also dependent on the load, the curves are shifting left as the load increases although the shifting of cut-off lag is not significant. In our experiments we can observe that the cut-off lag is always above the buffer size but typically not more than one decade above the buffer size. These findings give us simple practical engineering rule of thumb for estimating the range of relevant correlation time scale. A comprehensive study of these investigations with many traffic traces and queueing conditions is performed in our research but the detailed presentation of these results is beyond the scope of this paper.

2.4.3 Summary

We reported a traffic measurement experiment on SUNET ATM WAN. The recorded long cell traces that cover many time scales were used for LRD analysis and to investigate the question how LRD behaviour influences cell loss in different queueing environments. The impacts of correlations of different time scales were analysed by using external shuffling with different block sizes.

Our basic finding is that there is an upper time scale which determines the range of correlation of interest from a cell loss point of view. This time scale sensitively depends on the buffer size and slightly depends on the load but roughly speaking it is not bigger than ten times the buffer size. By using this simple engineering rule we have a fast approximation about the time scales of queueing interest. The detailed analysis and the development of an accurate approximation of the upper time scale is the topic of our present and future research.

2.5 Queueing performance of synthetic self-similar traffic

A queueing analysis [72] is presented in this section using Fractional Brownion Motion input traffic which is an exactly self-similar process.

2.5.1 FBM model

Fractional Brownian motion (fBm) has been extensively used to model self-similar traffic [4,5]. Consider a fBm process Z(t) with Hurst parameter H in [1/2,1). Let $A(t)$ denote the number of arrivals from a traffic stream over time $(0, t]$, given by

$$A(t) = mt + \sqrt{ma} Z_t$$

where $m > 0$ is the mean rate, and a is the coefficient of variance. Consider a single buffer of length B and service rate C.

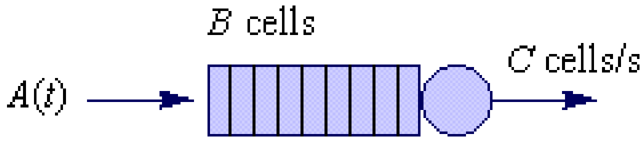


Figure 43: ATM Multiplexer

The queue length distribution may be approximated by a Weibull distribution given by Norros [97].

$$P(Q > B) \sim \exp\left\{ -\frac{(C-m)^{2H}}{2\kappa^2(H)am} B^{2-2H} \right\}$$

where $\kappa(H) = H^H (1-H)^{1-H}$

A further step can be made using the Bahadur-Rao theorem [30] to give the following for the queue length distribution given a single traffic source.

$$P(Q > B) \sim \frac{1}{\sqrt{2\pi}\sigma_{t^*}\theta_{t^*}} \exp\left\{ -\frac{(C-m)^{2H}}{2\kappa^2(H)am} B^{2-2H} \right\},$$

where $\theta_{t^*} = \dfrac{B + Ct^* - mt^*}{am(t^*)^{2H}}$, $\sigma_{t^*}^2 = am(t^*)^{2H}$

2.5.2 Cross-over point

For small buffers the self-similar traffic actually experiences a lower cell loss than Markovian traffic. The crossover point at which the self-similar traffic experiences the same cell loss as random traffic is dependent on the free capacity and the Hurst parameter. It is not (directly) dependent on the utilization or the variance. (The corresponding cell

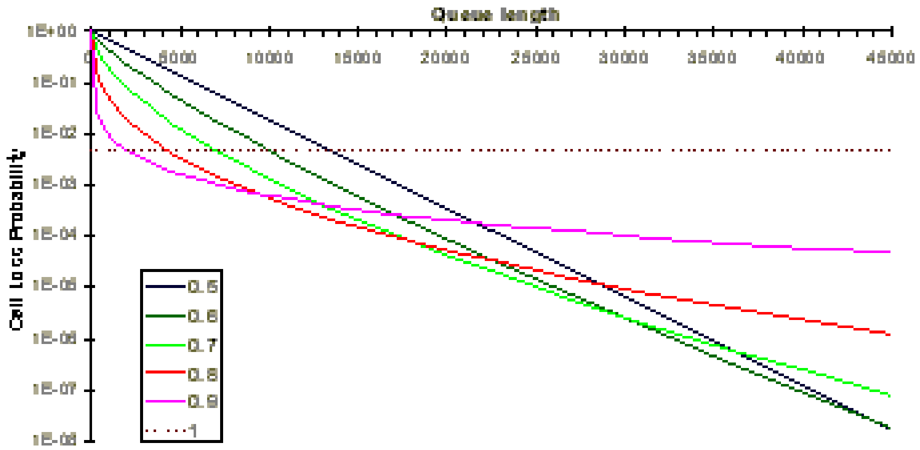


Figure 44: STM-1 link

loss at which the cross-over occurs however is dependent on utilization and the variance.) The value of the buffer size at which the cross-over occurs is given by

$$B = (C - m)(2H^H(1-H)^{1-H})^{\frac{2}{1-2H}} = (C-m)(4\kappa^2(H))^{\frac{1}{1-2H}}$$

Since $H \in [1/2, 1)$,

$$\frac{B}{C-m} \in (1/4, 1)$$

For example, consider an STM-1 link subject to self-similar traffic at 85 load (so m=132Mb/s). Assume the variance at a 1 second timescale is 25(Mb/s)^2. The cell loss for streams of different Hurst parameter is shown in Figure 44 for varying buffer size. It should be noted that this is an approximation for cell loss based on the buffer occupancy of an infinite queue.

The cross-over points where self-similar and Gaussian traffic have the same cell loss occur at the same queue size regardless of variance, though clearly the cell loss at this point is lower with smaller variance. Figure 45 shows how the buffer size at which the self-similar traffic has lower cell loss varies with the Hurst parameter and load.

The self-similarity only becomes an issue (in terms of experiencing worse QoS than Markov traffic models would predict) when the utilization is very high, or when the Hurst parameter is near to 1. Previous characterization suggests that most data traffic has a Hurst parameter around $0.7 - 0.85$. Suppose we have a traffic source, with known mean m, variance v at 1-second timescale, QoS 10^{-7}, and Hurst parameter H. Then the cross-over point occurs when

$$B^2 = \frac{1}{2}\gamma \ln(10)\nu(4\kappa^2(H))^{\frac{1}{1-2H}} \in \left(\frac{1}{8}\gamma\ln(10)\nu, \frac{1}{2}\gamma\ln(10)\nu\right)$$

The following graph shows the buffer size at which the cross-over point occurs as the variance changes, with a cell loss probability of 10^{-6}, 10^{-9} and 10^{-12}. If sources

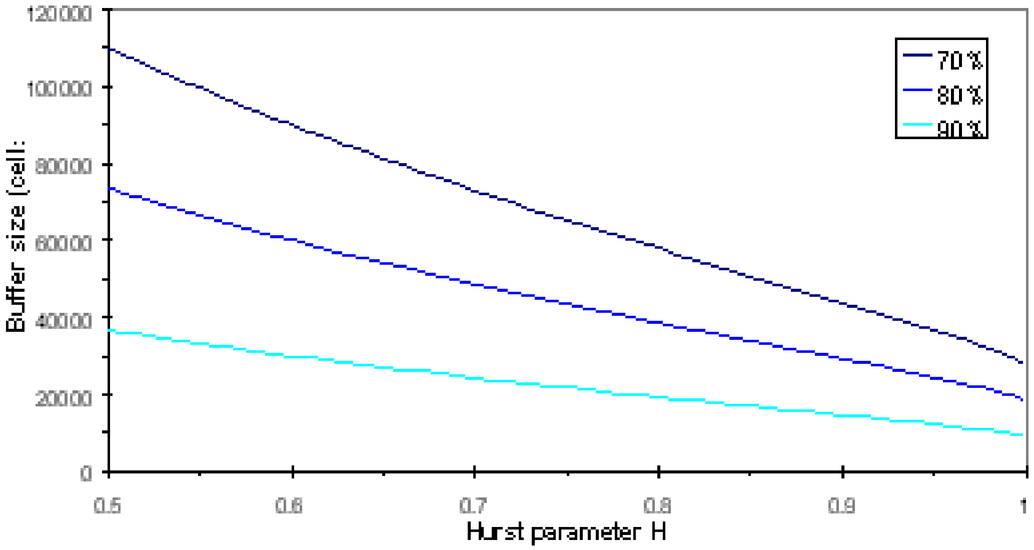


Figure 45: Cross-over point for STM-1 link

similar to the SMDS traffic mentioned above were multiplexed together, to give an 85 load on an STM-1 link, the variance at the 1-second timescale over the busy hour would be approximately $25(\mathrm{Mb/s})^2$. This corresponds to a minimum cross-over point with a buffer size of approximately 15,000-21,000 cells, depending on the CLR. Even if the buffer is twice this, traffic will only experience a higher cell loss rate than Markov traffic of the same variance if the Hurst parameter $H > 0.82$.

For buffers smaller than this, a Markov model would provide a pessimistic estimate for the cell loss, and hence over-dimension the network. Only if the buffer is larger than this would the self-similarity become an issue.

Given a switch with buffer size B and service rate C, let r be the utilization that can be achieved while maintaining a cell loss ratio (CLR) of 10^{-7}. Let v_1 be the variance of the traffic if it was scaled up to 155 Mb/s mean. Then for a utilization r, the variance is given by $v = v_1 r^{2H}$. Using the large deviation approximation, the maximum utilization that maintains a CLR of 10^{-7} is given by

$$\frac{1}{\rho} = 1 + \left[\frac{2\gamma(\ln 10)\kappa^2(H)\nu_1}{C^{2H}B^{2-2H}}\right]^{\frac{1}{2H}}$$

Figure 47 shows the utilization that may be achieved on an STM-1 link for various values of the Hurst parameter and varying buffer size. The variance of the traffic source is chosen such $v_1 = 25(\mathrm{Mb/s})^2$ using the notation above.

For an STM-1 link with traffic satisfying the above assumption regarding variance, and buffers of less than approximately $20,000$ cells, dimensioning according to Gaussian approximations will always err on the safe side. For a lower cell loss the cross-over point occurs at a larger buffer size.

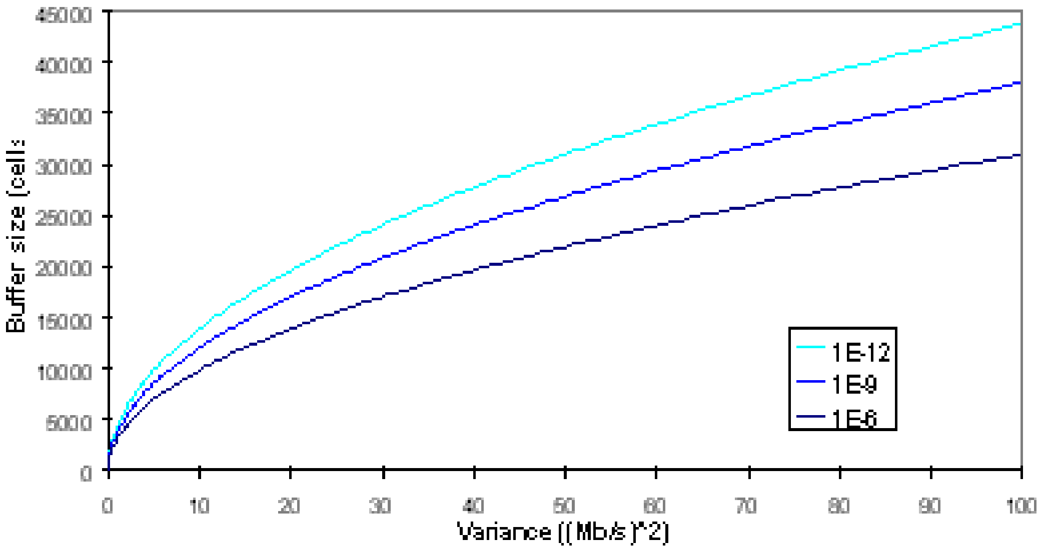


Figure 46: Minimum cross-over point for a given QoS

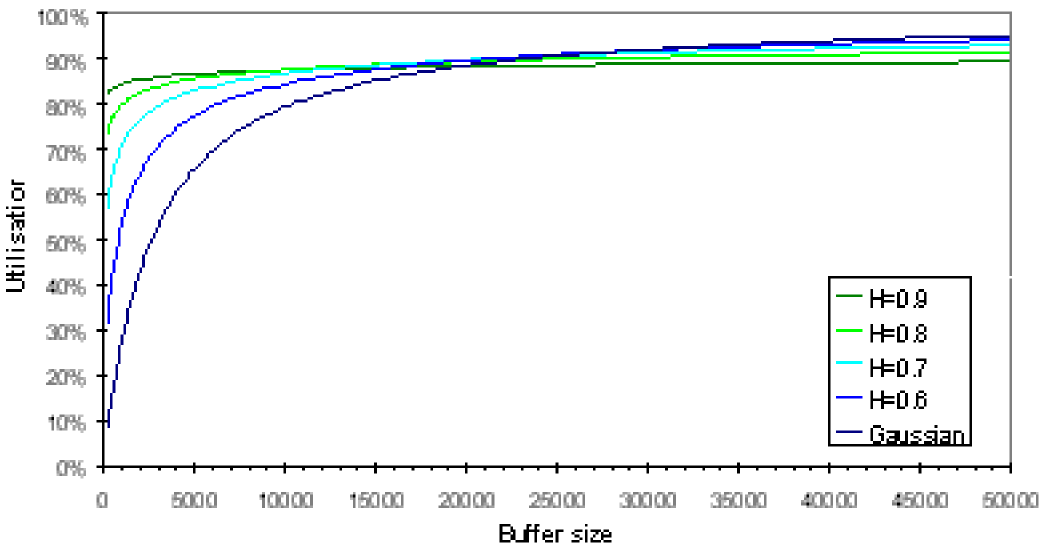


Figure 47: Maximum utilization to maintain 10^{-6} CLR

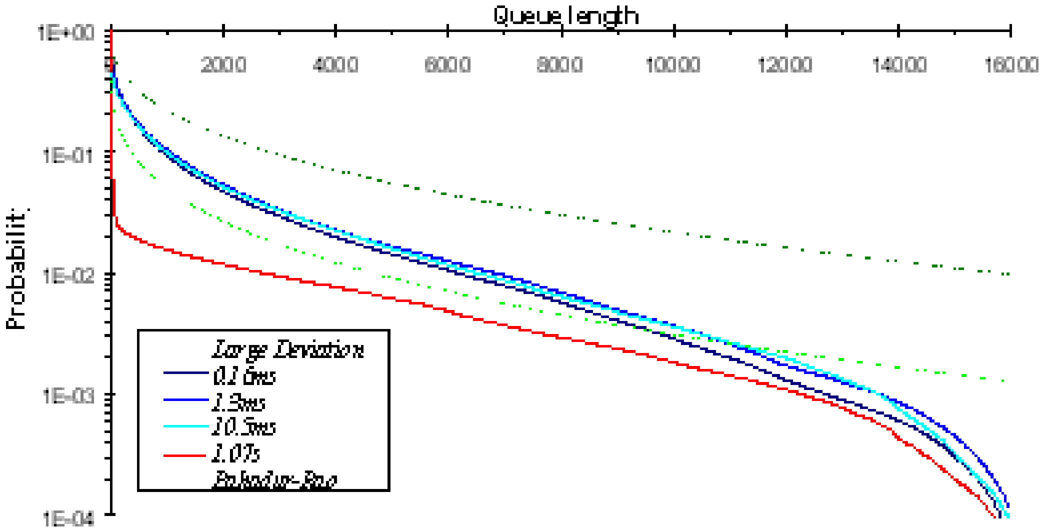


Figure 48: Cell loss using fGn model down to different timescales

2.5.3 Simulation

A self-similar background traffic generator has been implemented in Simula for use within an existing ATM network simulation model. The primary purpose of this was to model background data traffic, and to determine the effect of self-similar traffic on other traffic sources. The self-similar traffic on its own was found to fit well with the theory shown above, though a couple of observations were made. Self-similar traffic was simulated using fBm to determine the rates down to a millisecond timescale. At a finer timescale ($<$ 1ms), traffic is assumed to have a Poisson arrival process, with the rate determined by the fBm model. The simulation used a combination of methods to produce fractional Gaussian noise, firstly using the Discrete Time Fourier Transform method [99] to generate fGn rates at a 1-second timescale for the duration of the simulation, and secondly the Random Midpoint Displacement (RMD) algorithm [8] to calculate the rates at a sub-second timescale as they were required. This seemed to be fairly stable, in that changes to the timescale at which the Poisson arrival assumption is made had relatively little effect on the resultant cell loss rates, up to a point (see Figure 48). The lower than expected cell loss is due in part to input traffic being truncated at 155Mb/s in this experiment. Figure 48 shows the difference that was made by changing the timescale below which Poisson arrivals are assumed.

Figure 49 shows a comparison of simulation against theory for a single queue subject to a self-similar traffic stream. In this example, the link has capacity C=100 Mb/s, and the traffic source has m=90 Mb/s, v=16(Mb/s)^2, and $H = 0.8$. The queue length distribution for Gaussian traffic is also included. (The load on the link is deliberately high so that a significant number of cell losses might be seen by the simulation.)

The Large Deviation approximation appears to be a good approximation to the queue length distribution, however it is still an overestimate, and hence the buffer sizes discussed above at which self-similarity results in higher cell loss are an underestimate.

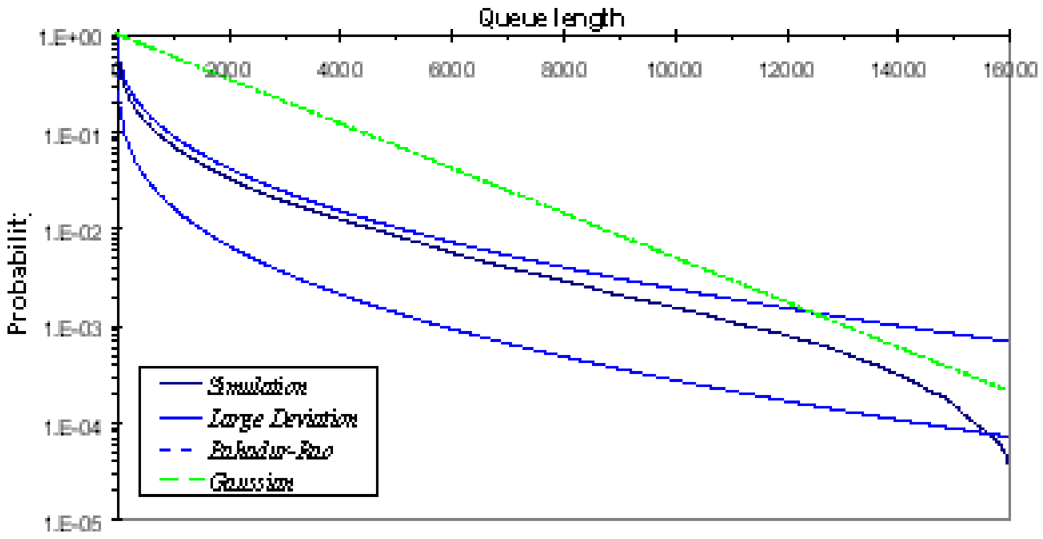


Figure 49: Comparison of simulation and theory

It should be noted that the simulation here is trace driven, and does not take account of real-time changes in the traffic as a result of controls detecting congestion in the network and adapting accordingly. Simulating such end-to-end controls for all traffic over an STM-1 link is not practical though, so we assume traffic has been shaped the way it is by congestion and multiplexing elsewhere in the network, and handle the resultant trace on the assumption that there is minimal cell loss. It should also be noted that despite a run-time of over 10^9 cell slots, the results still show a tail-off around 10^{-4} cell loss for H=0.8. Due to long-term fluctuations in the traffic, simulation of self-similar traffic requires very long run times to get reasonable accuracy. On the other hand, as H was decreased, high peaks in the traffic profile also caused significant problems for our simulation, increasing the run-time.

2.5.4 Summary

For typical buffer sizes, self-similar traffic actually experiences a lower cell loss than Markovian traffic with the same variance at a measurable timescale (say, 1 second). Only core network dimensioning has been considered here. The FGN approximation assumes free traffic, in that traffic is not controlled as a result of feedback concerning network utilization. Users are unlikely to make this free traffic assumption on their access class, on the basis that an access will often be picked either to handle the anticipated peak (for Real Time services), or to be 'sufficient' in some way (for Non-Real Time services), relying on feedback, say at the TCP layer, to control the traffic. The switches that are of concern regarding cell loss are therefore the core switches where traffic has been aggregated. No distinction has been made regarding VBR/ABR/UBR, and this section treats all of these together.

3 Spatial traffic characterization

The design of third generation mobile communication networks faces three major challenges: first, there is the tremendous world wide increase in the demand for mobile communication services. Second, the main resource in wireless systems, i.e. the frequency spectrum, is extremely limited. And third, new access technologies like *Space Division Multiple Access (SDMA)* and *Code Division Multiple Access (CDMA)* require new mobile network planning methods. Since these challenges are strongly interconnected, they can only be addressed by an *integrated concept*, cf. [113], in order to obtain an efficient, economic and optimal mobile network configuration.

The primary task of mobile system planning is to locate and configure the facilities, i.e. the base stations or the switching centers, and to interconnect them in an optimal way. To achieve an efficient and economic system configuration, the design of a mobile network has to be based on the analysis of the *distribution of the expected teletraffic demand* in the complete service area. In contrast, the traffic models applied so far for the demand estimation, characterize the teletraffic only in a single cell or they are too complex for practical use in the planning process. Therefore, the demand based design of mobile communication systems requires a traffic estimation and characterization procedure which is simple as well as accurate.

We first describe the traffic source models used so far in mobile network design and define a *geographical traffic model* which obeys the geographical and demographical factors for the expected teletraffic in a service region. Subsequently, we introduce the *demand node concept*. This is a novel technique for the representation of the spatial distribution of the teletraffic, which uses discrete points. We also outline a *traffic characterization procedure* which can provide a demand node distribution from publicly available geographical data. To generate the demand nodes, we introduce a *recursive partitional clustering* algorithm and validate the demand node concept by data from a cell structure of an operating mobile network. Finally, we outline how the demand concept can be applied for locating base stations [115].

3.1 Traffic estimation

In mobile communication networks the teletraffic originating from the service area of the system can be described mainly by two traffic models which differ by their view of the network. *a)* The *traffic source model*, which is also often referred to as the *mobility model*, describes the system as seen by the mobile unit. The traffic scenario is represented as a population of individual traffic sources performing a random walk through the service area and randomly generating demand for resources, i.e. the radio channel. An overview on these models is provided in Section 3.1.1. *b)* In contrast, the *network traffic model* of a mobile communication system describes the traffic as observed from the non-moving network elements, e.g. base stations or switches. This model characterizes the *spatial* and *temporal* distribution of the *traffic intensity* E, measured in *Erlangs*, in the two-dimensional service area. Both traffic models are used in mobile communication system design. Particularly the latter model is of principal interest when determining the location of the main facilities in a mobile network, i.e. the base stations and the switching centers. These components should be located close to the expected traffic in order to increase the system efficiency. We will focus in greater detail on this type of models.

3.1.1 Traffic source models

Due to their capability to describe the user behavior in detail, *traffic source models* are usually applied for the characterization of the traffic in an individual single cell of a mobile network. Using these models, local performance measures like *fresh call blocking probability* or *handover blocking probability* can be derived from the mobility pattern. Additionally, these models can be used to calculate the subjective quality-of-service values for individual users.

A widely used single cell model was first introduced by [51]. Their model assumes a uniformly distributed mobile user density and a non-directed uniform velocity distribution of the mobiles. Under this premise, performance values like the *mean channel holding time* and the *average call origination rate* in a cell can be computed.
[28] characterize the mobile phone traffic on vehicular highways by assuming a one-dimensional mobility pattern. They derive the performance values by applying a stationary flow model for the vehicular traffic. A similar one-dimensional highway model with a non-uniform density distribution was investigated by [64]. For the traffic characterization, fluid flow models with time-nonhomogeneous and time-homogeneous traffic have been used, as well as a approximative stochastic traffic model.
A limited directed two-dimensional mobility model was investigated by [36]. The model assumes a spatially homogeneous distribution of the demand and an isotropic mobility structure. [16] investigates a mobility model with a homogeneous demand distribution but assumes a non-uniform velocity distribution. The traffic orientation is non-directed and equally distributed. The application of these traffic source models in real network planning cases is strongly limited. Some models, like the highway model proposed by [64], give a deep insight on the impact of the terminal mobility on the cellular system performance, however they are rather complex to be applied in real network design. Other models, like the one suggested by [51], due to their simplification assumptions, can only be applied for the determination of the parameters in an isolated cell.

3.1.2 Traffic intensity

Since the mobile network planning process requires a comprehensive view of the expected load, a network teletraffic model has to be specified. Therefore, we define the *traffic intensity* function $E^{(t)}(x, y)$. This function describes the offered teletraffic, as seen by the fixed network elements, in a unit area element at location (x, y) and at time instant t. The coordinates (x, y) of the area element are integer numbers. Due to the definition given above, the traffic intensity function is a matrix of traffic values representing the demand from area elements in the service region, cf. Figure 50(b). The traffic intensity $E^{(t)}(x, y)$ can be derived from the location probability of the mobile units.

Under the premise that this probability $p_{\text{loc}}^{(t)}(\chi, \psi)$ is known, the average number of mobile units $\#\overline{mob}^{(t)}(x, y)$ in a certain area element at time t is:

$$\#\overline{mob}^{(t)}(x, y) = \int_{x}^{x+\Delta x} \int_{y}^{y+\Delta y} p_{\text{loc}}^{(t)}(\chi, \psi) d\chi \, d\psi \, . \tag{38}$$

Here, $p_{\text{loc}}^{(t)}(\chi, \psi)$ is the probability that, if the system is viewed from the outside, there is a mobile unit at location (χ, ψ). The location (χ, ψ) is a coordinate in $\mathbb{R}^2$ and $\Delta x \times \Delta y$ is the size of the unit area element.

Using the assumption that every mobile unit has the same *call attempt rate* $r(t)$ at time t, the traffic intensity $E^{(t)}(x, y)$ can be readily obtained:

$$E^{(t)}(x, y) = \#\overline{mob}^{(t)}(x, y) \, r(t) \, . \tag{39}$$

Since in reality it is almost impossible to directly calculate the location probability $p_{\text{loc}}^{(t)}(\chi, \psi)$ from the mobility model, the traffic intensity has to be derived from indirect statistical measures.

3.1.3 The geographic network traffic model

The offered traffic in a region can be estimated by the *geographical* and *demographical* characteristics of the service area. Such a demand model relates factors like *land use*, *population density*, *vehicular traffic*, and *income per capita* with the calling behavior of the mobile units. The model applies statistical assumptions on the relation of traffic and clutter type with the estimation of the demand. In the *geographic network traffic model*, the intensity $E_{geo}^{(t)}(x, y)$ is the aggregation of the traffic originating from these various factors:

$$E_{geo}^{(t)}(x, y) = \sum_{\text{all factors } i} \eta_i \cdot \delta_i^{(t)}(x, y), \tag{40}$$

where η_i is the traffic generated by factor i in an arbitrary area element of unit size, measured in *Erlangs per area unit*, and $\delta_i^{(t)}(x, y)$ is the assertion operator:

$$\delta_i^{(t)}(x, y) = \begin{cases} 0 & : \text{ factor } i \text{ is not valid at location } (x, y) \\ 1 & : \text{ factor } i \text{ is valid at location } (x, y) \end{cases} \, . \tag{41}$$

So far the planning of public communication systems uses geographic traffic models which have a large granularity. A typical *unit area size* is in the order of square kilometers, i.e. in public cellular mobile systems this is the size of *location areas*, cf. [44]. For the determination the positions of base stations a much smaller value is required. The locations of these facilities have to be determined within a spatial resolution of one hundred meters. An unit area element size in the order of $100m \times 100m$ is therefore indicated.

Traffic parameters The values for η_i, which are the traffic intensity originating from factor i per area element, can be derived from measurements in an existing mobile network and by taking advantage of the known causal connection between the traffic and its origin. A first approach is to assume a highly non-linear relationship. A general structure to model this behavior is to use a parametric exponential function. In the geographic model, proposed within this paper, the traffic-factor relationship is defined to be:

$$\eta_i = a \cdot b^{x_i} \tag{42}$$

where a is constant and b is the base of the exponential function. For the validation of Equation 42, presented in Section 3.3, a value of 10 has been used for the basis b.

To reduce the complexity of the parameter determination we introduce the normalization constraint:

$$\frac{E_{\text{total}}}{A_{\text{service area}}/a_{\text{unit element}}} = \sum_{\text{all factors}\, i} \eta_i \,, \tag{43}$$

where $A_{\text{service area}}$ is the size of the service area, $a_{\text{unit element}}$ is the size of an unit area element, and E_{total} is the total teletraffic in this region. The value of E_{total} can be measured in an operating cellular mobile network.

The structure of the geographical traffic model given in Equation40 and Equation 42 appears to be simple. However, it will be shown in Section 3.4 that this model is accurate enough to describe the traffic in cells of an operating mobile network. Moreover, due to its structure the model can easily be adapted to the proper traffic parameters. This capability enables its application for mobile system planning.

Stationary geographic traffic model The above proposed model $E_{geo}^{(t)}(x, y)$ includes also the temporal variation of the traffic intensity in the service area. Since communication systems must be configured in such a way that they can accommodate the highest expected load, the time index t is usually dropped and the traffic models are reduced to *stationary* models describing the peak traffic. The maximum load is the value of the traffic during the *busy hour*, cf. [95].

A pitfall for the network designer remains: the busy hour varies over time within the service area. In downtown areas the highest traffic usually occurs during the business hours, whereas in suburban regions the busy hour is expected to be in the evening. Therefore, the network engineer has to decide how to weight the different traffic factors, i.e. how to obey the different market shares of the various user groups in the traffic model of the network.

3.1.4 Traffic discretization

The core technique of the traffic characterization proposed in this paper is the representation of the spatial distribution of the demand for teletraffic by discrete points, called *demand nodes*. Demand nodes are widely used in economics for solving facility location problems, cf. [40].

Definition: A demand node represents the center of an area that contains a quantum of demand from teletraffic viewpoint, accounted in a fixed number of call requests per time unit.

The notion of demand nodes introduces a discretization of the demand in both space and demand. In consequence, the demand nodes are dense in areas of high traffic intensity and sparse in areas of low traffic intensity. Together with the time-independent geographic traffic model, the demand node concept constitutes a *static population model* for the description of the mobile subscriber distribution.

An illustration for the *demand node concept* is given in Figure 50: part (a) shows publicly available map data with land use information for the area around the city of Würzburg, Germany. The information was extracted from *ATKIS*, the official topographical cartographical data base of the Bavarian land survey office, cf. [7]. The depicted region has an extension of 15km $\times$ 15km. Figure 50(b) shows the traffic intensity distribution in this area, characterized by the traffic matrix: dark squares represent an expected high demand for mobile service, bright values correspond to a low teletraffic intensity. Part (d) of Figure 50 depicts a simplified result of the demand discretization. The demand nodes are dense in the city center and on highways, whereas they are sparse in rural areas.

3.2 Traffic characterization

3.2.1 Traffic characterization procedure

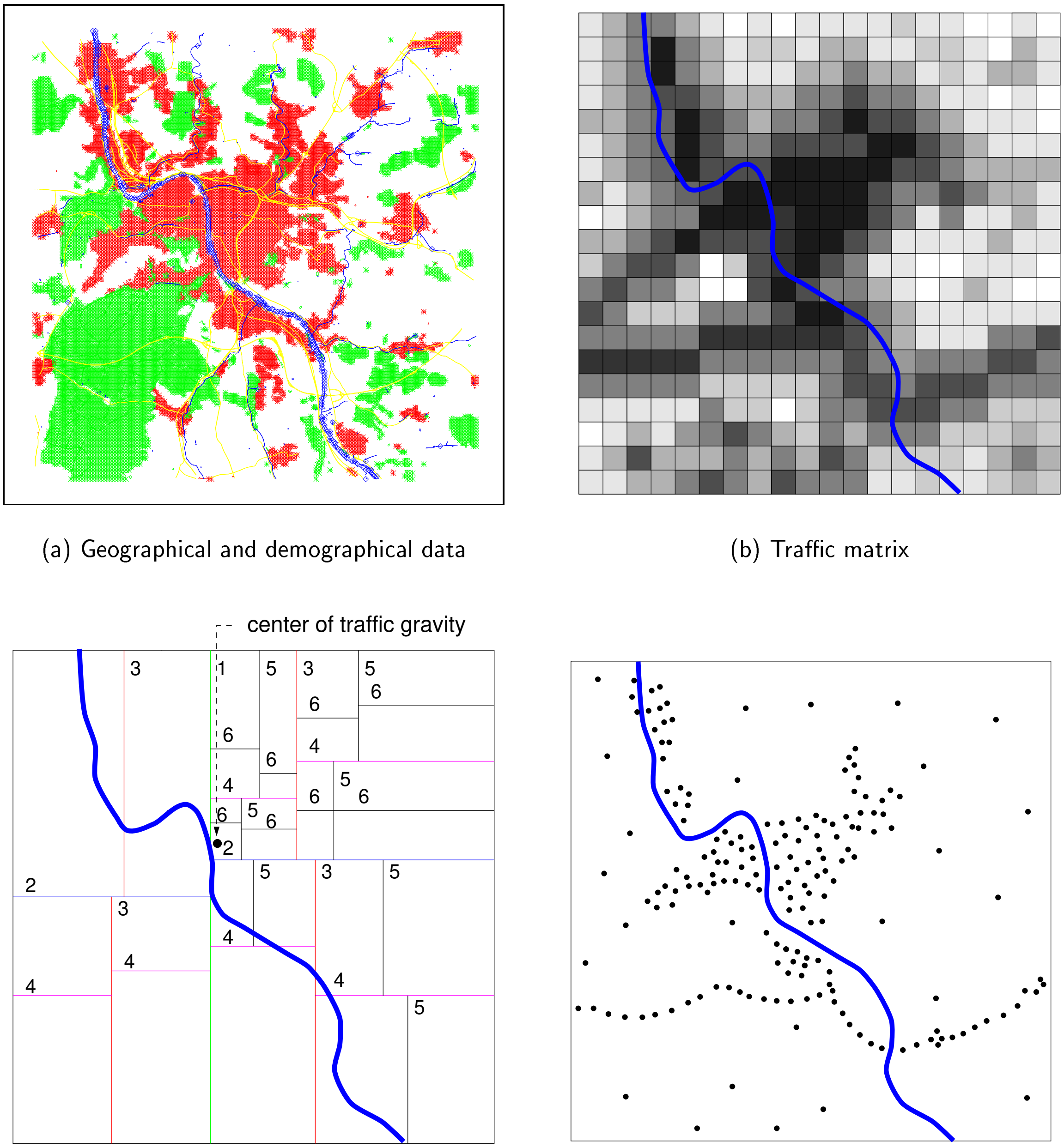
Based on the estimation method introduced in the previous section, the traffic characterization has to compute the spatial traffic intensity and its discrete demand node representation from *real world* data. In order to handle this type of data, the complete characterization process comprises four sequential steps:

Step 1 **Traffic model definition:**
Identification of traffic factors and determination of the traffic parameters in the geographical traffic model.

Step 2 **Data preprocessing:**
Preprocessing of the information in the geographical and demographical data base.

Step 3 **Traffic estimation:**
Calculation of the spatial traffic intensity in the service region.

Step 4 **Demand node generation:**
Generation of the discrete demand node distribution by the application of clustering methods.



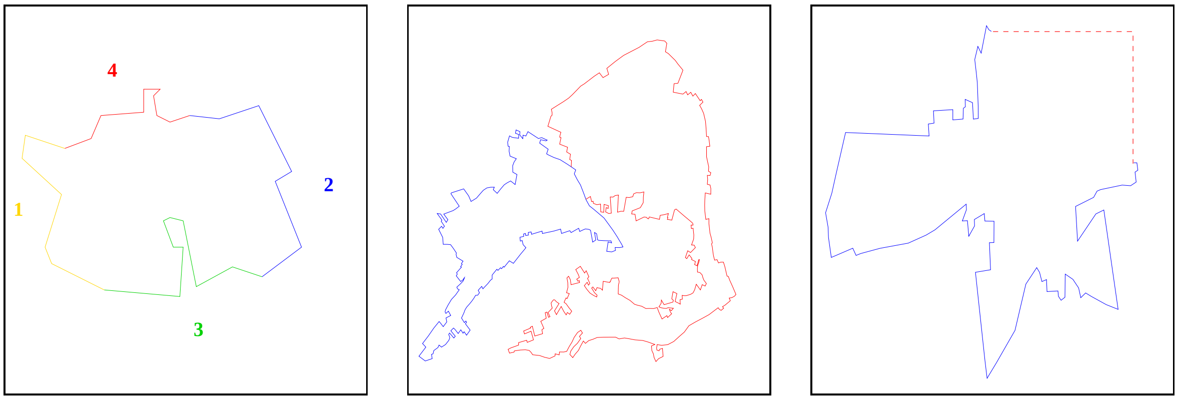
(a) Geographical and demographical data

(b) Traffic matrix

(c) Service area tessellation

(d) Demand node distribution

Figure 50: Demand node concept



(a) Unordered lines (b) Adjacent open and closed polygons (c) Closing lines

Figure 51: Dirty map information data

Traffic model definition The definition of geographical traffic model in *Step 1* of the characterization procedure is based on the arguments given in Section 3.1.3. A simple but accurate spatial geographic traffic model is the base for system optimization in the subsequent network design steps.

Data preprocessing The data preprocessing in *Step 2* is required since the data in geographical information systems are usually not collected with respect to mobile network planning. For example, ATKIS' main objective is to maintain map information. It uses a vector format for storing its drawing objects.

To determine the clutter type of a certain location, one has to identify the land type of the area surrounding this point. This requires the detection of the closed polygon describing the shape of this area. Since maps are mostly printed on paper, the order of drawing the lines of a closed shape doesn't matter, see Figure 51(a). To identify closed polygons, one has to check if every ending point of a line is a starting point of another one. If a closed polygon has been detected, the open lines are removed from the original base and replaced by its closed representation. Additionally, due to the map nature of the data, two adjacent area objects can be stored by a closed and an open polygon, see Figure 51(b). It also can happen that some data is missing, see Figure 51(c). In this case, line closing algorithms have to applied, cf. [63]. After the preprocessing step only closed area objects remain in the data base and the traffic characterization can proceed with the demand estimation.

Demand estimation *Step 3* of the traffic characterization process uses the geographical traffic model defined in *Step 1* for the estimation of the teletraffic demand per unit area element. The computed traffic values are stored in the *traffic matrix*. To obtain the traffic value on a certain unit area element, the procedure first determines the traffic factors valid for this element and then computes the matrix entry by applying Equation 40.

3.2.2 Demand node generation

The generation of the demand nodes in *Step 4* of the characterization process is performed by a *clustering method*. Clustering algorithms are distinguished into two classes, cf. [53]: *a)* the *Partitional Clustering* methods, which try to construct taxonomies between the properties of the data points, and *b)* the *Hierarchical Clustering* methods which derive the cluster centers by the agglomeration of input values.

The algorithm proposed for the demand generation is a recursive partitional clustering method. It is based on the idea to divide the service area until the teletraffic of every tessellation piece is below a threshold θ. Thus, the algorithm constructs a sequence of bisections of the service region. The demand node location is the center of gravity of the traffic weight of the tessellation pieces. The demand node generation algorithm is given in [115].

An example for the bisection sequence of the algorithm is shown in Figure 50(c). The numbers next to the partitioning lines indicate the recursion depth. To make the example more vivid, not every partition line is depicted in the example. The upper left quadrant of the Figure 50(c) shows only the lines until the recursion depth 3, the lower left part the lines until the depth 4, the lower right quarter the lines until depth 5 and the upper right quadrant of the region the lines until depth 6.

The partitional clustering algorithm of Algorithm 1 [115] is a fast but simple clustering method. However, its accuracy depends strongly on the quantization value θ, which gives only an upper bound for the traffic represented by a single demand node. Moreover, since the algorithm constructs a sequence of right-angled bisections, the shape of the tessellation pieces is always rectangular. To overcome these drawbacks, we investigate also hierarchical agglomerative clustering algorithms. These methods are able to obtain tessellation pieces of arbitrary shape and of a predefined traffic value.

3.3 Validation of the traffic estimation

To evaluate the capability of the traffic estimation and characterization of Section 3.2, the traffic approximation of this procedure was compared with the traffic distribution measured in cells of the GSM-based D1 system of the German network operator DeTeMobil. Figure 52 depicts the approximated cell boundaries of the D1 system superimposed on the land use of the investigated area around Würzburg.

The traffic estimation of the demand node concept was based on the geographical network model as defined by the Equations 40 and 42. For the validation, the model considered as the traffic factors the five clutter types which were available for this area in the ATKIS data base: *vehicular traffic*, *urban*, *open outdoor*, *water*, and *forest*. Table 4 shows the values of the exponents used for the calculation of η_i in Equation 42. The parameter a was calibrated from measurements and constant for every traffic factor i. The demand node representation of the estimated traffic in this region, generated by Algorithm 1, is depicted in Figure 53. As expected the demand nodes are dense in the city center and on highways and are sparse in rural areas.

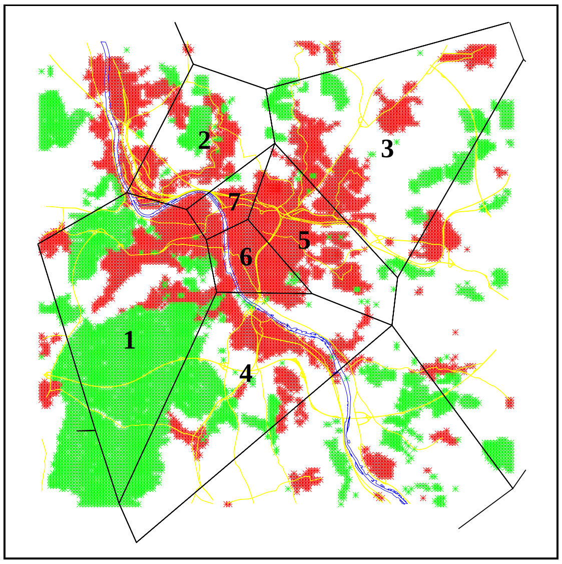


Figure 52: Cell boundaries

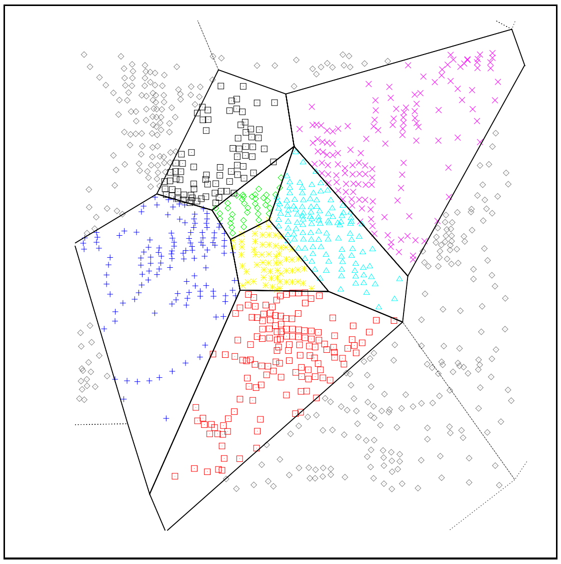


Figure 53: Demand node approximation

The share of the teletraffic of the cells in this area is shown in Figure 54. The solid line represents the proportion of each of the seven D1 cells of the measured total teletraffic. The dotted line in Figure 54 is the estimation of the geographic network traffic model. Both graphs are qualitatively almost the same for the cells with numbers 1, 2, 3, and 4. However, for the cells 5, 6, and 7 the estimation differs strongly from the measured distribution. The cause for the wrong approximation in these cells is the limited distinction of the traffic factors. Due to the use of ATKIS, the model does not distinguish between "urban" and "dense urban". However, the cells 5, 6, and 7 are located in the city center of the Würzburg. The high traffic demand due to the high user density in this area is not reflected in the model.

This example demonstrate that the geographical network traffic has the ability to estimate the traffic quite accurate (cf. cells 1, 2, 3, and 4). However, it has to be extended in some cases (cf. cells 5, 6, and 7).

clutter type	$x_i = log_b(\frac{\eta_i}{a})$
vehicular traffic	3
urban	2
open outdoor	1
water	0
forest	-1

Table 4: Parameter of the traffic clutter relationship

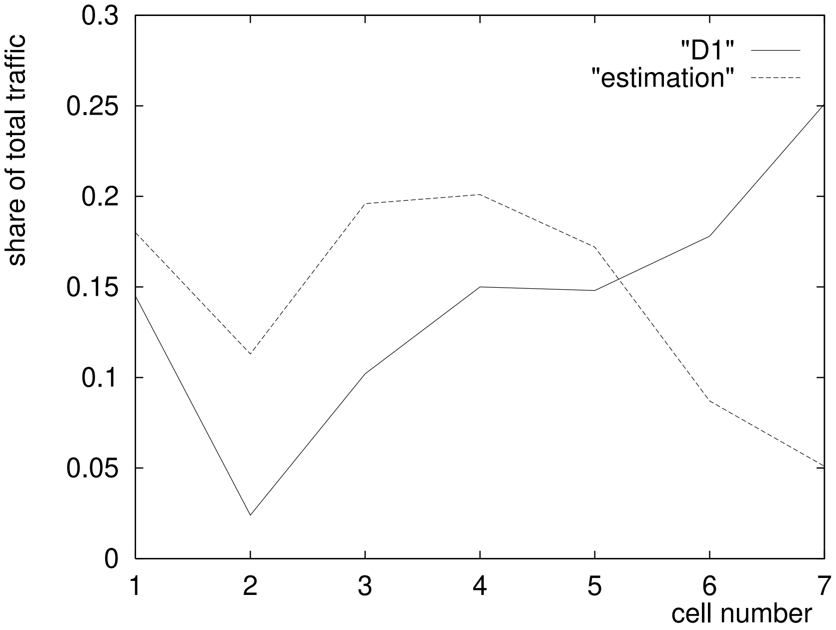


Figure 54: Cell traffic distribution

3.4 Demand based mobile network design

To prove the capability of the demand estimation and to show the feasibility of the integrated design concept, *ICEPT* - a prototype of a planning tool for cellular mobile networks was implemented at the University of Würzburg, cf. [114]. The tools' core components are the automatic network design algorithm *SCBPA (Set Cover Base Station Positioning Algorithm)* and a traffic characterization procedure as described in Section 3.2.

The SCBPA algorithm is a GREEDY heuristic which selects the optimal set of base stations that maximizes the proportion of *covered* traffic, i.e. the ration of the demand nodes which measure a pathloss on the forward/reverse link above the threshold of the link budget, cf. [113].

SCBPA was tested again on the topography around the city center of Würzburg. The task was to find the optimal locations of nine transmitters in this terrain. The result of the algorithm is depicted in Figure 55. The base station locations are marked by a $\diamond$ symbol. The lines indicate the convex hull around the set of demand nodes which are supplied by the base station. The SCBPA algorithm was able to obtain a 75% coverage of the teletraffic of the investigated area. The total computing time for the configuration, including the traffic characterization, was 4min on a SUN Ultra 1/170.

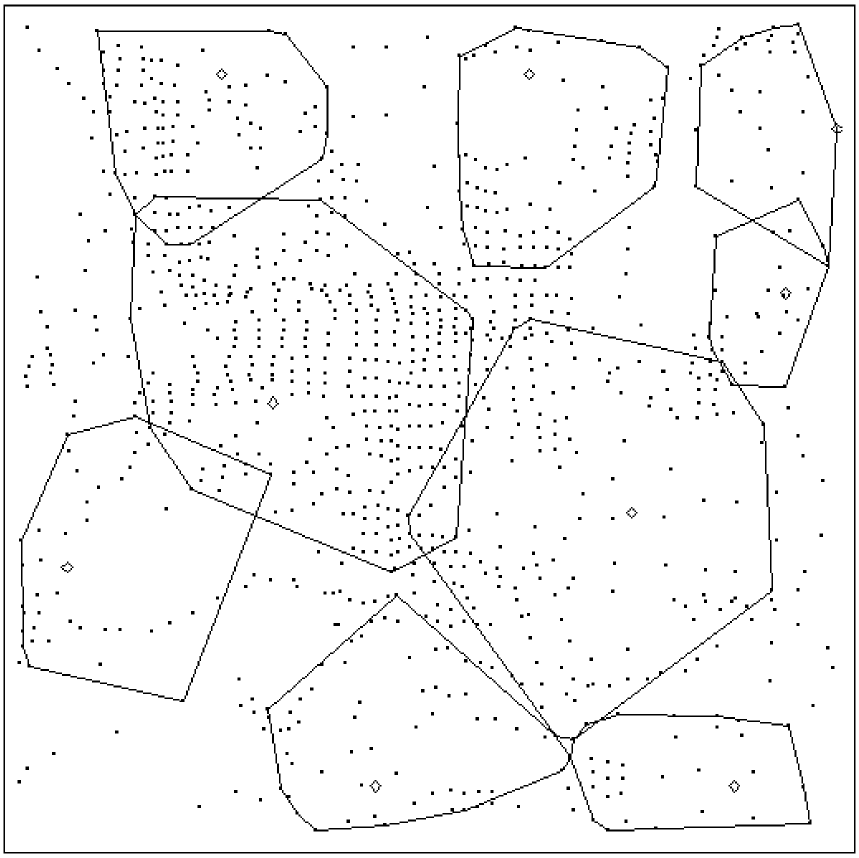


Figure 55: ICEPT planning result: base station locations

3.5 Summary

We presented a new method for the estimation and characterization of the expected teletraffic in mobile communication networks. The method considers the teletraffic from the network's viewpoint. Its traffic estimation is based on the *geographic traffic model*, which obeys the geographical and demographical factors for the demand for mobile communication services. For the spatial teletraffic characterization, a novel representation technique was introduced which uses the notion of discrete *demand nodes*. We demonstrated how the information in geographical information systems, like ATKIS, can be used to estimate the teletraffic demand in a service region and we validated the results with measurements from a real cellular network. Additionally we outlined how the discrete demand node representation enables the application of automatic mobile network design algorithms.

4 Traffic stream descriptors

An important teletraffic research topic is to find traffic descriptors which can support the accurate analysis of network performance. In this section we address the question of relevant arrival process characteristics that accurately determine queueing behaviour and also investigate the generalized peakedness as a measure of traffic burstiness.

4.1 Second order descriptors to characterize MAPs

The motivation for carrying out this work [3] has been the question of what statistical descriptors of the arrival process suffice to make an accurate prediction of queueing behaviour possible. This question has become increasingly interesting from a practical perspective due to the tremendous growth in real-life technical systems with complex non-renewal arrival processes. The relevance of this question has been particularly apparent in the area of modelling communication systems since findings in modern real life networks e.g. [37] and [62] have illustrated arrival process behaviour very different from what has previously been seen in communication networks.

It is well documented in the literature that in general queueing behaviour can NOT be accurately predicted on the basis of first and second order properties of the counts of the arrival process. In [2] and [34] it is shown that arrival processes, created by superposing Interrupted Poisson Processes (IPPs), with fixed first and second order properties of their counting processes can show drastically different queueing behaviour. Particularly illustrative is the work in [8] where it shown that even for two state Markov Modulated Poisson Processes (MMPPs), also known as Switched Poisson Processes (SPPs) [116], drastically different queueing behaviour can be exhibited for arrival processes with the same first and second order properties of the counts.

Sriram and Whitt have in [110] proposed to use properties of the interval process as arrival process descriptors. To provide tools for supplementing this work we in this paper derive expressions for the second order properties of the process of inter-arrival times (the interval process) for a Markovian Arrival Process (MAP) [71]. We derive a general formula for the Index of Dispersion of Intervals (IDI) and give a simplified formula for the interval covariances. Particular formulas for the special cases of the two state MAP and the SPP are given. It is shown that the fundamental rate, IDI and Index of Dispersion of Counts IDC completely characterize the stochastic behaviour of two state MAPs.

We introduce the notion of similar MAPs i.e. MAPs with parameter matrices which are similar. We show that several well known stochastical equivalences of two state MAPs can be expressed by similarity transformations of the MAPs. The valid region for these similarity transformations of 2 state MAPs are given in an appendix.

Finally we compare queueing behaviour of SPPs with the same rate and IDI to assess the meaningfullness of predicting queueing behaviour from first and second order properties of the interval process alone. Also we look further into the queueing behaviour of SPPs with the same rate and IDC drawing in part upon results from [8].

4.1.1 The MAP

The MAP [70], [71] and [96] is a Markov renewal process whose transition probability matrix $\mathbf{F}(*)$ is of the form

$$\mathbf{F}(x) \quad = \int_0^x e^{\mathbf{D}_0 u} du \mathbf{D}_1$$

where the matrices $\mathbf{D}_0 = [D_{0ij}]$ and $\mathbf{D}_1 = [D_{1ij}]$ are respectively a stable matrix and a non-negative matrix whose sum is an irreducible infinitesimal generator $\mathbf{D}$ with stationary probability vector $\vec{\pi}$. The fundamental rate of the process is given as $\lambda^\star = \vec{\pi}\mathbf{D}_1\vec{e}$ where $\vec{e}$ is a column vector of 1s. The interval-stationary probability vector $\vec{\phi}$ is found as the stationary probability vector of $\mathbf{F}(\infty) = (-\mathbf{D}_0)^{-1}\mathbf{D}_1$. It can readily be shown that $\vec{\phi} = (\lambda^\star)^{-1}\vec{\pi}\mathbf{D}_1$.

The MAP descriptors we are interested in here are the first and second order properties of the counting and the interval process. By the interval process we mean the process of succesive inter-arrival times. Throughout the text $N(t)$ will denote the counting process up to time t. By $N^e(t)$ we denote the equilibrium or time stationary version, while $N^\circ(t)$ denotes the interval stationary version.

We will make use of a number of double transforms in the following

$$\psi^e(t) = \int_0^\infty e^{-st} E\left(Z^{N^e(t)}\right) dt = \vec{\pi}\left(s\boldsymbol{I} - \boldsymbol{D}_0 - z\boldsymbol{D}_1\right)^{-1}\vec{e}$$

$$\psi^\circ(t) = \int_0^\infty e^{-st} E\left(Z^{N^\circ(t)}\right) dt = (\lambda^\star)^{-1}\vec{\pi}\boldsymbol{D}_1\left(s\boldsymbol{I} - \boldsymbol{D}_0 - z\boldsymbol{D}_1\right)^{-1}\vec{e}$$

Interval process results The interval process has not yet gained as much interest as the counting process when characterizing MAPs. We will demonstrate that there is a great potential in the interval process with respect to obtaining further insight in the behaviour of point processes. Especially when considering two state MAPs joint second order properties of counts and intervals determine the representation up to stochastical equivalence of the point process.

The IDI for a MAP Let X_n be a stationary sequence of inter-arrival times and let $S_n = \sum_{i=1}^n X_i$. The IDI is defined as $IDI(k) = \frac{(\lambda^\star)^2}{k} Var\{S_k\}$ i.e. as the ratio of the variance of S_n to the corresponding variance in case of a Poisson process [19] p. 71. It should be noted that while the interval stationary process in the context of IDI appears to be the most natural to consider other processes as the time stationary process might be meaningfull to consider. In the latter cases X_1 should be chosen meaningfully e.g. for the time stationary process X_1 should be the time from a random point to the next arrival and hence not an "inter-arrival" time.

Theorem 4.1 *The IDI for an interval stationary MAP - ($\mathbf{D}_0$,$\mathbf{D}_1$) is*

$$IDI(n) = 2\lambda^\star\vec{\pi}(\mathbf{I}+\mathbf{D}_0^{-1}\mathbf{D}_1+\mathbf{\Phi})^{-1}(-\mathbf{D}_0)^{-1}\vec{e}-1$$

$$-\tfrac{2}{n}\lambda^\star\vec{\pi}(\mathbf{I}-(-\mathbf{D}_0^{-1}\mathbf{D}_1)^n)(\mathbf{I}+\mathbf{D}_0^{-1}\mathbf{D}_1+\mathbf{\Phi})^{-2}(-\mathbf{D}_0^{-1}\mathbf{D}_1)(-\mathbf{D}_0)^{-1}\vec{e}$$

with

$$\mathbf{\Phi} = \vec{e}\vec{\phi}$$

The proof of this theorem is given in [3].

Recall $IDC(t)$

$$IDC(t) = 1-2\lambda^\star+\frac{2}{\lambda^\star}\vec{\pi}\mathbf{D}_1(\mathbf{\Pi}-\mathbf{D})^{-1}\mathbf{D}_1\vec{e}-\frac{2}{\lambda^\star t}\vec{\pi}\mathbf{D}_1(\mathbf{I}-e^{\mathbf{D}t})(\mathbf{\Pi}-\mathbf{D})^{-2}\mathbf{D}_1\vec{e}$$

where $\mathbf{\Pi} = \vec{e}\vec{\pi}$.

The fact that the asymptotic value of the IDI and the IDC are the same for a MAP (see e.g. [20] pp. 361 - 362) gives that the following equality must hold :

$$2\lambda^\star\vec{\pi}(\mathbf{I}+\mathbf{D}_0^{-1}\mathbf{D}_1+\mathbf{\Phi})^{-1}(-\mathbf{D}_0)^{-1}\vec{e}-2 =$$

$$\tfrac{2}{\lambda^\star}\vec{\pi}\mathbf{D}_1(\mathbf{\Pi}-\mathbf{D})^{-1}\mathbf{D}_1\vec{e}-2\lambda^\star$$

The two expressions in this interesting equality do not appear to be straightforward to derive from each other.

We have not been able to exploit this fact yet, altough there is a slight chance that some deeper insight could be gained. Most likely, however, the expressions can be derived from each other due to standard matrix analytic arguments.

Covariance function The covariance of intervals in the interval stationary version of the process can then be found as (following [35] p.153 with further simplifications following the lines in the derivation of the IDI).

$$\begin{aligned} Cov\{X_1,X_k\} = &\quad (\lambda^\star)^{-1}\vec{\pi}(((-\mathbf{D}_0^{-1})\mathbf{D}_1)^{k-1}-\mathbf{\Phi})(-\mathbf{D}_0)^{-1}\vec{e} \\ = &\quad (\lambda^\star)^{-1}\vec{\pi}((-\mathbf{D}_0^{-1})\mathbf{D}_1)^{k-1}(-\mathbf{D}_0)^{-1}\vec{e}-(\lambda^\star)^{-2} \end{aligned} \tag{44}$$

Results for the special case of a two state MAP The two state MAP is a 6 parameter model.

$$\boldsymbol{D_0} = \begin{bmatrix} -(r_1+\lambda_{11}+\lambda_{12}) & r_1 \\ r_2 & -(r_2+\lambda_{21}+\lambda_{22}) \end{bmatrix} \quad \boldsymbol{D_1} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} \tag{45}$$

with steady state probability vector $\vec{\pi} = (\frac{r_2+\lambda_{21}}{r_1+r_2+\lambda_{12}+\lambda_{21}}, \frac{r_1+\lambda_{12}}{r_1+r_2+\lambda_{12}+\lambda_{21}})$ and fundamental rate $\lambda^* = \vec{\pi}\boldsymbol{D_1}\vec{e} = \frac{(\lambda_{11}+\lambda_{12})(r_2+\lambda_{21})+(\lambda_{21}+\lambda_{22})(r_1+\lambda_{12})}{r_1+r_2+\lambda_{21}+\lambda_{12}}$.

In the following we use the notation $tr(\boldsymbol{Q})$ and $det(\boldsymbol{Q})$ for respectively the trace and the determinant of the matrix $\boldsymbol{Q}$.

Counting process results for the time stationary process The time stationary variance-time curve can be found as (following [69] p. 501 and with further simplifications).

$$V(t) = \quad (\lambda^\star - 2\delta^{-1}(\lambda^{\star 2} - \vec{\pi}\boldsymbol{D_1}^2\vec{e}))t + 2\delta^{-2}(1 - e^{\delta t})(\lambda^{\star 2} - \vec{\pi}\boldsymbol{D_1}^2\vec{e}) \tag{46}$$

where $\delta = -tr(\boldsymbol{D})$.

Interval process results for the interval stationary process The covariance between inter-arrival times seperated by $k-1$ intervals can be found as

$$Cov(k) = \frac{(\vec{\pi}\boldsymbol{D_1}^2\vec{e} - \lambda^{\star 2})}{\lambda^{\star 2}(\lambda^\star\delta + det(\boldsymbol{D}_1))}\left(\frac{det(\boldsymbol{D}_1)}{\lambda^\star\delta + det(\boldsymbol{D}_1)}\right)^k \tag{47}$$

where again $\delta = -tr(\boldsymbol{D})$.

The index of dispersion of intervals can be found as :

$$IDI(k) = \quad 1 - 2\frac{(\lambda^{\star 2} - \vec{\pi}\boldsymbol{D_1}^2\vec{e})}{\lambda^\star\delta}(1 - \frac{det(\boldsymbol{D}_1)}{k\lambda^\star\delta}(1 - (\frac{det(\boldsymbol{D}_1)}{\lambda^\star\delta + det(\boldsymbol{D}_1)})^k)) \tag{48}$$

During the straight forward but rather tedious derivations of the 2 state MAP results for the process of inter-arrival times it might be convenient to diagonalize the stochastic matrix $(-\boldsymbol{D_0})^{-1}\boldsymbol{D_1}$ as done e.g. in [3].

Remark:
The term $\lambda^\star\delta + det(\boldsymbol{D}_1) = det(\boldsymbol{D}_0)$ so that the geometric term in e.g. $Cov(k) = c_{cov}(\frac{det(\boldsymbol{D}_1)}{det(\boldsymbol{D}_0)})^k$ is the ratio $det(\boldsymbol{D}_1)$ to $det(\boldsymbol{D}_0)$.

It is evident that the first and second order properties of the counting process and the process of interarrival times can be maintained by fixing the 4 quantities: $\lambda^\star$, $\delta = -tr(\boldsymbol{D})$, $\vec{\pi}\boldsymbol{D_1}^2\vec{e}$ and $det(\boldsymbol{D_1})$, in a 2 state MAP. Fixing only the quantities $\lambda^\star$, $\delta = -tr(\boldsymbol{D})$ and $\vec{\pi}\boldsymbol{D_1}^2\vec{e}$ ensures that the first and second order properties of the time stationary counting process are maintained while fixing the quantities $\lambda^\star$, $\frac{(\lambda^\star)^2 - \vec{\pi}\boldsymbol{D_1}^2\vec{e}}{\delta}$ and $\frac{det(\boldsymbol{D_1})}{\delta}$ ensures that the first and second order properties of the interval stationary process are maintained.

4.1.2 Stochastic equivalence

The notion of stochastical equivalence (SE) is an important one when considering point processes generated by different MAPs. Two different MAPs will in general not be identical in any reasonable probabilistic sense when considering an arbitrary initial state. However, for certain combinations of initial conditions the point processes of several MAPs might very well have identical probabilistic behaviour. It appears to be most relevant to consider the time and interval stationary behaviour. The equivalence of the Interrupted Poisson Process (IPP) and the renewal process with an inter-arrival time distribution according to a 2 phased hyper-exponential, H_2, is the classical example of this construction. We give an important definition following the lines of [69].

Definition 4.1 *Two point processes are stochastically equivalent (SE) if for any $n \geq 1$ the joint distribution of the first n intervals agree.*

The joint Laplace-Stieltjes transform of the first n intervals in the interval stationary version of the process is given as

$$\Psi^i(s_1, \cdots, s_n) = (\lambda^\star)^{-1}\vec{\pi}\boldsymbol{D}_1(s_1\boldsymbol{I} - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1 \cdots (s_n\boldsymbol{I} - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1\vec{e}$$

Following the approach in [69] p.507 this expression can, for sufficiently large s_i, be expanded as

$$\Psi^i(s_1, \cdots, s_n) = (s_1 \cdots s_n)^{-1}(\lambda^\star)^{-1}\sum_{k_1=0}^{\infty} \cdots \sum_{k_n=0}^{\infty} \vec{\pi}\boldsymbol{D}_1(\boldsymbol{D}_0)^{k_1}\boldsymbol{D}_1 \cdots (\boldsymbol{D}_0)^{k_n}\boldsymbol{D}_1\vec{e}s_1^{-k_1} \cdots s_n^{-k_n}$$

For two MAPs parameterized by $(\boldsymbol{D}_0^x, \boldsymbol{D}_1^x)$ and $(\boldsymbol{D}_0^y, \boldsymbol{D}_1^y)$ to be stochastically equivalent by analyticity of the transforms we must have that

$$\vec{\pi}^x\boldsymbol{D}_1^x(\boldsymbol{D}_0^x)^{k_1}\boldsymbol{D}_1^x \cdots (\boldsymbol{D}_0^x)^{k_n}\boldsymbol{D}_1^x\vec{e} = \vec{\pi}^y\boldsymbol{D}_1^y(\boldsymbol{D}_0^y)^{k_1}\boldsymbol{D}_1^y \cdots (\boldsymbol{D}_0^y)^{k_n}\boldsymbol{D}_1^y\vec{e}$$

for all $n \geq 0$ and for all $k_1 \geq 0, \cdots, k_n \geq 0$

Here the requirements for stochastical equivalence have been established for the interval stationary version of the point processes. Recalling that $\vec{\pi}\boldsymbol{D}_1 = -\vec{\pi}\boldsymbol{D}_0$ it is easily verified that the interval stationary version of the point processes of two MAPs are stochastically equivalent if and only if the time stationary version of the point processes are stochastically equivalent.

Theorem 4.2 *Two two-state MAPs have the same rate, IDI and IDC if and only if their point processes are stochastically equivalent.*

The proof of this theorem is presented in [3].

Definition 4.2 *Two MAPs with parameter matrices $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ and $(\boldsymbol{S}_0, \boldsymbol{S}_1)$ are similar if there exists a similarity transformation such that $\boldsymbol{S}_0 = \boldsymbol{P}\boldsymbol{D}_0\boldsymbol{P}^{-1}$ and $\boldsymbol{S}_1 = \boldsymbol{P}\boldsymbol{D}_1\boldsymbol{P}^{-1}$, where $\boldsymbol{P}\vec{e} = \vec{e}$.*

Theorem 4.3 *Two similar MAPs have stochastically equivalent interval stationary processes.*

Proof: Consider the two similar MAPs $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ and $(\boldsymbol{S}_0, \boldsymbol{S}_1)$ with the similarity transformation matrix $\boldsymbol{P}$. The invariant probability vector for the time-stationary process $\vec{\pi}_S$ is given by $\vec{\pi}_D\boldsymbol{P}^{-1}$, correspondingly for the interval stationary version $\vec{\phi}_S = \vec{\phi}_D\boldsymbol{P}^{-1}$. By insertion, see definition 3.1, the requirements for stochastical equivalence are readily seen to be fulfilled.

□

We now give some examples of stochastically equivalent point processes derived from similar two state MAPs. In [3] we give the valid region for similarity transformations of a given two state MAP. In the examples we draw upon the results derived in [3].

Example 4.1 *The stochastical equivalence of the IPP and the H_2 renewal process is well known. In a MAP context we can express this by the similarity of the two MAPs*

$$\boldsymbol{D}_0 = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} \mu_1 p & \mu_1(1-p) \\ \mu_2 p & \mu_2(1-p) \end{bmatrix},$$

$$\boldsymbol{S}_0 = \begin{bmatrix} -\frac{(p\mu_1^2+(1-p)\mu_2^2)}{\mu_1 p+\mu_2(1-p)} & \frac{p(1-p)(\mu_1-\mu_2)^2}{\mu_1 p+\mu_2(1-p)} \\ \frac{\mu_1\mu_2}{\mu_1 p+\mu_2(1-p)} & -\frac{\mu_1\mu_2}{\mu_1 p+\mu_2(1-p)} \end{bmatrix}, \qquad \boldsymbol{S}_1 = \begin{bmatrix} \mu_1 p + \mu_2(1-p) & 0 \\ 0 & 0 \end{bmatrix},$$

which are related through $\boldsymbol{D}_i = \boldsymbol{P}^{-1}\boldsymbol{S}_i\boldsymbol{P}, \qquad i = 0, 1$ *with*

$$\boldsymbol{P} = \begin{bmatrix} p & 1-p \\ \frac{-\mu_2}{\mu_1-\mu_2} & \frac{\mu_1}{\mu_1-\mu_2} \end{bmatrix}, \qquad \boldsymbol{P}^{-1} = \begin{bmatrix} \frac{\mu_1}{\mu_1 p+\mu_2(1-p)} & \frac{(\mu_2-\mu_1)(1-p)}{\mu_1 p+\mu_2(1-p)} \\ \frac{\mu_2}{\mu_1 p+\mu_2(1-p)} & \frac{-(\mu_2-\mu_1)p}{\mu_1 p+\mu_2(1-p)} \end{bmatrix}$$

Another similarity transformation transforms the H_2 renewal process into a mixture of Generalized Erlang (GE)-distributions. The MAP $(\boldsymbol{T}_0, \boldsymbol{T}_1)$

$$\boldsymbol{T}_0 = \begin{bmatrix} -\mu_1 & (\mu_1-\mu_2)(1-p) \\ 0 & -\mu_2 \end{bmatrix}, \qquad \boldsymbol{T}_1 = \begin{bmatrix} \mu_1 p + \mu_2(1-p) & 0 \\ \mu_2 & 0 \end{bmatrix},$$

is related with $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ through $\boldsymbol{D}_i = \boldsymbol{Q}^{-1}\boldsymbol{T}_i\boldsymbol{Q}, \qquad i = 0, 1$ *with*

$$\boldsymbol{Q} = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{Q}^{-1} = \begin{bmatrix} \frac{1}{p} & \frac{-(1-p)}{p} \\ 0 & 1 \end{bmatrix}$$

Example 4.2 *In [69] it was shown that for all 2 state MMPPs (or SPPs) point processes there exists a stochastically equivalent Markov Switched Poisson Process (MSPP) point process. The MSPP can be parameterized as follows*

$$\boldsymbol{D}_0 = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} \mu_1 p_1 & \mu_1(1-p_1) \\ \mu_2 p_2 & \mu_2(1-p_2) \end{bmatrix},$$

Also here the stochastical equivalence can be expressed as a similarity transformation. Let the SPP be parameterized as follows

$$\boldsymbol{S}_0 = \begin{bmatrix} -r_1-\lambda_1 & r_1 \\ r_2 & -r_2-\lambda_2 \end{bmatrix}, \qquad \boldsymbol{S}_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

where

$$\lambda_1 = \frac{p_1\mu_1 + (1-p_2)\mu_2 + \sqrt{(p_1\mu_1 - (1-p_2)\mu_2)^2 + 4\mu_1\mu_2(1-p_1)p_2}}{2}$$

$$\lambda_2 = \frac{p_1\mu_1 + (1-p_2)\mu_2 - \sqrt{(p_1\mu_1 - (1-p_2)\mu_2)^2 + 4\mu_1\mu_2(1-p_1)p_2}}{2}$$

$$r_1 = \frac{((1-p_1)\mu_1 + p_2\mu_2)\lambda_1 - ((1-p_1)+p_2)\mu_1\mu_2}{\lambda_1 - \lambda_2}$$

$$r_2 = \frac{((1-p_1)+p_2)\mu_1\mu_2 - ((1-p_1)\mu_1 + p_2\mu_2)\lambda_2}{\lambda_1 - \lambda_2}$$

Then $\boldsymbol{D}_i = \boldsymbol{Q}^{-1}\boldsymbol{S}_i\boldsymbol{Q}, \qquad i = 0,1$ *with*

$$\boldsymbol{Q} = \begin{bmatrix} q_1 & 1-q_1 \\ q_2 & 1-q_2 \end{bmatrix}, \qquad \boldsymbol{Q}^{-1} = \frac{1}{q_1 - q_2}\begin{bmatrix} 1-q_2 & -(1-q_1) \\ -q_2 & q_1 \end{bmatrix}$$

where

$$q_1 = \frac{(1+p_2)\mu_2 - p_1\mu_1 + \sqrt{(p_1\mu_1 - (1-p_2)\mu_2)^2 + 4\mu_1\mu_2(1-p_1)p_2}}{2(\mu_2 - \mu_1)}$$

$$q_2 = \frac{(1+p_2)\mu_2 - p_1\mu_1 - \sqrt{(p_1\mu_1 - (1-p_2)\mu_2)^2 + 4\mu_1\mu_2(1-p_1)p_2}}{2(\mu_2 - \mu_1)}$$

Remarks:

- *As noted in [69] a stochastically equivalent MSPP point process for a given SPP point process can always be found while the converse is true only when the MSPP satisfies :* $p_1(1-p_2) \geq (1-p_1)p_2$.
- *With the parameterization used in this example it is clear that the stochastical equivalence of MSPP and SPP point processes can be seen as a generalization of the stochastical equivalence between the IPP and* H_2 *renewal process in the previous example. This special case is clearly obtained by setting* $p_1 = p_2 = p$.

Similarity is somewhat attractive since it is relatively easy to check whether two matrices are similar or not. On the other hand it is clear that similarity in general is not a necessary condition for stochastical equivalence of the interval stationary point processes induced by MAPs. The following example demonstrates the somewhat surprising result that even within the class of two state MAPs similarity is not a necessary condition in general.

Example 4.3 *If we consider the following two MAPs with stochastically equivalent stationary point processes*

$$\boldsymbol{D}_0 = \begin{bmatrix} -26.001600 & 2.159218 \\ 0.919591 & -16.998400 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} 16.001600 & 7.840782 \\ 4.080409 & 11.998400 \end{bmatrix},$$

$$\boldsymbol{D} = \begin{bmatrix} -10.000000 & 10.000000 \\ 5.000000 & -5.000000 \end{bmatrix}$$

$$\boldsymbol{S}_0 = \begin{bmatrix} -26.000000 & 2.000000 \\ 1.000000 & -17.000000 \end{bmatrix}, \qquad \boldsymbol{S}_1 = \begin{bmatrix} 16.000000 & 8.000000 \\ 4.000000 & 12.000000 \end{bmatrix},$$

$$\boldsymbol{S} = \begin{bmatrix} -10.000000 & 10.000000 \\ 5.000000 & -5.000000 \end{bmatrix}$$

It is easily verified that the two MAPs are not similar. We get

$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{-1}, \qquad \boldsymbol{S}_0 = \boldsymbol{Q}_0\boldsymbol{D}_0\boldsymbol{Q}_0^{-1}, \qquad \boldsymbol{S}_1 = \boldsymbol{Q}_1\boldsymbol{D}_1\boldsymbol{Q}_1^{-1}$$

with

$$\boldsymbol{Q} = \begin{bmatrix} 1.000000 & 0.000000 \\ 0.000000 & 1.000000 \end{bmatrix}, \qquad \boldsymbol{Q}_0 = \begin{bmatrix} 0.000000 & 0.222183 \\ 0.102158 & 1.000000 \end{bmatrix},$$

$$\boldsymbol{Q}_1 = \begin{bmatrix} 0.000000 & -1.999200 \\ -1.019694 & 1.000000 \end{bmatrix}$$

4.1.3 What do the IDI respectively the IDC tell us about queueing behaviour?

Sriram and Whitt [110] have suggested that the IDI might be a better overall characterization of a point process than the IDC with respect to queueing behaviour. Partly due to the lack of easily implementable formulae this track has not been extensively pursued. With the derivation of the IDI for the MAP we are able to perform some introductory steps in this direction. We have choosen the SPP (the special case of a two state MMPP) as one of the simplest non-renewal processes.

SPP results In this subsection we specialize general MAP results to the SPP. In MAP notation a SPP can be written as follows

$$\boldsymbol{D_0} = \begin{bmatrix} -(r_1+\lambda_1) & r_1 \\ r_2 & -(r_2+\lambda_2) \end{bmatrix} \quad \boldsymbol{D_1} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{49}$$

SPP results for the counting process The time-stationary covariance between the number of events in two timeslots of size Δt with $k-1$ timeslots between them is expressed by $(k > 0)$.

$$\gamma(k) = \tfrac{(\lambda_1-\lambda_2)^2 r_1 r_2 e^{-((r_1+r_2)(k-1)\Delta t)}}{(r_1+r_2)^4}(1 - 2e^{-((r_1+r_2)\Delta t)} + e^{-((r_1+r_2)2\Delta t)}) \tag{50}$$

The time-stationary variance time curve is given by (see e.g. [49])

$$\sigma^2(t) = (\tfrac{\lambda_1 r_2+\lambda_2 r_1}{r_1+r_2} + 2\tfrac{r_1 r_2(\lambda_1-\lambda_2)^2}{(r_1+r_2)^3})t - 2\tfrac{r_1 r_2(\lambda_1-\lambda_2)^2}{(r_1+r_2)^4}(1 - e^{-t(r_1+r_2)}) \tag{51}$$

SPP results for the interval process The results for the interval process are readily obtained from the two state MAP results in section 4.1.1.

The covariance of the SPP interval process X_n can in the interval-stationary version be found as

$$Cov\{X_1, X_k\} = \frac{r_1 r_2(\lambda_1-\lambda_2)^2}{(\lambda_1\lambda_2+\lambda_1 r_2+\lambda_2 r_1)(r_1\lambda_2+r_2\lambda_1)^2}\left(\frac{\lambda_1\lambda_2}{\lambda_1\lambda_2+r_1\lambda_2+r_2\lambda_1}\right)^{k-1},\ k>1 \tag{52}$$

The IDI for an SPP can be found as :

$$IDI(k) = 1+2\frac{r_1 r_2(\lambda_1-\lambda_2)^2}{(r_1\lambda_2+r_2\lambda_1)(r_1+r_2)^2}\left(1-\frac{\lambda_1\lambda_2}{k(\lambda_1 r_2+\lambda_2 r_1)}\left(1-\left(\frac{\lambda_1\lambda_2}{\lambda_1\lambda_2+r_1\lambda_2+r_2\lambda_1}\right)^k\right)\right) \tag{53}$$

Simple queueing experiments fixing the IDI respectively the IDC

Experiments with SPPs with fixed rate and IDC With four parameters available it is possible to find a continuum of SPPs with same rate and IDC. This has been done by Berger [8]. Following his approach it can be seen from formula (51) that fixing $\lambda^\star$, $k_1 = \frac{r_1 r_2(\lambda_1-\lambda_2)^2}{r_1+r_2}$ and $k_2 = r_1 + r_2$ a continuum of SPPs can be created by varying the ratio $r_{rat} = \frac{r_1}{r_2}$. It can be derived [8] that for a given ratio, λ_1 and λ_2 can be determined as follows.

$$\lambda_1 = \lambda^\star + \sqrt{\frac{k_1 r_{rat}}{k_2}}\ ,\ \lambda_2 = \lambda^\star - \sqrt{\frac{k_1}{k_2 r_{rat}}} \tag{54}$$

Starting with an IPP with $\lambda^\star = 0.3$ parameterized as $r_1 = 0.00074142$, $r_2 = 0.001779411$ and $\lambda_1 = 0.42499978257$ ($\lambda_2 = 0$) (same as in [8]) a continuum can be constructed. Here we consider ratios r_{rat}= 1,10,20,50,100,1000,10000, 100000. Here and in the remaining part of this paper we examine the tail probabilities in a single server queue with constant service time when assessing queueing behaviour. All queueing experiments are done with load $\rho = 0.3$.

For a general description of algorithms for calculating performance measures in the MAP/G/1 queue see e.g. [70] or [96].

Examining the tail probabilities arising with the different SPP arrival processes, figure 56, it is evident that trying to predict queueing behaviour on the basis of the rate and IDC is impossible.

Looking at formulae (54) it is clear that for every non-trivial SPP ($\lambda_1 \neq \lambda_2$), λ_1 can be made arbitrarily large for a model with same rate and IDC which makes it certain that queueing behaviour which conditionally overloads the queue always can be constructed since the intensity out of state 1 is bounded by $k_2 = r_1 + r_2$. However, it is also evident that the other intensity parameter $\lambda_2 \to \lambda^\star$ as $r_{rat} \to \infty$ and that the relative portion of the traffic that arrives in state 2 tends to 1 i.e. $\frac{\lambda_1 r_2}{\lambda_2 r_1} \to 0$ as $r_{rat} \to \infty$. These observations agree well with the queueing results in figure 56. As $r_{rat} \to \infty$ the slope of the tail becomes arbitrarily close to 0, however, simultaneously the asymptotic constant becomes smaller. It is thus certain that adverse queueing behaviour

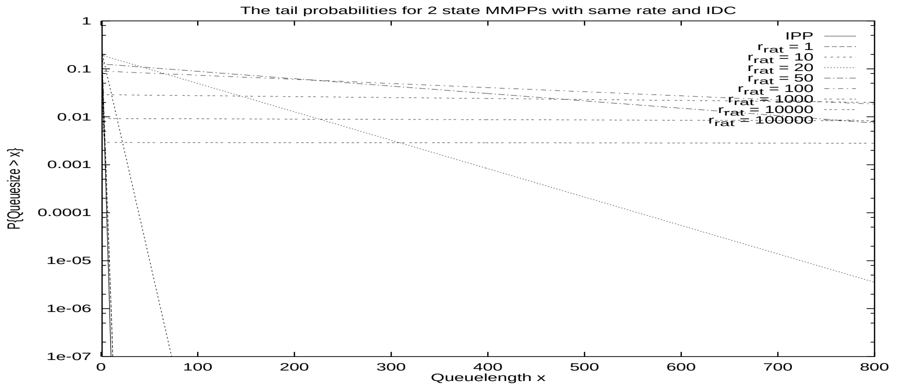


Figure 56: **Probability that the queue length exceeds x for a number of SPPs with fixed rate and IDC**

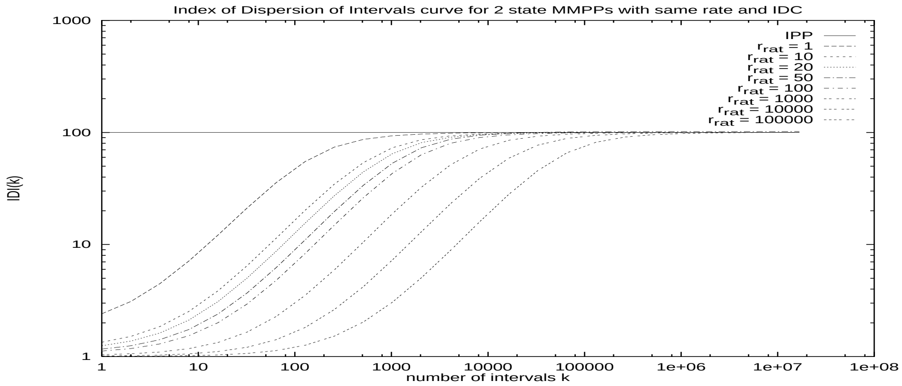


Figure 57: **IDI for a number of SPPs with fixed rate and IDC**

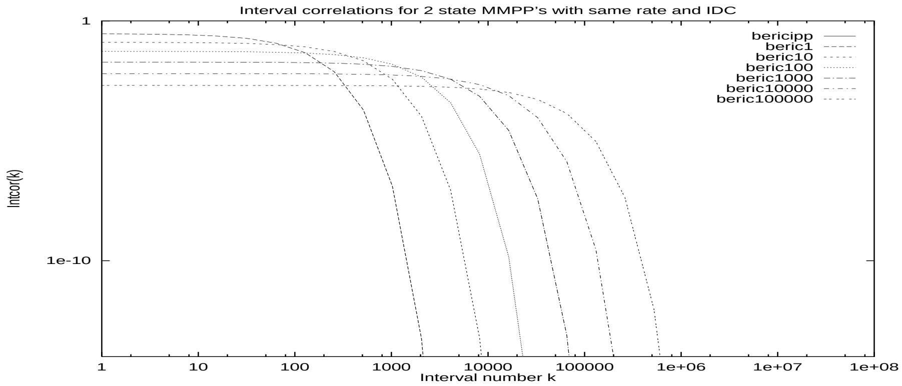


Figure 58: **Interval Correlations for a number of SPPs with fixed rate and IDC**

can be created, depending on the parameters it might, however, only affect such a small portion of the total traffic that it can be neglected e.g. an asymptotic constant below e.g. 10^{-9} when the slope is close to 0.

In figures 57 and 58 we display the IDI respectively the correlations in the interval process for the different models with same rate and IDC. Somewhat surprising perhaps it can be seen from figure 57 that it is NOT necessarily the model which reaches the asymptotic IDI value first which displays the worst queueing behaviour. From figure 58 it is evident that it is not only the magnitude of the correlations that are of significance but also the number of intervals over which they persist. In [110] and references therein it has been suggested to approximate non-renewal processes with a renewal process using a squared coefficient of variation found as a weighted sum of the IDI. From the results presented here it is evident that one should be very careful when employing such an approximation.

Experiments with SPPs with fixed rate and IDI Given the analytic expression for the IDI, formula (53), we look at different SPPs with fixed rate and IDI.

From formula (53) it can be seen that fixing $\lambda^\star$, $k_1 = \frac{r_1 r_2(\lambda_1-\lambda_2)^2}{(r_1+r_2)^3}$ and $k_2 = \frac{\lambda_1\lambda_2}{r_1+r_2}$ a continuum of SPPs can be created by varying the ratio $r_{rat} = \frac{r_1}{r_2}$.

$$\begin{aligned}
\lambda_1 &= \lambda^\star \frac{1+r_{rat}}{1+ar_{rat}} \\
\lambda_2 &= a\lambda^\star \frac{1+r_{rat}}{1+ar_{rat}} \\
r_1 &= \frac{ar_{rat}(\lambda^\star)^2(1+r_{rat})}{k_2(1+ar_{rat})^2} \\
r_2 &= \frac{a(\lambda^\star)^2(1+r_{rat})}{k_2(1+ar_{rat})^2} \\
&\text{where} \\
a &= 1+\frac{k_1(1+r_{rat})^2}{2k_2r_{rat}} - \sqrt{(\frac{k_1(1+r_{rat})^2}{2k_2r_{rat}})^2 + \frac{k_1(1+r_{rat})^2}{k_2r_{rat}}}
\end{aligned} \tag{55}$$

For comparison we choose the same rate and asymptote as above i.e. $\lambda^\star = 0.3$ and $\frac{k_1}{\lambda^\star} = 99.2$. Additionally we arbitrarily choose $k_2 = 1.2$. As in the previous case a continuum of SPPs can now readily be constructed. Here we consider ratios r_{rat}= 1,2,5,10,100,1000,10000, 100000.

From figure 59 it is evident that it also very difficult to predict queueing behaviour solely on the basis of the rate and IDI. However, as $r_{rat} \to \infty$ there seems to be convergence towards a worst case queueing behaviour which is unlike the case with fixed rate and IDC. Looking at figure 60 it is evident that the worst queueing behaviour is experienced for models which reach the asymptotic IDC value first. Looking at figure 61 it can be noted that for the counting process apparently short range correlation of large magnitude can be worse than long range correlations of low magnitude. This seems to be in contrast with claims that long range correlation of low magnitude are dominant with respect to queueing behaviour. From the formulae (55) it is possible to find theoretical support for these observations. Using a Taylor expansion and letting $r_{rat} \to \infty$ the following limit results can readily be found.

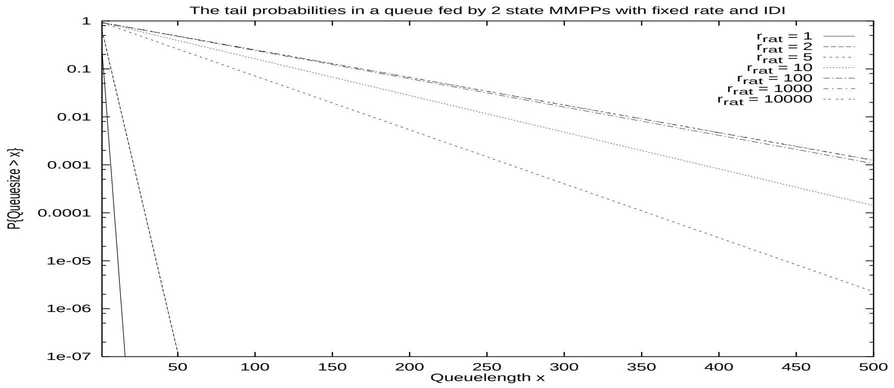


Figure 59: **Probability that the queue length exceeds x for a number of SPPs with fixed rate and IDI**

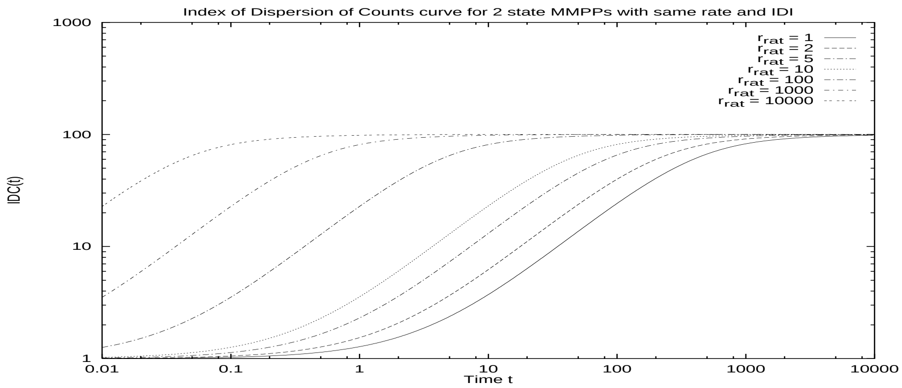


Figure 60: **IDC for a number of SPPs with fixed rate and IDI**

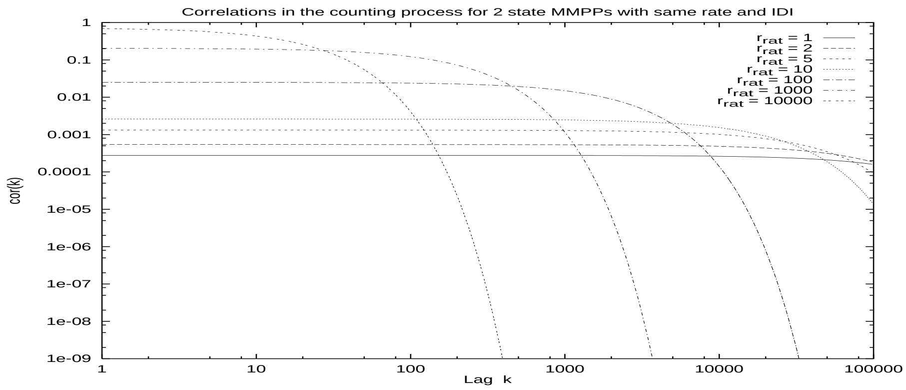


Figure 61: **Correlations in counting process (timewindow = 0.001) for a number of SPPs with fixed rate and IDI**

$$
\begin{aligned}
\lambda_2 &\to \frac{\lambda^\star k_2}{k_2+k_1} \quad \text{as } r_{rat}\to\infty \\
r_2 &\to \frac{(\lambda^\star)^2 k_1}{(k_2+k_1)^2} \quad \text{as } r_{rat}\to\infty \\
\frac{\lambda_1}{r_1} &\to \frac{k_1+k_2}{\lambda^\star} \quad \text{as } r_{rat}\to\infty
\end{aligned}
\tag{56}
$$

Additionally $\lambda_1 \to \infty$ as $r_{rat} \to \infty$ because $a \to 0$ as $r_{rat} \to \infty$. It is well known that for a SPP the number of arrivals during the sojourn time in a given state is geometrically distributed. For state i the probability of j arrivals, $p_i(j)$, is given by $p_i(j) = \frac{r_i}{r_i+\lambda_i}(\frac{\lambda_i}{r_i+\lambda_i})^j$ where r_i denotes the departure intensity from state i and λ_i the arrival intensity. Recalling the third limit result in formulas (56) it is clear that the number of arrivals during a sojourn time in state 1 tends to a fixed distribution as $r_{rat} \to \infty$ i.e. the following $p_1(j) = \frac{\lambda^\star}{\lambda^\star+k_1+k_2}(\frac{k_1+k_2}{\lambda^\star+k_1+k_2})^j$.

Since $r_1 \to \infty$ as $r_{rat} \to \infty$ it is evident that the total arrival process tends to a batch Poisson arrival process with batch arrival intensity $\lambda_B = \lambda^\star \frac{k_2+\lambda^\star}{k_2+k_1+\lambda^\star}$ and batch size distribution

$$
\begin{aligned}
p(1) &= (1-q) + q\frac{\lambda^\star}{\lambda^\star+k_1+k_2} \\
p(j) &= q\frac{\lambda^\star}{\lambda^\star+k_1+k_2}(\frac{k_1+k_2}{\lambda^\star+k_1+k_2})^{j-1} \text{ for } j \geq 2
\end{aligned}
$$

where $q = \frac{\lambda^\star k_1}{(k_2+k_1)(k_2+\lambda^\star)}$. Clearly the z transform of the batch size distribution $P(z)$ can be written as

$$
P(z) = (1-q)z + q\frac{\lambda^\star z}{\lambda^\star + (k_1+k_2)(1-z)} \tag{57}
$$

Approximating the tail of the queue seen by arrivals with the asymptotic tail yields $P\{Q_a > j\} \approx \alpha\sigma^j$. For a batch Poisson arrival process it is relatively easy to investigate the asymptotic tail behaviour of the queue using e.g. the results from [1]. First the asymptotic decay rate σ is determined from the following equation ([1] p. 119).

$$
\lambda_B(1-P(\sigma^{-1})) = \eta \text{ and } \phi(\eta) = \sigma^{-1} \tag{58}
$$

where $\phi(\eta) = E\{e^{\eta V}\}$ is the generating function for the service time distribution and η denotes the to σ corresponding exponential tail of the waiting time distribution ($0 < \sigma < 1$) and ($0 < \eta < \infty$). Without loss of generality we assume that the arrival process and service time distribution are scaled so that the mean service time is 1. The asymptotic constant α is then readily determined from :

$$
\alpha = \frac{(1-\rho)\lambda_B(1-P(0))}{\rho(\lambda_B P'(\sigma^{-1})\phi'(\eta) - 1)} \tag{59}
$$

where ρ denotes the load.

Remark:
We have here used the covariance of the number of arrivals in a slot of some size Δt formula (52). Due to the arbitrary size of the slot Δt it is not without problems to interpret this descriptor. It turns out that the curves in figure 61 are quite sensitive to the choice of Δt. From a theoretical perspective it would thus be more obvious to focus on the correlation of the rates. In applications, however, it is straightforward to estimate the correlation function of counts while estimation of the rate is a delicate topic.

4.2 Peakedness characterization

In this section we focus on peakedness as one of the most promising candidate measures of traffic burstiness [87, 88].

The simplest burstiness measures take only the first-order properties of the traffic into account. A set of candidates are the moments of the inter-arrival time distribution. In practice the peak to mean ratio and the squared coefficient of variation are the most frequently used first-order measures [98, 86].

Measures expressing second-order properties of the traffic are more complex. The autocorrelation function, the indices of dispersion [110, 47] and the generalized peakedness [26, 27] are the most well known measures from this class.

Moreover, there are a number of burstiness measures based on different concepts, e.g. we can use burst length measures [98, 111] or parameters of a leaky bucket for burstiness characterization [85]. By the concept of self-similarity the Hurst parameter and other fractal parameters are also candidates for burstiness measures [89, 62].

In this section we review the theory of generalized peakedness and further develop the basic concept by introducing the generalized peakedness in discrete time. The advantage of this approach is that it allows us to apply the general framework of peakedness for traffic engineering. We provide the computation of peakedness for a number of important discrete time models including the Markov modulated batch Bernoulli process and the batch renewal process. The relationship between IDC and peakedness is also presented. We discuss the challenges of measuring peakedness in practice. Moreover, we show a technique how Markov modulated traffic models can be fitted to a measured peakedness curve. Finally, the practical applicability of peakedness and our modeling technique are demonstrated by examples based on measured MPEG video, aggregated ATM and Ethernet traffic.

4.2.1 Peakedness measures

Peakedness of a traffic stream has been found a useful characterization tool in blocking approximations and in trunking theory [48]. It has been defined as the variance to mean ratio of the number of busy servers in an infinite hypothetical group of servers to which the traffic is offered, where the service times of the servers are independent and exponentially distributed with a common parameter.

Generalized peakedness Eckberg [26] extended this definition by allowing arbitrary service time distribution and defined *generalized peakedness* as a functional which maps holding time distributions into peakedness values. For a given complementary holding time distribution $F^c(x) = \mathrm{P}\{\text{holding time} > x\}$, Eckberg defines the peakedness functional $z\{F^c\}$ as the variance to mean ratio of the number of busy servers in a hypothetical infinite group of servers with independent holding times distributed according to F^c. The general definition provides a way to characterize the variability of an arrival stream with respect to a given service system.

Let us have a stationary arrival process S in continuous time with counting function $N(t) =$ the number of arrivals in $(0, t]$ for $t \geq 0$. The mean arrival intensity is denoted by $m = \mathbf{E}\{N(t)\}/t$, which is independent of t due to the stationarity of S.

Arrivals are allowed to come in batches of random size B. We define the batchiness parameter as $b = \mathbf{E}\{B^2\}/\mathbf{E}\{B\}$ which can be shown to be the mean size of a batch that an arbitrary arrival finds itself in. The differential process [19] $\Delta N(t)$ is defined for a fixed Δt as the number of arrivals in $(t, t + \Delta t]$, that is, $N(t + \Delta t) - N(t)$. We define the covariance density of the arrival process $k(s)$ for $s > 0$ as the covariance of the differential process as Δt goes to zero: $k(s) = \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\{\Delta N(t), \Delta N(t+s)\}}{(\Delta t)^2}$ which is independent of t due to the stationarity of S. For $s < 0$ we let $k(s) = k(-s)$.

We offer the arrival process S to an infinite server group where the service times are independent and have a complementary holding time distribution of $F^c(x)$ ($x \geq 0$; for $x < 0$, we define $F^c(x) = 0$), mean holding time of $1/\mu = \int_{-\infty}^{\infty} F^c(x)dx$ where μ is the service rate, and finally the autocorrelation of F^c is $\rho_{F^c}(x) = \int_{-\infty}^{\infty} F^c(s)F^c(s + x)ds$.

Denoting the number of busy servers at time t by $L(t)$, the generalized peakedness functional is defined as

$$z\{F^c\} = \frac{\mathrm{Var}\, L(t)}{\mathbf{E}\{L(t)\}}. \tag{60}$$

If the arrival stream is defined for the whole time axis $(-\infty, \infty)$, it is independent of t due to the stationarity of S. In practice, we never have an arrival process for an infinitely long time; in this case, we have to define the peakedness for a t which is large enough for the initial transient period in the service system to be negligible. (More precisely, $z\{F^c\} = \lim_{t \to \infty} \mathrm{Var}\, L(t)/\mathbf{E}\{L(t)\}$.)

With the notation introduced above, the peakedness of the arrival stream can be expressed in terms of the covariance density function as [26]

$$z\{F^c\} = 1 + \frac{\mu}{m} \int_{-\infty}^{\infty} (k(s) - m\delta(s))\rho_{F^c}(s)ds \tag{61}$$

where $\delta(s)$ is the Dirac delta function.

The important case of exponential service time simplifies to

$$z_{\exp}(\mu) = \frac{b + 1}{2} + \frac{1}{m} k^*(\mu) \tag{62}$$

where $k^*(\mu) = \int_{0+}^{\infty} k(s)e^{-\mu s}ds$, the Laplace transform of the covariance density function. Here we have the peakedness of a given arrival stream as a function of the service rate μ.

It is shown [26] (and is suggested by (62)) that the peakedness function $z_{\exp}(\mu)$ together with m determines $k(s)$ and therefore the pair $(z_{\exp}(\mu), m)$ is a complete second order characterization of the arrival process.

The peakedness function $z_{\exp}(\mu)$ can be used to compute the peakedness functional for a large class of holding time distributions as shown in [26]. The method is elaborated in [84] to give the peakedness functional for Coxian holding time distributions. The importance of Coxian holding times lies in the fact that any holding time distribution can be approximated with arbitrary accuracy by Coxian distributions. Eckberg also investigated the application of generalized peakedness in delay systems [27]. Eckberg's definition of generalized peakedness for point processes has been extended in [80, 81] to allow fluid flow models given by a rate function.

Peakedness in discrete time In order to use the peakedness measures in a B-ISDN framework, we now extend the peakedness concept for discrete time arrival streams.

We use the following notation: $w[i]$ is the number of arrivals at epoch i, where $i = \ldots -1, 0, 1, \ldots$. We assume the stationarity of $w[i]$. The first and second moments of $w[t]$ (independent of t) are denoted by m_1 and m_2. The covariance density of continuous time is replaced here by the autocovariance function $k[s] = \mathrm{Cov}\{w[i], w[i+s]\} = k[-s]$. (It is seen that $k[0] = m_2 - m_1^2$.)

The service time random variable T is also discrete and has the distribution $t[1], t[2], \ldots$ on positive integers. (It cannot take on zero value.) $\mu = 1/\mathbf{E}\{T\}$ is again the service rate, and it is easily shown that $1/\mu = \mathbf{E}\{T\} = \sum_{s=-\infty}^{\infty} F^c[s]$ where $F^c[x]$ is the complementary holding time distribution function: $F^c[x] = \sum_{u=x+1}^{\infty} t[u] = \mathrm{P}\{T > x\}$ if $x \geq 0$ and $F^c[x] = 0$ if $x < 0$. The autocorrelation function is now $\rho_{F^c}[x] = \sum_{s=-\infty}^{\infty} F^c[s]F^c[s+x]$. It is seen that $\rho_{F^c}[0] = \sum_{s=-\infty}^{\infty} (F^c)^2[s]$.

The traffic is offered to an infinite group of servers with independent identically distributed service times determined by $F^c[x]$. Each arrival takes a separate server. The peakedness of the arrival stream is defined as the variance to mean ratio of the number of busy servers in the infinite server group:

$$z\{F^c\} = \frac{\mathrm{Var}\, L[t]}{\mathbf{E}\{L[t]\}} \tag{63}$$

where $L[t]$ is the number of busy servers at time epoch t.

An important modification of the definition is to let the service time depend on the arrival epoch only (have a common service time for all $w[t]$ arrivals at epoch t). We call (in accordance with [81]) the peakedness value defined in this way the *modified* peakedness $\tilde{z}\{F^c\}$. As we have shown [83],

$$\tilde{z}\{F^c\} - z\{F^c\} = \left(\frac{m_2}{m_1} - 1\right)(1 - \rho_{F^c}[0]\mu). \tag{64}$$

that is, their difference is constant (cf. (35) in [81]). The first factor in the difference is zero if and only if the arrival stream has no simultaneous arrivals, the second factor is zero if and only if the holding time distribution is deterministic.

The importance of this modified definition lies in the fact that it gives a way to handle a whole batch of arrivals together, which can save a lot of computational effort in the case of measuring the peakedness for a general holding time distribution. However, in the case of geometric service times, the original definition of peakedness is easier to measure as shown in section 4.2.2. We will use the original definition of peakedness (Equation (63)) below.

We can express peakedness in terms of the autocovariance function $k[s]$ similarly to (61) as

$$z\{F^c\} = 1 + \frac{\mu}{m_1} \sum_{s=-\infty}^{\infty} \rho_{F^c}[s](k[s] - m_1\delta[s]). \tag{65}$$

The most important case in discrete time is the case of geometrically distributed holding times: $t[i] = \mu(1-\mu)^{i-1}$, $0 < \mu < 1$ (with $\mathbf{E}\{T\} = 1/\mu$ which justifies the notation).

In order to simplify the formulas, let us introduce the notation

$$K[s] = \begin{cases} \frac{2}{m_1}k[s] & \text{if } s > 0 \\ \frac{1}{m_1}k[0] & \text{if } s = 0 \end{cases}$$

and let its z-transform be $K^*(\omega) = \sum_{s=0}^{\infty} K[s]\omega^s$.

The peakedness function of the arrival stream with respect to geometric holding time distribution, as we derived in [83], is given by

$$z_{\text{geo}}(\mu) = 1 + \frac{K^*(1-\mu) - 1}{2-\mu} \tag{66}$$

Peakedness and IDC The widely used measure to characterize the variability of an arrival stream on different time scales is the index of dispersion for counts (IDC). It is defined as $I[t] = \frac{V[t]}{E[t]} = \frac{V[t]}{m_1 t}$ where $E[t]$ and $V[t]$ are the mean and variance of the number of arrivals in t consecutive epochs ($t = 1, 2, \ldots$).

The connection of IDC and peakedness for geometric holding times is, as we have shown [83]

$$z_{\text{geo}}(\mu) = 1 + \frac{\mu^2 \frac{d}{d\omega} I^*(\omega)|_{\omega=1-\mu} - 1}{2-\mu} \tag{67}$$

where $I^*(\omega)$ is the z-transform of $I[t]$.

We can use (67) to get asymptotic results which connect them [83]:

$$z_{\text{geo}}(0) = \frac{\lim_{s\to\infty} I[s] + 1}{2}, z_{\text{geo}}(1) = I[1] = \frac{\operatorname{Var} w[i]}{\mathbf{E}\{w[i]\}} \tag{68}$$

where the first equation is derived using the L'Hospital rule and the final value theorem, whereas the second equation is derived by the initial value theorem for $\frac{d}{d\omega}I^*$.

Peakedness of traffic models Next, we present the peakedness results for important traffic models. We consider discrete time models for the number of arrivals in consecutive epochs.

Batch Bernoulli process A very simple type of arrival stream model is the model with the number of arrivals in a time epoch be independent identically and generally distributed with mean m_1 and second moment m_2.

In this case, $k[i] = 0$ for all $i > 0$. Thus, $K^*(1-\mu) = K[0] = \frac{\mathrm{Var}\, w[i]}{\mathbf{E}\{w[i]\}}$ and $z_{\mathrm{geo}}(\mu) = 1 + \frac{\frac{\mathrm{Var}\, w[i]}{\mathbf{E}\{w[i]\}} - 1}{2-\mu}$ For the special case of Poisson batch arrivals, the distribution of arrivals in an epoch is Poissonian, thus $\frac{\mathrm{Var}\, w[i]}{\mathbf{E}\{w[i]\}} = 1$ which gives $z_{\mathrm{geo}}(\mu) = 1$.

The Poisson process can be considered as a reference process with respect to peakedness characterization. Batch arrival processes that are more bursty than the Poisson process have higher peakedness values, smoother processes have lower peakedness. (In the case of deterministic traffic, $z_{\mathrm{geo}}(\mu) = 1 - \frac{1}{2-\mu}$.)

Markov modulated batch Bernoulli process A very general Markovian process is the Markov modulated batch Bernoulli process (MMBBP). In this model, we have a discrete time Markov process as a modulating process. In each state of the modulating Markov-process, batch arrivals are generated according to a general distribution corresponding to the state.

Let $\mathbf{P}$ and $\mathbf{D}$ denote the transition probability matrix and the steady-state distribution vector of the modulating Markov process, respectively ($\mathbf{DP}=\mathbf{D}$). Let $\mathbf{M_1}$ and $\mathbf{M_2}$ be diagonal matrices corresponding to the first and second moments of the number of arrivals in the corresponding states. Let $\mathbf{e}$ be a vector of all ones and let $\mathbf{I}$ be the identity matrix.

We can express the mean number of arrivals as $m_1 = \mathbf{DM_1e}$ and the second moment as $m_2 = \mathbf{DM_2e}$. The autocovariance function of the arrival process is given by $k(i) = \mathbf{DM_1P}^i\mathbf{M_1e} - m_1^2$.

Using (66) we have derived the peakedness function as [83]

$$z_{\mathrm{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{2(1-\mu)\mathbf{DM_1P}(\mathbf{I}-(1-\mu)\mathbf{P})^{-1}\mathbf{M_1e} + m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \qquad (69)$$

A very important case of MMBBP is the Markov modulated Bernoulli process (MMBP); its peakedness curve is the special case of (69).

Markov modulated Bernoulli process As a special case of MMBBP, when the arrival process is Bernoulli in each state, we have a Markov modulated Bernoulli process. If the parameter of the Bernoulli process is p_i in state i, we have $\mathbf{M_1} = \mathrm{diag}\,[(]\, p_1, p_2, \ldots ...)$ and $\mathbf{M_2} = \mathbf{M_1}$.

The peakedness curve for geometrical holding times in this case is

$$z_{\mathrm{geo}}(\mu) = 1 + 2\frac{(1-\mu)\mathbf{DM_1P}(\mathbf{I}-(1-\mu)\mathbf{P})^{-1}\mathbf{M_1e}}{(2-\mu)m_1} - \frac{m_1}{\mu} \qquad (70)$$

Switched batch Bernoulli process Another important special case of MMBBP is the 2-state MMBBP (SBBP, switched batch Bernoulli process). Let us use the following notation: the transition matrix is $\mathbf{P} = \begin{bmatrix} 1-\alpha_1 & \alpha_1 \\ \alpha_2 & 1-\alpha_2 \end{bmatrix}$ and the steady state distribution is thus $\mathbf{D} = \frac{1}{\alpha_1+\alpha_2}(\alpha_2\ \alpha_1)$.

Denote $\gamma = 1-\alpha_1-\alpha_2$. In state 1, the first and second moments of the number of arrivals are $m_{1,(1)}$ and $m_{1,(2)}$, respectively; in state 2, the moments are $m_{2,(1)}$ and $m_{2,(2)}$.

The first and second moments of the number of arrivals are given by $m_1 = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(1)} + \alpha_1 m_{2,(1)})$, $m_2 = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(2)} + \alpha_1 m_{2,(2)})$. Let us also introduce the notation $m_* = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(1)}^2 + \alpha_1 m_{2,(1)}^2)$. Note that if the distribution of the batch size in a given state is deterministic, or if it is geometric or Bernoulli, we have $m_{i,(1)}^2 = m_{i,(2)}$ $(i=1,2)$ and thus $m_* = m_2$. If the batch distribution is Poisson, we have $m_* + m_1 = m_2$.

Using (69) and the possibility to explicitly compute the inverse of $\mathbf{I}-(1-\mu)\mathbf{P}$ in the 2-state case, we get

$$z_{\text{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{2}{m_1}\frac{(1-\mu)}{\mu}\left[m_* - \frac{(m_*-m_1^2)(1-\gamma)}{1-\gamma(1-\mu)}\right] + \frac{m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \tag{71}$$

and by (66) we get the peakedness curve.

It is interesting and important to note that the peakedness curve depends on the SBBP parameters only through m_1, m_2, m_*, γ. Therefore, we can get identical peakedness values for different SBBPs if these four parameters coincide.

Batch renewal process The batch renewal process is important to consider because of its ability to model the correlation structure of traffic [58]. The discrete time batch renewal process is made up of batches of arrivals, where the intervals between batches are independent and identically distributed random numbers, and the batch sizes are also independent and identically distributed, furthermore, the batch sizes are independent from the intervals between batches.

We use the following notation for the discrete time batch renewal process: a and b are the mean length of intervals between batches and the mean batch size, respectively. The first and second moments of the number of arrivals in an epoch is given by $m_1 = b/a$, and $m_2 = m_1 b(C_b^2+1)$ where C_b^2 is the squared coefficient of variation (variance to mean square ratio) of the batch size. The probability generating function of the distribution of time between batches is denoted by $A^*(\omega)$. ($A^*(\omega) = \sum_{s=1}^{\infty} a[s]\omega^s$ where $a[s]$ is the probability that the time between two consecutive batches is s.)

We have derived the peakedness for geometric holding times which is given by [83]

$$z_{\text{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{1+A^*(1-\mu)}{1-A^*(1-\mu)} - b + \frac{m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \tag{72}$$

If the distribution of time between batches follows a shifted generalized geometric distribution [58], that is, $a[t] = 1-\sigma$ if $t=1$ and $a[t] = \sigma\tau(1-\tau)^{t-2}$ if $t = 2,3,\ldots$, then its probability generating function is: $A^*(\omega) = \omega\left(1-\sigma+\frac{\sigma\tau\omega}{1-(1-\tau)\omega}\right)$ which makes the peakedness values easily computable.

Fitting traffic models to peakedness curves The peakedness shows the variability of the arrival stream with respect to different service holding times. It is of interest to investigate whether we can fit traffic models to peakedness curves based on measurements.

We outline here a fitting procedure based on the mean rate m_1 of the arrival traffic, the peakedness value at $\mu = 1$ and at three other points, μ_1, μ_2, μ_3. The model we fit to the peakedness curve is an interrupted batch Bernoulli process (IBBP): in one state of the modulating Markov process, the arrival number has a general distribution, in the other state, there are no arrivals.

First, by $z(1) = m_2/m_1 - m_1$, we get m_2. Introducing $\omega = 1 - \mu$, $\omega_i = 1 - \mu_i$ and using the notations of section 4.2.1, we can compute (using the values $K^*(\omega_i) = (z_{\text{geo}}(\mu_i) - 1)(\omega_i + 1) + 1$)

$$Y_i = Y(\omega_i) = m_1 \frac{1-\omega_i}{2\omega_i} \left(K^*(\omega_i) + m_1 \frac{1+\omega_i}{1-\omega_i} - \frac{m_2}{m_1} \right) \tag{73}$$

Using (71), $Y(\omega) = m_* - \frac{(m_* - m_1^2)(1-\gamma)}{1-\gamma\omega}$

Let us denote $\tilde{Y} = \frac{Y_1 - Y_2}{Y_2 - Y_3}$ which evaluates to $\tilde{Y} = \left(\frac{\omega_2-\omega_1}{\omega_3-\omega_2}\right)\left(\frac{1-\gamma\omega_3}{1-\gamma\omega_1}\right)$ and we get $\gamma = \frac{\tilde{Y}\frac{\omega_3-\omega_2}{\omega_2-\omega_1} - 1}{\tilde{Y}\frac{\omega_3-\omega_2}{\omega_2-\omega_1}\omega_1 - \omega_3}$ Once we have γ, we can obtain an estimation for m_* as $m_* = \frac{1}{3}\sum_{i=1}^{3} \frac{Y_i - \frac{m_1^2(1-\gamma)}{1-\gamma\omega_i}}{1 - \frac{1-\gamma}{1-\gamma\omega_i}}$ where we have on the right hand size an average for the known values ω_i, Y_i.

Then it is possible to fit an IBBP (no arrivals in state 2) as follows: $m_{1,(1)} = \frac{m_*}{m_1}, \alpha_2 = \frac{m_1(1-\gamma)}{m_{1,(1)}}, \alpha_1 = 1 - \gamma - \alpha_2, m_{1,(2)} = m_2 \frac{\alpha_1+\alpha_2}{\alpha_2}$. Given the first and second moments of the number of arrivals in state 1, we can use for example a generalized geometric distribution for modeling the batch size distribution. In this case, there are no arrivals with probability $1 - \varphi$, and there is a batch of arrivals with geometrically distributed size of parameter ψ. The moments are given by $m_{1,(1)} = \varphi/\psi$, $m_{1,(2)} = \varphi/\psi^2$ by which we can get φ, ψ for the model.

If it is possible to exactly fit an IBBP to the μ_i, $z_{\text{geo}}(\mu_i)$ pairs, the values that are summed in the equation for m_* are identical. If there is no IBBP that exactly fits the given peakedness values, m_* gives an estimation and the peakedness curve of the fitted IBBP model approximates the μ_i, $z_{\text{geo}}(\mu_i)$ pairs.

4.2.2 Generalized peakedness of real traffic

Measuring peakedness To measure the generalized peakedness of a traffic with a given holding time distribution, one can simulate the infinite server group. In discrete time, one can keep track of the first and second moment of the number of busy servers and compute the variance to mean ratio from them. The following points should be made about the estimation.

- We should take care of the initial phase of the simulation. If we have no prior knowledge about the traffic, we do not know what the mean number of busy servers will be. In this case, we can start from an empty system. The initial transient in the number of busy servers should be excluded from measurements.

- According to the definition, we should assign a server to each arrival, that is, assign a random holding time variable to every arrival in an epoch, which could involve a huge amount of computational effort. However, using the modified definition of peakedness and (64), we can reduce the computational effort by assigning only one random service time variable to all arrivals in an epoch.

- When the service time is geometric, we can minimize the computational effort by making use of the memoryless property. If at epoch t we have $L[t]$ busy servers, then at the next epoch we have $L[t+1] = L[t] + w[t+1] - D[t]$ where $D[t]$ is the number of departures from the service system at epoch t.

  The distribution of $D[t]$ is known to be binomial with parameters $L[t]$ and μ because each of the $L[t]$ servers finish service with probability μ. Therefore, in the measurement, it is enough to keep track of $L[t]$ together with the first and second moments of the previous $L[i], i \leq t$ values.

  This gives us the following procedure for computing the peakedness value for geometric holding time distribution with parameter μ:

  1. Reset $L_1 = 0$, $L_2 = 0$, L_{old} =initial value (see comments below);
  2. Set $L_{new} = L_{old} + w_{new} - d$ where d is a random number with distribution $binom(L_{old}, \mu)$ and w_{new} is the number of new arrivals in the next epoch;
  3. Set $L_1 = L_1 + L_{new}$, $L_2 = L_2 + L_{new}^2$;
  4. Set $L_{old} = L_{new}$ and loop back to 2 unless the measurement is over;
  5. Compute $l_1 = L_1/T, l_2 = L_2/T, z = l_2/l_1 - l_1$ where T is the length of the total measurement time.

  The setting of the initial value of L_{old} depends on the amount of a priori information that we have about the traffic. If we know the mean rate, we can set the initial L_{old} to its mean value determined by Little formula as m_1/μ. If we do not know the mean rate, we have to start from an empty system (initial $L_{old} = 0$) and simulate the service system without actually measuring (executing step 3) until the initial transient is over.

- An important advantage of using peakedness characterization is that we can measure peakedness by going through the traffic trace in only one sequence. This gives us the possibility of measuring peakedness for real-time traffic on the fly.

  Computing peakedness for one value of μ involves N cycles of the above procedure (where N is the total length of the measured traffic); if we want to measure peakedness at several μ values, we can easily implement the parallel execution of the procedure. In each cycle, we only have to compute a small number of additions and multiplications, and generate one binomially distributed random variable. Therefore, the complexity of the measurement is $O(N)$. The most time-consuming step in the measurement is the generation of the binomially distributed random number. We can reduce the computational cost of the measurement tremendously by approximating it with a normally distributed random number, for which precomputed look-up tables can be used.

- It is interesting to note that the measurement of peakedness involves randomness due to the simulation of servers, which means that for one sequence of traffic we could get different peakedness values. Experiments show that the peakedness measurements do not change significantly if we compute them more than once.

- The advantage of our approach compared to Eckberg's method for estimating peakedness for exponential holding times (cf. [27, 81]) is that our method does not neglect a lot of arrivals in the computation due to the selection of an arbitrary arrival.

Peakedness of video traffic Video traffic is a very important example of variable rate traffic. We investigated the application of peakedness measure for the characterization of variability of MPEG video traces [106]. In the MPEG coding scheme, the sequence of frames are divided into Groups of Picture (GOP), where each GOP is made up of so-called I, P and B frames. I frames are the largest because no prediction is used for coding them; P and B frames are smaller because one and two-directional prediction decreases the amount of information to be coded. The MPEG sequences that we considered had a GOP (Group of Pictures) length of 12 frames, a GOP pattern of IBBPBBPBBPBB, and frames capture frequency of 25 frames per second.

Figure 62(a) shows the peakedness curve of an an MPEG video trace of a movie (MrBean) as a function of the service rate μ. The mean service time of a server is therefore $1/\mu$ time epochs, where one time epoch is now 40ms. The solid curve is the peakedness function for the frame sequence (one frame corresponds to one epoch), whereas the dashed curve is the peakedness function for the GOP sequence (one GOP corresponds to 12 epoch so that is has the same time-length as the frame sequence) The scaling in the vertical axis is such that one arrival corresponds to one bit.

By decreasing the service rate, the service times become longer, and the number of busy servers in the infinite server group depends on the traffic properties on longer time scales. In this way, the peakedness curves show the variability of the traffic on different time scales, i.e. on the time scale of $1/\mu$.

Figure 62(a) shows that on short time scales, the variability of the frame sequence is much greater compared to the GOP sequence. But as we go to longer and longer time scales, the variability of the two sequences converge. What we can learn from this is that on longer time scales (for example, when dimensioning larger buffers), the statistical characteristics of GOP structure is less significant, and it is enough to consider the GOP sequence.

Figure 62(b) shows the peakedness curves for geometric service time distributions for five MPEG video GOP size traces. It gives us a relative comparison of the variability of different kinds of video sequences. (In this figure, one time epoch is set to one GOP which introduces a scaling compared to Figure 62(a).) The highest values of peakedness are exhibited by the MTV sequence, which is known to have lots of scene changes. Movie sequences show lower peakedness compared to the MTV sequence. The peakedness of a video conference sequence is found to be the smallest by orders of magnitude.

Figure 62(c) shows an IBBP fitted to an MPEG movie trace (MrBean, [106]). The solid line is the peakedness curve of the GOP sequence, the dashed line shows the peakedness curve of the fitted model. The circles show the peakedness values where the fitting was made. The points were chosen to represent the variability of the traffic on a long time scale (corresponding to the time scale of 1/0.01=100 epochs, here one epoch corresponds to 0.48 sec). As we can see, the model is able to capture the variability of the arrival stream on the investigated time scales.

Peakedness of aggregated ATM traffic We analysed the peakedness curve of an aggregated ATM traffic trace taken from the Finnish University and Research ATM WAN network (FUNET) [89]. The trace was approximately one hour long and consisted of the number of cell arrivals in each second. Figure 62(d) shows the peakedness curve of the measurement and two IBBPs fitted to it. The IBBP that was fitted at short time scale fits the measured peakedness curve well for shorter time scales, but it gives lower peakedness values for time scales longer than $1/0.05 = 20$sec. The other IBBP was fitted at a longer time scale; this model gives lower peakedness values for time scales shorter than 20sec.

Peakedness of Ethernet traffic Figure 62(e) shows the peakedness curve of IP packet traffic on an Ethernet [62] (we used 136 000 packet arrivals, about 428 seconds). Figure 62(e) shows the peakedness curve of aggregated ATM traffic [89] Two Markovian models (IBBPs) are fitted to the curves. The figure shows that the IBBPs are more variable on shorter time scales, but less variable on longer time scales.

Figure 62(e) and Figure 62(f) show the peakedness curve of an Ethernet traffic taken from the Bellcore measurements [62]. The measurement covers 1 million arrivals (approx. one hour). Figure 62(e) depicts peakedness on a lin-lin plot, Figure 62(f) is a log-log plot. We can investigate 5 different time scales in Figure 62(f). The interesting finding is that the peakedness increases linearly on the log-log plot as we decrease the rate (go to long time scales). Due to (68) and knowing that $\lim_{s\to\infty} I[s] = \infty$ if there is long range dependence (LRD) in the traffic, the peakedness diverges as the rate goes to zero. This observation of monotonicity in Figure 62(f) supports the presence of LRD assuming that the traffic stationarity assumption holds. It is important to note that *the peakedness curve can be used as an indicator of LRD.*

At different time scales we fitted simple Markovian models (IBBPs) to capture the peakedness curves in Figure 62(f). We can see that the burstiness scaling property of these models are not appropriate i.e. these models can cover a shorter range of time scales in burstiness than it would be necessary to follow the burstiness of the real traffic over all the investigated time scales.

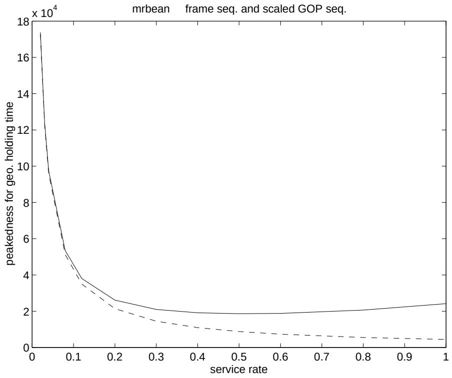
Our investigations of the aggregated ATM and Ethernet traffic indicate that simple Markovian models are not able to capture the burstiness characteristic of traffic over many time scales. For this case fractal traffic models seem to be more appropriate [89, 62]. However, for several practical cases we do not need to focus on *all* time scales but only on our working time scales (e.g. time scales of queueing) which can be efficiently modeled by Markovian models, too.

4.2.3 Summary

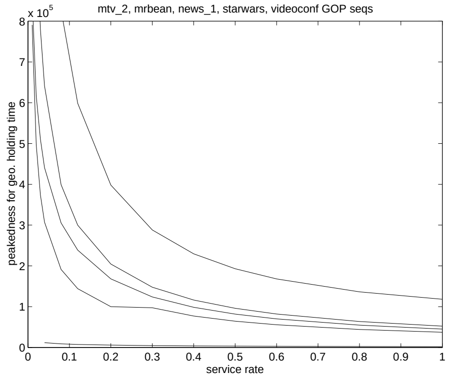
We have shown that peakedness can be used to characterize the bursty nature of traffic. Peakedness curves show the variability of traffic on different time scales and can be efficiently computed for real time traffic. We have extended the peakedness theory to discrete time and applied the peakedness characterization to variable rate video traffic, Ethernet traffic and aggregated ATM traffic as well as to the most important traffic models. We have shown that generalized peakedness can also be used for detecting long range dependence. We have also presented a new model fitting technique based on the concept of peakedness.

The basic idea of peakedness characterization is that we characterize traffic by its interactions with the service system. Although the traffic is usually offered to more complicated queueing systems, it is difficult to use complicated systems for characterization because it is very hard to handle them analytically. The infinite server group may be regarded as a compromise between generality and analytical tractability. Its generality is shown by the observation that peakedness gives a complete second order characterization, i.e. it contains all information about the correlation structure of the traffic.

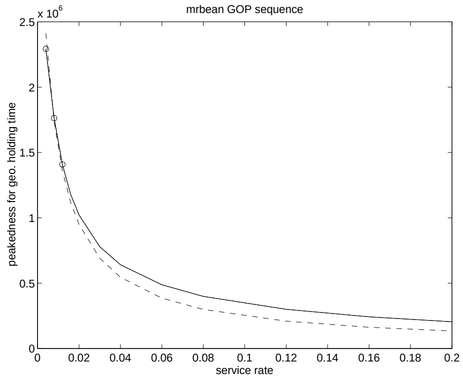
Most statistical measures, including peakedness, require the traffic to be stationary. However, the fact that we characterize the traffic by the reaction of a server system indicates that it is possible to extend it for non-stationary traffic which is problematic e.g. for IDC characterization. There are also other motivations for future research on peakedness measures: one is that, as we observed before, several Markovian arrival streams may have identical peakedness curves (because their correlation structure is identical); another motivation is that the peakedness characterization takes into account only the first and second moments of the arrival counts and in this way it fails to characterize the tail distributions sufficiently. Our future research will address these questions. The further development of peakedness theory including its extension to characterize non-stationary traffic are the topics of our future research.



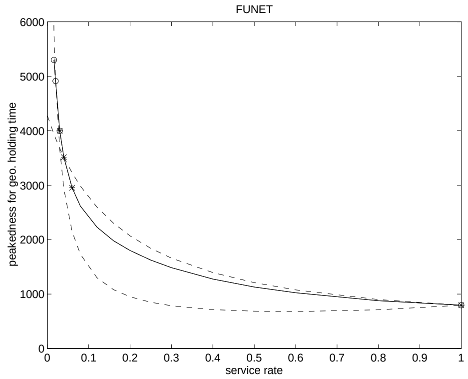
(a) Peakedness of the frame (solid) and GOP (dashed) sequence of MrBean MPEG video trace



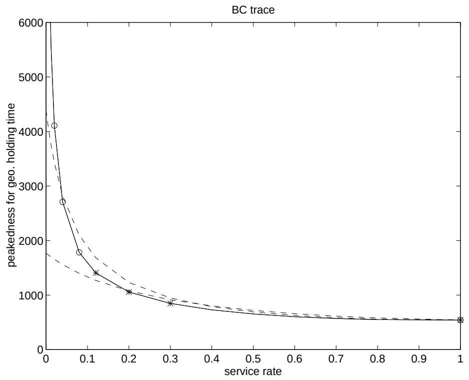
(b) Peakedness of MPEG GOP video sequences. From the uppermost down, the sequences are: TV (MTV), movie (MrBean), TV (News), movie (StarWars), video conference



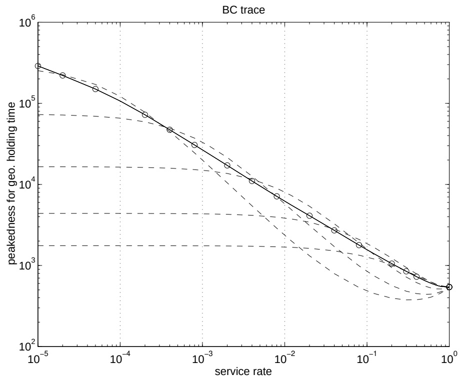
(c) Peakedness curve of a MPEG GOP movie trace (MrBean, solid line) and its IBBP model (dashed line).



(d) Peakedness of aggregated ATM traffic and Markovian models (dotted lines) fitted to it. IBBP models fitted at short (stars) and long (circles) time scales.



(e) Peakedness of IP packet trace on Ethernet (solid line) and Markovian models (dotted lines) fitted to it. IBBP models fitted at short (stars) and long (circles) time scales.



(f) Peakedness of Ethernet trace (solid) in log-log plot. On five time scales (separated by vertical lines) IBBP models are fitted (dashed).

Figure 62: Peakedness curves

5 Internet traffic characterization

The behaviour of Internet users has a major impact on the performance of networks and services. Especially telephone network operators are interested in descriptions of user generated Internet traffic to be able to adjust their current networks to the growing demand. Such traffic descriptions must be based on measurements of real traffic.

The modem pool at the University of Stuttgart offers the possibility to observe the users session behaviour in large scale and for a sufficiently long period. With around 4000 subscribed students and 800 subscribed staff members, the user population is large enough to draw conclusions on general user behaviour.

In this section we present the results of an evaluation of the modem pool log data from May to October 1997 with 369100 sessions [31]. The log data contain information on the login and holding times for access sessions. In contrast to many other publications concentrating on the packet level, we focus on this session level, i.e. the characteristics of the users dialup sessions without regarding the type and quantity of information transmitted during the sessions.

5.1 Session behaviour

The automatic monitoring of the user login times at the modem pool of the University of Stuttgart allows the evaluation of characteristic measures on session level. Since students and members of staff are assigned two separate modem pools, we distinguish between these two user groups. Note, that there is no way to specify the type of session 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 or a simple email retrieval 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 characterize the frequency and duration of a typical user session.

5.1.1 Holding time

Under the holding time of a session we understand the duration of the seizure of a modem. The mean holding time was around 21 minutes for students and 20 minutes for staff members. However the holding time shows a high variability. When comparing the students of different subjects the observed mean holding time varies from 8 to 31 minutes. Single users have even been online for several days. The maximum holding time was 11 days.

The holding time varies strongly during the course of the day. Figure 63 shows the holding time of sessions (of both user groups) associated with the time of the session start. Although this representation has to be regarded with caution (the average during the night is calculated from a relatively small number of calls), it allows the conclusion that long sessions start mainly during the night and early morning hours. Also the mean

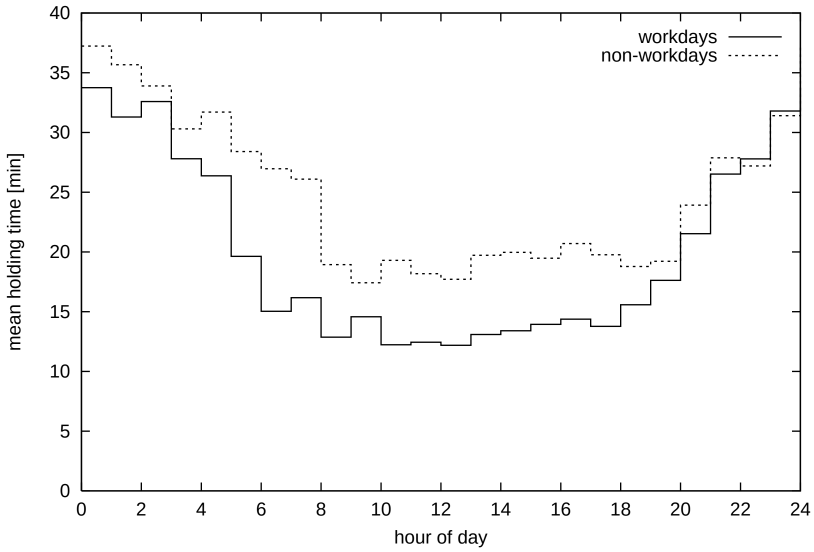


Figure 63: Mean holding time during the day

session length at night is significantly larger than during the day time. The average sessions in the mornings or during the day of non-workdays are longer.

In [94] Morgan reports a significant peak in holding time at 4 am. If the holding time would have been drawn associated with the session endings we would have obtained a peak at 4 am as well. This means that among sessions that end in the early morning, most have lasted for a long time and only few are of a short duration.

The high variability of the holding time is visible in the complementary cumulative distribution function (CCDF) which is depicted in Figure 64. The function shows the probability of a holding time being greater than the value on the horizontal axis. While there is a high probability for holding times of less than 2 hours the logarithmic presentation reveals that there is a small but not negligible probability for long sessions of 20 hours and more. This so-called ,heavy tail" is an indication for high variability of large values ([22]). The coefficient of variation of the holding time data is around 2.8 (i.e. the standard deviation is of the magnitude of 2.8 times the mean value).

5.1.2 Interarrival time

The interarrival time is the time between two consecutive session beginnings. This time can be measured either for sessions of individual users or for the sessions of all users. The first case leads to a description of the login frequency of a user while the latter one allows a description of the aggregate session arrivals as seen by the access provider.

For the aggregate traffic of 4900 subscribers the mean session interarrival time was 40 seconds. In contrast to the holding time discussed above, the interarrival time of aggregate sessions depends strongly on the number of participating users. The CCDF of the interarrival time of the aggregate traffic would be only a description of the behaviour of a group of 4900 users. To allow a comparison with other data we have scaled this

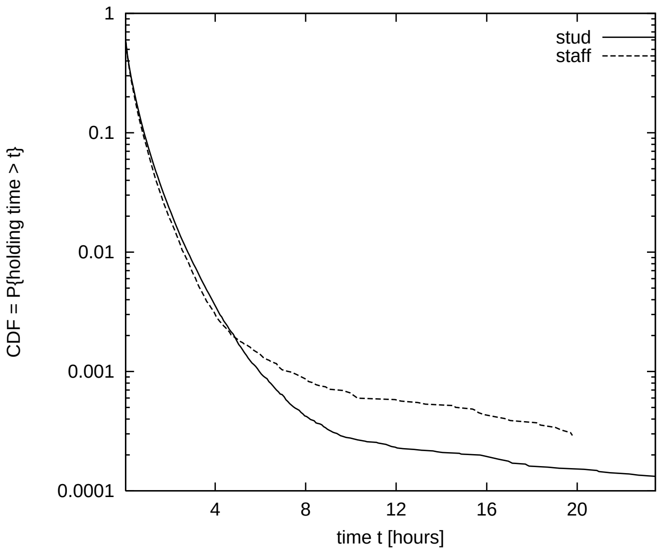


Figure 64: Complementary cumulative distribution function of the holding time

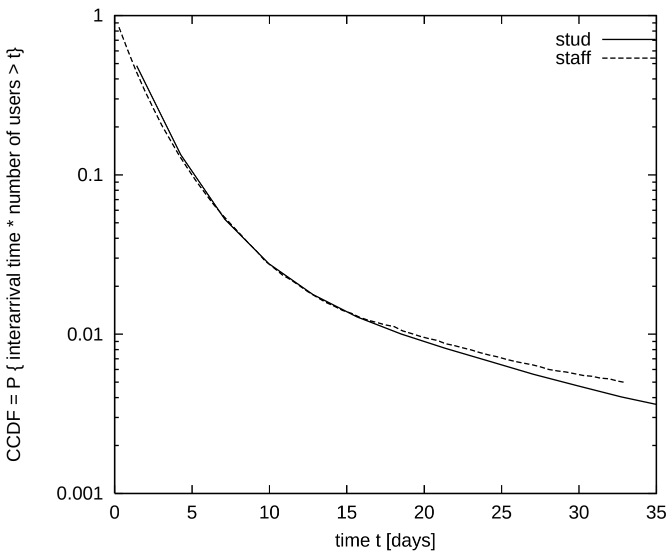


Figure 65: Complementary cumulative distribution function of the interarrival time

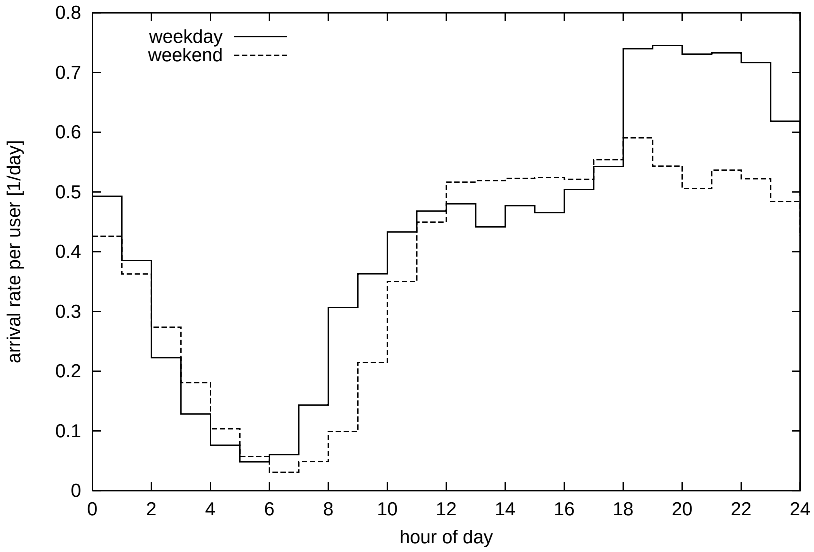


Figure 66: Mean arrival rate per user during the day

CCDF to describe the interarrival time per user (Figure 65). The scaled curves for 4100 students and 800 staff members fit amazingly well. Again a high variability is found for longer interarrival times which is indicated by the tail of the CCDF.

Figure 66 shows the mean session arrival rate per user during the course of the day for a typical weekday and on weekends. It is obvious that most sessions start during the afternoon and evening. In the early morning far less sessions are originated.

If the interarrival time is measured between sessions of individual users, a completely different distribution is received. Figure 67 shows the resulting complementary cumulative distribution function (average of all individual interarrival times of students as well as of members of staff). The curve shows characteristic steps in regular distances of 24 hours for both user groups. The same behaviour can also be found when regarding smaller user groups like students of certain fields. It is explained by the preferences of the users who tend to go online at a certain time of the day although not necessarily every day. Many users prefer for example to make use of cheaper telephone tariffs starting at 6 pm and 9 pm. Although this periodic behaviour is mostly due to telephone tariffs it is also caused by personal habits or work times of users.

Obviously there is a strong impact of the natural day and night shift, the working hours, individual preferences and tariffing schemes that leads to a periodic behaviour of end users. This periodicity may lead to unfortunate aggregation of traffic load but it may also be useful in influencing user behaviour in order to shift busy hours and balance network load.

5.1.3 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

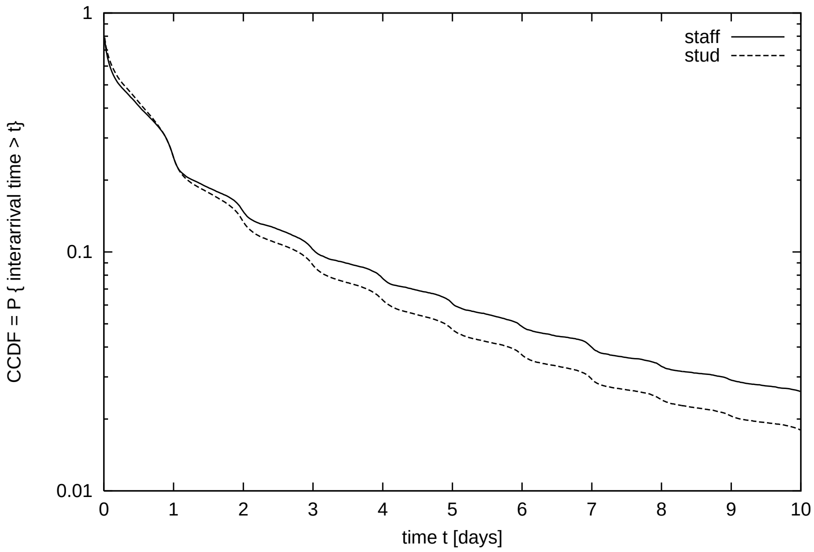


Figure 67: Complementary cumulative distribution function of the session interarrival time for individual users

power). Therefore we distinguish between traffic load related to user traffic (the seizure of the modem lines) and signalling load corresponding to the call arrival rate (see also Figure 66) to describe the actual load of the access network.

The average values for user traffic load and signalling traffic load for all days of the observed period are shown in the mean daily traffic profile in Figure 68. To allow a better comparison, both values are depicted in the same figure and are drawn in relation to their maximum values (the actual maximum values are 49 Erlang for the mean traffic load and 130 calls per hour for the mean arrival rate - note that only successful calls and no rejected calls could be detected at the modem pool). For this evaluation only the student modem pool with 56 modems and 4100 subscribers is taken into account.

The general shape of the traffic profile is almost complementary to the classic telephone traffic profile which has its busiest hours during the day and decreases at the end of the day. The shown traffic profile is certainly not typical for dialup sessions in general, but it is typical for the behaviour of home users who usually dial up after work hours.

The most obvious characteristics of the mean seizure are the two steps at 6 pm and 9 pm. As already mentioned above, these times mark the beginning of cheaper telephone tariffs during the observed period. A small peak is also visible right before 9 am when the expensive day tariff starts. On weekends and holidays the day tariff is the same as in the evening and morning hours and therefore no such steps are visible at 9 am and 6 pm. The user behaviour follows these tariffing scheme amazingly accurate.

The flat shape around midnight is due to the limited number of modems. It is interesting that there is a lot of traffic during late night and early morning hours. A comparison with the arrival rate indicates that this traffic is caused by few but long sessions (see Figure 63).

The time consistent busy hour for user traffic load is found at 11:45 pm to 0:45 am

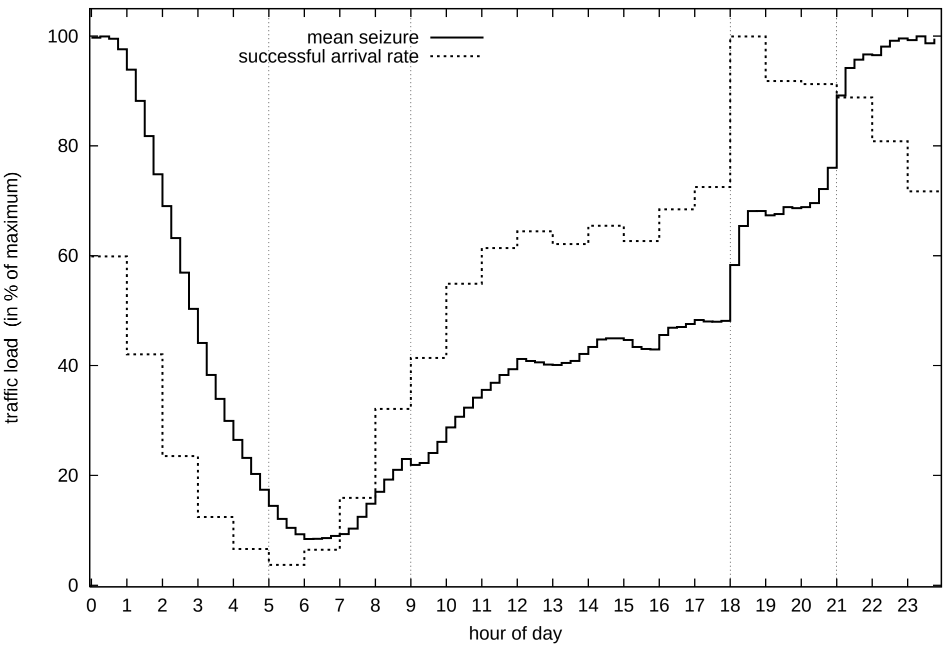


Figure 68: Mean daily traffic profile for traffic load and arrival rate (students only)

and for the arrival rate it is found at 6 pm. In [14] Bolotin points out, that these two busy hours are significantly shifted against each other for Internet traffic compared to telephone traffic. This phenomenon is caused by much longer holding times of around 20 minutes (compared to 3 minutes for classical telephone calls). For design and dimensioning of network components, traffic load has to be evaluated carefully for both busy hours.

5.2 Modelling dialup 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 to parameterize a random generator. This generator may then provide a simulation with values for the measure of interest according to that CCDF.

While classic telephone traffic is well described with negative exponentially distributed holding times and call interarrival times this is not true for Internet access session traffic any more. The high variability of the measures described above is not

Table 5: Cumulative distribution functions of Pareto, Weibull and the hyperexponential distributions

Pareto distribution	Weibull distribution	hyperexponential distribution
$P\{X \leq x\} = 1 - \left(\frac{k}{x}\right)^{\alpha}$ with $k > x$	$P\{X \leq x\} = 1 - e^{-\left(\frac{x}{\beta}^{\alpha}\right)}$	$P\{X \leq x\} = 1 - \sum_{i=1}^{k} p_i e^{-\lambda_i x}$ with $\sum_{i=1}^{k} p_i = 1$

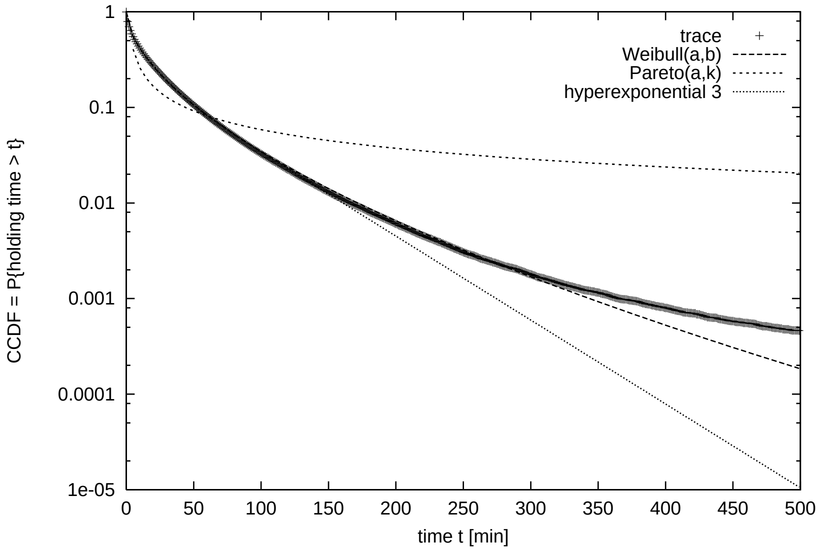


Figure 69: Fitting of several functions to the CCDF of the holding time

captured by this distribution. For the description of Internet traffic other distributions have been proposed like Pareto, hyperexponential or Weibull distributions (e.g. [21], [33], [73], [109]).

In the following sections these distributions are fitted to the empirical distributions of the traces and the resulting parameters are presented (see Table 5). We use the Marquardt-Levenberg algorithm to find a set of parameters for each function. This iterative algorithm performs a non linear least square fit.

5.2.1 Holding time

Figure 69 shows the results of the fitting operations to the CCDF of the session holding time. Obviously the Pareto distribution is not suited to describe this measure. The Weibull distribution, on the other hand, leads to a rather good fit and also the hyperexponential distribution describes the behaviour accurately to a certain point. The resulting parameters of the fit can be found in Table 6.

The table also shows the resulting mean value and the coefficient of variation of the parameterized functions in case they do exist. While the mean values for the Weibull and hyperexponential functions are close to the empirical mean, the variability of the real measure is still significantly higher.

Table 6: Resulting parameters of the fitting operation for the holding time

distribution	parameters	mean [min]	CoV
Trace	–	20.6	2.8
Pareto	$\alpha = 0.649$, $k = 1.265$	∞	∞
Weibull	$\alpha = 0.584$, $\beta = 12.544$	19.569	1.82
hyperexponential	$p_1 = 0.436$, $p_2 = 0.257$, $\lambda_1 = 0.072$, $\lambda_2 = 0.020$, $\lambda_3 = 0.73$	19.201	1.71

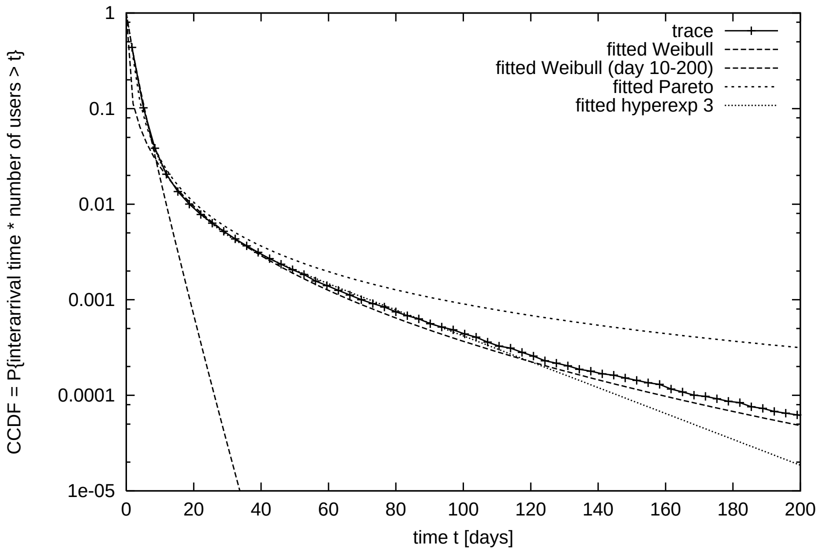


Figure 70: Fitting of several functions to the CCDF of the interarrival time

5.2.2 Interarrival time

As depicted in Figure 70 the scaled CCDF of the interarrival time (of the summary traffic of all subscribers) is best approximated by a Pareto or a hyperexponential distribution. The Weibull distribution can be fitted either to head or tail of the empirical distribution and is not well suited for its description. Table 7 shows the resulting parameters for all fitted functions as well as the corresponding mean values and the coefficients of variation.

Table 7: Resulting parameters of the fitting operation for the interarrival time

distribution	parameters	mean [min]	CoV
Trace	–	2.424	2.4
Pareto	$\alpha = 1.453$, $k = 0.964$	3.095	∞
Weibull	$\alpha = 0.879$, $\beta = 2.091$	2.230	1.272
hyperexponential	$p_1 = 0.926$, $p_2 = 0.064$, $\lambda_1 = 0.528$, $\lambda_2 = 0.137$, $\lambda_3 = 0.031$	2.524	2.040

For the interarrival time as for the holding time, the hyperexponential distribution is difficult to fit to the CCDF because of its numerous parameters. The choice of the starting values is significant and the characteristic heavy tail is not captured accurately. If a hyperexponential distribution is chosen to describe the traffic, a more systematic approach would be preferable as is suggested by Feldmann and Whitt in [33].

5.3 Summary

We have presented the traffic characteristics of dialup sessions monitored at the modem pool of the University of Stuttgart. The long holding times and the high variability of holding time and interarrival time can be found in other publications as well and seem to be typical for Internet traffic. We have also shown that the user behaviour is heavily influenced by the employed telephone tariffing scheme.

Finally a simple mathematical description of the holding time as well as of the interarrival time was presented. Note that the current modelling approach captures the overall behaviour of the traffic. As depicted in Figure 68 the traffic characteristics vary during the course of the day. Therefore, a model describing traffic load only during the busy hours of internet traffic as well as of telephone traffic would be helpful for network dimensioning.

We like to point out, that our results are based on empirical data of a special user group (students and university staff members) and might not describe general Internet traffic. Also the behaviour was strongly influenced by the telephone tariffing scheme in Germany and it should be mentioned that the fast Internet access itself was provided for free.

References

[1] J. Abate, G. L. Choudhury, and W. Whitt. Asymptotics for Steady-state Tail Probabilities in Structured Markov Queueing Models. *Commun. Statist. - Stochastic Models*, 10(1):99–143, 1994.

[2] A. T. Andersen, A. Jensen, and B. F. Nielsen. Modelling and performance study of packet-traffic with self-similar characteristics over several timescales with Markovian Arrival Processes (MAP). In I. Norros & J. Virtamo, editor, *Nordic Teletraffic Seminar, NTS 12, VTT Symposium 154*, pages 269–283, Espoo, Finland, August 1995. NTS, VTT.

[3] A. T. Andersen and B. F. Nielsen. On the use of second order descriptors to characterize MAPs. In *COST 257 TD(97)050*, September 1997.

[4] A.T. Andersen and B. F. Nielsen. On the implications of certain random permutations of inter-arrival times or counts in a point process. In *COST 257 TD(98)21*, January 1998.

[5] S. Appleby. Fractal telecommunications networks. *British Telecom Technol. J.*, 12(2):19–29, 1994.

[6] S. Appleby. Estimating the cost of a telecommunications network using the fractal structure of the human population distribution. *IEE proceedings: Communications*, 142(3):172–178, 1995.

[7] Amtliches Topographisches Kartographisches Informations System (in German); Bavarian land survey office, Munich, FRG, 1991.

[8] A.W. Berger AT&T. *On the Index of Dispersion for Counts for User Demand Modeling.* ITU, Madrid, Spain, 27-29 June 1994. Study Group 2, Question 17/2.

[9] J. Beran. *Statistics for Long-Memory Processes*, volume 61 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, New York, 1994.

[10] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-Range Dependence in Variable-Bit-Rate Video Traffic. *IEEE Transactions on Communications*, 43(2/3/4):1566–1579, 1995.

[11] C. Blondia. A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32, 1993.

[12] C. Blondia and F. Geerts. The correlation structure of the output of an ATM multiplexer. In *Proceedings of the fifth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks.* IFIP, 1997.

[13] D. C. Boes and J. D. Salas. Nonstationarity of the Mean and the Hurst Phenomenon. *Water Resources Research*, 14(1), 1978.

[14] V. Bolotin. New subscriber traffic variability patterns for network traffic engineering. In *Proc. of the 15th International Teletraffic Congress - ITC 15, Vol. 2*, pages 867–878, 1997.

[15] O.J. Boxma. Fluid queues and regular variation. *Performance Evaluation*, 27 & 28:699 – 712, 1996.

[16] E. Chlebus. Analytical grade of service evaluation in cellular mobile systems with respect to subscribers' velocity distribution. In *Proc. 8th Australian Teletraffic Research Seminar*, pages 90–101, 1993.

[17] D. B. H. Cline. Limit Theorems for the Shifting Level Process. *Journal of Applied Probability*, 20(2), 1983.

[18] D. R. Cox. *Statistics: An appraisal*, chapter Long-range dependence: a review, pages 55–74. Iowa State University Press, 1984.

[19] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen & Co Ltd, 1966.

[20] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Chapman and Hall, 1980.

[21] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. In *Proceedings of the ACM SIGMETRICS*, pages 160–169, 1996.

[22] M. Crovella and A. Bestavros. Performance characteristics of world wide web information systems. Tutorial at the SIGMETRICS'97, 1997.

[23] T. Daniels and C. Blondia. A discrete-time ATM traffic model with long-range dependence characteristics. In *COST257TD(97)39*, 1997.

[24] L. Devroye. The double kernel method in density estimation. *Ann. Inst. Henri Poincaré*, 25:533–580, 1989.

[25] N. Duffield, J. Lewis, and N. O'Connell. Statistical issues raised by the Bellcore data. In *Proc. 11th Teletraffic Symposium*, 1994.

[26] A. E. Eckberg, Jr. Generalized peakedness of teletraffic processes. In *ITC-10*, Montreal, 1983.

[27] A. E. Eckberg, Jr. Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness. In *ITC-11*, Kyoto, Japan, 1985.

[28] A. El-Dolil, W.-C. Wong, and R. Steele. Teletraffic performance of highway microcells with overlay macrocell. *IEEE Journal on Selected Areas in Communications*, 7(1):71–78, January 1989.

[29] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent. *IEEE/ACM Trans. on Networking*, 4(2):209–223, 1996.

[30] Z. Fan and P. Mars. Accurate approximation of cell loss probability for self-similar traffic in ATM networks. *Electronic Letters*, 32(19), September 1996.

[31] J. Färber, S. Bodamer, and J. Charzinski. Measurement and modeling of Internet traffic at access networks. In *COST257TD(98)43*, 1998.

[32] A. Feldmann, A. C. Gilbert, and W. Willinger. Datanetworks as cascades: Investigating the multifractal nature of Internet WAN traffic. In *Proceedings of the ACM/SIGCOMM'98*, Vancouver, Canada, September 1998.

[33] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models. *Performance Evaluation 31*, pages 245–279, 1998.

[34] K. W. Fendick and W. Whitt. Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue. *Proceedings of the IEEE*, 77(1):171–194, January 1989.

[35] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1993.

[36] G. J. Foschini, B. Gopinath, and Z. Miljanic. Channel cost of mobility. *IEEE Transactions on Vehicular Technology*, 42(4):414–424, November 1993.

[37] H. J. Fowler and W. E. Leland. Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management. *IEEE Journal On Selected Areas In Communications*, 9(7):1139–1149, September 1991.

[38] M. R. Frater, J. F. Arnold, and P. Tan. A New Statistical Model for Traffic Generated by VBR Coders for Television on the Broadband ISDN. *IEEE, Trans. Circuits & Systems for Video Technology*, 4(6):521–526, December 1994.

[39] M. W. Garrett. Contributions Toward Real-Time Services on Packet Switched Networks. Technical report, Columbia University, New York, 1993.

[40] A. Ghosh and S. L. McLafferty. *Location Strategies for Retail and Service Firms.* Heath, Lexington, MA, 1987.

[41] M. Grasse, M. Frater, and J. Arnold. Implications of non-stationarity of MPEG2. In *COST257TD(97)10*, 1997.

[42] M. Grasse, M. R. Frater, and J. F. Arnold. Statistics of Variable Bit Rate Video Coders with and without Motion Compensation. In *6th International Workshop on Packet Video*, Portland, Oregon, September 26-27, 1994.

[43] M. Grasse, M. R. Frater, and J. F. Arnold. Non-Stationarity in VBR Video Traffic. In *Proceedings of the Australian Telecommunication Networks & Applications Conference, Sydney*, 1995.

[44] S.M. Grasso, F.M. Mercuri, G. Roso, and D.M. Tacchino. DEMON: A forecasting tool for demand evaluation of mobile network resources. In *Proc. Networks '96*, pages 145–150, Sydney, Australia, November 1996.

[45] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. In *SIGCOMM'96*, August 1996.

[46] R. Grünenfelder and S. Robert. Which arrival law parameters are decisive for queueing system performance. In *ITC-14*, pages 377–386, Antibes Juan-les-Pins, June 1994.

[47] R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2), February 1991.

[48] H. Heffes and J. M. Holtzman. Peakedness of traffic carried by a finite trunk group with renewal input. *The Bell System Technical Journal*, 52(9):1617–1642, November 1973.

[49] H. Heffes and D.M. Lucantoni. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE Journal On Selected Areas In Communications*, 4(6):856–868, September 1986.

[50] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Transactions on Networking*, 4(3):301–317, 1996.

[51] D. Hong and S. S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, VT–35(3):77–92, August 1986.

[52] J. R. M. Hosking. Fractional Differencing. *Biometrika*, 68(1):165–176, 1981.

[53] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[54] J. Jormakka. On self-similar models for ATM traffic. In *Proc. ITC Specialists Seminar on Control in Communications*, pages 277–288, Lund, Sweden, September 1996.

[55] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 3. Charles Griffin & Company Ltd., London, 2nd edition, 1968.

[56] V. Klemeš. The Hurst Phenomenon: A Puzzle? *Water Resources Research*, 10(4), 1974.

[57] K. Kobayashi and Y. Takahashi. The tail probability of a gaussian fluid queue under finite measurement of input processes. In *Proc. Int. Conference on the Performance and Management*, pages 57–72, Tsukuba, Japan, November 1997.

[58] D. Kouvatsos and R. Fretwell. Batch renewal process: Exact model of traffic correlation. In *High Speed Networking for Multimedia Application*, pages 285–304. Kluwer Academic Press, 1996.

[59] T. G. Kurtz. *Stochastic Networks: Theory and Applications*, chapter Limit theorems for workload input models, pages 339–366. Clarendon Press, Oxford, UK, 1996.

[60] K. Laevens. (Heavy-tailed) on-off sources (i): traffic characteristics. In *COST257TD(97)09*, Leidschendam, jan 1997.

[61] K. Laevens. (Heavy-tailed) on-off sources (ii): superposition. In *COST257TD(97)48*, Leidschendam, jan 1997.

[62] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, February 1994.

[63] T. Leskien. Erfassung und Charakterisierung geographischer Daten für die Planung von Mobilfunksystemen (in German). Master's thesis, University of Würzburg, Institute of Computer Science, March 1997.

[64] K. K. Leung, W. A. Massey, and W. Whitt. Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications*, 12(8):1353–1364, October 1994.

[65] J. Lévy Véhel and R. Riedi. Fractional brownian motion and data traffic modeling: The other end of the spectrum. In *Fractals in Engineering 97*. Springer, 1997.

[66] S. Li and C. Hwang. Queue Response to Input Correlation Functions: Discrete Spectral Analysis. In *Proceedings of INFOCOM '92*. IEEE, May 1992.

[67] N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proceedings of INFOCOM'95*. IEEE, 1995.

[68] K. Lindberg. The FUNET network connection to foreign countries. *CSC News*, 7(4):19, December 1995.

[69] D. Liu and M. F. Neuts. Counter-Examples Involving Markovian Arrival Processes. *Commun. Statist. -Stochastic Models*, 7(3):499–509, 1991.

[70] D. M. Lucantoni. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Commun. Statist. Stochastic Models*, 7(1):1–46, 1991.

[71] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A Single-Server Queue with Server Vacations and a Class of Non-Renewal Arrival Processes. *Adv.Appl.Prob*, 22:676–705, 1990.

[72] R. Macfadyen. Self-similarity: Complication or distraction. In *COST 257 TD(98)001*, Rome, Italy, January 1998.

[73] B. Mah. An empirical model of http network traffic. In *Proceedings of the IEEE INFOCOM'97, Vol.2*, pages 592–600, 1997.

[74] B. Mandelbrot. Some Noises with 1/f Spectrum, a Bridge Between Direct Current and White Noise. *IEEE Transactions on Information Theory*, IT-13(2):289–298, April 1967.

[75] B. B. Mandelbrot. Possible refinement of the lognormal hypothesis concerning the distribution of energy dissipation in intermittent turbulence. In M. Rosenblatt and C. Van Atta, editors, *Statistical models and turbulence*, number 12 in Lecture notes in physics, pages 331–351. Springer, 1972.

[76] B. B. Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *J. Fluid. Mech.*, 64, 1974.

[77] P. Mannersalo, A. Koski, and I. Norros. Telecommunication networks and multifractal analysis of human population distribution. COST257TD(98)02, VTT Information Technology, 1998.

[78] P. Mannersalo and I. Norros. Multifractal analysis: a potential tool for characterizing teletraffic? COST257TD(97)32, VTT Information Technology, 1997.

[79] P. Mannersalo and I. Norros. Multifractal analysis of real ATM traffic: a first look. COST257TD(97)19, VTT Information Technology, 1997.

[80] B. L. Mark, D. L. Jagerman, and G. Ramamurthy. Application of peakedness measures to resource allocation in high-speed networks. In *Proceedings of ITC-15, Washington D.C., USA*, June 1997.

[81] B. L. Mark, D. L. Jagerman, and G. Ramamurthy. Peakedness measures for traffic characterization in high-speed networks. In *Proceedings of IEEE INFOCOM'97*, 1997.

[82] O. Melteig. Introduction to the PARASOL project. In *The 9th Nordic Teletraffic Seminar*, August 1990. Norwegian Telecom, Rearch Department.

[83] Gy. Miklós. Peakedness measures. Technical report, High Speed Networks Lab, Department of Telecommunications and Telematics, Technical University of Budapest, 1997.

[84] S. Molnár. *Evaluation of Quality of Service and Network Performance in ATM Networks.* PhD thesis, Technical University of Budapest, Department of Telecommunications and Telematics, 1995.

[85] S. Molnár, I. Cselényi, and N. Björkman. ATM traffic characterization and modeling based on the leaky bucket algorithm. In *IEEE Singapore International Conference on Communication Systems*, Singapore, November 1996.

[86] S. Molnár and Gy. Miklós. On burst and correlation structure of teletraffic models. In D. D. Kouvatsos, editor, *5th IFIP Workshop on Performance Modelling and Evalution of ATM Networks*, Ilkley, U.K., July 1997.

[87] S. Molnár and Gy. Miklós. Generalized peakedness in discrete time. In *COST 257 TD(98)052*, September 1998.

[88] S. Molnár and Gy. Miklós. Peakedness characterization in teletraffic. In *IFIP International Conference on Performance of Information and Communication Systems.* Lund, Sweden, 25-28 May, 1998.

[89] S. Molnár and A. Vidács. On modeling and shaping self-similar ATM traffic. In *Proceedings of ITC-15, Washington D.C., USA*, June 1997.

[90] S. Molnár and A. Vidács. How to characterize Hursty traffic? In *COST 257 TD(98)003*, Rome, Italy, January 1998.

[91] S. Molnár, A. Vidács, and I. Cselényi. Queueing performance in the presence of long-range dependence. In *COST 257 TD(98)051*, September 1998.

[92] S. Molnár, A. Vidács, and A. Nilsson. Bottlenecks on the way towards fractal characterization of network traffic: Estimation and interpretation of the hurst parameter. In *International Conference of the Performance and Management of Complex Communication Networks (PMCCN'97)*, Tsukuba, Japan, November 1997.

[93] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill, Tokyo, 3rd edition, 1974.

[94] S. Morgan. The internet and the local telephone network: Conflicts and opportunities. *IEEE Communications Magazine*, pages 42–48, January 1998.

[95] M. Mouly and M.-B. Pautet. *The GSM System for Mobile Communications.* published by the authors, ISBN: 2-9507190-0-7, 4, rue Elisée Reclus, F-91120 Palaiseau, France, 1992.

[96] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, volume 5 of *Probability: Pure and Applied*. Marcel Dekker, Inc, 1989.

[97] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks, September 1994.

[98] R. O. Onvural. *Asynchronous Transfer Mode Networks, Performance Issues*. Artech House, Boston, London, 1994.

[99] V. Paxson. Fast approximation of self-similar network traffic, April 1995.

[100] V. Paxson and S. Floyd. Why we don't know how to simulate the internet. In *Winter Simulation Conference*, Atlanta, December 1997.

[101] M. B. Priestley. *Spectral Analysis and Time Series*, volume 1. Academic Press, London, 1981.

[102] M. B. Priestley and T. Subba Rao. A Test for Non-stationarity of Time-series. *J. Roy. Statist. Soc. Ser. B*, 31:140–149, 1969.

[103] R. Riedi, M. Crouse, V. Ribeiro, and R. Baraniuk. A multifractal wavelet model with application to TCP network traffic. *IEEE Special Issue on Information Theory*, 1998. To appear.

[104] R. Riedi and J. Lévy Véhel. TCP traffic is multifractal: a numerical study. Inria research report, no. 3129, Project Fractales, INRIA Rocquencourt, 1997. submitted to IEEE Transactions of Networking.

[105] J. Roberts, U. Mocci, and J. Virtamo, editors. *Broadband network teletraffic*, volume 1155 of *Lecture notes in computer science*. Springer, 1997.

[106] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. In *Proceedings of the 20th Annual Conference on Local Computer Networks*, pages 397–406, Minneapolis, MN, 1995. ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/MPEG/.

[107] M. Roughan and D. Veitch. Measuring long-range dependence under changing traffic conditions, 1998. preprint.

[108] B. Ryu and A. Elwalid. The importance of long-range dependence of vbr video traffic in atm traffic engineering: Myths and realities. In *SIGCOMM'96*, August 1996.

[109] D. Shuang. Empirical model of www document arrivals at access link. In *CC International Communication Conference*, 1996.

[110] K. Sriram and W. Whitt. Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. *IEEE Journal On Selected Areas In Communications*, 4(6):833–846, September 1986.

[111] G. D. Stamoulis, M. E. Anagnostou, and A. D. Georgantas. Traffic source models for ATM networks: a survey. *Computer Communications*, 17(6), 1994.

[112] M. S. Taqqu, V. Teverosky, and W. Willinger. Is network traffic self-similar or multifractal? *Fractals*, 5:63–73, 1997.

[113] K. Tutschku, N. Gerlich, and P. Tran-Gia. An integrated approach to cellular network planning. In *Proc. Networks '96*, pages 185–190, Sydney, Australia, November 1996.

[114] K. Tutschku, K. Leibnitz, and P. Tran-Gia. ICEPT - An integrated cellular network planning tool. In *Proc. VTC 97*, Phoenix. USA, May 1997.

[115] K. Tutschku, T. Leskien, and P. Tran-Gia. Traffic estimation and characterization for the design of mobile communication networks. In *COST257TD(97)47*, 1997.

[116] M. H. van Hoorn and L. P. Seelen. The SPP/G/1 Queue: Single Server Queue with a Switched Poisson Process as Input Process. *OR Spektrum*, 5:205–218, 1983.

[117] A. Vidács, S. Molnár, and I. Cselényi. The impact of long range dependence on cell loss in an ATM wide area network. *Globecom '98*, November 1998.

[118] W. Vetterling W. Press, S. Teukoslky and B. Flannery. *Numerical Recipes in C (second edition)*. Cambridge University Press, 1994.

[119] W. Willinger, V. Paxson, and M. S. Taqqu. Self-similarity and heavy tails: structural modeling of network traffic. In J. Adler, R. E. Feldman, and M. S. Taqqu, editors, *A practical guide to heavy tails: statistical techniques and applications*. Birkhäuser, 1998.

[120] W. Willinger, M. Taqqu, and A. Erramilli. *A bibliographical Guide to self-similar traffic and performance modeling for modern high-speed networks*, chapter Stochastic Networks: Theory and Applications. 339–366, Oxford University Press 1996.

[121] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *Proc. of the ACM/Sigcomm'95*, 1995.

[122] S. Wittevrongel and H. Bruneel. Queue length and delay for statistical multiplexers with variable-length messages. In *GLOBECOM'94*, pages 1080–1084, San Francisco, November 1994.

[123] S. Wittevrongel and H. Bruneel. Effect of the on-period distribution on the performance of an atm multiplexer fed by on/off sources: an analytical study. In *PCN'95*, pages 33–47, Istanbul, October 1995.