The Pseudo Self-Similar Traffic Model: Application and Validation

Rachid El Abdouni Khayari Ramin Sadre
Boudewijn Haverkort
Laboratory for Performance Evaluation and Distributed Systems
Department of Computer Science, RWTH Aachen, 52056 Aachen, Germany

Alexander Ost
Ericsson Eurolab Germany
52134 Herzogenrath, Germany

Abstract

Since the early 1990's, a variety of studies has shown that network traffic, both for local- and wide-area networks, has self-similar properties. This led to new approaches in network traffic modelling because most traditional traffic approaches result in the underestimation of performance measures of interest. Instead of developing completely new traffic models, a number of researchers have proposed to adapt traditional traffic modelling approaches to incorporate aspects of self-similarity. The motivation for doing so is the hope to be able to reuse techniques and tools that have been developed in the past and with which experience has been gained.

One such approach for a traffic model that incorporates aspects of self-similarity is the so-called pseudo self-similar traffic model. *This model is appealing, as it is easy to understand and easily embedded in Markovian performance evaluation studies.*

In applying this model in a number of cases, we have perceived various problems which we initially thought were particular to these specific cases. However, we recently have been able to show that these problems are fundamental to the pseudo self-similar traffic model.

In this paper we review the pseudo self-similar traffic model and discuss its fundamental shortcomings. As far as we know, this is the first paper that discusses these shortcomings formally. We also report on ongoing work to overcome some of these problems.

1. Introduction

Extensive measurements in the 1990's have revealed the presence of long-term correlations, often denoted as *self-similarity, fractality* and *long-range dependency,* in network traffic. The seminal paper by Leland *et al.* [28] showed self-similarity in Ethernet traffic. Later, many others revealed similar properties in wide-area traffic, signaling traffic, high-speed network traffic and in multimedia and video traffic [6, 9, 10, 11, 15, 17, 18, 32]. Many studies have shown that ignoring the *self-similarity* in the analysis of systems leads in general to an underestimation of important performance measures [32]. Additionally, various studies have shown that the presence of self-similarity is generally associated with the presence of *heavy-tail distributions* for certain entities in the network, e.g., for WWW object-size distributions or silence period lengths [26, 27].

Considerable efforts have been undertaken to develop appropriate traffic models to evaluate the performance of systems underlying a self-similar workload, see for instance [5, 12, 27]. Instead of aiming at a complete new class of traffic models, for which little analysis means are yet known, many researchers have tried to capture the self-similarity of network traffic in more traditional Markovian models of some sort, like Markov modulated Poisson processes. The benefit of using this type of model is, among other things, the availability of a large number of techniques and tools for computing performance measures for systems underlying such a workload.

In this paper, after a brief introduction to self-similarity in Section 2, we focus on the so-called *pseudo self-similar traffic* (PSST) model, as introduced by Robert and Le Boudec [35, 36], in Section 3. This model is both simple and intuitively appealing, however, when applying this model in a number of cases, we have encountered various shortcomings, on which we report in Section 4. We then show that these shortcomings are not specific to our case studies, but instead that they are fundamental to the PSST model. As a result, *the PSST model in its current form, should not be used for modelling self-similar traffic*. We briefly touch upon a number of other Markovian models for self-similar traffic in Section 6, before we conclude the paper in Section 7.

2. Self-similarity

Self-similarity is an often-observed phenomenon in nature. It means that the basic structure of an object or observation can be found at diverse (time) scales. There are many related definitions in the literature to this term, cf. [5, 27, 28]; in this paper we adhere to the following definitions.

Definition 1 *A stochastic process $X = (X_t, t \geq 0)$, with $t \in T$ (the index set) is called second-order stationary (or weakly stationary) if*

1. *its expectation is constant over time, i.e., $E[X_t] = \mu, \quad$ for all $t \in T$, and*
2. *its covariance function γ is shift-invariant, i.e., $\gamma(X_{t_1+s}, X_{t_2+s}) = \gamma(X_{t_1}, X_{t_2}), \quad$ for all $s, t_1, t_2 \in T$.*

Definition 2 *An aggregated stochastic process $X^{(m)}$ is obtained from a stochastic process X by "averaging" over non-overlapping blocks of size m, that is, for $k = 1, 2, \cdots$:*

$$X_k^{(m)} = \frac{1}{m}\left(X_{km-m+1} + \cdots + X_{km}\right). \qquad (1)$$

Note that $X^{(m)}$ is weakly stationary if X is weakly stationary.

Definition 3 *A stochastic process $X = (X_t, t \geq 0)$ is called exactly self-similar with Hurst parameter H if*

$$X =_d m^{1-H} X^{(m)}, \qquad \text{for all } \ m = 1, 2, \cdots \qquad (2)$$

This definition implies that the aggregated process $X^{(m)}$ is related to X via a simple scaling relationship involving H in the sense of finite-dimensional distributions (denoted by $=_d$), cf. [27, Section 1.4.1.2].

Definition 4 *A stochastic process $X = (X_t, t \geq 0)$ is called exactly second-order self-similar if the aggregated processes $X^{(m)}$ has the same correlation structure as X, that is,*

$$r^{(m)}(k) = r(k), \quad \text{for all } \ m = 1, 2, \cdots \text{ and } k = 1, 2, \cdots,$$

where $r^{(m)}(k)$ denotes the autocorrelation function at lag k of the aggregated process $X^{(m)}$ and $r(k)$ denotes the autocorrelation function at lag k of the original stochastic process X.

Definition 5 *A process is called asymptotically second-order self-similar if*

$$r^{(m)}(k) \sim r(k), \quad m \to \infty.$$

Self-similar processes have the so-called property of *long-range dependency*, i.e., the autocorrelation function decays hyperbolically. This implies that $\sum_k r(k) \to \infty$. In contrast, *short-range dependency* implies an exponentially decaying autocorrelation function for which $\sum_k r(k) < \infty$.

The Hurst parameter defines *the degree of self-similarity* and expresses the rate of decay of the autocorrelation function. From (2) we obtain (for details, see [27, Section 1.4.1.2]):

$$\text{var}[X^{(m)}] \sim \alpha m^{-\beta}, \quad \beta = 2 - 2H, \quad 0 < \beta < 1. \quad (3)$$

There are various methods to identify self-similar processes, such as R/S-analysis, Whittle's maximum likelihood estimator, or the variance-time plot method [18, 28]. For the purpose of this paper, it suffices to use the variance-time plot method. This method estimates the Hurst parameter H from a graph of var$[X^{(m)}]$ versus m, plotted on a log-log scale. An example of such a variance-time plot is given in Figure 2, which will be discussed later. From (3), we derive that

$$\log(\text{var}[X^{(m)}]) \sim \log \alpha - \beta \log m,$$

so that β emerges as the negative gradient in the above mentioned plot. Using a linear regression technique on this plot, we can estimate β and, hence, H.

3. The pseudo self-similar traffic model

We describe the PSST model, as introduced in [35, 36], in Section 3.1 and discuss the computation of its parameters in Section 3.2. A continuous-time variant of the model is presented in Section 3.3

3.1. Model definition

Model description. The PSST model attempts to characterise traffic self-similarity by the use of a discrete-time Markov modulated Bernoulli process (MMBP), i.e., the discrete-time analog of a Markov modulated Poisson process. The modulating Markov chain has n states, numbered 0 through $n-1$, and its corresponding state transition diagram is depicted in Figure 1. Its one-step transition probability matrix is given as:

$$\mathbf{A} = \begin{pmatrix} \Sigma_0 & 1/a & 1/a^2 & \cdots & 1/a^{n-1} \\ q/a & \Sigma_1 & 0 & \cdots & 0 \\ (q/a)^2 & 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (q/a)^{n-1} & 0 & 0 & \cdots & \Sigma_{n-1} \end{pmatrix},$$

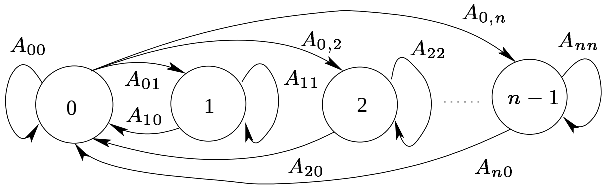


Figure 1. The state-transition diagram for the modulating Markov chain of the PSST model.

with $\Sigma_0 = 1 - \frac{1}{a} - \frac{1}{a^2} - \cdots - \frac{1}{a^{n-1}}$ and $\Sigma_i = 1 - (\frac{q}{a})^i$, for $i = 1, \cdots, n-1$.

At every discrete time step, a state transition, possibly a self-loop, takes place in the modulating chain. Only upon entry in state 0 a packet arrival takes place. As can be observed, the PSST model is completely specified by the three parameters q, n and a. This makes the model attractive, as it requires only three parameters to be set, e.g., based on some fitting procedure.

Notice that the parameters a and q need to fulfill certain conditions so that $\mathbf{A}$ is indeed a stochastic matrix describing a discrete-time Markov chain: $q, a > 0$, $q \leq a$ and a such that $0 \leq A_{0,0} \leq 1$.

In the sequel, we denote with $A^k_{i,j}$ the entry in row i and column j of $\mathbf{A}^k$. We furthermore define $N = (N_t, t \in \mathbb{N})$ as the discrete-time stochastic process describing the number of arrivals over time, as described by the MMBP.

Moments. Using the notation and terminology of the MMPP cookbook [38], we can derive the following results for the first and second moment of the number of arrivals N in an interval of length 1, i.e., per discrete time step:

$$E[N] = \underline{\pi}\mathbf{\Lambda}\underline{e} \quad \text{and} \quad E[N^2] = \underline{\pi}\mathbf{\Lambda}^2\underline{e}, \tag{4}$$

where

- $\underline{\pi}$ is the steady-state solution of the ergodic DTMC given by $\mathbf{A}$, that is, $\underline{\pi} = \underline{\pi}\mathbf{A}$ and $\sum_i \pi_i = 1$; it can easily be shown (by substitution) that
$$\underline{\pi} = (\pi_0, \cdots, \pi_{n-1}) = \frac{1 - 1/q}{1 - 1/q^n}\left(1, \frac{1}{q}, \cdots, \frac{1}{q^{n-1}}\right);$$
- $\underline{e} = (1, 1, \cdots, 1)^T$, a column vector of just 1's;
- and the $n \times n$-matrix $\mathbf{\Lambda}$ has the simple form:
$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

Notice that (4) is in a form typical for Markov modulated arrival processes [38]. However, given the above explicit expressions for $\underline{\pi}$ and $\mathbf{\Lambda}$, we can generalise and reduce (4) as follows. The k-th moment of the number of packet arrivals per unit time is given as:

$$E[N^k] = E[N] = \frac{1 - 1/q}{1 - 1/q^n} = \pi_0. \tag{5}$$

Note that the first moment of N can also easily be derived in the following way: an arrival takes place with probability 1, whenever the current state i is occupied (with happens, on the long run, with probability π_i) and the next state is state 0 (which happens with transition probability $A_{i,0}$). Hence, we have $E[N] = \sum_{i=0}^{n-1} \pi_i A_{i,0}$ which, after simple manipulations, indeed yields π_0.

Aggregated process. The m-aggregated process $N^{(m)} = (N_t^{(m)}, t \geq 0)$ is introduced, defined as the average number of arrivals over m successive intervals (preceding t):

$$N_t^{(m)} = \frac{1}{m}(N_{t-m+1} + N_{t-m+2} + \cdots + N_t), \quad t > m.$$

Since N is second-order stationary, we obtain for the first moment of $N_t^{(m)}$:

$$E[N_t^{(m)}] = E[N^{(m)}] = E[N].$$

For the variance of $N_t^{(m)}$, we follow the definition, that is, var$[N_t^{(m)}] =$ var$[N^{(m)}] = E[(N^{(m)})^2] - E[N^{(m)}]^2$, which can be reduced to [35, 36]:

$$\text{var}[N^{(m)}] = \frac{1}{m}E[N^2] - E[N]^2 + \frac{2}{m^2}\sum_{i=1}^{m-1}(m-i)\underline{\pi}\mathbf{\Lambda}\mathbf{A}^i\mathbf{\Lambda}\underline{e}. \tag{6}$$

The matrix $\mathbf{\Lambda}\mathbf{A}^i\mathbf{\Lambda}$ is an $n \times n$ matrix consisting completely of zeroes except for the non-zero entry $A^i_{0,0}$ in the upper left corner. An explicit expression for the autocorrelation of the m-aggregated process at lag k, that is, $r^{(m)}(k)$, cannot be easily obtained.

Finally, note that the expressions for $E[N]$ and var$[N^{(m)}]$ in [36] contain typographical errors.

3.2 Computation of the parameters n, q and a

In this section we assume that the expectation $E[N]$, the variance var$[N^{(m)}]$, and the Hurst parameter H of the process under study are known, that is, they have been obtained from a trace using some estimation procedure. Given these (required) workload parameters, we describe how the model parameters n, q and a for the PSST can be computed. Note that the iterative recipe given below has been proposed in [36]; its simplicity makes it attractive to use. It is not the aim of the current paper to improve on this scheme.

Computation of n: The value for n is chosen by experience. It is suggested that values around $n = 6$ give good results in most cases [35, 36]. We used similar values in our experiments [12, 3, 31].

Computation of q: The Newton iterative method is used to solve the non-linear equation (5) in order to compute q from a known estimate for $E[N]$ and given n.

Computation of a: Assume that the Hurst parameter H of the measured workload has been estimated from the log-log plot of var$[N^{(m)}]$ against m, for instance using a least-squares fitting procedure. From (6) we see that var$[N^{(m)}]$ depends, via the entry $A^i_{0,0}$ in the summation, on the actual value of a. Thus, implicitly a function $\mathcal{V}(a)$ is defined that yields, for given a, the function of var$[N^{(m)}]$ against m. Hence, for a starting value $\hat{a}$, we can estimate the negative gradient of log var$[N^{(m)}]$ against log m, giving $\hat{\beta}$ and $\hat{H}$ (an estimate for H). If $\hat{H}$ differs from the measured value for H, we compute a next estimate for $\hat{a}$, using an interval splitting procedure, and iterate until we have achieved the desired accuracy. We do not address the issue of uniqueness of the found value for a.

We illustrate this procedure in Section 4.

3.3 Continuous-time variant

In this section we derive an alternative representation of the PSST model as a phase-type renewal process [30]. We use that representation for transforming the approach to the continuous-time domain, so that numerical analysis tools for the the evaluation of continuous-time Markov chains can be applied as well.

Phase-type representation. As the PSST model generates packet arrivals only upon entering state 0, the PSST arrival process forms a renewal process. Due to the Markovian nature of the modulating process, its renewal times can be described as time to absorption (towards state 0) in an absorbing Markov chain, i.e., as a phase-type distribution. That phase-type distribution is easily obtained by replacing all transitions to state 0 by transitions to the phase-type distribution's absorbing state, and by setting the initial probability for state 0 to 1, resulting in the representation (for background on this notation, see [30, Eq. (2.2.8)]):

$$\mathbf{T} = \begin{pmatrix} 0 & 1/a & 1/a^2 & \cdots & 1/a^{n-1} \\ 0 & \Sigma_1 & 0 & \cdots & 0 \\ 0 & 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Sigma_{n-1} \end{pmatrix},$$

with initial probability vector $\underline{\alpha} = (1, 0, \ldots, 0)$. Using standard results for phase-type distributions, higher moments of the PSST interarrival time distribution can easily be obtained. Due to its phase-type nature, we refer to the corresponding arrival process as pseudo self-similar phase-type process (PSSP).

Continuous-time variant. The phase-type PSSP representation of the PSST can be used to derive a continuous-time analogue of the PSST model by transforming the discrete-time phase-type distribution to the continuous-time domain, replacing $\mathbf{P}$ by a transition rate matrix $\mathbf{Q} = [q_{i,j}]$. In [3, 12], two approaches were investigated for obtaining $\mathbf{Q}$: (i) by matching the mean sojourn times in each state of the phase-type distribution, thereby setting $\mathbf{Q} := \mathbf{P} - \mathbf{I}$, or (ii) by matching the sojourn times' probability distribution functions in the discrete- and continuous-time domains at integer points, requiring $1 - e^{q_{i,i}t} = 1 - (1 - p_{i,i})^t$, thus $q_{i,i} = \ln(p_{i,i})$, and choosing the remaining rates to reflect the original transition probabilities. As the latter approach changes the PSSP's traffic intensity, it has not been followed further, however.

Using the first approach to derive a continuous-time PSSP representation, matrix-geometric tools for the numerical analysis of continuous-time Markov chains can be employed to evaluate queueing models of PSSP|MAP|1 type, i.e., a PSSP as arrival process, a single server and Markov-modulated service times. We resorted to the the tool SPN2MGM [21, 22] to perform several case studies, where we observed a number of peculiarities (see next section) that led to the current paper.

4. Application

In Section 4.1 we show how well the PSST model performs when matching its parameters to data obtained from measurements. We then apply, in Section 4.2, the PSST model in queueing analysis.

4.1 Fitting the PSST model to traces

We have applied the PSST models in a number of case studies [3, 12, 31]. For the experiments in the current paper, we have used two different data trace:

- The access traces from the **RWTH sunsite webserver** already examined in [3]. It contains the log entries of approximately 640,000 accesses to the server and was collected during two weeks in 1998. For that trace, the mean time between arriving requests was 1.459 seconds and we computed a Hurst parameter $H = 0.927$.

- The **DEC WWW access logfiles** [13] with approximately 15$\cdot$10^6 events, having a mean interarrival time of about 0.0761 seconds and a Hurst parameter $H = 0.99$.

First of all, for both traces we invariably found that with moderate n (always less than 10), we could fit $E[N]$ as well as the Hurst parameter H accurately, i.e., with relative errors smaller than 0.1%, using the procedure outlined in Section 3.2. The PSSP-parameters found for PSSPs with n=4, 6, 8 and 10 states are listed in Table 1. For both traces, notice the large values of a that appear; these result in Markovian models that are stiff, and hence, diffucult to solve numerically.

	RWTH		DEC	
n	q	a	q	a
4	3.10658	6304.83	0.260535	4.53e+06
6	3.17248	6173.44	0.477608	4.35e+06
8	3.17859	6173.44	0.615499	4.25e+06
10	3.17921	6173.44	0.705391	4.15e+06

Table 1. Parameters of the PSSPs.

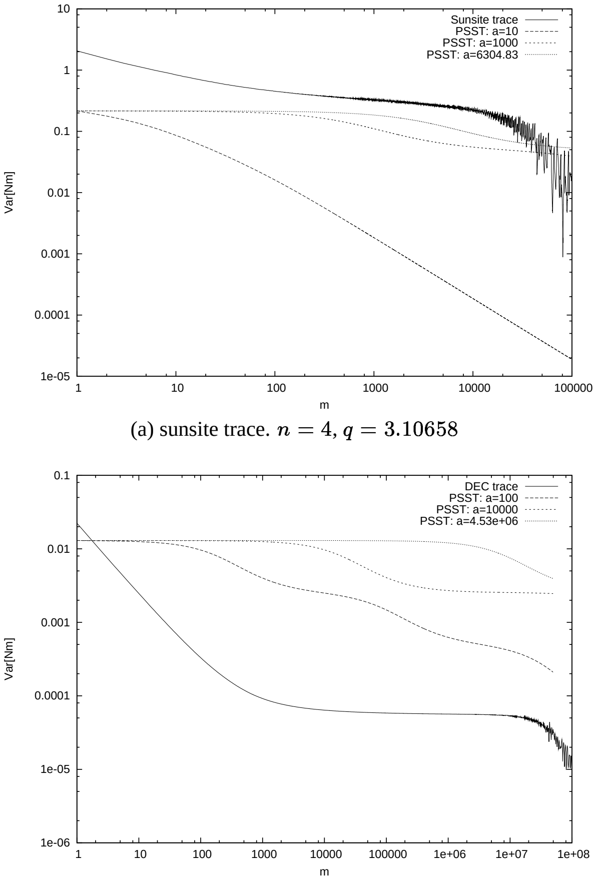
We also invariably observed a rather large difference between the absolute value of var$[N^{(m)}]$ of the model, in comparison with similar metrics directly obtained from the measurements. As an example of such a difference, observe Figure 2, in which we graphically display var$[N^{(m)}]$ against m (on a log-log scale), for both traces and their corresponding PSST models.

The curves obtained for $a = 6304.83$ (sunsite trace) resp. a=4.53e+06 (DEC trace) are nearly parallel to the curves for the real traces, i.e., the Hurst parameter is well estimated, but the absolute values differ considerably. Thus, even though the PSST model allows for a good fit of H and $E[N]$, a good fit for var$[N^{(m)}]$ is *not* necessarily the case. As already described in [3, 31], we think the bad variance fit is responsible for the bad queueing performance predictions made with the model.

4.2 The PSST model in queueing analyses

In this section we investigate to which extent PSSPs are an appropriate substitute for real traffic as far as the performance measures obtained in a model-based evaluation are concerned. This question is particularly interesting since such a modeling-oriented validation has not been accomplished for the original approach suggested in [35, 36].

Our investigation is based on a comparison study for a simple queueing system. In order to minimize the impact of the service process on our study, we consider a system where arriving jobs are subject to exponential services. Taking real traffic measurements as starting point, we first derive performance figures using trace-driven simulation. We then compare these results to a numerical analysis of the PSSP|M|1 model that results from matching the trace with a PSSP.



(a) sunsite trace. $n = 4$, $q = 3.10658$

(b) DEC trace. $n = 4$, $q = 0.260535$

Figure 2. Comparison between the effective value of the variance var$[N^{(m)}]$ (from the traces) and the results obtained from the PSST models as a function of the parameter m for different values of a.

Tool Environment. For the simulation study, a custom-developed simulation package has been used which is capable to process roughly 10^6 events/second.

For the numerical analysis of the PSSP|M|1 model, we developed an infinite-state stochastic Petri net model (iSPN) [31]. iSPNs have been designed to simplify the specification of Markovian quasi-birth-and-death processes, which arise in many queueing scenarios (all MAP|MAP|1 models, and thus also the PSSP|M|1 model, are covered). The key difference to normal SPNs is the introduction of an unbounded place that can hold an arbitrary number of tokens, and which essentially keeps track of the number of

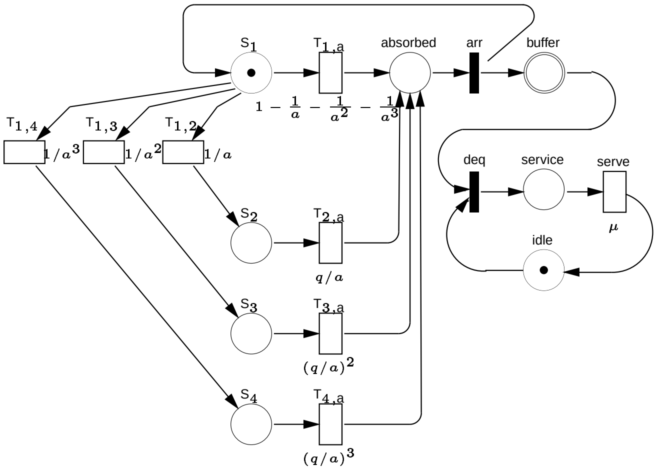


Figure 3. The iSPN for a PSSP|M|1 system with arrivals according to a 4-state PSSP.

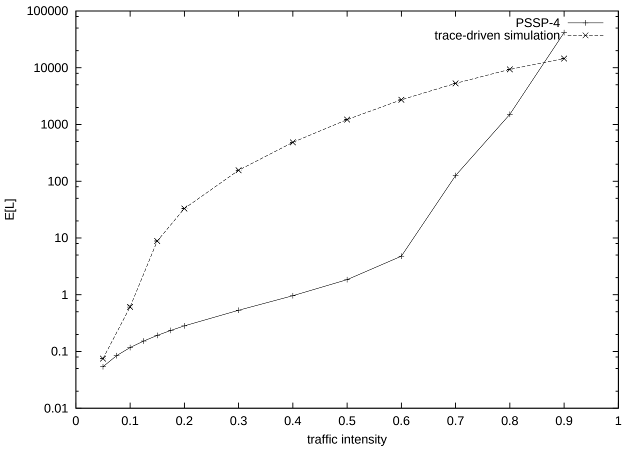


Figure 4. The mean buffer size at different saturation levels.

jobs waiting to be served. The iSPN for a PSSP|M|1 model, with $n = 4$, is shown in Figure 3; the iSPN clearly resembles the modulating Markov chain in Figure 1.

For the specification and evaluation of the PSSP|M|1 iSPN model, the tool SPN2MGM has been used. SPN2MGM allows to derive reward-based performance measure from iSPNs by applying efficient solution algorithms to solve the underlying quasi-birth-and-death Markov chains. For further details on iSPNs, we refer to [21, 22, 31].

Numerical Results. We first note that the exponential decay of transition rates in a PSSP phase-type distribution quickly leads to stiff Markov chains; this is especially the case for large values of a. In many cases, ill-conditioned boundary equations did not allow to numerically derive the QBD's steady-state solution, even when the QBDs were small. For the examples in Table 1, the minimum and maximum transition rates in the PSSP differ by factors as large as $2.5 \cdot 10^{11}$ (sunsite trace; $N = 4$) and $8.4 \cdot 10^{60}$ (DEC trace; $N = 10$). Only for the first parameter combination reliable numerical results could be obtained.

For that parameter combination, Figure 4 depicts the mean number of queued customers E$[L]$ in the PSSP|M|1 system for different traffic intensity levels (obtained by varying the service rate). It can be observed that the PSSP model yields much too optimistic results most of the time, with mean buffer occupancies being orders of magnitude smaller than those derived by the trace-driven simulation. Furthermore, the PSSP model is not capable to capture the impact of the traffic intensity on the mean buffer size in a qualitatively correct manner; at higher traffic intensities, the effect of the traffic intensity on the mean buffer length becomes stronger for PSSPs, which is in contrast to the simulation case.

Except for extremely large traffic intensities, this dramatic discrepancy can be explained by looking at the queue-length distribution for both trace simulation and PSSP. For an example traffic intensity level of 0.6, it can be observed in Fig. 5(a) that the probability of small queue lengths is much larger in the PSSP case than for the real trace, except for queue lengths 0 and 1. In strong contrast to this, for queue lengths larger than about 100 (see Fig. 5(b)), the probabilities quickly approach zero in the PSSP case, but they remain much larger for the original trace. Evidently, the heavy tail of the queue length distribution in the simulation case leads to much higher mean queue lengths than in the PSSP case. We have found the same behaviour for other traces as well [8]. Note that the observed slow decay of the queue length distribution tail for self-similar traffic is in accordance with the theoretical results in [7, 14].

As second reason for the largely differing results we found that the PSSP approach is only capable to match the Hurst parameter and the first moment of the interarrival time distribution; however, higher-order statistics usually heavily impact the performance of a queueing system as well. This can be clearly seen when examining the transient behaviour of the PSSP|M|1 system.

Transient behaviour. To examine its transient behaviour, we also simulated the iSPN for the PSSP|M|1 system. Figure 6 exemplarily shows the evolution of the queue length for the PSSP|M|1 system as well as for the trace-driven $\cdot$|M|1 queue as recorded during one simulation run at traffic intensity 0.75. It seems that the PSSP is able to match the

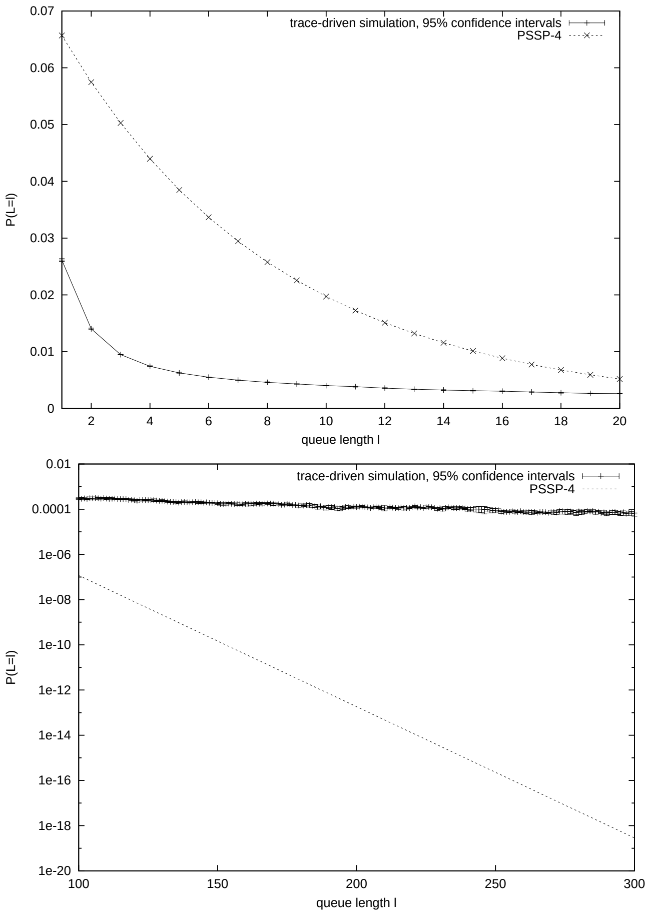


Figure 5. Queue length distribution of the PSSP$|$M$|$1 sample system at traffic intensity 0.6 for queue lengths l (a) up to 20, and (b) over 100.

general character of the trace data: both curves show sudden increases of the queue length stemming from bursts in the arrival rate. However, the curves also show that the variations produced by the PSSP are much smaller and shorter. Although both queues experience the same traffic intensity, the trace-driven simulation results in peak queue lengths ten times higher than found in the PSSP simulation. We think that this directly follows from the mismatch of higher-order statistics by the PSSP model.

5 Formal validation

In this section we analyse the PSST model in more detail in order to find the cause for the differences in variance and the misleading queueing performance as illustrated in the previous section.

By simplifying (6), using the fact that $E[N^k] = E[N]$ for all $k \geq 1$ (cf. (5)), and exploiting the special structure of $\mathbf{\Lambda}$, we obtain an efficient method to compute var$[N^{(m)}]$ for various values of m:

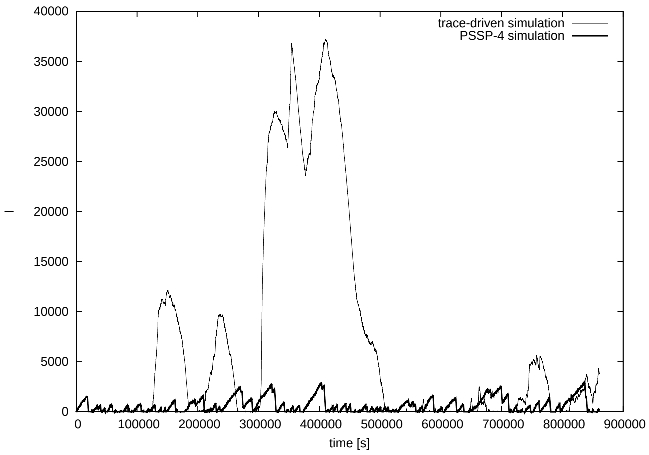


Figure 6. Queue length evolution over time at traffic intensity 0.75.

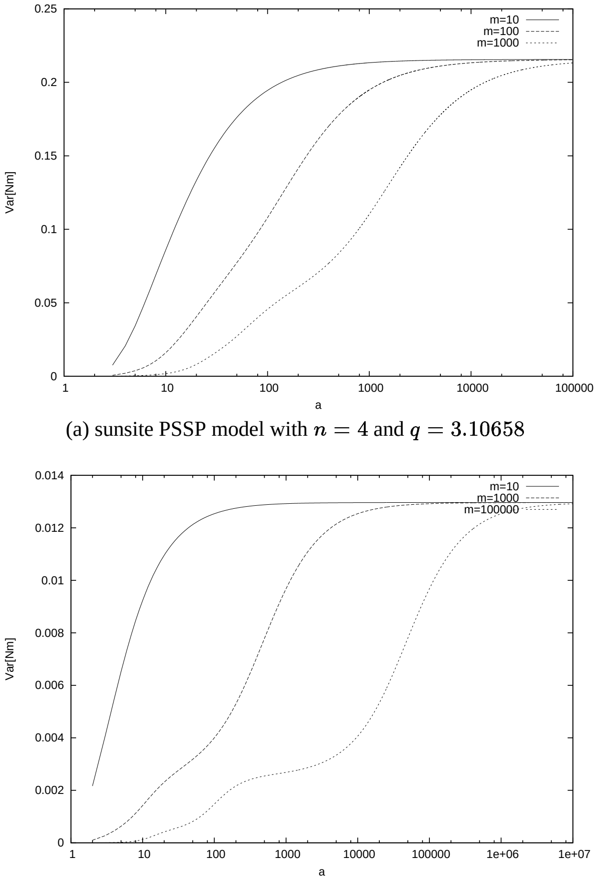
$$\begin{aligned}
\text{var}[N^{(m)}] &= \frac{1}{m}E[N^2] - E[N]^2 \\
&\quad + \frac{2}{m^2}E[N]\sum_{i=1}^{m-1}(m-i)A_{0,0}^{i} \\
&= \frac{E[N]}{m}\left(1 + \frac{2}{m}\sum_{i=1}^{m-1}(m-i)A_{0,0}^{i}\right) \\
&\quad - E[N]^2. \qquad (7)
\end{aligned}$$

Note that in order to compute var$[N^{(m)}]$ we only need to compute the first column of the matrices $\mathbf{A}^i$ ($i = 1, \cdots, m-2$) and not the complete matrices $\mathbf{A}^i$, for all $i = 1, \cdots, m-1$. This implies an important reduction in computational complexity, especially seen in light of the fact that var$[N^{(m)}]$ has to be computed repeatedly in the iterative procedure to compute a.

Let us now discuss the relation between var$[N^{(m)}]$, a and m. To do so, we first consider var$[N^{(m)}]$ as a function of a for three fixed values for m as given in Figure 7 (where a change of a incurs a change of H, but that does not bother us at this point).

As can be observed, var$[N^{(m)}]$ monotonously increases with a, however, seemingly towards an upper bound somewhere between 0.20 and 0.25 for the sunsite PSSP model resp. 0.13 for the DEC PSSP model. It can indeed be proven that this upperbound exists and that it is reached for $a \to \infty$. To do so, we proceed in two steps:

1. We first prove var$[N^{(m)}] \leq E[N] - E[N]^2$.
   Since $\mathbf{A}$ is a stochastic matrix we always have $0 \leq$



(a) sunsite PSSP model with $n = 4$ and $q = 3.10658$

(b) DEC PSSP model with $n = 4$ and $q = 0.260535$

Figure 7. var$[N^{(m)}]$ as a function of a.

$A^i_{0,0} \leq 1$. Equation (7) shows that var$[N^{(m)}]$ is linearly dependent on $A^i_{0,0}$, with positive coefficient $m - i$, so that we obtain an upper bound for var$[N^{(m)}]$ by setting $A^i_{0,0}$ to 1:

$$\begin{aligned}\text{var}[N^{(m)}] &\leq \frac{E[N]}{m}\left(1 + \frac{2}{m}\sum_{i=1}^{m-1}(m-i)\cdot 1\right) \\ &\quad - E[N]^2 \\ &= \frac{E[N]}{m}\left(1 + \frac{2}{m}\frac{m(m-1)}{2}\right) - E[N]^2 \\ &= E[N] - E[N]^2.\end{aligned}$$

2. We now prove $\lim_{a\to\infty}$ var$[N^{(m)}] = E[N] - E[N]^2$. Using Equation (7), we rewrite the right-hand side of it as follows:

$$\frac{E[N]}{m}\left(1 + \frac{2}{m}\sum_{i=1}^{m-1}(m-i)\lim_{a\to\infty} A^i_{0,0}\right) - E[N]^2,$$

in which we can limit the term $A^i_{0,0}$ by 1, so that we obtain

$$\frac{E[N]}{m}\left(1 + \frac{2}{m}\sum_{i=1}^{m-1}(m-i)\cdot 1\right) - E[N]^2.$$

Using the same transformation as before, we can rewrite this into $E[N] - E[N]^2$.

We thus have proved the following theorem.

Theorem 1 *In the PSSP model, as defined in Section 3, var$[N^{(m)}]$ is bounded (from above) by $E[N] - E[N]^2$. The limiting value is reached for $m \to \infty$.*

The limiting value of var$[N^{(m)}] = E[N] - E[N]^2$ can be further reduced using (5) as follows:

$$\begin{aligned}\lim_{a\to\infty}\text{var}[N^{(m)}] &= E[N] - E[N]^2 \\ &= \frac{E[N]}{q}\left(\frac{1 - 1/q^{n-1}}{1 - 1/q^n}\right) \\ &\approx \frac{E[N]}{q}, \qquad (8)\end{aligned}$$

where the approximation can be understood from the fact that the ratio $\left(\frac{1-1/q^{n-1}}{1-1/q^n}\right)$ is close to 1, for $q \geq 2$ and n not too small. This fact is illustrated in Figure 8 in which we show this ratio (upper three curves), as a function of q, for three values of n. In the same figure, we also show for the same three values of n, $E[N]$ as a function of q (lower three curves). Figure 8 should now be read as follows. For a given scenario, pick the value $E[N]$ that has to be modelled on the y-axis. Find the corresponding value of q to this value. For the thus found value of q, the ratio can be read from the set of upper curves. As is clear from the figure, for $E[N] \geq 0.65$, we find $q \geq 3$, so that the ratio is very close to 1.

The upper bound for var$[N^{(m)}]$ is of the form $U(x) = x - x^2$, with $x = E[N]$. The function $U(x)$, a parabola, achieves its maximum at $x = \frac{1}{2}$, being $U(\frac{1}{2}) = \frac{1}{4}$. Moreover, in our case $x = E[N]$, which takes values between 0 and 1. Hence, var$[N^{(m)}]$ is always non-negative, as required. However, if $E[N] \neq \frac{1}{2}$, $U(E[N]) < \frac{1}{4}$. In the example at hand, we have $E[N] = 0.685$, so that var$[N^{(m)}]$ must be bounded by $U(0.685) = 0.216$, which is indeed the case, as can be observed in Figure 7. Thus, we have proven the following theorem.

Theorem 2 *In the PSSP model, var$[N^{(m)}] \leq \frac{1}{4}$.*

Recalling that $E[N]$ can be interpreted as the traffic intensity generated by the PSST model, we note that for large traffic intensities, that is, $E[N]$ close to 1, the maximum achieved variance var$[N^{(m)}]$ becomes smaller and smaller.

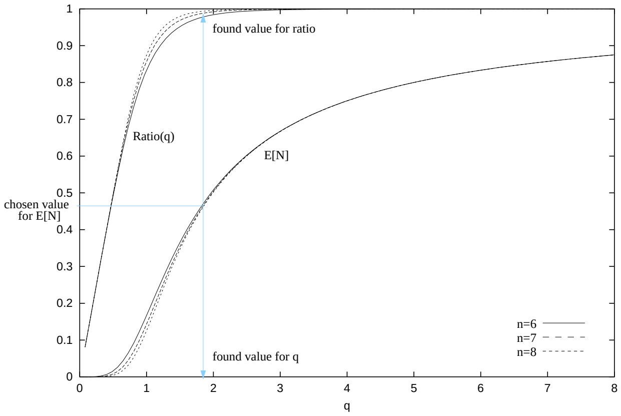


Figure 8. Ratio $\frac{1-1/q^{n-1}}{1-1/q^n}$ (upper three curves) and $E[N]$ (lower three curves) as a function of q, for $n = 6, 7$ and 8.

In conclusion, we have formally derived that even though the PSST model does allow for the correct fitting of both $E[N]$ and H, values of var$[N^{(m)}]$ larger than $\frac{1}{4}$ can not be obtained with it. Moreover, under high load conditions where $E[N]$ comes close to 1, var$[N^{(m)}]$ will have to approach 0. For these reasons, we suggest not to use the PSST model in any modelling study.

6. Alternative models and fitting procedures

In the previous section we have shown that the PSST model has its shortcomings in modelling self-similar traffic. Nevertheless, the idea of adequately modelling self-similar traffic using Markovian models remains appealing. In this section, we briefly touch upon a number of recent approaches in this direction, without aiming at completeness.

Early work on fitting MMPP's. As early as 1986, Heffes and Lucantoni reported on a procedure to match the four parameters of a 2-state MMPP (Markov modulated Poisson processes) to interarrival measurements [23]. As parameters to be matched they proposed the mean arrival rate, the variance to mean ratio of the number of arrivals in the interval $[0, t)$ and in the long-term interval $[0, t)_{t\to\infty}$, as well as the third moment of the number of arrivals in a limited interval. The authors report good queueing performance when the model is applied for packetized voice sources.

In this context also the work of Meier-Hellstern and Fisher on fitting Markov-modulated arrival processes should be mentioned [38, 29]. In [29] a maximum-likelihood optimisation is used for the transition density matrix.

In 1991, Gusella reported on similar work, in which he proposed to use the index of dispersion for counts and the index of dispersion for intervals as parameters to match [20].

Notice that these papers did appear before the notions of self-similarity, long-range dependence and heavy-tail distributions had become apparent in the context of network traffic modelling.

Successive superposition of two-state MMPP's. The fitting procedure developed by Andersen and Nielsen bases on matching of second-order properties of counts (number of arrivals in a certain interval) between the model and the measurement data [4]. The model comprises superpositioning of a number of two-state MMPPs. The authors suggest that, typically, the use of four two-state MMPP's (leading to a 16-state MMPP) suffices to model highly-variable traffic with long-range dependencies. The method matches five traffic characteristics: the mean arrival rate of the process, the lag-1 correlation, the Hurst parameter, the number of MMPP's to be superposed and the number of time scales to be taken into consideration. Although the model seems to be acceptable for describing second-order traffic properties, it does not appear to be suitable to predict queueing behaviour; indeed, the authors argue that the number of the employed traffic descriptors might not be sufficient for that. Furthermore, the suggested fitting procedure has the drawback that the number of parameters to be fitted can grow beyond five, the number of traffic characteristics being matched. This leaves open the problem of how to deal with the remaining degrees of freedom.

Using hyper-exponentials. Feldmann and Whitt proposed a model and a fitting procedure for handling heavy-tailed distributions, based on a mixture of hyper-exponentials [1]. Such distributions could be used to model independent interarrival times. We recently reported on an alternative fitting procedure for this type of (interarrival time) distribution [33], based on the EM-algorithm [37]. Even when our fitting procedure is more costly, it allows a far better fit, especially with respect to higher-order characteristics (variance, skewness). Furthermore, when using these distributions for describing service times in queueing models, very good performance predictions have been made (when compared to trace-driven simulations). We still need to extend this work towards interarrival time distributions; in that case, we will also have to take into account the correlations between successive arrivals.

Separate treatment of short- and long-range dependence. The fitting method by Horvath, Rozsa and Telek is based on the superpositioning of a phase-type renewal process and an interrupted Poisson process, in oder to capture both long- and short-range dependence [2]. The traffic descriptors to be fitted are the arrival rate, the index of dispersion for counts $I(t) = \text{var}[N_t]/E[N_t]$ (for two different values of t), and the Hurst parameter. To approximate the heavy-tailed distribution of the interarrival times, a hyperexponential distribution is proposed (to be fitted with the algorithm of Feldmann and Whitt) [1].

Unfortunately, their results did not yet show a good fit for the traffic statistics nor for the queue length distribution (in the analysis of a $\cdot|D|1$-queue).

MMPP exhibiting multifractal behaviour. Horvath and Telek recently proposed [25] the use of a special MMPP with a symmetric n-dimensional cube form, thereby employing so-called Haar wavelet theory [34]. In fact, the compostion of the proposed MMPP structure is similar to the generation of the Haar wavelet transform [24]. Starting at the largest considered time scale, with an arrival rate equal to 1, the model is generated in an iterative fashion. At the next (finer) time scale, a new cube is generated so that the structure of the MMPP remains unchanged as well as the behaviour at the previously addressed time scales. At level n, the MMPP comprises 2^n states and has $n+2$ parameters, n of which are computed by minimising the relative errors of the second moment of Haar wavelet coefficients. The other two remaining parameters are determined by observing the "best-looking" fit. The proposed model seems to be useful for approximating the fractal behaviour of the considered trace. Furthermore, good results have been observed for the queue length distribution, in particular for higher utilizations.

7. Conclusions and outlook

In this paper we have analysed and validated the recently-proposed pseudo self-similar traffic model [36]. Earlier work, cf. [3, 31] only showed empirically that this model is less suitable for use in queueing model evaluations, even when it does correctly characterise the traffic intensity as well as the Hurst parameter. In the current paper, we have extended our empirical evaluation of this model, thus making our earlier claim more firm.

Furthermore, we have formally analysed the PSST model and shown, in a case-independent fashion, a major shortcoming of it, being its inability to capture the variance of the traffic process adequately.

In a broader context, the aim of this paper has been to show that even when traffic models appear to perform well in a number of cases studies (using a number of traces), these models still have to undergo a thorough analysis in order to establish what can, and what cannot be described with them. Too often a model is presented as a good one, on the basis of only a small number of cases (if more than one at all).

As the current paper shows, developing traffic models dealing adequately with self-similarity remains a big challenge. One of the problems for all the developed models is their generalisation to other scenarios than those explicitly tested for. In the current paper, we have formally proved that a proposed models does have certain limitations. When proposing new models, it is required to prove that "such limitations" are not present; it is currently unclear, however, which limitations, and which not, need to be considered in such a proof. Other unresolved issues are the selection of the proper traffic characteristics for the fitting procedures to work with, and the desired quality measures for the fitted model (at the traffic streeam level, or at the level of queue performance).

We finally note that in some recent studies, the issues of heavy-tailness and long-range dependencies are questioned again [16, 19]. Indeed, these authors claim that not so much the tail-behaviour of, for instance, file-size distributions is important for system performance, but rather their "waist-behaviour". Similarly, it is claimed that TCP does only generate strong correlation structures over a limited range of time scales. These new insights seem to suggest that there is still room for Markovian models, as we have recently shown [33].

Acknowledgements. The work of Rachid El Abdouni Khayari and Ramin Sadre is supported by the german science foundation DFG in project HA 2966/2 ("Planung moderner Kommunikationsnetze unter Berücksichtigung realitätsgetreuer Verkehrsstrukturen").

We thank the anonymous reviewers for their constructive comments that helped improving the paper. We thank Peter Kemper (University of Dortmund) for presenting a prelimianry version of this paper at the International Performance and Dependability Symposium 2002 in Washington, DC, USA.

References

[1] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–258, 1998.

[2] A. Horvath, G.I. Rozsa, and M. Telek. A MAP fitting method to approximate real traffic behavior. In *Proc. 8th IFIP Workshop on Performance Modelling of ATM & IP Networks*, pages 33/1–12, Illkley, UK, 2001.

[3] A. Ost and B. R. Haverkort. Modeling and evaluation of pseudo self-similar traffic with infinite-state

stochastic Petri nets. In M. Ajmone Marsan, J. Quemada, T. Robles, and M. Silva, editors, *Formal Methods and Telecommunications*, pages 120–136. Prensas Universitarias de Zaragoza, 1999.

[4] A. T. Andersen and B. F. Nielsen. An application of superpositions of two-state Markovian sources to the modelling of self-similar behaviour. *IEEE Journal on Selected Areas in Communications*, 16(5):719–732, 1998.

[5] B. F. Nielsen. Modelling long-range dependent and heavy-tailed phenomena by matrix analytic methods. In G. Latouche and P. Taylor, editors, *Advances in Algorithmic Methods for Stochastic Models*, pages 265–278. Notable Publications, Inc., 2000.

[6] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43(2–4):1566–1579, February/March/April 1995.

[7] S.C. Borst, O.J. Boxma, and R. Núñez-Queija. Heavy tails: The effect of the service discipline. In T. Field, P.G. Harrison, J. Bradley, and U. Harder, editors, *Computer Performance Evaluation, Lecture Notes in Computer Science 2324*, pages 1–30. Springer-Verlag, 2002.

[8] D. Brocker. Messung und Modellierung komplexer Verkehrsstrukturen in Hochgeschwindigkeitsnetzen. Master's thesis, RWTH Aachen, Lehr- und Forschungsgebiet Verteilte Systeme, September 1998.

[9] T. Chiotis, F. Stanatelopoulos, and B. Maglaris. Traffic source models for realistic ATM performance modelling. In *Proc. 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, UK, July 1997.

[10] T. Chiotis, C. Stathis, and B. Maglaris. The impact of self-similarity on the statistical multiplexing of MPEG video data. In *Proc. 6th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, UK, July 1998.

[11] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

[12] D. Brocker. Messung und Modellierung komplexer Verkehrsstrukturen in Hochgeschwindigkeitsnetzen. Master's thesis, RWTH Aachen, Department of Computer Science, Germany, 1998.

[13] Digital Equipment Cooperation. Digital's Web Proxy Traces. ftp://ftp.digital.com/pub/DEC/traces/proxy.

[14] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Math. Proc. Cam. Phil. Soc.*, 118:363–374, 1995. Available at `ftp://www.stp.dias.ie/DAPG/dapg9330.ps`.

[15] D. E. Duffy, A. A. McIntosh, M. Rosenstein, and W. Willinger. Statistical analysis of CCSN/SS7 traffic data from working subnetworks. *IEEE Journal on Selected Areas in Communications*, 12(3):544–551, 1994.

[16] D. R. Figueiredo, B. Liu, V. Misra, and D. Towsley. On the Autocorrelation Structure of TCP Traffic. *Computer Networks Journal Special Issue on "Advances in Modeling and Engineering of Long-Range Dependent Traffic", to appear*, 2002.

[17] 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, 1991.

[18] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proc. ACM SIGCOMM '94*, volume 24 of *Computer Communications Review*, pages 269–280, London, October 1994.

[19] W. Gong, Y. Liu, V. Misra, and D. Towsley. On the Tails of Web File Size Distributions. In *Proceeding of the thirty-Ninth Annual Allerton Conference on Communication, Control and Computing*, October 2001.

[20] R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2):203–211, 1991.

[21] B. R. Haverkort. SPN2MGM: Tool support for matrix-geometric stochastic Petri nets. In *Proceedings of the 2nd International Computer Performance and Dependability Symposium*, pages 219–228. IEEE Computer Society Press, 1996.

[22] B. R. Haverkort and A. Ost. Steady-state analysis of infinite stochastic Petri nets: A comparison between the spectral expansion and the matrix-geometric method. In *Proceedings of the 7th International Workshop on Petri Nets and Performance Models*, pages 36–45. IEEE Computer Society Press, 1997.

[23] H. Heffes and D.M. Lucantoni. A Markov modulated characterisation of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, 1986.

[24] A. Horvath and T. Telek. Markovain Modeling of Real Traffic: Heuristic Phase Type and MAP Fitting of Heavy Tailed and Fractal Like Samples. In M. C. Calzarossa and S. Tucci, editors, *Proceeding of Performance Evaluation of Complex Systems: Techniques and Tools*, pages 405–434, 2002.

[25] A. Horvath and T. Telek. A Markovian Point Process Exhibiting Multifractal Behaviour and its Application to Traffic Modeling. In *Proceeding of MAM4*, Adelaide, Australia, 2002.

[26] K. Park, G. Kim, and M. Crovella. On he relationsship between file sizes, transport protocols, and self-similar network traffic. In *IEEE International Conference on Network Systems*, pages 171–180, 1996.

[27] K. Park and W. Willinger. *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, 2000.

[28] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. In *Proc. ACM SIGCOMM '93*, volume 23 of *Computer Communications Review*, pages 183–193, October 1993.

[29] K.S. Meier-Hellstern. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research*, 29:370–377, 1987.

[30] M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins University Press, 1981.

[31] A. Ost. *Performance of Communication Systems: A Model-Based Approach with Matrix-Geometric Methods*. Springer, March 2001.

[32] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[33] R. El Abdouni Khayari, R. Sadre, and B.R. Haverkort. Fitting world-wide web request traces with the EM-Algorithm. In R. van der Mei and F. Huebner-Szabo de Bucs, editors, *Proceedings of SPIE*, volume 4523, pages 211–220, Denver, USA, August 2001.

[34] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A Multifractal Wavelet Model with Application to Network Traffic. *IEEE Transactions on Information Theory*, 45:992–1018, April 1999.

[35] S. Robert. *Modélisation markovienne du trafic dans les réseaux de communication*. PhD thesis, EPFL, Lausanne, Switzerland, 1996.

[36] S. Robert and J.-Y. Le Boudec. New models for pseudo self-similar traffic. *Performance Evaluation*, 30:57–68, 1997.

[37] S. Asmussen and O. Nerman. Fitting phase-type distributions via the EM algorithm. In *Symposium i Anvendt Statistik, Copenhagen*, pages 335–346, 1991.

[38] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.