Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DE ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Miguel LOPEZ
AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Ph.D. (Electrical Engineering)
GRADE - DEGREE

School of Information, Technology and Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

On Network Resource Allocation using α-Stable Long Range Dependant Traffic Models

D.Makrakis
DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

L. O. Barbosa
CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

C. Huang

A. Miri

S. Pierre

O. Yang

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

On Network Resource Allocation using α-Stable Long-Range Dependent Traffic Models

by

Miguel López Guerrero

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering
at
University of Ottawa
Ottawa, Ontario
May, 2004

Thesis Directors:
Dr. Dimitrios Makrakis
Dr. Luis Orozco Barbosa

Your file Votre référence
ISBN: 0-494-01732-5
Our file Notre référence
ISBN: 0-494-01732-5

To my parents
Teresa Guerrero Basurto and Víctor López Zúñiga

Abstract

Recent studies suggest that networks should be designed taking into account the long-range dependence and high-variability properties of the traffic they carry. It has been proven in the past that these two statistical properties can be properly represented using traffic models based on α-stable self-similar stochastic processes. Assuming this traffic modeling approach, in this dissertation we propose and evaluate some techniques for resource allocation. We propose suitable envelope processes for the levels of bandwidth demand which allows us to develop static resource allocation schemes. The proposal is based on a generalization, to the α-stable case, of the concept of probabilistic envelope processes, which have been previously defined for simpler models. It is shown that, with this approach, we can simply and effectively deal with much of the argued complexity encountered in α-stable models and develop techniques for proper dimensioning of network elements. From our analysis it is concluded that the presence of heavy tails in the distribution of the traffic process has a severe impact on the requirements of network resource. For instance, the multiplexing gain is negatively affected, which directly impacts the scale economies expected by service providers. In order to cope with these issues, dynamic resource allocation is also considered. A dynamic prediction-based resource allocation method is introduced and evaluated. It is shown that it significantly improves network utilization over static resource allocation schemes in trade for some signaling and processing overhead. Although other schemes based on prediction have been proposed, we use a novel linear prediction algorithm for symmetric fractional stable noise. This approach is intended for some traffic classes whose marginal distribution exhibits a heavy tail. The linear prediction algorithm we use was recently introduced by other researchers, but has not been studied in detail. Therefore, its performance evaluation is also carried out. In addition to this study on the prediction-based approach, a dynamic resource allocation scheme based on envelope processes is also introduced and evaluated. We conclude that when α-stable models are properly used and interpreted, they let us accurately represent network traffic and therefore design and analyze reliable resource allocation mechanisms.

Acknowledgements

The development of this work got an enormous benefit from the expertise and experience of my thesis advisors, Dr. Dimitrios Makrakis and Dr. Luis Orozco Barbosa. From whom I also received a lot of encouragement and support. The constructive comments provided by Dr. Changcheng Huang (Carleton University), Dr. Samuel Pierre (École Polytechnique de Montréal) and Dr. Oliver Yang contributed to considerably improve this work. The fruitful discussions with Dr. José R. Gallardo López (CICESE, Mexico) are also greatly appreciated.

This work was made possible (in part) by the financing mechanisms of the Consejo Nacional de Ciencia y Tecnología (CONACYT). I would like to acknowledge the collective effort and commitment of all people working at this Mexican institution.

Additional funding for the completion of this work came from the following Canadian sponsors: the National Capital Institute on Telecommunications (NCIT), the Natural Sciences and Engineering Research Council (NSERC) and the Communications and Information Technology Ontario (CITO).

I thank them all.

Contents

List of Figures

List of Tables

Acronyms

H-sssi	Self-similar process with index H and stationary increments
iid	Independent and identically distributed
FSM	Fractional stable motion
FSN	Fractional stable noise
LFSM	Linear fractional stable motion
LFSN	Linear fractional stable noise
Log-FSM	Log-fractional stable motion
Log-FSN	Log-fractional stable noise
$S\alpha S$	Symmetric α-stable

Mathematical notation

$X \sim f_X(x)$	The random variable X has the distribution $f_X(x)$
$a_t \sim b_t$	$\lim_{t\to\infty} a_t/b_t = 1$
$\cong$	Approximately equal
$\triangleq$	Equal by definition
$\stackrel{d}{=}$	Equality in distribution
$\mathbf{Z}$	The set of integer numbers
$\mathbf{R}$	The set of real numbers
j	$\sqrt{-1}$ (unless otherwise specified)
$Im(x)$	Imaginary part of x
$max(S)$	The maximum value in S
$min(S)$	The minimum value in S
$Re(x)$	Real part of x
$sign(x)$	1 if $x>0$, 0 if $x=0$ and -1 if $x<0$
$sup(S)$	The supremum of a set S
$N(m,s^2)$	Normal distribution of mean m and variance s^2
$S_\alpha(\sigma,\beta,\mu)$	Stable distribution with index α, scale parameter σ, skewness parameter β and shift parameter μ.

Chapter 1

Introduction

The immense number of complex interactions, taking place in today's networks, makes the planning and control of telecommunications systems a difficult task. Traffic generated by different applications is transferred by diverse protocols coexisting in heterogeneous physical links. In addition to the issue of complexity, factors such as increasing network size, dynamic routing, changing congestion levels and the evolution of the hardware and applications seem to indicate that whatever can be said at a certain instant, about the network operation, will not be valid at a different point in time. These factors prevent us from defining what could be considered a typical situation. Accordingly, traffic controls must be designed considering a wide spectrum of realistic scenarios. Use of oversimplifications in the analysis and design of traffic control algorithms may be misleading, resulting in inability to satisfactorily meet the requirements of the traffic sources.

In [29] Floyd *et al.* tried to identify a set of invariant elements across the Internet. Two of them are of particular importance in the case of aggregate traffic modeling: the self-similar property of several traffic classes and the heavy-tailed distribution of the levels of network activity. Although the impact of these statistical characteristics on network performance is still subject of some controversy (e.g., see [80]), it is undeniable that available performance analyses strongly suggest that they may be responsible in part for the congestion commonly encountered in the current infrastructure.

Several issues arise with better understanding of the statistical properties of network traffic and their relevance. One of them is that there is an evident necessity of using more realistic traffic models in the simulation of telecommunications networks, so that their real performance can be observed. Another consequence is that new or enhanced traffic controls must be developed and evaluated using more accurate traffic models. Traffic controls developed assuming earlier traffic models with finite-variance distributions (i.e., more smoothly varying) are expected to underestimate the resource demand. This is the motivation behind this dissertation: the development of traffic control mechanisms that take into account the real behavior of network traffic.

1.1 Hypothesis to be tested

It is assumed that models based on α-stable self-similar stochastic models are more suitable to represent aggregate network traffic when compared with more traditional approaches based on Gaussian processes. This is a valid assumption since it has been satisfactorily proven in the past [33] [46] that α-stable self-similar stochastic processes are able to simultaneously and parsimoniously capture the following two properties of aggregate network traffic: long-range dependence and heavy tails. We theorize that since these two relevant statistical properties of network traffic are properly captured using this modeling approach, it can be effectively used in the development of more reliable schemes for resource allocation as compared with currently used approaches. This is an important issue since use of effective resource allocation techniques will allow the network to support applications with stringent quality of service requirements and therefore, improve customer satisfaction.

1.2 Methodology

Resource allocation policies can be classified into two different categories: static and dynamic. Static approaches are easy to implement and are usually encountered in network technologies supporting quality of service guarantees. In turn, dynamic allocation schemes achieve higher network utilization, but they incur in some signaling overhead which makes them difficult to deploy.

In this work some resource allocation techniques are proposed and evaluated assuming traffic models based on α-stable self-similar stochastic processes. For static reservations it is proposed to compute the amount of bandwidth to be allocated using the concept of a probabilistic envelope process. This idea has been explored by other researchers assuming more traditional models; but in the case of the α-stable ones, the corresponding envelopes have not been defined. Therefore, first we propose a suitable envelope process and then we use it in the context of bandwidth allocation. The main difference with respect to previously defined envelopes is that our proposal takes into account the large deviations occurring in a traffic process. Therefore, with our approach we expect to find improvements in performance when the traffic process under consideration exhibits heavy tails. The evaluation is carried out through computer simulations and direct computation of some performance metrics.

For dynamic reservations two approaches are considered in order to estimate future bandwidth requirements in a traffic process. One is based on the use envelope processes and the other one uses the prediction algorithm described by Gallardo *et al.* in [32]. However, this prediction method has not been thoroughly studied and a number of issues arise when implementing it. For instance, the selection of an appropriate order for the predictor remains as an important issue to be satisfactorily justified. In order to address this and other issues, we carry out its performance evaluation through extensive simulations since it becomes very difficult to analyze by other means. We then use these techniques in some strategies for dynamic bandwidth allocation and evaluate their performance.

It is worth emphasizing that in this work we do not focus on a particular network platform; we rather provide techniques that can be implemented in diverse technologies.

1.3 Organization of the manuscript

Chapter 2 covers most of the concepts and methods used through this dissertation. It is included with a twofold purpose, it is intended to be used as a quick and handy reference while reading other sections of the document and it also allows us to introduce the notation used in this work.

Chapter 3 provides a literature review on the topic of traffic modeling and queueing analysis using fractional stable processes. In this section we make use of available results to show that the long-range dependence and high-variability properties of network traffic have a significant impact on the performance of a queueing system. This is an important matter since it justifies our interest in: a) the use of these traffic models to conveniently represent the two aforementioned statistical properties of network traffic and, b) the development of more effective mechanisms for resource allocation than the ones currently used.

In Chapter 4 we develop the topic of static resource allocation. As previously mentioned, the approach we follow is based on the use of probabilistic envelopes for traffic models based on fractional stable processes. This sort of envelope has been defined for simpler models and some analysis methods are already available. We build on this knowledge and generalize related concepts and methods as to consider α-stable processes.

Dynamic resource allocation is the topic to be developed in the second part of the dissertation. This service policy can be implemented through several techniques, but the approaches we use are based on envelope processes and on linear prediction of symmetric fractional stable noise. The linear prediction algorithm under consideration is thoroughly evaluated in Chapter 5 since it has just been superficially studied and related studies are scarce. In Chapter 6 we propose and evaluate some strategies for dynamic resource allocation.

In addition to several conclusions given at the end of some chapters, we end this dissertation in Chapter 7 with some comments, conclusions and recommendations for future work.

1.4 Research contributions

In a succinct way the contributions of this dissertation can be listed as follows:

- Previous works on traffic modeling with fractional stable processes used exclusively either: non-skewed [33] or totally skewed to the right [46] distributions. Instead of assuming a fixed skewness in the distribution, in this work it is shown that the degree of asymmetry on the traffic is an additional factor that can be properly modeled using general α-stable distributions.
- Recently defined envelope processes for some traffic models take into consideration the long-range dependence property of network traffic, but not the heavy tail of its marginal distribution. In this work, the concept of envelope processes is generalized one step further in order to consider discrete and continuous long-range dependent traffic models exhibiting heavy tails. Available methods of analysis and design using envelope processes are also extended.

- Due to the scarce availability of related studies, a thorough experimental study on the performance of the prediction algorithm introduced by Gallardo *et al.* in [32] was carried out. Whereas this algorithm was found to be robust under likely parameter estimation errors, it was also found to be inaccurate in some regions of the parameter space. Some optimization procedures are proposed and tested. It is also concluded that information about long-range dependencies in the traffic is of marginal usefulness for linear prediction.
- Some strategies for dynamic bandwidth allocation are proposed and experimentally evaluated.

Parts of this dissertation have been published in conference proceedings [60] [61] [62] [63] (see also the joint contributions [41] and [42]) and in a journal article [59].

Chapter 2

Background concepts

2.1 Heavy-tailed distributions

The distribution of a random variable X is called heavy-tailed if [56][1]

$$P\{X > x\} \sim L(x)x^{-\alpha} \text{ as } x \to \infty \text{ and } 0 < \alpha < 2, \tag{1}$$

where $L(x)$ is slowly varying at infinity (i.e., $\lim_{x\to\infty} L(ax)/L(x) = 1$ for all $a > 0$). Note that the smaller the α, the more significant the amount of probability mass in the tail of the distribution which means that the random variable can take values far away from the median with non-negligible probability. It can be easily proven that when α<1 the distribution does not have a mean value. Furthermore, when α<2 it has infinite variance.

2.2 α-stable random variables

α-stable random variables are examples of distributions that can exhibit heavy tails. The term stable comes from the fact that addition of independent α-stable random variables, with the same tail decay (i.e., common index α in Equation (1)), results in a random variable that is also α-stable with the same tail behavior. Following the notation used in [81], $S_\alpha(\sigma,\beta,\mu)$ stands for an α-stable distribution with index of stability α ($0<\alpha\leq 2$), scale parameter σ

---

[1] We use the notation $a_t \sim b_t$ to represent that the quantities a_t and b_t are asymptotically proportional (i.e., $\lim_{t\to\infty} a_t/b_t = 1$).

($\sigma \geq 0$), skewness parameter β ($-1 \leq \beta \leq 1$) and shift parameter μ ($\mu \in \mathbf{R}$). An α-stable random variable X can be defined in terms of its characteristic function

$$\Phi_X(\theta) = E[exp(j\theta X)] = \begin{cases} exp\left[j\mu\theta - \sigma^\alpha |\theta|^\alpha \cdot \left(1 - j\beta \cdot sign(\theta) \cdot tan\left(\frac{\pi\alpha}{2}\right)\right)\right] & \text{if } \alpha \neq 1, \\ exp\left[j\mu\theta - \sigma|\theta| \cdot \left(1 + j\beta \cdot \frac{2}{\pi} \cdot sign(\theta) \cdot ln|\theta|\right)\right] & \text{if } \alpha = 1. \end{cases} \tag{2}$$

In particular, a random variable X is symmetric α-stable ($S\alpha S$) if and only if $X \sim S_\alpha(\sigma, 0, 0)$ [2]. In this case (2) simplifies to

$$\Phi_X(\theta) = exp\left(-\sigma^\alpha |\theta|^\alpha\right). \tag{3}$$

Generally speaking, α-stable random variables lack of closed form expressions for their probability density functions. An important exception is the setting α=2 which leads to the Gaussian distribution. The other known exceptions are the Cauchy (α=1, β=0) and Lèvy (α=1/2, $\beta = 1$) distributions [81 p. 9].

2.2.1 Some relevant properties of α-stable distributions

Shift and multiplication by a constant. Let $X \sim S_\alpha(\sigma, \beta, \mu)$ and let a be a real constant. The following properties can be easily demonstrated [81 p. 11]:

$$X + a \sim S_\alpha(\sigma, \beta, \mu + a), \tag{4}$$

$$\begin{cases} a \cdot X \sim S_\alpha\left(|a|\sigma, sign(a)\beta, a\mu\right) & \text{if } \alpha \neq 1, \\ a \cdot X \sim S_1\left(|a|\sigma, sign(a)\beta, a\mu - \frac{2}{\pi} a(ln|a|)\sigma\beta\right) & \text{if } \alpha = 1 \text{ and } a \neq 0. \end{cases} \tag{5}$$

"Stability." Let X_1 and X_2 be independent random variables such that $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, for $i = 1, 2$. From their characteristic functions it can be demonstrated that [81 p. 10]

[2] We use the notation $X \sim f_x(x)$ to represent that the random variable X has the distribution $f_x(x)$.

$$X_1 + X_2 \sim S_\alpha(\sigma, \beta, \mu) \text{ with } \sigma = \left(\sigma_1^\alpha + \sigma_2^\alpha\right)^{1/\alpha},\ \beta = \frac{\beta_1\sigma_1^\alpha + \beta_2\sigma_2^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha} \text{ and } \mu = \mu_1 + \mu_2. \quad (6)$$

Existence of moments. Let us consider an α-stable random variable $X \sim S_\alpha(\sigma,\beta,\mu)$ with $0 < \alpha < 2$. The existence of moments for X is conditioned upon the value of its index of stability α [81 p.18] [70 pp. 22-23] as follows:

$$E[|X|^p] < \infty \text{ for any } p \in (0,\alpha) \text{ and } E[|X|^p] = \infty \text{ for } p \geq \alpha. \quad (7)$$

Tail behavior. For $0 < \alpha < 2$ the tail probabilities of $X \sim S_\alpha(\sigma,\beta,\mu)$ behave according to the following asymptotic behavior [81 p. 16]:

$$\begin{cases} lim_{x\to\infty}\, x^\alpha P\{X > x\} = C_\alpha \dfrac{1+\beta}{2}\sigma^\alpha \\ lim_{x\to\infty}\, x^\alpha P\{X < -x\} = C_\alpha \dfrac{1-\beta}{2}\sigma^\alpha \end{cases} \quad (8)$$

where $C_\alpha = \dfrac{1-\alpha}{\Gamma(2-\alpha)cos(\pi\alpha/2)}$ for $\alpha \neq 1$ and $C_\alpha = \dfrac{2}{\pi}$ for $\alpha = 1$. This property shows that the α-stable random variable, with $\alpha < 2$, is indeed heavy-tailed.

Dispersion. A useful quantity derived from the parameters of a stable distribution is its dispersion γ. It is obtained as

$$\gamma = \sigma^\alpha \ . \quad (9)$$

Let again X_1 and X_2 be independent α-stable random variables and a, b real constants. From (5) and (6) it can be easily proven that if a random variable X is defined as $X = aX_1 + bX_2$ then its dispersion is given by

$$\gamma_X = \gamma_1|a|^\alpha + \gamma_2|b|^\alpha \quad (10)$$

where γ_1 and γ_2 are the dispersions of X_1 and X_2, respectively.

Codifference. Let X_1 and X_2 be two jointly $S\alpha S$ random variables, a measure of the statistical dependence between them can be obtained through their codifference $\tau_{X_1 X_2}$, defined as [81 p. 103]

$$\tau_{X_1,X_2} = \sigma_{X_1}^{\alpha} + \sigma_{X_2}^{\alpha} - \sigma_{X_1-X_2}^{\alpha} \tag{11}$$

where σ_{X_1}, σ_{X_2} and $\sigma_{X_1-X_2}$ are the scale parameters of X_1, X_2 and $(X_1\text{-}X_2)$ respectively.

2.2.2 Parameter estimation for α-stable distributions

Several estimation methods for the parameters of a stable distribution are currently available. Although they are based on diverse principles, a recent performance comparison among some representative techniques [6] concluded that there is not a single approach consistently superior to the others. The most commonly used methods are based on sample fractiles (e.g., [26]) or on the sample characteristic function (see [70 Chapter 5] for a survey). The method selected for this work is a combination of both and relies on a procedure introduced by Koutrouvelis in [54] and [55], which has shown consistently good results over large portions of the parameter space [6]. As it will be explained, Koutrouvelis' method is based on the following two relations resulting from applying some transformations to the characteristic function (2):

$$ln(-Re(ln(\Phi(\theta)))) = \alpha\, ln|\theta| + \alpha\, ln(\sigma) \tag{12}$$

and for $\alpha \neq 1$

$$Im(ln(\Phi(\theta))) = \mu\theta + \beta\sigma^{\alpha}|\theta|^{\alpha}\, sign(\theta) tan\left(\frac{\pi\alpha}{2}\right). \tag{13}$$

Assume that given N samples of an α-stable random variable $(X_i; i = 1,\cdots,N)$ with $1 < \alpha \leq 2$, we wish to find the estimates $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ and $\hat{\mu}$ for the parameters of its distribution.

Since we are considering the case $\alpha > 1$, the mean of the random variable exists (see Equation (7)) and it equals the shift parameter μ [81 p. 19]. Therefore, we can use the sample mean to find $\hat{\mu}$. An initial estimate for the scale parameter $\hat{\sigma}_i$ can be obtained through the sample fractile method assuming a symmetric distribution for the data (see Appendix C for a description of the method). Once the parameters μ and σ have been estimated, it is convenient to normalize the sample data by transforming each data point X_i to $X_i^{'} = (X_i - \hat{\mu})/\hat{\sigma}_i$. The distribution of the set of transformed points has approximately zero mean and unitary scale parameter, but retains the same values for α and β (see Equations (4) and (5)). It is interesting to observe that under these conditions, Equations (12) and (13) can be simplified as follows[3]:

$$ln(-Re(ln(\Phi(\theta)))) \cong \alpha\, ln|\theta| \tag{14}$$

and

$$Im(ln(\Phi(\theta))) \cong \beta|\theta|^{\alpha} sign(\theta) tan\left(\frac{\pi\alpha}{2}\right). \tag{15}$$

When we know that the sample data were drawn from a symmetric distribution, we can use (14) to extract the remaining parameter. However, in the general case we have to consider that the estimate $\hat{\sigma}_i$ provided by the fractile method is just an initial estimate and has to be adjusted in order to take into account the degree of skewness in the sample data. Using (12) instead, observe that this equation fits the linear model $y = \alpha x + b$ with $y = ln(-Re(ln(\Phi(\theta))))$, $x = ln|\theta|$, slope α and y-intercept $b = \alpha\, ln(\sigma)$. Therefore, given a set of ordered pairs $\{(x_k, y_k); k = 1, \cdots, K\}$, a linear regression model let us determine $\hat{\alpha}$ and $\hat{\sigma}$.

[3] We use the notation $a \cong b$ to represent that the quantities a and b are approximately equal.

Samples of x can be calculated using an appropriate set $\{\theta_k; k = 1, \cdots, K\}$. The corresponding samples of y can be obtained using the sample characteristic function:

$$\hat{\Phi}(\theta_k) = \frac{1}{N} \sum_{i=1}^{N} exp(j\theta_k X_i'). \tag{16}$$

The authors in [53] proposed that if normalized data are used, θ can be sampled using as few as ten evenly spaced points within the interval [0.1, 1]. It is worth mentioning that since we are using normalized data in (16), the resulting estimate for the scale parameter has to be re-scaled by the initial estimate $\hat{\sigma}_i$.

Whereas in the symmetric case the previously described procedure determines all the parameters of the distribution (see [52]), an additional step is required to determine β in the general case. Observe that (15) also fits a linear model $y = \beta x$ with $y = Im(ln(\Phi(\theta)))$ and $x = |\theta|^{\alpha} sign(\theta) tan(\pi\alpha/2)$. Therefore, $\hat{\beta}$ can also be extracted from a linear regression model on a set of sample points $\{(x_k, y_k); k = 1, \cdots, K\}$ with θ sampled as previously explained.

2.3 Self-similarity

A stochastic process exhibits self-similarity if it preserves its statistical properties under judicious scaling over time or space. Following the standard definition of self-similarity [81 p. 311], a continuous time process $\{X(t), t \in \mathbf{R}\}$ is called self-similar with index $H > 0$ (H-ss) if for any $a > 0$ the finite-dimensional distributions of $\{X(at), t \in \mathbf{R}\}$ are the same as the finite-dimensional distributions of $\{a^H X(t), t \in \mathbf{R}\}$. In other words, for any $d \geq 1$, $t_1, t_2, \ldots, t_d \in \mathbf{R}$ and $a > 0$, the random vectors $(X(at_1), X(at_2), \ldots, X(at_d))$ and $(a^H X(t_1), a^H X(t_2), \ldots, a^H X(t_d))$ have the same distribution. As a consequence of this, for a fixed time index, the marginal distributions of a self-similar process obey the following scaling property:

$$X(at) \overset{d}{=} a^H X(t) \tag{17}$$

where $\overset{d}{=}$ means equality in distribution. Note that $X(at)$ and $X(t)$ do not have the same distribution and therefore, a process satisfying (17) cannot be stationary.

The notion of self-similarity can also be applied in the context of stochastic processes with discrete time index resulting in closely related (but not equivalent) definitions. Let $Y = \{Y_j, j \in \mathbf{Z}\}$ represent a (wide-sense) stationary time series. For each $m = 1,2,3...$ let $Y^{(m)} = \{Y_j^{(m)}, j \in \mathbf{Z}\}$ represent a new process obtained by aggregating the original one Y over non-overlapping blocks of size m, i.e., $Y_j^{(m)} = Y_{jm-m+1} + \ldots + Y_{jm}$. It is said that Y is self-similar if [19] [75] [86],

$$Y_j^{(m)} \overset{d}{=} m^H Y_j \ ; \ m \geq 1. \tag{18}$$

Self-similarity is often investigated through the behavior of the autocorrelation function. The process Y is called *exactly second order self-similar* if $Y^{(m)}/m$ possess the same autocorrelation structure as the original process for any $m \geq 1$ and *asymptotically second order self-similar* if this property arises when $m \to \infty$ [16] [56]. Alternative, not equivalent, definitions of self-similarity have also been proposed [86].

We end this section by introducing the notation H-sssi to denote a process that is self-similar with index H and has stationary increments.

2.4 Long-range dependence

The standard notion of long-range dependence involves the tail behavior of the autocorrelation function of (wide-sense) stationary time series [92]. Assume a stationary time series $Y = \{Y_j, j \in \mathbf{Z}\}$ such as the unitary increments of an H-sssi process $X = \{X(t), t \in \mathbf{R}\}$ (i.e., $Y_j = X(j+1) - X(j)$). The discrete time process Y is said to exhibit long-range dependence if its autocorrelation function $\rho(k)$ has the following decaying behavior [8]:

$$\rho(k) \sim L(k) \cdot k^{-d}, \text{ as } k \to \infty \tag{19}$$

where $0 < d < 1$ and $L(k)$ is slowly varying at infinity. It is important to observe that (19) also implies that the autocorrelation function is not summable, i.e., $\sum_k \rho(k) = \infty$. The presence of the long-range dependence property also implies that the spectral density, defined as $f(\lambda) = \sum_k \rho(k) e^{ik\lambda}$, diverges at the origin like a power function, i.e. $f(\lambda) \to \infty$ as $\lambda \to 0$. In contrast, a process shows *short-range dependence* if its autocorrelation function follows an exponential decay [8], i.e., $\rho(k) \sim a^k$ with $0 < a < 1$, as $k \to \infty$. It can be seen that in this case the autocorrelation function is summable, i.e., $\sum_k \rho(k) < \infty$.

It is important to mention that even though the terms "(second order) self-similarity" and "long-range dependence" are not equivalent, they are frequently used interchangeably in the literature. This is due to the fact that both refer to the tail behavior of the autocorrelations. Furthermore, the increments of a finite variance H-sssi process exhibit long-range dependence as long as $1/2 < H < 1$ with the parameter d in (19) related with the index of self-similarity H through the expression [56][1 p. 48]

$$H = 1 - d/2. \tag{20}$$

Note that the long-range dependence of a process has been defined in terms of the behavior of its autocorrelation function and spectral density. However, when we consider processes with α-stable marginal distributions (e.g., the increments of an α-stable H-sssi stochastic process), second order moments can only be obtained for the Gaussian case (i.e., $\alpha = 2$). For $\alpha < 2$ second order moments are non-existent (see Equation (7)) and therefore, correlations and spectral densities cannot be obtained. Nevertheless, the sample autocorrelation function converges in probability for a large class of processes with marginal distributions of infinite variance. More explicitly, let $\{Y_j = \sum_{k=-\infty}^{\infty} c_k U_{j-k},\ j \in \mathbf{Z}\}$ be a stationary moving average process where $\{U_k\}$ are iid α-stable random variables with $0 < \alpha < 2$ and the real sequence $\{c_k\}$ satisfies some mild conditions as specified in [21]. Theorem 4.2 in [21] can be used to affirm that the normalized sample autocorrelation function $\hat{\rho}(k)$ computed from N consecutive samples $\{Y_j, j = 1,\dots,N\}$ as

$$\hat{\rho}(k) = \frac{\sum_{j=1}^{N-k} Y_j Y_{j+k}}{\sum_{j=1}^{N} Y_j^2} \quad ; k = 0,\dots,N-1, \tag{21}$$

converges in probability, as the sample size increases, to the autocorrelation function defined by $\rho(k) = \sum_{j=-\infty}^{\infty} c_j c_{j+k} \Big/ \sum_{j=-\infty}^{\infty} c_j^2$ corresponding to an analogous process where the $\{U_k\}$ have finite variance (i.e., α=2).

The absence of autocorrelation has motivated alternate definitions of long-range dependence for infinite-variance processes (see [12]). The asymptotic behavior of the codifference (see Equation (11)) can also be used to study long-range dependence in the case of $S\alpha S$ processes. In the skewed α-stable case, there are some other quantities closely related to the codifference that can be used instead [81 p. 213] [93]. Using this criterion, a stationary process $\{Y_j, j \in \mathbf{Z}\}$ is long-range dependent, in the codifference sense, if

$$\tau(k) \sim L(k)k^{-d}, \text{ as } k \to \infty \tag{22}$$

with $\tau(k) = \tau_{Y_k,Y_0}$, $0 < d < 1$ and $L(k)$ as defined for (19).

2.5 Fractional stable motions and fractional stable noises

We will consider the following two self-similar processes with index H and stationary increments (H-sssi): linear fractional stable motion (LFSM or $\{L_{\alpha,H}(t),\ t \in \mathbf{R}\}$) and Log-fractional stable motion (Log-FSM or $\{\Lambda_\alpha(t),\ t \in \mathbf{R}\}$) [4]. For these two processes the statistical dependences among samples of their increment process decay to zero as a power function (see [81 pp. 367-370]). The marginal distributions of these processes are not restricted to be Gaussian, they can be α-stable (i.e., let $\{X(t), -\infty < t < \infty\}$ be one of the considered α-stable H-sssi process, for a fixed t we have that $X(t) \sim S_\alpha(\sigma_t, \beta_t, 0)$). A LFSM process is defined as [81 p. 343]:

$$L_{\alpha,H}(t) = \int_{-\infty}^{\infty} f(t,x) \cdot M(dx)\ ,\ -\infty < t < \infty \tag{23}$$

where $M(dx)$ is an α-stable random measure on $\mathbf{R}$ with Lebesgue control measure and skewness intensity $\beta(x)$ with $\beta(\cdot)=0$ if $\alpha=1$ (let us assume a constant intensity $\beta(\cdot)= \beta$). In addition, the kernel function $f(t,x)$ is given by

[4] The term "fractional" was proposed by Mandelbrot in [65] to denote those sequences whose experimentally obtained sample spectral densities follow a power law near the origin (see [50] for an in-depth explanation). Such behavior has been observed in a large number of phenomena occurring in diverse areas such as physics, engineering, economics, music, weather, traffic and hydrology [91].

$$f(t,x) = a[(t-x)_+^{H-1/\alpha} - (-x)_+^{H-1/\alpha}] + b[(t-x)_-^{H-1/\alpha} - (-x)_-^{H-1/\alpha}] \tag{24}$$

where $H \neq 1/\alpha$,

$$z_+ = \begin{cases} z & ; z \geq 0 \\ 0 & ; z < 0 \end{cases} \qquad \text{and} \qquad z_- = \begin{cases} 0 & ; z \geq 0 \\ -z & ; z < 0 \end{cases}.$$

The parameters a and b are arbitrary real constants with $|a| + |b| > 0$, although some specific values and relations between them are frequently used. For instance, the process is called *balanced* when $a = b$, *well-balanced* when $a = b = 1$ and *anti-balanced* if $a = -b$.

By analogy with the case α=2, it is considered that when H>1/α the increment process of LFSM displays long-range dependence. Since the parameter H is restricted to lie within the interval (0, 1), then α must be greater than 1 for the long-range dependence property to be present.

In turn, a Log-FSM process is defined as [51] [89]:

$$\Lambda_\alpha(t) = \int_{-\infty}^{\infty} f(t,x) \cdot M(dx), \quad -\infty < t < \infty \tag{25}$$

where $M(dx)$ is again an α-stable random measure as previously defined for (23) and the kernel function is:

$$f(t,x) = l(t-x) - l(x) \tag{26}$$

where

$$l(z) = \begin{cases} ln|z| & \text{for } z \neq 0, \\ 0 & \text{for } z = 0. \end{cases}$$

For the Log-FSM process, its parameters satisfy the relation H=1/α, where 1<α<2.

A fractional stable noise (FSN) sequence $\{Y_j, j \in \mathbf{Z}\}$ represents the increment during the *j-th* time interval of unit length of an α-stable H-sssi process. From the previous definitions we can see that linear fractional stable noise (LFSN) can be obtained from LFSM as follows:

$$Y_j = L_{\alpha,H}(j+1) - L_{\alpha,H}(j) \;,\; j \in \mathbf{Z}. \tag{27}$$

Similarly, Log-fractional stable noise (Log-FSN) can be generated from Log-FSM by

$$Y_j = \Lambda_\alpha(j+1) - \Lambda_\alpha(j) \;,\; j \in \mathbf{Z}. \tag{28}$$

We will use the term FSN to denote either LFSN or Log-FSN. It can be generalized that

$$Y_j = \int_{-\infty}^{\infty} [f(j+1,x) - f(j,x)] \cdot M(dx) \stackrel{\Delta}{=} \int_{-\infty}^{\infty} g(j,x) \cdot M(dx) \;,\; j \in \mathbf{Z} \tag{29}$$

where $\stackrel{\Delta}{=}$ means equal by definition. Figures 2.1a and 2.1b show the integration kernel of two LFSN processes with different values of d=H-1/α. The considered cases are: the well-balanced and an anti-balanced (a=-b=1 in (24)). As observed on the figures, with close values between H and 1/α, the integration kernel of the LFSN tends to produce degenerate processes. In these cases, the kernel of the well-balanced LFSN tends to a function which is equal to zero everywhere, except at the singularities j and j+1, leading to the generation of a null process. In turn, the kernel of the anti-balanced process becomes zero everywhere, except in the interval (j,j+1), leading to the generation of uncorrelated noise. Figure 2.1c shows the kernel of the Log-FSN.

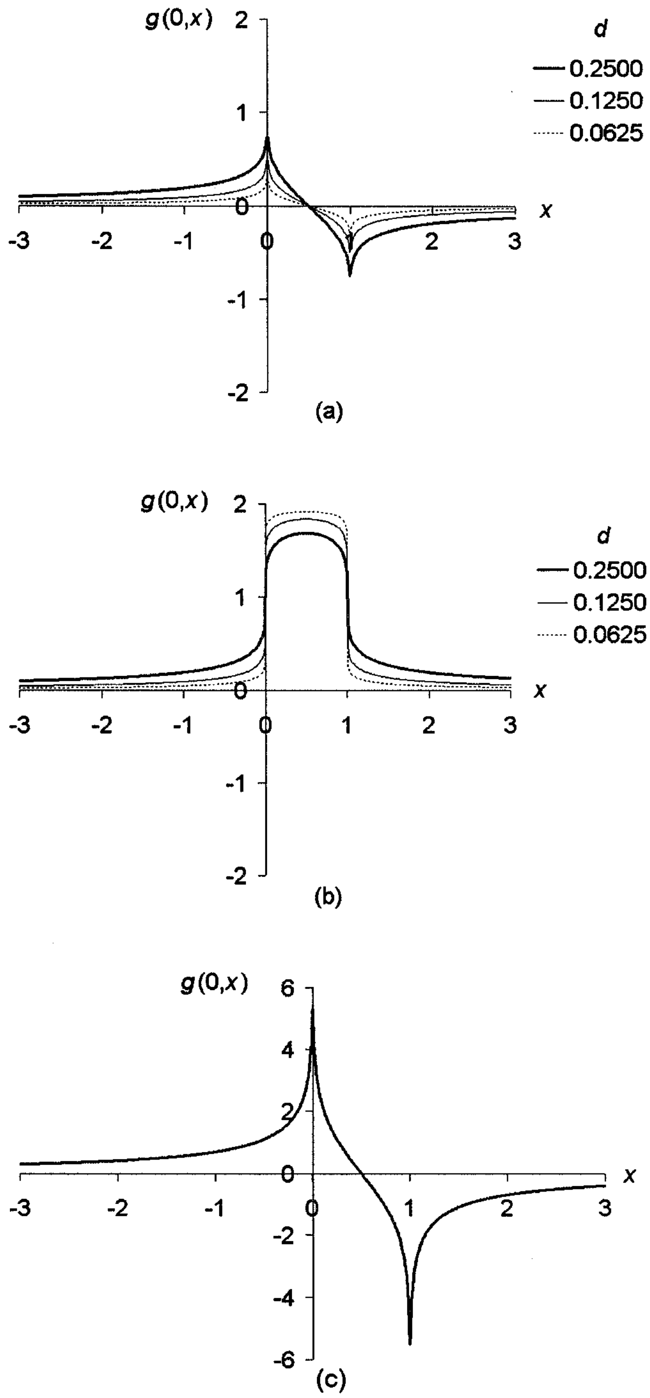


Figure 2.1 Kernel at j=0 for: a) well-balanced FSN (a=b=1), b) anti-balanced FSN (a=-b=1) and c) Log-FSN

The stationarity property of the increments a fractional stable motion $\{X(t), t \in \mathbf{R}\}$ let us determine the distribution of its FSN sequence $\{Y_j, j \in \mathbf{Z}\}$ since $Y_j = X(j+1) - X(j) \overset{d}{=} X(1) - X(0)$. But $X(0)$=0 a.s., thus:

$$Y_j \overset{d}{=} X(1) \sim S_\alpha(\sigma_1, \beta_1, 0). \tag{30}$$

Renaming $\sigma_Y = \sigma_1$ and $\beta_Y = \beta_1$ we can use the notation $Y_j \sim S_\alpha(\sigma_Y, \beta_Y, 0)$. The process

$$\{Y_{j(std)} = Y_j / \sigma_Y \ , \ j \in \mathbf{Z}\} \tag{31}$$

is called *standard* fractional stable noise and is normalized in the sense that the scale parameter of its marginal distribution equals 1.

Fractional Brownian motion and Brownian motion can be obtained by setting $\alpha = 2$ in (23) and (25) respectively, from which in turn we can obtain fractional Gaussian noise and white Gaussian noise when we apply (27) and (28), respectively.

2.6 Synthesis of symmetric fractional stable noise

In the case of FSN processes with symmetrical marginal distributions (29) can be restated (in distribution) as the following discrete time moving-average expression (see [32] for details):

$$Y_j = \sum_{k=-\infty}^{\infty} \phi(j-k) \cdot U_k \ , \ j \in \mathbf{Z} \tag{32}$$

where $\{U_k\}$ is a set of *standard* independent and identically distributed $S\alpha S$ random variables, i.e. $U_k \sim S_\alpha(1,0,0)$, and the coefficients $\phi(j\text{-}k)$ are given by

$$\phi(j-k) = \left[\int_k^{k+1} |g(j,x)|^\alpha \cdot dx \right]^{1/\alpha} > 0. \tag{33}$$

It can be observed from (32) that a symmetric FSN process with parameters H and α can be generated by a linear filter driven by symmetric α-stable noise. The dispersion of Y_j can be determined by applying the result stated by (10) into (32). Thus we obtain:

$$\gamma_Y = \sigma_Y^\alpha = \sum_{k=-\infty}^{\infty} \phi^\alpha(k) \tag{34}$$

where we have taken into consideration that the dispersion of U_k results in a unitary value. Observe that γ_Y does not depend on j.

Chapter 3

Aggregate network traffic modeling

Many researchers tried to determine the intrinsic characteristics of network traffic during the past several years. As a result of this effort, a number of studies concluded that traditional models, based on short-range dependent processes, cannot describe the behavior of real network traffic (e.g., [76]). The work by Leland *et al.* reported in their seminal paper [56] demonstrated that Ethernet LAN network traffic is rather self-similar in nature, with statistical dependences decaying so slowly that they span across several orders of magnitude. Since then, subsequent researchers have found and tried to explain the occurrence of these statistical properties in several types of network traffic. For instance, in [8] it is concluded that long-range dependence is an inherent feature of nearly all variable-bit-rate (VBR) video sequences. In [19] it is shown that aggregate WWW traffic exhibits self-similar characteristics, at least when the demand is high enough. Individual connections in a wireless network have been observed to follow a scaling behavior consistent with long-range dependence over several time scales [87].

Several traffic models have been proposed with the objective of reproducing the statistical behavior of network traffic. As mentioned in [8] there are two known classes of stochastic models that can capture long-range dependence in a flexible and parsimonious manner: fractional ARIMA processes and the increments of self-similar processes. Having more parameters, the former allows for simultaneous modeling of the short-range and long-range behavior of a time series [56]. With fewer parameters the latter allows for the *parsimonious* modeling of long-range dependencies only [56] (see also some results reported in [36]). However, use of this modeling approach is strongly motivated by the fact that self-similar processes naturally arise as a result of the interactions occurring in the network [92], allowing for a phenomenological justification of the model. In this work, only this type of models will be considered.

Some of the most relevant self-similar models presented in the past are based on fractional Brownian motion. In this context it is worth mentioning the model proposed by Norros in [71] (see also [72]) which was later satisfactorily justified by Willinger *et al.* [92] in terms of the aggregate effect of the traffic generated by individual connections.

In addition to the slow decay of correlations over several time scales, recent statistical analyses showed that some aggregate traffic classes also contain higher variability than the one assumed by Gaussian models meaning that they may take on sporadic large values (relative to their mean value) with non-negligible probability [6]. Recent models like the ones introduced in [33] and [46][5] propose the use of stable non-Gaussian stochastic processes to better capture the statistical properties of an aggregate traffic process. These models are a generalization of the ones presented in [71] and [92], in the sense that, rather than limiting the marginal distribution of the process to be Gaussian, α-stable distributions are now used, which allows for a better agreement between the burstiness of the artificial process and that of real traffic.

Use of α-stable stochastic processes in traffic modeling can be justified by means of the Generalized Central Limit Theorem (see [20 p. 94], [70 pp. 3, 22] and [81 Definition

[5] Parts of the work reported in [46] can also be found in [47] and [49].

1.1.5]). This mathematical law states that the statistics of a normalized sum of a large number of independent and identically distributed random variables converge to a function that in the limit must be a stable distribution. The distribution of the referred random variables can have either finite or infinite variance. The well-known Central Limit Theorem can be considered as a particular case with summands of finite variance, leading to a sum with a Gaussian distribution. Thus, α-stable random variables can be used to represent the statistics of aggregate traffic (i.e., resulting from the aggregation of many individual sources with or without finite variance) entering into a network node.

In this section we provide a survey on traffic models based on α-stable self-similar stochastic processes and their effect on queueing performance. The section ends with some comments on the use of this modeling approach.

3.1 Use of α-stable processes in aggregate traffic modeling

Define $A^*(t)$ as the actual number of arrivals in a traffic process during the time interval $[0,t]$. This process can be modeled using α-stable self-similar stochastic processes as follows:

$$A(t) = mt + aX(t)\ ,\ t \geq 0 \tag{35}$$

(i.e., for any $d \geq 1$ and a set $\{t_1, t_2, \cdots, t_d\} \in [0,t]$ the random vectors $(A^*(t_1),\ldots,A^*(t_d))$ and $(A(t_1),\ldots,A(t_d))$ are intended to have the same distribution) where the parameter m is the mean number of arrivals per time unit, $a > 0$ is a scaling factor and $\{X(t), t \geq 0\}$ is a fractional stable motion. Although this model is suitable in some analyses, it may be of limited usefulness in practical applications since it is based on fractional stable motion, which is a non-stationary and continuous process. For simulation and modeling purposes its increment process, being a stationary sequence, may be more appropriate. Accordingly, define W_j^* as the actual number of arrivals during the time interval $(j, j+1]$ in a traffic process. This discrete-time process can be modeled as a sequence $\{W_j\}$ (i.e, as previously explained for (35), the finite-dimensional distributions of $\{W_j^*, j = 0,1,2,\cdots\}$ and $\{W_j, j = 0,1,2,\cdots\}$ are intended to be the same) defined as follows:

$$W_j = m + aY_j \;\; ; \;\; j = 0,1,2,\ldots \tag{36}$$

where the parameters m and[6] $a > 0$ have the same meaning as in (35) and $\{Y_j\}$ is a fractional stable noise. Gallardo *et al.* in [33] proposed this model with symmetric distributions about m to represent network traffic exhibiting symmetric fluctuations around its mean value (e.g. aggregate MPEG-coded video traffic [33]). A properly modified version of (36), using symmetric marginal distributions for the FSN process, was also proposed in [33] for several other traffic classes with asymmetric fluctuations. In these cases the traffic pattern exhibits prominent upward deviations from its mean value; but the downward deviations are rather small (e.g., Ethernet and WWW traffic [33]). A different approach to represent network traffic with such asymmetry was independently developed and evaluated by Karasaridis *et al.* in [46]. This one is again based on the model described by (36) but uses a kernel function with parameters $a = 1, b = 0$ (see Equation (24)) and totally skewed to the right distributions[7] (i.e., skewness parameter $\beta = 1$ in Equation (2)). Some additional work

[6] In this work the scaling factor a (also appearing in (35)) is defined as the ratio of the scale parameter of the traffic process to the scale parameter of the fractional stable noise in its model.

[7] It is worth emphasizing that the term "totally skewed to the right" does not refer to the shape of the probability density function. This terminology has caused some confusion and in some papers it has been

with this model has been carried out by Garroppo *et al.* [36]. These proposals proved to fit different levels of variability and long-range correlations in aggregate traffic better than Gaussian self-similar models. With respect to these existing proposals, in this work we add a bit of modeling flexibility by letting the skewness parameter β to take on any value within its valid range. Otherwise stated, in this work the traffic model under consideration can be expressed in the form of (35) or (36).

3.2 Survey on queueing analysis with self-similar traffic models

A number of performance analyses, scattered in the literature, have shown the effect of the relevant traffic characteristics on network performance. Among them, of paramount importance is the analysis of a queueing system, representing a network node, since a number of performance metrics are related somehow with the buffer requirements. In this section available analyses are used to determine how these statistical characteristics of network traffic affect the performance of a queueing system.

Consider a single-server queueing system with an infinite buffer. Let C represent the service rate measured in packets per time slot, W_j the number of arrivals during the time slot

---

mistakenly implied or explicitly stated that α-stable random variables with β=1 have a probability density function which is skewed to the right and that therefore, its support is mostly for positive values. This was believed to be a useful characteristic to represent inherently non-negative processes, such as a packet arrival stream. However, with α>1 (the common case in traffic modeling), these statements are not correct since the actual skewness in the probability density function is to the left and, assuming a zero shift, a considerable amount of probability mass is concentrated in the negative part (e.g. see the graphs in [81 p. 36]). This by no means demerits the usefulness of this type of random variables since, properly shifted and scaled, they can accurately represent the statistical properties of network traffic as it will be shown in the sequel (see also [46]). Reference [81 p. 13] explains the rationale behind the naming convention for these random variables.

$(j, j+1]$ and q_j the buffer occupancy at time j (i.e., the beginning of the time slot). Given an initial condition (assume $q_0 = 0$) the following well-known Lindley recursion describes the evolution of the buffer occupancy:

$$q_{j+1} = \left(q_j + W_j - C\right)^+ \;\; ; \; j = 0, 1, 2, \ldots \tag{37}$$

where

$$z^+ = \begin{cases} z & ; z \geq 0 \\ 0 & ; z < 0 \end{cases}.$$

Let A_n represent the value of the cumulative arrival process obtained as follows:

$$A_n = W_0 + W_1 + \ldots + W_{n-1} \;\; ; \; n = 1, 2, 3, \ldots \tag{38}$$

with $A_0 = 0$. Equation (37) can be solved for q_j in terms of the cumulative arrival process (see for instance [83 p. 264]):

$$q_j = \max_{0 \leq i \leq j}\left(A_j - A_i - C \cdot (j - i)\right); \; j = 0, 1, 2, \ldots . \tag{39}$$

It has been shown [64] that if the input sequence $W = \{W_j, j = 0, 1, 2, \ldots\}$ is stationary and ergodic with $E\{W\} < C$, then the system will eventually reach statistical equilibrium, i.e. $q_j \xrightarrow[j]{} q$ for some $\mathbf{R}^+$-valued random variable q. This random quantity can be defined as follows:

$$q = \sup_{0 \leq n}\left(A_n - Cn\right) \tag{40}$$

which is similar to the Reich formula [78] (see also [7] and [71]) We are interested in the distribution of q, in particular in the probability that the buffer occupancy exceeds a certain level x, i.e.,

$$P(q > x) = P\left(\sup_{0 \le n}(A_n - Cn) > x\right). \tag{41}$$

Taking into consideration the inequality [23]

$$P\left(\sup_{0 \le n}(A_n - Cn) > x\right) \ge \sup_{0 \le n} P(A_n - Cn > x) \tag{42}$$

we have that

$$P(q > x) \ge \sup_{0 \le n} P(A_n - Cn > x). \tag{43}$$

Several definitions for a cumulative arrival process $\{A(t), t \ge 0\}$ have been proposed and analyzed in this context (consider that $A_n = A(t = n)$). Here some pertinent results with self-similar input processes will be summarized.

Gaussian input traffic

In [71] Norros carried out the analysis of the queueing system (40) with a traffic model described by

$$A(t) = mt + \sqrt{am}X(t) \tag{44}$$

where $\{X(t), t \in \mathbf{R}\}$ is a normalized fractional Brownian motion with Hurst parameter *H*, $H \in [1/2, 1)$ (which can be considered as fractional stable motion with index of stability $\alpha = 2$); *m* is the mean input rate with $m \in (0, C)$ and *a* is a variance coefficient with *a*>0. It was shown in [71] that this traffic model satisfies the previously mentioned conditions for statistical equilibrium. Using a normalized service rate (i.e., *C*=1) Norros demonstrated that the tail behavior for the queue length has a lower bound approximately given by

$$P(q > x) \ge \exp\left(-\left[\frac{1}{2am(1-H)^2}\left(\frac{(1-m)(1-H)}{H}\right)^{2H}\right]x^{2(1-H)}\right). \tag{45}$$

This expression is of the form $exp\left(-(x/b)^c\right)$ with $b > 0$, $c > 0$. This implies that the tail has a lower bound that follows a Weibullian distribution [25 p. 154]. As an illustration of the implications of this result, Figure 3.1 shows the tail behavior for a fixed value of network utilization and different degrees of self-similarity (a unitary variance coefficient was considered). It can be seen that the presence of this property in network traffic has a tremendous impact in the buffer occupancy when compared with the exponential decay (i.e., $P\{q > x\} \sim e^{-bx}$ as $x \to \infty$ and $b > 0$) predicted using short-range dependent traffic models [3] [8].

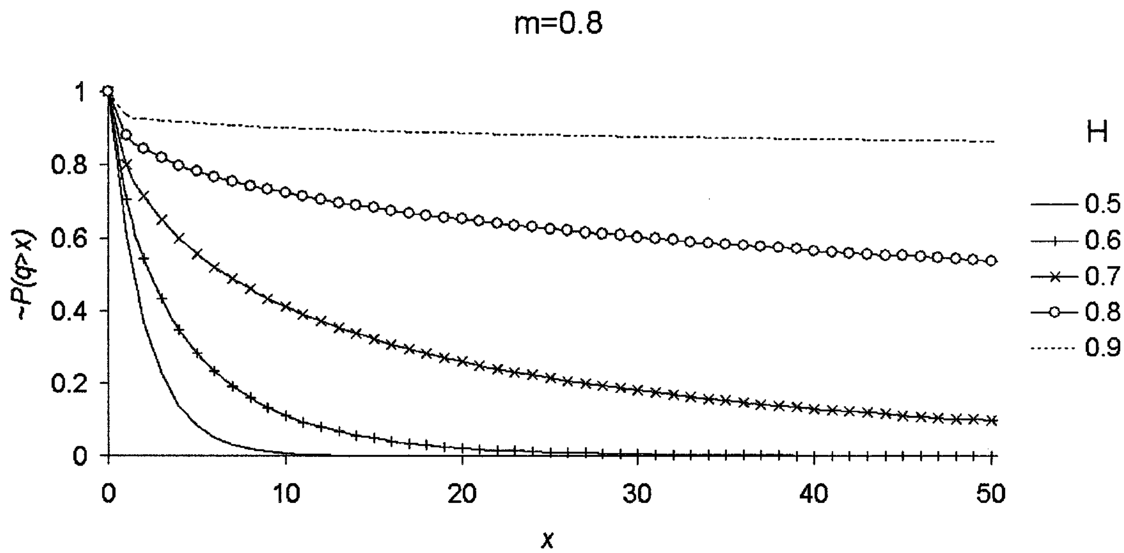


Figure 3.1 Tail probability for the buffer occupancy (lower-bound values) with fractional Brownian input.

α-stable non-Gaussian input traffic

Let us now consider the queueing system (40) for the case $1 < \alpha < 2$. Starting from the traffic model given by Equation (35) the following cumulative arrival process gives the number of packets arriving during the time interval $[0,t]$:

$$A(t) = mt + aX(t) \quad (46)$$

where $\{X(t), t \geq 0\}$ is a normalized fractional stable motion; the parameter $m \in (0, C)$ is the mean input rate and a is a scaling factor. Let us review the conditions for statistical equilibrium for this model. It is evident that the sequence derived from the unitary increments of (46) (i.e., $W_j = A(j+1) - A(j) = m + aY_j$, j=0,1,...) is a stationary process because the model is based on fractional stable motion, which has stationary increments. As for the ergodic properties of the sequence, let us first consider the symmetric case. Theorem 2 in [11] provides the sufficient and necessary condition for mixing[8] of stationary *SαS* processes with representation in the form of Equation (29). The referred condition is that the codifference (see Equation (11)) between two sample points of the process must tend to zero as the distance between them goes to infinity. This condition holds true for every *SαS* moving average, thus it can be concluded that they are mixing [90 p. 328] [40 p. 302] and therefore ergodic [79] (see also [81 p. 580]). With these results, ergodicity can be guaranteed in the case of Y_j having symmetric marginal distributions (symmetric LFSN and Log-FSN, whose marginal distributions are symmetric, see Appendix A). Regrettably, we cannot use the previous result in the case of skewed LFSN processes since the codifference is only defined for symmetric processes. However, a suitable generalization of this concept is also

[8] A stationary mixing sequence $\{Y_j, j = 0, 1, 2, \ldots\}$ satisfies $\lim_{n \to \infty} P(\{Y_0 \in A\} \cap \{Y_n \in B\}) = P\{Y_0 \in A\} P\{Y_0 \in B\}$ for all $A, B \in \sigma\{Y_j\}$ [31 pp. 563-564]. This property can be thought of as a kind of asymptotic independence.

available. Let us consider a strictly stationary stochastic process $Y=\{Y(t);t\in \mathbf{R}\}$ and the following quantity [81 p.213][9]:

$$I(\theta_1,\theta_2;t)=-\ln E\left\{e^{j(\theta_1 Y(t)+\theta_2 Y(0))}\right\}+\ln E\left\{e^{j\theta_1 Y(t)}\right\}+\ln E\left\{e^{j\theta_2 Y(0)}\right\}. \tag{47}$$

It has been proven (see references in [81 p. 213]) that if $\lim_{t\to\infty} I(\theta_1,\theta_2;t)=0$ for all θ_1 and θ_2, then the process Y is mixing. Since this is the asymptotic behavior observed for skewed linear fractional stable noise [81 pp. 367 and 370 Remark 1], we conclude that it is mixing and therefore ergodic. Thus, from Birkhoof's Ergodic Theorem [37]:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n}W_j=m+a\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n}Y_j=m+aE\{X(1)\}=m+a\mu=m<C \tag{48}$$

and statistical equilibrium can be guaranteed.

Assuming a large value for x and using a normalized service rate (i.e. C=1), the following lower bound for the tail can be demonstrated from (43) following a derivation similar to the one shown in [33] (see also [46]):

$$P(q>x)\geq C_\alpha\frac{1+\beta}{2}\left(\frac{((1-m)/H)^H}{a(1-H)^{1-H}}\right)^{-\alpha}x^{-\alpha(1-H)} \tag{49}$$

where β is the skeweness parameter of the distribution of Y_j and the constant C_α (not to be confused with the service rate C), is given by

$$C_\alpha\stackrel{\Delta}{=}\frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}. \tag{50}$$

Equation (49) shows that the tail of the queue distribution shows at the most a hyperbolic decay (i.e., $\varepsilon\sim x^{-\upsilon}$, as $x\to\infty$, $0<\upsilon<1$). Note that the tail is so heavy that the mean value

[9] Equation (47) provides a measure of the statistical dependence between the two sample points of a stationary stochastic process $\{Y(t),t\in\mathbf{R}\}$ and is related with the codifference since $\tau_{Y(t),Y(0)}=-I(1,-1;t)$ [81 p.213].

of the buffer occupancy does not exist, regardless of the traffic intensity. This striking behavior is more demanding in terms of network resources than the Weibullian decay obtained by Norros for the Gaussian case. This result, first presented by Gallardo *et al.* in [33], encompasses and generalizes several others found in diverse papers (e.g., see references in [33]). It is consistent with earlier works that have shown that high variability gives rise to a non-existent mean waiting time [92].

Figure 3.2 below shows an example of the effect of the index of stability α on the tail of the queue length distribution. It was obtained from Equation (49) considering symmetric distributions (i.e., β=0), a unitary scaling factor a, and fixed values for both m and H. The slow decay rate observed in the tail of the queue length distribution is an indication of the buffer inefficiency problem, i.e. use of large buffers in the system does not substantially help to decrease losses.

We close this section emphasizing that (49) was determined using the asymptotic tail behavior of α-stable distributions, therefore it is expected to be more accurate for large values of x. It is also worth mentioning that some work on upper bound expressions for the tail of the queue length distribution has also been carried out by Karasaridis *et al.* and reported in [44], [45] and [48]. In addition to these findings, recent investigations have found approximations or more accurate expressions for $P(q > x)$ considering Gaussian (e.g., see [73] and references therein) or stable non-Gaussian self-similar input processes [68].

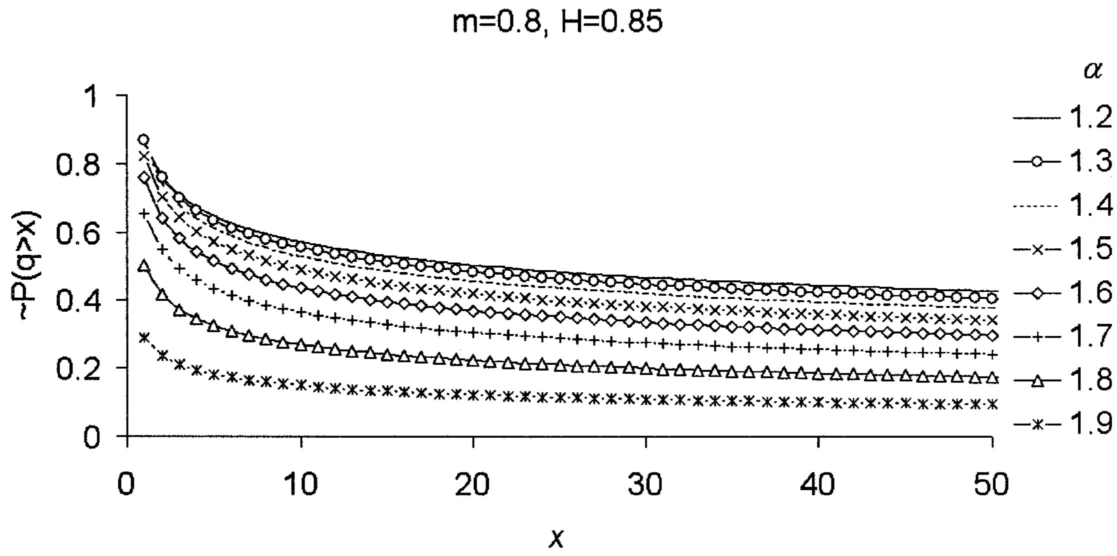


Figure 3.2 Tail probability for the buffer occupancy (lower-bound values) with fractional stable input.

3.3 Some comments on the use of idealizations in traffic modeling

Long-range dependence, self-similarity and infinite variance are mathematical idealizations that can be used to model the behavior of network traffic. However, this has to be said no without caveats since some misconceptions may arise from the erroneous interpretation of these statistical properties. For instance, a traffic model based on a self-similar process displays self-similarity at all time scales; however, to affirm that network traffic shows the same behavior can be considered as an overstatement since this cannot be validated from finite data sets. As another example, it can be mentioned that wrong conclusions can be drawn from the misuse of infinite variance models in performance analysis. For instance, assume that we take samples of a highly varying traffic being carried by a link of certain capacity. This traffic process most likely cannot be modeled using Gaussian processes, but it can be represented using α-stable processes with α<2. Using the results from the previous section, one can be tempted to affirm that when this traffic arrives to a network node, regardless of the outgoing link utilization and storage capacity, the arrival process will certainly overflow the buffer. However, when the capacity of the outgoing link equals or exceeds the one of the ingoing link (the one we used to take measurements), we can affirm that there is no queue build-up in this node at all. This apparent contradiction comes from the use of modeling idealizations. Use of α-stable distributions allows us to model sporadic large deviations, but their magnitude in a real system is always limited by maximum link capacities. This example illustrates the importance of both, a careful use of idealizations in traffic modeling and the identification of the scales (time and amplitude) in which a model applies.

Chapter 4

Static resource allocation through probabilistic envelope processes

Estimation of the resource requirements of network traffic using α-stable models is not a straightforward task. For instance, the concept of effective bandwidth [83 p. 261] cannot be applied because the moment generating function for these models is not exponentially bounded [33]. In this work it is proposed to use envelope processes[10] in a probabilistic sense in order to represent levels of network activity (i.e., the traffic process is guaranteed to remain under the envelope with a certain probability) and estimate the resource requirements. This approach allows us to represent network traffic in a simple manner, facilitating the analysis and design of network elements.

Probabilistic envelope processes have been defined for more traditional models and usually consist of the mean value of the process plus some multiple of its standard deviation. Consider, for instance, a cumulative arrival process $A(t)$ modeled as the superposition of the expected number of arrivals up to time t plus a Brownian motion $\{B(t),\ t \in \mathbf{R}\}$ of variance

[10] In our naming convention we use the notion of "envelope process" by Chang [13], but in a probabilistic sense. Chang uses this term to denote a deterministic function $\hat{A}(t_1,t_2)$ that bounds a cumulative arrival process $A(t_1,t_2)$, i.e. $A(t_1,t_2) \le \hat{A}(t_1,t_2)$. This terminology has also been adopted in other works (e.g. [83 p. 264] [30]). The concept of probabilistic envelope process has also been referred as "effective envelope" in [10].

$s^2 t$. In other words, $A(t)=mt+B(t)$, where m is the mean number of arrivals per time unit. A probabilistic envelope process $\hat{A}(t)$ for this model can be defined as

$$\hat{A}(t) = mt + K_G s t^{1/2} \tag{51}$$

where K_G determines the probability that the process will exceed its stochastic boundary as explained below. Taking into consideration that the marginal distributions of $A(t)$ are $N(mt,s^2t)$ [39 p. 1], this definition is motivated by the fact that $P\{A(t) > \hat{A}(t)\}$ becomes $P\{X > K_G\}$ where $X \sim N(0,1)$. In other words, we can determine K_G from the well-known Gaussian distribution. This concept has also been applied to other traffic models. Fonseca *et al.* in [30] proposed a "fractional Brownian motion envelope process" to delimit a cumulative arrival process $A(t)$ modeled using a fractional Brownian motion of variance $s^2 t^{2H}$ as follows:

$$\hat{A}(t) = mt + K_G s t^H \tag{52}$$

where m and K_G have the same meaning as in (51) and $A(0)=0$.

In the general case of α-stable stochastic processes, up to the author's best knowledge, suitable probabilistic envelope processes have not been defined yet. In this case some new analytical difficulties arise, the most evident limitation being that, as previously mentioned, α-stable processes generally lack second order moments. Consequently, the standard deviation cannot be obtained and therefore, alternative statistical descriptors for the variability of the process must be considered. In this section we address this issue and propose probabilistic envelope processes suitable for some traffic models based on α-stable self-similar stochastic processes.

4.1 Probabilistic envelope process for a traffic model based on fractional stable motion

Consider again the traffic model based on a fractional stable motion $\{X(t), t \geq 0\}$ as defined by (35) (i.e., $A(t) = mt + aX(t)$; $t \geq 0$ where m is the mean number of arrivals per time unit and a>0 is a scaling factor). The proposed envelope for this model is given by the expression

$$\hat{A}(t) = m_A(t) + K\sigma_A(t) \tag{53}$$

where the parameters $m_A(t)$ and $\sigma_A(t)$ are the mean and scale parameter of the process $A(t)$ at time t respectively, and K is a parameter that determines the probability that the process $A(t)$ surpasses its envelope $\hat{A}(t)$. In our proposal we use the scale parameter instead of the standard deviation because, except for the case α=2, the latter cannot be obtained for α-stable distributions.

In order to obtain the components of the envelope, let us determine the marginal distributions of $A(t)$. Observe from (35) and basic properties of α-stable random variables (see Equations (4) and (5)) that, for a fixed t, $A(t) \sim S_\alpha(|a|\sigma_t, \beta_t, mt)$. Therefore, the mean of $A(t)$ at time t is

$$m_A(t) = mt \tag{54}$$

(for an α-stable random variable X, with $1 < \alpha \leq 2$, $\mu = E\{X\}$, see [81 p. 19]) and its corresponding scale parameter is given by:

$$\sigma_A(t) = |a| \cdot \sigma_t . \tag{55}$$

Recall that σ_t represents the scale parameter of the process $X(t)$ at time t. This parameter can be determined taking into consideration that for an α-stable H-sssi process the following property holds [81 p. 340]:

$$\sigma_t = \sigma_1 |t|^H \tag{56}$$

where σ_1 has to be computed for the specific type of process under consideration. In the case of LFSM [81 p. 345] [11]

$$\sigma_1^\alpha = \left(|a|^\alpha + |b|^\alpha\right)\int_0^\infty \left((1+x)^{H-1/\alpha} - x^{H-1/\alpha}\right)^\alpha dx + \int_0^1 \left|a(1-x)^{H-1/\alpha} - bx^{H-1/\alpha}\right|^\alpha dx \tag{57}$$

whereas for the Log-FSM it is demonstrated in Appendix A that

$$\sigma_1^\alpha = 2\int_0^\infty |l(1+x) - l(x)|^\alpha dx + \int_0^1 |l(1-x) - l(x)|^\alpha dx. \tag{58}$$

Finally, the parameter K in (53) is uniquely determined by fixing an overshoot probability ε (i.e., probability that the cumulative arrival process $A(t)$ will surpass its envelope process $\hat{A}(t)$), as evidenced observing that:

$$P\{A(t) > \hat{A}(t)\} = P\{A(t) > m_A(t) + K\sigma_A(t)\} = P\left\{\frac{A(t) - m_A(t)}{\sigma_A(t)} > K\right\} = \varepsilon. \tag{59}$$

It is important to observe that shifting $A(t)$ by $m_A(t)$ and scaling the result by $1/\sigma_A(t)$ is in fact a normalization procedure. Thus,

$$\text{if } X = \frac{A(t) - m_A(t)}{\sigma_A(t)} \text{ then } X \sim S_\alpha(1, \beta_t, 0). \tag{60}$$

[11] In Equations (57) and (61) the parameters a and b correspond to the ones in (24).

The skewness parameter β_t is not directly used in the envelope; but it is required in order to determine the appropriate distribution of X from where K can be related with the probability ε. In the case of LFSM this parameter is given by [81 p. 345]

$$\beta_t = \frac{\beta}{\sigma_1^\alpha}\left(a^{<\alpha>} - b^{<\alpha>}\right)\int_0^\infty \left((1+x)^{H-1/\alpha} - x^{H-1/\alpha}\right)^\alpha dx + \int_0^1 \left(a(1-x)^{H-1/\alpha} - bx^{H-1/\alpha}\right)^{<\alpha>} dx \quad (61)$$

where β is the skewness intensity of the α-stable random measure in (23) and $z^{<p>} = |z|^p \cdot sign(z)$ (i.e., signed power, see [81 p. 87]). For the Log-FSM we have that $\beta_t = 0$ (see Appendix A).

4.2 Probabilistic envelope process for a traffic model based on fractional stable noise

Let us recall that the traffic model (36) (i.e., $W_j = m + aY_j$; $j = 0,1,2,\ldots$) is based on a fractional stable noise $\{Y_j, j = 0,1,2,\cdots\}$ with parameters m and a representing the mean number of arrivals per time unit and a scaling factor respectively. For this model, the following envelope process $\hat{W}$ is proposed

$$\hat{W} = m + K\sigma_W \quad (62)$$

where $\sigma_W = a\sigma_Y$. The scale parameter σ_Y can be computed from equations (57) and (58) (recall that we are using the notation $\sigma_Y = \sigma_1$, see Section 2.5). In the particular case where the model uses fractional stable noise with symmetrical marginal distributions we can use (34) and claim that[12]

[12] For an in-depth analysis of this approach using symmetric processes see Appendix B.

$$\sigma_W = a \cdot \left[\sum_{k=-\infty}^{\infty} \phi^{\alpha}(k) \right]^{1/\alpha} . \tag{63}$$

The parameter K is uniquely determined by fixing the overshoot probability ε :

$$P\{W_j > \hat{W}\} = P\{W_j > m + K\sigma_W\} = P\left\{ \frac{W_j - m}{\sigma_W} > K \right\} = \varepsilon . \tag{64}$$

It can be observed that

$$\text{if } Y = \frac{W_j - m}{\sigma_W} \text{ then } Y \sim S_\alpha(1, \beta_W, 0). \tag{65}$$

Unfortunately, in general, there are no closed form expressions for the distribution of an α-stable random variable that would allow us to exactly determine K analytically from the probability ε . Some asymptotic expressions are available (see [81 p. 16]), but they only hold for large values of *K*. Formulae for power series expansion of stable density functions are also available [70 p. 14] and have the appealing feature of being appropriate for real-time computations, but they are also prone to inaccuracies due to numerical precision [70 p. 14]. In this work, stable densities are computed using the inverse Fourier transform of the characteristic function [57 p.146]. This procedure is quite accurate [70 p. 19], although computationally intense. Using this method, the characteristic function of Y can be numerically inverted in order to obtain its probability density function and then a table of values of K versus ε (i.e., K versus $P\{Y > K\}$) can be generated. As an example, the tail probabilities of four *SαS* random variables were obtained using this method and are shown in Figures 4.1 and 4.2.

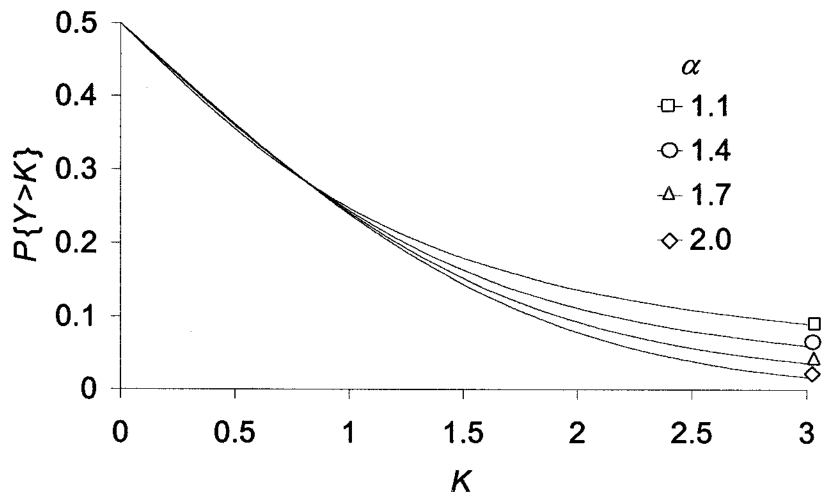


Figure 4.1 Tail probability of some *SαS* random variables in the range [0,3].

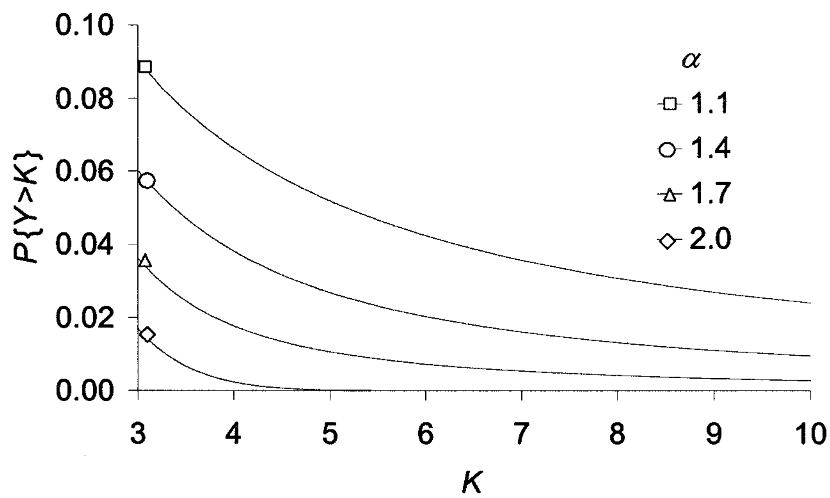


Figure 4.2 Tail probability of some *SαS* random variables in the range [3,10].

4.3 Static resource allocation

Available queueing analyses with α-stable self-similar models suggest that the buffer requirements are considerably increased by the presence of long-range dependence and high-variability properties in the network traffic and that this situation is not alleviated by reasonable increases in the buffer size [33][46][71][72]. Since an indiscriminate increase in buffer sizes is not a viable option, in order to provide quality of service guarantees, the deployment of high-speed links is required along with the development of techniques for proper dimensioning of network elements. In this section we address this issue using probabilistic envelope processes and traffic models based on α-stable self-similar stochastic processes.

4.3.1 Bandwidth allocation with a loss constraint

Let us consider the previously described probabilistic envelope for the increments of an arrival process. This envelope can be used by a constant bandwidth allocation policy in a very straightforward manner. This is due to the fact that the envelope can be thought of as the amount of bandwidth required to reduce, to an arbitrarily selected value, the percentage of time in which the input traffic exceeds the allocation. With such allocation policy, the magnitude of the actual data loss can be easily estimated (see the derivation in Appendix D). This technique is illustrated with the following case study.

WAN traffic

In this example we use the file BC-Oct89Ext [95], which contains one of the traces collected by researchers at Bellcore. This trace contains the time stamp and size of 1E6 external packets (packets heading out of Bellcore's facilities to the Internet). This represents 122,797.83 seconds of WAN traffic. From this file we obtained the number of packets per 10-second time slot, resulting in 12,279 records. The trace shows a high level of variability and is shown in Figure 4.3.

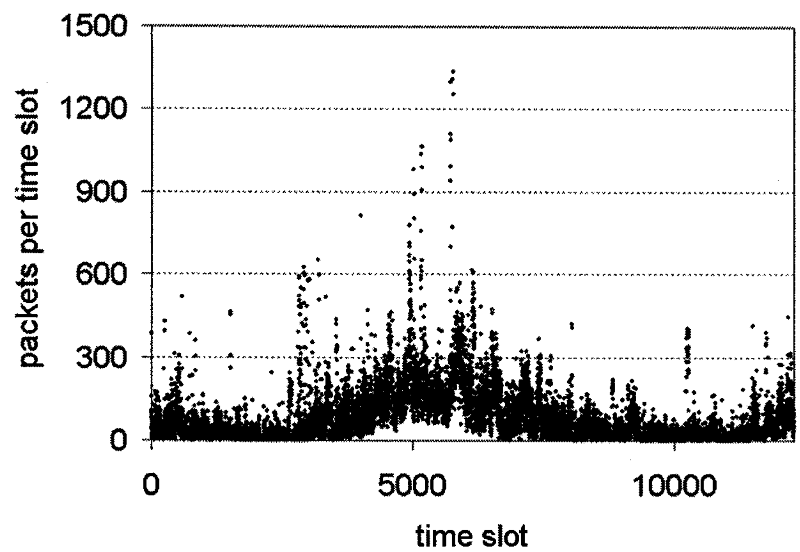


Figure 4.3 WAN traffic trace.

This trace has a sample mean m=81.43 [packets/time_slot] and standard deviation s=99.02 [packets/time_slot]. We used the procedure described in Section 2.2.2 to fit the sample data with stable parameters. The scale parameter was found to be $\sigma_W = 56.23$ [packets/time_slot] and the index of stability $\alpha = 1.61$. The estimate for the skewness parameter exceeded its maximum valid value. This may be due to a number of factors such as some limitations of the parameter estimation method in this region of the parameter space, or to the fact that the data is not drawn from an α-stable distribution but we are rather fitting the data into a stable law. We set it to the maximum permissible value (i.e. $\beta_W = 1$) which corresponds to the case of totally skewed to the right α-stable random variables. Since α >1, we can consider the sample mean to be an estimate of the shift parameter.

Assume we want to transfer this traffic stream through a network node while maintaining maximum loss guarantees. In order to determine the amount of bandwidth to be allocated, we need to find the envelope process that probabilistically bounds the amount of packets arriving in a time slot. From the traffic characterization phase we can determine that the envelope process is $\hat{W} = 81.43 + 56.23K$. As previously explained, K is selected according to the required quality of service parameter ε from a table of K vs. $\varepsilon = P\{Y > K\}$ where $Y \sim S_{1.61}(1,1,0)$. The envelope is implemented through a dropper section that detects and drops the packets exceeding the allocated system capacity per time slot.

Figure 4.4 shows the results. The solid line corresponds to the experimentally obtained fraction of slots with losses for different values of K in the range[13] [0, 10]. Each point of this line was obtained by dividing the number of times the traffic samples exceeded a particular bandwidth allocation (envelope process with a specific K) by the total number of samples in the traffic trace. Observe that this curve closely matches the tail probability of the random variable $S_{1.61}(1,1,0)$ (line with circles) indicating that, due to the good fit, K can be reliably selected from the 1.61-stable distribution. For comparison purposes we are also including the tail probability of a Gaussian random variable (dashed line) of zero mean and unitary scale parameter. It is interesting to observe that the asymmetry in the distribution of the data and its heavy tail prevents us from using the Gaussian distribution for determining K except for a few values around 1.5.

[13] We do not consider values of K less than 0 since we restrict our attention to the cases where the amount of bandwidth to be allocated is at least the mean value of the traffic process.

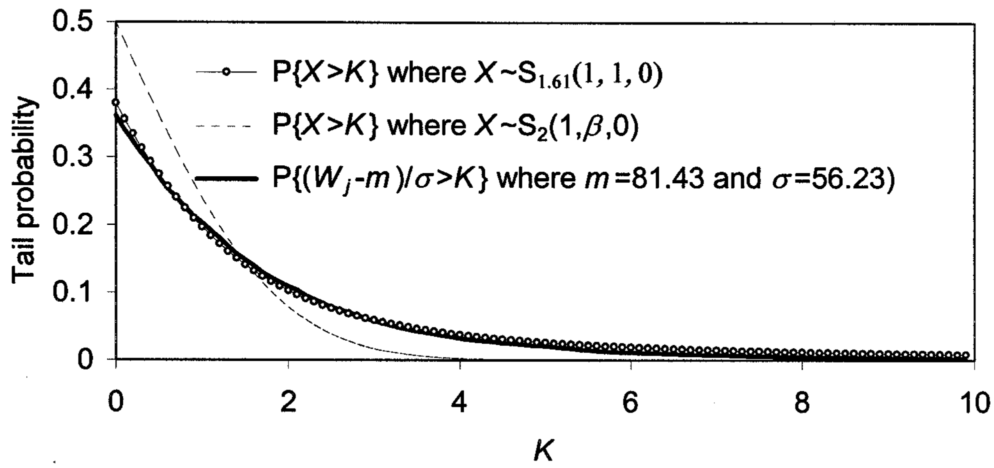


Figure 4.4 Experimental overshoot probability for a case study with WAN traffic and comparison with the tail probability of the Gaussian and 1.61-stable distributions.

The expected magnitude of losses can be computed as indicated in Appendix D. As an example of this, the solid line appearing in Figure 4.5 was obtained from Equation (D.7). It depicts the fraction of data loss expected to be achieved by the previously described service policy. For comparison purposes, the graph also includes results using a probabilistic envelope process that assumes a Gaussian marginal distribution in the traffic (dashed line). The parameters of this envelope come from the sample mean and sample standard deviation[14], i.e. $\hat{W}_j = 81.43 + 70.02K$.

---

[14] The sample standard deviation has to be scaled down by $\sqrt{2}$ in order to obtain the scale parameter required by the probabilistic envelope process. This is due to the fact that, for a Gaussian distribution its scale parameter σ and standard deviation s are related by the expression $\sigma^2 = \frac{1}{2}s^2$ [81 p. 20].

From Figure 4.5 it is observed that the model using the 1.61-stable distribution predicts a slower decay of the data loss with increasing bandwidth than the Gaussian assumption. This is due to the fact that the model with α=1.61 takes into consideration the heavy tail of the traffic and the other does not. Therefore, they lead to substantially different assumptions on the levels of data loss occurring in the system.

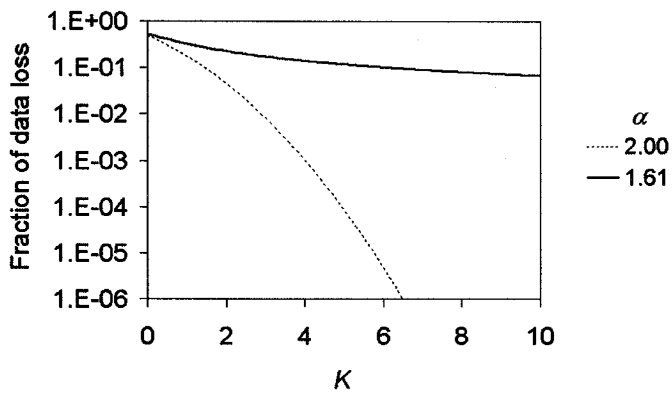


Figure 4.5. Expected fraction of data loss using α-stable and Gaussian models (WAN traffic).

4.3.2 Resource allocation with loss and delay constraints

Consider a single-server queueing system of constant-service rate C with cumulative arrival process $A(t)$. Assume a service policy where the bits or packets exceeding a probabilistic envelope process $\hat{A}(t)$ are discarded before entering into the queueing system. Denote by $q^*(t)$ the resulting buffer occupancy and let the time origin correspond to the beginning of a busy period of length T (i.e., $T = \{t : t \geq 0, q^*(t) > 0\}$). An upper bound for the buffer occupancy $q^*(t)$ during the busy period is given by

$$\hat{q}(t) = \hat{A}(t) - C \cdot t \ , \ t \in T . \tag{66}$$

We can analyze the envelope for the buffer occupancy as to determine the maximum buffer requirements of the input process. It is appropriate to mention that this methodology was first used by Fonseca *et al.* in [30] to analyze a queueing system with fractional Brownian input. In this work we generalize this analysis to the α-stable case. Let us define

$$q_{max} = max\{\hat{q}(t)\}. \tag{67}$$

In order to carry out the maximization procedure implied by (67) it is useful to use (53) and its related expressions (54), (55) and (56) and restate (66) as

$$\hat{q}(t) = K \cdot |a| \cdot \sigma_1 \cdot t^H - (C - m) \cdot t \ , \ t \in T. \tag{68}$$

Setting the first derivative of (68) to zero yields the time t^* for which an extreme value is reached. This derivation shows that

$$t^* = \left(\frac{K \cdot |a| \cdot H \cdot \sigma_1}{C - m} \right)^{\frac{1}{1-H}} \ , \ C > m. \tag{69}$$

The second derivative of (68) confirms that $\hat{q}(t)$ takes on its maximum at the value given by (69). Now the maximum of the envelope can be determined by

$$q_{max} = \hat{q}(t)\big|_{t=t^*}. \tag{70}$$

After some algebra implied by (70), the resulting expression for $q_{\max}$ is

$$q_{max} = (1 - H) \cdot \left(\frac{H^H \cdot K \cdot |a| \cdot \sigma_1}{(C - m)^H} \right)^{\frac{1}{1-H}}. \tag{71}$$

Recall that the parameter K let us reduce to an arbitrary small value the percentage of time in which the input traffic exceeds the allocation and therefore, reduce data loss. However, due to the exponent in (71) this might result in very large values for $q_{\max}$.

Although the magnitude of data loss can be accurately computed as explained in Appendix D, at this point we can provide an approximate relation (which has also been implicitly used in [30]). Let us denote by $q(t)$ the buffer occupancy in the system if the input traffic was not limited by the envelope process. The following equation would describe the evolution of the buffer occupancy over the busy period as previously defined:

$$q(t) = A(t) - C \cdot t \ , \ t \in T . \tag{72}$$

If we consider network operation in a state with high traffic intensity so that the buffer is non-empty most of the time, we can approximate[15] the probability $P\{q(t) > \hat{q}(t)\}$ with $t \in T$ by the probability $P\{A(t) > \hat{A}(t)\} = \varepsilon$. Under this assumption, we can claim that $P\{q(t) > \hat{q}(t)\} \cong \varepsilon$ with $t \in T$. Thus, because q_{max} is the maximum value of $\hat{q}(t)$, we have that $P\{q(t) > \hat{q}(t)\} \geq P\{q(t) > q_{max}\}$ and therefore $\varepsilon \geq P\{q(t) > q_{max}\}$ with $t \in T$. In addition, ε, being approximately the probability that the buffer occupancy exceeds q_{max}, can be used, in turn, to approximate the loss probability in a finite buffer system of size q_{max} if it is "small enough" [83 Chapter 4].

[15] Note that, using (66) and (72), the probability $P\{q(t) > \hat{q}(t)\}$ with $t \in T$ can be transformed to $P\{A(t) - Ct > \hat{A}(t) - Ct | q^*(t) > 0\} = P\{A(t) > \hat{A}(t) | q^*(t) > 0\}$. In addition, from the theorem on total probability [57 p. 52] we have that $P\{A(t) > \hat{A}(t)\} = P\{A(t) > \hat{A}(t) | q^*(t) = 0\} P\{q^*(t) = 0\} + P\{A(t) > \hat{A}(t) | q^*(t) > 0\} P\{q^*(t) > 0\}$. Therefore, the probability $P\{A(t) > \hat{A}(t) | q^*(t) > 0\}$ can be approximated by $P\{A(t) > \hat{A}(t)\} = \varepsilon$ under the assumption that the network traffic is so intense that $P\{q^*(t) = 0\} \cong 0$.

Under previous assumptions, (71) can be used to find the bandwidth R required to guarantee that the buffer occupancy in the system does not exceed a buffer size B with probability 1-ε, i.e.,

$$R = m + B^{(H-1)/H}\left(K \cdot |a| \cdot \sigma_1\right)^{1/H} H(1-H)^{(1-H)/H} \tag{73}$$

At this point it is worth mentioning that Equation (73) generalizes some results previously found by other researchers. This result is consistent with the one obtained by Fonseca *et al.* [30 Equation (5)] and by Stathis *et al.* [84 Equation (12)] using fractional Brownian motion (this process can be regarded as a particular case with α=2). However, an important difference should be pointed out. Whereas for the Gaussian case (i.e., α=2) small values of ε are achieved with relatively small values of K, this is not the case as α reduces its value (thus the tail becomes "heavier", see Figures 4.1 and 4.2) where significantly larger values of K are required in order to achieve similar values for ε. Because of the exponent affecting K in (73), this situation may severely increase the bandwidth requirements of an aggregate traffic flow exhibiting heavy tails when compared to the Gaussian case.

The following example illustrates how the ideas contained in this section are used for resource provisioning in a network node transporting aggregate video traffic.

QoS-aware video communications

In the proposed case study we consider a traffic flow coming from the aggregation of several MPEG video sequences obtained from the University of Würzburg in Germany [96]. The resulting sequence $\{W_j^*, j=0, 1, 2, \ldots\}$ is shown in Figure 4.6, it consists of 2627 samples with a sample mean m=3,655,758 [bits/s]. The remaining parameters for the model described by (36) were estimated as follows. We used the method described in Section 2.2.2 to determine the parameters of the marginal distribution of the sequence, yielding a scale parameter $\sigma_W = 281{,}615$ [bits/s] and an index of stability $\alpha = 1.97$. The index of stability resulted very close to 2, thus the value of β_W is rather irrelevant (see Equation (2)), we set it to zero. The parameters that describe the evolution of the process in time were determined as described below. In order to estimate the Hurst parameter, we took into account that the scale parameter of an aggregate of W_j^* over T time slots follows the law $\sigma_T = T^H \sigma_1$ (see Equation (56), here we are using the notation $\sigma_{aggregation\_level}$). Therefore, H was found as the slope of a linear regression model of $\log(T)$ vs. $\log(\sigma_T)$ yielding a value $H = 0.91$. It was also concluded that the sample normalized autocorrelation function of this sequence was more appropriately represented using an anti-balanced model (a=-b=1 in Equation (24)).

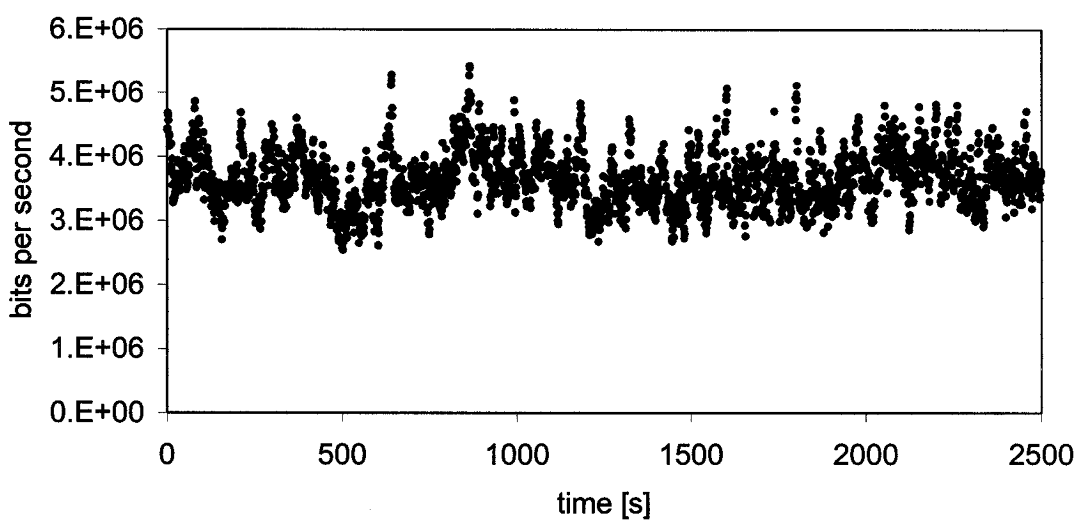


Figure 4.6 Aggregate video traffic trace.

Assume we want to determine the amount of resources required to transfer such traffic class through a network node with quality of service guarantees specified in terms of maximum tolerated delay and maximum percentage of losses. From the delay specification, the value of the maximum buffer size B can be determined. The fraction of data loss can be assumed to be approximately equal to the probability that the buffer occupancy in an infinite-buffer system exceeds B if it is small [83 Chapter 4], which is the usual case in modern networks. In turn, this buffer overflow probability can be approximated by the probability ε under the conditions previously explained.

In order to apply (73), we need to determine the parameter K by looking up on a table of K versus $\varepsilon = P\{Y > K\}$ where $Y \sim S_{1.97}(1,0,0)$. In this example let us consider the following values for ε: 0.01, 0.001, 1E-4 and 1E-5, which results in the values for K shown in the second row of Table 4.1.

Table 4.1 Probabilities $\varepsilon=P\{Y>K\}$ and resulting values of K from a distribution $Y \sim S_{1.97}(1,0,0)$

ε	0.01	0.001	1E-4	1E-5
K	3.39	5.06	12.89	40.76

For a given buffer size B, Equation (73) can be used to determine the amount of bandwidth R to be allocated to this traffic class as to guarantee, with probability 1-ε, that the buffer occupancy will not exceed B. Figure 4.7 shows the resulting bandwidth requirements under different buffer sizes for the considered cases. As depicted in the figure, for relatively high levels of losses, moderate increases in the allocated bandwidth results in a much more effective way to reduce losses than large increases in the buffer size. However, due to the heavy tail nature of the traffic process, in order to achieve relatively low levels of losses more significant increases in bandwidth allocation are required.

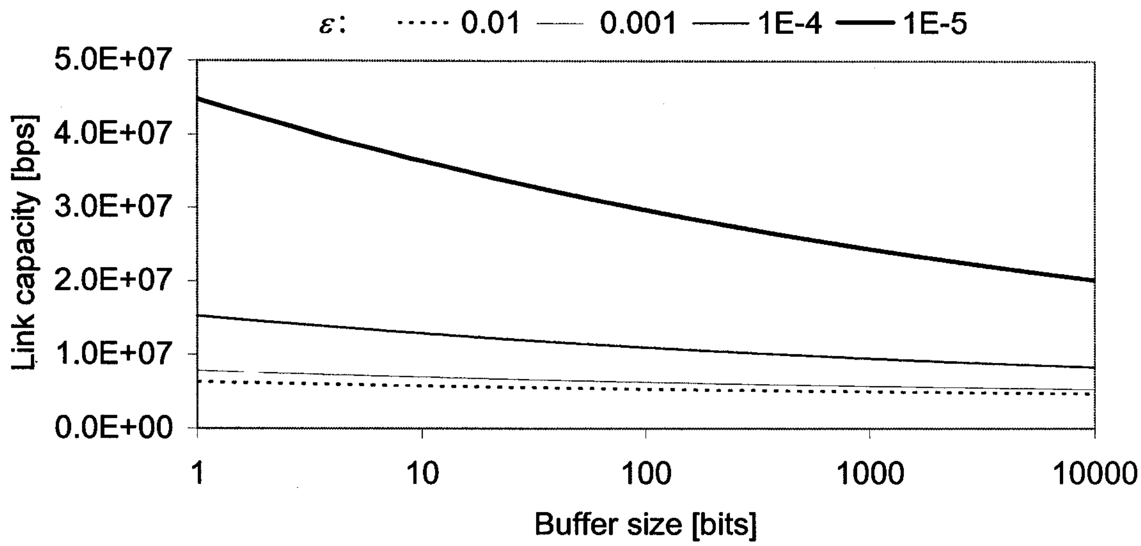


Figure 4.7 Resource provisioning chart for a case study with aggregate video traffic.

4.3.3 Probabilistic envelope processes and connection admission control

Consider an aggregate of N independent flows, where the number of arrivals for the i-th flow in the interval $[0,t]$ is modeled as

$$A^{(i)}(t) = m^{(i)}t + a^{(i)}X^{(i)}(t) \tag{74}$$

where $X^{(i)}(t)$ is fractional stable motion (either LFSM of Log-FSM), $m^{(i)}$ is its mean input rate and $a^{(i)}>0$ is its scaling factor. Assume that the FSM process has symmetrical α-stable marginal distributions (i.e., for a fixed t we have that $X^{(i)}(t) \sim S_\alpha\left(\sigma_t^{(i)},0,0\right)$). This does not affect the validity of the issues that we will discuss in this section, but some expressions are more readable and can be easily generalized. Using again the basic properties of α-stable random variables (see Equations (4) and (5)) it is observed that at time t the marginal distribution of each flow is $A^{(i)}(t) \sim S_\alpha\left(a^{(i)}\sigma_t{}^{(i)},0,m^{(i)}t\right)$, therefore the total cumulative arrival process

$$A(t) = \sum_{i=1}^{N} A^{(i)}(t) \tag{75}$$

is distributed as $A(t) \sim S_\alpha\left(\left(\sum_{i=1}^{N}\left(a^{(i)}\sigma_t^{(i)}\right)^\alpha\right)^{1/\alpha},0,\sum_{i=1}^{N} m^{(i)}t\right)$ at time t (see Equation (6)). Now assume that the flows are identically distributed. In this case, the marginal distributions of $A(t)$ become

$$A(t) \sim S_\alpha\left(N^{1/\alpha}a\sigma_t,0,Nmt\right). \tag{76}$$

It is observed that whereas the mean of the aggregate process increases proportionally to N, its scale parameter grows as $N^{1/\alpha}$ with $1 < \alpha \leq 2$. Taking into consideration that these two parameters can be used in an envelope process to determine the resource allocation for the aggregate, the previous result implies that the bandwidth requirements grow slower than N. To further investigate this issue, let us define the multiplexing gain as $NR^{(1)} / R^{(N)}$ where $R^{(1)}$ and $R^{(N)}$ are the bandwidth requirements of 1 and N flows, respectively. Based on Equation (76), let us propose the following probabilistic envelope process for an aggregate of N iid flows (see Equation (53)):

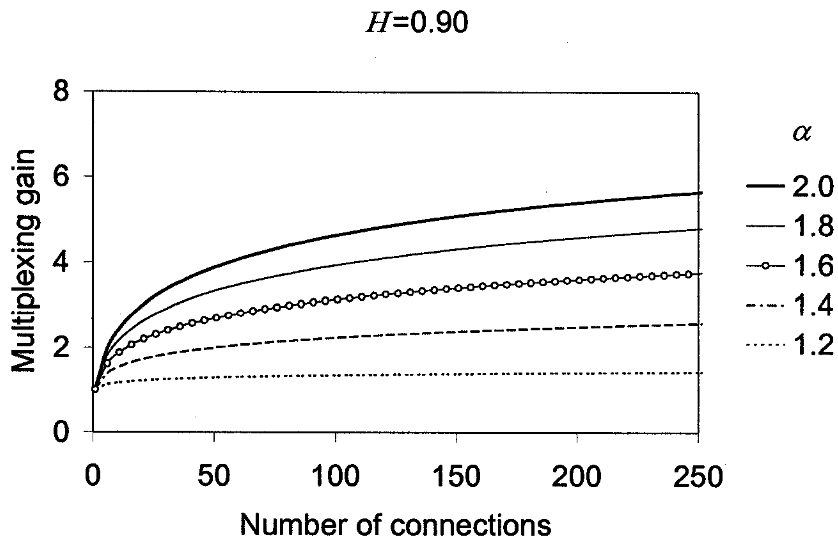
$$\hat{A}^{(N)}(t) = Nmt + KN^{1/\alpha}a\sigma_t. \tag{77}$$

Let us consider the bandwidth allocation policy where data exceeding the envelope process are simply dropped. Note that when (77) applies to time intervals of unitary length it becomes a probabilistic envelope for the rate and then, the bandwidth requirements can be determined as $R^{(1)} = \hat{A}^{(1)}(1)$ and $R^{(N)} = \hat{A}^{(N)}(1)$. Therefore, the multiplexing gain can be computed as $\left(N\cdot\hat{A}^{(1)}(1)\right)/\hat{A}^{(N)}(1)$ yielding $(Nm+KNa\sigma_1)/\left(Nm+KN^{1/\alpha}a\sigma_1\right)$. Since $1<\alpha\leq 2$, it is immediately observed that the maximum achievable multiplexing gain corresponds to the Gaussian case and that it decreases as the marginal distributions of the input process exhibit heavier tails.

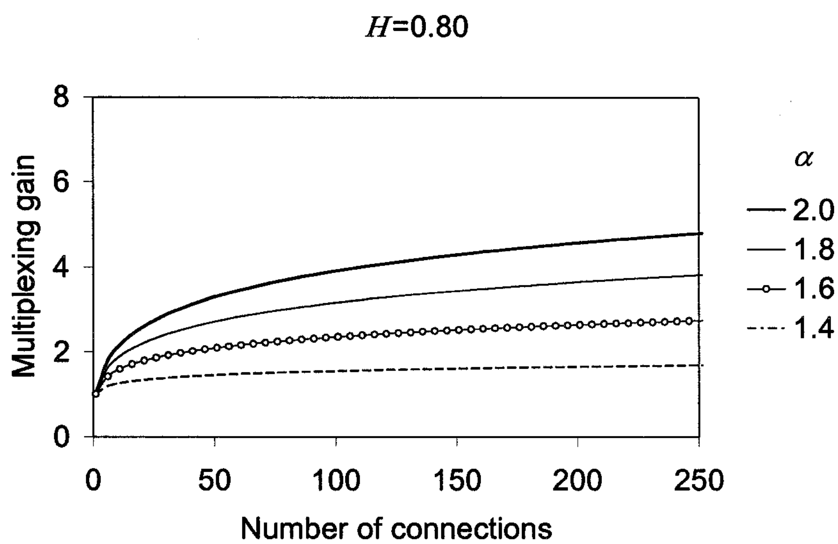
For the bandwidth allocation policy, where the incoming packets are stored in a queue, Equation (73) becomes

$$R^{(N)}=Nm+B^{(H-1)/H}\left(K\cdot N^{1/\alpha}|a|\cdot\sigma_1\right)^{1/H}H(1-H)^{(1-H)/H} \tag{78}$$

and the multiplexing gain can be computed as $\left(N\cdot R^{(1)}\right)/R^{(N)}$. In this case the effect of the parameters on the multiplexing gain is not so evident. When the input traffic is modeled using the Gaussian self-similar process it has been observed in [30] that the multiplexing gain increases with increasing values in variance, Hurst parameter H and number of sources N. Due to the similitude of the equations, these observations are also valid in the α-stable non-Gaussian case (take into consideration that the scale parameter replaces the standard deviation). In contrast, increasing thickness of the tail of the marginal distribution, expressed through the value of α, negatively affects the multiplexing gain. This is exemplified in Figures 4.8 (a) and 4.8 (b) with a case study using typical values for the Hurst parameter. Although the actual values of the multiplexing gain shown in the case study depend on the selected parameter values, it can be generalized that the heavier the tail of the marginal distribution is (lower values of the index α), the smaller the multiplexing gain becomes. In other words, whereas multiplexing of scaled up processes increases the multiplexing gain, increasing the probability of occurrence of sporadic large fluctuations has the opposite effect.



(a)



(b)

Figure 4.8 Examples of multiplexing gain for the bandwidth allocation scheme with loss and delay constraints. Case study with the following parameters: m=10, $a \cdot \sigma_1$=100, K=1, B=10.

With the results contained in this section we propose the following connection admission control, similar to the ones proposed in the past for Gaussian models (e.g., [30], [84]). During the connection-establishment phase, the model parameters of a new traffic flow are provided to the network by its source along with the desired quality of service specified in terms of the maximum tolerated delay and maximum percentage of losses. From the maximum delay D_{max}, the network determines the value of the maximum buffer size B, since it is reasonable to require that $B \leq RD_{\max}$ where R is the link rate. If a cost function relating the link rate and buffer size is also available, the pair (B,R) that minimizes the cost can be uniquely determined (e.g., see [22], [27], [77]). Other criteria, like minimum bandwidth (e.g., [88]), can alternatively be taken into consideration. The fraction of data loss can be assumed to be approximately equal to ε under the conditions previously explained. The new connection is accepted only if the required capacity computed through (78) can be accommodated within the available resource. If only loss constraints are to be satisfied, Equation (77) with t=1 can be used instead.

4.4 Conclusions

In this section we have provided some experimental evidence supporting the claim that traffic models based on α-stable self-similar stochastic processes are more suitable to represent aggregate traffic flows than Gaussian models. This is due to the fact that some traffic classes exhibit asymmetry and heavy tails in their marginal distributions, which are properly captured by α-stable random variables, but not by the Gaussian one. This suggests that use of Gaussian models in some cases may be questionable. They may lead to an erroneous estimation of bandwidth requirements, which results in unrealistic assumptions during the design, analysis and dimensioning of network elements and the consequent lack of quality of service perceived by end-users.

Although it is not hard to accept that traffic modeling based on α-stable self-similar stochastic processes allows for more accurate modeling, practical use of these models is currently limited by increased computational cost, limited mathematical tractability and scarce availability of performance analyses. The absence of closed form expressions for the probability density functions of α-stable distributions and their general lack of second order statistics have limited their applicability in traditional problems of teletraffic engineering. We addressed this issue and showed that, in spite of the increased complexity, these models can be effectively used in network analysis and control.

Assuming a traffic model based on α-stable self-similar stochastic processes, we proposed an extension to the α-stable case of the concept of probabilistic envelope processes, previously defined for more traditional models. We represented the bandwidth demand imposed by the traffic model with an envelope process and showed that with this approach we can simply and effectively deal with much of the argued complexity encountered in α-stable models and develop techniques for proper dimensioning of network elements. We concluded that the presence of heavy tails in the distribution of the traffic process negatively affects some elements of network design (buffer sizes, link rates and multiplexing gain) when compared with Gaussian assumptions.

Chapter 5

Linear prediction

α-stable long-range dependent processes allow for parsimonious modeling of aggregate network traffic. This characteristic is of great usefulness in performance evaluation and system design. An immediate reason is that these models can be used for the creation of realistic conditions in simulated environments. A more subtle reason is that modeling of the statistical dependencies among samples of a traffic flow can be exploited in order to improve the performance of applications and the network as a whole. For instance, since some statistical information about the future of a network traffic process is contained in measurements from the past and present, future bandwidth requirements can be estimated using prediction techniques. This knowledge can be used to propose efficient resource management mechanisms.

Development of applications based on the prediction of α-stable processes is currently limited by computational cost, limited mathematical tractability and scarce availability of performance analysis. The most evident limitation is that the value of the parameter α conditions the existence of some moments, which in turn limits the applicability of traditional statistical procedures. As a result of this, traditional criteria for selecting the best predictor based on second order moments cannot be applied. Instead, several authors have proposed to use alternative criteria like the minimization of the expected absolute prediction error or the minimization of the dispersion of the error distribution [18].

Based on the minimum dispersion criterion, a novel method for computing the prediction coefficients of a linear predictor for α-stable long-range dependent processes was introduced by Gallardo *et al.* in [32]. Although the performance of this algorithm has not been thoroughly studied, it is beginning to find some applications in proposals for traffic and QoS control (e.g., [34], [35]).

5.1 A method of linear prediction for symmetric fractional stable noise

A linear combination of N past samples of a symmetric fractional stable noise can be used to estimate a sample Y_j of the process as follows:

$$\hat{Y}_j = \sum_{i=1}^{N} a_i \cdot Y_{j-i} . \tag{79}$$

The prediction coefficients $\{a_i, i = 1,2,\cdots,N\}$ are constants that are computed in such a way as to minimize a metric related to the prediction error. The computation of these coefficients in the case of α-stable random processes is not trivial and is still object of ongoing research. The linear space generated by the observations of the process $\{Y_j\}$ is a Banach space (with $1 \le \alpha < 2$). It lacks the useful properties for the linear estimation problem that are present in the Hilbert space corresponding to the linear space generated by samples of processes with second-order statistics (including Gaussian processes) [70 pp. 98, 157].

As previously mentioned, the minimum dispersion criterion, proposed in [85], can be used as a suitable criterion to obtain the set of prediction coefficients. In order to apply this criterion, we need to find the dispersion of the prediction error distribution and then determine the set of coefficients that minimizes such dispersion function. To this end, let us use the constructive definition of fractional stable noise (see Equation (32)) to obtain the prediction error e_j as follows:

$$e_j = Y_j - \hat{Y}_j = \sum_{k=-\infty}^{\infty}\left(\phi(j-k) - \sum_{i=1}^{N} a_i \cdot \phi(j-i-k)\right) \cdot U_k . \tag{80}$$

In order to obtain the dispersion of the prediction error γ_e we can apply the property stated by (10) into (80). Hence, the objective function is

$$\gamma_e = \sum_{k=-\infty}^{\infty}\left|\phi(k) - \sum_{i=1}^{N} a_i \cdot \phi(k-i)\right|^{\alpha} . \tag{81}$$

The values of the parameters $\{a_i, i = 1,\ldots,N\}$ must be defined in such a way as to minimize γ_e. For this purpose, a number of multidimensional optimization methods can be used. In this work, we focus on the algorithm proposed in [32] by Gallardo *et al.* It is briefly explained below for clarity. Let us depart from the partial derivative of γ_e with respect to a generic coefficient a_m. It is given by

$$\frac{\partial}{\partial a_m}\gamma_e = -\alpha \cdot \sum_{k=-\infty}^{\infty}\left[\phi(k) - \sum_{i=1}^{N} a_i \cdot \phi(k-i)\right]^{\langle \alpha - 1 \rangle} \cdot \phi(k-m) \quad \text{for} \quad m \in \{1,2,\ldots,N\} \tag{82}$$

where we used the signed power function [81 p. 87] defined as

$$z^{\langle p \rangle} = |z|^p \cdot sign(z). \tag{83}$$

It is worth mentioning that the positiveness of the second derivative of the equation for γ_e (not shown) indicates a concavity upwards. In order to solve the set of non-linear equations

$$\frac{\partial}{\partial a_m}\gamma_e = 0 \quad ; m \in \{1,2,\ldots,N\} \tag{84}$$

it was proposed in [32] to use a function $\psi(k)$ such as

$$\psi(k) \stackrel{\Delta}{=} \phi(k) - \sum_{i=1}^{N} b_i \cdot \phi\ (k-i) \tag{85}$$

so that the set of simultaneous non-linear equations (84) can be approximated by the linear set

$$\sum_{k=-\infty}^{\infty} \psi(k)\cdot\phi(k-m)=0 \quad ; m\in\{1,2,\ldots,N\}. \tag{86}$$

If we define $R(m)=\sum_{k=-\infty}^{\infty}\phi(k)\phi(k-m)$, it is easy to see from (86) that the coefficients $\{b_i, i=1,\ldots,N\}$ in (85) satisfy

$$R(m)=\sum_{i=1}^{N} b_i R(m-i), \quad m\in\{1,2,\ldots,N\}. \tag{87}$$

The values of the prediction coefficients can be obtained considering the function $\psi(k)$ in the context of (82). Thus, it can be demonstrated that

$$\sum_{i=1}^{N} a_i\cdot\phi(k-i)=\phi(k)-[\psi(k)]^{\langle 1/(\alpha-1)\rangle} \quad ; k\in\{1,2,\ldots,N\}. \tag{88}$$

Summarizing, the method introduced in [32] consists in solving the set of linear equations (87) in order to to find the set $\{b_i, i=1,\ldots,N\}$ which in turn let us obtain (through the function $\psi(k)$) the prediction coefficients $\{a_i, i=1,\ldots,N\}$ by solving a second set of linear equations given by (88).

Some tacit assumptions have been made in the method for linear traffic prediction. An important one is that the network traffic is considered to be stationary so that the parameters that describe its statistical behavior do not change. These model parameters are needed for the computation of the prediction coefficients so that they are also assumed to be available beforehand.

5.2 Performance evaluation of the prediction algorithm

A number of issues arise during the development of applications based on the prediction algorithm proposed by Gallardo *et al.* in [32]. For instance, no studies have been carried out in order to determine the sensitivity of the prediction with respect to uncertainties in the traffic model parameters. This is an important matter since the method makes use of parameters extracted from a set of samples of the actual process to be predicted. However, due to a number of factors, like the finite nature of the data set from which the parameters are extracted, there is always a discrepancy between actual and estimated values. Another issue that requires further studies comes from the processing-efficient but approximate solution for some equations in the algorithm. The effect of such approximations on the performance of prediction-based applications has been largely overlooked. Finally, it can be mentioned that determination of the appropriate order for the predictor remains as an important issue to be satisfactorily justified. Up to the author's best knowledge, these problems and their solutions have just been superficially studied or have not been addressed at all. In order to provide useful elements for a better use of this algorithm, in this work a performance study is presented consisting of three parts: a) an analysis of the impact on the linear prediction technique of possible inaccuracies in its parameters, b) a study of the reliability of the solution provided by the method for computing the prediction coefficients and, c) the effect of the predictor order on the prediction error distribution.

5.2.1 Sensitivity study

The performance of the previously described prediction algorithm was extensively investigated with both, standard[16] LFSN and standard Log-FSN under different combinations of values for their parameters. Recall that FSN processes, with symmetric marginal distributions, are completely defined by the values of their index of stability α and Hurst parameter H. In order to cover the range of common values for these parameters, the sensitivity of the algorithm was explored in the vicinity of the valid points formed by the combinations between $\alpha \in \{1.3, 1.6, 1.9\}$ and $H \in \{0.80, 0.85, 0.90\}$. They correspond to different levels of variability (high, medium and low) and the typical range of values for long-range dependence reported in the literature (e.g. [56]).

The evaluation method was carried out as follows. Given a combination of α and H, the sensitivity of the algorithm to the value of H was evaluated considering a certain estimate (i.e. $\hat{H}$) and the actual value of α. The corresponding set of purposely non-optimal prediction coefficients was computed and the dispersion of the prediction error was determined using (81). The summation in (81) was truncated including a large number of the most significant terms $\{\phi(k), |k| \leq 10{,}000\}$, which ensured a reliable computation of the dispersion. The function $\phi(k)$ is monotonically decreasing in $|k|$ and the truncated terms were relatively negligible (i.e., for the most significant truncated terms $\phi(k)/\phi(0)$ was typically in the range of 10^{-5} to 10^{-7}). The sensitivity of the prediction to the value of α was determined with the same procedure: using the actual value of H, some values for α

[16] Consistent use of normalized noise processes (see Equation (31)) turns out to be important for a proper presentation of results.

were proposed (i.e. $\hat{\alpha}$), the corresponding prediction coefficients were computed and the effect on the dispersion of the error was observed.

In all cases the order of the predictor was set to 50. This rather high order was chosen based on the experimental observation that, at this order of magnitude, increments in the order of the predictor do not produce a significant decrease in the dispersion of the prediction error distribution (this issue will be further commented in Section 5.2.3). This was done this way in order to determine the dispersion without introducing additional variability in the measurements by a small observation window size.

We start by examining how inaccuracies in the value of $\hat{H}$ affect the prediction when using an anti-balanced model ($a = -b = 1$ in (24)). The results from the tests are shown in Figure 5.1. It is observed that the linear prediction is quite robust to significant differences between the actual and estimated values of the Hurst parameter, as long as they are not too close to the point $\hat{H} = 1/\alpha$. This high level of dispersion is a consequence of the negligible correlation among consecutive samples of the process generated by the kernel of the anti-balanced process near the point $H = 1/\alpha$ (see Section 2.5).

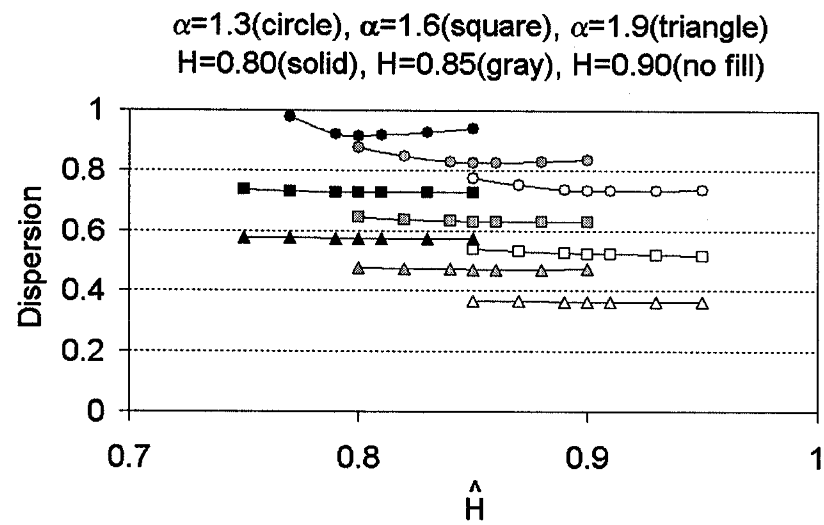


Figure 5.1 Effect of discrepancies between H and $\hat{H}$ on the dispersion of the prediction error (anti-balanced process.)

In the second set of experiments we evaluated the robustness of the prediction algorithm under discrepancies between α and its estimated value $\hat{\alpha}$. The obtained results are shown in Figure 5.2. It is clearly noticeable that the curves show a concavity upwards with their deepest point around the value where α and $\hat{\alpha}$ coincide. These graphs also show that relatively small diferencies do not have a significant effect on the performance of the prediction. However, as $\hat{\alpha}$ gets farther from that point, and closer to $1/H$, the prediction error rapidly increases.

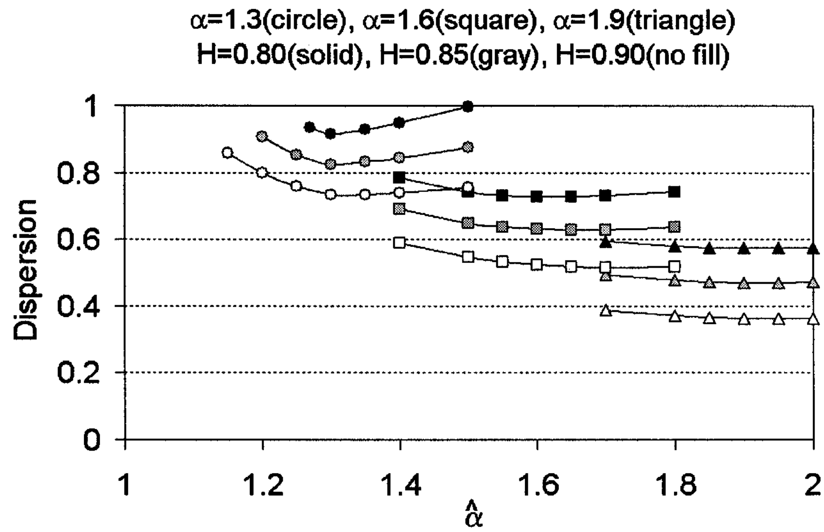


Figure 5.2 Effect of discrepancies between α and $\hat{\alpha}$ on the dispersion of the prediction error (anti-balanced process.)

The results obtained with the well-balanced process (i.e., $a = b = 1$) are shown in Figures 5.3 and 5.4. In this case the high level of dispersion observed near the point $\hat{H} = 1/\alpha$ is due to the fact that the kernel of the balanced process degenerates into a null process at the point $H = 1/\alpha$ (see Section 2.5). One difference worth mentioning with respect to the anti-balanced case occurred in the tests of sensitivity to $\hat{\alpha}$ where the minimum dispersion point had an appreciable shift from the optimal value. We found that this difference was completely due to an inherent lack of accuracy of the algorithm during the computation of the prediction coefficients, which was made more evident with the well-balanced case. A general solution to this problem will be discussed in the following section.

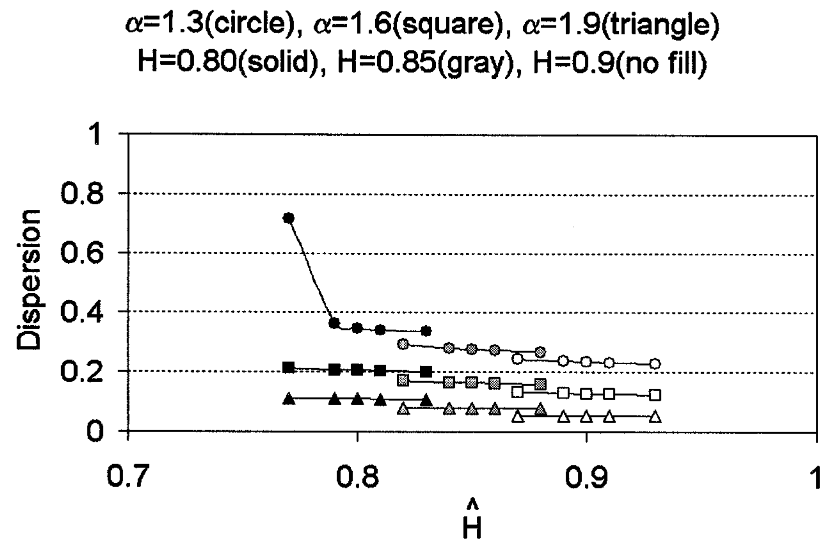


Figure 5.3 Effect of discrepancies between H and $\hat{H}$ on the dispersion of the prediction error (well-balanced process.)

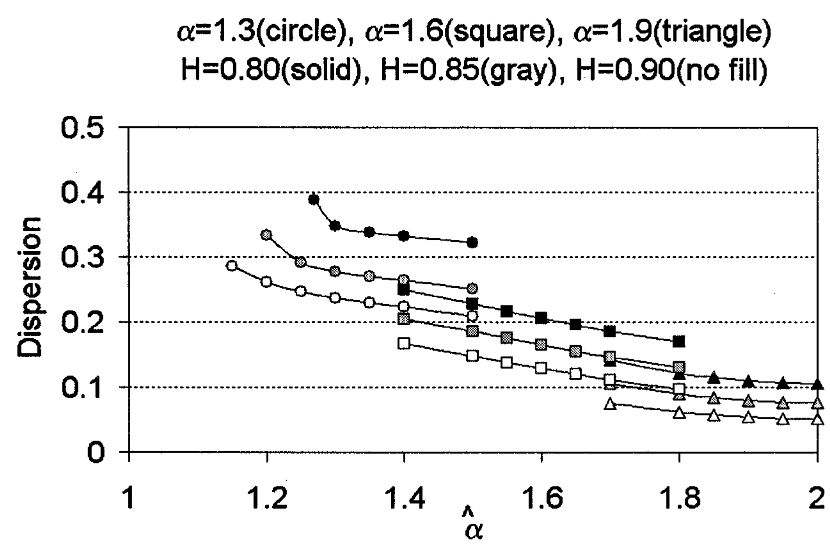


Figure 5.4 Effect of discrepancies between α and $\hat{\alpha}$ on the dispersion of the prediction error (well-balanced process.)

As to the Log-FSN case, one can only test the sensitivity of the predictor to $\hat{\alpha}$ since its parameters are related by the expression α=1/H. The corresponding results are shown in

Figure 5.5. Although the results were similar with the ones from the previous cases (including a shift in the point of minimum dispersion with respect to the optimal value for low values of the index of stability), one difference worth noting is the increasing trend in dispersion with increasing values of α. This is explained by the fact that as this parameter increases, the statistical dependencies decrease. Recall that at the point α=2 the process becomes white Gaussian noise.

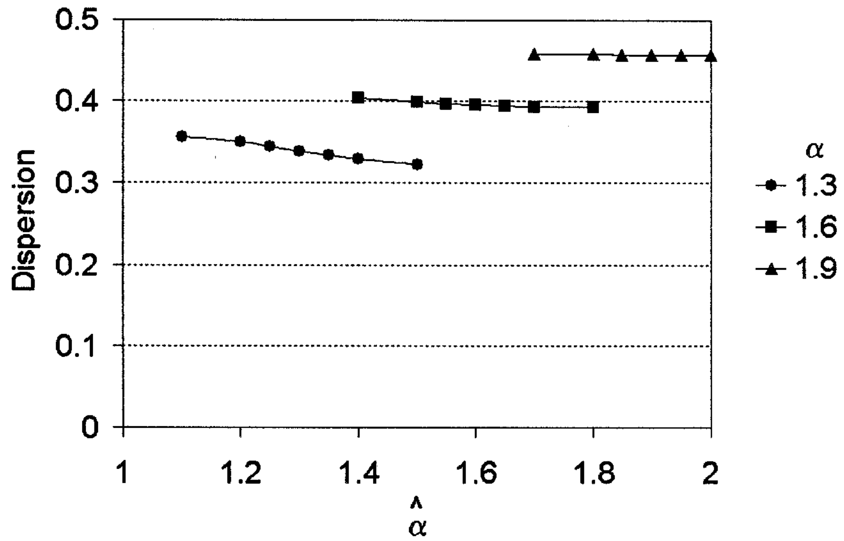


Figure 5.5 Effect of discrepancies between α and $\hat{\alpha}$ on the dispersion of the prediction error (Log-FSN process.)

5.2.2 Accuracy considerations

The set of prediction coefficients $\{a_i, i=1,2,\cdots,N\}$ was determined by solving with an approximation the set of N simultaneous non-linear equations described by (84). From the results presented in the previous section, it can be inferred that these values usually provide acceptable performance under some likely discrepancies between estimated and actual values for the statistical descriptors of the traffic process. However, such procedure may produce a set of coefficients that do not make (84) exactly zero and therefore the minimization of the dispersion of the error is not optimal. If more accuracy is needed, an additional optimization procedure will be required. Our proposal is to use the method proposed by Gallardo *et al.* [32] to get an initial estimate for the prediction coefficients and then work with these values in such a way as to get closer to the point that minimizes the dispersion.

In order to find the set of N prediction coefficients that better minimizes (81) several methods can be implemented. A straightforward approach is to use the well-known *Steepest Descent Method*, which consists in finding the minimum of a function along the direction given by the negative of the gradient. In our case the gradient of the dispersion function can be directly computed from (82) since in this case it is a vector whose m^{th} component is the partial derivative of the dispersion with respect to a_m. Once a new estimate for the minimum is found, the gradient of the function is computed again at this point and we proceed along the new direction. This procedure is repeated iteratively until an accuracy criterion is satisfied. Preliminary results showed that even though this method indeed produces a set of coefficients with better performance in terms of dispersion, it has the drawback of being excessively slow. This is due to the fact that the gradient is not a global property but a local one. One consequence of this fact is that the successive search directions, generated by this method, produce a zigzagging path towards the minimum and therefore slow convergence. A general solution for this problem was proposed by Fletcher and Reeves in [28]. They proposed to use a suitable linear combination of the current gradient and the preceding search direction. This combination may point more closely towards the actual minimum and produce faster convergence. This method is known as the *Conjugate Gradient Method*.

Using the conjugate gradient method we determined the amount of improvement in performance when an optimized set of prediction coefficients was used for the prediction procedure (see [61] for implementation details). For several combinations of α and H (where $\alpha>1/H$) the difference in performance with and without optimization was assessed, based on the values for the dispersion of the prediction error. In all cases the order of the predictor was set to 50 and, when optimizing the set of prediction coefficients, the number of iterations in the optimization procedure was heuristically set to 20. Each iteration consisted of the following steps. Firstly, the *interval of uncertainty* (i.e., the interval containing the minimum) was determined by exploring the dispersion function until three points were found for which the middle one had the least of the three values. This procedure was performed in steps of one tenth one of the most significant coefficients. Secondly, a one-dimensional technique for obtaining a local minimum was performed until the interval of uncertainty was reduced to a fraction of the original step size. Finally, by using the two endpoints of the interval and its middle point, an estimate for the minimum was computed by using inverse parabolic interpolation.

Figures 5.6, 5.7 and 5.8 show the obtained results. It can be seen that in all cases the percentage of improvement tends to zero at high values of the index of stability α. This can be explained by the fact that when α is 2 (Gaussian case) the method for computing the prediction coefficients becomes an exact method. In this region an additional optimization procedure is not recommended since the computational overload, in our tests, counted on average for about 65% of the total CPU time required to compute the prediction coefficients. The decreasing trend in improvement when α approaches one might be due to the fact that, in this region, the exponent of (82) approaches to zero, which minimizes the effect of the values that the prediction coefficients take.

Observe that the cases with high levels of improvement match with the ones observed in the previous section where appreciable shifts from the optimal point were evident.

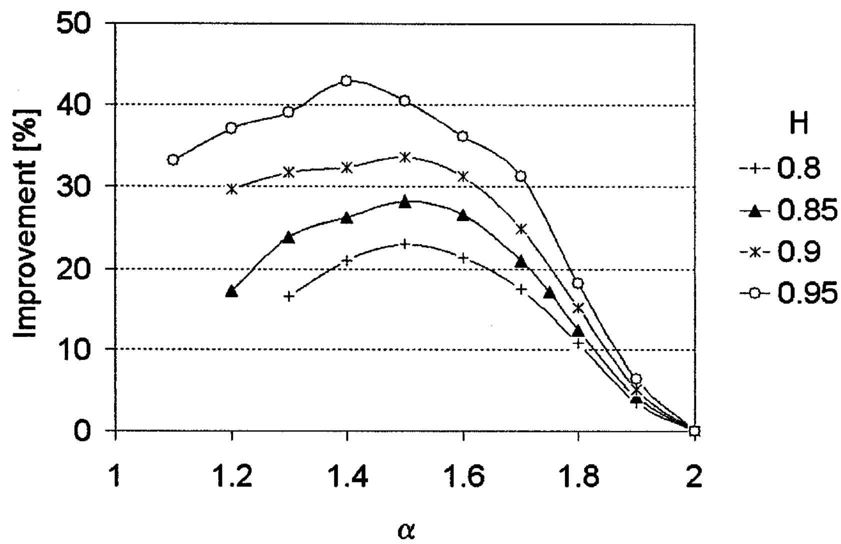


Figure 5.6 Percentage of improvement in dispersion with optimized coefficients (well-balanced process.)

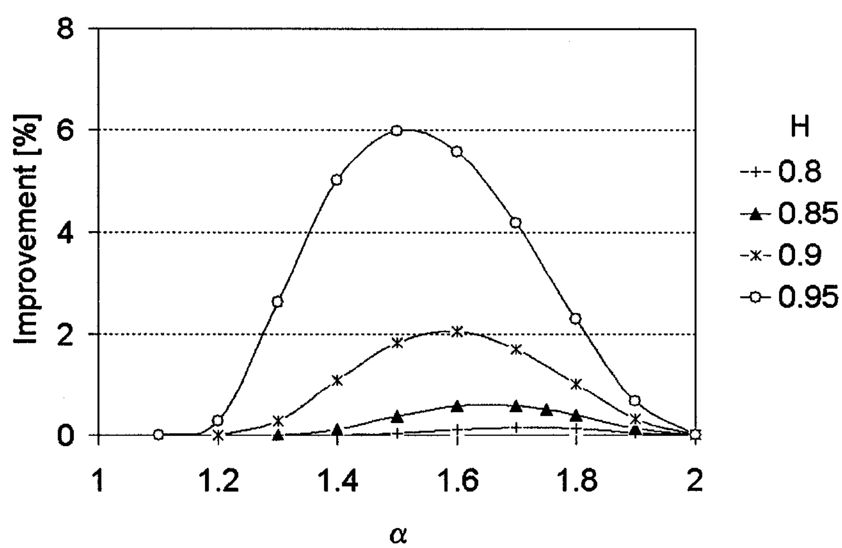


Figure 5.7 Percentage of improvement in dispersion with optimized coefficients (anti-balanced process.)

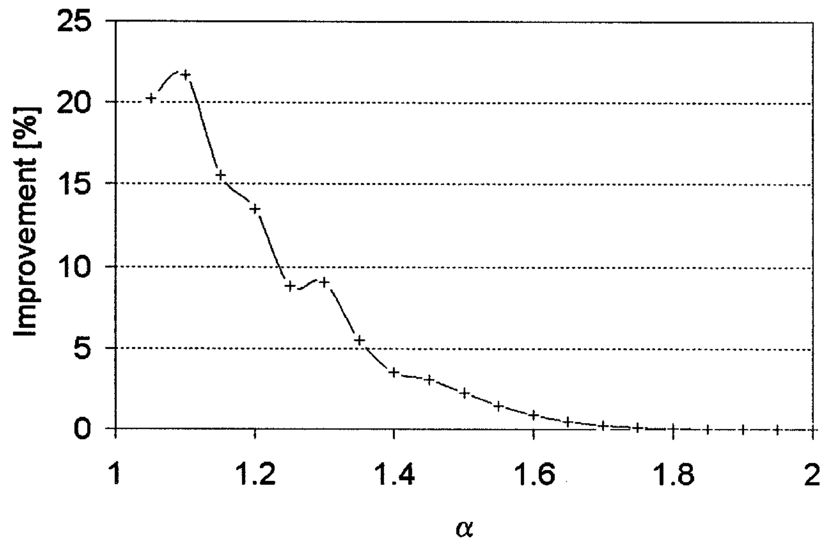


Figure 5.8 Percentage of improvement in dispersion with optimized coefficients (Log-FSN process.)

5.2.3 Filter order selection

The effect of the predictor order on the magnitude of the dispersion of the error was investigated using standard FSN processes and diverse combinations in their parameters α and H. Figures 5.9, 5.10 and 5.11 below show the results for the well-balanced FSN, anti-balanced FSN and Log-FSN respectively. These figures confirm our intuitive notion that the larger the predictor order, the smaller the error (dispersion) is going to be. However, what is interesting is the fast decay rate of the dispersion for the first decades whereas for larger orders there is a nearly negligible improvement. Taking into consideration that the complexity of the predictor will also increase, it is hard to justify the use of large predictors in order to take into account long-range correlations.

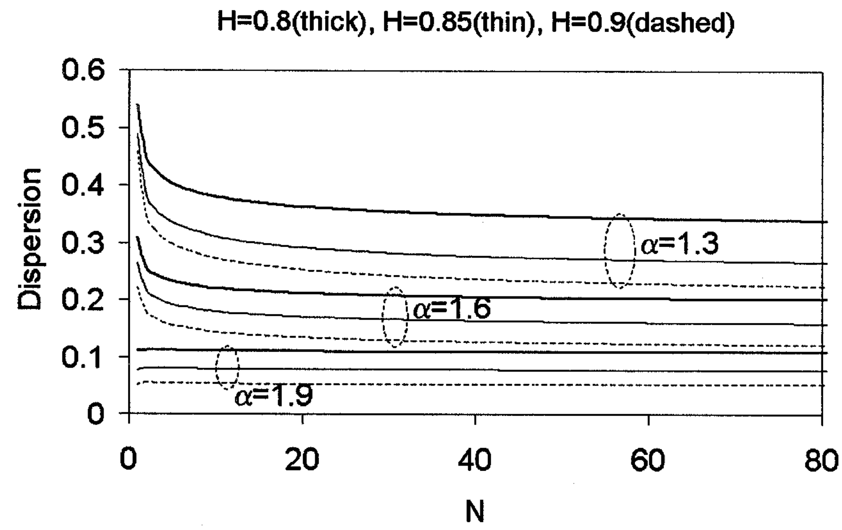


Figure 5.9 Dispersion of the prediction error with different predictor orders (standard well-balanced process.)

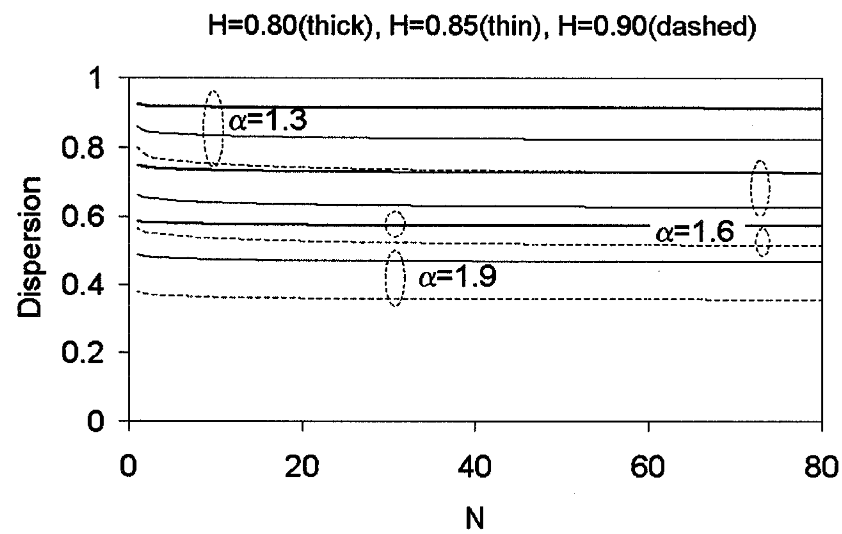


Figure 5.10 Dispersion of the prediction error with different predictor orders (standard anti-balanced process.)

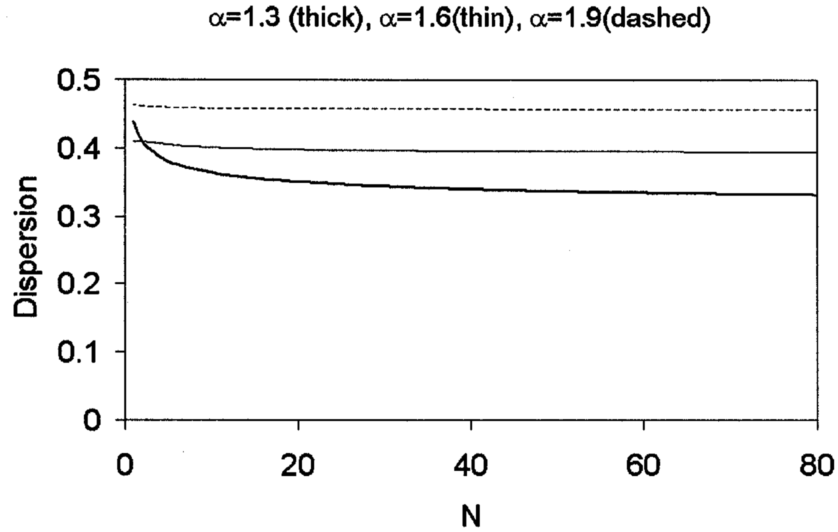


Figure 5.11 Dispersion of the prediction error with different predictor orders (standard Log-FSN process.)

These results support the observation that long-range correlations are of marginal usefulness in linear prediction. This statement is also supported by the following two observations. First, by the fact that past samples of the traffic process, falling out of the observation window set by the order of the filter, cannot be taken into consideration. Consequently, their statistical information about long-range correlations is lost. Second, additional observations on the relative magnitudes of the prediction coefficients led us to conclude that far samples are expected to provide negligible contributions to the estimation. Therefore, even if we increase the observation window size as to include past samples, providing information about long-range correlations, it is expected that their contributions to the estimated value will be rather negligible.

5.3 Conclusions

Through extensive simulations in this section we investigated how diverse factors affect the performance of the prediction algorithm proposed by Gallardo *et al.* in [32]. A summary of the main findings follows.

We evaluated the robustness of the prediction method under parameter estimation errors. Our results show that for different levels of variability (i.e. different values of α) and long-range correlations (determined by the Hurst parameter H), likely discrepancies between their actual and extracted values do not prevent the prediction algorithm from working properly. It was found that its performance is more influenced by the estimated variability of the process. The case when α and $1/H$ have close values deserves special attention since significant errors can be avoided by using a suitable model from the family of FSN processes.

We also found that under certain circumstances, the prediction coefficients obtained from a direct implementation of the method proposed in [32] do not produce an optimum performance. However, they can be optimized with a numerical method if more accuracy is needed. When incorporating an additional numerical procedure for improving the accuracy of the linear prediction, the amount of gain in performance depends on both, the Hurst parameter H and the index of stability α. Significant improvements in the dispersion of the error are achieved for high values of the Hurst parameter and middle values of α (within the valid range).

In the last set of experiments we evaluated the effect of the order of the predictor on the predicted values. As our intuition dictated, it was observed that the dispersion of the error is strongly dependent on the values of recent past samples of the process and in a minor scale on samples from the far past. In addition, the slowly decaying rate of the dispersion with respect to increases in the order of large predictors contrasts with the fast decay observed with small orders. This observation led us to conclude that, taking into consideration the additional computational cost, it is not advisable to use a relatively large order in an attempt to capture long-range correlations. Consequently, we conclude that long-range correlations are of marginal importance in linear prediction. We emphasize the consistency of this result with previous studies by Grossglauser [38] that show that there is a "correlation horizon", specific to the system under study, after which the impact of correlations becomes negligible for all practical purposes. Furthermore, some other researchers claim that only short-range correlations are important for traffic prediction [74].

We believe that with the elements herein provided we can better determine the applicability of the prediction technique to particular applications.

Chapter 6

Dynamic resource allocation

In a scenario where different connections share a link, the network must provide enough resources to each of them and simultaneously satisfy the expected quality of service. To this end, a constant bandwidth allocation policy, i.e. to provide a constant-but-enough amount of resources to each stream, can be easily deployed in network technologies supporting quality of service guarantees. However, this approach is not the most efficient in terms of multiplexing gain. Alternatively, dynamic allocation of resources can be implemented so that unused bandwidth can be redistributed among those connections that require it. When low demanding streams become more active, a redistribution of resources must take place so that they can use their share of the total available resources.

A substantial amount of research has been done in this area yielding relevant contributions worth mentioning. In [58] the authors observed that the performance of a finite buffer is mainly determined by the low frequency component of the input process. Using this result, in [14] it was proposed to extract this band from a compressed video stream, track its bandwidth requirements and dynamically change the bandwidth allocation as to ensure proper resource provisioning. In [4] the authors introduced, categorized, and evaluated several dynamic bandwidth allocation policies. Although diverse conclusions can be drawn from this work, it is worth calling up their observation stating that a simple dynamic bandwidth allocation scheme that takes into consideration the heuristic "the future is similar to the near past" yields fair results. In [2] Adas observed that when a dynamic bandwidth allocation scheme is implemented using traffic prediction, only the prediction errors leading to underestimation, need to be buffered. This results in small queues since the error process, in his tests, resembled white noise or at most a short-memory process. Therefore, smaller delays with higher network utilization can be achieved.

Recently, dynamic bandwidth allocation has been the object of extensive research in the context of video transmission (e.g. [5], [66]) as well as in diverse areas such as satellite networks [43], burst switching [69], Differentiated Services [82], IP over Synchronous Optical Networks (SONET) [94] and others.

In spite of the interest in this approach for bandwidth allocation, problems resulting from signaling overhead and system latency have limited its deployment. The selection of an appropriate reallocation frequency leads to a trade-off with the network utilization. The more frequent the reallocation is, the higher the network utilization becomes, but the higher the signaling overhead. This reallocation frequency is primarily limited by processing time and system latency (i.e., the time the system takes to redistribute the bandwidth). In other words, periodic reallocation is only feasible if the sum of the processing time plus the system latency is smaller than the desired time between successive bandwidth adjustments. Additionally, the particularities of a network technology also limit the frequency of changes. Nevertheless, several strategies with varying degrees of effectiveness can be proposed to reduce the changes. For instance, some bandwidth adjustments can be avoided if the difference between successive values is negligible [4] (other techniques are explained in [2] and [14]).

In the first part of this section it is shown that dynamic allocation schemes supplemented with prediction techniques can potentially lead to the development of efficient algorithms with reasonable overhead. It is worth emphasizing that the prediction procedure under consideration assumes modeling of aggregate network traffic using fractional stable noise. Up to the author's best knowledge, this particular approach has not been explored before. In the second part of this section we propose a dynamic bandwidth allocation algorithm based on the concept of envelope processes. This approach let us reduce the number of renegotiations by properly determining the amount of bandwidth to be allocated over a window of arbitrary length. These schemes are described and experimentally evaluated.

6.1 Dynamic bandwidth allocation based on linear prediction

In this approach the amount of resources to be individually allocated in future time slots is estimated based on the predicted bandwidth demand of the traffic flows. It is assumed that the statistical behavior of aggregate traffic streams is described through α -stable long-range dependent stochastic processes. Accordingly, each traffic process has to be modeled using an underlying fractional stable noise process $\{Y_j\}$. As previously mentioned, the following model has been proposed [33] for aggregate video traffic:

$$W_j^* \stackrel{d}{=} m + aY_j \;\; ; \;\; j = 0, 1, 2, \ldots \tag{89}$$

where W_j^* represents the actual number of bytes arriving during the j-th time interval of unit length. The fractional stable noise process $\{Y_j\}$ is symmetric with index of stability α and Hurst parameter H. The parameter m is the mean of the number of bytes per unit time and

a is a scaling factor. A slightly modified version of the previous model was proposed in [33] for Ethernet and WWW traffic:

$$W_j^* \overset{d}{=} m + a(|Y_j| - \mu_Y) \; ; \; j = 0, 1, 2, \ldots \tag{90}$$

where μ_Y is the mean value of $|Y_j|$. The model components m, a and Y_j have the same meaning as in Equation (89). The values for all the model parameters have to be estimated off-line from actual traffic samples.

Once the traffic modeling has been done, the prediction coefficients $\{a_i\}$ have to be computed for each traffic class as previously explained in Chapter 5. As indicated by Equation (79), these coefficients can be used to estimate future values of the process Y_j from the observations $\{Y_j^*\}$, which are extracted from actual past samples of the traffic process $\{W_j^*\}$. An additional parameter $L\,(L \geq 0)$ can be used to represent a time lag that due to a number of factors (e.g., processing time or signaling) would prevent an immediate allocation of the proper amount of bandwidth as follows:

$$\hat{Y}_j = \hat{Y}_{k+L} = \sum_{i=1}^{N} a_i \cdot Y_{k-i}^* \tag{91}$$

where $\hat{Y}_j$ is the estimated value of the j-th sample of the underlying FSN process. Finally, future bandwidth requirements can be forecasted as:

$$\hat{W}_j = m(1+p) + a\hat{Y}_j \tag{92}$$

or

$$\hat{W}_j = m(1 + p) + a(|\hat{Y}_j| - \mu_Y) \tag{93}$$

in the case of Ethernet and WWW traffic. The parameter p controls the QoS (in terms of loss probability) by over-allocating additional bandwidth. For instance, in the case of aggregate video traffic it can be observed from (89), (92) and (64) that:

$$P\left\{W_j^* > \hat{W}_j\right\} = P\left\{\frac{a}{m}(Y_j - \hat{Y}_j) > p\right\} = \varepsilon . \tag{94}$$

The effect of the over-allocation can be observed through the percentage of downgraded data (i.e., data that either, could not be transmitted or were given a lower priority). This percentage can be computed as:

$$\%downgraded\_data = \left(1 - \frac{\sum_j min(W_j^*, \hat{W}_j)}{\sum_j W_j^*}\right) \times 100 \cdot \tag{95}$$

Although the percentage of downgraded data can be reduced with additional bandwidth, the over-allocation required to cope with the large excursions caused by the burstiness of the traffic process may lead to a significant waste of resources. Therefore, the proper operating point will be a trade-off between the desired values of QoS and efficiency in the use of resources. We can quantify efficiency as follows:

$$\eta\% = \frac{\sum_j min\left(\hat{W}_j, W_j^*\right)}{\sum_j \hat{W}_j} \times 100 \tag{96}$$

A periodic bandwidth allocation scheme, like the one previously described, redistributes the total available capacity among the set of active traffic streams at regular intervals.

6.1.1 Experimental evaluation with unconstrained resource

In this section, through computer simulations, we evaluate the performance of the dynamic prediction-based bandwidth allocation scheme. For our experiments we considered two types of network traffic: aggregate MPEG-1 video and moderately bursty Ethernet LAN traffic. Although the two traffic streams share the same link, they are not restricted in the amount of bandwidth they can take. This approach was followed in order to observe the effect of some parameters on the performance of the system without introducing additional randomness generated by constrained resource and contenting requests of users, which could result in refusals of additional resource provision because it has been reserved by other connections (this issue will be addressed in the following section). Considering the suggested models and parameter extraction procedures described in [33], the video traffic class was modeled using (89) with the following parameters: m=3,655,758 [bps], a=295,431 and a standard well-balanced Linear FSN with $\alpha = 1.95$ and $H = 0.90$. Similarly, Ethernet traffic was modeled using (90) with parameters m=1,105,424 [bps], a=839,624, μ_Y=1.1575 [bps] and standard well-balanced FSN process with $\alpha = 1.95$ and $H = 0.83$. With this information, the prediction coefficients were computed as explained in Chapter 5 considering a predictor of order 50 in both cases. Since for both types of traffic $\alpha \sim 2$, there was no need to apply an additional optimization procedure. It is worth mentioning that although both traces have the same value for their index of stability α, the Ethernet trace showed a higher peak-to-mean ratio (which is captured by the other model parameters). Each experiment was performed eleven times using Equations (95) and (96) to obtain the metrics of performance.

In our first set of simulations we investigated the amount of improvement in performance we expect to obtain when a dynamic prediction-based resource allocation technique is used instead of a constant bandwidth allocation. The following values for the over-allocation parameter p (see Equations (92) and (93)) were used: 0, 0.01, 0.03, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4 and 0.5, resulting in increased requirements of average bandwidth per time slot. The corresponding simulation results (sample mean values) are shown in Figures 6.1 to 6.4. In the second part of the experiment, the previously obtained values of average bandwidth were allocated to the traffic traces during the whole simulation time in order to observe the performance of a constant allocation policy. These results are also included in the figures. It is observed that for most cases the dynamic allocation outperforms the static scheme. However, an efficiency comparison between the two techniques shows that for the video traffic class both techniques show nearly comparable performance. This is mainly due to the low burstiness of this traffic trace (small peak-to-mean ratio). However, the difference becomes noticeable at higher levels of burstiness as evidenced by the values of efficiency obtained with the Ethernet trace. It is clear that in order to obtain a comparable performance with the dynamic approach, the static technique must allocate a large amount of bandwidth, in order to be able to cope with highly varying traffic.

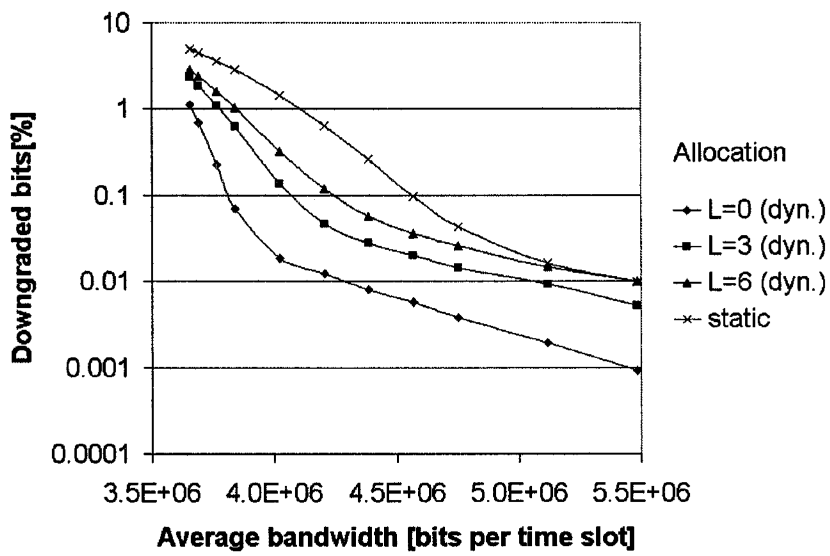


Figure 6.1 Percentage of downgraded bits for the video class under static and dynamic bandwidth allocation policies.

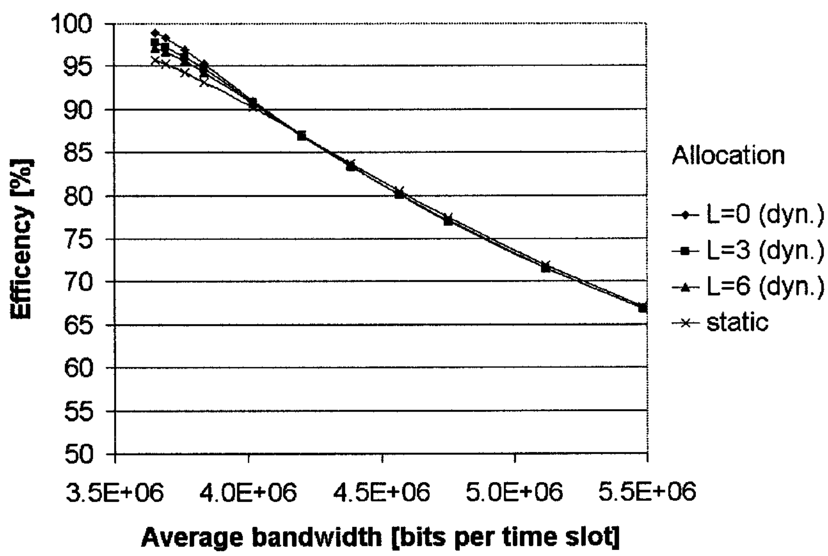


Figure 6.2 Efficiency in the use of allocated capacity for the video class under static and dynamic bandwidth allocation policies.

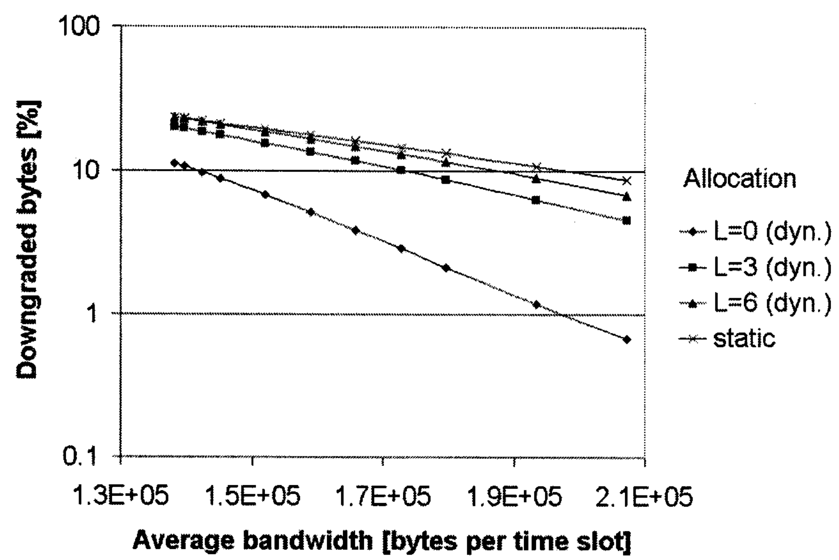


Figure 6.3 Percentage of downgraded packets for Ethernet traffic under static and dynamic bandwidth allocation policies.

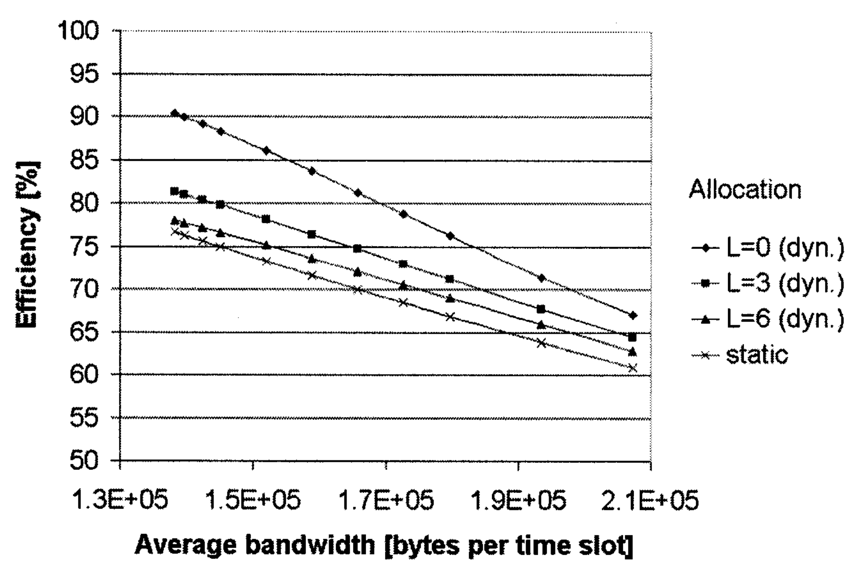


Figure 6.4 Efficiency in the use of allocated capacity for Ethernet traffic under static and dynamic bandwidth allocation policies.

Furthermore, Figures 6.1 and 6.3 show the benefits of using a prediction-based dynamic bandwidth allocation scheme. Note that the y-axis in these two figures use a logarithmic scale. For instance, from Figure 6.1 it is clear that when L = 0 and for a percentage of downgraded data of 0.01, there is a significant reduction, around 20%, on the amount of bandwidth required with respect to a static bandwidth allocation. From the figure, it is also clear that as L increases, the bandwidth requirements increase for a given percentage of downgraded data.

In our second set of simulations we explore the effect of the prediction-related parameters on the system performance. We ran several tests under varying values of the parameters L and p (see Equations (91), (92) and (93)). The simulation results (sample mean values) for video traffic are shown in Figures 6.5 and 6.6 whereas the ones corresponding to Ethernet traffic are shown in Figures 6.7 and 6.8. These results clearly exemplify the trade-off between the quality of service (in terms of percentage of downgraded data) and efficiency. For instance, for video traffic, it is observed that when using a prediction lag of 9, the percentage of downgraded data is reduced from 1 to nearly 0 (the actual values are: 0.9391 ± 0.0284 and 0.0073 ± 0.0061 using 95% confidence intervals) when the margin of over-allocation is increased from 5% to 25% of the mean value of the traffic process. However, this comes not without a cost. There is a drop in efficiency from 95% to 80% (the actual values are: $(94.3118 \pm 0.1989)\%$ and $(79.9391 \pm 0.3972)\%$ using 95% confidence intervals).

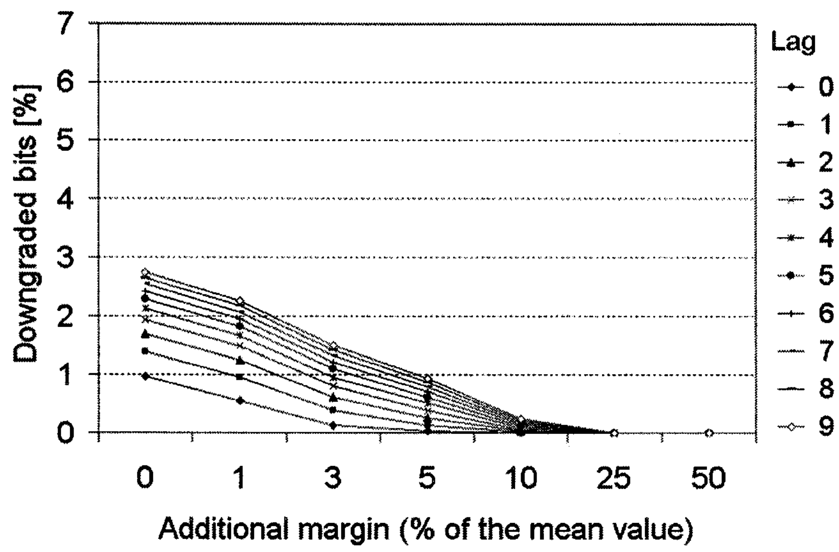


Figure 6.5 Percentage of downgraded bits for the video class under different time lags.

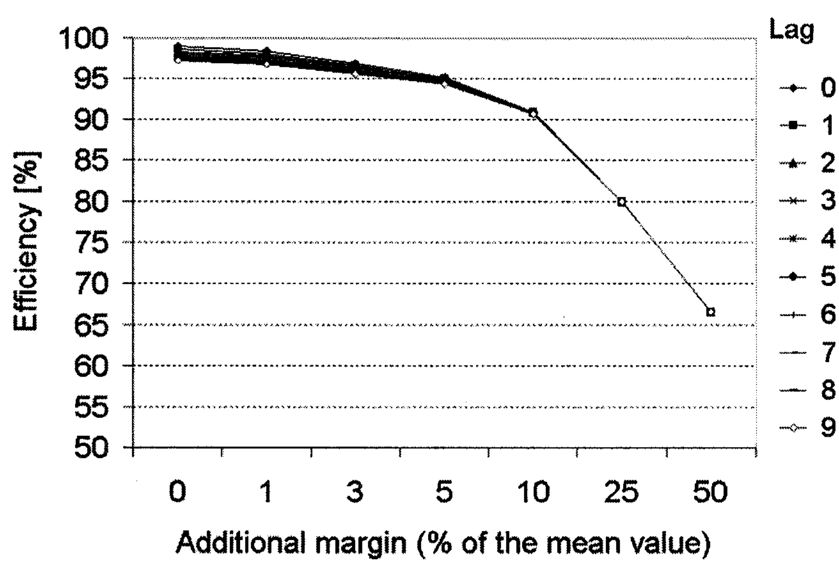


Figure 6.6 Efficiency in the use of allocated capacity for the video class under different time lags.

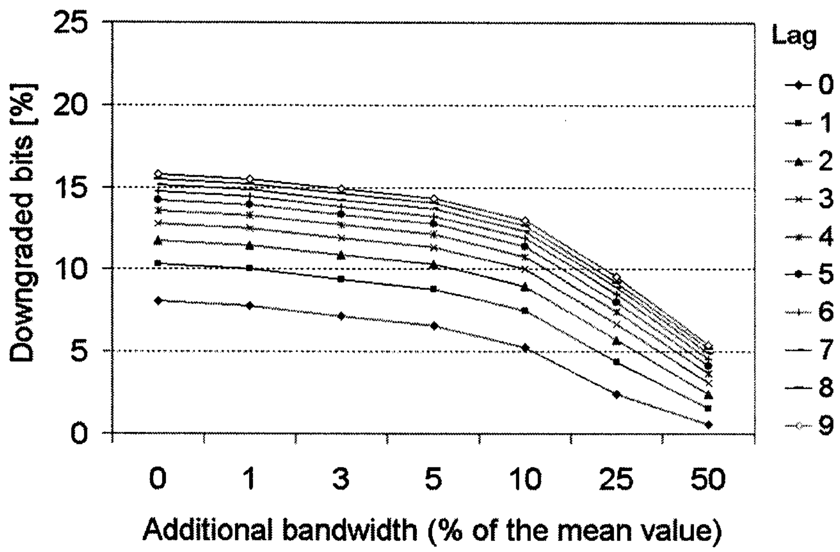


Figure 6.7 Percentage of downgraded packets for Ethernet traffic under different time lags.

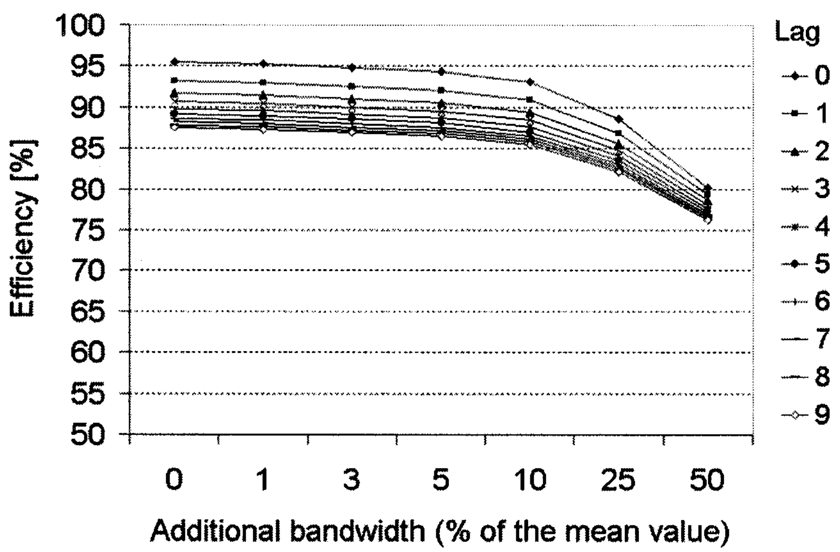


Figure 6.8 Efficiency in the use of allocated capacity for Ethernet traffic under different time lags.

6.1.2 Experimental evaluation with constrained resource

Let us now assume that the traffic streams considered in the previous section share a link of limited bandwidth C. This resource is periodically distributed using linear prediction to estimate the bandwidth requirements of individual traffic classes. If enough resources are available in order to satisfy the total bandwidth requirement, the allocation will take place as previously explained. However, this strategy has to be complemented with a way to cope with the case when the estimated bandwidth demand exceeds the link capacity. It is proposed to define in advance a set of weights $\{w_i; i=1,\ldots,N\}$ for the N contending traffic classes and use it to split the link capacity in these cases. This set satisfies the following property:

$$\sum_{i=1}^{N} w_i = 1 \tag{97}$$

This way, under heavy traffic situations each traffic class is guaranteed a share $w_i C$ of the available bandwidth.

Under light-load conditions the performance of this bandwidth allocation scheme is similar to the one observed in the previous section. This is due to the fact that in these cases there is little or null interaction among different traffic streams and the bandwidth allocation is primarily done using prediction. However, with heavier loads the situation changes and the distribution of resources relies more on the set of weights. This is illustrated in the following case study where we consider again the two traffic classes described in the previous section, i.e. aggregate video and Ethernet traffic. Assume that these traffic aggregates share a link of 7Mbps and are assigned the following weights: $w_1 = w_2 = 0.5$. Let us consider the same parameter values used in the previous section (i.e., prediction lag in the range 0 to 9 and percentages of over-allocation 0, 1, 3, 5, 10, 25 and 50) in order to be able to directly compare the impact of the contention handling scheme on the system performance. Although the parameters for the bandwidth allocation algorithm for each traffic class are independent of each other, the tests were run under common conditions of prediction lag and over-allocation. Each experiment was performed eleven times obtaining the sample mean values described below. Figure 6.9 shows the values of link utilization achieved for the possible combinations of the parameters. It is observed that larger prediction lags result on lower figures for the link utilization due to the diminished accuracy of the predicted values. Over-allocation may also increase link utilization up to a point where there is no more data available to make use of the resource or to a point where the estimated bandwidth requirements exceed the available capacity most of the time precluding the use of bandwidth allocation based on prediction in favor of the allocation based on the set of weights. This is the situation shown in Figure 6.9 for an over-allocation of 50% of the mean value. The simulation results for the video class shown in Figures 6.10 and 6.11 clearly let us discern between the region where the curves are mainly influenced by the prediction-based allocation and the one where they are primarily affected by the bandwidth allocation based on the set of weights. Finally, Figures 6.12 and 6.13 show the simulation results corresponding to Ethernet traffic which in our simulations had a lower mean value than video traffic. It is observed that, for large values of over-allocation, the performance of the system is independent of the prediction-related parameters. It is worth mentioning that, in this region, the policy for bandwidth allocation is mainly static and therefore, the percentage of losses can be estimated using envelope processes as previously explained.

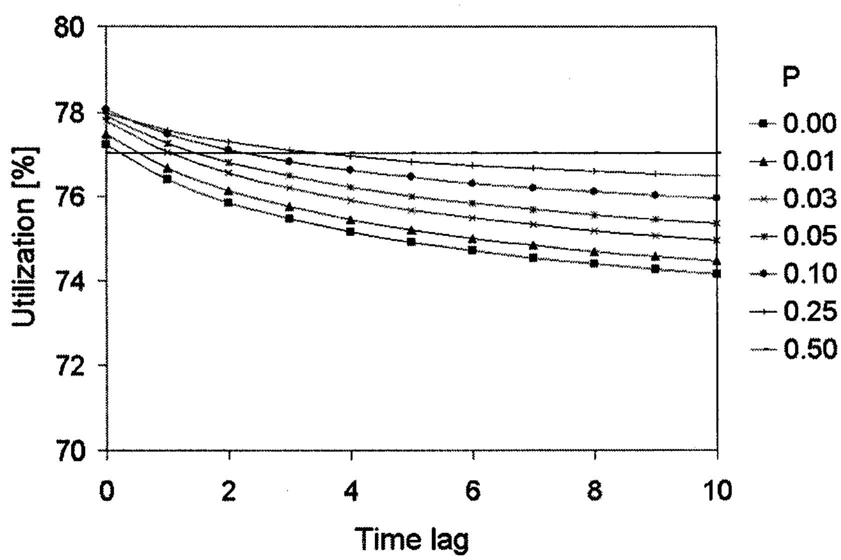


Figure 6.9 Link utilization achieved in a link of 7Mbps with video and Ethernet traffic

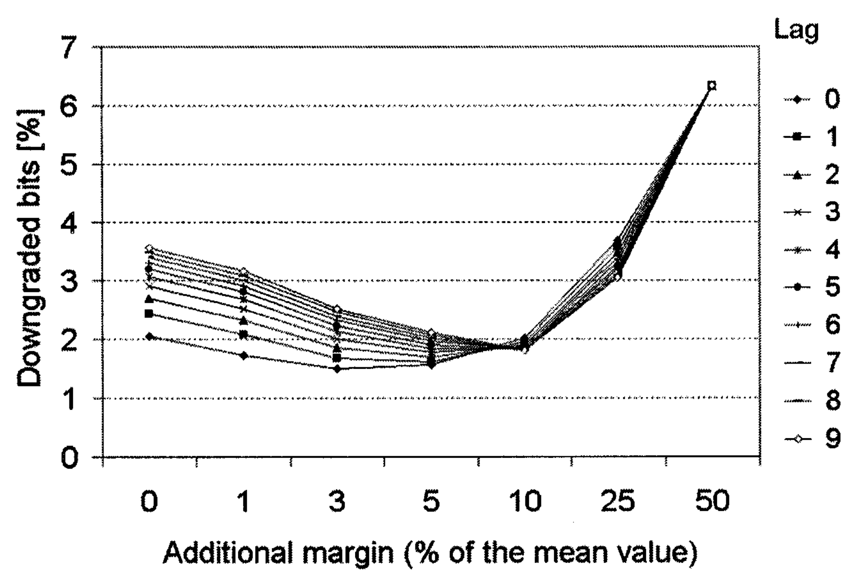


Figure 6.10 Percentage of downgraded bits for video traffic in a shared link of 7Mbps.

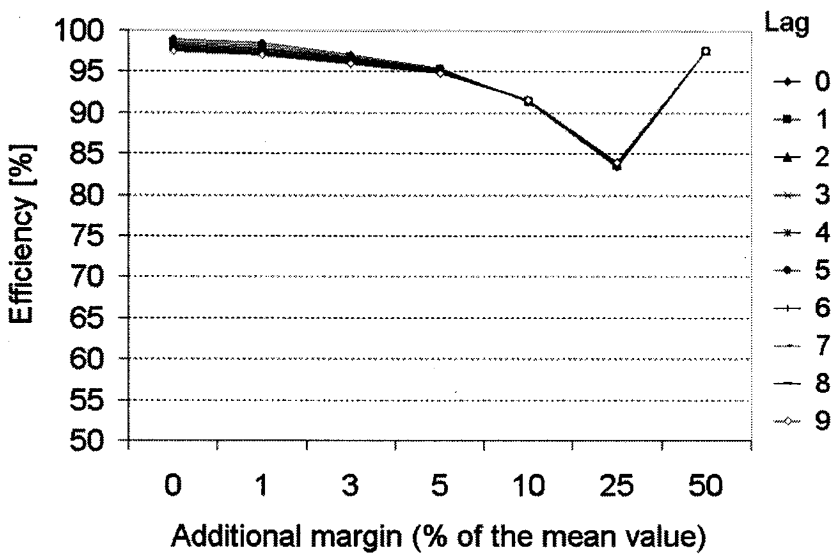


Figure 6.11 Efficiency in the use of allocated bandwidth for video traffic in a shared link of 7Mbps.

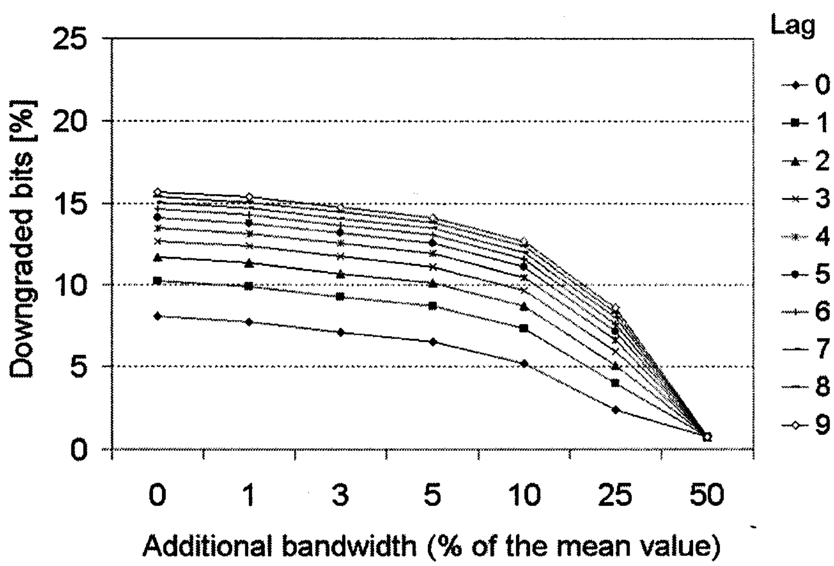


Figure 6.12 Percentage of downgraded data for Ethernet traffic in a shared link of 7Mbps.

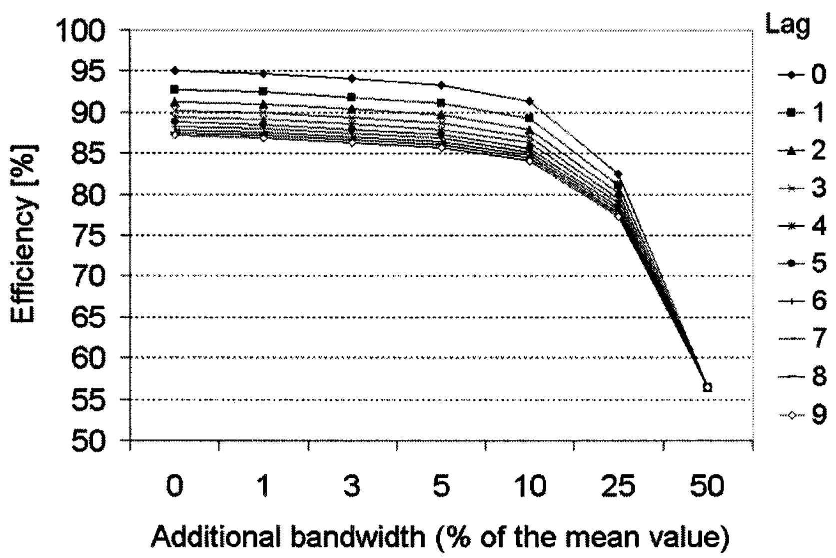


Figure 6.13 Efficiency in the use of allocated bandwidth for Ethernet traffic in a shared link of 7Mbps.

6.2 Dynamic bandwidth allocation based on envelope processes

Let us recall that, in previous sections, probabilistic envelopes for α-stable processes have been used to determine the amount of long-term bandwidth allocation required by a traffic stream. The parameters of the envelope are extracted from samples of the process and they are assumed to be valid during the actual data transmission. These procedures (i.e., parameter extraction and bandwidth allocation) can be periodically executed in order to exploit knowledge of local statistics of a traffic process and achieve more efficient use of the network resource. This idea can be used to propose an envelope-based dynamic bandwidth allocation algorithm as described below.

Let us again represent with $\{W_j^*, j = 0,1,\ldots\}$ the actual number of arrivals of a traffic process over time intervals of unitary length. Also let (36) be the model for this process (i.e., $W_j^* \overset{d}{=} m + aY_j$, where m is the mean number of arrivals per time slot, a is a scaling factor and $\{Y_j\}$ is fractional stable noise). Recall that, for this model, the proposed envelope is defined as $\hat{W} = m + K\sigma_W$ (see Equation (62)), where m has the same meaning as in the traffic model, σ_W is the scale parameter of the traffic process and K is a parameter controlling the magnitude of the data loss.

In the proposed algorithm, future bandwidth requirements are estimated based on local envelope processes as follows. Traffic samples are collected during a measurement time (T_{meas}) and processed (extraction of the parameters of the envelope process) in order to determine the proper bandwidth allocation that satisfies a loss constraint. The service rate is updated with this value and remains effective during the following T_{alloc} time units. This algorithm is executed in a cyclic way by collecting the samples over a sliding window of size T_{meas}, which advances T_{alloc} time units each time to start a new cycle. Let us consider that $T_{alloc} \le T_{meas}$, so that successive cycles can overlap. We consider the processing time to be negligible in comparison to T_{alloc} and T_{meas}.

Experimental evaluation

The proposed dynamic bandwidth allocation algorithm will be compared to a more conventional approach that assumes a Gaussian distribution in the traffic process. It operates in the same way as our proposed algorithm, but determines the bandwidth allocations using the following local Gaussian envelope process (see [15] and [24]):

$$\hat{W} = m + K_G s \tag{98}$$

where m is again the mean number of arrivals per time unit, s is its standard deviation and K_G is selected as to guarantee that the traffic will not exceed the envelope with probability $1-\varepsilon$, i.e. $P\{Y > K_G\} = \varepsilon$ where $Y \sim N(0,1)$. The parameters m and s are estimated using the sample mean and sample standard deviation, respectively. It is worth mentioning that, being the Gaussian distribution a particular case of the family of the α-stable distributions, the envelope (98) can be transformed into an equivalent envelope process parameterized in terms of the mean and scale parameter[17].

The metrics of performance to be considered in the evaluation include the percentage of downgraded data and efficiency (see Equations (95) and (96)). The expected value of the magnitude of downgraded data has been analyzed in Appendix D based on the assumption that traffic characterization procedures and a long-term bandwidth allocation have been carried out before the actual data transmission. In contrast, the algorithms considered in this section use local parameter estimation and short-term allocations and therefore, some differences in performance are expected. Nevertheless, the formulae derived in Appendix D can be used to establish a point of reference.

In addition to the metrics of performance mentioned above, let us consider the ratio of the average allocation to the mean value of the input process. Let us denote this metric as r_{over} and compute it as follows:

$$r_{over} = \frac{\sum_j \hat{W}_j}{\sum_j W_j^*} \tag{99}$$

---

[17] This transformation can be carried out taking into consideration that, for a Gaussian random variable, the scale parameter σ and standard deviation s are related by the expression $\sqrt{2}\sigma = s$ [81 p. 318]. Using this condition, it is trivial to prove that the Gaussian envelope as defined by (98) and the envelope $\hat{W} = m + K\sigma_W$ with $P\{Y > K\} = \varepsilon$ and $Y \sim S_2(1,\cdot,0)$ are equivalent if $K = \sqrt{2}K_G$. The estimation methods for the standard deviation and the scale parameter are not the same, though.

The simulations will be carried out with video traffic generated from the model $W_j = m + aY_j$ using the following parameters: m=3,655,758 [bps], a=295,431 and standard well-balanced LFSN with $\alpha = 1.95$ and $H = 0.90$ [33]. We will consider the following values for the probability ε that the process exceeds its envelope: 0.1, 0.01 and 0.001. These probabilities uniquely determine K in our proposed algorithm and the value of K_G in the Gaussian envelope. The required values of the parameter K were obtained from the inversion of the characteristic function of the random variable $S_{1.95}(1,0,0)$. The parameter K_G was obtained directly from tables (e.g., see Table 3.4 in [57]). These values are shown in Tables 6.1 and 6.2 along with the expected magnitude of downgraded data (losses) for each case (computed using Equation (D.7) from Appendix D). In addition to the values of the parameter K_G of the Gaussian envelope, Table 6.2 also includes, in parentheses, the values of the parameter K ($K = \sqrt{2}K_G$) that would be used in an equivalent envelope process using α-stable parameters. The parameters m, σ and s are estimated in each cycle of the algorithm, whereas the value of the parameters K and K_G was kept constant for a particular value of ε.

From the values shown on Tables 6.1 and 6.2 it can be inferred that for $\varepsilon = 0.1$ both algorithms will produce similar performance (both have similar values for K, which indicates that they will allocate a similar amount of bandwidth). As shown on the tables, in order to achieve smaller values in the probability ε, the parameter K has to be increased. However, due to the different decay rate of the tail of the α-stable and Gaussian distributions the increase rate for K is slower for the Gaussian envelope. Therefore, it is expected that for $\varepsilon = 0.01$ and $\varepsilon = 0.001$ the difference in performance will be more noticeable. Furthermore, from the values shown on the tables, it can be observed that the levels of the expected downgraded data may be significantly different. The Gaussian assumption predicts a faster decay of the amount of downgraded data than the α-stable model. In order to explore this issue a bit further, Equation (D.7) was sampled with additional values of K obtaining the curves shown in Figure 6.14. As depicted on the figure, even though the values of α seem to be relatively close values, they may lead to substantially different assumptions on the levels of downgraded data occurring in the system. This is due to the fact that the model with α=1.95 has a heavy tail and the other does not.

Table 6.1 Values for the parameter K corresponding to different values of the probability ε ($P\{Y > K\} = \varepsilon$ where $Y \sim S_{1.95}(1,0,0)$) and estimated fraction of downgraded data.

	ε =0.1	ε=0.01	ε =0.001
K	1.82	3.45	5.83
Downgraded_data [%]	0.6568	0.1049	0.0405

Table 6.2 Values for the parameter K_G corresponding to different values of the probability ε ($P\{Y > K_G\} = \varepsilon$ where $Y \sim N(0,1)$) and estimated fraction of downgraded data.

	ε =0.1	ε=0.01	ε =0.001
K_G	1.2815 (K=1.8123)	2.3263 (K=3.2899)	3.0902 (K=4.3702)
Downgraded_data [%]	0.5536	0.0393	0.0032

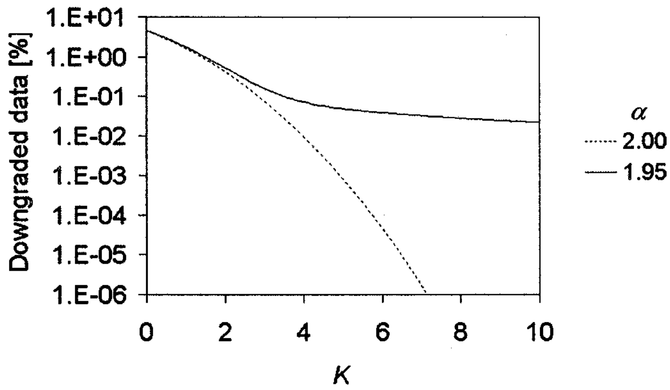


Figure 6.14. Fraction of downgraded data as a function of the parameter K of the envelope processes.

In the first set of simulations we let the measurement window (T_{meas}) take several values from 1 to 5 minutes whereas the allocation time (T_{alloc}) is set to a fixed value of 1 minute. The corresponding results (sample mean values) are shown in Figures 6.15 to 6.17.

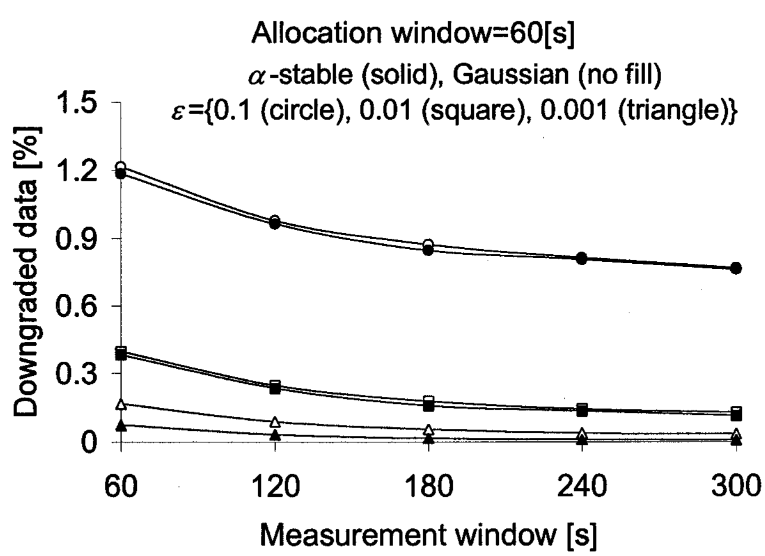


Figure 6.15 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of downgraded data (tests with video traffic and different sizes for the measurement window).

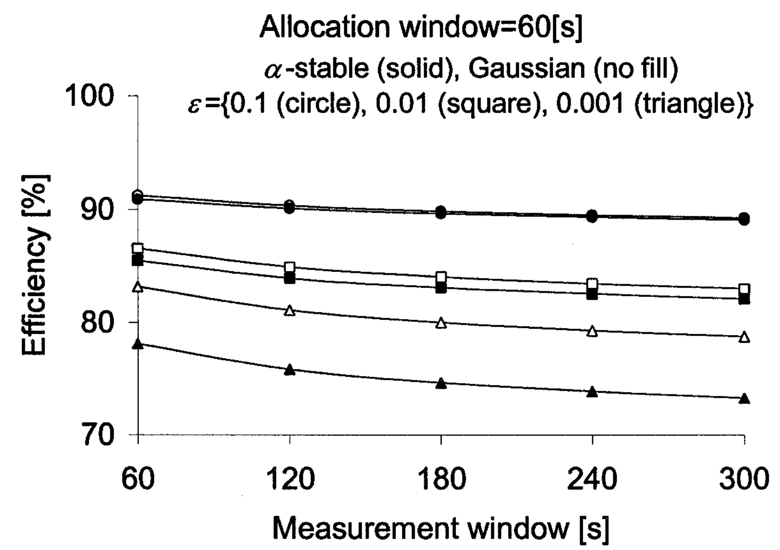


Figure 6.16 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of efficiency (tests with video traffic and different sizes for the measurement window).

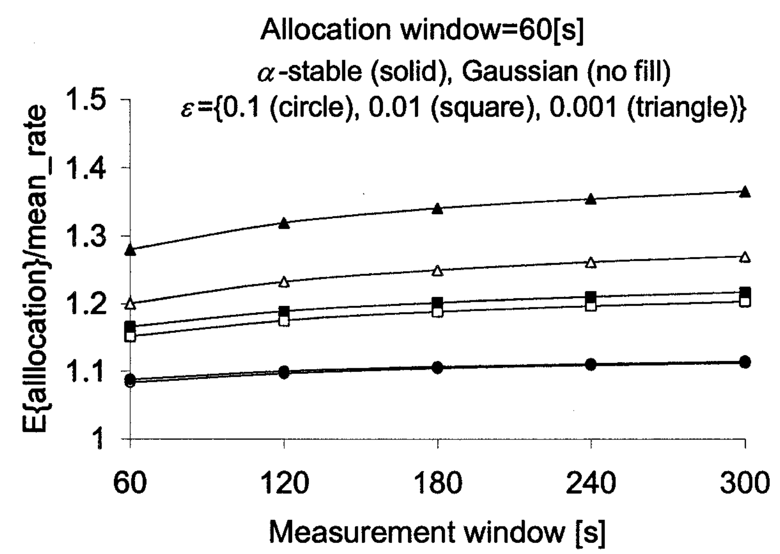


Figure 6.17 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of the ratio of the average allocation to the mean input rate (tests with video traffic and different sizes for the measurement window).

It is observed that as the measurement window increases, the percentages of downgraded data achieved by our proposed algorithm asymptotically decrease to the expected values shown in the second row of Table 6.1. This is due to the fact that larger estimation windows allow us to obtain more accurate estimates for the parameters of the envelope. A similar trend is observed in the simulations with the Gaussian envelope, but the expected percentage of downgraded data shown in the second row of Table 6.2 underestimated the magnitude of the actual losses. This is more noticeable at more stringent conditions of data loss. For instance, for ε=0.001 the expected percentage of data loss (see Table 6.2) with the Gaussian envelope is 0.0032, but the experimentally obtained value was higher, (0.0351 $\pm$0.0109)% using 95% confidence intervals. In contrast, with the envelope based on the α-stable model, the expected value (see Table 6.1) was 0.0405% and the experimental result was (0.0088 $\pm$0.0041)%, again with 95% confidence intervals.

In the second set of experiments we set the measurement window to 5 minutes and let the allocation window vary from 1 to 5 minutes. The simulation results are shown in Figs. 6.18 to 6.20.

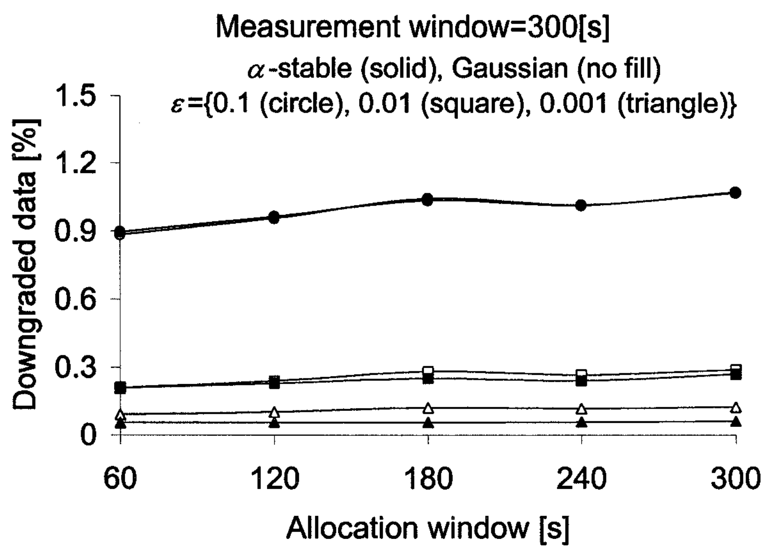


Figure 6.18 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of downgraded data (tests with video traffic and different sizes for the allocation window).

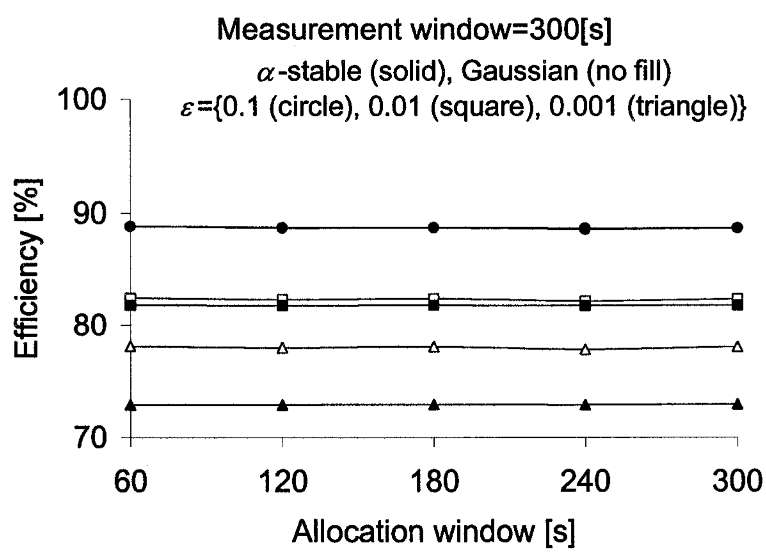


Figure 6.19 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of efficiency (tests with video traffic and different sizes for the allocation window).

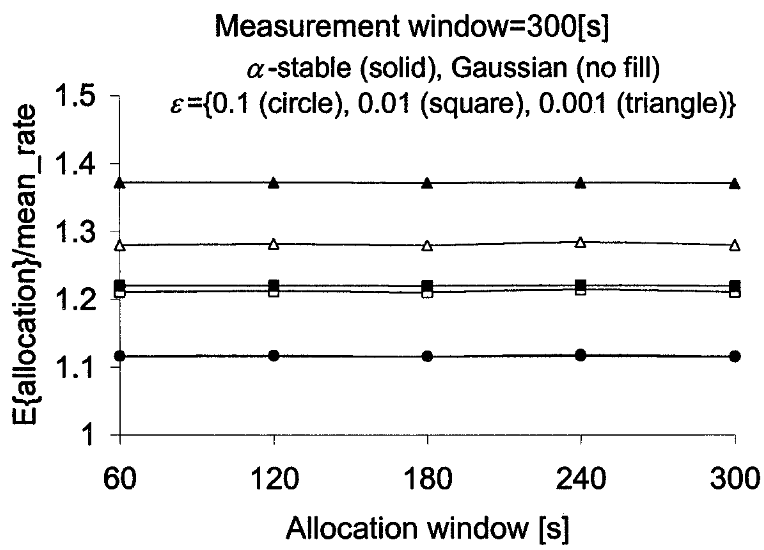


Figure 6.20 Performance of the envelope-based dynamic bandwidth allocation algorithms in terms of the ratio of the average allocation to the mean input rate (tests with video traffic and different sizes for the allocation window).

Recall that the size of the allocation window is directly related to the reallocation frequency, the larger its size, the less frequent the reallocations occur. Figure 6.18 shows that there is an increasing trend in downgraded data for larger sizes of the allocation window. This is due to the diminished responsiveness of the algorithms since the mean bandwidth allocation remains nearly constant (see Figs. 6.19 and 6.20). The performance of the algorithms in terms of percentage of downgraded data was again more closely predicted using our proposed envelope than the Gaussian one.

From these tests it is concluded that the dynamic bandwidth allocation algorithm based on local α-stable envelope process results in a more reliable way to provide stringent quality of service guarantees when the input traffic exhibits heavy tails. Further refinements in the algorithm are possible, such as on-line estimation of the index of stability, and are left for future research.

6.3 Conclusions

We have argued that a static allocation of network resources can be highly inefficient. Based on this premise, in this section, we have explored dynamic bandwidth allocation schemes. Our approach is based on the modeling of network traffic using FSN processes and appropriate estimation methods for future bandwidth requirements (suitable to the statistical properties of the samples of FSN processes). It was shown that dynamic prediction-based bandwidth allocation algorithms can significantly improve network utilization over static resource allocation schemes with minimum overhead since even low order linear predictors can achieve acceptable performance. We also explored dynamic bandwidth allocation algorithms based on the concept of probabilistic envelope processes, used in previous sections of this work for static allocations. The basic idea is to carry out short-term static allocations, as opposed to a single long-term constant bandwidth allocation, in order to increase the efficiency on network usage. Our proposed algorithm can arbitrarily reduce the frequency of renegotiations in trade for some over-allocation. Use of envelope processes for α-stable processes resulted to be appropriate to provide stringent quality of service guarantees to network traffic exhibiting heavy tails.

Chapter 7

Thesis conclusion and suggestions for future research

Increasing awareness of the relevance of long-range dependence and high variability properties of network traffic will certainly have an impact on the future network infrastructure. For instance, in this work it was shown (using our proposed envelope processes) that due to the large deviations happening in the traffic process, the amount of network resource required, in order to achieve a small percentage of losses, is significantly larger than the one expected using more traditional models that do not take into account the high variability encountered in some traffic classes. One can be tempted to predict that if quality of service guarantees are to be provided, future networks will consist of over-provisioned links which will remain lightly loaded most of the time. On the other hand, it was also shown that the heavy tails in the traffic process negatively affect the multiplexing gain, which in turn directly impacts the scale economies expected by service providers. As pointed out by the authors in [17], this interaction between customer satisfaction and cost factors becomes both a business and a technological problem. Whereas the effect of this interaction on the future network infrastructure may lead to merely speculative analysis at this point in time, we can be certain that with the limited resources provided by the current infrastructure an alternate way to satisfy increased quality of service expectations is to implement more efficient resource allocation mechanisms. In this work this issue was addressed by using dynamic bandwidth allocation based on linear prediction of symmetric α-stable stochastic processes. With this approach, it was shown how one can substantially improve the network utilization in trade for additional complexity and signaling overhead.

It can be concluded that although resource allocation schemes can be developed based on different policies (constant, dynamic or a combination of both) and the balancing of diverse factors (simplicity, cost, efficiency, customer satisfaction and others), proper use of traffic models based on α-stable long-range dependent stochastic processes let us introduce realistic assumptions during their analysis and design.

The series of issues arising during the development of this work motivate a great deal of research in diverse-but-interrelated directions. We observed, for instance, that in the past few years the importance of short-range correlations has been largely neglected in traffic modeling. Although it is true that short-range correlations do not have an important effect on some metrics of performance evaluation, in this work it has been concluded that they can play an important roll in the performance of linear prediction. This result suggests that use of models that simultaneously capture short and long-range correlations is highly desirable in the evaluation of resource allocation schemes based on prediction.

In addition to the proper modeling of correlations, accurate modeling of network traffic is complemented by properly capturing the statistical properties of the marginal distributions. As previously mentioned, existing models based on self-similar stochastic processes have tried to address this issue by using α-stable random variables in order to capture the heavy-tail decay of the marginal distributions. However, they have used exclusively either symmetric distributions [33] or totally skewed to the right random variables [46]. In this work it has been suggested to consider the degree of skewness in the process as an additional parameter, worth of being taken into consideration. This approach allows for more accurate traffic modeling, but it is dampened by the scarce availability of related analysis using skewed distributions. It seems that we are just beginning to explore the rich parameter space that α-stable self-similar processes have to offer. Whereas symmetric processes have been favored by practitioners, comparable works using skewed processes are rather scarce. Therefore, methods for synthesis, parameter estimation and prediction of skewed processes have been less exhaustively studied. For instance, most of the available prediction methods (including the one we used in this work) are intended for symmetric processes. Development of comparable prediction methods for skewed processes is left as future work.

It can be finally mentioned that adaptive linear prediction of α-stable processes has not been applied to network traffic, although some algorithms for adaptive linear prediction of symmetric α-stable time series have already been proposed (e.g., [9] and [70 Section 8.5]). Applications using an adaptive approach would have a high level of flexibility, but it would also introduce some new problems such as convergence and adaptation time. This is also left as material for future research.

References

1. P. Abry, P. Flandrin, M. S. Taqqu and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger editors, pp. 39-88. USA: John Willey & Sons, 2000.
2. A. M. Adas, "Using adaptive linear prediction to support real-time VBR video under RCBR network service model," *IEEE/ACM Transactions on Networking*, vol. 6, no.5, pp. 635-644, October 1998.
3. R. G. Addie, M. Zukerman and T. D. Neame, "Broadband traffic modeling: simple solutions to hard problems," *IEEE Communications Magazine*, vol. 36, no. 8, August 1998.
4. Y. Afek, M. Cohen, E. Haalman and Y. Mansour, "Dynamic bandwidth allocation policies," in *1996 IEEE INFOCOM*, pp. 880-887.
5. M. Ashibani, D. Mashao and B. Nleya, "Dynamic bandwidth allocation protocol for hierarchically compressed video in ATM networks," in Proc. of *ICC 2002*, pp. 2247-2252.
6. S. Bates and S. McLaughlin, "The estimation of stable distribution parameters from teletraffic data," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 865-870, March 2000.
7. V. E. Beneš, *General Stochastic Processes in the Theory of Queues*, Addison-Wesley, 1963.

8. J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, "Long range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43 (2/3/4), pp. 1566-1579, 1995.

9. J. S. Bodenschatz and C. L. Nikias, "Symmetric alpha-stable filter theory," *IEEE Transactions on Signal Processing*, vol. 45, no.9, September 1997.

10. R. Boorstyn, A. Burchard, J. Lieberherr and C. Oottamakorn, "Effective envelopes: statistical bounds on multiplexed traffic in packet networks," in Proceedings of *IEEE INFOCOM 2000*, vol. 3, pp. 1223-1232.

11. S. Cambanis, C. D. Hardin Jr. and A. Weron, "Ergodic properties of stationary stable processes," *Stochastic Processes and their Applications*, vol. 24, pp. 1-18, 1987.

12. O. Cappé, E. Moulines, J. C. Pesquet, A. Petropulu and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Processing Magazine*, vol. 19, no.3, pp. 14-27, May 2002.

13. C. S. Chang, "Stability, Queue Length and Delay of Deterministic and Stochastic Networks," *IEEE Transactions on Automatic Control*, vol. 39, pp. 913-931, May 1994.

14. S. Chong, S. Q. Li and J. Ghosh, "Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM," *IEEE Journal in Selected Areas in Communications*, vol. 13, no. 1, pp. 12-23, January 1995.

15. C. N. Chuah, L. Subramanian, R. H. Katz and A. D. Joseph, "QoS provisioning using a clearing house architecture," in *International Workshop on Quality of Service (IWQoS)*, 2000.

16. D. R. Cox, "Long-Range Dependence: a Review," in *Statistics: An Appraisal*, H. A. David and H. T. David, editors. Iowa State University Press, pp. 55-74, 1984.

17. D. Clark and W. Lehr, "Provisioning for bursty Internet traffic: implications for industry and Internet structure," in *Workshop on Internet Service Quality Economics*, 1999.

18. D. B. R. Cline and P. J. Brockwell, "Linear prediction of ARMA processes with infinite variance," *Stochastic Processes and its Applications*, vol. 19, pp. 281-296, 1985.

19. M. Crovella and A. Bestravos, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835-846, 1997.

20. M. E. Crovella and L. Lipsky, "Simulations with heavy–tailed workloads," in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger editors, pp. 89-100. USA: John Wiley & Sons, 2000.

21. R. Davis and S. Resnick, "Limit theory for moving averages of random variables with regularly varying tail probabilities," *The Annals of Probability*, vol. 13, no. 1, pp. 179-195, February 1985.

22. G. Doubrovina, M. Falkner and M. Devetsikiotis, "Optimal cost traffic shaping with self-similar input sources," in Proc. of *GLOBECOM'99*, pp.1616-1622.

23. N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications", *Mathematical Proceedings of the Cambridge Philosophical Society*, no. 118, pp. 363-374, 1995.

24. N. G. Duffield, P. Goyal and A. Greenberg, "A flexible model for resource management in virtual private networks," in *Proceedings of ACM SIGCOMM*, pp. 95-108, 1999.

25. M. Evans, N. Hastings and B. Peacock, *Statistical Distributions*, second edition. New York, USA: John Wiley & Sons, 1993.

26. E. F. Fama and R. Roll, "Parameter estimates for symmetric stable distributions," *Journal of the American Statistical Association*, vol. 66, no. 334, pp. 331-338, June 1971.

27. M. Falkner, M. Devetsikiotis and I. Lambadaris, "Minimum cost traffic shaping," *IEEE Communications Letters*, vol. 3, no. 9, pp. 257-259, September, 1999.

28. R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *Computer Journal*, vol. 7, pp. 149-154, 1964.

29. S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, pp. 392-403, August 2001.

30. N. L. S. Fonseca and G. S. Mayor, "On the equivalent bandwidth of self-similar sources," *ACM Transactions on Modeling and Computer Simulation*, vol. 10, no. 2, pp. 104-124, April 2000.

31. B. Fristedt and L. Gray, *A Modern Approach to Probability Theory*, USA: Birkhäuser, 1997.

32. J. R. Gallardo, D. Makrakis and L. Orozco-Barbosa, "Prediction of alpha-stable long-range dependent stochastic process," *Advances in Performance Analysis,* vol. 3, no. 1, pp. 31-42, 2000.

33. J. R. Gallardo, D. Makrakis and L. Orozco-Barbosa, "Use of alpha-stable self-similar stochastic processes for modeling traffic in broadband networks," *Performance Evaluation Journal*, vol. 40, no. 1-3, pp. 71–98, March 2000. Also in *Proc. of SPIE's Symposium on Voice, Video and Data Communications* within the Performance and Control of Network Systems program, 1998.

34. J. R. Gallardo and D. Makrakis, "Dynamic predictive weighted fair queueing for differentiated services," in *ICC 2001*, vol. 8, pp. 2380-2384.

35. J. R. Gallardo, D. Makrakis and M. Angulo, "Dynamic resource management considering the real behavior of aggregate traffic," *IEEE Transactions on Multimedia*, vol. 3, pp. 177-185, June 2001.

36. R. G. Garroppo, S. Giordano, S. Porcarelli and G. Procissi, "Testing α-stable processes in modelling broadband teletraffic," in *ICC 2000*, pp. 1615-1619.

37. B. V. Gnedenko, *The Theory of Probability* (translated from the Russian by B.D. Seckler). New York, USA: Chelsea Pub. Co., 1968.

38. M. Grossglauser, "On the relevance of Long-Range Dependence in Network Traffic," *IEEE/ACM Transactions on Networking*, vol. 7, no. 5, pp. 392-403, October 1999.

39. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, USA: John Wiley and Sons, 1985.

40. A. Janicki and A. Weron, *Simulation and Chaotic Behavior of α-Stable Stochastic Processes*. New York , USA: Marcel Dekker, 1994.

41. Y. Jia, M. López Guerrero, O. Kabranov, D. Makrakis and L. Orozco Barbosa, "Design and testbed implementation of adaptive MPLS-DiffServ enabled virtual

private networks," in Proc. of *Canadian Conference in Electrical and Computer Engineering, 2003* (*IEEE CCECE 2003*), vol. 2, pp. 965-968.

42. Y. Jia, M. López Guerrero, O. Kabranov, D. Makrakis and L. Orozco Barbosa, "Dynamic resource allocation in QoS-enabled / MPLS supported virtual private networks and its Linux based implementation," in Proc. of *Canadian Conference in Electrical and Computer Engineering, 2002* (*CCECE 2002*), vol. 3, pp. 1448-1454.

43. Z. Jiang, Y. Li and V. C. M. Leung, "A predictive demand assignment multiple access protocol for broadband satellite networks supporting Internet applications," in Proc. of *ICC 2002*, pp. 2973-2977.

44. A. Karasaridis and D. Hatzinakos, "A non-Gaussian self-similar process for broadband heavy-traffic modeling," in Proc. of *GLOBECOM'98*, vol. 5, pp. 2995-3000.

45. A. Karasaridis and D. Hatzinakos, "Bandwidth allocation bounds for α-stable self-similar Internet traffic models," in *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics (HOS '99)*, pp. 214-218.

46. A. Karasaridis and D. Hatzinakos, "Network heavy traffic modeling using α-stable self-similar processes," *IEEE Transactions on Communications*, vol. 49, no. 7, pp. 1203-1214, July 2001.

47. A. Karasaridis and D. Hatzinakos, "On the modeling of network traffic and fast simulation of rare events using α-stable self-similar processes," in *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics (HOS '97)*, pp. 268-272, 1997.

48. A. Karasaridis and D. Hatzinakos, "Resource allocation issues for long-tailed LRD Internet WAN traffic," in Proc. of *GLOBECOM'99*, vol. 2., pp. 1495-1499.

49. A. Karasaridis and D. Hatzinakos, "Webcasting and Web network traffic measurements and modeling with α-stable self-similar processes," in *1997 IEEE Workshop on Nonlinear Signal and Image Processing* (*NSIP'97*).

50. N. J. Kasdin, "Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 802-827, May 1995.

51. Y. Kasahara, M. Maejima and W. Vervaat, "Log-fractional stable processes," *Stochastic Processes and their Applications*, vol. 30, pp. 329-339, 1988.

52. S. M. Kogon and D. G. Manolakis, "Signal modeling with self-similar α-stable processes: the fractional Lévy stable motion model," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 1006-1010, April 1996.

53. S. M. Kogon and D. B. Williams, "On the characterization of impulsive noise with α-stable distributions using Fourier techniques," in Proc. of *29th Asilomar Conference on Signals, Systems and Computers*, pp. 787-791, 1995.

54. I. A. Koutrouvelis, "An iterative procedure for the estimation of the parameters of stable laws," *Communications in Statistics: Simulation and Computation*, vol. 10, pp. 17-28, 1981.

55. I. A. Koutrouvelis, "Regression-type estimation of the parameters of stable laws," *Journal of the American Statistical Association*, vol. 75, no. 372, pp. 918-928, December 1980.

56. W. E. Leland, M. S. Taqqu, W. Willinger and C. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1-15, 1994.

57. A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Second Edition, Addison-Wesley, 1994.

58. S. Q. Li, S. Chong. C. Hwang and X. Zhao, "Link capacity allocation and network control by filtered input rate in high speed networks," in Proc. of *IEEE INFOCOM'93*, pp.744-750, 1993.

59. M. Lopez-Guerrero, J. R. Gallardo, L. Orozco-Barbosa and D. Makrakis, "On the dynamic allocation of resources using linear prediction of aggregate network traffic," *Computer Communications*, vol. 26, no. 12, pp. 1341-1352, July 2003.

60. M. Lopez-Guerrero, L. Orozco-Barbosa and D. Makrakis, "On the use of probabilistic envelope processes in resource allocation," in Proc. of *Canadian Conference in Electrical and Computer Engineering, 2003* (*CCECE'03*), vol. 2, pp. 993-996, 2003.

61. M. Lopez-Guerrero, J. R. Gallardo, D. Makrakis and L. Orozco-Barbosa, "Optimizing linear prediction of network traffic using modeling based on fractional stable noise," in *ICII2001-Beijing*, vol. 2, pp. 587-592, 2001.

62. M. López-Guerrero, J. R. Gallardo, D. Makrakis and Luis Orozco-Barbosa, "Sensitivity of a network traffic prediction algorithm," in *PACRIM'01*, vol. 2, pp. 615-618, 2001.

63. M. López-Guerrero, L. Orozco-Barbosa and D. Makrakis, "Stochastic upper bounds for aggregate network traffic with modeling based on fractional stable noise," in *Networks 2002*, pp. 345-356.

64. R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proceedings of the Cambridge Philosophical Society*, vol. 58, pp. 497-520, 1962.

65. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422-437, October 1968.

66. P. Markov and H. Mehrpour, "Dynamic bandwidth allocation for streaming MPEG coded video over ATM networks," in Proc. of *ICC 2002*, pp. 2325-2329.

67. G. Mayor and J. Silvester, "Time scale analysis of an ATM queueing system with long-range dependent traffic," in Proc. of *INFOCOM'97*, vol. 1, pp.205-212.

68. T. Mikosch and G. Samorodnitsky, "Ruin probability with claims modeled by a stationary ergodic stable process," *The Annals of Probability*, vol. 28, no. 4, pp. 1814-1851, 2000.

69. D. Morató, J. Aracil, L. A. Díez, M. Izal and E. Magaña, "On linear prediction of Internet traffic for packet and burst switching networks," in Proc. of *International Conference on Computer Communications and Networks 2001* (*ICCCN 2001*), pp. 138-143.

70. C. L. Nikias and N. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. New York, USA: John Wiley & Sons, 1995.

71. I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387-396, 1994.

72. 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, pp. 953-962, August 1995.

73. I. Norros, "Queueing behavior under fractional Brownian traffic," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger editors, pp. 1-38. USA: John Wiley & Sons, 2000.

74. S. A. M. Östring and H. Sirisena, "The influence of long-range dependence on traffic prediction," in *ICC'2001*, vol. 4, pp. 1000-1005.

75. K. Park and W. Willinger, "Self-similar network traffic: an overview," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger editors, pp. 1-38. USA: John Wiley & Sons, 2000.

76. V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, June 1995.

77. G. Procissi, A. Garg, M. Gerla and M. Y. Sanadidi, "Token bucket characterization of long-range dependent traffic," *Computer Communications*, vol. 25, pp. 1009-1017, 2002.

78. E. Reich, "On the integrodifferential equation of Takács. I," *Annals of Mathematical Statistics*, vol. 29, pp. 563-570, 1958.

79. J. Rosiński and T. Żak, "The equivalence of ergodicity and weak mixing for infinitely divisible processes," *Journal of Theoretical Probability*, vol. 10, no.1, pp. 73-86, 1997.

80. Z. Sahinoglu and S. Tekinay, "On multimedia networks: self-similar traffic and network performance," *IEEE Communications Magazine*, vol. 37, no. 1, pp. 48-52, January 1999.

81. G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*. New York, USA: Chapman & Hall, 1994.

82. H. Shimonishi, I. Maki, T. Murase and M. Murata, "Dynamic fair bandwidth allocation for DiffServ classes," in Proc. of *ICC 2002*, pp. 2348-2352.

83. M. Schwartz, *Broadband Integrated Networks*. New Jersey, USA: Prentice Hall PTR, 1996.

84. C. Stathis and B. Maglaris, "Modeling the self-similar behavior of network traffic," *Computer Networks*, vol. 34, pp. 37-47, 2000.

85. B. W. Stuck, "Minimum error dispersion linear filtering of scalar symmetric stable processes," *IEEE Transactions on Automatic Control*, vol. AC-23, no. 3, pp. 507-509, June 1978.

86. M. S. Taqqu, V. Teverovsky and W. Willinger, "Is network traffic self-similar or multifractal?," *Fractals*, vol. 5, no. 1, pp. 63-73, 1997.

87. O. Tickoo and B. Sikdar, "On the impact of IEEE 802.11 MAC on traffic characteristics," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 189-203, February 2003.

88. S. Valaee, "A recursive estimator of worst-case burstiness," *IEEE/ACM Transactions on Networking*, vol. 9, no. 2, pp. 211-222, April 2001.

89. W. Vervaat, "Sample path properties of self-similar processes with stationary increments," *The Annals of Probability*, vol. 13, no.1, pp.1-27, 1985.

90. A. Weron, "Stable processes and measures: a survey," in *Probability Theory on Vector Spaces III*, D. Szynal and A. Weron, eds., pp. 306-364, within the series *Lecture Notes in Mathematics,* vol. 1080. New York, USA: Springer, 1983.

91. W. Willinger, M. S. Taqqu and A. Erramilli, "A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks," in *Stochastic Networks: Theory and Applications*, F.P. Kelly, S. Zachary and I. Ziedins editors, pp. 339-366. Oxford: Clarendon Press, 1996.

92. W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71-86, February 1997.

93. X. Yang, A. P. Petropulu and J. C. Pesquet, "Estimating long-range dependence in impulsive traffic flows," in Proc. of *International Conference in Acoustics, Speech and Signal Processing 2001* (*ICASSP'01*), vol. 6, pp. 3413-3416.

94. J. Zeng and N. Ansari, "Periodic bandwidth allocation based on virtual queue occupancy," in Proc. of *ICC 2002*, pp. 2242-2246.

95. http://ita.ee.lbl.gov/html/contrib/BC.html

96. http://www-info3.informatik.uni-wuerzburg.de/MPEG/traces/

Appendix A

The marginal distributions of Log-FSM

The marginal distributions of a Log-fractional stable motion (Log-FSM) process are briefly discussed here for clarity purposes since the author could not find a suitable reference. Recall from Equations (25) and (26) that Log-FSM can be obtained as follows:

$$\Lambda_\alpha(t) = \int_{-\infty}^{\infty} \left(l(t-x) - l(x)\right) \cdot M(dx) \quad ; \quad -\infty < t < \infty \tag{A.1}$$

where $M(dx)$ an α-stable random measure on **R** with Lebesgue control measure and constant skewness intensity β. For fixed t>0 the stochastic integral (A.1) yields an α-stable random variable i.e.,

$$\Lambda_\alpha(t) \sim S_\alpha(\sigma_t, \beta_t, 0) \tag{A.2}$$

From the properties of the α-stable integral [81 p. 117] we have that σ_t and β_t in (A.2) are given by:

$$\sigma_t^\alpha = \int_{-\infty}^{\infty} \left| l(t-x) - l(x) \right|^\alpha dx \tag{A.3}$$

and

$$\beta_t = \frac{\beta}{\sigma_t^\alpha} \int_{-\infty}^{\infty} \left[l(t-x) - l(x) \right]^{<\alpha>} dx \tag{A.4}$$

where $z^{<p>} = |z|^p \cdot sign(z)$. Consider the scale parameter of the marginal distribution given by (A.3). Let us split the integration interval into the subintervals $(-\infty,0]$, $[0,t]$ and $[t,\infty)$, and then apply the definition of $l(\cdot)$ (see Equation(26)) as to obtain:

$$\sigma_t^\alpha = \int_{-\infty}^{0^-} \left| ln|t-x| - ln|x| \right|^\alpha dx + \int_{0^+}^{t^-} \left| ln|t-x| - ln|x| \right|^\alpha dx + \int_{t^+}^{\infty} \left| ln|t-x| - ln|x| \right|^\alpha dx \tag{A.5}$$

so that the first and third integrals in (A.5) can be transformed into:

$$\int_{-\infty}^{0^-} \left| ln|t-x| - ln|x| \right|^\alpha dx = \int_{0^+}^{\infty} \left| ln|t+x| - ln|x| \right|^\alpha dx \tag{A.6}$$

$$\int_{t^+}^{\infty} \left| ln|t-x| - ln|x| \right|^\alpha dx = \int_{0^+}^{\infty} \left| -ln|x| + ln|t+x| \right|^\alpha dx \tag{A.7}$$

Therefore, (A.5) can be expressed as:

$$\sigma_t^\alpha = \int_{0^+}^{\infty} \left| ln|t+x| - ln|x| \right|^\alpha dx + \int_{0^+}^{t^-} \left| ln|t-x| - ln|x| \right|^\alpha dx + \int_{0^+}^{\infty} \left| ln|t+x| - ln|x| \right|^\alpha dx \tag{A.8}$$

From (A.8) we obtain (58) when we set $t = 1$. The proof for Equation (57) is obtained in a similar way.

Now consider the skewness parameter of the marginal distribution given by (A.4). In this case it is convenient to split the integration interval into the subintervals $(-\infty, 0]$, $[0, t/2]$, $[t/2, t]$ and $[t, \infty)$. Following similar procedures to the ones previously used, it is a matter of simple algebra to show that the resulting integrals cancel out each other yielding a skewness parameter $\beta_t = 0$, which indicates that the marginal distributions of the Log-FSM are symmetric.

Appendix B

Probabilistic envelope process for a traffic model based on symmetric fractional stable noise

Let us consider a particular case of the traffic model given by (36) with the number of arrivals during the interval $(j,j+1]$ given by:

$$W_j = m + aY_j \ ; \ j = 0,1,2,\ldots \tag{B.1}$$

where m is the mean number of arrivals per time unit, a is a scaling factor and Y_j is fractional stable noise with symmetric marginal distributions.

Let A_n represent a cumulative arrival process with $A_0 = 0$ at the beginning of an observation window, i.e.,

$$A_n = W_0 + W_1 + \ldots + W_{n-1} \ ; \ n = 1,2,3,\ldots \tag{B.2}$$

Using (B.1) this process can also be restated as:

$$A_n = nm + a(Y_0 + Y_1 + \ldots + Y_{n-1}) \ ; \ n = 1,2,3,\ldots \tag{B.3}$$

Observe that the summation of samples of fractional stable noise implied by (B.3) becomes in fact fractional stable motion with discrete time index.

In order to provide an upper bound for A_n the following envelope process $\hat{A}_n$ is proposed:

$$\hat{A}_n = m_A(n) + K\sigma_A(n) \tag{B.4}$$

where $m_A(n)$ and $\sigma_A(n)$ are the mean and scale parameter of the cumulative arrival process at time n. The parameter K determines the probability that the process will surpass its envelope. In order to determine the parameters of the envelope process it is useful to apply the definition given by Equation (32) and transform (B.3) into:

$$A_n = nm + \sum_{k=-\infty}^{\infty} U_k \left[a \sum_{j=0}^{n-1} \phi(j-k) \right] \quad ; \; n = 1,2,3,\ldots \tag{B.5}$$

Now we apply expectation to (B.5) to obtain its mean:

$$m_A(n) = E\{A_n\} = nm + \sum_{k=-\infty}^{\infty} E\{U_k\} \left[a \sum_{j=0}^{n-1} \phi(j-k) \right] = nm \tag{B.6}$$

In (B.6) we have used the fact that for the *SαS* random variables under consideration, with $1 < \alpha \le 2$, their mean equals the shift parameter (see [81] p. 19), which in turn is zero.

Using Equation (10) into (B.5) it can be easily shown that the dispersion of the process $\gamma_A(n)$ is given by:

$$\gamma_A(n) = \sum_{k=-\infty}^{\infty} \left| a \sum_{j=0}^{n-1} \phi(j-k) \right|^{\alpha} \gamma_{U_k} \tag{B.7}$$

After some simple algebraic steps, and taking into consideration that the dispersion γ_{U_k} results in a unitary value, we obtain the following expression for the scale parameter of the process A_n:

$$\sigma_A(n) = [\gamma_A(n)]^{1/\alpha} = |a| \cdot \left[\sum_{k=-\infty}^{\infty} \left| \sum_{j=0}^{n-1} \phi(j-k) \right|^{\alpha} \right]^{1/\alpha} \tag{B.8}$$

Finally, the parameter K is uniquely determined by fixing the overshoot probability ε:

$$P\{A_n > \hat{A}_n\} = P\{A_n > m_A(n) + K\sigma_A(n)\} = P\left\{ \frac{A_n - m_A(n)}{\sigma_A(n)} > K \right\} = \varepsilon \tag{B.9}$$

Note that

$$\text{if } X = \frac{A_n - m_A(n)}{\sigma_A(n)} \text{ then } X \sim S_\alpha(1,0,0) \tag{B.10}$$

Therefore, from the distribution of X we can determine K using the probability ε. As previously mentioned, this distribution can be obtained numerically from the characteristic function. The following case study is provided in order to illustrate the concepts described in this section.

Probabilistic envelope processes for aggregate video traffic

In the set of experiments we used traces consisting of 10,000 samples of symmetric balanced fractional stable noise with parameters α =1.95 and H=0.90. As explained in [33] this particular combination can be used to generate synthetic aggregate video (MPEG-1) traffic. We considered several observation window sizes in the range from 1 to 50. The value of K that satisfies an appropriate probability ε had to be determined from a tailored table of K versus ε (i.e, K versus $P\{X > K\}$ where $X \sim S_{1.95}(1,0,0)$). In order to generate such table, the characteristic function was numerically inverted to obtain a punctual representation of the probability density function. Table B.1 shows the theoretical overshoot probability ε and Table B.2 shows the simulation results in terms of the sample mean (95% confidence intervals are also included).

It can be observed that the envelope process let us accurately specify a stochastic upper bound for the values of the synthetic trace, regardless of the observation window size.

Table B.1 Theoretical overshoot probability ε from the distribution $S_{1.95}(1,0,0)$.

K	1	2	3	5
ε	0.2401	0.0808	0.0199	0.0016

Table B.2 Experimentally observed overshoot probability ε.

Window size	K			
	1	2	3	5
1	0.2396 (±0.0401)	0.0795 (±0.0250)	0.0201 (±0.0118)	0.0016 (±0.0014)
2	0.2398 (±0.0402)	0.0794 (±0.0250)	0.0201 (±0.0118)	0.0016 (±0.0014)
3	0.2397 (±0.0403)	0.0793 (±0.0252)	0.0201 (±0.0119)	0.0016 (±0.0014)
5	0.2394 (±0.0405)	0.0795 (±0.0253)	0.0202 (±0.0120)	0.0016 (±0.0015)
10	0.2394 (±0.0410)	0.0794 (±0.0257)	0.0202 (±0.0122)	0.0015 (±0.0014)
20	0.2395 (±0.0419)	0.0792 (±0.0260)	0.0201 (±0.0125)	0.0016 (±0.0015)
30	0.2388 (±0.0423)	0.0778 (±0.0265)	0.0198 (±0.0129)	0.0014 (±0.0015)
50	0.2388 (±0.0431)	0.0778 (±0.0271)	0.0202 (±0.0142)	0.0016 (±0.0017)

Appendix C

Method of sample fractiles

Let $X_1,\ldots,X_N$ be N samples of a random variable X. We can obtain the corresponding order statistics $X_{(1)},\ldots,X_{(N)}$ by placing the sample values in ascending order i.e., $X_{(1)} \leq \ldots \leq X_{(N)}$. Recalling that for a cumulative distribution function $F(x)$ its f ($0<f<1$) fractile x_f is defined by $F(x_f)=f$, an estimator for the f fractile $\hat{x}_f$ is simply given by $X_{(N+1)f}$. However, in order to avoid spurious skewness in the estimate a correction has to be made as described below [70].

Assuming that $0 \leq i \leq N$ and $\frac{2i-1}{2N} \leq f \leq \frac{2i+1}{2N}$ then

$$\hat{x}_f = X_{(i)} + \left(X_{(i+1)} - X_{(i)}\right)\frac{f-q(i)}{q(i+1)-q(i)} \tag{C.1}$$

where

$$q(i) = \frac{2i-1}{2N}$$

with $\hat{x}_f = X_{(1)}$ for i=0 and $\hat{x}_f = X_{(1)}$ for i=N.

In [26] the authors proposed to estimate the scale parameter σ of an α-stable distribution through the appropriate sample fractiles. In the case of standard symmetric α-stable distributions they observed that the 0.72 fractile is in the interval 0.827±0.003 for $1 \leq \alpha \leq 2$, therefore it was suggested to estimate the scale parameter σ as:

$$\hat{\sigma} = \frac{1}{2(0.827)}\left(\hat{x}_{0.72} - \hat{x}_{0.28}\right) \tag{C.2}$$

Due to its computational simplicity and robustness over a large portion of the parameter space, some methods use Equation (C.2) in order to provide initial estimates for the scale parameter of skewed α-stable distributions [54][55].

Appendix D

Fraction of downgraded data produced by constant envelope-based bandwidth allocation

Let us again represent with $\{W_j^*, j = 0, \ldots, N-1\}$ a sequence of samples of a traffic process and let $\{W_j = m + \sigma_W Y_j\}$ be its model (i.e., for any integer $d \geq 0$ with $j + d \leq N-1$, the vectors $(W_j^*, \ldots, W_{j+d}^*)$ and $(W_j, \ldots, W_{j+d})$ have the same distribution), where m is the mean number of arrivals per time slot, σ_W is the scale parameter of the traffic process and $\{Y_j\}$ is a *standard* fractional stable noise. The probabilistic envelope process for this traffic model is given by $\hat{W} = m + K\sigma_W$, where m and σ_W have the same meaning as above and K is a parameter that controls the probability ε that a traffic sample exceeds its envelope, i.e. $P\{W_j^* > \hat{W}\} = \varepsilon$.

Consider a constant bandwidth allocation policy where data exceeding the envelope are downgraded (dropped or given a lower priority). We can quantify the fraction of downgraded data as indicated by (95) or by using the following equivalent way:

$$Downgraded\_data = \frac{\sum_{j=0}^{N-1} max\left(0, W_j^* - \hat{W}\right)}{\sum_{j=0}^{N-1} W_j^*} \tag{D.1}$$

Let us apply expectation to (D.1) and make use of the traffic model in order to estimate this metric of performance. Thus,

$$E\{Downgraded\_data\} = E\left\{\frac{\sum_{j=0}^{N-1}\max\left(0, W_j - \hat{W}\right)}{\sum_{j=0}^{N-1} W_j}\right\} \tag{D.2}$$

Observe that the denominator in (D.2) can be expressed in terms of the components of the traffic model as $Nm + \sigma_W(Y_0 + ... + Y_{N-1})$. Recall that the fractional stable noise processes considered in this work are ergodic with zero mean and therefore, a summation of a large number of their samples tends to zero. Assuming a large *N*, the denominator becomes a constant and we can approximate the computation of (D.2) as follows:

$$E\{Downgraded\_data\} \cong \frac{1}{Nm}\sum_{j=0}^{N-1} E\left\{max\left(0, W_j - \hat{W}\right)\right\} \tag{D.3}$$

where we have also made use of the linear property of the expectancy operator. Since the traffic model is a stationary sequence, the expected value on the right hand side of (D.3) does not change from sample to sample and the following equation holds:

$$E\{Downgraded\_data\} \cong \frac{1}{m} \cdot E\left\{\max\left(0, W - \hat{W}\right)\right\} \tag{D.4}$$

Without loss of generality, in (D.4) we have eased the notation by making use of a random variable W representing the marginal distribution of the traffic model (i.e., $W \overset{d}{=} W_j$). Equation (D.4) implies that:

$$E\{Downgraded\_data\} \cong \frac{1}{m}\left(\int_{-\infty}^{\hat{W}} 0 \cdot f_W(W)dW + \int_{\hat{W}}^{\infty}\left(W - \hat{W}\right)f_W(W)dW\right) \tag{D.5}$$

where $f_W(W)$ is the marginal distribution of $\{W_j\}$. Therefore,

$$E\{Downgraded\_data\} \cong \frac{1}{m}\int_{\hat{W}}^{\infty}\left(W-\hat{W}\right)f_W(W)dW \tag{D.6}$$

Let us express the integral in (D.6) in terms of the components of both, the envelope process and the traffic model. In order to ease the notation, let us denote the marginal distribution of the fractional stable process $\{Yj\}$ with a random variable Y (i.e., $Y \overset{d}{=} Y_j$). From previous definitions, given at the beginning of this section, we can make the following substitutions: a) the limits of the integral can be re-expressed in terms of Y, considering that the limit $W=\hat{W}$ implies that $Y=K$ and $W\to\infty$ implies that $Y\to\infty$; b) the factor $\left(W-\hat{W}\right)$ becomes $\sigma_W(Y-K)$; c) the probability density function[18] $f_W(W)=(1/\sigma_W)f_Y((W-m)/\sigma_W)$ becomes $f_Y(Y)/\sigma_W$ and d) $dW=\sigma_W dY$. Thus, we finally obtain:

$$E\{Downgraded\_data\} \cong \frac{\sigma_W}{m}\int_{K}^{\infty}(Y-K)f_Y(Y)dY \tag{D.7}$$

Equation (D.7) can be simplified a bit further if we re-express it as:

$$E\{Downgraded\_data\} \cong \frac{\sigma_W}{m}\left(\int_{K}^{\infty}Yf_Y(Y)dY - K\int_{K}^{\infty}f_Y(Y)dY\right) \tag{D.8}$$

The second integral on the right hand side of (D.8) is, in fact, the probability ε, since $P\left\{W_j^* > \hat{W}\right\} = P\left\{Y_j > K\right\} = \varepsilon$. Therefore, we can conclude that:

$$E\{Downgraded\_data\} \cong \frac{\sigma_W}{m}\left(\int_{K}^{\infty}Yf_Y(Y)dY - K\varepsilon\right) \tag{D.9}$$

[18] This function is obtained from the density function of Y_j through the well-known formula to determine the probability density function of a linear function of a random variable (see [57 p. 121]).

Although there are no closed-form expressions for the α-stable probability density function $f_Y(Y)$, the integrals in (D.7) or (D.9) can be computed numerically.