INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA**

UMI[®]
800-521-0600

A Cost Minimization Methodology for Shaping of Self-Similar Traffic

by

Galina Doubrovina, (MSc (Geophysics), PhD (Phys. & Math.))

Moscow State University, Moscow, Russia

**A thesis submitted to
the Faculty of Graduate Studies
in partial fulfillment of
the requirements for the degree of**

Master of Information and Systems Science

**Faculty of Engineering
Department of Systems and Computer Engineering**

Carleton University

March 22, 1999

Ottawa, Ontario

© copyright

1999, Galina Doubrovina

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

395 Wellington Street
Ottawa ON K1A 0N4
Canada

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-43357-9

Canada

The undersigned recommended to the Faculty of Graduate Studies and Research
acceptance of the thesis

**A Cost Minimization Methodology for Shaping of Self-Similar
Traffic**

submitted by Galina Doubrovina

in partial fulfillment of the requirements for
the degree of Master of Information and Systems Science

Prof. M. Devetsikiotis, Thesis Supervisor

Chair, Department of Systems and
Computer Engineering

Carleton University

Abstract

This work addresses the problem of determining the parameters of ATM network traffic shaper which is represented as a single server queue with finite buffer and deterministic service rate. In our scheme, the user wishes to gain access to the network with certain cell loss probability guaranteed when the input traffic is self-similar. The user has to provide the network with mean cell rate, variance, degree of self-similarity (Hurst parameter) of the submitted traffic and the desirable QoS. The network is able to provide a connection if the traffic is shaped by the user or by the network, and if sufficient resources are available to accommodate the shaped traffic stream. The network forces the user to declare his/her traffic parameters as close to reality as possible in order not to pay extra money. This allows the network to get more revenue by admitting more users. We formulate a minimization problem to determine the optimal parameters for the traffic shaper and use techniques of non-linear programming to obtain the solution. In our approach, traffic shaping involves minimization of the connection cost in terms of the network resources, such that access to the network is obtained and the desired QoS is provided by the connection. The algorithm of the proposed scheme is described and its workability is shown for self-similar traffic.

Acknowledgments

I would like to express sincere appreciation to Prof. M. Devetsikiotis for his patient direction and warm support for this work. His guidance and encouragement is very important during the course of the thesis. I am extremely grateful to Matthias Falkner for very helpful comments and advices he gave me during my work on this thesis.

I wish to thank my husband, son and parents for their support and encouragement during this research.

I gratefully acknowledge the funding provided by the Natural Sciences and Engineering Research Council of Canada.

Table of Contents

Acceptance sheet.....	ii
Abstract.....	iii
Acknowledgments.....	iv
Table of Contents.....	v
List of Tables.....	vii
List of Figures.....	viii
List of Acronyms.....	ix
Chapter 1.	
Introduction	1
1.1. Background.....	1
1.2. Contributions of this Thesis.....	4
1.3. Thesis Organization.....	5
Chapter 2.	
Self-Similar Traffic and Related Issues	6
2.1. Introduction.....	6
2.2. Self-Similar Traffic Streams: Background.....	7
2.3. Self-Similar Models, Fractional Gaussian Noise Generation.....	9
Chapter 3.	
Theoretical Approximation of Cell Loss Probability for Self-Similar Traffic in ATM Networks	14
3.1. Quality of Service: Definitions.....	14
3.2. Cell Loss Probability: Analytical Expressions.....	16

Chapter 4.	
State of the Art in Connection Admission Control.....	22
Chapter 5.	
A Brief Review of Network Pricing Schemes.....	33
5.1. Pricing Communication Services.....	33
5.2. Pricing in ATM networks.....	37
5.3. The Role of Price in Connection Establishment Process.....	40
Chapter 6.	
Optimal Cost Traffic Shaping Model: Problem Statement and Solution.....	45
6.1. Problem Statement.....	45
6.2. Problem Solution.....	50
Chapter 7.	
Optimal Cost Traffic Shaping Model: Experimental Results.....	55
7.1. Simulation.....	55
7.2. Minimization Results.....	59
7.3. Combined Minimization and Simulation Results.....	65
Chapter 8	
Conclusions and Recommendations for Future Research.....	70
8.1. Conclusions.....	70
8.2. Recommendations for Future Research.....	72
References.....	74

List of Tables

Table 1: Analytical Verification of the Minimization Problem.....	61
Table 2: Minimization Results (the Optimal Solution is not Included).....	63
Table 3: Minimization Results for Different Feasible Regions Spliced in the Service-Dimension.....	64
Table 4: Minimization Results for Different Feasible Regions Spliced in the Buffer-Dimension.....	65
Table 5: A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{\max} = 1.0 \times 10^{-1}$).....	66
Table 6: A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{\max} = 1.0 \times 10^{-2}$).....	67
Table 7: A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{\max} = 1.0 \times 10^{-3}$).....	67
Table 8: A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{\max} = 1.0 \times 10^{-4}$).....	67

List of Figures

Figure 6.1. Contour Plot of the Feasible Region.....	50
Figure 7.1. Single Server Queue Scheme.....	55
Figure 7.2,a. Empirical and Analytical Autocorreleation Curves ($H = 0.65$).....	56
Figure 7.2,b. Empirical and Analytical Autocorreleation Curves ($H = 0.75$).....	57
Figure 7.3,a. Simulated and Analytical CLP (vs. Buffer Size).....	58
Figure 7.3,b. Simulated and Analytical CLP(vs. Service Rate).....	59
Figure 7.4. Value of the Cost Function Restricted to the Feasible Region.....	62
Figure 7.5, a. Simulated CLP ($CLP_{\max}=1.0 \times 10^{-2}$, $H = 0.65$).....	68
Figure 7.5, b. Simulated CLP ($CLP_{\max}=1.0 \times 10^{-2}$, $H = 0.75$).....	69
Figure 7.5, c. Simulated CLP ($CLP_{\max}=1.0 \times 10^{-2}$, $H = 0.85$).....	69

List of Acronyms

ATM	Asynchronous Transfer Mode
B-ISDN	Broad-Band Integrated Services Digital Networks
CAC	Call Admission Control
CER	Cell Error Ratio
CDF	Cumulative Distribution Function
CDV	Cell Delay Variation
CLP	Cell Loss Probability
CLR	Cell Loss Rate
CMR	Cell Misinsertion Rate
CTD	Cell Transfer Delay
CVDT	Cell Variation Delay Tolerance
F-ARIMA	Fractional Autoregressive Integrated Moving-Average
FBM	Fractional Brownian Motion
FGN	Fractional Gaussian Noise
IDC	Index of Dispersion of Counts
IID	Independent Identically Distributed
ISDN	Integrated Services Digital Network
ITU-T	ITU Telecommunication Standardization Sector
KKT	Karush-Kuhn-Tucker
LAN	Local Area Network

LRD	Long-Range Dependence
MCR	Minimum Cell Rate
NLP	Non-Linear Programming
OBJ	Objective Function
PCR	Peak Cell Rate
QP	Quadratic Programming
QoS	Quality of Service
SQP	Sequential Quadratic Programming
SCR	Sustained Cell Rate
SECBR	Severely-Errored Cell Block Ratio
SMDS	Switched Multimegabit Data Service
SRP	Spatial Renewal Process
UNI	User-to Network Interface
VBR	Variable Bit Rate
VC	Virtual Circuit
VP	Virtual Path
WAN	Wide Area Network

Chapter 1

Introduction

1.1. Background

ATM attracts attention as a technology that solves not only tough network problems of today, but promises the speed and flexibility to answer demands of the future as well [1, 2, 3]. ATM evolved as an attempt to combine both fundamental approaches to switching (packet switching and circuit switching) into a single integrated switching mechanism - the grand unified network. It has received much attention because of its high capacity, bandwidth scalability, and its ability to support multi-service traffic. ATM has been designed from the outset to carry voice, video and bursty data traffic on a single common protocol. Moreover, ATM-based networks will have to provide a guaranteed performance or Quality of Service (QoS) to many of traffic classes, and it is not enough anymore for networks to just provide a connection from a source to a destination on a best-effort basis. The actual QoS performance metrics of interest can vary from one application to another, but usually include such measures as cell loss, delay and delay jitter guarantees and call blocking constraints guarantees. Some criteria for call acceptance/rejection must exist in order to preserve the network from getting into a congested state [1, 3].

Call Admission Control (CAC) algorithms are used to prevent the network from suffering congestion. To be efficient, call admission control should be based on a traffic contract between the user and the network. The user is expected to characterize the traffic stream he/she wants to transmit and a desired QoS level. This will give the network an indication

of the anticipated load of the connection. The network then determines its resources availability. If sufficient resources are available to carry the connection without degrading the QoS of existing traffic streams, the call is admitted to the network. A specific set of user-declared traffic characterizations and the corresponding QoS are then registered in the traffic contract. The user's traffic stream is subsequently policed to ensure that the actual traffic stream does not violate the traffic contract. Furthermore, the user may shape its traffic stream to comply with the traffic contract.

Though many different CAC schemes were proposed which lead to high utilization of the network, in practice, a peak rate allocation scheme is often used, which does not permit to utilize the available network resources efficiently. When parties set up a connection to transmit the information streams while guaranteeing the agreed QoS, they should also agree in prices. If the price for the required QoS is too high for the user, there is no sense for him/her in connection with the network. Instead, he/she has to decide if he/she wishes to pay more or sacrifice the QoS.

While pricing schemes for peak bandwidth allocation may be trivial – just a constant charge per time unit or data unit, these schemes would be extremely inefficient for other types of bandwidth allocation, which are likely to be introduced in everyday practice in the near future. Obviously enough, multimedia traffic (considered to be soon the predominant type of traffic in ATM networks) can not use peak bandwidth allocation if a reasonable measure of efficiency has to be achieved. Moreover, network users wish to minimize their costs while network providers wish to increase their revenues. Thus, it is

desirable to design such a pricing scheme which would penalize users for requesting and not actually utilizing excessive network resources or, putting it differently, encouraging users to request an optimal amount of resources when negotiating a traffic contract. This will enable network providers to serve more users and maximize revenues while maintaining requested QoS.

This work presents such traffic shaping methodology for self-similar user traffic. It extends the methodology proposed in [4] to cover self-similar user traffic. Given the user's desired QoS and the network resources availability, our proposed procedure determines the least-cost parameters for a traffic shaper that guarantees access to the network while satisfying the QoS constraints. In [4] this is done for on/off sources, and the QoS constraints are formulated using an effective bandwidth notation.

In our work, the traffic shaping problem is solved for self-similar sources and interpreted as trying to find the minimum cost parameters of the traffic shaper which enable the user to gain access to the network while satisfying the QoS requirements of the connection. The user provides basic characteristics of the traffic stream, such as the mean rate, the variance, and the degree of self-similarity (Hurst parameter). Furthermore, the user specifies a maximum tolerated cell loss probability ($CLP_{\max}$).

The network supplies absolute upper bounds on the available resources as well as the form of the cost function by which the user is charged. A cost minimization problem has then to be solved either by the user or by the network to determine the minimum cost

parameters of a traffic shaper which will guarantee access to the network while satisfying the QoS demands.

By choosing the appropriate form of the cost function, the network forces the user to declare his/her traffic parameters as close to reality as possible in order not to pay extra money. This allows the network to get more revenue by admitting more users.

We used the proposed technique for ATM network, but it can be used for any QoS-oriented (connection-oriented) network.

1.2. Contributions of This Thesis

The major steps of research for this thesis are:

1. Study of self-similar traffic concepts.
2. Analysis of existing call admission control schemes.
3. Analysis of existing pricing schemes for data networks.
4. Selection of a self-similar traffic generator and adapting it for solving of our problem.
5. Study of different analytical approaches for self-similar traffic characterization (such as cell loss probability, cell delay distribution, etc.).
6. Study and analysis of different non-linear optimization techniques in order to choose the most suitable one and adapt it to our problem.
7. Development of a methodology for determining traffic shaper parameters during connection admission procedure.
8. Verification of proposed methodology using both simulation and analytical results.

1.3. Thesis Organization

The remaining chapters of the thesis are organized as follows:

Chapter 2: Introduces the concepts of self-similar traffic.

Chapter 3: Presents theoretical approximation of cell loss probability for self-similar traffic, that is partially used later in our work.

Chapter 4: Reviews existing pricing schemes for data networks.

Chapter 5: Describes call admission control algorithms for different classes of sources.

Chapter 6: Formulates the minimization problem to determine the minimum cost traffic shaping parameters for a user with self-similar input traffic.

Chapter 7: Presents practical results we have obtained using the proposed scheme that show the feasibility of our traffic shaping methodology.

Chapter 8: Gives the conclusions and provides suggestions for future research.

Chapter 2

Self-Similar Traffic and Related Issues

2.1. Introduction

Asynchronous Transfer Mode (ATM) network was adopted as the transfer mode of broad-band integrated services digital networks (B-ISDN) and so is one of the most promising solutions for the next generation networks. ATM is a high-speed network that must support applications with diverse traffic characteristics and quality of service (QoS) requirements. Accurate traffic characterization and modeling are very important for making the use of ATM network resources more efficient [1, 2, 3].

Studies of traffic measurements for LAN traffic, WAN traffic and variable bit rate (VBR) traffic [5, 6, 7, 8, 9] have shown that the actual network traffic is long range dependent (LRD) and would be better modeled by self-similar (“fractal”) rather than traditional Poisson-based process. A bibliographical guide to the papers on self-similar traffic is given in [10].

If traffic were Poisson, its characteristic burst length would be smoothed by averaging over a long enough time scale, but measurements of real traffic indicate that significant traffic burstiness is presented on a wide range of time scales. Traffic that is bursty on many or all time scales can be described statistically using a notion of *self-similarity*, which is a property associated with fractals - objects whose appearance is unchanged regardless of the scale at which they are viewed [11]. Contrary to the common belief that

multiplexing traffic streams tends to produce smoothed-out aggregate traffic with reduced burstiness, aggregating self-similar traffic streams can intensify burstiness rather than diminish it. Since a self-similar process has observable bursts on all time scales, it exhibits *long-range dependence* (LRD).

Self-similarity itself is characterized in different ways: a spectral density that diverges at the origin, a nonsummable autocorrelation function (indicating long-range dependence), an index of Dispersion of Counts (IDC) that increases monotonically with the sample time T , etc. [8]. The parameter designed to indicate the degree of self-similarity and characterize self-similar processes is the so-called Hurst parameter H ($0.5 < H < 1$), named after H.E.Hurst who studied long-term storage in water reservoirs [12].

As $H \rightarrow 1$, the degree of self-similarity increases. The importance of long-range dependence in network traffic was observed in studies such as [8], which showed that packet loss and delay behavior is radically different in simulations using real traffic data rather than traditional network models.

2.2. Self-Similar Traffic Streams: Background

A detailed discussion on self-similarity and long-range dependency is given in different papers by many authors (see for example, [5, 8, 9, 11, 13, 14, 15, 16, 17, 18, 19]). We will briefly summarize the definitions given in these papers. All the details and proofs the reader can find in [18,19].

Consider a second-order stationary real-number stochastic process $X = (..., X_{-1}, X_0, X_1, ...)$ of discrete argument (time) $t \in I_{-\infty} = \{..., -1, 0, 1, ...\}$. Denote by $\mu = E[X_t] < \infty$ and $\sigma^2 = \text{var}[X_t] < \infty$ - the mean and the variance of X_t , respectively. Denote by

$$r(k) = \frac{E[(X_{t+k} - \mu)(X_t - \mu)]}{\sigma^2}, \quad k \in I_0 = \{0, 1, 2, ...\},$$

the autocorrelation function of the process X . The mean μ , the variance σ^2 , and the autocorrelation function $r(k)$ do not depend on time t , and $r(k) = r(-k)$. [see 8, 13, 18, 19].

Definition 1. A process X is called *exactly second-order self-similar* with parameter

$$H = 1 - \frac{1}{2}\beta, \quad 0 < \beta < 1, \text{ if its autocorrelation function is}$$

$$r(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}] = g(k), \quad k \in I_1 = \{1, 2, \dots\}.$$

The process X with $r(k) = g(k)$ is called *self-similar* because it does not change its correlation structure with averaging. This means that the averaged over blocks of m process $(..., X_{-1}^{(m)}, X_0^{(m)}, X_1^{(m)}, ...)$,

where

$$X_t^{(m)} = \frac{1}{m}(X_{tm-m+1} + \dots + X_{tm}), \quad m \in I_1, \quad t \in I_{-\infty}$$

has the autocorrelation function $r^{(m)}(k)$ such that

$$r^{(m)}(k) = r(k) = g(k).$$

Definition 2. A process X is called *asymptotically second-order self-similar* with

$$\text{parameter } H = 1 - \frac{1}{2}\beta, \quad 0 < \beta < 1, \text{ if for all } k \in I_1, \quad \lim_{m \rightarrow \infty} r^{(m)}(k) = g(k).$$

This means that X is asymptotically second-order self-similar if, after averaging over blocks of m and with $m \rightarrow \infty$, its correlation structure tends to that of exactly second-order self-similar process.

The important property of X , which guarantees that X is asymptotically second-order self-similar with parameter $H = 1 - \frac{1}{2}\beta$, is $\lim_{k \rightarrow \infty} \frac{r(k)}{k^{-\beta}} = c$, where $0 < c < \infty$ is a constant and $0 < \beta < 1$ [18]. This limiting equation gives the definition of *long-range dependence* (LRD) process (see [5]). A process with long-range dependence has an autocorrelation function $r(k) \sim k^{-\beta}$ as $k \rightarrow \infty$, where $0 < \beta < 1$. Thus, the autocorrelation function of such a process decays hyperbolically (as compared to the exponential decay exhibited by traditional traffic models) [11].

2.3. Self-Similar Models and Fractional Gaussian Noise Generation

The need in conducting of simulation for self-similar traffic has motivated creation of appropriate models. There are many stochastic models that exhibit the self-similar property, but the most widely studied are the exactly self-similar fractional Gaussian noise (FGN) and the asymptotically self-similar fractional autoregressive integrated moving-average (F-ARIMA) processes. Associated with FGN is fractional Brownian motion (FBM), which is the integrated version of FGN (i.e. an FBM process is the sum of FGN increments) [5, 20, 21, 22, 23, 24, 25, 26, 27, 28]. These models require only a few parameters to be specified such as mean, variance and Hurst parameter, and they produce natural looking traffic. FGN and F-ARIMA models differ in their assumptions about the

short-range dependencies in data: FGN assumes that there is no short-range dependence, while F-ARIMA can model both long time dependence and short time dependence at the same time [29].

A fractional Gaussian noise process $X = \{X_k: k = 1, 2, \dots\}$ is a stationary Gaussian process with mean $m = E[X_k]$, variance $\sigma^2 = E[(X_k - m)^2]$, and autocorrelation function

$$\begin{aligned} r(k) &= 1, & k=0 \\ r(k) &= \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}), & k = 1, 2, 3, \dots \end{aligned} \quad (2.1)$$

Therefore, if $\frac{1}{2} < H < 1$, FGN is exactly second-order self-similar with Hurst parameter H [30].

FGN can be viewed as a reasonable first approximation of LRD processes, since it can be derived from a special type of central limit theorem applied to LRD processes [31]. The down side of this model is that it is computationally demanding [25, 32]. For example, Hosking's method, which can be used to generate FGN, generates sample traces with $O(n^2)$ computational effort with respect to sample path length. In practice this has limited simulation runs to a maximum of about 5×10^5 samples [33]. Techniques for effective generation of approximate FBM processes include the Random Midpoint Displacement algorithm that allows to generate long traces from FBM processes accurately and quickly [34], and the aggregation of on/off sources produced by chaotic map [35]. The generation of FGN by aggregated Markov-Gauss processes is described in [36].

T.Taralp et al. [33] proposed an efficient and simple technique to generate a good approximation of FGN traffic. The technique is based on the Spatial Renewal Process (SRP). The SRP process can exactly match any given marginal distribution, and at least in principal, can even match sub-exponential autocorrelation functions that produce LRD-traffic [33]. The SRP model is fit to the steady-state distribution and autocorrelation requirements that characterize an FGN process. Another advantage of the SRP model is that it is computationally inexpensive in comparison with another LRD-traffic models: it generates sample traces with $O(n)$ effort with respect to path length, has very limited memory requirements and generates data “on-line” [33]. These were the reasons why SRP-FGN generator was chosen in our work for LRD-traffic simulation.

The Spatial Renewal Process, introduced by Jelencovic, Lazar and Semret [37] is a subclass of sub-exponentially time dependent stochastic processes. Sub-exponentiality means that the autocorrelation function of the stationary process decays not exponentially for large lags but in a slower polynomial or “Pareto-like” decay. This implies that the time dependence is long-range with a Hurst parameter greater than 0.5 and an infinite autocorrelation function energy. Consider a point process with independent identically distributed (IID) inter-renewal time sequence $\{T_n\}$ described by the CDF function $F_T(t)$. This process plays a role of background process in the SRP model and is responsible for the time dependence structure. A second process, independent of the first, is a sequence of IID values $\{X_n\}$, which are distributed according to the desired steady-state marginal distribution of the traffic. The SRP process is a chain of renewal periods [33]. The marginal distribution of the output process is identical to the marginal of X since the

processes X_n and T_n are independent of one another, so any marginal can be chosen for X output process. The autocorrelation function $r(t)$ of the output process is related to the SRP renewal distribution function $F_T(t)$ as follows:

$$1 - r(t) = \mu^{-1} \int_0^t (1 - F_T(u)) du, \quad t \geq 0 \quad (2.2)$$

$$-\frac{d}{dt}r(t) = \mu^{-1}(1 - F_T(t)), \quad r(0) = 1, \quad t \geq 0 \quad (2.3)$$

$$\text{where } \mu = \int_0^\infty (1 - F_T(u)) du$$

The autocorrelation function $r(t)$ must be a convex function for all times $t \geq 0$ in order to be sure that the corresponding renewal distribution function $F_T(t)$ is a valid CDF function. The autocorrelation function of Pareto, exponential, or piecewise linear form can be chosen, provided that the following constraints are met:

$$r(t) = 1, \quad t = 0$$

$$\frac{d}{dt}r(t) \leq 0, \quad \forall t \geq 0 \quad (2.4)$$

$$\frac{d^2}{dt^2}r(t) \geq 0, \quad \forall t \geq 0 \quad (2.5)$$

The autocorrelation function $r(k)$ of an exactly second-order self-similar FGN process with Hurst parameter $0.5 < H < 1$ (see (2.1)) was redefined as a continuous function of t for compatibility with the SRP model [33]:

$$r(t) = \frac{1}{2}(2 + (2^{2H} - 4)t), \quad 0 \leq t < 1 \quad (2.6)$$

$$r(t) = \frac{1}{2}((t+1)^{2H} - 2t^{2H} + (t-1)^{2H}), \quad t \geq 1$$

Applying the formulas (2.1) - (2.6) with $\mu = \text{const}$ chosen such that $F_T(0) = 0$ to a renewal distribution function $F_T(t)$ gives the follows [33]:

$$F_T(t) = (1 + \mu)(2^{2H-1} - 2), \quad 0 \leq t < 1 \quad (2.7)$$

$$F_T(t) = (1 + \mu)H((t+1)^{2H-1} - 2t^{2H-1} + (t-1)^{2H-1}), \quad t \geq 1$$

$$\mu = -(2^{2H-1} - 2)^{-1}$$

We calculated the SRP renewal function $F_T(t)$ with different Hurst parameters ($H = 0.65$, $H = 0.75$, $H = 0.85$, $H=0.95$) using equations (2.7), approximated by a linear piecewise function and loaded into the SRP generator program. The background X_n IID sequence was generated according to an $N(0,1)$ distribution because the steady-state distribution of an FGN process is Gaussian. The $N(0,1)$ CDF was also approximated with a linear piecewise function for use by the generator. To take advantage of the averaging effects of aggregation, 25 independent SRP streams (individually producing FGN) were generated and combined. The output was computed by summing the traces and re-normalizing the sample variance back to one.

The FGN process is fully characterized by its autocorrelation function, given that its steady-state distribution is Gaussian and the covariance property of jointly Gaussian random variables applies [38]. So, when the autocorrelation function produced by the SRP generator closely follows the theoretical autocorrelation function (2.1) and replicates an $N(0,1)$ steady-state distribution, the output is FGN.

Chapter 3

Theoretical Approximation of Cell Loss Probability for Self-Similar Traffic in ATM Networks

3.1. Quality of Service: Definitions

Quality of service is an important issue for ATM networks, in particular because of real-time traffic such as audio and video. When a virtual circuit is established, both the user and the network have to agree on a contract defining the service. To make it possible to have concrete traffic contracts, the ATM standard defines a number of QoS (Quality of Service) parameters whose values the user and network can negotiate. For each QoS parameter, the worst case performance for each parameter is specified, and the network is required to meet or exceed it. Tannenbaum [39] (as well as ATM Forum) considers the following QoS parameters:

I. Parameters that specify how fast the user wants to send the traffic:

PCR (Peak Cell Rate) is the maximum rate at which the sender is planning to send cells;

SCR (Sustained Cell Rate) is the expected or required cell rate averaged over a long time interval;

MCR (Minimum Cell Rate) is the minimum number of cells/sec that the user considers acceptable. If the network can not guarantee to provide enough bandwidth for this, it must reject the connection.

CVDT (Cell Variation Delay Tolerance) tells how much variation will be present in cell transmission times.

II. Parameters which describe network characteristics and are measured at the receiver (there are negotiable):

CLR (Cell Loss Ratio) (or **CLP (Cell Loss Probability)**) measures the fraction of the transmitted cells that are lost or are delivered so late as to be useless;

CTD (Cell Transfer Delay) is the average transit time from source to destination;

CDV (Cell Delay Variation) measures how uniformly the cells are delivered.

III. Parameters which specify the network characteristics (there are generally not negotiable):

CER (Cell Error Ratio) is the fraction of cells that are delivered with one or more bits wrong;

SECBR (Severely-Errored Cell Block Ratio) is the fraction of N-cell blocks of which M or more cells contain an error;

CMR (Cell Misinsertion Rate) is the number of cells/sec that are delivered to the wrong destination on account of an undetected error in the header.

The main QoS parameters which the user and the network can negotiate and that are the part of the user-network contract, are CLR (CLP) and CDV. If the network has not enough resources to satisfy these parameters specified by the user in the contract, the user's call can not be admitted by the network. So the satisfaction of these QoS constraints is an important part of the call admission control schemes. In order to specify these constraints for the self-similar traffic, they have to be available theoretically. There

are several papers [40, 41, 42, 43, 44, 45, 46, 47, 48] which present the methods for estimating QoS parameters such as cell loss probability, buffer overflow probability, cell delay distribution in ATM networks, considering self-similar behavior of the arrival process.

3.2. Cell Loss Probability: Analytical Expressions

H. Saito, K. Shiimoto in [49, 50, 51] consider the upper bound of cell loss probability without assuming cell arrival process, based on the average number of cells ψ that arrive during a fixed interval, and the maximum number of cells ξ that arrive in the same fixed interval. The buffer size is determined such that the maximum delay under first come, first served discipline can be satisfied. The output buffer size is K , and the maximum admissible delay in the output buffer T is assumed to satisfy the relationship $K = TC/L$, where C - the bandwidth of a transmission link, and L - the length of the cell. Assuming that cells of n calls are transmitted by a transmission link and the probability that j cells of the i^{th} call arrive during τ is $p_i^*(j)$, $i = 1, \dots, n$; $j = 0, 1, \dots$, it was shown that the cell loss probability is upper-bounded by the following equation:

$$CLP \leq \frac{\sum_{k=0}^{\infty} [k - C\tau / L]^+ p_1^* \dots p_n^*(k)}{\sum_{k=0}^{\infty} k p_1^* \dots p_n^*(k)},$$

where $C\tau / L \leq K + 1$, $*$ is the convolution operation, and $[x]^+ = x$, if $x > 0$, and $[x]^+ = 0$, otherwise.

For $(n+1)^{\text{th}}$ call let $\theta_{n+1}(j) = 1 - \psi_{n+1} / \xi_{n+1}$, $j = 0$,

and $\theta_{n+1}(j) = 0$, otherwise.

It can be shown that the estimation of cell loss probability after the new, $(n+1)^{\text{th}}$, call acceptance, is as follows:

$$\text{CLP}_e \leq \frac{\sum_{k=0}^{\infty} [k - C\tau / L]^+ p_e * \theta_{n+1}(k)}{(\sum_{k=0}^{\infty} k p_e(k) + \tau a_{n+1} / L)}, \text{ where } p_e = (p_e(0), p_e(1), \dots) \text{ is the estimate}$$

of $\{p_1^* \dots p_n^*(k)\}$, the probability of the number of cells arriving from connected sources [49]. It can be expected that the cell loss probability is upper-bounded by the estimate CLP_e when the errors in estimating p_e are sufficiently small.

N. Likhanov, B. Tsybakov, N. Georganas [17, 18, 19] base the considered self-similar models on the framework of some known marked point process [14, 52]. The basis model used is called source process and was obtained by reducing to some extent the generality of the marked point process. Four models of self-similar input traffic, which are particular cases of the source process, are considered: the process with constant source rate, the process with non-random source rate, the process with random source rate and the process with stationary sources. A lower bound to the overflow probability of a finite ATM buffer and a lower bound to the cell loss probability are obtained. It is shown that both probabilities can not decrease faster than hyperbolically when the buffer size increases, in contrast to the traditional Markov models, where the decreasing is faster than exponential.

Some other papers [26, 53, 42, 43] use the Gaussian approximation for the self-similar input traffic to compute the cell loss probability. Below we use propositions from these

papers to get the expression for the cell loss probability for a queuing system with fractional Gaussian noise (FGN) arrival process.

Let $V(t)$ be the virtual waiting time process of the system, which can represent the number of cells in the queue up to time t , assuming an infinite buffer:

$$V(t) = \sup_{s \leq t} (A(t) - A(s) - C(t - s)), \quad (3.1)$$

where $t \in (-\infty, \infty)$, $A(t)$ - the arrival process, C – service rate.

I. Norros [53] shows that the arrival process $A(t)$ can be understood as an analogue to and a generalization of a diffusion approximation for a Poisson process. A proposed diffusion approximation for the number of arrivals in the interval $(0, t]$, considering a heavy traffic Poisson process, is as following:

$$A(t) \cong mt + \sqrt{m} B(t), \quad (3.2)$$

where $m > 0$ - the mean input rate, $B(t)$ - standard Brownian motion ($H = 0.5$).

This follows from

$$\frac{A(t) - mt}{\sqrt{m}} \cong B(t), \quad (3.3)$$

which can be proved by the Central Limit Theorem [54].

Extending (3.2) to the case of self-similar (FGN) process, as suggested in [53], the arrival process $A(t)$, representing the number of arrivals up to time t , can be approximated by

$A(t) \cong mt + \sqrt{am} Z(t)$, where $t \in (-\infty, \infty)$, $Z(t)$ is a normalized fractional Brownian motion with zero mean and variance $|t|^{2H}$, $a > 0$ is a variance coefficient, H is a Hurst parameter of $Z(t)$, $C > m$ is a service rate.

Let $\Delta(n)$ denote the increment of the process $A(t)$ in the time interval $((n-1)T, nT]$. Then

$$\Delta(n) = A(nT) - A((n-1)T) = mT + \sqrt{ma} [Z(nT) - Z((n-1)T)] = mT + \sqrt{ma} G(nT) \quad (3.4)$$

where $G(n)$ - FGN with zero mean and a variance equal to unity.

From (3.4) follows that

$$E[\Delta(n)] = mT, \quad (3.5)$$

$$\text{var}[\Delta(n)] = amE[G^2(nT)] = am|T|^{2H}E[G^2(n)] = am|T|^{2H}. \quad (3.6)$$

For the unit of time interval ($T = 1$)

$$\text{var}[\Delta(n)] = am \quad (3.7)$$

From (3.4) - (3.6) is concluded that $\Delta(n)$, the increment of $A(t)$, is also a FGN with Hurst parameter H , mean mT and variance $ma|T|^{2H}$.

Putting what was obtained for the arrival process to (3.1), we have:

$$V(t) = \sup_{s \leq t} [(m - C)(t - s) + \sqrt{am} (Z(t) - Z(s))] \quad (3.8)$$

Proposition [42, 26]. Given a fractional queuing system with MCR m , variance coefficient a , Hurst parameter H , service rate C and finite buffer size K , the approximation of the cell loss probability is:

$$\Pr(V(t) > K) \cong \exp\left[\frac{-(C - m)^{2H}}{2am[(1 - H)^{1-H} H^H]^2} K^{2-2H}\right], \quad (3.9)$$

$$\text{or } e^{-\frac{(C-m)^{2H}}{2am[(1-H)^{1-H} H^H]^2} K^{2-2H}}.$$

Proof of this proposition is given in [53, 42].

In [26] the cell loss probability approximation is also done for the case when the input process is the composition of N independent FGN processes with different Hurst parameters. The arrival process $A(t)$ was considered as superposition of N independent processes:

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

The obtained cell loss probability in this case is:

$$\Pr(V(t) > K) \geq \max_{t \geq 0} \left[Q \left(\frac{K + (C - \sum_{i=1}^N m_i)t}{\sqrt{\sum_{i=1}^N a_i m_i t^{2H_i}}} \right) \right], \quad (3.10)$$

where $\sigma^2 = \sum_{i=1}^N a_i m_i t^{2H_i}$ and $m = \sum_{i=1}^N m_i$ are the variance and the mean of $A(t)$ respectively.

In [43] the same problem has been addressed, but for a large number N of homogeneous self-similar input traffic streams. The approach is based on the Bahadur-Rao theorem for sums of IID random variables [56]. The proposed approximation of cell loss probability is as follows:

$$\Pr(V(t) > K) \approx \frac{1}{\sqrt{2\pi N} \sigma_i \theta_i} \exp \left[-N^{2H-1} \frac{(c-m)^{2H}}{2am[H^H(1-H)^{1-H}]^2} K^{2-2H} \right], \quad (3.11)$$

where $\theta_i = \frac{k + Ct^* - mt^*}{\sigma^2 (t^*)^{2H}}$; $t^* = \frac{Hk}{(1-H)(C-m)}$; $\sigma_i^2 = \sigma^2 (t^*)^{2H}$; $\sigma^2 = am$ - variance

coefficient; k and c - the amounts of buffer space and bandwidth per source respectively, so $K = N \times k$ and $C = N \times c$.

The approximation (3.11) gives a cell loss probability of about 1 to 1.5 orders closer to the magnitude obtained by simulation, if N , the number of sources, is at least more than twenty.

Chapter 4

State of the Art in Connection Admission Control

Since ATM is a connection-oriented service, before a user starts to transmit over an ATM network, a connection has to be established. This is done at call setup. The main goal of this procedure is to establish a path between the sender and the receiver. This path may involve one or more ATM switches and corresponding resources, allocated to the new connection on each of these switches. When a new connection request is received by the network, the call admission procedure is executed to decide whether to accept or reject the call. Connection admission usually depends on a traffic contract between the user and the network. The user provides the traffic parameters and the QoS he/she wants to have, which are registered in the contract. If the network has enough resources to satisfy new user requirements without affecting the QoS provided to existing connections, a call is accepted; otherwise it is rejected. The user's traffic is policed by the network to ensure that the actual traffic stream does not violate the traffic contract. Two main questions have to be answered: how can the amount of bandwidth required by a new connection be determined, and how can the network be sure that the QoS required by existing connections are not affected by the new connection. Both questions should be taken into account in Connection Admission Control algorithms.

H. Perros and K. Elsayed [44] classify CAC schemes as non-statistical allocation, or peak bandwidth allocation, and statistical allocation.

Peak bandwidth allocation, as follows from its name, requires that a peak bandwidth be reserved at the output port for the specific source, independent of whether or not the source transmits continuously at the peak rate. The advantage of peak bandwidth allocation is that it is easy to decide whether to accept a new connection or not, because it does not need any knowledge of the distribution of the arrival process, only knowledge of the peak rate of the new connection is required. Roughly speaking, the new connection is accepted if the sum of the peak rates of all the existing connections plus the peak rate of the new connection is less than the capacity of the output link. However, because cells from different connections may be interleaved and may momentarily arrive faster than expected, momentarily exceeding the peak rate, it is better to allocate at a peak rate slightly higher than that requested. The obvious disadvantage of peak allocation is that unless connections transmit at peak rates, the output port will be grossly underutilized.

In statistical allocation, allocated bandwidth is less than the peak rate of the source. As a result, the sum of all peak rates may be greater than the capacity of the output link. Statistical allocation gives economy when dealing with bursty sources, but it is difficult to carry out effectively, and the decisions must be made on the fly.

The problem of call admission control has received considerable attention in the literature, and a variety of different call admission schemes were proposed.

In [57] K. Steenhaut et al. consider two different approaches of CAC: the parametric approach, in which the CAC assumes that the cell arrival processes for the contending

connections are completely known via a parametric description of the arrival distribution; and the non-parametric approach, in which the CAC does not know the complete cell arrival process, although it knows some of its characteristics. The *Non-Parametric Approach* [58] may use the peak and average cell rates of connections and does not require any knowledge of the distribution of the arrival process. The call admission procedure is based on an estimate of the blocking probability $U(n, r)$ when there are n multiplexed connections onto the link. If $U(n, r)$ is conservative (i.e., $P_{CLR} \leq U(n, r)$), the call admission procedure is based on first calculating $U(n+1, r)$, which is the CLR after the new connection request is superposed together with n already established connections. If $U(n+1, r)$ is greater than the desired loss ratio, then the new connection is accepted; otherwise it is rejected.

In the “*Gaussian Approximation*” method, each connection is characterized by its average bit rate and standard deviation. With n multiplexed connections, the total bandwidth required by n connections c_0 is determined so that the probability that the instantaneous aggregate bit rate exceeds c_0 is less than a given value, assuming that the aggregate bit rate distribution is Gaussian. This algorithm has a limited use because the aggregate bit rate distribution can be accurately considered as Gaussian only if the stationary distributions of individual connections are themselves Gaussian [58]. But in ATM networks the standard deviations of connections often differ significantly unless only connections with similar characteristics are multiplexed together. Furthermore, in this algorithm all connections are treated as if they have the same cell loss requirements, although in reality they may differ significantly. Also the buffer size is not taken in

consideration in this method, and so the method does not fully explore the amount of statistical gain that can be achieved.

The “*Fast Buffer/Bandwidth Reservation/Allocation*” scheme was devised for the transmission of bursty sources. A bursty source in this scheme is characterized by a two-state Markovian model. The main idea behind this scheme is the following. When a virtual circuit is established, the path through the network is set up and the routing tables are appropriately updated, but no resources are allocated to the virtual circuit. When a source is ready to transmit a burst, at that moment the network attempts to allocate the necessary resources for the duration of the burst. At the end of the active period all reserved resources are released. This process is repeated through the duration of the connection. To decide if a new connection can be multiplexed with the existing connections on a link, the probability of requiring more buffer slots/bandwidth than are available is calculated and referred to as the excess demand probability. If the excess demand probability is greater than a predefined value, the new connection is rejected; otherwise it is accepted [58].

D. Tranchier et al. [59] propose a fast bandwidth allocation protocol for VBR sources whose peak bit rate is less than 2% of the link capacity. In [60, 61] J. Turner introduces a fast reservation scheme where buffer space is allocated rather than bandwidth. B. Doshi and H. Heffes [62, 63] propose a fast buffer allocation scheme for long file transfers.

In the “*Equivalent Capacity*” method, a connection is also characterized by a two-state model in which the flow of bits is generated at a peak rate during the active period whereas no bits are generated during the silent period. Unlike the other techniques discussed so far, this method is based on a flow model [64]. Assuming each source is characterized as a two-state on/off model and the duration at each state is exponentially distributed and independent of each other, the amount of bandwidth required by a connection (referred to as the equivalent capacity) is estimated, given the desired buffer overflow probability. To capture the effect of multiplexing, the Gaussian approximation is used together with the equivalent capacities.

Some studies [64, 65] have indicated the inaccuracy of equivalent capacity methods in some situations. In [66] various approaches for equivalent capacity are compared and some modifications to enhance the accuracy of the scheme are proposed. To overcome the shortcomings of the equivalent capacity for multiplexers that can achieve a substantial statistical gain even with small or no buffers, a method combining Chernoff bounds and equivalent capacity approximation is proposed in [67]. This, however, does not solve all the problems with inaccuracy of equivalent capacity approximation in some other cases. On-line evaluating of equivalent capacity has been discussed in [68, 69]. In [69] N. Duffield uses maximum entropy as a method for characterizing traffic sources and their equivalent capacity. In [70] C.-S. Chang and J. Thomas introduce a calculus for evaluating source equivalent capacity at output of multiplexers and upon demultiplexing or routing. V. Kulkarani et al. [71] consider the equivalent capacity vector for two-

priority on/off source. Some other works (e.g. [40, 72, 73, 74, 75]) are also related with equivalent capacity scheme.

The “*Flow Approximation to Cell Loss Rate*” technique considers a Variable Bit Rate (VBR) source alternating between active and silent periods characterized only by its peak p_i and average m_i bit rates. No assumptions on the distributions of the two periods are required, except that the probability of source i being active (idle) is equal to $\frac{m_i}{p_i(1 - m_i / p_i)}$. This method requires a minimal amount of information to characterize a

source while taking into consideration the burstiness of traffic, but it has two drawbacks: first, it does not use the buffer size in the decision making, thereby restricting the amount of statistical gains that can be achieved, and second, the procedure does not distinguish between the cell loss requirements of individual connections.

The “*Heavy Traffic Approximation*” is based on the asymptotic behavior of the tail of the queue length in an infinite capacity queue with constant service times and a Markovian cell arrival process governed by a probability matrix $P(z)$. The probability $P(\text{queue length} > I) = p$ can be used to make call admission decisions. In [76] K. Sohraby proposes an approximation for bandwidth allocation based on the asymptotic behavior of the tail of the queue-length distribution. The arrival process is general enough to model the superposition of n sources into a queue. If the probability p with the new connection included in the superposed process is less than or equal to the desired loss ratio, then the new connection is accepted; otherwise it is rejected.

Some call admission schemes have been proposed based on an “*Upper Bound for the Cell Loss Probability*” techniques (e.g. [26, 50, 42, 45, 46, 47, 48]). In [50] Saito proposes an upper bound based on the average number of cells that arrive during a fixed interval, and the maximum number of cells that arrive in the same fixed interval. The fixed interval is taken to be a half of the maximum admissible delay in a buffer. A new call is accepted if the resulting upper bound is less than the admissible cell loss probability; otherwise it is rejected.

S. Giordano et al. in [26] and W.-Y. Fu, H. Latchman in [42] present a new approach for CAC which employs the Hurst parameter as an index of burstiness suitable to characterize the self-similar nature of multimedia traffic. These algorithms also assume the knowledge of the mean and the variance of the process. The approximation of the upper bound of the cell loss probability is obtained for the finite size queue and FGN traffic model, and the bandwidth required to accept a connection is computed. If the required bandwidth is less than the available bandwidth, the connection is accepted, otherwise it is rejected. It is shown that the proposed CAC algorithms achieve a larger number of admitted connections than the peak rate allocation schemes.

The CAC algorithm from [77] can be considered as a “*Delay-Based*” algorithm. Its mechanism approximates (through simulation) the end-to-end delay distribution under a particular network model on which CAC is based.

Some CAC algorithms fall in the so-called “*Delay-and-Loss-Based*” category [78]. The CAC algorithm from [79] is based on the time framing strategy to provide guarantees in terms of delay, delay jitter and the upper bounds of the burst and cell loss ratio. The algorithm from [80] assumes that cell arrivals to each input port form a simple Bernoulli process and formulates the CAC problem as a nonlinear combinatorial optimization problem. In [78] J. Hah and M. Yang propose the CAC algorithm which uses a novel estimation method, called quasi-linear dual class correlation (QLDC). All heterogeneous traffic calls are initially categorized into various classes. According to the number of calls in each traffic class, QLDC conservatively and precisely estimates the cell delay and cell loss ratio for each traffic class via simple vector multiplication.

Several CAC schemes are based on the notion that a source is only allowed to transmit up to a maximum number of bits (or cells) within a fixed period of time. This period is known by different names, such as frame and time window. So these schemes are sometimes called “*Time Windows*” methods [44]. In [81] S. Golestani proposes a mechanism by which for each connection the number of cells transmitted on any link in the network is bounded. Each connection can only transmit on a link up to a fixed number of cells per frame; thus the total number of cells transmitted by all connections on the same link is upper bounded. The proposed mechanism gives the possibility to control the number of cells such that for each connection the number of cells per frame transmitted on an outgoing link cannot exceed its upper bound. This scheme requires buffering. The “*Time Windows*” scheme is also proposed in [82]. In [83] F. Vakil and R. Singh describe a node-to-node flow control mechanism. For each connection, the

transmitting node can only transmit up to a certain number of cells every fixed time period. The number of cells can be transmitted is specified by the receiving node, and can be dynamically modified. This method also requires buffering.

Another class of the call admission control schemes is “*Dynamic CAC*”. In this case, the bandwidth allocated to a connection is dynamically adjusted every fixed period of time. Such schemes are investigated in [41, 49, 84, 85, etc.]. In [41, 49] CAC schemes are presented which dynamically shape the traffic on the link and use the burstiness associated with traffic sources to effect dynamic assignment of bandwidth. In the proposed schemes, call acceptance is based on an on-line evaluation of the upper bound on cell loss probability, which is derived from the estimated distribution of the number of calls arriving. Using these schemes, the negotiated quality of service will be assured when there is no estimation error and the number of calls is large.

In [80, 86, 87] CAC is formulated as an optimization problem where particular reward function is optimized. Y.S. Lin and J.R. Yee in [80] consider the problem of determining admission control as a combinatorial optimization problem where the objective is the maximization of the total reward for the admitted sessions where the reward for each session reflects its priority.

A different approach for call admission control has been proposed in [88]. The paper presents a simple and robust ATM call admission control scheme. A source is characterized by its peak rate and cell delay variation tolerance. Acceptance decisions are

based on whether the current load is less than a precalculated threshold, and Bayesian decision theory are used for the choice of threshold. In this model buffers are used for cell-scale congestion, while no buffers are used for burst-level congestion.

CAC schemes for virtual paths have been examined in [57, 89, 90].

In [91] three CAC algorithms (equivalent capacity method, the heavy traffic approximation, and the non-parametric approach) are compared. Based on the results of this comparison, the equivalent capacity method achieves the highest level of statistical multiplexing while guaranteeing the desired application CLR at the link. The heavy traffic approximation becomes effective as the ratio of the buffer size to the burst length increases. If the buffer size is small, then it coincides with the peak bandwidth allocation (i.e., deterministic allocation). It appears that the non-parametric approach is less sensitive to buffer size. As the buffer size goes to infinity (which is not possible in practice), both the equivalent capacity and the heavy traffic approximation converge to the same level of statistical multiplexing. It was also observed that, of the three methods, the non-parametric approach is the most sensitive to CLR, followed by the heavy traffic approximation and the equivalent capacity. As expected, the amount of statistical multiplexing increases as the CLR requirement of the source increases in all three methods. Relatively, in each method, the amount of statistical multiplexing increases the most with the non-parametric approach and the least with the equivalent capacity method.

None of described CAC schemes consider the CAC as a cost minimization problem, where the user tries to establish a connection that satisfies the QoS requirements at a minimum cost. This is done in [4] and in our work.

Call admission alone is not sufficient to prevent congestion, because the users may not stay within the connection parameters negotiated at the call setup phase. So the network must ensure that sources stay within those parameters. Traffic policing or usage parameter control function exists for this purpose. A policing function should detect a non-conforming source as quickly as possible and take appropriate actions to minimize the potentially harmful effect of the excess traffic. Various policing mechanisms have been proposed (e.g. Leaky Bucket scheme, Window-Based Techniques) [58, 68].

Chapter 5

A Brief Review of Network Pricing Schemes

5.1. Pricing Communication Services

The network user's objective is to get the best QoS possible while paying as little money as possible. If the benefit for the user exceeds the cost charged by the network, the user will use that service. Maximizing user benefit might be unreasonable if the network objective is to maximize its profit. But if the network objective is the best service of the users while preventing congestion and degradation of service qualities, charging mechanisms can play an important role in congestion control, resource allocation, and, as a result, better services for users. Traditionally, communication services have been charged using flat pricing schemes, where users are charged a fixed price irrespective of the transmission volume or duration [92]. While this scheme is desirably simple for users, it fails to take advantage of price as a mechanism for congestion control.

A number of researches consider the problem of charging in the networks. S. Jordan and H. Jiang in [93] also give a review of existing pricing schemes for computer networks which is partially used below.

MacKie-Mason [94] introduces a market-pricing scheme for each packet transmitted on the Internet, to control congestion and improve network efficiency. In his scheme each user submits a bid for each packet he/she wants to transmit. The network transmits all packets whose bids exceed the cut-off price. The network sets the cut-off price equal to

the equilibrium price, where demand meets capacity, or to the marginal cost of transmitting one more packet, whichever is applicable. The marginal cost includes non-congestion cost for the network to transmit the packet, and congestion cost. The congestion cost depends on delay and loss of packets during the network congestion. And, in one's turn, the delay depends on the utilization of the resources, which is defined to be total use divided by capacity. So, the whole price depends on the utilization level of the network and on user's demands. In this case it is in the user's interests to provide the true parameter values of his/her traffic stream to the network, since he/she can not manipulate the price to his/her own advantage. In [95] J. MacKie-Mason and H. Varian consider the issue of how this pricing scheme affects industry structure and performance. It is supposed that each producer uses a two-part tariff for pricing: a subscription/attachment fee per user per month, and a usage fee. The considered utility maximization problem for consumer is how to choose the network resource to use and how much to use it. It is found that if the resource is provided in a competitive market with connect fees and usage fees, the equilibrium price and capacity will maximize net social benefits. If there is a monopoly provider, the profit-maximizing usage price could be either higher or lower than their socially optimal price, depending on the value that different users put on the resource.

In [96] a priority pricing policy for multiple service disciplines in computer networks is considered. A service discipline is defined as a function of the user's requests for assigned network services. The network objective is to maximize total user's benefit by setting the right prices. The problem of how to choose the right prices for the network

and the right priorities for the users is treated as a game problem. The decision corresponds to the Nash equilibrium of the game. Nash equilibrium is an equilibrium point in a game where the strategy chosen by each player is the best strategy, given the choices of all other players. Using the simulation for various network configurations it is found that there is a price range for each priority level that will result in a Nash implementation. However, simulations are time-consuming, so more research is required to determine how the optimal pricing can be obtained without them. This approach is similar to the one by MacKie-Mason, described above, because both of them try to allocate network resources among users depending on how much the user values a particular service. The priority pricing scheme is effective when the number of service disciplines is very low.

J. Walrand and P. Varaiya in [97] consider four charging mechanisms: a fixed or access charge, a usage charge, a congestion charge, and a service quality charge. A *fixed* charge is a monthly subscription fee for access to the network. The access may be limited to a certain period of time each day, or it may be unlimited. The fee payment does not depend on how many connections the user actually makes. A *usage* charge depends on the amount of use. It is based on each connection or call the user makes. A *congestion* charge depends on the amount of traffic or load that the network is carrying at the time of the user's connection: the charge is higher when the network is congested, and there is no congestion charge when it is noncongested. The *quality* charge reflects the difference in resource use: the better QoS the user requires, the more network resources are used, the higher quality charge. So theoretically a network user should pay a four-part bill

consisting of a fixed charge for the additional fixed network costs, a usage charge depending on the cost of resource used, a congestion charge that limits less-valued connections, and a quality charge for additional resources needed for higher QoS.

In [98] Parris studies the effect of per-packet pricing, set-up pricing and peak load pricing on network performance, as measured by revenue, call-blocking probability, link utilization, and cost recovery. Simulation results show that network revenue first increases and then decreases while blocking probability only decreases as the per-packet price increases. Set-up pricing causes blocking probability to decrease compared to per-packet pricing. Peak load pricing raises network revenue and decreases blocking probability by smoothing out the network usage.

In practice the situation is not exactly the same as in theory. Pricing schemes in telephone companies include fixed charges, usage charges that depend on distance and call duration, and congestion charges. Pricing schemes for data networks are different. For example, a local area network usually belongs to a single enterprise that provides free access to its members. But as the use and cost of LANs have increased, to cover the cost, a fixed, internal charge per host or per user or department is often imposed. More recent network services such as ISDN, SMDS and Frame Relay include a form of usage charge.

5.2. Pricing in ATM networks

ATM traffic is transmitted in small fixed-length packets (cells). The cells are sent through shared transmission lines and routed by switches with shared buffers. The current ATM network architecture for resource management contains two levels: virtual circuit (VC) and virtual path (VP) [99]. A virtual path is a group of connections sharing a common path from source to destination. Virtual circuits are the individual connections within a virtual path level. Virtual circuits can be grouped into circuit bundles and share bandwidth and buffers based on three variables: path, QoS, and source type. When a call arrives, the network checks whether a virtual path for the source-destination pair exists. If it exists, the network checks if there is enough capacity within the VP for a new VC to accommodate this call. When the call is admitted, cells are delivered in order. If the VP does not exist or there is not enough capacity for the VC within the existing VP, the network can either reject the call, or request more resources (if the VP exists), or create a new VP (if no VP exists). The main network control mechanisms such as routing, resource reservation, congestion control act on the VP level. To avoid congestion and degradation of QoS, admission control and flow control are used. Allocation of bandwidth and buffers for different traffic streams with different QoS should be done in an efficient and fair way, especially when user's demand exceeds the network's capacity. Establishment of the connection for connection-oriented services in high-speed networks such as ATM typically requires a traffic contract between the user and the network. Charging mechanisms and resource management for ATM networks are closely connected, because, on the one hand, charging affects user's behavior and, as a result,

traffic characteristics, and, on the other hand, resource management parameters (e.g. QoS parameters), can be used for charging mechanism.

H. Saito [100] classifies the charging principles into two categories: usage-sensitive and time-sensitive charging. In packet switched networks *usage-sensitive* charging means that the user pays according to the number of packets he/she sends. Even in ATM networks, the user likely prefers usage-sensitive charging. But, for example, the telephone network uses time-sensitive charging. *Time-sensitive* charging means that the user pays according to the length of time the call connection is held open, regardless of the actual time spent talking. Time-sensitive charging does not require usage monitoring, and so it is simpler. If an ATM network uses usage-sensitive charging, its resource management must measure the actual usage of resources assigned to each connection. For example, resource management must use the statistics of actually offered or carried cells or bursts because the ITU-T standard and the ATM Forum UNI specification for the source traffic descriptor only specify the upper bounds on cell rate statistics and cannot specify the actual number of cells. The amount of resources assigned to each connection is determined by the connection admission control (CAC) based on a source traffic descriptor. If this amount is fixed during the whole duration of a call irrespective of the number of generated cells, the time-sensitive charging scheme should be used.

In [101] S. Low and P. Varaiya consider an iterative negotiation process between the user and the network. It is supposed that the network's objective is to maximize the user's benefit, and the user's objective is to maximize his/her consumer surplus by requesting

the best combination of network resources and having satisfying QoS. The network posts prices for its resources (buffer and bandwidth) on each link, and the user decides how much resources to request. Depending on availability of the resources, the network posts the new prices, and so on. Finally the iterative process converges to the optimal solution. There is no need for traffic policing because the network does not need to know detailed individual user traffic characterizations, it is up to the user to choose necessary resources for the desired QoS.

A similar resource-based pricing scheme is proposed in [102]. The composite prices are considered as a weighted sum of reserved network resources such as buffer, bandwidth, CPU time, and delay. But this method requires traffic policing because it allows some sharing among users.

Murphy et al. [103] suggest a distributed pricing policy to allocate bandwidth on VP and VC levels to achieve maximal network efficiency which is defined as a difference between total user benefit and cost. The cost function is convex in the VP's utilization and includes some congestion cost. By setting the price equal to its marginal cost, the network maximizes its objective, and the user chooses the amount of bandwidth based on the posted service price to maximize his/her profit.

In [104, 105, 106] Murphy et al. consider a responsive pricing scheme to control network congestion. In case of high (low) network utilization, the network increases (decreases) the prices. Adaptive users (who are able and willing to influence their network traffic in

response to pricing signals issued by the network), reduce (increase) the offered traffic in response and increase the network efficiency this way.

Kelly [107, 108, 109] proposes a pricing scheme, using an effective bandwidth notion. The pricing scheme given by Kelly is obtained under the assumption of on/off sources. The tariff is described as follows: the network represents a user's effective bandwidth as a function $B(M)$ of the user's input traffic mean rate; before the call admission, the network requires the user to announce a value as his estimate of M . The user then is charged the amount $a_m + b_m M$, where M is the actual measured value of the mean rate, and a_m is a charge per unit time, b_m is a charge per cell carried, are chosen so that $a_m + b_m M$ is tangent to $B(M)$ at m . Kelly proves that declaring the true mean rate ($m = E[M]$), the user minimizes the expected cost per unit time, assuming $B(M)$ is concave and that the minimal expected charge is equal to the user's effective bandwidth. Such a pricing scheme encourages the user to tell the truth about his/her estimate of the input traffic parameters, and so it does not require policing.

5.3. Pricing and Connection Establishment Process

The connection establishment for connection-oriented networks is typically based on a traffic contract between the user and the network. At the beginning of negotiations, the user characterizes the traffic streams that will be transmitted and the desirable QoS. In its turn, the network determines the resource availability. If enough resources are available, the parties agree to set up connections to transmit the agreed traffic streams with agreed QoS guaranteed, and at agreed prices.

In [93] the authors discuss the role of prices in combining user characterization, network resource allocation, and contract negotiation to form a complete connection establishment process. At the beginning of the distributed negotiation process (first stage), both user and network know their objectives. A user has information about his traffic stream and desirable QoS. On the other hand, the network determines its available capacity. The second stage involves negotiations between multiple network and user agents, in which the user chooses the amount of resources needed for his/her desired application, and at market-clearing prices the network and user agree to set up connections to transmit the agreed information in a manner to guarantee the agreed QoS at agreed prices. The notion of effective bandwidth is used in this work as a user traffic characterization and a pricing base, and the total network efficiency is measured by total user benefit. A hierarchical and distributed negotiation structure is proposed under which only hierarchically adjacent and geographically local network entities communicate with each other. The hierarchical structure consists of VCs. Contract negotiation and network maximization are distributed and performed at three levels: user - circuit bundle, circuit bundle - virtual path, virtual path - physical trunk. Each circuit bundle, VP and trunk maximizes its total user benefit as a supplier as it negotiates with lower network levels by setting the right prices. At each level, an equilibrium price is reached and total user benefit is locally maximized. At the first level, negotiation takes place between a user and a circuit bundle; in this case a circuit bundle is a supplier, and the user is a consumer. The user, who wishes to set up a VC, negotiates with the circuit bundle at short intervals periodically. The user declares his/her demands for bandwidth under the desired QoS, and the circuit bundle determines

its availability. At the second level, the VP is a supplier of bandwidth while its circuit bundles are consumers; the market clearing price is again determined by equating demand and supply. The circuit bundle demands bandwidth from the VP at longer intervals. Finally, at the highest level, the physical trunk is a supplier and the VP is a consumer, which demands a bandwidth from the physical trunk; market clearing prices are determined by equating demand and supply at even longer intervals. Consequently, all these negotiations and computations of optimal resource allocations are performed between the user and the network in a distributed manner. The user can incorporate the QoS parameters into its demand because the demand at each step is formulated in terms of effective bandwidth, while effective bandwidth is considered as composed of two parts: mean rate and burstiness. If a user is charged an amount equal to effective bandwidth multiplied by price, this charge includes a fixed price per packet plus an amount based on burstiness. The user could maximize his/her consumer surplus by choosing the optimal effective bandwidth depending on the posted price per effective bandwidth unit and defined as a function of the received QoS of the connection.

In [4, 110], a cost based connection establishment mechanism providing different QoS procedure is proposed. Given the user's desired QoS and the network resources availability, this procedure determines the least-cost parameters for a traffic shaper, which guarantees access to the network while satisfying the QoS constraints. This procedure is done for on/off sources, and the QoS constraints are formulated using an effective bandwidth notation. The connection admission control problem is formulated as a minimization problem for finding the minimum cost values for (σ, ρ) of the policing

function or the traffic shaping function which provide the access to the network for the user and satisfying the QoS requirements of the connection. The notion of (σ, ρ) - constraint [111] gives an upper bound for an arrival process. Let A_n denote the number of cell arrivals per time slot n in a discrete time model. A given arrival process $\{A_n\}$ is (σ, ρ) - constrained in the time interval (k, m) , if

$$\sum_{n=k}^m A_n \leq \rho(m - k + 1) + \sigma ,$$

where $k, m \in \mathbb{Z}^+$

Here ρ and σ can be interpreted as a rate and the burstiness of the bounded process respectively. Note that any arrival process $\{A_n\}$ is bounded above by an infinite family of (σ, ρ) - constraints.

During the CAC procedure, the user declares values for σ and ρ for the input traffic. The objective function of the minimization problem is a cost function depending on effective bandwidth B , σ and ρ . The constraints representing the budget, buffer and bandwidth capacity available, and desirable cell loss probability and the probability of the delay exceeding some values are all functions of B , σ and ρ . It is a network responsibility to determine the appropriate form of the cost function to guarantee a solution [4].

Our work is based on the same principles of the connection establishment process. We address the problem of determining the parameters of a traffic shaper modeled as a single server queue with finite buffer size and deterministic service rate [112]. Self-similar input

traffic is considered in our work. The user and the network prepare the contract, in which they agree on QoS and prices. The user provides basic characteristics of his/her traffic stream and the desired QoS. This gives the network an indication of the anticipated load of the connection. The network then determines its resource availability. If sufficient resources are available to carry the connection, the cost minimization problem is solved to determine the minimum cost parameters of a traffic shaper which will guarantee access to the network while satisfying the QoS demands. After this process, during the connection, the network can police the user traffic and implement flow control, service priority mechanisms and packet multiplexing disciplines.

Chapter 6

Optimal Cost Traffic Shaping Model: Problem Statement and Solution

6.1. Problem Statement

Assume that a user wishes to transmit a self-similar traffic stream with known mean cell rate (m), variance (a) and Hurst parameter (H). The user demands a guarantee on the cell loss probability (CLP_{max}). The network is able to provide this guarantee if the user agrees to voluntarily shape the traffic using a single server queue with finite buffer size K and deterministic service rate C , and if K and C are within the region $K_{min} \leq K \leq K_{max}$ and $C_{min} \leq C \leq C_{max}$ prescribed by the network. This guarantee is obtained by provisioning a virtual circuit with buffer size K and service rate C along the source-destination path. Depending on the services provided by the network, the traffic shaping could alternatively be done by the network, rather than by user.

Under these assumptions, the user can solve the cost minimization problem, the goal of which is to determine the resource requirements for the service rate C and the buffer size K of the traffic shaper as the solution to the following minimization problem:

Minimize

$$Cost = f(K, C) \quad (OBJ)$$

S.t.

$$C_{min} \leq C \leq C_{max} \quad (C1)$$

$$K_{min} \leq K \leq K_{max} \quad (C2)$$

$$CLP(K, C) \leq CLP_{max} \quad (C3)$$

We do not consider the budget constraint because it is an objective function constraint.

In order to incorporate constraint (C3), an expression for the cell loss probability (CLP) as a function of the buffer size and service rate of the traffic shaper is required.

Giordano et al. [26] and Fu [42] determine the following analytical form for the probability that the queue size Q exceeds level K when a single self-similar traffic stream with Hurst parameter H is fed into a single server queue with deterministic service rate C (see Chapter 3):

$$\Pr(Q > K) = e^{-\frac{(C-m)^{2H}}{2am[(1-H)^{1-H}H^H]^2}K^{2-2H}} \quad (6.1)$$

To facilitate numerical minimization, we assume that the objective function (OBJ) is convex in the decision variables C and K . Through our experiments a quadratic cost function of the form $Cost = c_0(C-c_l)^2 + k_0(K-k_l)$ with $c_0 = 1$; $c_l = 0$; $k_0 = 1$; $k_l = 0$ was used. The values of the coefficients c_0 and k_0 are set in accordance with the price per unit of buffer and bandwidth respectively. We thus model a quadratic cost function with global minimum at the origin. Since the cost function is monotonic increasing for $K > 0$ and $C > 0$, the user is charged an increasing amount per additional resource demanded. After that the minimization problem becomes as follows:

Minimize

$$Cost(C, K) = c_0(C-c_l)^2 + k_0(K-k_l)^2 \quad (OBJ)$$

S.t.

$$C_{min} \leq C \leq C_{max} \quad (C1) \quad .(6.2)$$

$$K_{min} \leq K \leq K_{max} \quad (C2)$$

$$CLP \leq CLP_{max} \quad (C3)$$

where $c_0 = 1$; $c_1 = 0$; $k_0 = 1$; $k_1 = 0$ are the coefficients of the quadratic cost function; C_{min} , C_{max} , K_{min} , K_{max} - the minimum and the maximum values of service rate and buffer size; CLP - cell loss probability function (6.1).

Or, rewriting it in a standard form, we have:

$$\underset{x \in R^2}{\text{Minimize}} \quad f_0(\bar{x}) = c_0(x_1 - c_1)^2 + c_2(x_2 - c_3)^2$$

S.t.

$$f_1(\bar{x}) = c_{min} - x_1 \leq 0$$

$$f_2(\bar{x}) = x_1 - c_{max} \leq 0$$

$$f_3(\bar{x}) = k_{min} - x_2 \leq 0$$

$$f_4(\bar{x}) = x_2 - k_{max} \leq 0$$

$$f_5(\bar{x}) = \exp\left[\frac{-(x_1 - m)^{2H} x_2^{2-2H}}{2am((1-H)^{1-H} H^H)^2}\right] - CLP_{max} \leq 0$$

where x_1, x_2 - current service rate and buffer size respectively, c_{min} , c_{max} , k_{min} , k_{max} - minimal and maximum service rate and buffer size respectively, c_0 , c_1 , c_2 , c_3 - constant coefficients, m - mean rate, a - variance, H - Hurst parameter, CLP_{max} - maximal permissible cell loss probability.

Trying to apply the Karush-Kuhn-Tucker (KKT) method [113, 114] to solve this problem we obtain the following.

From Lagrangian function:

$$\begin{aligned} L(\bar{x}, \bar{u}) = & (c_0(x_1 - c_1)^2 + c_2(x_2 - c_3)^2) + u_1(c_{\min} - x_1) + u_2(x_1 - c_{\max}) + u_3(k_{\min} - x_2) \\ & + u_4(x_2 - k_{\max}) + u_5(\exp[-\frac{(x_1 - m)^{2H} x_2^{2-2H}}{2am((1-H)^{1-H} H^H)^2}] - CLP_{\max}), \end{aligned}$$

where u_1, u_2, u_3, u_4, u_5 – Lagrange multipliers.

Write down the KKT conditions:

Gradient conditions

$$\begin{aligned} \frac{\partial L}{\partial x_1} = & 2c_0(x_1 - c_1) - u_1 + u_2 - \frac{u_5}{am} (\exp[-\frac{(1-H)^{-2+2H} H^{-2H} (x_1 - m)^{2H} x_2^{2-2H}}{2am}] * \\ & (1-H)^{2H-2} H^{1-2H} (x_1 - m)^{2H-1} x_2^{2-2H}) \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial x_2} = & 2c_2(x_2 - c_3) - u_3 + u_4 - \frac{u_5}{2am} (\exp[-\frac{(1-H)^{2H-2} H^{-2H} (x_1 - m)^{2H} x_2^{2-2H}}{2am}] * \\ & (2-2H)(1-H)^{2H-2} H^{-2H} (x_1 - m)^{2H} x_2^{1-2H}) \end{aligned}$$

Feasibility

$$f_1(x) = c_{\min} - x_1 \leq 0$$

$$f_2(x) = x_1 - c_{\max} \leq 0$$

$$f_3(x) = k_{\min} - x_2 \leq 0$$

$$f_4(x) = x_2 - k_{\max} \leq 0$$

$$f_5(x) = \exp[-\frac{(x_1 - m)^{2H} x_2^{2-2H}}{2am((1-H)^{1-H} H^H)^2}] - CLP_{\max} \leq 0$$

Orthogonality

$$u_1(c_{\min} - x_1) = 0$$

$$u_2(x_1 - c_{\max}) = 0$$

$$u_3(k_{\min} - x_2) = 0$$

$$u_4(x_2 - k_{\max}) = 0$$

$$u_5(\exp[-\frac{(x_1 - m)^{2H} x_2^{2-2H}}{2am((1-H)^{1-H} H^H)^2}] - CLP_{\max}) = 0$$

Nonnegativity

$$u_1 \geq 0, u_2 \geq 0, u_3 \geq 0, u_4 \geq 0, u_5 \geq 0$$

Next, check whether the functions are convex.

$$f_0(\bar{x}) = c_0(x_1 - c_1)^2 + c_2(x_2 - c_3)^2$$

$$\frac{\partial f_0}{\partial x_1} = 2c_0(x_1 - c_1); \quad \frac{\partial f_0}{\partial x_2} = 2c_2(x_2 - c_3); \quad \frac{\partial^2 f_0}{\partial x_1^2} = 2c_0; \quad \frac{\partial^2 f_0}{\partial x_2^2} = 2c_2;$$

$$\frac{\partial^2 f_0}{\partial x_1 \partial x_2} = \frac{\partial^2 f_0}{\partial x_2 \partial x_1} = 0$$

$$\text{Hessian matrix } H = \begin{bmatrix} 2c_0 & 0 \\ 0 & 2c_2 \end{bmatrix}$$

leading principal minors:

$$|2c_0| > 0, \quad \begin{vmatrix} 2c_0 & 0 \\ 0 & 2c_2 \end{vmatrix} = 4c_0c_2 > 0$$

other principal minors:

$$|2c_2| > 0$$

The function $f_0(\bar{x})$ is strictly convex.

The functions $f_1(\bar{x}), f_2(\bar{x}), f_3(\bar{x}), f_4(\bar{x})$ are linear and thus convex by the definition of a convex function. It is seen from the contour plot on fig. 6.1 that the function $f_5(\bar{x})$ is concave. These facts guarantee that the problem has a global minimum.

Fig. 6.1. Contour Plot of the Feasible Region

6.2. Problem Solution

Since this KKT system does not have an analytical solution, one of the numerical methods of nonlinear programming has to be applied to solve the problem. The theory of nonlinear programming is a huge area, but among the methods that are widely used and actually commercially implemented are genetic algorithms, quadratic programming, geometric programming and different kinds of gradient methods [113, 115].

Genetic algorithms use a semi-random heuristic, and are applicable when the solution space is extremely large, solutions can be expressed as strings, a “score” or object function value can be assigned to any string. It is obviously not our case.

Geometric programming is used when an objective function and each of the constraints

have the form: $f(\bar{t}) = \sum_{i=1}^N u_i(\bar{t})$ with $u_i(\bar{t}) = c_i t_1^{a_{i1}} t_2^{a_{i2}} \dots t_m^{a_{im}}$, where each component t_j of

the variable $\bar{t}$ is positive, the exponents a_{ij} are arbitrary real numbers, and the coefficients c_i are positive (so called “posynomial functions”) [113], and it is not applicable in our case.

Quadratic programming (QP) is mostly used when an objective function is quadratic; the objective function is a concave and we maximize, or the objective function is a convex and we minimize; all constraints are linear.

When the objective function is convex and constraints form a concave/convex set (or, an objective function is concave and constraints form a convex/concave set), gradient based methods will find a global minimum (global maximum).

In our case we need to find the global minimum of a quadratic objective function, which is convex, and under both linear and non-linear, convex and concave inequality constraints. So we are interested in some “universal” NLP method. The chosen optimization routine, which satisfies these requirements, is called DONLP2. That routine

was developed by Peter Spelucci from Technical University at Darmstadt, Germany (see [116] for details of DONLP2). It is suitable for small and medium sized problems and is presently restricted to 1000 unknowns and 2000 constraints.

DONLP2 implements a sequential quadratic programming (SQP) method for general nonlinear optimization problems including nonlinear equality and inequality constraints (generally referred to as the NLP problem).

The purpose of DONLP2 is minimization of an (in general nonlinear) differentiable real function f subject to (in general nonlinear) inequality and equality constraints g, h .

$$f(x) = \min_{x \in S}$$

$S = \{x \in \mathbb{R}^n : h(x) = 0, g(x) \geq 0\}$, where g and h are vector valued functions of dimension NG and NH .

Bound constraints are integrated in the inequality constraints g . These might be identified by a special indicator (“gunit”) in order to simplify calculation of its gradients and also in order to allow a special treatment, known as the gradient projection technique. Also fixed variables might be introduced via h in the same manner.

As long as the gradients of the constraints currently considered as binding are linearly independent, DONLP2 employs an equality constrained QP to determine a direction of descent and a full inequality constrained QP with constraint regularization otherwise. Regularization is achieved by adding an artificial variable for each single constraint and penalizing these constraints.

Bound constraints are treated in a special way by DONLP2. The initial point is corrected automatically to fit the bounds. Thereafter any succeeding point is projected with respect to the bounds. The presence of bound constraints is indicated using gunit-array. For example, for the objective function we have:

`o8gri_1.gunit[0*3] = -1;` (which means that the function is of general form)

`o8gri_1.gunit[0*3+1] = 0;`

`o8gri_1.gunit[0*3+2] = 0;`

and for the $C \geq C_{\min}$ constraint we have:

`o8gri_1.gunit[1*3] = 1;`(which means that the function depends linearly on $X[1]$ (i.e. C_{opt})

`o8gri_1.gunit[1*3+1] = 1;` (because C is $X[1]$ in our program)

`o8gri_1.gunit[1*3+2] = 1;` (because $X[1]$ has the coefficient equal to 1).

The input parameters for the minimization program are the follows:

a - variance of the input traffic,

m - mean rate of the input traffic,

H - Hurst parameter of the input traffic,

$CLP_{\max}$ - maximum cell loss probability (the desirable QoS),

c_0, c_1, k_0, k_1 - the coefficients of the objective function (cost function) parabola,

$C_{\min}, C_{\max}$ - minimum and maximum values of service rate,

$K_{\min}, K_{\max}$ - minimum and maximum values of buffer size.

This scheme is also applicable under relaxed assumptions. For linear cost function the solution to the minimization problem may simplify to direct analytical techniques, such as the Lagrangian method.

Chapter 7

Optimal Cost Traffic Shaping Model: Experimental Results

7.1. Simulation

We considered the simulation of a single buffer queue with a single deterministic server and both a single LRD arrival process and several LRD arrival processes with different parameters (mean, variance and Hurst). For practical purposes the case of a user with a single input process at a time looks more reasonable, so experimental results for the case of a single traffic stream at a time (Fig. 7.1) are considered below.

Fig. 7.1. Single Server Queue Scheme

This work has two main parts: theoretical/simulation results and minimization results. The theoretical part of the minimization problem is described in Chapter 6. Here we present the experimental results. The simulation has been done for: a) validation of the analytical formula: to compare the results on cell loss probability, obtained with the *analytical formula* (5.1) (which gives the approximation of CLP) and used for the constraints in the minimization problem, with the *simulation results* in order to be sure that the cell loss probability obtained during the simulation does not exceed the maximum cell loss probability given for the minimization problem; b) testing the minimization algorithm for different degrees of self-similarity.

In our simulation the generator described in [33] was used (see Chapter 2). It generates Fractional Gaussian Noise traffic (approximately) and is based on a Spatial Renewal Process. We checked this generator by calculating the mean, the variance and the autocorrelation function for the test traces with different Hurst parameters and different number of random generator seed. Figs. 7.2,a-b represent the plots with theoretical autocorrelation functions and experimental ones obtained from the test traces with different Hurst parameters to illustrate the validity of traffic generator.

(a) $H = 0.65$

(b) $H = 0.75$

Fig. 7.2. Empirical and Analytical Autocorrelation Curves

The calculations of cell loss probability have been conducted for different values of the random number generator seed - to make sure that the modeling results do not depend on the particular run of the random number generator. The results obtained for different values of the random number generator seed were very close to each other.

We performed several simulation runs to verify the validity of the analytical formula (5.1) used to characterize the user's constraint on the CLP of the self-similar arrival stream. For this purpose, we generated a single FGN traffic stream with Hurst parameter $H = 0.85$, mean cell rate $m = 500$, variance $a = 2$ using the SRP model described in Chapter 2. We fed this stream into a single server queue with varying service rates and

buffer sizes. Fig. 7.3,a shows the results for the CLP with buffer size K varying from 1 to 200 cells and fixed service rate $C = 540$ cells/s. Fig 7.3,b shows a similar comparison for a fixed buffer size $K = 15$ cells and service rate C varying from 505 to 835 cells /s.

As one can see from Figs. 7.3,a-b, for both cases the simulated CLP is between 1.5 to 3 orders of magnitude lower than the calculated CLP from (5.1). Furthermore, the analytical and simulated CLP functions are sufficiently parallel, indicating that the decay rate of the CLP is correctly captured by the formula. For these reasons we are confident to use the formula (5.1) as an upper bound for the CLP constraint (C3) in the user's minimization problem (5.2).

(a) CLP vs. Buffer Size

(b) CLP vs. Service Rate

Fig. 7.3. Simulated and Analytical CLP

7.2. Minimization Results

To verify the proposed minimization algorithm, we studied the sensitivity of the minimum cost solution to the buffer size K and the service rate C of the traffic shaper with respect to the constraints (C1) and (C2) of the minimization problem (5.2). We base the minimization analysis on the underlying SRP traffic stream with the following parameters: $m = 500$; $a = 2$; $H = 0.85$. In this case, the number of arrivals is $A(t) \cong mt + \sqrt{am} Z(t)$, where $t \in (-\infty, \infty)$, $Z(t)$ is a normalized fractional Brownian motion with zero mean and variance $|t|^{2H}$, $a > 0$ is a variance coefficient, H is a Hurst parameter of $Z(t)$, $C > m$ is a service rate (see Chapter 3). Since we have chosen $m = 500$ and $a = 2$, $\sqrt{am} = \sqrt{1000} \approx 33$; if $t = 1$ (for example), $Z(t)$ varies from 1 to 2 or 3. So the range of cell

arrivals varies approximately from $(500 - 70)$ to $(500 + 70)$. We also assumed the user's maximum tolerated CLP as $CLP_{\max} = 1 \times 10^{-6}$.

In our first experiment, we removed the boundary constraints (C1) and (C2) on K and C which are typically dictated by the network. This provides us with the minimum cost traffic shaping parameters to satisfy the required CLP for the user. We then applied the non-linear programming algorithm DONLP2 to the problem formulation and obtained the following results, denoted by C_{opt} , K_{opt} and Cost_{opt} :

$$C_{\text{opt}} = 609.03 \text{ cells/s; } K_{\text{opt}} = 108.31 \text{ cells; } \text{Cost}_{\text{opt}} = 382650.$$

This result was verified analytically using Mathematica by obtaining the set of pairs (C, K) for which the CLP constraint (C3) of the minimization problem (5.2) is satisfied, and then restricting the objective function (OBJ) to this set. Table 1 was obtained for service rate and buffer size varying from $(C_{\text{opt}} - 10)$ to $(C_{\text{opt}} + 10)$ and from $(K_{\text{opt}} - 10)$ to $(K_{\text{opt}} + 10)$. We computed the analytical CLP using equation (5.1) for different values of C and K and verified that the obtained CLP satisfies constraint (C3). If the pair (C, K) did satisfy this constraint, we calculated the value of the cost function using the objective function (OBJ) expression. Otherwise, we declared the pair (C, K) as infeasible, indicating "N/A" in the cost column of the Table 1. As before, $m = 500$; $a = 2$; $CLP_{\max} = 1.00 \times 10^{-6}$.

Table 1
Analytical Verification of the Minimization Problem

C	K	CLP	COST
599	98	$1.14 \times 10^{-5} > \text{CLP}_{\max}$	N/A
599	108	$8.17 \times 10^{-6} > \text{CLP}_{\max}$	N/A
599	118	$5.96 \times 10^{-6} > \text{CLP}_{\max}$	N/A
605	107	$2.47 \times 10^{-6} > \text{CLP}_{\max}$	N/A
605	109	$2.3 \times 10^{-6} > \text{CLP}_{\max}$	N/A
609	98	$1.51 \times 10^{-6} > \text{CLP}_{\max}$	N/A
609	107	$1.05 \times 10^{-6} > \text{CLP}_{\max}$	N/A
609	109	9.8×10^{-7}	382762
609.03	108.31	1×10^{-6}	382650
619	98	1.75×10^{-7}	392765
619	118	7.19×10^{-8}	397085

Fig. 7.4 also verifies the optimal solution obtained from the non-linear programming algorithm DONLP2. It shows the value of the cost function for all points that satisfy the CLP constraint where K and C are in the ranges $108 \leq K \leq 110$ and $608.5 \leq C \leq 610$.

Note that the cost increases strictly with K and C within the restricted region.

Fig.7.4. Value of the Cost Function Restricted to the Feasible Region

In our second experiment, we imposed constraints (C1) and (C2) of the minimization problem (5.2) for different values of $C_{\min}$, $C_{\max}$, $K_{\min}$ and $K_{\max}$. We chose the values for the constraints to create feasible regions excluding the optimal solution (K_{opt} , C_{opt}). Furthermore, we moved the feasible region further and further away from the optimal solution. In this way, we expected the minimization algorithm to converge to a boundary solution, satisfying the constraints (C1) and (C2).

The results in Table 2 indicate that the minimization algorithm indeed converges to a boundary solution, satisfying the lower bound conditions of the constraints (C1) and (C2). As expected, the value of the cost function rises the further we depart from the optimal solution.

Table 2
Minimization Results (the Optimal Solution is not Included)

$C_{\min}$	$C_{\max}$	$K_{\min}$	$K_{\max}$	C	K	Cost
620	720	105	205	620	105	395000
720	820	205	305	720	205	560000
920	1000	200	300	920	200	886000

In our next experiment, for validation of the minimization algorithm, we fixed the range of the buffer size K to be $10 \leq K \leq 5000$, thus allowing the optimal solution K_{opt} to be reached. However, we now spliced the C -dimension to create feasible regions that mostly excluded the optimal value of C .

Table 3 also confirms the validity of our minimization algorithm. For the range $540 \leq C \leq 600$, the solution for C is at the upper bound $C_{\max}$. For all ranges of $C_{\min} > C_{\text{opt}}$, the solution for C is at the lower bound $C_{\min}$. Moreover, the results indicate how buffer is traded in for service to minimize the cost. As the solution for the service rate C increases due to the constraints on $C_{\min}$ and $C_{\max}$, the value for the buffer size K decreases. The network forces the user to demand a service rate that is higher than the one the user would select without constraint (C1). In return, the user demands less buffer size while still satisfying the CLP constraint (C3). Note that the value of the cost function is higher for all ranges where the optimal solution C_{opt} is not included. Only in the range $550 \leq C \leq 650$ the optimal solution $C_{\text{opt}} = 609.03$ is included, and the cost function is at its global minimum. Furthermore, since we have chosen a quadratic cost function increasing

strictly in C and K , the value of the objective function also increases strictly, the further we depart from the optimal solution.

Table 3
Minimization Results for Different Feasible Regions Spliced in the Service-Dimension

$C_{\min}$	$C_{\max}$	C	K	Cost
540	560	560.00	3195.93	10527597
560	580	580.00	626.05	728343
580	600	600.00	176.79	391254
600	620	609.03	108.31	382650
620	640	620.00	62.92	388358
640	660	640.00	26.26	410290
660	680	660.00	12.32	435752
680	700	680.00	10.00	442500

We repeated the last experiment allowing the service rate C to lie in the range $540 \leq C \leq 1000$ (thus including the optimal value C_{opt}). We now spliced the K -dimension to regions of width 20, thus excluding the optimal value of K_{opt} for most of the resulting feasible regions. These regions are constructed both above and below the optimal value K_{opt} , thus again forcing the algorithm to converge to the upper bounds of $K_{\min}$ and $K_{\max}$ respectively. The results of this experiment are presented in the Table 4.

Table 4
Minimization Results for Different Feasible Regions Spliced in the Buffer-Dimension

$K_{\min}$	$K_{\max}$	C	K	Cost
40	60	621.01	60.00	389253
60	80	615.02	80.00	386649
80	100	610.58	100.00	382805
100	120	609.03	108.31	382650
120	140	607.08	120.00	382942
140	160	604.20	140.00	384662
160	180	601.78	160.00	387735

Table 4 illustrates that the expected boundary solutions are indeed reached by our minimization algorithm. For ranges of K which lie below the optimal value of $K_{\text{opt}} = 108.031$, the solution converges to the upper bound $K_{\max}$ of the feasible region for K . For ranges of K which lie above the optimal value K_{opt} , the solution converges to the lower bound $K_{\min}$. The buffer/bandwidth tradeoff is seen again. As the network forces the user to demand more buffer than K_{opt} , the service rate required for satisfying of the CLP constraint decreases. The cost function is again higher for all regions that exclude the optimal solution $(K_{\text{opt}}, C_{\text{opt}})$.

7.3. Combined Minimization and Simulation Results

We conducted further simulation tests. In the first test, the service rate and the buffer size were varied from $C_{\text{opt}} - 10$ ($\cong 599$) to $C_{\text{opt}} + 10$ ($\cong 619$) and from $K_{\text{opt}} - 10$ ($\cong 98$) to $K_{\text{opt}} + 10$ ($\cong 118$); as before, the mean cell rate $m = 500$ cells/s, the variance $a = 2$ and the

$CLP_{\max} = 1.0 \times 10^{-6}$. The duration of the simulation was 10^6 . In all cases the cell loss probability obtained during the simulation was 0.0 because the events of losses under given service rates and buffer sizes are rare. For this reason, in the next test the $CLP_{\max}$ was increased.

Next, our algorithm was tested for different degrees of self-similarity. In this case, we generated traffic streams with Hurst parameters $H = 0.65$, $H = 0.75$, $H = 0.85$ and $H = 0.95$, mean cell rate $m = 500$ cells/s and variance $a = 2$. We fixed the maximum tolerated CLP at $CLP_{\max} = 1.0 \times 10^{-1}$; $CLP_{\max} = 1.0 \times 10^{-2}$; $CLP_{\max} = 1.0 \times 10^{-3}$ and $CLP_{\max} = 1.0 \times 10^{-4}$, and the allowable ranges for C and K at $520 \leq C \leq 620$ and $10 \leq K \leq 500$ respectively. We used the same cost function as above and applied the minimization algorithm DONLP2. The results of this experiment for different values of $CLP_{\max}$ are shown in Tables 5, 6, 7 and 8.

Table 5
A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{\max} = 1.0 \times 10^{-1}$)

Hurst	C_{opt}	K_{opt}	Cost	CLP_{sim}
0.65	522.80	80.89	279864	9.4×10^{-4}
0.75	531.11	74.27	287596	8.2×10^{-4}
0.85	541.79	63.19	297533	6.5×10^{-4}
0.95	556.57	40.66	311422	4.6×10^{-4}

Table 6

A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{max} = 1.0 \times 10^{-2}$)

Hurst	C_{opt}	K_{opt}	Cost	CLP_{sim}
0.65	534.67	99.98	295870	8.9×10^{-5}
0.75	546.04	91.68	306565	7.2×10^{-5}
0.85	560.60	77.60	320290	5.2×10^{-5}
0.95	580.61	49.78	339586	3.0×10^{-5}

Table 7

A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{max} = 1.0 \times 10^{-3}$)

Hurst	C_{opt}	K_{opt}	Cost	CLP_{sim}
0.65	544.15	113.91	309075	7.0×10^{-6}
0.75	557.88	103.83	322008	6.0×10^{-6}
0.85	575.31	87.45	338632	5.0×10^{-6}
0.95	599.18	55.86	362140	2.0×10^{-6}

Table 8

A Comparison Between Simulated and Analytical CLP for Input Streams with Different Hurst Parameters ($CLP_{max} = 1.0 \times 10^{-4}$)

Hurst	C_{opt}	K_{opt}	Cost	CLP_{sim}
0.65	552.42	124.89	320768	1.0×10^{-6}
0.75	568.04	113.60	335578	1.0×10^{-6}
0.85	587.83	95.46	354657	0.0
0.95	614.88	60.86	381778	0.0

As can be seen from Tables 5, 6, 7 and 8, that, in this case, the optimal C_{opt} and K_{opt} are not at one of the boundary conditions C_{min} , C_{max} , K_{min} or K_{max} . This indicates that the solution represents the global minimum cost that satisfies the CLP constraint (C3). Furthermore, as could be expected, the minimum cost solution increases with the Hurst parameter, indicating that *traffic with a higher degree of self-similarity requires more resources from the network to guarantee the specified CLP*.

To show that the simulated CLP is indeed below the $CLP_{max}=1.0 \times 10^{-2}$, the simulation was conducted for service rate values from 520 to 620 and buffer sizes from 50 to 150 with step size equal to 10 for both service rate and buffer size. The value of mean cell rate and variance were the same as in previous tests, and the values of Hurst parameter were $H = 0.65$, $H = 0.75$ and $H = 0.85$. The results are shown on Figs. 7.5,a-c, from which one can see that simulated cell loss probability is at least two times less than CLP_{max} .

(b) $H = 0.75$

(c) $H = 0.85$

Fig. 7.5. Simulated CLP ($CLP_{\max} = 1.0 \times 10^{-3}$)

Chapter 8

Conclusions and Recommendations for Future Research

8.1. Conclusions

We presented a traffic shaping methodology for self-similar user traffic.

The issue of how to determine the parameters for a traffic shaper during connection admission control was addressed.

In our approach, traffic shaping involves the minimization of the cost of the connection in terms of the network resources (bandwidth and buffer), such that access to the network is obtained and the desired QoS is provided by the connection.

The problem of traffic shaping is formulated for self-similar input streams, when the input stream parameters such as mean cell rate, variance and degree of self-similarity (Hurst parameter) are given by the user, and the user's demand on maximum tolerated cell loss probability is guaranteed. In its turn, the network provides the upper bounds on the available resources, such as bandwidth and buffer, as well as the form of the cost function by which the user is charged.

The analytical expression for an upper bound of the user's constraint on cell loss probability as a function of service rate and buffer size of the traffic shaper for the self-similar arrival stream was chosen. Several simulation runs were performed to verify the validity of the analytical formula chosen. It was shown that simulated CLP is between 1.5 and 3 orders of magnitude lower than CLP calculated with analytical formula.

The problem was formulated as a cost minimization problem. The quadratic shape of the cost function was chosen, and it was shown that the problem had a global minimum.

Different tests were conducted which verified the validity of proposed minimization algorithm.

The method for solving the considered non-linear programming problem was proposed and tested.

The proposed traffic shaping scheme was tested for a number of input streams with different Hurst parameters.

The conducted simulation runs showed that the behavior of cost minimization problem solution for different Hurst parameters was, as could be expected from the theory: the minimum cost solution increases with Hurst parameter, indicating that traffic with a higher degree of self-similarity requires more resources from the network to guarantee a specified CLP.

8.2. Recommendations for Future Research

The presented traffic shaping methodology can be extended to include delay constraint to the minimization problem while retaining the assumption that the input traffic exhibits self-similarity.

Further development of self-similar traffic theory could lead to using alternative forms for representing the CLP for self-similar traffic streams, provided the CLP remains a function of the decision variables (bandwidth and buffer). This would lead to a better approximation of cell loss probability.

Since it was shown that simulated CLP is between 1.5 and 3 orders of magnitude lower than the CLP calculated with analytical formula, it would be a good idea to compare the simulation results with the real data results, obtained for the same traffic parameters. And if the results would be close enough, the analytical formula for CLP could be multiplied by the constant coefficient which is equal to $\min[\frac{CLP_{simulated}}{CLP_{analytical}}]$. It would result in saving user's money and network's resources.

The coefficients of the cost function (the objective function in our minimization problem) could be changed. If, for example, buffer price is greater than bandwidth price, the variable in the cost function standing for buffer size could be multiplied by larger coefficient than the one used for bandwidth.

The proposed methodology could be extended for the case of multiple input sources, which would permit to use the benefits of statistical multiplexing.

REFERENCES

1. M. Prycker. *Asynchronous Transmission Mode; Solution for Broadband ISDN*. 3^d ed. London. Prentice Hall. 380 pp.1995.
2. M. Sexton. *Broadband Networking, ATM, SDH, and SONET*. Boston. Massachusetts. Artech House. 589 pp. 1997.
3. O. Kyas. *ATM Networks*. 2^d ed. London. International Thompson Computer Press. . 433 pp. 1996.
4. M.Falkner, M.Devetsikiotis, I.Lambadaris. *A Cost-Based Connection Admission Control Procedure*. In: Proc. IEEE ICC, June 1999, to appear.
5. J.Beran. *Statistics for Long-Memory Processes*. Chapman and Hall, New York. 315 pp. 1994.
6. J.Beran, R.Sherman, M.S.Taqqu, W.Willinger. *Long-Range Dependence in Variable-Bit-rate Video Traffic*. In: IEEE Trans. Comm. Vol. 43, No. 2-4, pp. 1566-1579. 1995.
7. W.E.Leland, M.S.Taqqu, W.Willinger, D.V. Wilson. *On the Self-Similar Nature of Ethernet Traffic*. In: Proc.ACM Sugcomm'93. San Francisco. CA, pp. 183-193. 1993.
8. W.E.Leland, M.S.Taqqu, W.Willinger, D.V.Wilson. *On the Self-Similar Nature of Ethernet Traffic (Extended Version)*. In: IEEE/ACM Trans. on Networking. Vol. 2, No. 1, pp. 1-15. 1994
9. V.Paxson, S.Floyd. *Wide-Area Traffic: the Failure of Poisson Modeling*. In: IEEE/ACM Trans. on Networking. Vol. 3, No. 3, pp. 226-244. June 1995.

10. W.Willinger, M.S.Taqqu, A.Erramilli. *A Bibliographical guide to Self-Similar Traffic and Performance Modelling for Modern High-Speed Networks*. In: F.P.Kelly, S.Zachary, I.Ziedins, Eds. *Stochastic Network: Theory and Applications*. Clarendon Press (Oxford University Press). Oxford, pp. 336-366. 1996.
11. M.Crovella, A.Bestavros. *Self-Similarity in World Wide Web Traffic Evidence and Possible Causes*. In: Proc. of the 1996 ACM SIGMETRICS International Conference on Measurements and Modeling of Computer Systems, pp. 160-169. May 1996.
12. H.E.Hurst. *Long-Term Storage Capacity of Reservoirs*. In: Trans. Amer. Soc. Civil Eng. Vol. 116, pp. 770-799. 1951.
13. D.R.Cox. *Long-Range Dependence: a Review*. In: Statistics: An Appraisal, Proc. 50th Annivesary Conf. Iowa State Statistical Library. H.A.David, H.T.David, Eds. Iowa State Univ. Press, pp. 55-74. 1984.
14. D.R.Cox, V.Isham. *Point Processes*. London/New York. Chapman and Hall. 188 pp. 1980.
15. W.Willinger, M.S.Taqqu, W.Leland, D.Wilson. *Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements*. In: Statistical Science. Vol. 10, No. 1, pp. 67-85. Feb. 1995.
16. J.Beran, R.Sherman, M.S.Taqqu, W.Willinger. *Variable-Bit-Rate Video Traffic and Long-Range Dependence*. In: IEEE/ACM Trans. on Communications. Vol. 43, No. 2-4, pp. 1566-1579. 1995.
17. N.Likhanov, B.Tsybakov, N.D.Georganas. *Analysis of and ATM Buffer with Self-Similar (Fractal) Input Traffic*. In: Proc. IEEE INFOCOM'95. Boston. MA, pp. 985-992 1995.

18. B.Tsybakov, N.D.Georganas. *On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound and Cell Delay Distribution*. In: IEEE/ACM Trans. on Networking. Vol. 5, No. 3, pp. 379-409. 1997.
19. B.Tsybakov, N.D.Georganas. *Self-Similar Traffic and Upper Bounds to Buffer-Overflow Probability in an ATM Queue*. In: Performance Evaluation. 32, pp. 57-80. 1998.
20. J.Beran. *Statistical Methods for Data with Long-Range Dependence; with discussion*. In: Statistical Science. Vol. 7, No. 4, pp. 404-427. 1992.
21. G.Samorodnitsky, M.Taqqu. *Stable Non-Gaussian Random Processes*. Chapman and Hall. 656 pp. 1994.
22. M.Garrett, W.Willinger. *Analysis, Modeling and Generation of Self-Similar VBR Video Traffic*. In: Proc. of SIGCOMM'94. Sept. 1994.
23. P.Brockwell, R.Davis. *Time Series: Theory and Methods*. In: Springer Series in Statistics. Springer-Verlag. 2nd ed. 577 pp. 1991.
24. V.Paxson. *Fast, Approximate Synthetic of Fractional Gaussian Noise for Generating Self-Similar Network Traffic*. In: ACM SIGCOMM. Computer Communication Review, pp. 5-18. 1997.
25. C.Huang, M.Devetsikiotis, I.Lambadaris, A.R.Kaye. *Modeling and Similation of Self-Similar Variable Bit RateCompressed Video: A Unified Approach*. In: Proc of SIGCOMM'95. Cambridge. MA USA, pp. 114-125. 1995.
26. S.Giordano, M.Pagano, R.Pannocchia, F.Russo. *A New Call Admission Control Scheme Based on the Self Similar Nature of Multimedia Traffic*. In: IEEE ICC, pp. 1612-1618. 1996.

27. W.Lau, A.Erramilli, J.L.Wang, W.Willinger. *Self-Similar Traffic Parameter Estimation: A Semi-Parametric Periodogram-Based Algorithm*. In: Proc. IEEE Globecom, pp. 2225-2231. 1995.
28. A.Adas, A.Mukherjee. *On Resource Management and QoS Guarantees for Long Range Dependent Traffic*. In: Proc. IEEE Infocom, pp. 779-787. 1995.
29. J.R.M.Hosking. *Fractional Differencing*. *Biometrika*. Vol. 68, No. 1, pp. 165-176. 1981.
30. C.Huang, M.Devetsikiotis, I Lambadaris, A.R.Kaye. *Fast Similation for Self-Similar Traffic in ATM Networks*. In: Proc. IEEE ICC, pp. 438-444. 1995.
31. B.B.Mandelbrot, J.R.Walls. *Computer Experiments with Fractional Gaussian Noises*. In: Water Resources Research. 5, pp. 228-267. 1969.
32. J.Gordon. *Long Range Correlation in Multiplexed Pareto Traffic*. In: Proc. of International IFIP-IEEE Conf. on Broadband Communications, pp. 28-39. 1996.
33. T.Taralp, M.Devetsikiotis, I.Lambadaris, A.Bose. *Efficient Fractional Gaussian Noise Generation Using the Spatial Renewal Process*. In: Proc. IEEE ICC. 1998.
34. W.Lau, A.Erramilli, J.L.Wang, W.Willinger. *Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and its Properties*. In: Proc. IEEE ICC, pp. 466-472. 1995.
35. P.Pruthi, Erramilli. *Heavy-Tailed ON/OFF Source Behavior and Self-Similar Traffic*. In: Proc. IEEE ICC, pp. 445-450. 1995.

36. G.Mayor, J.Silvester. *Simulation of an ATM Queueing System with a Fractional Brownian Noise Arrival Process*. In: Proc. IEEE ICC, pp. 1607-1611. 1996.
37. P.R.Jelenkovic, A.A.Lazar, N.Semret. *Multiple Time Scales and Subexponentiality in MPEG Video Streams*. In: Proc. of International IFIP-IEEE Conf. on Broadband Communications, pp. 64-75. 1996.
38. A.Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley Publishing Company. Don Mills. Ont. 2^d ed. 596 pp. 1994.
39. A.Tannenbaum. *Computer Networks*. 3^d ed. Upper Saddle River. N.J. Prentice Hall PTR. 813 pp. 1996.
40. R.Guerin, H.Ahmadi, M.Naghshineh. *Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks*. In: IEEE Journal on Selected Areas in Communications. Vol. 9, pp. 968-981. Sept. 1991.
41. S.Srinidhi, W.Thesling, V.Konangi. *An Adaptive Scheme for Admission Control in ATM Networks*. In: Computer Networks and ISDN Systems. 29, pp. 569-582. 1997.
42. W-Y.Fu, H.Latchman. *A Self-Similarity Based Call Admission Control Scheme for Output Buffered ATM Switches*. In: 7th IEEE Workshop on Computer-Aided Modeling, Analysis and Design of Communication Networks and Links (CAMAD'98). Rio de Janeiro. August 1998.

43. Z.Fan, P.Mars. *Accurate Approximation of Cell Loss Probability for Self-Similar Traffic in ATM Networks*. In: Electronics Letters. Vol. 32, No. 19, pp. 1749-1751. Sept. 1996.
44. H.Perros, K.Elsayed. *Call Admission Control Schemes: A Review*. In: IEEE Communications Magazine, pp. 82-91. Nov. 1996.
45. C.Rasmussen et al. *Source-Independent Call Acceptance Procedures in ATM Networks*. In: IEEE JSAC. Vol. 9, pp. 351-358. 1991.
46. P.Castelli, E.Cavallero, A.Tonetti. *Policing and Call Admission Problems in ATM Networks*. In: A.Jensen, V.B.Iversen, Eds. Teltraffic and Datatraffic in a Period of Change. North-Holland, pp. 847-852. 1991.
47. B.T.Doshi. *Deterministic Rule Based Traffic Descriptors for Broadband ISDN: Worst Case Behavior and Connection Acceptance Control*. In: Proc. 14th International Teletraffic Congress (ITC'94), pp. 591-600. 1994.
48. A.Elwalid, D.Mitra, R.H.Wentworth. *A New Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node*. In: IEEE JSAC. Vol. 13, pp. 1115-1127. 1995.
49. H.Saito, K.Shiomoto. *Dynamic Call Admission Control in ATM Networks*. In: IEEE Journal on Selected Areas in Communications. Vol. 9, No. 7, pp. 982-989. Sept. 1991.
50. H.Saito. *Call Admission Control in an ATM Network Using Upper Bound of Cell Loss Probability*. In: IEEE Trans. in Communications. Vol. 40, pp. 1512-1521. 1992.

51. H.Saito. *New Dimensioning Concept for ATM Networks*. In: Proc. 7th International Teletraffic Cong. Sem. 1990.
52. K.Mattes, J.Kerstan, J.Mecke. *Infinitely Divisible Point Processes*. Wiley. London. 1978.
53. I.Norros. *A Storage Model with Self-Similar Input*. In: Ququing Systems. Vol. 16, pp. 387-396. 1994.
54. J.Walrand. *An Introduction to Queuing Networks*. Englewood Cliffs. N.J. Prentice-Hall, 384 pp. 1988.
55. N.G.Duffield, N.O'Connell. *Large Deviations and Overflow Probabilities for the General Single-Server Queue, with Applications*. In: Mathematical Proc. of Cambridge Philosophical Soc. Vol. 118, pt. 2, pp. 363-374. Sept. 1995.
56. A.Dembo, O.Zeitouni. *Large Deviations Technique and Applications*. Jones and Barlett. Boston. 360 pp. 1992.
57. K. Steenhaut, K. Degieter, W. Brissinck, E. Dirckx. *Scheduling and Admission Control Policies: a Case Study for ATM*. In: Computer Networks and ISDN Systems. Vol. 29, pp. 539-554. 1997.
58. R. Onvural. *Asynchronous R Transfer Mode Networks*. 2nd ed., 535 pp. 1995.
59. D.Tranchier et al. *Fast Bandwidth Allocation in ATM Networks*. Tech. Rep. CNET-Lannion. 1992.
60. J. Turner. *A Proposed Bandwidth Management and Congestion Control Scheme for Multicast ATM Networks*. Tech. Rep. In: Comp. and Commun. Res. Ctr. Washington Univ. 46 pp. 1991.

61. J. Turner *Bandwidth Management in ATM Networks Using Fast Buffer Reservation*. In: Proc. Australian Broadband Switching and Services Symp. Melbourne. Australia. July 15-17. pp. 279-290. 1992.
62. B. Doshi, H. Heffes. *Overload Performance of an Adaptive Buffer-Window Allocation Scheme for a Class of High Speed Networks*. In: A.Jensen, V.B.Iversen, Eds. *Teltraffic and Datatrafic in a Period of Change*. North-Holland, pp. 441-446. 1991.
63. B.Doshi, H.Heffes. *Performance of an in-Call Buffer-Window Reservation/Allocation Scheme for Long File Transfers*. In: IEEE JSAC. Vol. 9, pp. 1013-1023. 1991.
64. G. Choudhury, D. Lucantoni, W. Whitt. *On the Effective Bandwidths for Admission Control in ATM Networks*. In: Proc. 14th Int. Teletraffic Congress (ITC'94), pp. 411-420. 1994.
65. K. Elsayed, H. Perros. *Analysis of and ATM Statistical Multiplexer with Heterogenous Markovian On/Off Sources and Applications to Call Admission Control*. In: J. High Speed Networks. 1997.
66. K. Rege. *Equivalent Bandwidth and Related Admission Criteria for ATM Systems – a Performance Study*. In: Int. J. Commun. Sys. Vol. 7, pp. 181-197. 1994.
67. A. Elwalid. *Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing*. In: IEEE JSAC. Vol. 13, pp. 1004-1016. 1995.

68. G. De Veciana, G. Kesidis, J. Walrand. *Resource Management in Wide-Area ATM Networks Using Effective Bandwidth*. In: IEEE JSAC. Vol. 13, pp. 1081-1090. 1995.
69. N. Duffield et al. *Entropy of ATM Traffic Streams: a Tool for Estimating QoS Parameters*. In: IEEE JSAC. Vol. 13, pp. 981-990. 1995.
70. C.-S. Chang, J. Thomas. *Effective Bandwidth in High Speed Networks*. In: IEEE JSAC. Vol. 13, pp. 1091-1100. 1995.
71. V. Kulkarani, L. Gun, P. Chimento. *Effective Bandwidth Vector for Two-Priority ATM Traffic*. In: Proc. INFOCOM'94, pp. 1056-1064. 1994.
72. F. Kelly. *Effective Bandwidths at Multi-Class Queues*. In: Queuing Sys. Vol. 9, pp. 5-16. 1991.
73. R. Gibbens, P. Hunt. *Effective Bandwidth for Multi-Type UAS Channel*. In: Queuing Systems Vol. 9, pp. 17-26. 1991.
74. G. Kesidis, J. Walrand, C.-S. Chang. *Effective Bandwidth for Multiclass Markov Fluids and Other ATM Sources*. In: IEEE Trans. in Networking. Vol. 1, No. 4, pp. 424-428. Aug. 1993.
75. R. Guerin, L. Gun. *A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet-Switched Networks*. In: Proc. INFOCOM'92, pp. 1-12. 1992.
76. K. Sohraby. *On the Asymptotic Behavior of Heterogeneous Statistical Multiplexer with Applications*. In: Proc. INFOCOM'92, pp. 839-847. 1992.

77. D. Yates et al. *On pre-Session End-to-End Delay and the Call Admission Problem for Real-Time Applications with QoS Requirements*. In: ACM SIGCOMM'93, pp. 2-12. 1993.
78. J. Hah, M. Yuang. *Estimation-Based Call Admission Control with Delay and Loss Guarantees in ATM Networks*. In: IEE Proc.-Communications. Vol. 144, No. 2, pp. 85-92. Apr. 1997.
79. A. Dailianas, A. Bovopoulos. *Real-Time Admission Control Algorithm with Delay and Loss Guarantee in ATM Networks*. In: IEEE INFOCOM'94, pp. 1065-1072. 1994.
80. F. Lin, J. Yee. *A Real-Time Distributed Routing and Admission Control Algorithm for ATM Networks*. In: IEEE INFOCOM'93, pp. 792-801. 1993.
81. S. Golestani. *Congestion-Free Communication in Broadband Packet Networks*. In: IEEE Trans. in Communications. Vol. 39, pp. 1802-1812. 1991
82. T. Faber, L. Landweber. *Dynamic Time Windows: Packet Admission Control with Feedback*. In: Proc. SIGCOMM'92, pp. 124-135. 1992.
83. F. Vakil, R. Singh. *Shutter: a Flow Control Scheme for ATM Networks*. In: Proc. 7th ITC Spec. Sem. Morristown. NJ. Oct. 1990.
84. T. Tedijanto, L. Gun. *Effectiveness of Dynamic Bandwidth Management Mechanisms in ATM Networks*. In: Proc. INFOCOM'93, pp. 358-367. 1993.
85. R. Bolla et al. *An Integrated Dynamic Resource Allocation Scheme for ATM Networks*. In: Proc. INFOCOM'93, pp. 1288-1297. 1993.
86. L. Gun, V. Kulkarni, A. Narayanan. *Bandwidth Allocation and Access Control in High-Speed Networks*. In: Methodologies for High Speed Networks. R.

- Onvural, H. Perros, G. Pujolle, Eds. *Annals. Of Oper. Res.* Vol. 49, pp. 161-183. 1994.
87. S. Evans. *Optimal Resource Management and Capacity Allocation in a Broadband Integrated Services Network*. P. King, I. Mitrani, R. Pooley, Eds. In: *Performance'90*. Elsevier, pp. 159-173. 1990.
88. R. Gibbens, F. Kelly, P. Key. *A Decision-Theoretic Approach to Call Admission Control in ATM Networks*. In: *IEEE JSAC*. Vol. 13, pp. 1101-1113. 1995.
89. Y. Sato, K. Sato. *Evaluation of Statistical Cell Multiplexing Effects and Path Capacity Design in ATM Networks*. In: *IEICE Trans. in Communications*. E75-B, pp. 642-648. 1992.
90. K. Sato, S. Ohta, I. Tokizawa. *Broadband ATM Network Architecture Based on Virtual Paths*. In: *IEEE Trans. in Communications*. Vol. 38, pp. 1212-1222. 1990.
91. K.M. Elsayed, H.G. Perros. *Call Admission Control in High Speed Networks*. Tech. Report. Comp. Science Department. North Carolina State University. 1994.
92. L.Anania, R.J.Solomon. *Flat – the minimalist price*. In: *Internet Economics*. L.W. McKnight and J.P.Bailey, Eds. MIT Press. Cambridge, pp. 91-118. 1997.
93. S.Jordan, H.Jiang. *Connection Establishment in High-Speed Networks*. In: *IEEE Journal on Selected Areas in Communications*. Vol. 13, No. 7, pp. 1150-1161. Sept. 1995.

94. J.K.MacKie-Mason, H.R.Varian. *Pricing the Internet*. In: Proc. 2nd International Conf. On Telecommunication Systems Modeling and Analysis. Nashville. TN, pp. 378-393. March 1994.
95. J.K.MacKie-Mason, H.R.Varian. *Pricing Congestible Network Resources*. In: IEEE Journal on Selected Areas in Communications. Vol. 13, No. 7, pp. 1141-1148. Sept. 1995.
96. R.Cocchi, D.Estrin, S.Shenker, L.Zhang. *Pricing in Computer Networks: Motivation, Formulation, and Example*. In: IEEE/ACM Trans. Networking. Vol. 1, pp. 614-627. Dec. 1993.
97. J.Walrand, P.Varaiya. *High-Performance Communication Networks*. Morgan Kaufmann. 535 pp. 1996.
98. C.Parris, S.Keshav, D Ferrari. *A Framework for the Study of Pricing in Integrated Networks*. In: Int. Comput. Sci. Inst. Berkley. CA. Technical Report TR-92-016. Mar.1992.
99. J.Burgin, D.Dorman. *Broadband ISDN Resource Management: the Role of Virtual Paths*. In: IEEE Communications Magazine. Vol. 29, pp. 44-48. Sept. 1993.
100. H.Saito. *Resource Management and Charging in ATM Networks*. In: Computer Networks and ISDN Systems. Vol. 28, pp. 641-644. 1996.
101. S.Low, P.Varaiya. *A New Approach to Service Provisioning in ATM Networks*. In: IEEE Trans. Networking. Vol. 1, pp. 547-553. 1993.
102. C.Parris, D.Ferrari. *A Resource Based Pricing Policy for Real-Time Channels in a Packet-Switching Network*. Preprint.

103. J.Murphy, L.Murphy, E.C.Posner. *Distributed Pricing for Embedded ATM Networks*. In: International Teletraffic Congress ITC-14. Antibes. France. June 1994.
104. L. Murphy, J. Murphy, J. MacKie-Mason. *Feedback and Efficiency in ATM Networks*. In: Proc. IEEE ICC. Ilkley. England, pp. 378-393. July 1995.
105. J. MacKie-Mason, L. Murphy, J. Murphy. *Responsive Pricing in the Internet*. In: Internet Economics. L.W. McKnight and J.P. Bailey, Eds. MIT Press. Cambridge. Massachusetts, pp. 279-303. 1997.
106. L. Murphy, J. Murphy. *Pricing for ATM Network Efficiency*. In: Proc. Of the 3^d International Conference on Telecommunication Systems Modeling and Analysis. Nashville. USA, pp. 349-356. March 1995.
107. F.Kelly. *Tariffs and Effective Bandwidth in Multiservice Networks*. In: International Teletraffic Congress ITC-14, pp. 401-410. 1994.
108. F.Kelly. *Notes on Effective Bandwidths*. In: Stochastic Networks: Theory and Applications. Royal Statistical Society Lecture Notes Series. 4. Oxford University Press, pp. 141-168. 1996.
109. F.Kelly. *On Tariffs, Policing and Admission Control for Multiservice Networks*. In: Operations Research Letters. 15. Elsevier, pp. 1-9. 1994.
110. M. Falkner, M. Devetsikiotis, I. Lambadaris. *Minimum Cost Traffic Shaping: a User's Perspective on Connection Admission Control*. Submitted for Presentation at IEEE ICC'99. Vancouver, Canada. June 1999.

111. R.L.Cruz. *A Calculus of Network Delay. Part 1: Network Elements in Isolation*. In: IEEE Trans. On Information Theory. Vol. 37, pp. 114-131. 1991.
112. G. Doubrovina, M. Falkner, M. Devetsikiotis. *Optimal Cost Traffic Shaping with Self-Similar Input Sources*. Submitted for Presentation at GLOBECOM'99. Rio de Janeiro, Brazil. Dec. 1999.
113. J.G.Ecker, M. Kupferschmid. *Introduction to Operations Research*. John Wiley & Sons, 509 pp. 1988.
114. D. Wismer, R. Chattergy. *Introduction to Nonlinear Optimization: a Problem Solving Approach*. New York: North-Holland. 395 pp. 1978.
115. D. Luenberger. *Linear and Nonlinear Programming*. 2nd ed. Addison-Wesley. 491 p. 1984.
116. P.Spelucci. *DONLP2 Users Guide*. TUD, Germany. 22 pp.