# Chapter 3
# Traffic Characterization

Sándor Molnár
*Budapest University of Technology and Economics, Hungary*

Contributors:
Sándor Molnár (Budapest University of Technology and Economics, Hungary), Ilkka Norros, (VTT, Finland), and Bo Friis Nielsen (Tech. University of Denmark, Denmark)

## 3.1 Introduction

Understanding the nature of traffic, identifying its characteristics and building practical models are vital for the teletraffic engineering of today's packet switched networks. New observations of measured traffic call for new approaches (e.g. multifractal characterization) and the ever changing services and protocols of the Internet trigger particular models (e.g. WWW models).

This chapter overviews the traffic characterization and modeling activities of the COST-257 project. Section 3.2 presents different models developed for packet traffic including discrete time models, Gaussian models and multifractal models. Models related to particular services (internet dial-up traffic) and protocols (WWW traffic, TCP traffic) are summarized in section 3.3. Section 3.4 discusses the statistical methods related to traffic models and, finally section 3.5 addresses the issue of traffic generation.

For more details see the chapter on "Source Characterization in Broadband Networks" in the COST-257 Interim Report [177] and the referred COST TDs.

## 3.2 Models for packet traffic

### 3.2.1 Discrete time models

Discrete time queueing models are of inherent interest in modelling communication systems. Many such systems are timed or synchronous with an advance of certain information units at every time slot or clock cycle. Models of this type are especially important when modelling ATM systems but are also relevant for the modelling of other systems. A general reference to these types of systems can be found in [29]. The work within this area in the COST-257 project has partly focused on general models and partly focused on models dedicated to the modelling of highly variable packet traffic.

In [136] a framework for the modelling and analysis of a superposition of ON-OFF sources in discrete time is considered. The sources are allowed to have a general distribution while successive ON and OFF periods are independent. Second order descriptors for the process of generated cells and results for the superposition of $N$ of these ON-OFF sources are derived. The superposition of $N$ sources of this kind is then analysed under the assumption that the duration of the OFF periods is geometric.

The work in [136] primarily addresses light-tailed sources. In [137] the case of sources with heavy-tailed ON or OFF sources is examined. A special form of hyper-geometric distribution was applied to model the heavy tails of these distributions. When considering a certain limiting construction with heavy-tailed ON periods and geometrically distributed OFF periods the model can be seen as a discrete analogue of the continuous M/G/$\infty$ system which has been used for the modelling of long-range dependence in a continuous setting. An outline of a possible queueing analysis is given but no numerical results are obtained.

The paper [52] is related. The approach followed in this paper leads to a different way of obtaining a discrete version of the M/G/$\infty$ arrival model. In each time slot there is a Poisson distributed number of arrivals. Each of these arrivals consists of a batch which has a size described by a Pareto distribution. This process is fed into a single server queue and it is shown that the queue length distribution asymptotically behaves like a power law distribution. It is shown by numerical simulation that the asymptotic behaviour is a very good approximation over the whole range of values for the buffer occupancy.

A process obtained by a superposition of an infinite number of simple ON-OFF sources is analysed in [51]. It is shown that the resulting process exhibits long-range dependence and by matrix analytic methods it is argued that the queue is unstable for certain values of the parameters in the sequence of ON-OFF sources.

Another arrival model, not directly related to highly variable traffic is described in [173]. This work characterises the output process of a discrete time queue with deterministic service fed by a batch renewal process. A joint transform for the duration of the idle and busy periods in successive busy cycles is derived.

### 3.2.2 Gaussian traffic models

By the Central Limit Theorem, the sum of a large number of "small" independent random variables has an approximately Gaussian (normal) distribution. Thus, one can expect that an aggregated teletraffic stream consisting of a large number of individual communications can be reasonably well modelled by a Gaussian stochastic process.

A continuous time Gaussian process with stationary increments is a process $(A_t : t \in \mathbb{R})$ such that for any $t_1, \ldots, t_k$ and $s$, the random vector $(A_{t_1} - A_s, \ldots, A_{t_k} - A_s)$ has a multivariate Gaussian distribution independent of $s$. For normalization, we set $A_0 \overset{\text{a.s.}}{=} 0$. Such a process is characterized by a number $m$ and a symmetric non-negative function $v(t)$ such that

$$\mathrm{E}A_t = mt, \quad \mathrm{Var}\,(A_t) = v(t).$$

The covariance function of $A$ is then given by

$$\mathrm{Cov}\,(A_s, A_t) = \Gamma(s,t) = \frac{1}{2}(v(s) + v(t) - v(t-s)). \tag{3.1}$$

In teletraffic modelling, $m$ is called the mean rate, and $v(t)$ is the cumulative variance function. If $A$ and $B$ are two independent processes, we have $v_{A+B}(t) = v_A(t) + v_B(t)$. This means that it is very simple to deal with complicated traffic mixes with different dependence structures, if the Gaussian approximation is good enough.

The general case includes processes whose paths have unbounded variation. Basic examples include Brownian motion, used in diffusion approximations to queues, and fractional Brownian motion, a self-similar process used as a model for long-range dependent traffic. On the other hand, it is often interesting to model traffic by a Gaussian process with smooth paths, i.e. as a Gaussian fluid process. In that case, $A$ can be written as

$$A_t = \int_0^t \Lambda_s \,\mathrm{d}s,$$

where $\Lambda$, the rate process, is a stationary Gaussian process. For more details, see [3].

Queueing theory with Gaussian input is discussed in subsection 6.3.4.

Other COST-257 work related to Gaussian traffic models: [277] presents a new estimation technique in the case of fBm traffic. A new simulation technique for general Gaussian traffic is described in [194]. A "handbook" of fBm formulae was compiled in [195]. A summary of COST-257 work with Gaussian traffic is included in the hypertext version of this document.

### 3.2.3 Multifractal analysis and modelling

Multifractal analysis is a mathematical discipline that studies a singular measure $\mu$ through local Hölder exponents

$$\alpha(x) = \lim_{\epsilon \downarrow 0} \frac{\log \mu(B(x,\epsilon))}{\log \lambda(B(x,\epsilon))} \quad \text{(if the limit exists)}, \tag{3.2}$$

where $\lambda$ denotes Lebesgue measure and $B(x,\epsilon)$ is an open ball. Multifractal analysis is also a data analysis method that studies the statistics of approximate, finite resolution Hölder exponents computed from the data.

The data, interpreted in this case as a measure $\mu$ on time interval $[0,t]$, are said to possess multifractal scaling, if the partition function

$$S_k(\eta) = \sum_{i=1}^{2^k} \mu\left(\left(\frac{i-1}{2^k}, \frac{i}{2^k}\right]\right)^{\eta}$$

is linear with respect to $k$ in log-log-plot for $\eta$ in some relatively large interval $[\eta_{\min}, \eta_{\max}]$. Riedi and Levy Vehel [220] observed that network traffic has, at least in their data, very good multifractal scaling properties. They computed so-called multifractal spectra, which showed that the character of short time scale burstiness of real traffic differs from models used that far, including the self-similar ones that can be called "monofractal". This also inspired some work in COST-257.

The multifractal scaling property was indeed found in data from the Finnish University Network, see [158, 157]. The latter TD shows also that popular traffic models yield in multifractal analysis different spectra than real traffic.

In hope to find a logically simple prototype model for traffic with multifractal characteristics, the following class of processes was defined and analysed in [159]. Let $\Lambda_t$ be a stationary Markov process taking only finitely many

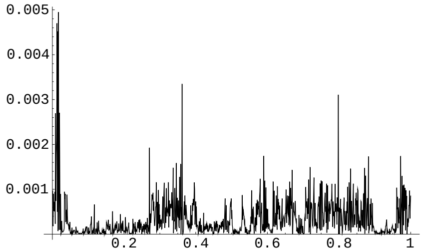


Figure 3.1: Realization of increments of $A^{(7)}$ when $\Lambda$ is a two-state Markov process with transition intensities $\nu_1 = 2$, $\nu_2 = 1/2$, transmission rates $S_1 = 1/3$, $S_2 = 7/6$, and $b = 4$. The resolution is 0.001.

values that are all positive and such that $\mathrm{E}\Lambda_t = 1$, and let $b > 1$. Define the increasing processes

$$A_t^{(n)} = \int_0^t \prod_{i=0}^{n} \Lambda_{b^i t}^{(i)} \,\mathrm{d}t. \tag{3.3}$$

If there exists a non-degenerate limit process as $n \to \infty$, this process has very well-behaving multifractal properties, besides having stationary increments. As regards the existence, it is proved in [159] that $A^{(\infty)}$ exists as an $L^2$ limit of $A^{(n)}$ if and only if $b > \mathrm{E}\Lambda_t^2$. As an example what these processes look like, figure 3.1 shows the increments of a realization of $A^{(7)}$.

# 3.3 Models for particular services and protocols

## 3.3.1 Models for Internet dial-up traffic

A popular way of accessing the Internet is to use the public telephone network with modems or ISDN dial-up connections. In [74, 75] scalable models were developed for dimensioning purposes based on comprehensive traffic measurements and analysis study of dial-up traffic. It has been found that dial-up sessions are much longer than classical telephone sessions, and their length depending on the tariffing scheme. The session holding time and session inter-arrival time exhibit high variability over the whole day. However, for modeling purposes during busy hours a renewal process with a hyperexponential distribution with only the first two moments matched is an adequate model with respect to loss behaviour.

### 3.3.2 Models for Internet user traffic

The modelling of Internet traffic is one of the main targets of recent teletraffic research. The contribution of the World Wide Web (WWW) has experienced an extreme increase in the last decade and its expected future dominance called for a number of studies in COST-257 [119, 271, 151].

The characteristics of recorded WWW traffic was analyzed in [119] considering the IP and session arrival statistics. This work also studies the impact of WWW traffic when ATM is the underlying transport mechanism. Based on the traffic characterization results a simulation study has been carried out to investigate the queueing behaviour. The main finding of this study is that variations on long time scales are not present in the queueing system if dynamic capacity allocation is used. More precisely, if the capacity of the server is increased with an amount proportional to the mean rate of the new TCP connection, the long-term correlations disappear when this TCP connection is opened. From a traffic modelling point of view this is advantageous since the application of short-range dependent models (e.g. Markovian models), which are analytically more tractable than long-range dependent models (e.g. fractal models), is adequate in these cases. However, the practical application of such a scheme is still an unsolved issue since the information needed for optimal capacity allocation is unknown until the connection is closed.

In [271] a simple model for a WWW session is constructed based on a measurement and analysis study on a local Ethernet segment. It turned out that both the distribution of the response size and the distribution of the inter-response time have the heavy-tailed property and both quantities are well modelled by the Pareto distribution. Moreover, the samples of both values are found to be independent. The model is validated by the simulated data transmission over an ATM link utilizing VBR service category. The validation of this model also shows that it exhibits stronger short-term dependencies and lacks the long-range dependencies.

A study of different Internet services is presented in [151]. The traffic, which can be categorized to be stream or elastic traffic, generated by different services (e.g. encyclopedia, tele-conference, tele-education, video player, virtual world) is characterized based on the characteristics of relevant events (sessions, connections, transactions, packets). The study provides guidelines about the appropriate choice of distributions and parameters for building appropriate traffic models.

### 3.3.3 Models of TCP traffic

Teletraffic theory has been quite succesful in modelling communication systems during the 20th century. The success is partly due to the possibility of separating the modelling of telecommunication systems as a number of arrival streams that can be treated as input to rather complex service systems. The very nature of TCP invalidates this approach due to the inherent feedback from the network model to the source model. This together with the increasing importance of the Internet has created a demand for traffic models specifically addressed to TCP traffic. Ideally such a model should reflect the feedback mechanism which is introduced by the flow control algorithm. This research field is still quite immature in spite of its importance.

Some work in the area has been performed within the COST-257 project. The work reported in [35] partially addresses the feed back problem by modelling first order effects. The models consist of a number of Markovian TCP source models and a TCP network model. The latter is a single server queue with limited buffer space and deterministic server. The probability of loss is calculated with a Poisson process as input process. The TCP source model is an approximate model of TCP Tahoe with slow start and congestion avoidance. The calibration of the two models is done by adjusting the segment loss probability. The segment loss probability influences the total load offered by the TCP model and that load in turn specifies the loss probability in the network model. The two models are solved iteratively until the segment loss probability in the two models agree.

## 3.4 Statistical methods related to traffic models

The applicability of source models is ultimately tested by comparison with measurements. The statistical problem of actually fitting parameters to source models is generally quite challenging. For this reason a significant amount of work within this area focus on special properties of the traffic processes typically but not restricted to first and second order properties of the measured traffic and the abstract models.

Several methods for estimation of the important parameter $H$ related to self-similar traffic models are discussed in [180]. These methods are validated through the analysis of a number of traces obtained by measurements in the ATM network of Telecom Finland. The conclusion of the paper is that the estimation of the Hurst parameter is highly sensitive but more important is that the Hurst parameter probably is meaningful only when considering pure self-similar traffic (that is traffic well described by fractional Brownian motion).

The study [179] addresses the importance of long-range dependence with respect to network engineering. The study is performed by several simulation experiments based on data obtained from measurements of the ATM network of Telecom Finland. The conclusion of the paper is that there is a certain time scale of importance, the size of which is related to the buffer size of the system to be engineered. The actual scaling factor is left for further research.

A related study [8] questions the usefulness of certain random permutations that have been used [67] for establishing the effect of long-range dependence for queueing systems. Two kinds of random permutations, the so called internal and external shuffling are considered. It is shown in [8] that the former more or less preserves both short and long term correlations while the latter reduces short term correlations significantly. The results indicate that it is very difficult to use shuffling as a means to investigate the importance of long-range dependence on queueing performance.

One important second order descriptor of an arrival process is the peakedness functional [60, 161]. This descriptor is analysed in detail for discrete time processes in [178]. The refinement necessary for discrete time is performed and peakedness is calculated for a number of important discrete time models. The potential for the application of peakedness as a powerful descriptor is illustrated by several examples of real traces.

A standard method in performance modeling is to analyse a system with an aggregated stream, which is obtained from a superposition of simple ON-OFF sources. An inherent problem of this approach is the dimension of the aggregated system. This problem has been addressed in [248]. The approach is to fit the spectral properties of the superposition with a model of lower dimension. The approach is well known in continuous time but this work modifies the method for discrete time. The method is applied to the superposition of MPEG video sources and the results obtained are validated through experiments. The results of this experiment are quite promising.

Most traffic models have an emphasis on second order properties of the counting process. There are several equivalent ways of describing second order properties like the Index of Dispersion for Counts, the autocorrelation function of the intensity, the Palm function, peakedness, and spectral analysis of the autocorrelation function. The results of [7] analyse the sensitivity of other properties when the point process is used as the input process to a queue. It is shown that neither second order properties of counts nor of intervals are sufficient to predict queueing behaviour. It is concluded that in order to rely on second order information it is necessary to have more knowledge as can be obtained from measurements.

The investigation of modem users of a dial-in modem pool is carried out in [273, 274]. It is found that session duration and asymmetry between up and down stream traffic volumes are surprisingly independent for varying modem speed. The traffic volume and the obtained data rates were found to increase proportionally to the access speed. This study is relevant in the design of asymmetric access methods like ADSL, since it shows that the degree of asymmetry must not be too high.

## 3.5 Traffic generation algorithms

A fast algorithm called $\text{RMD}_{mn}$ for generating Gaussian traffic was presented in [194] and implemented as a freely distributable C program in the case of fractional Brownian motion. The idea is to improve the well-known Random Midpoint Displacement algorithm (see figure 3.2), exact only for Brownian motion, by conditioning to the (at most) $m$ nearest already generated neighbor values on the left and $n$ on the right. Experiments showed that $m$ and $n$ can be quite small, for example 4 gives excellent traces of fBm. For a general

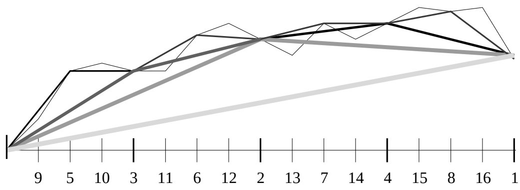


Figure 3.2: Random Midpoint Displacement algorithm for generating the path of a Gaussian process top-down. The numbers show the order of generated points.

Gaussian process with stationary increments, the algorithm takes as input the variance function (see page 45). Although the algorithm has a top-down character, a version proceeding indefinitely in time has been designed and made available. The memory requirement grows only logarithmically w.r.t. the number of generated points.