Proceedings of the 13th ITC Specialist Seminar on IP Measurement, Modeling and Management, Monterey, CA, September 2000, pp. 21-1-21-10

Evaluation of Effective Bandwidth Schemes for Self-Similar Traffic

Stefan Bodamer
University of Stuttgart, Inst. of Communication Networks and Computer Engineering
E-mail: bodamer@ind.uni-stuttgart.de

Joachim Charzinski
Siemens AG Munich
Information and Communication Networks
E-mail: joachim.charzinski@icn.siemens.de

Abstract

In this paper, different approximations for the effective bandwidth of self-similar traffic streams are reviewed. Among those, classical approaches originally based on Markovian models are regarded in the context of self-similar traffic. On the other hand, a solution is considered that explicitly takes the long-range dependent character of the traffic stream into account by using a fractional Brownian motion model. Furthermore, we propose an effective bandwidth scheme that provides a combination of approaches from both domains. In order to achieve an objective comparison of the different schemes, an M/Pareto fluid burst has been chosen as a common traffic model, and the corresponding traffic parameters needed by the different schemes have been derived. A burst level simulation of the same traffic model serves as another reference. The evaluation shows that the accuracy of the results obtained using the different schemes depends very much on the traffic parameters.

1. Introduction

Variable bit rate connections offer the chance of exploiting a statistical multiplexing gain – an effect describing that in order to transport the traffic of a number of variable bit rate connections, less than the sum of all peak rates is needed if a small loss or delay is acceptable. The effective bandwidth is one way of characterizing the resource requirements of a variable rate connection. It has originally been introduced in the context of connection admission control for ATM networks but as a measure of resource consumption it can also be applied in optimal charging or network dimensioning tasks. The latter is the context in which this paper has been written.

Most of today's data traffic is carried over the Internet's transmission control protocol TCP. This protocol performs a flow control by which each connection adapts its bandwidth to the maximum fair share available. This makes it difficult to use classical dimensioning methods or quality of service (QoS) parameters [5]. However, even under these circumstances, a possible dimensioning target can be to adjust the capacity of one link in a network such that this link is definitely (e.g., with respect to a certain packet loss probability) not the bottleneck for TCP connections using it.

Several approaches for determining effective bandwidths have been derived for Markovian or other short-range dependent traffic. On the other hand, Internet traffic has been found to exhibit significant amounts of self-similarity and long-range dependence [6, 11, 16, 20] due to an extremely high variability of burst durations [20]. It has been shown that a high amount of self-similarity leads to greatly increasing queue lengths as compared to short-range dependent traffic [8]. Our goal is to compare different classical effective bandwidth schemes with others explicitly considering self-similarity and to find the parameter regions in which these schemes can be used to give accurate estimates for the bandwidth needed by a traffic stream.

In Section 2, we introduce the M/Pareto traffic model as a burst scale model where Pareto distributed bursts of traffic arrive at negative exponentially distributed interarrival instants and derive parameters for characterizing this traffic. Section 3 gives an overview of some effective bandwidth schemes, indicating how they can be applied in the M/Pareto context. In addition, a new approach combining formulas for two burst scale multiplexing phenomena is presented. Finally, a parameter study carried out in Section 4 is used to compare the different effective bandwidth formulas, discussing their individual strengths and weaknesses using the M/Pareto traffic model as a common basis. Fluid flow burst scale simulation results for the same traffic model are used as a reference.

2. Traffic Model

A superposition of many ON/OFF sources with heavy-tailed ON or OFF durations has been suggested as traffic model that captures the long-range dependence effects of network traffic [20]. Among the class of heavy-tailed distributions the Pareto distribution has turned out to be most appropriate one for modelling in many cases [6]. Increasing the number of ON/OFF sources with Pareto distributed ON durations and decreasing the relative contribution of each source results in an M/Pareto model. In [12], Neame, Zukerman and Addie show that this is a quite appropriate model for long-range dependent traffic streams that can be well matched to a measured trace.

In the context of this paper an M/Pareto fluid burst model is assumed where bursts comprising a certain amount of fluid arrive according to a Poisson process with rate λ. The fluid arrival rate during a burst is denoted by h. The distribution of the burst size B (i.e. the amount of fluid arriving in a burst) follows a Pareto distribution with minimum value k and shape parameter α:

$$\mathrm{P}(B \le s) = 1 - \left(\frac{k}{s}\right)^{\alpha}, \quad s \ge k \tag{1}$$

Special interest is given to the range $1 < \alpha \le 2$ for the shape parameter leading to finite mean but infinite variance of the burst size. In this case traffic generated by the M/Pareto model is asymptotically self-similar with Hurst parameter

$$H = \frac{3-\alpha}{2}, \quad \frac{1}{2} \le H < 1 \tag{2}$$

The mean burst size $b = \mathrm{E}[B]$ is given by $b = k \cdot \alpha / (\alpha - 1)$. The mean rate of the total traffic stream is $m = \lambda \cdot b$.

The self-similar behaviour of the M/Pareto process becomes obvious when regarding the cumulated arrival process A_t, i.e. the fluid arriving in an interval of length t. The variance of A_t can be obtained by

$$\mathrm{VAR}[A_t] = 2 \cdot \lambda \cdot h \cdot \int_0^t du \int_0^u dv \int_{vh}^{\infty} \mathrm{P}(B > s) ds \tag{3}$$

Repeated integration leads to the following expression:

$$\mathrm{VAR}[A_t] = \begin{cases} \lambda h^2 \cdot \left(\frac{\alpha}{\alpha-1} \cdot \frac{kt^2}{h} - \frac{t^3}{3}\right) & 0 \le t \le \frac{k}{h} \\ c_1 \cdot t^{3-\alpha} + c_2 \cdot t + c_3 & t > \frac{k}{h} \end{cases} \tag{4}$$

where the constants are given by

$$c_1 = \frac{-2 \cdot \lambda \cdot h^2 \cdot \left(\frac{k}{h}\right)^{\alpha}}{(1-\alpha) \cdot (2-\alpha) \cdot (3-\alpha)}$$

$$c_2 = -\lambda \cdot k^2 \cdot \frac{\alpha}{2-\alpha}$$

$$c_3 = \lambda \cdot \frac{k^3}{h} \cdot \frac{\alpha}{3 \cdot (3-\alpha)} \tag{5}$$

The same result (however in a different presentation) is given in [1, 12].

If t approaches infinity only the first term of the expression for $t > k/h$ keeps relevant, i.e. the variance increases with t^{2H} in the limiting case:

$$\mathrm{VAR}[A_t] \to c_1 \cdot t^{2H}, t \to \infty \tag{6}$$

If m, b and H are used instead of λ, k and α we get

$$\mathrm{VAR}[A_t] \to \frac{m \cdot h^{2H-1} \cdot \left(\frac{2-2H}{3-2H} \cdot b\right)^{2-2H}}{(3-2H) \cdot (2H-1) \cdot H} \cdot t^{2H} \tag{7}$$

Another interpretation of this result is that the variance of the average rate observed within an interval of length t approaches $c_1 \cdot t^{2H-2}$ for $t \to \infty$. That means it decreases very slowly if the Hurst parameter is close to 1.

If the cumulated arrival process in the case of a finite variance burst size (corresponding to short-range dependent traffic) is regarded for comparison we obtain

$$\mathrm{VAR}[A_t] \to m \cdot \frac{\mathrm{E}[B^2]}{\mathrm{E}[B]} \cdot t \tag{8}$$

for $t \to \infty$, i.e. the variance of the average rate in an interval of length t decreases with t^{-1}.

3. Effective Bandwidth

3.1 Definition

The notion of effective bandwidth provides a measure of the resource requirements of a traffic stream with certain quality of service (QoS) constraints. Statistical properties of the traffic stream have to be considered as well as system parameters (e.g., buffer size, service discipline) and the traffic mix. The terms equivalent bandwidth and equivalent capacity are often used as synonyms for effective bandwidth.

A mathematical framework for effective bandwidth has been defined based on the general expression [10]

$$\alpha(s, t) = \frac{1}{st} \cdot \log \mathrm{E}[\exp(s \cdot A_t)] \tag{9}$$

which depends on the space parameter s and the time parameter t. Effective bandwidths for various types of traffic models have been derived from this definition [10]. The problem with respect to a practical usage of these expressions, however, is to find appropriate values for s and t, which depend on the QoS requirements and the system parameters. As this may become a rather complex task [7] we restrict ourselves to approximate expressions which can be derived independently of (9).

In the following, different approximations of the effective bandwidth of an M/Pareto traffic stream are presented. The fluid is assumed to be the input of a FIFO server with capacity C, which has to be defined according to the effective bandwidth of the traffic stream. The QoS constraint of all presented methods is the overflow probability, i.e. the probability that queue length Q exceeds some threshold x.

3.2 Rate Envelope Multiplexing (REM) Approximation

A rather simple method to approximate the effective bandwidth is rate envelope multiplexing (REM) [17]. Only the current total fluid arrival rate is considered and compared with the link rate neglecting the effect of buffering. Therefore it is also called bufferless approach. Another term is stationary approximation [9].

The attractive feature of this approach is that it is independent of the burst size distribution type. Only mean and peak rate are relevant. Therefore it can be directly applied to the M/Pareto traffic model introduced in Section 2.

The probability that the total arrival rate R exceeds the link rate C can be determined if the rate distribution is known. In the case of an M/G fluid burst model the exact rate distribution is given by a Poisson distribution. If the ratio m/h is large enough, a Gaussian distribution provides a reasonably good approximation.

Assuming that R is approximately distributed according to a Gaussian distribution with mean m and variance $\sigma^2 = m \cdot h$, the probability of R exceeding C can be determined:

$$\mathrm{P}(R > C) \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{\frac{C-m}{\sigma}}^{\infty} \exp\left(-\frac{y^2}{2}\right) dy \qquad (10)$$

Using a rough approximation of the Gaussian distribution as done in [9] provides a further simplification:

$$\mathrm{P}(R > C) \approx \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(C-m)^2}{2 \cdot \sigma^2}\right) = \varepsilon \qquad (11)$$

An expression for the effective bandwidth is obtained by solving (11) for C [9]:

$$C = m \cdot \left(1 + \sqrt{-2\ln\varepsilon - \ln(2\pi)} \cdot \sqrt{\frac{h}{m}}\right), \ \varepsilon < \frac{1}{\sqrt{2\pi}} \qquad (12)$$

It has to be remarked that the approximation of the Gaussian distribution that leads to (11) and (12) is not very accurate. However, we observed that ε according to (11) matches the queueing probability $\mathrm{P}(Q > 0)$, which is larger than $\mathrm{P}(R > C)$ in general, quite well. From this one can conclude that (11) may be seen as an upper bound of $\mathrm{P}(Q > x)$. Therefore, (12) can be interpreted as a strictly conservative solution.

3.3 Fluid Flow (FF) Approximation

While the solution presented in the previous section neglects buffering, the approach denoted as fluid flow approximation in [9] uses the queue length distribution in the case of exponentially distributed burst size. Basic results have been obtained by Anick, Mitra and Sondhi [2]. Assuming that traffic is generated by a superposition of ON/OFF sources with exponentially distributed phase durations, they are able to calculate the distribution of queue length Q in an infinitely large buffer by solving a system of differential equations. If the buffer threshold x is reasonably large, $\mathrm{P}(Q > x)$ is well approximated by a single exponential term corresponding to the dominant eigenvalue in the underlying equation system. If the number of sources goes to infinity maintaining a constant aggregate mean rate the following result for the M/M burst traffic model is obtained:

$$\mathrm{P}(Q > x) = \mathrm{P}(Q > 0) \cdot \exp\left(-(1-\rho) \cdot \frac{x}{b}\right) \qquad (13)$$

where $\rho = m/C$ is the system load and b denotes the mean burst size. As the calculation of $\mathrm{P}(Q > 0)$ is reasonably complex, Guérin et al. suggest to make the simplifying assumption $\mathrm{P}(Q > 0) \approx 1$ [9]. This corresponds to approximating $\mathrm{P}(Q > x)$ by the conditional probability $\mathrm{P}(Q > x | Q > 0)$ which leads to

$$\mathrm{P}(Q > x) \approx \exp\left(-\left(1 - \frac{m}{C}\right) \cdot \frac{x}{b}\right) = \varepsilon \qquad (14)$$

The effective bandwidth is obtained by solving (14):

$$C = \frac{m}{1 + \frac{b}{x} \cdot \ln\varepsilon}, \quad x/b > -\ln\varepsilon \qquad (15)$$

The result is independent of the peak rate h. Although the derivation is based on the presumption of an exponentially distributed burst size it can principally also be interpreted as a rough approximation for the effective bandwidth of an M/Pareto traffic stream with equal mean burst size thereby simply neglecting the heavy tail effect.

3.4 Fractional Brownian Motion (FBM) Approximation

A very basic traffic model that is able to capture the effect of long-range dependence is the fractional Brownian motion (FBM) model [13]. The cumulated arrival process is described by

$$A_t = mt + \sqrt{m \cdot a} \cdot Z_t \qquad (16)$$

where m and a denote the mean arrival rate and the variance coefficient (which is not the same as the coefficient of variation), respectively. The random variable Z_t repre-

Table 1: Measured FBM parameters

parameter	Bellcore 1	Bellcore 2	ADSL	local ISP
m	2279 kbit/s	12.3 kbit/s	10.5 kbit/s	8.76 kbit/s
$\hat{a}$	262.8 kbit s	68.6 kbit s	440 kbit s	38 kbit s
H	0.78	0.86	0.915	0.88

sents a normalised FBM with Hurst parameter $H \in [1/2, 1)$. Z_t is mainly characterised by zero mean and variance t^{2H} for $t > 0$. Therefore the variance of A_t is given by

$$\mathrm{VAR}[A_t] = m \cdot a \cdot t^{2H}, \ t > 0 \tag{17}$$

Additional properties of the FBM model are, e.g., discussed in [13, 14, 17].

In [13, 14], Norros presents an approach to obtain an approximation for the queue length distribution in an unlimited buffer fed by an FBM traffic stream and emptied with service rate C. Using a scaling law for the fractional Brownian storage Norros finds that the distribution of queue length Q roughly follows a Weibull distribution[1]. Like in the derivation of (14) in Section 3.3 $\mathrm{P}(Q > 0) \approx 1$ is assumed, i.e. $\mathrm{P}(Q > x)$ is approximated by $\mathrm{P}(Q > x | Q > 0)$. Then the following formula for the complementary queue length distribution is obtained:

$$\mathrm{P}(Q > x) \approx \exp\left(-\frac{(C-m)^{2H} \cdot x^{2-2H}}{2 \cdot H^{2H} \cdot (1-H)^{2-2H} \cdot a \cdot m}\right) \tag{18}$$

Solving for C yields an expression for the effective bandwidth [14]:

$$C = m \cdot \left(1 + (-2\kappa(H)^2 \ln\varepsilon)^{\frac{1}{2H}} \cdot a^{\frac{1}{2H}} \cdot x^{-\frac{1-H}{H}} \cdot m^{-\frac{2H-1}{2H}}\right) \tag{19}$$

with $\kappa(H) = H^H \cdot (1-H)^{1-H}$.

In the following, we refer to (19) as the Norros formula. The result is also is in accordance with the solution found in [10] using the general effective bandwidth definition given in (9). A discussion of effects of the parameters in the Norros formula can be found in [15].

The variance coefficient a as well as m and H may be determined by the evaluation of measurements. In a variance time plot (see Fig. 1 for an example) where $y = \log(\mathrm{VAR}[A_t]/V_u^2)$ is drawn over $x = \log(t/t_u)$ (t_u and V_u denote some time and volume unit, e.g., $t_u = 1$ s and $V_u = 1$ kbit, respectively) they define the regression line $y = r(x)$ for large values of x:

$$r(x) = \log(m \cdot \hat{a}/V_u^2) + 2H \cdot x \tag{20}$$

So $\hat{a} = a \cdot t_u^{2H}$, which is measured, e.g., in $\mathrm{kbit} \cdot s$, can be derived from the intersection of the regression line with the vertical line at $t = t_u$.

This measurement-based approach, however, will only lead to valid results for a very long measurement period. The empirical variance obtained during a short measurement interval may significantly differ from the expected variance in a long-term sense. The parameters $\hat{a}$ and H derived from the empirical variance as well as the mean rate may be useless in this case. As we will point out Section 4 even several millions of bursts may be too less in the case of high values of the Hurst parameter.

Typical values of m, $\hat{a}$ and H are listed in Table 1. While the first two parameter sets are obtained from different Ethernet traffic measurements at Bellcore [11] as given in [14], the latter two result from newer HTTP traffic measurements in an ADSL-based access network and at a local ISP [4].

An application of the Norros formula to the traffic model specified in Section 2 can be achieved by mapping the M/Pareto model to an FBM model. As pointed out in [14] this can be done by equalling the mean and the variance of the corresponding cumulated arrival processes. Regarding the variance of the M/Pareto process it is appropriate to take only the limiting term for $t \to \infty$ according to (7) into account. Equalling (7) and (17) then yields

$$a = \frac{h^{2H-1} \cdot \left(\frac{2-2H}{3-2H} \cdot b\right)^{2-2H}}{(3-2H) \cdot (2H-1) \cdot H} \tag{21}$$

as an expression for the variance coefficient of the M/Pareto process. The convergence behaviour of $\mathrm{VAR}[A_t]$ for increasing values of t depends very much on the Hurst parameter (Fig. 1). While the variance converges very fast to the term expressed in (7) for H close to 1, no convergence is achieved for $H = 1/2$. Therefore the variance coefficient of the M/Pareto process according to (21) tends to infinity for $H \to 1/2$. Furthermore, Fig. 2 shows that the value of a extremely depends on the Hurst parameter as well as on the peak rate and the mean burst size. Therefore a variation of the Hurst parameter alone without adapting the variance coefficient is not useful. Similar observations have been made by Veitch and Abry in [19].

[1] A similar result is obtained by Brichet et al. in [3] in the case of heavy traffic and heavy-tailed ON/OFF sources. Tsybakov and Georganas on the other hand obtain a hyperbolic decay of the overflow probability [18].

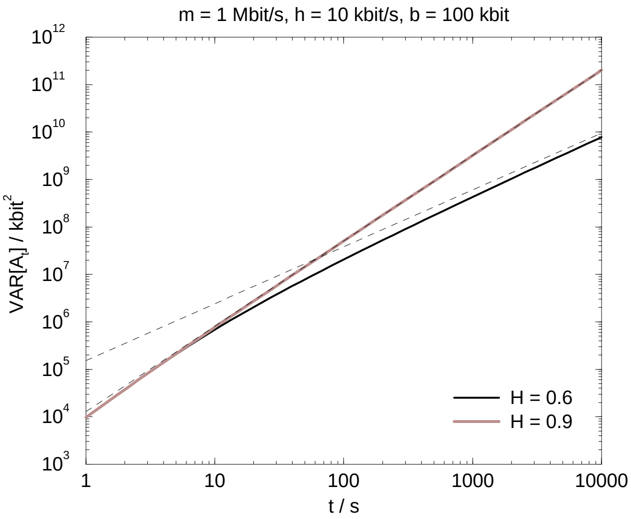

Fig. 1: Asymptotic behaviour of VAR$[A_t]$ for the M/Pareto arrival process

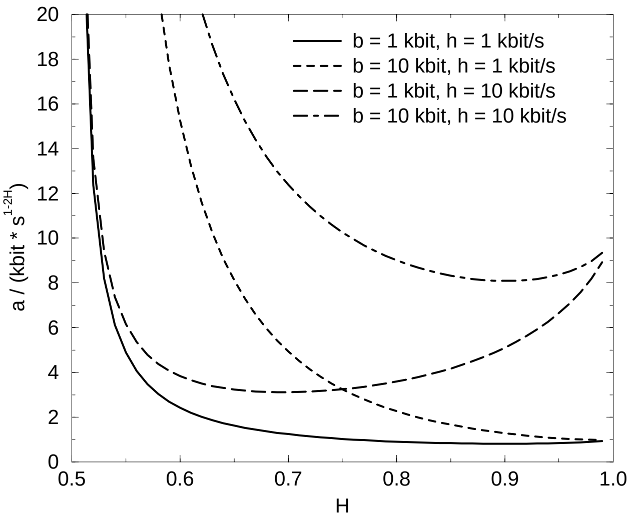

Fig. 2: Dependence of variance coefficient *a* for the M/Pareto arrival process on Hurst parameter

If a in the Norros formula is substituted by the variance coefficient of the M/Pareto process as given in (21) the following expression for the effective bandwidth of an M/Pareto traffic stream is obtained:

$$C = m \cdot \left(1 + \chi(H) \cdot (-2\ln\varepsilon)^{\frac{1}{2H}} \cdot \left(\frac{x}{b}\right)^{\frac{H-1}{H}} \cdot \left(\frac{h}{m}\right)^{\frac{2H-1}{2H}}\right) \qquad (22)$$

Therein, $\chi(H)$ is used as an abbreviation for

$$\chi(H) = 2^{\frac{1-H}{H}}(3-2H)^{-\frac{3-2H}{2H}} \cdot H^{\frac{2H-1}{2H}} \cdot (1-H)^{\frac{2-2H}{H}} \cdot (2H-1)^{-\frac{1}{2H}} \qquad (23)$$

A mapping to a short-range dependent burst arrival model ($H = 0.5$) can be done in the same way. If the burst size is assumed to be exponentially distributed the variance coefficient is given due to (8) by

$$a = \frac{\mathrm{E}[B^2]}{\mathrm{E}[B]} = 2b \qquad (24)$$

This leads to relatively simple expressions for the overflow probability (using $\rho = m/C$) and the effective bandwidth of an M/M/∞ burst traffic stream, respectively:

$$\mathrm{P}(Q > x) \approx \exp\left(-\frac{1-\rho}{\rho} \cdot \frac{x}{b}\right) \qquad (25)$$

$$C = m \cdot \left(1 - \ln\varepsilon \cdot \frac{b}{x}\right) \qquad (26)$$

Note that these results are similar to those obtained in Section 3.3 but slightly different.

3.5 Combined Method

A simple ad hoc approach to find an effective bandwidth formula that considers both rate envelope multiplexing as well as rate sharing for self-similar traffic is to combine the formulas derived in Section 3.2 and Section 3.4. While ε_{REM} according to (11) has been pointed out to be a reasonably well approximation for $\mathrm{P}(Q > 0)$, ε_{FBM} according to (18) has actually been derived as an expression for $\mathrm{P}(Q > x | Q > 0)$. Therefore the product of the results of (11) and (18) should deliver a good approximation for $\mathrm{P}(Q > x)$.

$$\begin{aligned}\varepsilon &= \varepsilon_{REM} \cdot \varepsilon_{FBM} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(C-m)^2}{2\sigma^2} - \frac{(C-m)^{2H} \cdot x^{2-2H}}{2 \cdot \kappa(H)^2 \cdot a \cdot m}\right) \qquad (27)\end{aligned}$$

Some transformations lead to an equation of the form

$$K_1 \cdot (C-m)^2 + K_2 \cdot (C-m)^{2H} + K_3 = 0 \qquad (28)$$

with constants

$$\begin{aligned}K_1 &= \frac{1}{2mh} \\ K_2 &= \frac{x^{2-2H}}{2 \cdot H^{2H} \cdot (1-H)^{2-2H} \cdot a \cdot m} \\ K_3 &= \ln(\varepsilon \cdot \sqrt{2\pi}) \qquad (29)\end{aligned}$$

Equation (28) can be solved for C using a simple numerical method. If, e.g., Newton's method with initial value $C_0 - m = h$ is used, only very few iterations are required to find a reasonably exact solution.

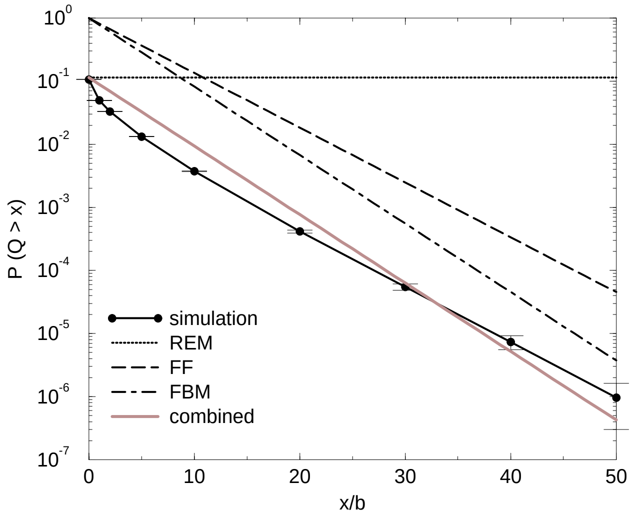

Fig. 3: Queue length ccdf for exponentially distributed burst size with $C/h = 50$ and 80% load

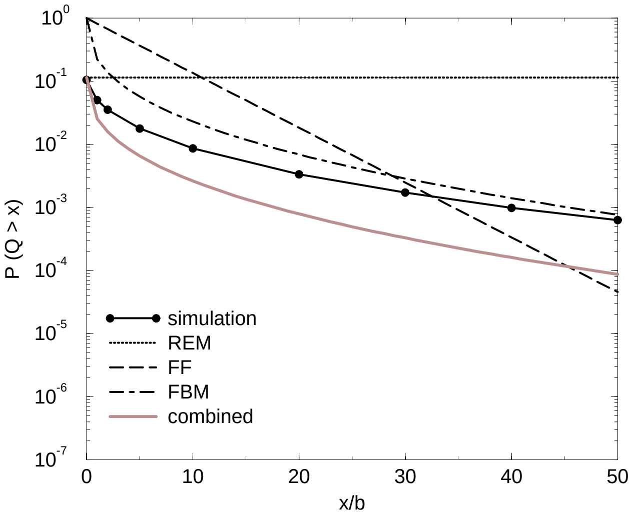

Fig. 4: Queue length ccdf for Pareto (H = 0.8) distr. burst size with $C/h = 50$ and 80% load

4. Comparative Evaluation

In this section, the effective bandwidth results obtained by the previously presented approaches are compared with each other and with simulations. The simulator directly implements the fluid burst model, i.e. only the burst level is considered in the simulations. This has mainly two advantages as compared to a packet level simulation. First, simulation time is reduced by a factor which is in the order of the number of packets per burst. Second, packet level effects, which are not considered by the previously presented effective bandwidth schemes, are omitted. The M/Pareto fluid burst model discussed in Section 2 is used in the simulation to obtain reference results. The analytical results are used as given in Section 3 for the REM, FF, FBM and combined approaches under M/Pareto traffic.

Simulation of self-similar traffic generally has to be regarded cautiously as the statistical significance may increase only very slowly with simulation time. An example may help to clarify this. In the case of short-range dependent traffic the expected standard deviation of the mean rate measured during simulation is reduced by a factor of 10 if the simulation time is prolonged by a factor of 100. If traffic is self-similar with Hurst parameter H, however, simulation duration has to be increased by a factor of $10^{1/(1-H)}$, i.e. by a factor of 10^{10} in the case of H= 0.9, to achieve the same goal. We observed that the measured mean input rate was significantly different (deviation of up to 5%) from the value specified in the simulation configuration for H = 0.9 although a huge number of bursts (10^8, in some cases 10^9) were considered in the simulations. As the effective bandwidth presented below is related to the mean rate as measured during simulation, however, the results can still be seen as reasonable estimates for the real values even in the case of H close to 1.

The example above leads us to a further remark. Although widely concealed the problem of statistical significance also occurs when performance studies on the basis of traces are performed as presented in many papers. If we assume that an average burst arriving on a network link consists of 10 packets (which surely is a reasonable number) a simulation of 10^8 bursts comprises 10^9 packets. To get the same number of packets a measurement on a fully loaded 100 Mbit/s would have to last for more than 11 hours (assuming an average packet size of 500 bytes). Within such a long period, however, the arrival process can no longer be assumed to be stationary due to diurnal variations. Therefore, traces used for performance studies presented in literature are usually much shorter, typically in the order of 10^6 packets [14]. This leads to a much lower statistical significance as compared to our simulations.

Now we first take a look on the complementary queue length distribution. In Fig. 3 and Fig. 4 the results for exponentially and Pareto (H = 0.8) distributed burst size are depicted, respectively. The offered load is 80% in both cases. A value of 40 is used for the ratio m/h. Note that m denotes the total mean rate of the traffic stream, i.e. m/h represents the mean number of simultaneously active bursts. The REM approximation is independent of the buffer threshold and returns a constant value equal to $\mathrm{P}(Q > 0)$. The FF approximation is able to follow the exponential decay of $\mathrm{P}(Q > x)$ in the exponential case. If the input traffic is self-similar, however, this method consequently underestimates the slope of the curve. The FBM

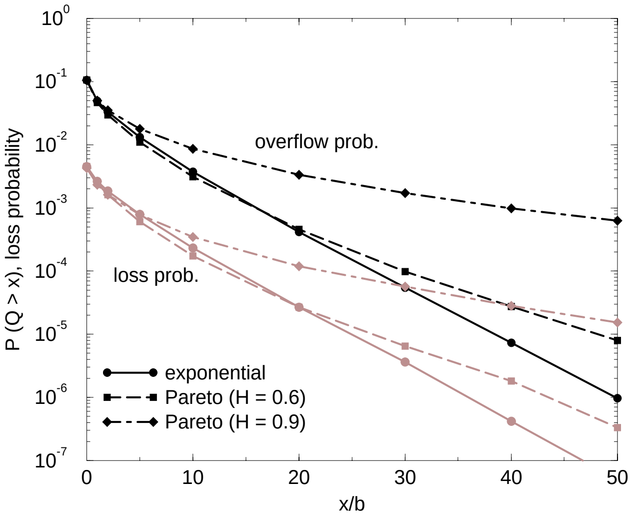

Fig. 5: Difference between overflow and loss probability for different burst size distributions

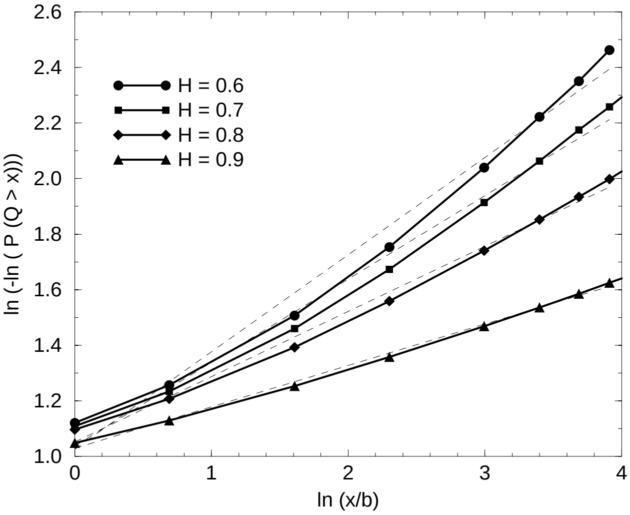

Fig. 6: Test of Q for Weibull distribution

approximation using (18) as well as (21) for the M/Pareto and (24) for the M/M case, respectively, can follow the shape of $\mathrm{P}(Q > x)$ quite well in both the self-similar and the non self-similar case. Both figures reveal, however, that the decay is slightly overestimated. This has also a negative influence on the result according to the combined method which underestimates the overflow probability, especially in the case $H = 0.8$. Unlike the FBM approximation the combined method matches the overflow probability quite well if the buffer size is very small.

The convex shape of the overflow probability curve in Fig. 4 has already indicated that the queue length distribution may be well described by a Weibull distribution as assumed in Section 3.4. But we wanted to have a closer look on that in order to see whether this property really holds in the case of M/Pareto input traffic. Therefore, we have drawn $\ln(-\ln \mathrm{P}(Q > x))$ over $\ln(x/b)$ which should give a straight line if Q follows a Weibull distribution. As obvious from Fig. 6 this is true with good accuracy at least for H close to 1. If the Hurst parameter is small, however, slight deviations can be observed.

As mentioned before all effective bandwidth schemes presented in Section 3 have in common that they use the overflow probability instead of the loss probability as QoS measure. In order to show the difference between both measures we also made simulations of a finite buffer queue. The loss probability results for buffer size x are depicted in Fig. 5 together with those of the complementary queue length distribution in the infinite buffer case. Again a ratio of $m/h = 40$ and an offered load of 80% were chosen. One can see that the loss probability curves principally have a very similar shape as compared to the overflow probability curves. This holds for different degrees of self-similarity. However, the loss probability is about one order of magnitude smaller in this case. So one can conclude that it makes a significant difference whether the effective bandwidth is related to loss or to overflow probability.

In the following, we compare the analytical results for the effective bandwidth as given by (12), (15), (22) and (27) for the different methods with those obtained by simulation for different degrees of self-similarity. The simulation results have been found by subsequently performing simulations of the queueing model with varying service rate. Like the analytical results the simulation results are based on the overflow probability ε as QoS measure using a value of 10^{-2} in all cases. The overflow probability is related to a buffer threshold of 10 and 50 times the mean burst size in the right and the left figures, respectively. We refer to this as the medium and large buffer case. The mean burst size as well as the peak rate during a burst are kept constant while the mean rate is varied.

In Fig. 7 and Fig. 8 the effective bandwidths related to the total mean rate m are drawn over m/h for the case of exponentially distributed burst size. The figures show that the effective bandwidth values according to FBM and FF approximation are proportional to the total mean rate. The effective bandwidth based on the REM approximation, on the other hand, remains unchanged if the buffer threshold is increased. While the REM solution turns out to be quite accurate in the medium buffer case (at least for higher mean rates) it is the worst solution if a large buffer is assumed. The Norros formula always yields a lower effective bandwidth than the FF approximation. This generally leads to a higher network utilisation. However, in cases of low mean rates the Norros formula obviously underestimates the required service rate. The combined method turns out to be the most accurate one for medium buffer

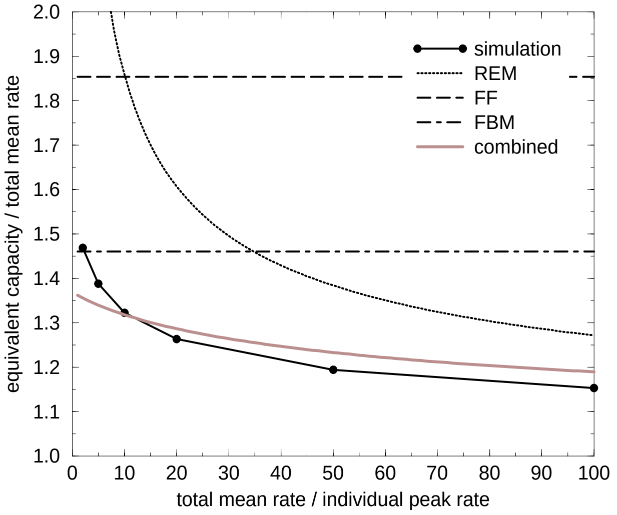

Fig. 7: M/M traffic, *x/b* = 10

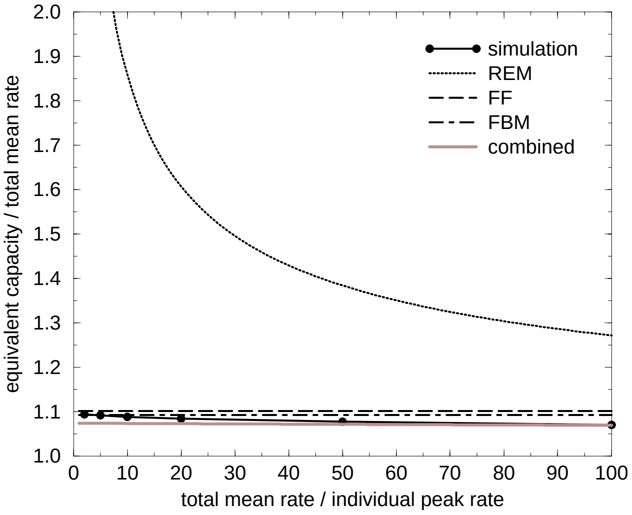

Fig. 8: M/M traffic, *x/b* = 50

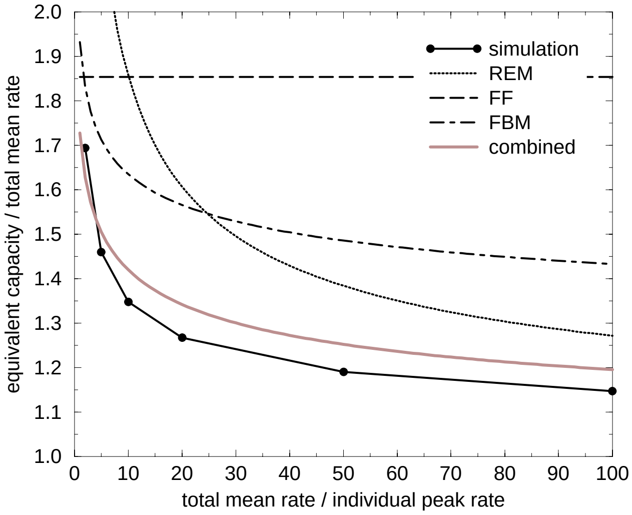

Fig. 9: M/Pareto traffic with *H* = 0.6, *x/b* = 10

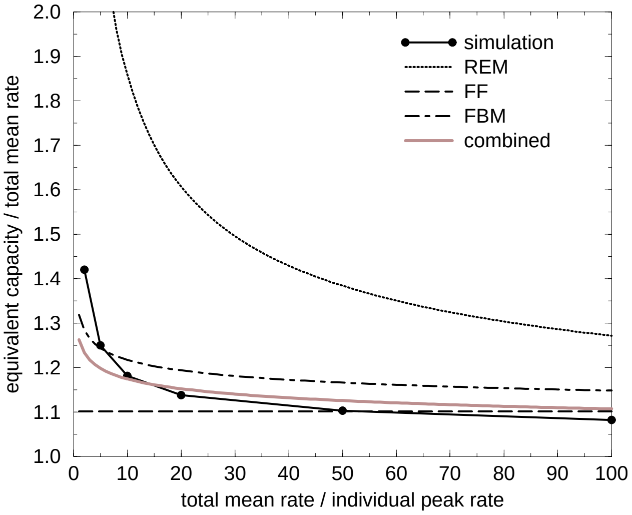

Fig. 10: M/Pareto traffic with *H* = 0.6, *x/b* = 50

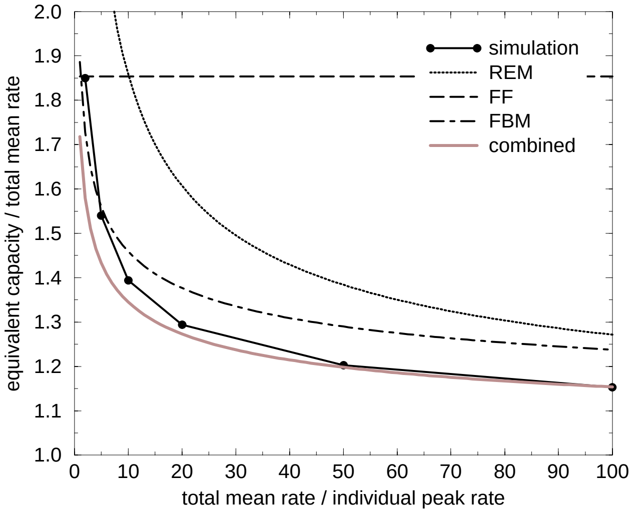

Fig. 11: M/Pareto traffic with *H* = 0.7, *x/b* = 10

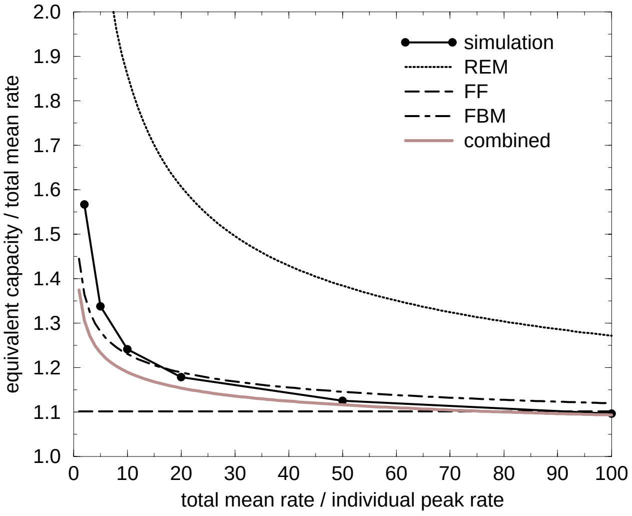

Fig. 12: M/Pareto traffic with *H* = 0.7, *x/b* = 50

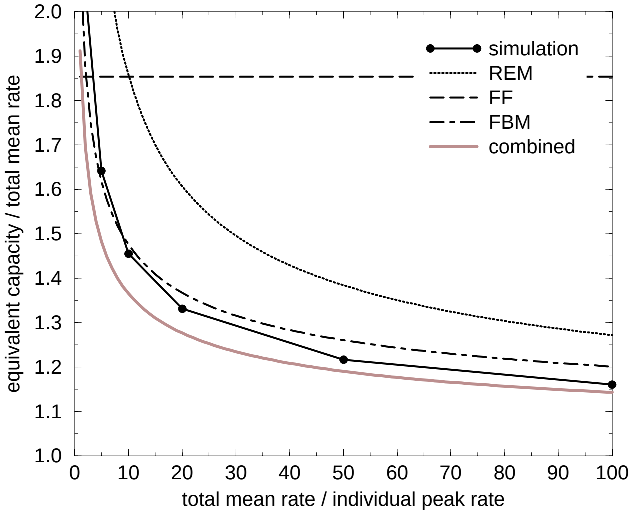

Fig. 13: M/Pareto traffic with *H* = 0.8, *x/b* = 10

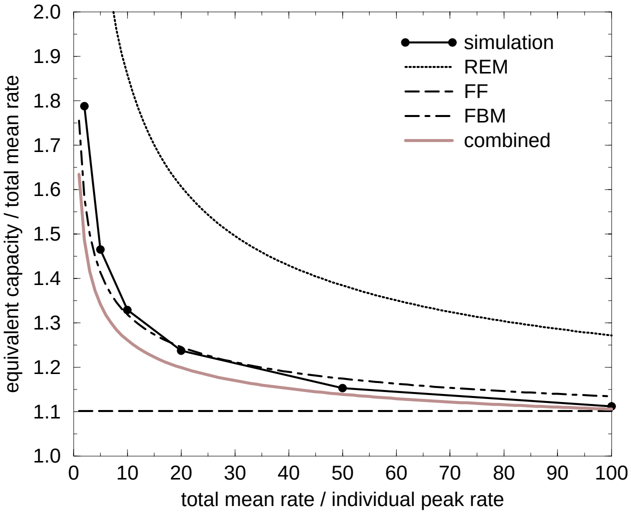

Fig. 14: M/Pareto traffic with *H* = 0.8, *x/b* = 50

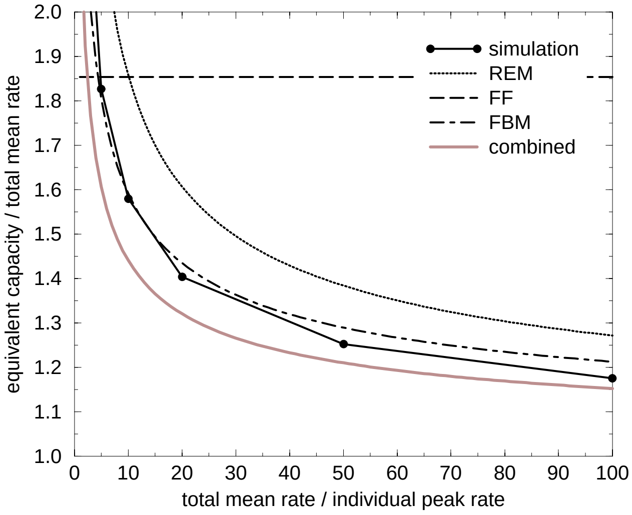

Fig. 15: M/Pareto traffic with *H* = 0.9, *x/b* = 10

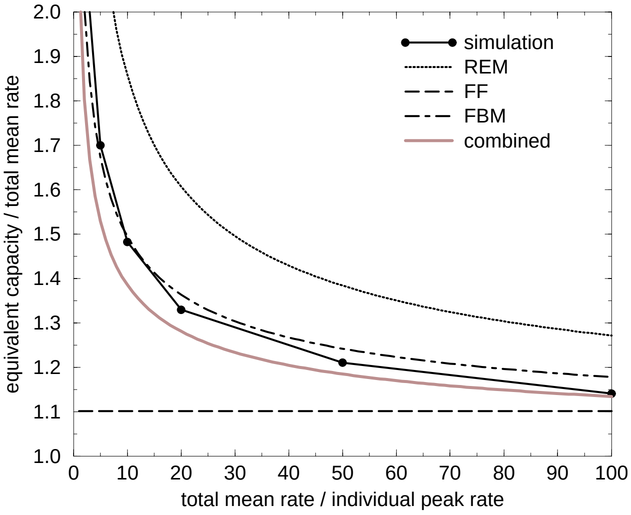

Fig. 16: M/Pareto traffic with *H* = 0.9, *x/b* = 50

size. However, the effect of underestimating the effective bandwidth in some parameter regions caused by the contribution of the FBM approximation is also increased. This becomes obvious especially in the large buffer case.

In the next figures the effective bandwidth formulas are applied to self-similar traffic based on the M/Pareto fluid burst model with Hurst parameter $H = 0.6$ (Fig. 9 and Fig. 10), $H = 0.7$ (Fig. 11 and Fig. 12), $H = 0.8$ (Fig. 13 and Fig. 14), and $H = 0.9$ (Fig. 15 and Fig. 16), respectively. The results for self-similar traffic show that the FF approximation significantly underestimates the required service rate if the buffer size is large. The REM approximation on the other hand is always a conservative solution. It becomes more accurate when the Hurst parameter is increased even for large buffers.

The FBM approximation is able to give quite accurate results if the Hurst parameter is large enough. In the case of $H = 0.6$, however, it is overly conservative and gives an effective bandwidth even greater than that provided by the REM approximation. This effect is caused by the slow convergence of the M/Pareto process to an FBM with respect to the variance for lower values of the Hurst parameter (Fig. 1). On the other hand, the FBM approximation underestimates the required capacity over the whole range of H if m/h is low.

The combined method suffers from this underestimation effect of the FBM approximation. It is still conservative for $H = 0.6$ over a wide range of mean rates. In this case it turns out to be the most accurate one. If, however, the Hurst parameter is increased, the combined method underestimates the effective bandwidth over a wide range of m/h. For very high total mean rates it can be shown by simulations that it becomes more accurate than the FBM approximation for any value of the Hurst parameter.

5. Conclusions

The M/Pareto traffic model has been used as a traffic model exhibiting long-range dependence to produce a consistent set of parameters that allows the comparison of different effective bandwidth schemes, namely the rate envelope multiplexing, fluid flow and fractional Brownian motion approximations. The latter explicitly includes long-range dependence effects. A fourth method, combining the results of rate envelope multiplexing and fractional Brownian motion approaches, was developed in an attempt to find a scheme that captures both effects.

A comparison of analytical with burst scale fluid simulation results has shown that none of the approaches is able to give exact results for the effective bandwidth over the full range of parameters. The fractional Brownian motion is generally problematic as it underestimates the bandwidth needed by traffic streams with small mean rates. This weakness is also present in the combined method. On the other hand, the FBM model overestimates the bandwidth needed by short-range dependent or weakly self-similar traffic if the total mean rate is relatively high. The combined method yields more accurate results in this parameter range. For highly self-similar traffic, the FBM approximation gives quite accurate results, but the combined method is even more accurate if the mean rate is very high. In cases of small and medium buffer sizes or high Hurst parameter, the REM method is a reasonably accurate approximation, even though it is completely insensitive to the burst size distribution.

References

[1] R. Addie, P. Mannersalo, I. Norros: "Performance Formulae for Queues with Gaussian Input." *Proceedings of the 16th International Teletraffic Congress (ITC 16)*, Edinburgh, UK, June 1999, pp. 1169-1178.

[2] D. Anick, D. Mitra, M. Sondhi: "Stochastic Theory of a Data-Handling System with Multiple Sources." *Bell System Technical Journal*, Vol. 61, No. 8, Oct. 1982, pp. 1871-1894.

[3] F. Brichet, J. Roberts, A. Simonian, D. Veitch: "Heavy Traffic Analysis of a Storage Model with Long Range Dependent On/Off Sources." *Queueing Systems*, Vol. 23, 1996.

[4] J. Charzinski: "Internet Client Traffic Measurement and Characterisation Results." *Proceedings of the 13th International Symposium on Services and Local Access (ISSLS 2000)*, Stockholm, June 2000.

[5] J. Charzinski: "Fun Factor Dimensioning for Elastic Traffic." Accepted for presentation at *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, Sep. 2000.

[6] M. E. Crovella, A. Bestavros: "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes." *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, Dec. 1997, pp. 835-846.

[7] R. J. Gibbens, Y. C. Teh: "Critical time and space scales for statistical multiplexing." *Proceedings of the 16th International Teletraffic Congress (ITC 16)*, Edinburgh, UK, June 1999, pp. 87-96.

[8] M. Grossglauser, J.-C. Bolot: "On the relevance of long-range dependence in network traffic." *IEEE/ACM Transactions on Networking*, Vol. 7, No. 5, Oct. 1999, pp. 629-640.

[9] R. Guérin: "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks." *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, Sep. 1991, pp. 968-981.

[10] F. Kelly: "Notes on effective bandwidths." In *Stochastic Networks: Theory and Applications*, Eds.: F. P. Kelly, S. Zachary, I. Ziedins, Clarendon Press, Oxford, 1996, pp.141-168.

[11] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson: "On the Self-Similar Nature of Ethernet Traffic (Extended Version)." *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, Feb. 1994, pp. 1-15.

[12] T. D. Neame, M. Zukerman, R. G. Addie: "A practical approach for multimedia traffic modeling." *Proceedings of the 5th International Conference on Broadband Communications (BC '99)*, Hong Kong, Nov. 1999, pp. 73-82.

[13] I. Norros: "A storage model with self-similar input." *Queueing Systems*, Vol. 16, No. 2, 1994, pp. 387-396.

[14] I. Norros: "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks." *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, Aug. 1995, pp. 953-962.

[15] A. Patel, C. Williamson: *Statistical Multiplexing of Self-Similar Traffic: Theoretical and Simulation Results*, University of Saskatchewan, Department of Computer Science, April 1997, http://www.cs.usask.ca/faculty/carey/papers/statmuxing.ps.

[16] V. Paxson, S. Floyd: "Wide Area Traffic: The Failure of Poisson Modeling." *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, June 1995, pp. 226-244.

[17] J. Roberts, U. Mocci, J. Virtamo (Eds.): *Broadband network teletraffic: Final Report of Action COST 242*. Springer, 1996.

[18] B. Tsybakov, N. D. Georganas: "On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound, and Cell Delay Distribution." *IEEE/ACM Transactions on Networking*, Vol. 5, No. 3, June 1997, pp. 397-409.

[19] D. Veitch, P. Abry: "A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence." *IEEE Transactions on Information Theory*, Vol. 45, No. 3, April 1999, pp. 878-897.

[20] W. Willinger, M. S. Taqqu, R. Sherman, D. V. Wilson: "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level." *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, Feb. 1997, pp. 71-86.