ON THE EFFECTIVE BANDWIDTH FOR RESOURCE MANAGEMENT IN ATM NETWORKS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND ELECTRONİCS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Tijani Chahed

June 1997

ON THE EFFECTIVE BANDWIDTH FOR RESOURCE MANAGEMENT IN ATM NETWORKS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND

ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Tijani Chahed

June 1997

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Erdal Arıkan(Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Tuğrul Dayar

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Ülkü Gürler

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

ON THE EFFECTIVE BANDWIDTH FOR RESOURCE MANAGEMENT IN ATM NETWORKS

Tijani Chahed
M.S. in Electrical and Electronics Engineering
Supervisor: Prof. Dr. Erdal Arıkan
June 1997

The aim of this work is to present a synthesis view of the concept and theory of the effective bandwidth, a measure that captures the statistical characteristics of time-varying ATM sources, to cover some of its applications, to investigate its limitations and to determine the extent of its effectiveness.

For the effective bandwidth to be a viable tool, its on-line, real-time application to network management tasks is a major concern. In this line of thought, on-line traffic characterization and resource management in ATM networks are the main topics of our work.

Since optimal resource management in ATM networks relies essentially on the ability of the network design to perform statistical multiplexing and account for the gains, this case is investigated and a novel, optimal connection admission control (CAC) algorithm is suggested.

Keywords : ATM, Effective Bandwidth, On-Line Resource Management.

ÖZET

ATM AĞLARINDA KAYNAK YÖNETİMİ İÇİN EŞDEĞER BANT GENİŞLİĞİ ÜZERİNE

Tijani Chahed
Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Erdal Arıkan
Haziran 1997

Eşdeğer bant genişliği, zamanla değişen Eşzamansız Aktarım Modu (ATM) kaynaklarının istatiksel karakteristiklerinin bir ölçüsüdür. Bu çalışmada etkin bant genişliği kavramı ve teorisinin bir sentezi sunulmakta, daha sonra uygulamalarından bahsedilmekte, sınırlamaları araştırılmakta ve etkinliği ölçülmektedir.

Eşdeğer bant genişliğinin etkin bir gereç olabilmesi için, ağ yönetimi işinin hatta ve gerçek zamanda uygulanabilmesi büyük önem taşır. Bu bağlamda, çalışmamızın ana konusu Eşzamansız Aktarım Modu ağlarında hatta trafik karakterizasyonu ve kaynak yönetimidir.

Eşzamansız Aktarım Modu ağlarında optimal kaynak yönetimi, ağ tasarımının istatiksel çoğullama ve kazançları dikkate alma yeteneğiyle yakından ilişkilidir. Bu durum araştırılmakta, yeni ve optimal bir bağlantı kabul kontrolu mekanizması önerilmektedir.

Anahtar Kelimeler : Eşzamansız Aktarım Modu, Eşdeğer Bant Genişliği, Hatta Kaynak Yönetimi.

ACKNOWLEDGEMENT

I express my deep gratitude to Dr. Erdal Arıkan for making all this happen.

Sir, I tried to thank you all the way through, but if I didn't, let me do it here. Thank you, Sir.

I am indebted to Dr. Atalar for giving me the oppurtunity of undertaking this work.

I thank the numerous authors who provided me with precious references and invaluable discussions. Special thanks to Dr. Lucantoni, Dr. Chang, Dr. Kesidis, and Dr. Mark, to name just a few.

I thank the readers Dr. Dayar and Dr. Gürler for their effort, kindness, and time.

I am grateful to my family and my friends for their care and support.

I thank any person who in one way or another made this work come true.

Needless to say, none of this would have been possible without my mother, *naturally.*

Tijani C.

to Safia
to the memory of Fathi

Contents

List of Figures

Chapter 1

Introduction

The trend of current and recent developments in telecommunication networks is towards high-speed digital networks, notably the ATM-based B-ISDN, which are expected to support heterogeneous classes of traffic, each usually requiring a different quality of service (QoS). This environment, though flexible enough for supporting existing and future services, presents a dynamic nature that poses complex resource management problems when trying to achieve efficient use of network resources. Efficiency and higher resource utilization are met through statistical multiplexing. The latter, in turn, requires a complete characterization of the traffic. Hence, traffic characterization has an important role to play in the design and control of networks that would be able to cope with bursty multimedia traffic and guaranteed QoS.

Work has been done on a measure that captures the characteristics of the source, including its burstiness, and the service requirements. This measure is known as the *effective bandwidth.*

A measure of the performance of a traffic descriptor such as the effective bandwidth may be the cell loss probability (CLP) at the buffer. Hence, the problem of determining CLPs in queuing systems is also crucial in the development of emergent technology of networks using ATM. However, as ATM networks

would support bursty multimedia traffic, often multiplexing a large number of input streams, an exact analysis and calculation of the CLPs is intractable. In the presence of very small CLP values of the order of 10^{-9} and reasonably large buffer sizes, as in an ATM context, a large deviation approach (Appendix A) may be adopted, and the effective bandwidth acts as an asymptotic approximation to the tail probabilities of loss.

To have a better view and understanding of the effective bandwidth concept, we provide a brief overview that covers the motivations behind such a theory and the major developments that have been reached so far.

The traditional circuit-switched network assumes that each call of class j requires an amount of capacity α_{jk} at link k. The capacity requirement α_{jk} is constant throughout the connection. The theory of such networks is rich, and well-understood. However, what if a call's resource requirements are to vary randomly over the call's duration ?

Hui [25,26] investigated the case of a simple model of unbuffered resources. He assigned to each time-varying source of class j a measure of the real, true bandwidth requirement α_j which he termed the effective bandwidth of the resource at each source of class j. The effective bandwidth α_j depends on the characteristics of the source of class j such as its burstiness, and on the degree of statistical multiplexing possible at the resource. Kelly [27] generalized the notion of the effective bandwidth to certain models of a buffered resource. This would enable to carry the insights available in the circuit-switched networks to these new emerging ATM networks.

Gibbens and Hunt [21] defined the effective bandwidth for a uniform arrival, uniform service (UAS) model. Its calculation is found to be quick, simple and efficient. The multi-service network reduces to a circuit-switched network.

Guerin et al. [23] proposed a computationally simple approximation for the effective bandwidth (also referred to as the equivalent capacity in the literature) for individual and multiplexed connections. They derived an exact computational procedure for the effective bandwidth, though intractable in an ATM networks context.

In [5], Chang developed a general framework for the theory of the effective bandwidth through the large deviation principle. He defined a calculus for network operations based on the effective bandwidth. The importance of his work is that it made possible to extend the single queue results to the network case.

El Walid and Mitra [20] derived the effective bandwidth for general Markovian traffic sources. The effective bandwidth emerged as an explicitly-identified and a simply-computed measure. Its computation depends only on the source characteristics and not on the system. This lead to a decentralized estimation for the measurements.

Later, the effective bandwidth, through a heuristic approach, was proven to exist for multiclass Markov fluids and other ATM sources (constant rate memoryless sources, discrete time Markov sources, Markov fluids and Markov-modulated Poisson processes) [29]. It was regarded as the fixed rate at which each source is transmitting, in a small CLP context. It does not depend on the number of sources sharing the buffer nor on the model parameters of the other sources.

In [17], the effective bandwidth was defined as the minimum bandwidth required by the connection to accommodate its desired QoS. Their work introduced an on-line resource management application, namely traffic monitoring, that makes use of the effective bandwidth concept.

Recently, Kelly [28] presented a synthesis work on the effective bandwidth theory. He stressed the unifying role of the concept, as a summary of the statistical characteristics of sources over time and space, as a limit and an approximation for models of multiplexing under QoS constraints and as a basis for simple and robust scheme for resource management applications such as connection admission control (CAC) mechanisms for a poorly characterized traffic.

It is important to note that, while the effective bandwidth theory is well-defined and understood, very few attempts have been made so as to clear the way to a functional use of it. An application of the theoretical tool to the real world problems is essential. Our main concern, in this work, is to investigate procedures that make use of the effective bandwidth concept for on-line, real-time resource management in ATM networks. Moreover, the effective bandwidth

analysis is generally addressed on a single source model basis. However, resource management in ATM networks is not optimal unless statistical multiplexing is exploited and the gains achieved therein are accounted for. Hence, merging input streams together on a single link and performing network applications on a multiplexed basis is especially of concern.

Our work is organized as follows. Chapter 2 is devoted to a detailed literature review of the effective bandwidth concept. We cover the theory of the effective bandwidth along with some illustrative examples, introduce several traffic descriptors and examine the major resource management applications. Chapter 3 describes and simulates two schemes for on-line, real-time, ATM source characterization. Specifically, based on a model suggested by De Veciana et al. [17], we estimate, via on-line measurement, the performance of a source. An application of this estimation procedure, namely traffic monitoring, is illustrated for a Markovian arrival process. The second scheme, suggested by Mark et al. [32], enables characterization of general types of sources with no restrictive assumptions on the type of offered traffic load. This methodology is also studied and its improved features over the former are emphasized. Once sources are characterized, resource management issues can be addressed in either of two ways: Sources may be either treated individually or on a multiplexed basis. In Chapter 4 we present our work for the multi-source case. We focus on CAC using both the individual and the multiplexed approaches. Since efficiency and optimal utilization of network resources are met through statistical multiplexing, our work focuses on merging several input streams together on a single link and account for the gains achieved therein. Based on this new approach, we suggest a novel, optimal CAC scheme. Owing to the inherent shortcomings of the effective bandwidth concept, Chapter 5 summarizes the major limitations and criticism addressed to this approximation and states our concluding remarks.

Chapter 2

Effective Bandwidth: Review, Theory and Applications

2.1 Major Works on Subject

We present a detailed literature review that covers the motivations behind the effective bandwidth theory, its aims, the major developments that have been reached so far and their underlying limitations.

As mentioned in the Introduction, the traditional circuit-switched network assumes that each link k has a capacity C_k and that each call of class j requires an amount of capacity α_{jk} at link k. The network is able to carry n_j calls of class j if $\sum_j n_j \alpha_{jk} \leq C_k$. The theory of such networks is rich and well-understood. However, what if a call's resource requirements are to vary randomly over the calls duration ?

Hui [25,26], for a simple model of unbuffered resources and using the large deviation approximation (Appendix A), showed that the probability of resource overload can be held below a desired level by imposing that the number of calls

n_j accepted from sources of class j satisfies

$$\sum_j n_j \alpha_j \leq C, \tag{2.1}$$

where C is the capacity of the resource and α_j is the effective bandwidth of the resource at each source of class j. The effective bandwidth α_j depends on the characteristics of the source of class j such as its burstiness and on the degree of statistical multiplexing possible at the resource. The obtained results were asymptotically exact.

Kelly [27] generalized the notion of the effective bandwidth, additive over sources of different classes, to certain models of a buffered resource. Introducing finite size buffers is important for they disable traffic fluctuations for the streams sharing the resources. Most importantly, this would enable to carry the insights available in the circuit-switched networks to these new emerging ATM networks. The models considered were the $M/G/1$ and the $GI/G/1$ queues. Kelly concluded that this concept is important for source classification and admission control. The major drawback is in overload, i.e., for large resource capacity C and larger effective offered load, the accepted sources would be biased towards those with low effective bandwidth.

In [21], Gibbens and Hunt defined the effective bandwidth for the uniform arrival, uniform service (UAS) model. They considered heterogeneous ON/OFF sources which alternate between exponentially distributed periods of transmission at the peak rate and idleness. The effective bandwidth is found to be a measure that depends only on the mean bandwidth, burstiness of the source and the channel. Its calculation is found to be quick and efficient. The multi-service network reduces to a circuit-switched network.

The work found in [21] was based on [2]. Therein, the authors analyzed the statistical multiplexing of N sources of a single type onto a communication channel. So, two generalizations had to be investigated: (i) Case of more than one single channel type, and (ii) case of a network of such channels. To investigate the first issue, one could think of solving a set of single-type source problems. However, this turned out to be wrong. Instead, for m different types of sources,

an effective bandwidth is defined as the set

$$\alpha(B,p) = \{\mathbf{N} = (N_1, N_2, ..N_m) : P_N(\text{queue length} \geq B) \leq p\}, \quad (2.2)$$

where B is the buffer size and p is the cell loss probability (CLP). For p of the order of 10^{-9}, the large deviations approach is used and $\alpha(B,p)$ is approximated by

$$\alpha(B,p) = \{\mathbf{N} = (N_1, N_2, ..N_m) : \sum_i n_i\alpha_i \leq C\}, \quad (2.3)$$

where α_i is the effective bandwidth of type-i source. This approximation would make it possible to analyze the two issues cited above. Specifically, the author distinguished two cases. For large buffer size B, $log(p)/B$ tends to $\xi \in [-\infty, 0]$, approximation (2.3) was shown to be exact, and for an ON/OFF type-i source with ON and OFF periods exponentially distributed with means μ_i^{-1} and λ_i^{-1}

$$\alpha_i(\xi) = \frac{\xi\gamma_i + \mu_i + \lambda_i - \sqrt{(\xi\gamma_i + \mu_i - \lambda_i)^2 + 4\lambda_i\mu_i}}{2\xi}, \quad (2.4)$$

where γ_i is the constant rate of transmission of source of type i while in ON state.

Note that, $\xi = 0$ corresponds to

$$\alpha_i(0) = \frac{\lambda_i\gamma_i}{\lambda_i + \gamma_i}. \quad (2.5)$$

That is, the effective bandwidth is taken as the mean rate, which is a lower bound. The only constraint on the system is that it be stable. This is due to the long time elapsed before the buffer size is exceeded, and sources have been through their ON and OFF periods many times. The other case is for $\xi = -\infty$. Then,

$$\alpha_i(-\infty) = \gamma_i. \quad (2.6)$$

This corresponds to the peak rate. This happens when the source is so bursty, that a peak rate bandwidth allocation is needed. Hence, the measure $\zeta = 1 - e^{\xi} = 1 - p^{1/B} \in [0,1]$ is the asymptotic system burstiness, i.e., the larger ζ becomes the more susceptible the channel is to burstiness. The fact that ζ depends on the channel only, and not on the offered load, is of prime importance in the design of ATM networks.

For small buffer size B, the buffer overflows if

$$\sum_i \gamma_i n_i > c. \tag{2.7}$$

The rare event is no longer that the queue has built up, but rather that an unusually large number of sources are ON at once. For $B = 0$, the exact acceptance set is given by the Bernoulli model, and the UAS model turns out more accurate than the large deviation approximation. For B small, but non-zero, the difference between the two approximations is large, but gets smaller as B is increased. As a conclusion, the effective bandwidth is valid for finite buffer size B, with different levels of accuracy.

Guerin et al. [23] proposed a computationally simple approximation for the effective bandwidth (also referred to as the equivalent capacity in the literature) for individual and multiplexed connections. This unified metric would turn out to be useful for bandwidth management, routing and call control procedures. It should be emphasized that their methodology can provide an exact computational procedure for the effective bandwidth, but its complexity makes it intractable in real-time network traffic applications. Thus, an approximation is more appropriate. The first approximation is based on the fluid-flow model (continuous-time Markov chain) and the second is the stationary approximation for the bit rate distribution. In most cases where the statistical multiplexing effect is significant, the distribution of the stationary bit rate can be accurately approximated by the Gaussian distribution. However it should be emphasized that the second approximation does not hold in the case of sources with long burst periods (as for the asymptotic approximation on the large buffer size). The authors argued that even an exact model does not provide a correct measure of the CLP seen by connections, as it is unable to fully capture the impact of the interactions within the network.

In [4], Chang developed a general framework for the theory of the effective bandwidth. A general form of the effective bandwidth was derived using the large deviation theorem. The following calculus for network operations was found: (i) The effective bandwidth of multiplexed independent sources is less than the sum of the effective bandwidths of the sources, (ii) the distribution of delays and

queue lengths is bounded by exponential tails if the effective bandwidth is less than the channel capacity, (iii) the effective bandwidth of the output is less than the effective bandwidth of the input, (iv) the effective bandwidth of a routed process from a departure process can be derived from the effective bandwidth of the departure process and the effective bandwidth of the routed process. The importance of these rules is that they make possible to extend the single queue results to the network case.

In [5], an intuitive derivation for the effective bandwidth was given. To achieve this goal, and instead of investigating the queue behavior at the packet level, time and space were scaled, yielding the approximation of ATM networks by stochastic fluid network, with external inputs given by the Gibbs distribution. The tail distribution of queue lengths in fluid networks could then be solved by simple minimization problems.

El Walid and Mitra [20] derived the effective bandwidth for general Markovian traffic sources, with no restrictions on their dimensionality, homogeneity or time reversibility. They defined the effective bandwidth as the maximal real eigenvalue of a matrix directly obtained from the source parameters and the admission criterion (network resources and service requirements).

The source is characterized by (M, λ), where M is the irreducible infinitesimal generator of a Markov chain and λ_s is the source constant rate at state s. Let $G(B)$ be the stationary $Pr(X \geq B)$, where X is the random buffer content and $G(B)$ the overflow probability for a buffer of size B. The admission criterion is $G(B) \leq p$. The asymptotic regime states that p tends to 0 and B tends to infinity. In this case $\log(p/B)$ tends to $\xi \in [-\infty, 0]$. The effective bandwidth depends on (M, λ) , and on the buffer and overflow probability only through ξ. The effective bandwidth α turned out to be the maximal real eigenvalue of the matrix

$$\left[\Lambda - \frac{1}{\xi} M\right], \tag{2.8}$$

where $\Lambda = diag(\lambda)$. For k sources characterized by $(M^{(k)}, \lambda^{(k)})$, $\alpha = \sum \alpha^{(k)}$. As for the point processes case, The effective bandwidth α is the maximal real

eigenvalue of the matrix,

$$\left[\frac{1}{e^{\xi}}\Lambda - \frac{1}{1 - e^{\xi}}M\right]. \tag{2.9}$$

Hence, the effective bandwidth is an explicitly-identified, a simply-computed measure. Its computation depends only on the source characteristics, and not the system, which leads the way to a decentralized estimation for the measurements. It has correct properties at the natural asymptotic regime (CLP tends to zero and buffer size tends to infinity) of the small loss probabilities. The authors considered both Markov-modulated fluid and point processes (such as Markov-modulated Poisson or phase renewal processes). Most importantly, the effective bandwidth notion acts as a bridge to the familiar circuit-switched network design. Numerically, the effective bandwidth provides a conservative bound on the acceptance region, and contrary to the mean and peak rates, it provides an effective basis for admission control. Finally, in the presence of a leaky bucket, the effective bandwidth decreases, by time, from the maximal effective bandwidth of the source to a lower value.

Later, the effective bandwidth, through a heuristic approach, was proven to exist for multiclass Markov fluids and other ATM sources (constant-rate memoryless sources, discrete-time Markov sources, Markov fluids and Markov-modulated Poisson processes) [29]. It was regarded as the fixed rate at which each source is transmitting, in a small CLP context. It does not depend on the number of sources sharing the buffer nor on the model parameters of the other sources. Again, the buffer size was assumed infinite and the source was subject to the same assumptions as in [5].

Gun et al. [24] investigated the effective bandwidth vectors for multiclass traffic multiplexed on a partitioned buffer. Their model included a single source generating j different classes of traffic and various independent, heterogeneous and Markov-modulated fluid sources multiplexed into a single buffer. They defined an effective bandwidth vector corresponding to the j QoS classes. In their analysis, they used the property that if the sum of the effective bandwidth of all the sources multiplexed onto the buffer is less than the speed of the channel to remove the data from that buffer, then the QoS is satisfied for all the sources.

Moreover, this work is valid for sources with close QoS requirements. In practice, when sources have different QoS, they all line up to the stringent QoS. This results in sub-optimal resource allocation. The solution can be to classify traffic into multiple classes according to their QoS and to assign a separate buffer for each. However, this method has many drawbacks. It is better to adopt a shared buffer scheme. Yet, the effective bandwidth approximation is again based on the large buffer asymptotics. This can significantly overestimate (in the case of sources that are more bursty than Poisson) or underestimate (in the case of sources less bursty than Poisson) the number of sources that can be multiplexed on a trunk. The effective bandwidth vector is not additive, provides an approximation and not a bound and leads, numerically, to sub-optimal buffer thresholds.

In [17], the effective bandwidth was defined as the minimum bandwidth required by a connection to accommodate its desired QoS. Based on large deviations, an effective bandwidth approximation was found to mimic the tail probabilities of loss. For a buffered link with capacity c cells/sec, supporting a stationary and ergodic arrival packet stream $A(0,t)$, let X be the buffer's stationary workload. QoS aims to limit the likelihood of large delays or to ensure that CLP is small. Based on the large deviation theorem, the tail probability of loss is approximated by

$$P(X > B) \leq p \equiv e^{-B\delta} \ll 1, \tag{2.10}$$

where δ is a measure of the stringency of the QoS requirement and B is the buffer size. The effective bandwidth function is defined by

$$\alpha(\delta) = \delta^{-1} \lim_{t\to\infty} t^{-1} \log\left[E\{e^{\delta A(0,t)}\}\right]. \tag{2.11}$$

The derivation is as follows. Under mild conditions for both continuous and discrete-time arrival processes, and for all $\delta \geq 0$,

$$\alpha(\delta) < c \iff \lim_{B\to\infty} B^{-1} \log[P(X > B)] \leq -\delta, \tag{2.12}$$

or equivalently,

$$P(X > B) = \exp\left[-B\alpha^{-1}(c) + o(B)\right]. \tag{2.13}$$

It was used as a traffic descriptor for traffic monitoring. The traffic considered was one that tolerates statistical QoS guarantees, and again the large

deviation assumptions were made. On its viability as a traffic descriptor, it turned out to be efficient, reasonably accurate, robust and simple in the sense of implementation and performance. On its accuracy and the relaxation of the stationarity assumptions, it turned out not to capture all the statistical multiplexing gains. One way to solve this is to use refined asymptotics [9] (Section 2.3.1), or to combine the effective bandwidth with a zero-buffer approximation. As for the quasi-static approximation (the assumption that the system reaches steady state), it could not be relaxed. The effective bandwidth approximation is: $P(X > B) \approx Aexp(-BI) \approx exp(-BI)$, where $I = \alpha^{-1}(c)$. Thus A is approximated as 1. However, in reality A may be large, or small. Determining A may be prohibitive and would make bandwidth management based on the effective bandwidth inefficient. A remedy to this would be to measure the extent of the statistical multiplexing gain (A) and change the admission criterion (δ) [30]. But, this solution requires a complete specification of the entire effective bandwidth characteristics.

In [15], the effective bandwidth was derived for stationary sources. It has been computed as a function of the mean rate, index of dispersion and the buffer size. Specifically, for a stationary process X_n with kth order autocovariance $\gamma(k)$ and spectral density function $f(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k) \exp(i\omega k)$, the index of dispersion γ is defined as

$$\gamma = \pi f(0) = \sum_{k=-\infty}^{\infty} \gamma(k). \tag{2.14}$$

The assumption is that the process should be purely indeterministic without periodicity or long-term dependencies. As for the effective bandwidth and for N_i stationary, independent inputs satisfying the Gartner-Ellis theorem (Appendix B), each with mean μ_i and index of dispersion γ_i, the effective bandwidth α_i associated with each source is

$$\alpha_i = \mu_i + \frac{\delta\gamma_i}{2} + o(\delta), \tag{2.15}$$

where δ is a measure of the stringency of the QoS requirements. Moreover, formulae for Markov-modulated fluid (two-state fluid, $M/M/\infty$, n-state fluid) were also given therein. These formulae have been tested numerically in [16] and yielded good results.

A simple, intuitive overview on the effective bandwidth theory was presented in [4]. The approach was through the large deviation theory and the Laplace method of integration. It includes: (i) Identification of the energy function, entropy function and the effective bandwidth function, (ii) a calculus of the effective bandwidth function, (iii) bandwidth allocation and buffer management, (iv) traffic descriptors and (v) envelope processes and conjugate processes, a method for fast simulations and bounds. The effective bandwidth is regarded as the minimum bandwidth to satisfy the corresponding QoS. For a source modeled as a constant-rate fluid with rate λ on a period of time t and probability density function $f(\lambda, t)$, the tail distribution of the queue is given by the integral of $f(\lambda, t)$. Using the large deviation theorem, $f(\lambda, t)$ is shown to have the Gibbs distribution, i.e.,

$$f(\lambda, t) = \exp(-t\Lambda^*(\lambda)). \tag{2.16}$$

$\Lambda^*(\lambda)$ is the energy function obtained from the Legendre transformation of $\Lambda(\theta)$, i.e.,

$$\Lambda^*(\lambda) = \sup(\theta\lambda - \Lambda(\theta)), \tag{2.17}$$

and $\Lambda(\theta)$is the entropy function obtained from the Gartner-Ellis limit of a source, i.e.,

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log\left[E\{e^{\theta A(0,t)}\}\right]. \tag{2.18}$$

Then,

$$\alpha^*(\theta) = \frac{\Lambda(\theta)}{\theta}. \tag{2.19}$$

is the effective bandwidth function. Note that $\alpha^*(\theta)$ is increasing in θ, it tends to the average rate as θ tends to 0 and tends to the peak rate as θ tends to infinity. The calculus of the effective bandwidth is then derived heuristically. It mainly shows that the effective bandwidth of multiplexing independent arrivals is additive. As for the application of the theory to networks, the results are asymptotic, in the sense that they yield very loose bounds for finite buffer sizes. The independent input assumption makes it difficult to analyze networks with loops.

Recently, Kelly [28] presented a synthesis work on the effective bandwidth

theory. He stressed the unifying role of the concept, as a summary of the statistical characteristics of sources over time and space, as a limit and an approximation for models of multiplexing under QoS constraints and as a basis for simple and robust scheme for tariffing and connection admission control mechanisms for poorly characterized traffic. The material found in the next section is inspired mainly from this work [28].

2.2 Effective Bandwidth Theory

In what follows, an alternative definition of the effective bandwidth is given, along with some of its major properties and examples. Later, the theory is extended to some common multiplexing models. Note that the effective bandwidth, defined next, is associated with *stationary* sources.

2.2.1 Definition

Kelly [28] defines the effective bandwidth in terms of two free parameters, s and t, representing the space and time scales, respectively. The appropriate choice of s and t will depend on the characteristics of the resource (capacity, buffer size, traffic mix, scheduling policy, etc.).

Let $X[0,t]$ be the amount of work arriving from a source over a time interval $[0,t]$ and let $X[0,t]$ have stationary increments. The effective bandwidth is defined as

$$\alpha(s,t) = \frac{1}{st} \log\left[E\{e^{sX[0,t]}\}\right]. \tag{2.20}$$

2.2.2 Properties

Among others, the following properties are distinguished as the major ones:

1. If $X[0,t]$ has independent increments, then $\alpha(s,t)$ does not depend on t.

2. If $X[0,t] = Xt$ for $t > 0$, then $\alpha(s,t) = \alpha(st,1)$, i.e., $\alpha(s,t)$ depends on s and t through the product st only. Otherwise, $\alpha(s/t,t)$ is strictly decreasing in t.

3. Additive Property - If $X[0,t] = \sum_i X_i[0,t]$, where $(X_i[0,t])_i$ are independent, then

$$\alpha(s,t) = \sum_i \alpha_i(s,t). \tag{2.21}$$

4. For any fixed value of t, $\alpha(s,t)$ is increasing in s, and lies between the mean and the peak of the arrival rate measured over an interval of length t, i.e.,

$$\frac{EX[0,t]}{t} \leq \alpha(s,t) \leq \frac{\bar{X}[0,t]}{t}. \tag{2.22}$$

2.2.3 Examples

To have a better appreciation of the effective bandwidth concept, Kelly [28] derived some closed form expressions for some common source types. These are presented here along with some illustrative figures.

Periodic Sources

This model can be used to describe arrivals of constant-rate sources. Let B units of workload be produced at times $Ud + nd$, $n = 0, 1, \ldots$, where U is uniformly distributed on the interval $[0,1]$. The effective bandwidth function is given by

$$\alpha(s,t) = \frac{B}{t}\lfloor\frac{t}{d}\rfloor + \frac{1}{st}\log\left[1 + (\frac{t}{d} - \lfloor\frac{t}{d}\rfloor)(e^{Bs} - 1)\right]. \tag{2.23}$$

Note that

$$\lim_{t\to 0}\alpha(s,t) = \frac{e^{Bs} - 1}{ds}. \tag{2.24}$$

An illustration of Eqn. (2.23) is given in Fig. 2.1.

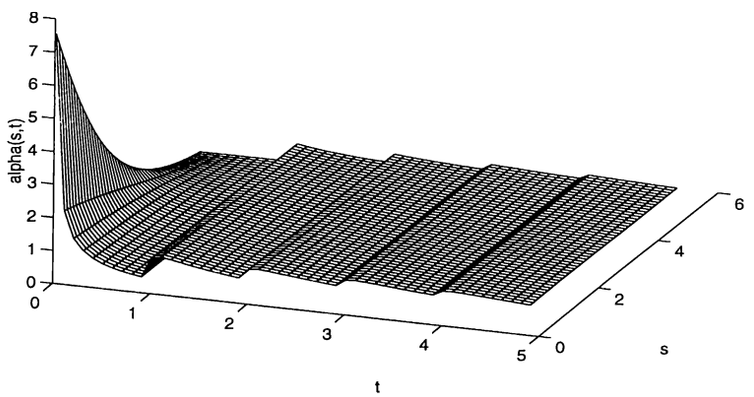

Figure 2.1: The effective bandwidth for periodic sources. The parameters $B = d = 1$. A single unit of workload is produced at the end of every unit interval with random phase. The effective bandwidth is seen to grow over intervals shorter than the period of the source

Fluid Sources

Let a stationary fluid source be described by a two-state Markov chain. The transition rate from state 1 to state 2 is λ and the transition rate from state 2 to state 1 is μ. While in state 1, workload is produced at a constant rate h, and no workload is produced at state 2. The effective bandwidth function is given by

$$\alpha(s,t) = \frac{1}{st} \log \left[\left(\frac{\lambda}{\lambda+\mu}, \frac{\mu}{\lambda+\mu} \right) \exp \left[\begin{pmatrix} -\mu + hs & \mu \\ \lambda & -\lambda \end{pmatrix} t \right] \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right]. \quad (2.25)$$

Note that

$$\lim_{t\to\infty} \alpha(s,t) = \frac{1}{2s} (hs - \mu - \lambda + \sqrt{(hs - \mu + \lambda)^2 + 4\mu\lambda}). \quad (2.26)$$

In general, for a stationary source described by a finite Markov chain with stationary distribution π and q-matrix Q, with the workload produced at rate h_i while in state i,

$$\alpha(s,t) = \frac{1}{st} \log[\pi \exp[(Q + hs)t]1], \quad (2.27)$$

where $h = diag(h_i)_i$.

Note that

$$\lim_{t\to\infty} \alpha(s,t) = \frac{1}{s}\,\phi(s), \tag{2.28}$$

where $\phi(s)$ is the largest real eigenvalue of the matrix $Q + hs$. An illustration of Eqn. (2.25) is given in Fig. 2.2.

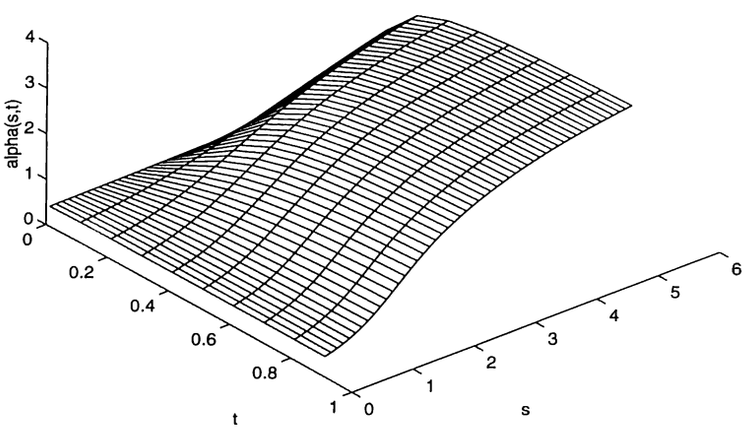

Figure 2.2: Effective bandwidth of an on/off fluid source. The parameters $\lambda = 1$, $\mu = 9$, and $h = 10$. The effective bandwidth approaches the mean rate $\lambda h/(\lambda+\mu)$, as s or t approaches zero.

Gaussian Sources

Consider $X[0,t] = \lambda t + Z(t)$, where $Z(t)$ is normally distributed with zero mean, then,

$$\alpha(s,t) = \lambda + \frac{s}{2t}\,Var[Z(t)]. \tag{2.29}$$

In the heavy traffic case, $Var[Z(t)] = \sigma^2 t$.

In general, Z is a fractional Brownian motion with Hurst parameter $H \in (0,1)$, and $Var[Z(t)] = \sigma^2 t^{2H}$. Then,

$$\alpha(s,t) = \lambda + \frac{\sigma^2 s}{2} t^{2H-1}. \tag{2.30}$$

If $H < 1/2$, $\lim_{t\to\infty} \alpha(s,t)$ is finite and does not depend on s; if $H = 1/2$, $\lim_{t\to\infty} \alpha(s,t)$ is finite but depends on s; and if $H > 1/2$, $\alpha(s,t)$ grows as a fractional power of t. The latter case exhibits long range order. An illustration of Eqn. (2.30) is given in Fig. 2.3.

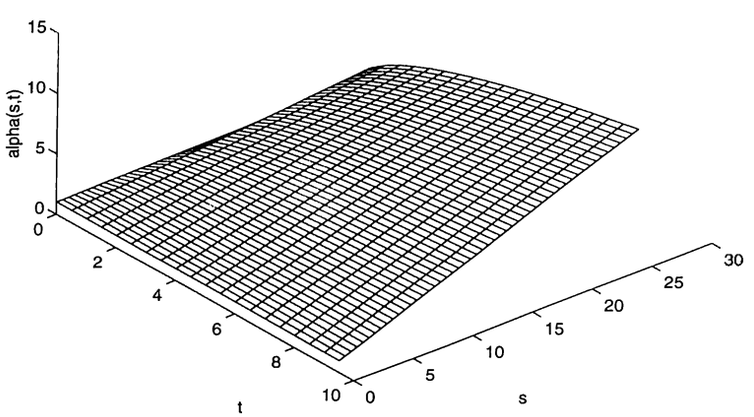

Figure 2.3: The Effective bandwidth of a Gaussian source. The parameters are $H = 0.75$, $\lambda = 1$, and $\sigma^2 = 0.25$. The long-range order is indicated by the continued growth of the effective bandwidth with large t.

General ON/OFF Sources

The source in this case alternates between long periods in ON state with effective bandwidth $\alpha_1(s,t)$, and long periods in OFF state where it produces no workload. Let p be the proportion of time spent in the ON state. For values of t small compared to periods spent in ON and OFF states, the effective bandwidth is

$$\alpha(s,t) = \frac{1}{st} \log\left[1 + p\left(\exp(st\alpha_1(s,t)) - 1\right)\right], \tag{2.31}$$

where $\alpha_1(s,t)$ is given by (2.23). An illustration of Eqn. (2.31) is given in Fig. 2.4.

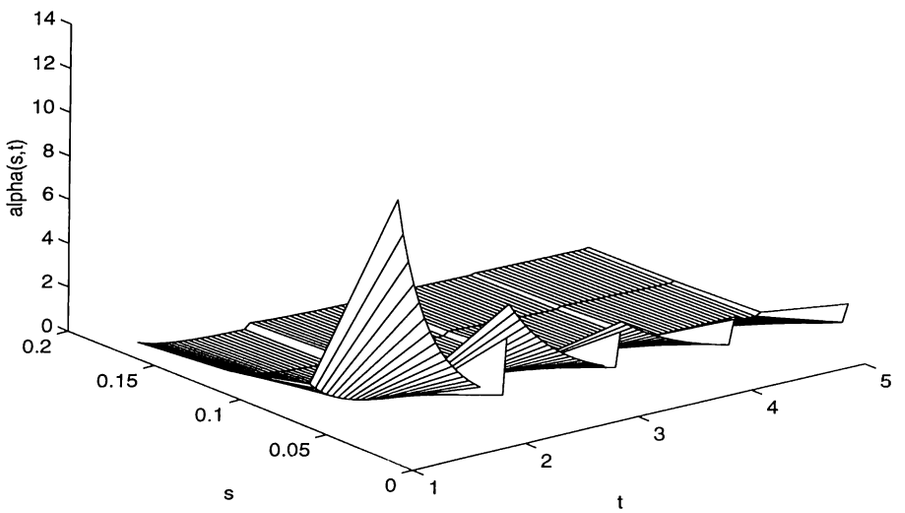

Figure 2.4: The effective bandwidth of an ON/OFF periodic source. The parameters are $p = 0.05$, $B = 2$, and $d = 1$. The increase of the effective bandwidth as t either increases towards the interval over which the source remains ON or OFF, or decreases below the period of the source.

2.2.4 Multiplexing Models

Owing to the importance of multiplexing in achieving efficiency and optimality in resource management in ATM networks, Kelly [28] suggests the following multiplexing scheme.

Let the arrival process be as follows:

$$X[0,t] = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji}[0,t], \tag{2.32}$$

where $(X_{ji}[0,t])_{ji}$ are independent processes with stationary increments whose distribution may depend on j, but not on i. Let there be a resource to cope with the aggregate arriving stream. n_j is the number of sources of type j and $\alpha_j(s,t)$ is the effective bandwidth of a source of type j. Hence,

$$\alpha(s,t) = \sum_{j=1}^{J} n_j \alpha_j(s,t). \tag{2.33}$$

For several multiplexing models to come, we would be interested in the relationship between the constraint

$$\sum_{j=1}^{J} n_j \alpha_j(s^*, t^*) \leq C^*, \tag{2.34}$$

for choices of (s^*, t^*, C^*) and the acceptance region defined as the set of vectors $(n_1, n_2, ..., n_J)$ for which a given performance, in terms of queuing delay or buffer overflow, is guaranteed.

Bufferless Models

Let

$$X = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji}, \tag{2.35}$$

where X_{ji} are independent random variables with scaled logarithmic moment generating functions

$$\alpha_j(s) = \frac{1}{s} \log \left[E\{e^{sX_{ji}}\} \right]. \tag{2.36}$$

X_{ji} may be regarded as the instantaneous arrival load at a bufferless resource of capacity C, so that $\alpha_j(s) = \lim_{t \to 0} \alpha_j(s/t, t)$. Chernoff's bound yields

$$\log P(X \geq C) \leq \log \left[E\{e^{s(X-C)}\} \right] = s\,(\alpha(s) - C), \tag{2.37}$$

where $\alpha(s) = \sum_j n_j \alpha_j(s)$. Hence, $\log P(X \geq C) \leq -\gamma$ is satisfied within the set,

$$\mathcal{A} = \{n : \inf_s \left[s\,(\sum_j n_j \alpha_j(s) - C) \right] \leq -\gamma\} \quad \text{for} \quad n \geq 0. \tag{2.38}$$

The half space touching at a point n^* on the boundary of region $\mathcal{A}$ is

$$\sum_j n_j \alpha_j(s^*) \leq C - \frac{\gamma}{s^*}, \tag{2.39}$$

where s^* satisfies the infimum in (2.38) with n replaced by n^*. This is a conservative bound on the non-linear acceptance region for a bufferless model. Applying the Chernoff theorem, we get

$$\lim_{N\to\infty} \frac{\mathcal{A}\ (\gamma N\ ,\ CN)}{N} = \mathcal{A}; \tag{2.40}$$

that is, as the number of sources increases and the tail probability decreases, the approximation made to lead to the region $\mathcal{A}$ becomes more accurate.

M/G/1 Models

In this section, (2.34) will emerge as the linear limiting form, conservative bound, on the acceptance region for a buffered model with general input processes.

Let $X_{ji}[0,t]$ have independent increments, $\alpha_j(s) = \alpha_j(s,t)$ and $\alpha(s) = \alpha(s,t)$. Let Q be the stationary workload in the queue and C the capacity of the server. The buffer size is infinite. The Pollaczek-Khinchin formula is

$$E[e^{sQ}] \ = \ \frac{C\ -\ \alpha(0)}{C\ -\ \alpha(s)}. \tag{2.41}$$

Let κ be a finite constant such that $\alpha(\kappa) \ = \ C$, then the Cramer's estimate gives

$$P(Q \geq B) \ \approx \ \frac{C\ -\ \alpha(0)}{\kappa\ \alpha'(\kappa)}\ e^{-\kappa B} \qquad \text{as} \quad B \to \infty. \tag{2.42}$$

For $P(Q \geq B) \ \leq \ -\gamma, n \ \in \ \mathcal{A}(\gamma\ ,\ B)$, Cramer's estimate implies

$$\lim_{N\to\infty} \mathcal{A}\ (\gamma N\ ,\ BN) \ = \ \mathcal{A}, \tag{2.43}$$

where

$$\mathcal{A} \ = \ \{n\ :\ \sum_j n_j\ \alpha(\frac{\gamma}{B}) \ \leq \ C\}. \tag{2.44}$$

So far, we were concerned with the proportion of time the buffer occupancy exceeded a level B in a queue with an infinite buffer. Next, we will consider the case of a finite buffer size where the workload exceeding this level would be lost.

Finite Buffers

Let the proportion of the workload lost in the finite buffer be $L(B)$. From (2.42), we deduce

$$L(B) \approx \frac{C\ (C\ -\ \alpha(0))^2}{\kappa\ \alpha'(\kappa)\ \alpha(0)}\ e^{\kappa B} \quad \text{as} \quad B \to \infty. \tag{2.45}$$

For $\log L(B) \leq -\gamma$, $n \in \mathcal{A}_{prop}(\gamma, B)$ and

$$\lim_{N\to\infty} \mathcal{A}\ (\gamma N\ ,\ BN) = \mathcal{A}. \tag{2.46}$$

Mean Delays

From the Pollaczek-Khinchin formula (2.41), it follows that $EQ = \frac{\alpha'(0)}{(C\ -\ \alpha(0))}$ and the constraint $EQ \leq L$ is satisfied if and only if

$$\sum_{j=1}^{J} n_j \left[\alpha_j(0) + \frac{\alpha_j'(0)}{L}\right] \leq C. \tag{2.47}$$

Buffer Asymptotic Models

For models more general than the $M/G/1$, tail probabilities decay exponentially. To see this, consider an arrival stream $X[0,t]$ with stationary and ergodic increments. Let Q be the stationary-distributed workload of a queue with capacity C and finite buffer size B. Let

$$\lim_{t\to\infty} \alpha(s,t) = \alpha(s). \tag{2.48}$$

Let there be a finite constant κ such that $\alpha(\kappa) = C$ and $\alpha'(\kappa)$ is finite. Then,

$$\lim_{B\to\infty} \frac{1}{B} \log\ P(Q \geq B) = -\kappa. \tag{2.49}$$

Deterministic Multiplexing

For a server of capacity C and a buffer of finite capacity B, the condition under which the buffer capacity is never exceeded is that $n \in \mathcal{A}$, where

$$\mathcal{A} = \bigcap_{0<t<\infty} \mathcal{A}_t, \tag{2.50}$$

is the intersection of the linearly constrained regions

$$\mathcal{A}_t = \{n : \sum_j n_j \alpha_j(\infty, t) \leq C + \frac{B}{t}\}. \tag{2.51}$$

2.3 Applications of Effective Bandwidth Theory

As the effective bandwidth formulation is rather simple and robust, the theory has many applications in high-speed digital networks, such as ATM. The main applications we will encounter in this section are the ones dealing with resource management in ATM networks. Resource management is an important issue in the design and implementation of such networks, for it leads the way to achieve optimality in their use, maximize their efficiency, enhance their reliability and thus ensure their viability.

2.3.1 Effective Bandwidth as Traffic Descriptor

To derive the bandwidth requirement of a service, one needs to characterize its traffic pattern, either by a set of traffic descriptors or by a mathematical model. The latter is known to be intractable in an ATM networks context. Traffic descriptors, such as the effective bandwidth, can be used to approximate the given traffic. The QoS can be formulated in terms of the CLP in the queue, i.e., for a traffic workload X and a buffer capacity B, the CLP is $P(X > B)$, or in terms of the delay experienced by the packets (or cells) in the queue, that is, $P(W > t)$ for W being the waiting time. Clearly, these two measures of performance are related. In what follows, we present the main traffic descriptors.

Three-parameter Approach

In Guerin et al. [23], traffic was characterized by three parameters: peak rate λ_p, average burst period b and the utilization ρ (i.e., fraction of time the source is active). These parameters were then mapped to an ON/OFF Markov fluid (fluid-flow approximation). For a finite buffer size B, the equivalent capacity (effective bandwidth), α, is related to the overflow probability p by

$$p = \beta \exp\left(-\frac{B(\alpha - \rho\lambda_p)}{b(1 - \rho)(\lambda_p - \alpha)\alpha}\right), \tag{2.52}$$

where β is

$$\beta = \frac{(\alpha - \rho\lambda_p) + p\rho(\lambda_p - \alpha)}{(1 - \rho)\alpha}. \tag{2.53}$$

Typically, β is taken as 1. For the single source case, the equivalent capacity is given as

$$\alpha = \frac{(1/p)b(1-\rho)\lambda_p - B + \sqrt{[(1/p)b(1-\rho)\lambda_p - B]^2 + 4B(1/p)b\rho(1-\rho)\lambda_p}}{2(1/p)b(1 - \rho)}, \tag{2.54}$$

Extending the result to the multiple source case, the equivalent capacity is simply

$$C_F = \sum_{i=1}^{N} \alpha_i, \tag{2.55}$$

where α_i are determined from (2.54).

Note that for connections with long burst periods and low utilization, β is significantly different from 1. However, multiplexing a number of connections with long burst periods can be approximated by a stationary bit rate distribution (stationary approximation). The equivalent capacity C_s is given by

$$C_S = m + p'\sigma, \tag{2.56}$$

with

$$p' = \sqrt{-2ln(p) - ln(2\pi)}, \tag{2.57}$$

where m is the mean aggregate bit rate ($m = \sum_{i=1}^{N} m_i$) and σ is the standard deviation of the aggregate bit rate ($\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$).

Combining the two approximations, the equivalent capacity C is simply the minimum between C_F and C_S, namely

$$C = \min(m + p'\sigma , \sum_{i=1}^{N} \alpha_i). \tag{2.58}$$

Four-parameter Approach

In [5], a four-parameter traffic descriptor is obtained. These four parameters are related to the average rate, asymptotic variance, peak rate and average burst duration of the source. The derivation is as follows: the effective bandwidth function may be expanded at $\theta = 0$ and $\theta = \infty$ to give

$$\alpha^*(\theta) = \eta_1 + \eta_2\theta + o(\theta^2) \qquad \text{as } \theta \to 0, \tag{2.59}$$

$$\alpha^*(\theta) = \eta_3 - \frac{\eta_4}{\theta} + o(\frac{1}{\theta^2}) \qquad \text{as } \theta \to \infty, \tag{2.60}$$

where η_1 is average rate, $2\eta_2$ is the asymptotic variance, η_3 is the peak rate and η_4 is the average burst period.

Let $I(c)$ be the inverse function of the effective bandwidth function $\alpha^*(\theta)$, i.e., $\theta = I(c)$ is the solution of $a^*(\theta) = c$. $I(c)$ has a zero at η_1 and a pole at η_3. Hence,

$$I(c) \approx \frac{c - \eta_1}{(c - \eta_3)(\beta_1 c + \beta_2)}, \tag{2.61}$$

where

$$\beta_1 = -\frac{1}{\eta_4} + \frac{\eta_2}{(\eta_3 - \eta_1)^2}, \tag{2.62}$$

and

$$\beta_2 = \frac{\eta_1}{\eta_4} + \frac{\eta_2\eta_3}{(\eta_3 - \eta_1)^2}. \tag{2.63}$$

So, the queue length distribution can be approximated by

$$Pr(X > B) \approx exp(-I(c)B). \tag{2.64}$$

The model considered assumes infinite buffer size at each queue, and to apply the large deviation theorem, the source is assumed to be jointly stationary and ergodic, among other assumptions so as to satisfy the Gartner-Ellis theorem. It should be pointed out that the latter theorem is valid for finite dimensional random vectors instead of stochastic processes; moreover, the large deviation principle is only valid for independent, identically distributed random variables. The numerical results therein show that the effective bandwidth approximation holds for the case of positively correlated sources and negatively correlated ones. It is worth pointing out that these parameters η_i's and especially the asymptotic variance η_2 are hard to measure in real life applications.

First Order Exponent

The effective bandwidth is the traffic descriptor of predilection in our work. Recall from Eqn. (2.11) that the effective bandwidth function as proposed by De Veciana et al. [17] is as follows. For a stationary and ergodic arrival stream $A(0,t)$ and a stationary workload X, the effective bandwidth has been approximated by

$$\alpha(\delta) \;=\; \delta^{-1} \lim_{t\to\infty} t^{-1} \log\left[E\{\exp[\delta A(0,t)]\}\right]. \tag{2.65}$$

In considerable generality, the cell loss tail probabilities are asymptotically exponential, i.e., for a buffer workload X and finite buffer size B

$$P(X > B) \approx a e^{-bB} \quad \text{as} \quad B \to \infty, \tag{2.66}$$

where a is a positive constant called the asymptotic constant, and b is a positive constant termed the asymptotic decay rate. The effective bandwidth approximation assumes $a = 1$.

Here, the QoS-constraint matches the CLP to the first order, that is, in the exponent. Recall from Section 2.1 that this result has been obtained using the asymptotic approximation.

Refined Asymptotics

A refined three-term approximation is given in [9]. It has the form

$$P(W > x) = a_1 e^{-b_1 x} + a_2 e^{-b_2 x} + a_3 e^{-b_3 x}, \tag{2.67}$$

where a_1 and b_1 are the asymptotic constant and asymptotic decay rate as given by (2.66), while a_2, a_3, b_2 and b_3 are chosen to match the probability of delay $P(W > 0)$, and the first moments EW, $E(W^2)$ and $E(W^3)$, with b_2 and b_3 required to satisfy $b_1 \leq \min(b_2, b_3)$.

2.3.2 Bandwidth Management

One of the major applications of the effective bandwidth theory is bandwidth allocation. Bandwidth allocation and management falls in two major categories: (i) Independent and (ii) dynamic bandwidth allocation [4].

Independent Bandwidth Allocation

Let $\{c(t)\}$, $t \geq 0$, be an independent arrival process, so $\{c(t)\}$ is called an independent bandwidth allocation sequence. Consider an arrival process $a(t)$ with an energy function $\Lambda_a(\theta)$, so

$$Pr(q(\infty) \geq x) \approx e^{-\theta^* x}, \tag{2.68}$$

if θ^* is the unique solution of

$$\Lambda_a(\theta) + \Lambda_c(-\theta) = 0. \tag{2.69}$$

Let $\Lambda_c^*(\alpha)$ be the entropy function of $c(t)$ and α_c^0 be the global minimum of $\Lambda_c^*(\alpha)$, i.e., $\Lambda_c^*(\alpha_c^0) = 0$. The best bandwidth allocation is if the constant-rate bandwidth allocation sequence $c(t) = \alpha_c^0$. In view of the Legendre transformation,

$$\Lambda_c(-\theta^*) \geq -\theta^* \alpha - \Lambda_c^*(\alpha), \tag{2.70}$$

the last two equations yield

$$\alpha^*(\theta^*) = \frac{\Lambda_a^*(\theta^*)}{\theta^*} \leq \alpha_c^0 + \frac{\Lambda_c^*(\alpha_c^0)}{\theta^*} = \alpha_c^0, \tag{2.71}$$

where $\alpha^*(\theta^*)$ is the effective bandwidth function. Since it is increasing, a larger decay rate is met if the constant-rate bandwidth allocation sequence $c(t) = \alpha_c^0$. Hence, the optimal independent bandwidth allocation sequences are the constant-rate sequences.

Dynamic Bandwidth Allocation

Let the above defined sequence $\{c(t)\}$ depend on the arrival process. Specifically, let $\mu(\gamma)$ for $0 \leq \gamma \leq 1$ be a dynamic bandwidth allocation sequence with $\mu(\gamma)$ being the amount of bandwidth allocated when the buffer occupancy is γx. Since the optimal independent bandwidth allocation sequences are the constant-rate sequences, adding randomness to $\mu(\gamma)$ does not improve the performance of the tail distribution of the queue length. Hence, $\mu(\gamma)$ can be taken as a deterministic function.

Let $\mu(\gamma)$ be a piecewise linear function with $\mu(\gamma) = c_1$ for $\gamma \leq \gamma_1$, and $\mu(\gamma) = c_2$ for $\gamma > \gamma_1$. Let the arrival process $a(t)$ have the energy function $\Lambda_a(\theta)$, the entropy function $\Lambda_a^*(\alpha)$ and the effective bandwidth function $\alpha^*(\theta)$. For the buffer to exceed x at time t, the buffer should exceed $\gamma_1 x$ at some time $t_1 \leq t$. It then takes $t - tt_1$ to build up another $(1-\gamma_1)x$ cells. Hence, the problem can be separated to two parts: (i) The buffer to build up $\gamma_1 x$ cells with capacity c_1, (ii) the buffer to build up $(1-\gamma_1)x$ cells with capacity c_2. Only when these two events happen, the buffer exceeds x. Hence,

$$Pr(q(\infty) \geq x) \approx e^{-x(\gamma_1\theta^* + (1-\gamma_1)\theta_2^*)}, \tag{2.72}$$

where θ_i^* are the unique solutions of $\alpha^*(\theta) = c_i$. In general,

$$Pr(q(\infty) \geq x) \approx e^{-x\int_0^1 \theta_\gamma^* d\gamma}, \tag{2.73}$$

where θ_γ^* is the unique solution of $\alpha^*(\theta) = \mu(\gamma)$. Hence, the effective bandwidth for this problem is $\alpha^*(\bar{\theta})$, where

$$\bar{\theta} = \int_0^1 \theta_\gamma^* d\gamma. \tag{2.74}$$

2.3.3 Admission Control

Admission control is an important issue to be considered. Its main advantages are: (i) Prevention of network congestion, (ii) limiting calls and hence, guaranteeing a QoS.

The sources to share the resources are time-varying. Using the effective bandwidth approximation, they would be regarded as constant-rate sources with the rate being their effective bandwidth. So, the network layer and the higher layers need not be changed and this minimizes the impact of the new high-speed networks on the existing ones, such as the circuit-switching networks.

There have been some schemes proposed for connection acceptance control and we will focus on the most important ones.

Using Effective Load

Connection acceptance control is primarily concerned with expectation of the future QoS. In [28], an effective and robust use of prior declarations and empirical averages is made. The key idea was to use the prior declarations to choose a linear function that bounds the effective bandwidth function. Once this is done, the connection acceptance control would be based on relatively simple measurements needed to evaluate this function. Specifically, let

$$Z = E\{e^{sX[\tau,\tau+t]}\}. \tag{2.75}$$

So the effective bandwidth function is simply

$$\alpha(Z) = \frac{1}{st}\log(Z). \tag{2.76}$$

Before a call's admission, the user is required to enter a value z. The tangent to the curve $\alpha(Z)$ at $Z = z$ is

$$f(z;Z) = a(z) + b(z)Z. \tag{2.77}$$

Suppose that the resource has already accepted connections $1, 2, \ldots, I$, each with coefficients $(a(z_i), b(z_i))$, chosen by the users at the time of the connection. Suppose that the resource measures the load $X_i[\tau, \tau + t]$ produced by connection i over a period of length i and let $Y_i = \exp(sX_i[\tau, \tau + t])$. We may define an effective load on the resource as

$$\sum_{i=1}^{I} (a_i + b_i). \tag{2.78}$$

The connection acceptance control is, then, to accept or reject a connection according to the most recently calculated effective load.

Using Traffic Descriptors

In [5], two approaches for admission control are considered: (i) Case of QoS formulated in terms of loss probability, (ii) case of unknown parameters.

If QoS is expressed in terms of the loss probability, then before entering to the network, the service proposes to the network controller a service request including the source address, the destination address, the four-parameter traffic descriptor and the QoS. Using the effective bandwidth formulae, the loss probabilities are estimated and a decision as to accept the call or not is made.

In case of unknown traffic descriptor parameters, service may be partitioned into classes that have certain traffic descriptor parameters and satisfy a given QoS. A service to be accepted needs only to provide its class and a decision is made thereafter.

Admission Policies

In [18], the author investigates how different admission policies might decrease the effective bandwidth of a source. The policies considered are memoryless (i.e., they reject or set to lower priority a fraction of the arriving load). Conditions

are found so as to determine the admission policy that yields the lowest effective bandwidth (Proposition 4.1 in [18]).

Sophisticated Scheduling Policies

The admission control schemes considered so far assume a shared buffer. However, the case of distributed buffers reveals a challenging issue in admission control, since it requires the use of some sophisticated scheduling policies, such as the generalized processor sharing (GPS) or its ATM-oriented version, the packetized general processor sharing (PGPS). PGPS [17] is a work conservative policy that fairly guarantees a minimum bandwidth to a given FIFO.

One way to decide as to admit a connection or not is to calculate the spare capacity at a given buffer i at a PGPS node. The spare capacity is given by [17]

$$c - N_i\alpha_i(\delta_i) - \min[(1 - \phi_i c), \sum_{j \neq i} N_j \alpha_j(\delta_j)], \tag{2.79}$$

where c is the buffer service rate in cells/sec, N_i is the number of independent sources at buffer i, δ_i is a measure of the stringency of the QoS requirements and $\alpha_i(\delta_i)$ is the effective bandwidth of the source i.

2.3.4 Tariffing

In a practical setting, tariffing (or charging) is an important issue to be considered. An intuitive idea is to charge a service proportional to product of the effective bandwidth and the time used by the service.

2.3.5 Example of Charging Mechanism

Kelly [28] uses the same function as for the admission control, to characterize the tariffs $f(z; Z)$, namely

$$f(z;Z) = a(z) + b(z)Z. \tag{2.80}$$

defined as the tangent to the effective bandwidth curve $\alpha(Z)$ at $z = Z$. Note that z is entered by the user. The latter is assumed risk-neutral and willing to minimize the expected cost of the connection per unit time.

Malicious User

One important aspect is to be kept in mind. In most of the tariffing schemes, the inputs are assumed independent. However, Chang [5] states the case of malicious users. A user could send a file in a round robin fashion through 100 sessions instead of a single one. The 100 sessions are obviously not independent. The charge based on the effective bandwidth may come out less since they are assumed independent. One way to circumvent this, the network controller should compute the effective bandwidth of the 100 sessions under the maximum dependence assumption, in order to derive the worst case analysis.

2.3.6 Traffic Monitoring

Once a call is accepted, it is important to make sure that the connection is conforming to the negotiated traffic contract it is accepted upon. Traffic monitoring performs this function.

In [17], the authors propose a scheme of on-line monitoring with admission control to exploit unknown statistical multiplexing gains and thus increase utilization. This work will be illustrated in the next chapter.

2.3.7 Other Applications

The other applications of the effective bandwidth theory include: routing, source classification, fast simulation of intree networks, etc, to state just a few.

Chapter 3

On-Line Resource Management in ATM Networks: Source Characterization

3.1 Preview

As mentioned earlier, in ATM networks, the traffic pattern is expected to be heterogeneous, including data, voice, video, image and any combination of these. The traffic sources are bursty, with variable QoS requirements depending on the specific application of interest. Traffic workload would be emitting at several Megabyte/sec rates and cell loss probabilities (CLPs) are in the order of $10^{-4} - 10^{-9}$.

By resource management we refer to the optimal allocation of network resources, such as link capacity and buffer spaces, to the input traffic streams so as to accommodate their desired QoS. Resource management aims at achieving optimality in the use of the network, maximizing its efficiency, enhancing its reliability and thus ensuring its viability.

To perform a resource management task, input traffic needs to be characterized. Traffic characterization plays an important role in the design, control and management of networks that would be able to cope with bursty multi-media

traffic with guaranteed QoS. The effective bandwidth acts as a summary of the statistical characteristics of sources over time and space and thus enables to characterize the input traffic.

We have seen that the effective bandwidth is defined as the minimum bandwidth that needs to be allocated to a source so as to satisfy its negotiated QoS. In a small CLP context, it can be regarded as the fixed rate at which a bursty source would be transmitting. This unified metric characterizes, to a good approximation, the traffic that needs to be dealt with. It is shown to depend only on the source parameters and not the system. The effective bandwidth provides some intuitive yet important properties. Of great importance is its additivity and the fact that it lies between the mean and the peak rates of the source. On its calculation, it is simple and explicitly-identified. As a traffic descriptor, it is accurate, efficient and robust. Its applications cover most of resource management network tasks. The effective bandwidth theory is well-defined, rich and understood.

Nevertheless, very few attempts have been made so as to map the theory into direct measurement and estimation procedures. For the effective bandwidth to be a viable tool, *on-line*, real-time estimation schemes are essential. Our main concern, in this work, is to point at two procedures that make use of the effective bandwidth concept for on-line, real-time resource management in ATM networks.

In this chapter, the single source case is considered. The effective bandwidth is used to characterize, to an accurate approximation and via on-line estimation procedures, the underlying source. Based on the large deviation principle, the effective bandwidth is an asymptotic approximation to the tail probabilities of loss. A resource management application, namely traffic monitoring, is also illustrated.

Specifically, De Veciana et al. [17] proposed a 'virtual buffer model' that they used to perform real-time traffic monitoring. We adopt their model to estimate, via on-line measurement, the performance of a source. The idea is to introduce a virtual buffer at the user network interface (UNI) and estimate the relevant parameters of concern, for instance, CLP. Let us note that this work assumes Markovian albeit general arrival processes.

While this work was in progress, a new scheme by Mark et al. [32] was suggested. It enables characterization of general types of sources with no restrictive assumptions on the type of offered traffic load. This methodology is also studied and the underlying improvements over the previous scheme are emphasized.

Henceforth, our work is organized as follows. In this chapter, we review the 'virtual buffer model' for traffic characterization and see its application in traffic monitoring. Then, we present and simulate the new scheme and point out to its improved features.

In the next chapter, we investigate the case of several sources that are to compete for network resources, we show the importance of merging input streams together on a single link to account for statistical multiplexing gains and we suggest an on-line resource management application, namely, connection admission control (CAC).

3.2 First Order Exponent Approach

Recall from Section 2.3.1 that the proposed traffic descriptor is the effective bandwidth. It depends on both the statistical nature of the stream of cells and the nature of the required QoS constraint.

It has been shown [17] that for a buffered link with capacity c cells/sec supporting a stationary and ergodic arrival packet stream $A(0,t)$, the CLP is approximated by

$$P(X > B) \approx Ae^{-b\delta} \approx e^{-b\delta}, \tag{3.1}$$

where B is a reasonably large ATM buffer size, X is the buffer stationary workload and δ is a measure of the stringency of the QoS requirement. Explicitly $\delta = -\log(p)/B$, where p is the desired CLP.

The effective bandwidth function $\alpha(.)$ is shown in Fig. 3.1 and is given by

$$\alpha(\delta) = \delta^{-1} \lim_{t \to \infty} t^{-1} \log E\{\exp[\delta A(0,t)]\}. \tag{3.2}$$

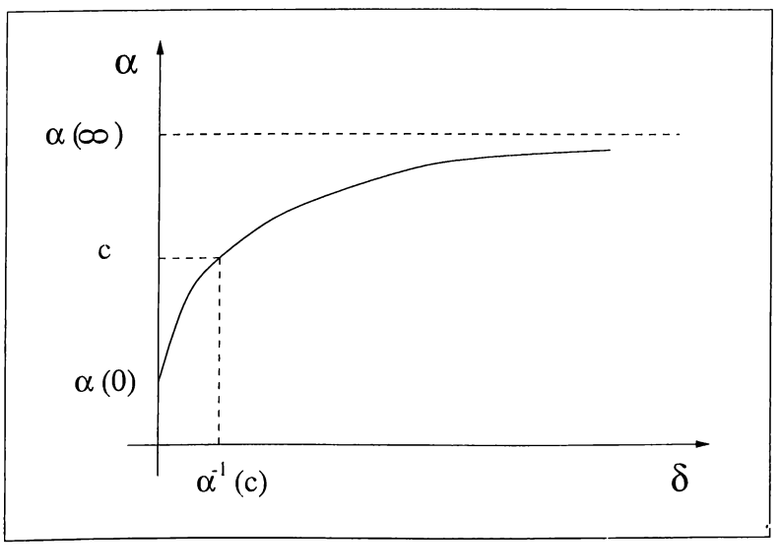

Figure 3.1: Effective Bandwidth Function

The effective bandwidth function $\alpha(.)$ is non decreasing in δ, with $\alpha(0)$ and $\alpha(\infty)$ being the mean and the peak rates of the source, respectively.

The δ constraint on X given by (3.1) matches that in (3.2) to the first order, i.e., in the exponent, for large buffer sizes. Hence, to the first order, the effective bandwidth is the minimum bandwidth required by a connection to accommodate its desired QoS.

3.2.1 Virtual Buffer Method

The on-line, real-time estimation of the effective bandwidth characteristic is done as follows. The source is connected to a virtual buffer at the UNI. An estimation probe is used to estimate the different parameters of interest so as to measure the CLP.

Model

The model proposed to accomplish this task is depicted in Fig. 3.2.

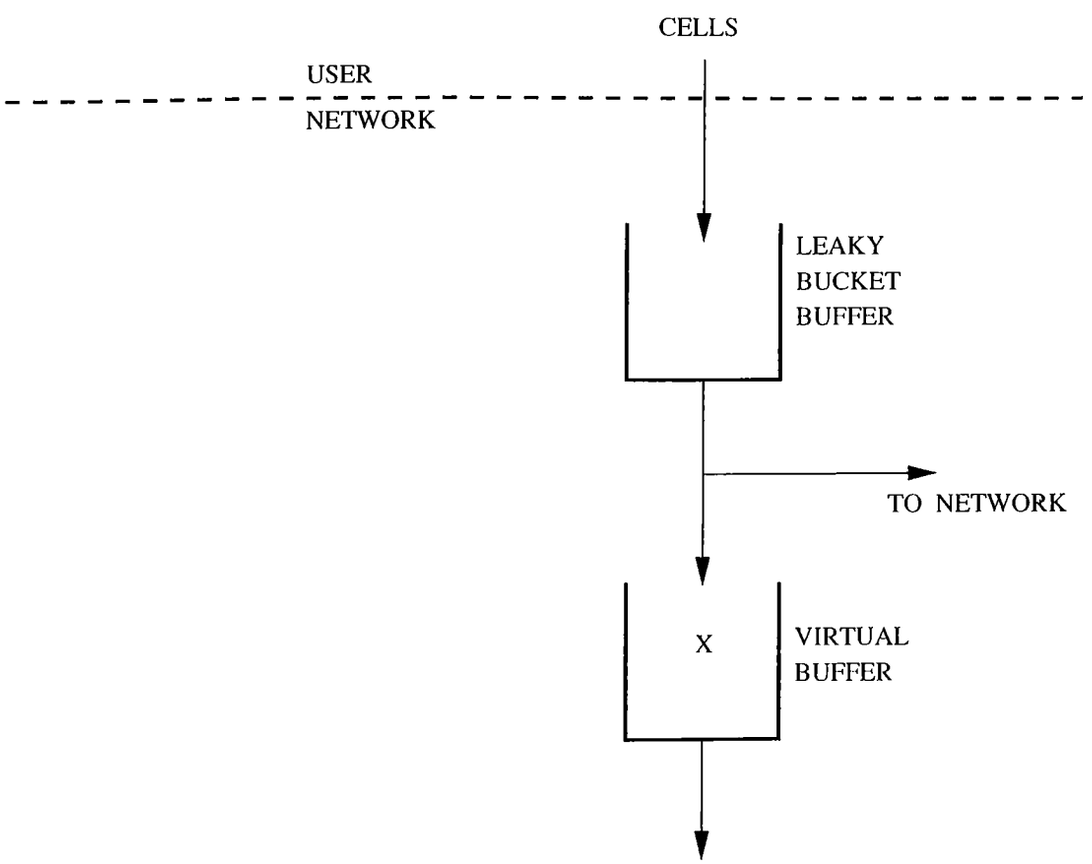

Figure 3.2: Virtual Buffer Model

Algorithm

The virtual buffer is assigned a deterministic service rate c and is used to estimate the QoS $\alpha^{-1}(c)$. To estimate $\alpha^{-1}(c)$, we have $P(X > B) = \exp[-B\alpha^{-1}(c)+o(B)]$, where X is distributed as the steady state of the workload of the virtual buffer. We take B_1 such that $P(X > B_1)$ is not too large and assume $P(X > B) = Ae^{-bI}$ for $b \geq B_1$, where A and I are the quantities to be estimated. Note that $I = \alpha^{-1}(c)$. The buffer workload is monitored over time and the empirical distribution $\pi(.)$ of the workload beyond B_1 is obtained.

A and I are chosen to minimize the Kullback-Leibler distance between π and $p(b) = A\exp(-bI)$ for $b \geq B_1$, and are given by

$$I = \log\left[1 + \frac{1 - \Pi(B_1 - 1)}{\sum_{b=B_1}^{\infty} b\pi(b) - B_1(1 - \Pi(B_1 - 1))}\right], \tag{3.3}$$

and

$$A = [1 - \Pi(B_1 - 1)]\exp(B_1 I), \tag{3.4}$$

where $\Pi(B_1 - 1) = \sum_{b=1}^{B_1-1} \pi(b)$.

3.2.2 Numerical Simulations

The source used in these simulations is an ON/OFF continuous time Markov chain, with peak rate $\lambda_p = 60.0$ and transition probabilities $P_{0,1} = 0.9$, $P_{1,0} = 0.1$. We choose $\delta = 0.6$, which corresponds to a buffer size $B = 7$ and a CLP $p = 10^{-4}$. Throughout this work, all simulations are run on OPNET, short for OPtimized Network Engineering Tools, a sophisticated workstation-based environment for the modeling and simulation of communication systems, protocols and networks.

Estimating I

A plot of an estimate of I versus time is given in Fig. 3.3.

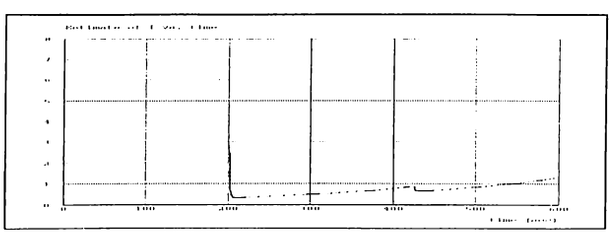

Figure 3.3: Estimate of I versus time

Estimating A

A plot of an estimate of A versus time is given in Fig. 3.4.

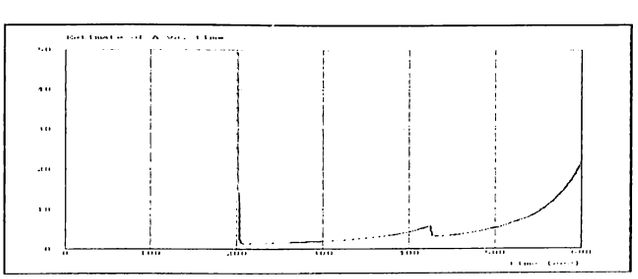

Figure 3.4: Estimate of A versus time

Estimating CLP

Fig. 3.5 shows the true CLP as measured in the simulations at the virtual buffer, an estimate of CLP corresponding to $CLP = Ae^{-IB}$, which we call the *true* asymptote and denoted by $C\hat{L}P$ and an estimate of the CLP as given by $CLP = e^{-IB}$, as given by the effective bandwidth approximation and denoted by $C\tilde{L}P$. The values obtained are the following. The true CLP=0.00022568, the effective

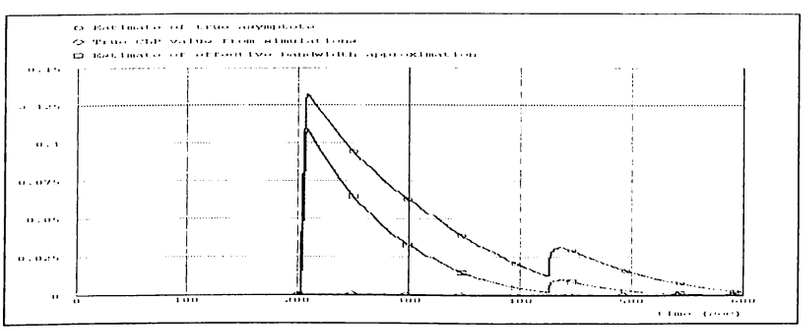

Figure 3.5: Estimate of True CLP, $C\hat{L}P$ and $C\tilde{L}P$ vs. time

bandwidth CLP=0.00009945 and the true asymptote CLP=0.00215182.

3.2.3 Comments

From the plots of the true CLP and the two estimates, we see that the effective bandwidth approximation, namely $CLP = e^{-IB}$ performs quite well. Actually, the true value of CLP lies between the two approximations. Moreover, it has been shown [2] that in the case of a single source model, this approximation works just fine. However, A may take arbitrarily large or small values which may render the approximation bad.

3.3 Traffic Monitoring

As ATM provides bandwidth on demand, peak rate policing will not suffice to ensure QoS and fairness. Consequently, connections violating agreed upon traffic descriptors must be throttled into compliance. An approach to accomplish this is given for a fixed value of δ.

3.3.1 Model

The model used to accomplish this task is depicted in Fig. 3.6.

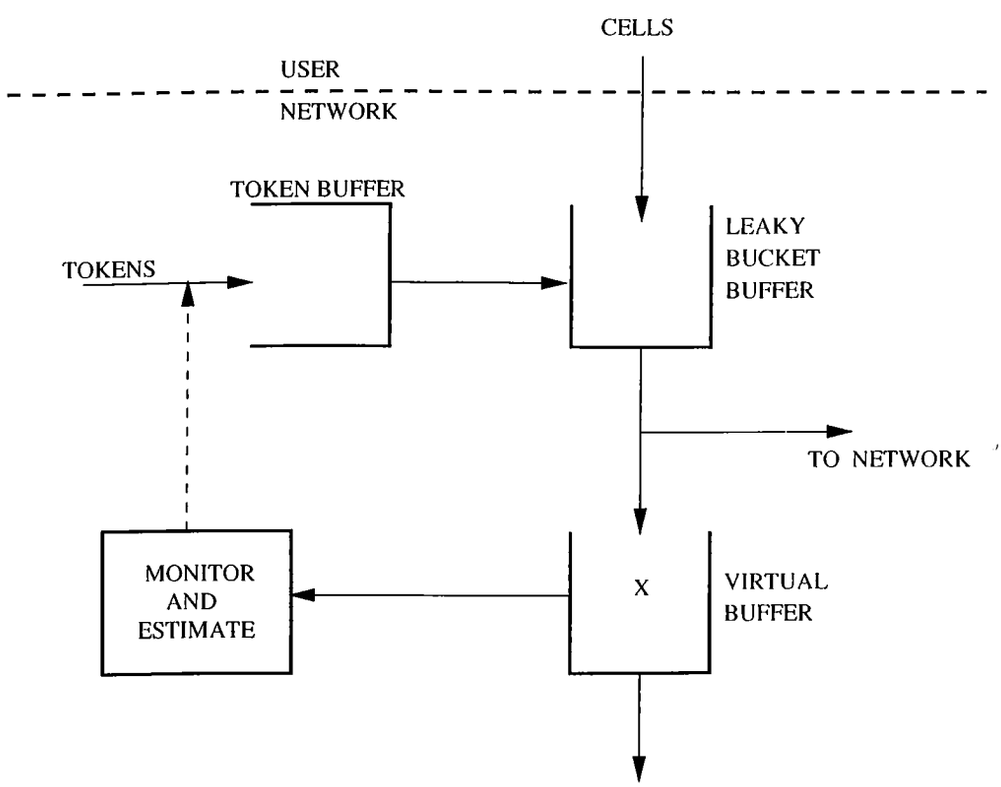

Figure 3.6: Virtual Buffer Model for Traffic Monitoring

3.3.2 Algorithm

Let β be the user-specified effective bandwidth and α be the true effective bandwidth. Let R be the token arrival rate. and let γ_R be the effective bandwidth of the departure process from the leaky bucket to the network. Initially, $R = \beta(\infty)$. So, the connection is, at first, largely unaffected by the leaky bucket. The virtual buffer is used to obtain an estimate of I over time. A user is said to have violated his effective bandwidth descriptor β at δ if

$$\gamma_{R(\delta)} > \beta(\delta) \qquad \text{or} \qquad I_t < \delta. \tag{3.5}$$

To enforce the traffic descriptor, we adjust the token rate R. If violation occurs, R is set to $\beta(\delta)$, so as to make the process entering the network compliant. R will be set again to $\beta(\infty)$ at the earliest time t such that $I_t > \delta$, i.e., compliance.

3.3.3 Numerical Simulations

The source used in these simulations is an ON/OFF continuous time Markov chain, with peak rate $\lambda_p = 60.0$ and transition probabilities $P_{0,1} = 0.9$, $P_{1,0} = 0.1$. We choose $\delta = 2$ which corresponds to a buffer size $B = 2$ and CLP $p = 10^{-4}$. B_1 is taken as 2. Tokens are generated at a constant rate depending on whether the source is compliant or not.

Case of Source in Violation

If no traffic monitoring is performed, the source violates the negotiated contract and the corresponding plot of I versus time is given in Fig. 3.7.

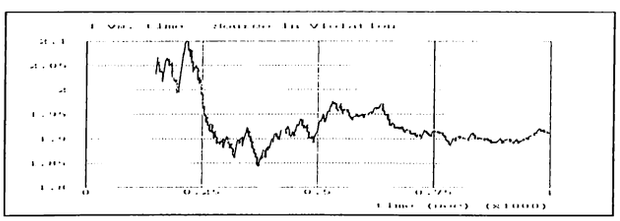

Figure 3.7: Estimate of I vs. time - Source in Violation

Note that in this case, A is different from 1 and is given in Fig. 3.8.

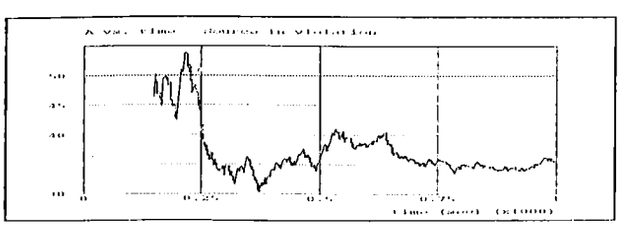

Figure 3.8: Estimate of A vs. time - Source in Violation

Case of Source in Compliance

Performing the traffic monitoring algorithm throttles the source into compliance and, in this case, the estimate of I versus time is given in Fig. 3.9.

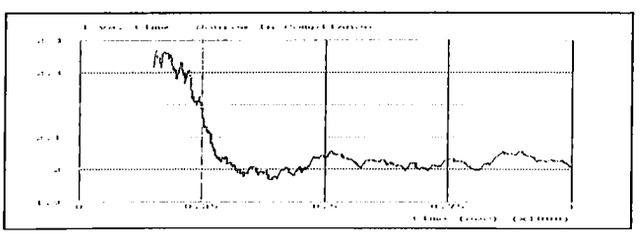

Figure 3.9: Estimate of I vs. time - Source in Compliance

Again, A takes values different from 1 as shown in Fig. 3.10.

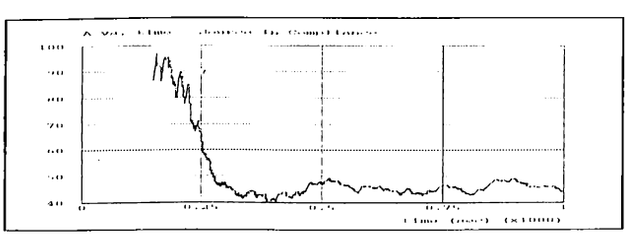

Figure 3.10: Estimate of A vs. time - Source in Compliance

3.3.4 Comments

Clearly, traffic monitoring is an important preventive procedure that aims to enforce the negotiated contract and force the user into compliance. By preventive, we refer to the ability of traffic monitoring to ensure that no congestion would be encountered in the network. Based on the above mentioned algorithm, the user is indeed throttled into compliance. Let us note again that the effective bandwidth approximation that assumes $A = 1$ may be misleading as A can take arbitrarily large or small values.

3.4 Approach Revisited

Owing to the importance of including the asymptotic constant, the restrictive Markovian assumptions on the arrival process and to the complexity of the expressions involving a large number of parameters that require in turn a large time-window for real-time estimation, the previous method has many shortcomings. A new scheme for traffic characterization has been suggested by Mark et al. [32]. This new model along with the derived expressions holds for a general arrival process. No restrictive assumptions are imposed on the source model. Moreover, the asymptotic constant is included in the computations.

In this new scheme, a finite-size leaky bucket with deterministic leak rate μ is also connected to the source. The overall model is a $G/D/1/B$ queue, with B being the buffer size. The measure of the source performance is the complementary waiting time distribution (CWTD), i.e.,

$$P(W > t) = ae^{-bt}, \tag{3.6}$$

where a, now, stands for the asymptotic constant and b is the asymptotic decay rate. a is a function of μ, the leaky bucket service rate which ranges between λ_m and λ_p, the mean and the peak rates of the source, respectively.

Mark et al. [32] have shown that

$$a = P(W > 0). \tag{3.7}$$

As for b, it is also a function of μ and is given by

$$b = \frac{a\mu}{\mu\tau_r a + q}, \tag{3.8}$$

where τ_r and q, both functions of μ, stand for the number of cells in the the queue and the remaining waiting time of the cell in service as seen by the next arrival, respectively.

In this line of thought, plots for a and b versus μ fully characterize the source. For a fixed buffer size B, the effective bandwidth is, then, the minimum value of μ that corresponds to the values of a and b that yield the desired QoS.

3.4.1 New Scheme

The user is to specify the peak rate λ_p, the mean rate λ_m and a measure of the desired QoS, such as the CLP. The network resources (buffer capacity B and service rate μ) ought to be allocated so as to meet the desired QoS.

As stated earlier, a measure for the QoS may be the CLP. Explicitly, a cell is lost if

$$X > B, \tag{3.9}$$

where X is a counter of the buffer workload.

This condition (3.9) is equivalent to

$$W > B/\mu, \tag{3.10}$$

where W is the waiting time (including time in service) of an arriving cell.

The CLP is given by [32]

$$P(W > B/\mu) = ae^{-bB/\mu}. \tag{3.11}$$

Another performance criterion may be formulated in terms of the delay experienced by any cell in the leaky bucket, i.e.,

$$P(D > d_{max}) < \epsilon_D. \tag{3.12}$$

The CWTD, standing for the probability that D exceeds a delay bound d_{max}, is given by

$$P(D > d_{max}) = P(W > B/\mu + d_{max}) = ae^{-bd_{max}}e^{-bB/\mu}. \tag{3.13}$$

Finally, note again that this scheme and the derived approximations hold for a general arrival process. No restrictive assumptions are imposed on the source model.

3.4.2 Implementation Issues

In this section, we address some implementation issues related to the estimation and measurement of several parameters related to Eqns. (3.7) and (3.8), re-written below, for convenience.

$$a = P(W > 0), \tag{3.14}$$

$$b = \frac{a\mu}{\mu\tau_r a + q}. \tag{3.15}$$

Algorithm

We allow some N_i arrivals to pass ($i = 1, \ldots, M$).

Let S_i, Q_i and T_i denote the number in service, the number in queue and the remaining time for the customer in service, respectively.

An estimate for a is

$$\hat{a} = \frac{1}{M}\sum_{i=1}^{M} S_i.$$

An estimate for q is

$$\hat{q} = \frac{1}{M}\sum_{i=1}^{M} Q_i.$$

And an estimate for τ_r is

$$\hat{\tau_r} = \frac{1}{\hat{a}M}\sum_{i=1}^{M} T_i.$$

Model

The model used to carry out this simulation is shown in Fig. 3.11.

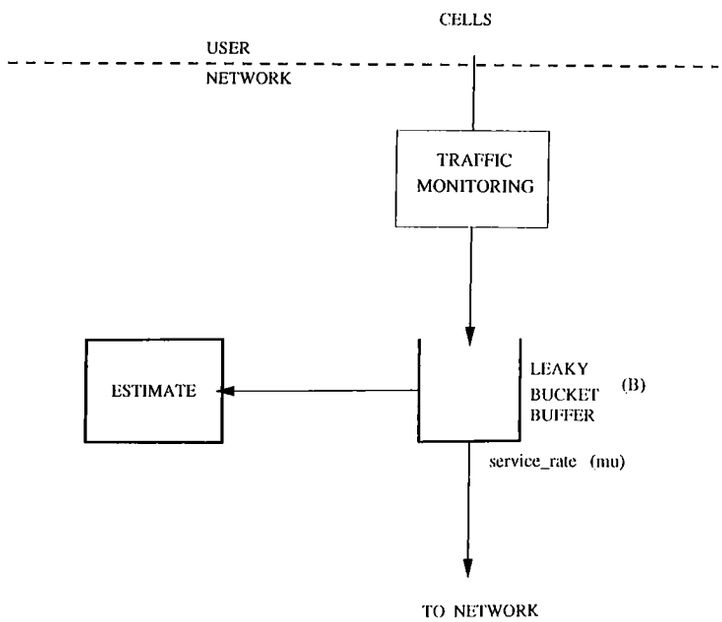

Figure 3.11: Model for New Scheme

3.4.3 Numerical Simulations

In this section, we simulate numerically the algorithm suggested in the previous section.

Since plots of a and b versus μ are at sake, we modify the model in Fig. 3.11 to one in which the source is directly connected to five fictitious buffers each with an individual service rate value, namely μ_i (for $i = 1, 2, \ldots, 5$) and finite buffer size B. Note that the only condition on μ_i is to be within the mean and the peak rates of the source, i.e., $\lambda_m \leq \mu_i \leq \lambda_p$, for $i = 1, 2, \ldots, 5$.

Hence the modified model is as in Fig. 3.12.

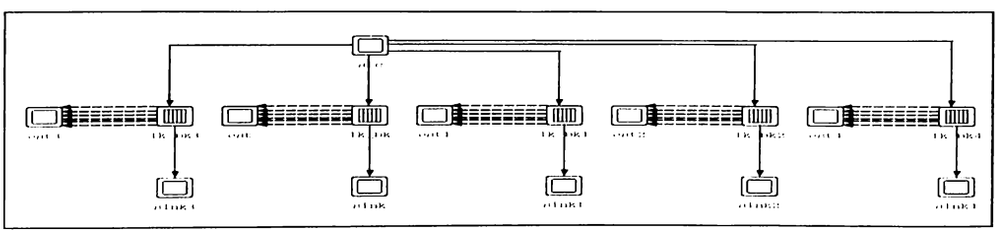

Figure 3.12: Modified Model for New Scheme

As stated earlier, the user is to specify the peak rate λ_p, the mean rate λ_m and either the CLP p or the maximum delay bound d_{max}. For a fixed buffer size B, the problem is to allocate a minimum service rate μ so as to guarantee the negotiated QoS. The source is an ON/OFF continuous-time Markov chain with the ON and OFF periods distributed according to Erlang-k distributions. The peak rate $\lambda_p = 60$ cells/sec, the mean rate $\lambda_m = 15.0$ cells/sec and $k = 1$. The user also asks for a CLP not exceeding $p = 10^{-4}$ and a maximum delay bound not exceeding $d_{max} = 0.2$. The buffer size B is fixed to 20.

Estimating a

To fully characterize the source, we need to obtain, among others, a plot of an estimate of a versus μ. Fig. 3.13 shows $\hat{a}$ versus μ.

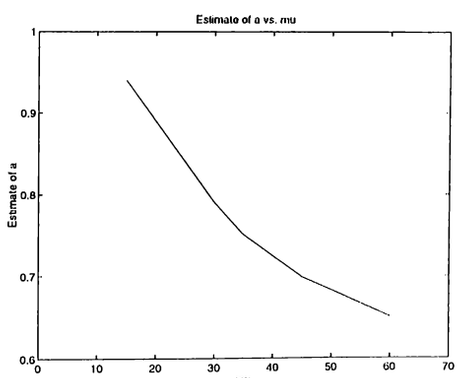

Figure 3.13: Estimate of a versus μ

Estimating b

Once an estimate for a is found, one needs to find an estimate of b as a function of μ. Fig. 3.14 shows $\hat{b}$ versus μ.

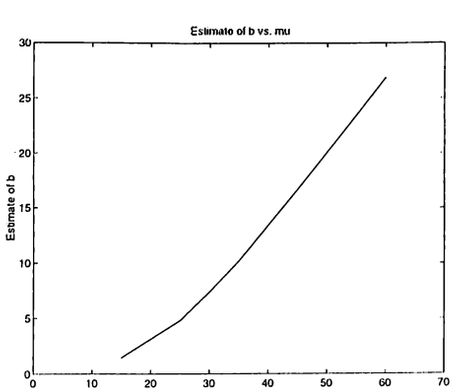

Figure 3.14: Estimate of b versus μ

Estimating CLP

An estimate of the CLP, $\hat{p}$, is measured according to the above mentioned procedure through the formula:

$$P(W > B/\mu) = ae^{-bB/\mu}. \tag{3.16}$$

To check the accuracy of (3.16), we compare the estimated value $\hat{p}$ to the true value p obtained directly from the simulations.

Fig. 3.15 shows p and $\hat{p}$ vs. μ.

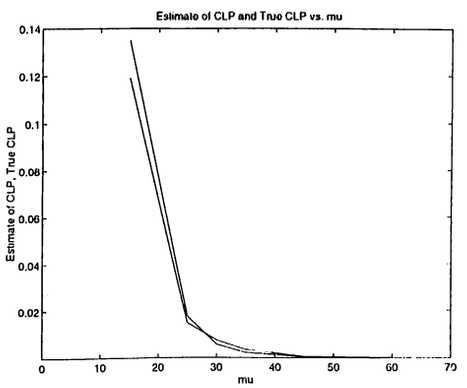

Figure 3.15: Estimate of p and $hatp$ versus μ

Fig. 3.16 magnifies p and $\hat{p}$ vs. μ, for small values only.

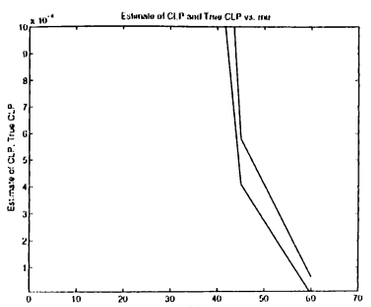

Figure 3.16: Estimate of p and $\hat{p}$ vs. μ - Magnified Figure

Estimating CWTD

In the case of the QoS given in terms of a maximum delay bound, the formula to be used to estimate the CWTD is:

$$P(D > d_{max}) = ae^{-bd_{max}}e^{-bB/\mu}. \tag{3.17}$$

Herein, we check the extent to which this estimation is correct. We plot an estimate $\hat{p_d}$ and the true value p_d of the CWTD. For values of μ close to the mean rate of the source, our estimate of the CWTD is somehow biased. Subtracting the mean delay from the estimate of the CWTD yields an unbiased estimate of the CWTD at those critical values.

Fig. 3.17 shows p_d, $\hat{p_d}$ and the unbiased estimate of the CWTD versus μ.

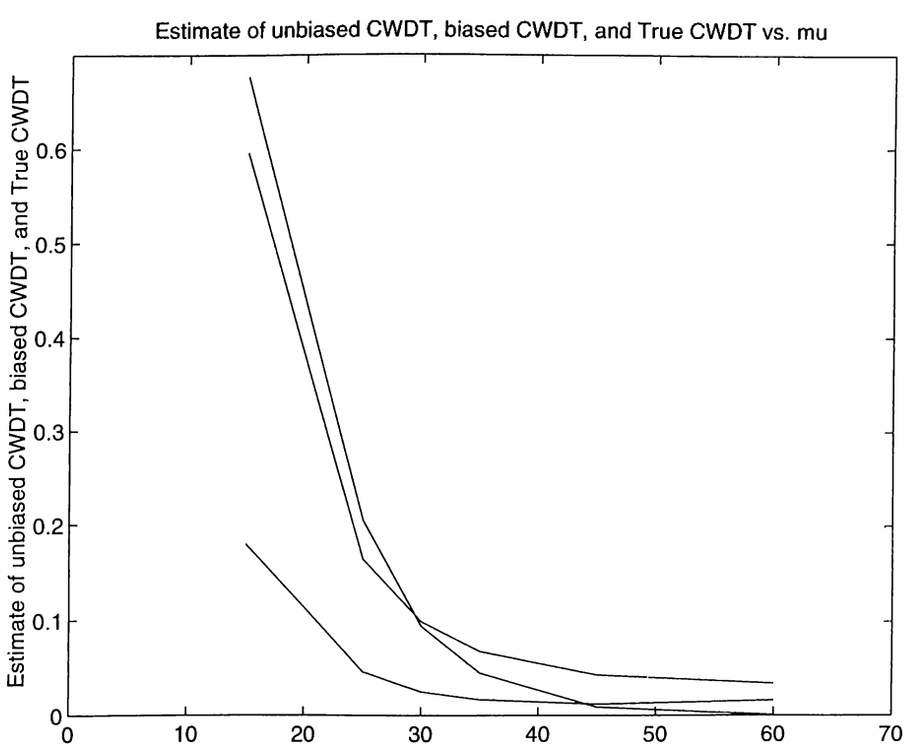

Figure 3.17: Estimate of p_d and $\hat{p_d}$ versus. μ

3.4.4 Comments

We have seen a scheme that allows the estimation, via on-line measurements, of parameters that characterize the source. This enables an optimal allocation of the resources so as to guarantee the negotiated QOS.

As can be seen, a is a decreasing function of μ and b is an increasing function of μ. Once estimates for a and b versus μ are obtained for a fixed buffer size B, a minimum value of μ can be found so as to meet the QoS of a connection.

The suggested estimates for CLP and CWTD perform well and approximate the true values to a good extent, mainly for large values of μ. At values of μ close to the source mean rate, we could implement an unbiased estimate of the CWTD and restore any lost accuracies.

It is important to note that the above work has been done for a general ON/OFF source and still holds for an even more general source model.

3.5 Discussion of Results

In this chapter, we have seen on-line, real-time schemes to characterize the sources of traffic. The first is shown to perform well in traffic monitoring. Traffic monitoring is an important network management task for it prevents further traffic congestion in the network. However, this scheme does not account for the asymptotic constant which it assumes to be 1.

The effective bandwidth approximation is

$$P(X > B) \approx ae^{-bB} \approx e^{-bB}. \tag{3.18}$$

Actually, as long as a single source is considered, this approximation performs quite well. For very small CLP values, approximation (3.18) has been shown [9] to hold well for values of a larger than 10^{-4} and smaller than 10^{4}, which is the case in a wide range of problems. Whenever a is out of this range, it should be

included in the computations. Besides, the real value of CLP (obtained directly from our simulations) lies between the effective bandwidth approximation and the true asymptote. The main limitation of this scheme is that it assumes Markovian arrival processes only.

As of the second scheme, it does not impose any restrictive assumptions on the arrival process. The latter is assumed to be a general one. Moreover, the asymptotic constant is included. In addition, this new scheme evaluates the performance of a source in terms of both CLP and CWTD. Though the latter may display some bias at values of the buffer service rate close to the mean rate of the source due to the highly variable delay (jitter) of the arrival process in these regions, an unbiased estimate could be reached. In this sense, this new methodology outperforms the previous one.

The case of multi-input streams each characterized by an individual effective bandwidth and that are to compete for network resources is especially of interest. As the new scheme enables direct on-line measurement of general source characterization, it seems promising for resource management applications. In this line of thought, the next step would be to consider its application to CAC which is the main topic of the next chapter.

Chapter 4

On-Line Resource Management in ATM Networks: Multi-Input Case

4.1 Preview

So far and in the last chapter, we have seen on-line, real-time schemes to characterize the traffic. The effective bandwidth and its asymptotic approximation of the tail probabilities of loss and/or delay was used to perform this task. Once a source is characterized, resource management network applications can be addressed. The new picture is one in which several sources, each with an individual effective bandwidth and QoS requirements, are to compete for the network resources, including mainly link capacity and buffer space.

There are two main approaches in addressing these application issues.

Once a source is characterized, network tasks, such as connection admission control (CAC), can be straightforward. As far as CAC is concerned, a source is, *simply*, admitted to the network if enough resources can be found so as to meet its desired QoS requirements. However, this is to do without exploiting any statistical multiplexing gains.

It is important to note that with this approach sources are treated in isolation from each other and on an individual basis. Resources are not optimally managed, for no statistical multiplexing gain is exploited. Resource management in ATM networks relies essentially on the capability of the network design to perform statistical multiplexing and account for the encountered gains. And this is the *raison d'être* of merging input streams together on a single link. It is a practice that allows statistical multiplexing of the input streams, hence, achieving gains. Based on this second approach and owing to the dynamic nature of the multiplexing procedure, network applications pose complex problems.

In this chapter, among the different network applications and within the resource management context, CAC is considered. Clearly, this amounts to admitting or rejecting a call connection request to the network, judging upon available resources and offered workload.

In the first section, we show a CAC procedure that is performed on an individual basis, i.e., input traffic streams are not multiplexed together on a single link. Though this practice is sub-optimal, its simplicity is rather attractive. In the next section, we treat the case of merging several input streams together and study the resultant performance. Based on this approach, we suggest a novel CAC scheme that accounts for the statistical gains and helps to achieve optimality in resource management.

4.2 Connection Admission Control

In this section, we present a CAC algorithm that operates on an individual basis. Each source is characterized by an individual effective bandwidth α_i and individual QoS requirement δ_i.

4.2.1 Model

For CAC, the model in Fig. 4.1 may be used. For the sake of illustration, two sources are considered in this model. An extrapolation to a larger number of sources is straightforward.

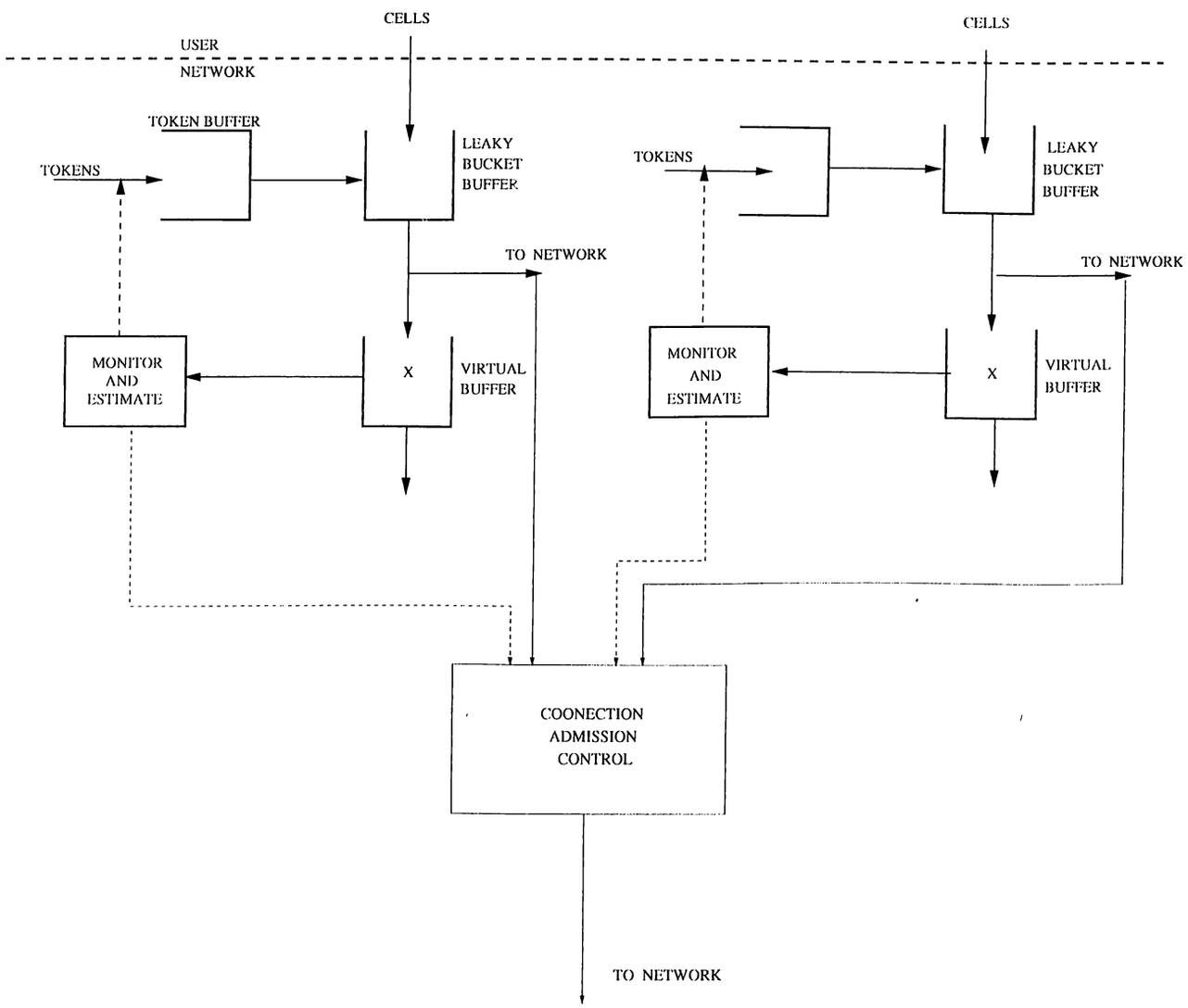

Figure 4.1: Model for CAC

Note that, at this stage, sources are assumed already monitored or being monitored, so no further comments on traffic monitoring would be made.

4.2.2 Algorithm

We recall that the CLP has been approximated by $P(X > B) = Ae^{-IB}$, where X is the buffer workload, B is the buffer size and I is the inverse function of the effective bandwidth, i.e., $I = \alpha^{-1}(c)$. Furthermore, the effective bandwidth approximation assumes $A = 1$.

The connections that are to be eventually accepted to the network are led to a buffered link with buffer size B and QoS capacity Δ. A clear definition of Δ is as follows. As the QoS of a connection is defined in terms of $\delta = -\frac{\log p}{B}$, where p is the desired CLP and B is the buffer size, for a maximum number of connections n, each with best performance CLP measure $p_i = 10^{-9}$, $i = 1, \ldots, n.$, $\Delta = \sum_{i=1}^{n} -\frac{\log p_i}{B} = \frac{9n}{B}$. Hence, a call would be accepted if it does not exceed this capacity Δ. For the effective bandwidth approximation, i.e., $A = 1$, this amounts to

$$I_i < \Delta.$$

For the asymptotic approximation, i.e, including the asymptotic constant A, the condition to be satisfied is

$$I_i + \log(\tfrac{A_i}{B}) < \Delta.$$

Note that both approximations are additive and a call is accepted when a spare capacity can be allocated to it without disturbing the previous accepted calls. In the effective bandwidth approximation, a new call is accepted if

$$I_i < \Delta - \sum_j I_j. \tag{4.1}$$

For the asymptotic approximation, a new connection is accepted if

$$I_i + \log(\frac{A_i}{B}) < \Delta - \sum_j \Big(I_j + \log(\frac{A_j}{B})\Big). \tag{4.2}$$

4.2.3 Numerical Simulations

In these simulations, two sources are to apply for admission. The case of a larger number of sources is treated in the same fashion. A source that is not admitted is kept tuned until a spare capacity is found and hence, the connection is accepted.

The following four cases sum up the major scenarios that can be encountered and compare the effective bandwidth approximation (A is not included) and the true asymptote approximation (A is included).

Case I

For the first case, $\Delta = 20.0$.

The following results are obtained,

```
set 1 -

Delta=20.0

time = 500.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.189294 : Accepted
src 2 : I = 2.776793 : Accepted
Results for True Asymptotic Approximation
src 1 : I = 2.189294 : A = 66.075301 : Accepted
src 2 : I = 2.776793 : A = 233.217714 : Accepted

time = 1000.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.039051 : Accepted
src 2 : I = 2.952725 : Accepted
```

```
Results for True Asymptotic Approximation
src 1 : I = 2.039051 : A = 47.247815 : Accepted
src 2 : I = 2.952725 : A = 335.917177 : Accepted
```

Comments

In this case, the capacity Δ is large enough to accommodate both calls.

Case II

For the second case, $\Delta = 3.0$.

The following results are obtained,

```
set 2 -

Delta=3.0

time = 500.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.189294 : Accepted
src 2 : I = 2.776793 : Rejected
Results for True Asymptotic Approximation
src 1 : I = 2.189294 : A = 66.075301 : Rejected
src 2 : I = 2.776793 : A = 233.217714 : Rejected

time = 1000.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.039051 : Accepted
src 2 : I = 2.952725 : Rejected
Results for True Asymptotic Approximation
```

```
src 1 : I = 2.039051 : A = 47.247815 : Rejected
src 2 : I = 2.952725 : A = 335.917177 : Rejected
```

Comments

In this case, the capacity Δ is too small to admit the second connection.

Case III

For the third case, $\Delta = 4.8$

The following results are obtained,

```
set 3 -

Delta=4.8

time = 500.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.189294 : Accepted
src 2 : I = 2.776793 : Rejected
Results for True Asymptotic Approximation
src 1 : I = 2.189294 : A = 66.075301 : Rejected
src 2 : I = 2.776793 : A = 233.217714 : Rejected

time = 1000.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.039051 : Accepted
src 2 : I = 2.952725 : Rejected
Results for True Asymptotic Approximation
src 1 : I = 2.039051 : A = 47.247815 : Rejected
src 2 : I = 2.952725 : A = 335.917177 : Rejected
```

```
time = 1500.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.13346 : Accepted
src 2 : I = 2.535677 : Accepted
Results for True Asymptotic Approximation
src 1 : I = 2.039051 : A = 47.247815 : Rejected
src 2 : I = 2.952725 : A = 335.917177 : Rejected
```

Comments

Note that at the beginning connection 2 was not accepted. It may be that the source was not well monitored yet. Some time later, the source was such that it could satisfy the spare capacity and hence was admitted to the network.

Case IV

For the fourth case, $\Delta = 5.0$.

The following results are obtained,

```
set 4 -

Delta=5.0

time = 500.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.189294 : Accepted
src 2 : I = 2.776793 : Accepted
Results for True Asymptotic Approximation
src 1 : I = 2.189294 : A = 66.075301 : Rejected
src 2 : I = 2.776793 : A = 233.217714 : Rejected
```

```
time = 1000.000000

Results for Effective Bandwidth Approximation
src 1 : I = 2.039051 : Accepted
src 2 : I = 2.952725 : Accepted
Results for True Asymptotic Approximation
src 1 : I = 2.039051 : A = 47.247815 : Rejected
src 2 : I = 2.952725 : A = 335.917177 : Rejected
```

Comments

Clearly, taking A into consideration may affect the results. A connection that has been accepted based on the effective bandwidth approximation may be rejected according to the asymptotic approximation.

4.2.4 Comments

For the virtual buffer method, traffic monitoring is very efficient and leads to very good results. As for CAC, when the effective bandwidth approximation is used (A is set to 1), this may mislead the decision. Including A is important for the accuracy of the CAC.

Clearly, the effective bandwidth approximation may show some weakness in the case of large A (underestimation of link capacity), or small A (overestimation of link capacity). Hence, including A may restore invaluable inaccuracy.

Yet, this CAC scheme, though sub-optimal as it does not account for statistical multiplexing gains, is rather simple and with the inclusion of A yields quite satisfactory results.

4.3 Multiplexing Input Streams - CAC revisited

Optimal resource management in ATM networks relies on the ability of the management procedure to account for statistical multiplexing and exploit the gains found therein. In CAC, for instance, treating sources on an individual basis and in isolation yields poor resource management results as it does not profit from any multiplexing gains. This leads us to the importance of merging the input streams together on a single link and proceed with the network management on a multiplexed basis.

4.3.1 Theoretical Preliminaries

The effective bandwidth α is defined as the minimum bandwidth a source is allocated so as to satisfy a given QoS. Likewise, we define a new measure, the *effective buffer size* β, as the minimum buffer size to be allocated to a source so as to satisfy QoS. Resource management reduces to finding the optimal yet feasible pair (α, β) that would meet the negotiated QoS.

Based on work found in [29], expressions for β are derived and are written here with their respective derivations.

In the case of the QoS measure p given in terms of CLP, we recall that CLP is given as

$$P(W > B/\mu) = ae^{-bB/\mu}, \tag{4.3}$$

where W, B and μ stand for the waiting time, buffer size and buffer service rate, respectively. a and b, both functions of μ, are the parameters to be estimated.

Setting $P(W > B/\mu) = p$ and solving for B yields

$$B_{clp} = \frac{\mu}{b} log(\frac{a}{p}). \tag{4.4}$$

In the case of the QoS given in terms of a delay experienced by a packet

passing through the leaky bucket, CWTD is given as

$$P(W > B/\mu + d_{max}) = ae^{-bd_{max}}e^{-bB/\mu}, \tag{4.5}$$

for a maximum tolerated delay d_{max}.

Again, setting $P(W > B/\mu + d_{max}) = p$ and solving for B gives

$$B_{cwdt} = \frac{\mu}{b}[log(\frac{a}{p}) - bd_{max}]. \tag{4.6}$$

The method to characterize a source suggested in [32] and tested in Chapter 3, Section 3.4, proceeds as follows. The user declares a mean rate λ_m, a peak rate λ_p and a measure for the stringency of the required QoS, p. The latter can be formulated in terms of CLP and/or CWTD.

The source is observed over a window of time during which the declared λ_m and λ_p are checked for and the source is characterized. By the latter, we mean that the source is given various fictitious link capacities μ ranging from λ_m to λ_p and for each μ, estimates of the CLP and/or CWTD are obtained. Based on these estimations, a certain QoS measure corresponds to a pair (α, β).

One may believe that once sources are characterized, network tasks, such as CAC, are straightforward. A source is, simply, admitted to the network if a resource pair (α, β) can be found. However, this is to do without exploiting any statistical multiplexing gains.

4.3.2 Algorithm

Our method is, therefore, formed of two stages. At a first stage, we obtain, for every source that is to compete for the resources and in an on-line fashion, the optimal characterization pair (α, β). We, then, sum up the individual α's and β's which yields the following total parameters B_t and C_t

$$B_t = \sum_i \beta_i, \tag{4.7}$$

and

$$C_t = \sum_i \alpha_i.^1 \tag{4.8}$$

We multiplex the input streams together and offer the resulting workload to a single buffer with buffer size B_{mux} and link capacity C_{mux} satisfying

$$B_{mux} \geq B_t, \tag{4.9}$$

and

$$C_{mux} \geq C_t. \tag{4.10}$$

Our results, presented in the next subsection, show that for a fixed C_t satisfying Eqn. (4.8), B_t is less than $\sum_i \beta_i$. Equality (4.7) holds for a total buffer size B'_t satisfying

$$B'_t < B_t, \tag{4.11}$$

which is the sought for statistical multiplexing gain.

Hence, CAC, the second stage of our work, comes into the picture. Keeping track of the current buffer occupancy B_c and link capacity used $C_c = C_t$ already allocated to some connections, a call with parameters (α, β) is admitted if

$$\beta \leq B_{mux} - B_c, \tag{4.12}$$

and

$$\alpha \leq C_{mux} - C_c. \tag{4.13}$$

At first, B_c is initialized to B_t as given by Eqn. (4.7). Then, B_c is regularly updated, via the on-line estimation procedure, to B'_t, the real (effective) buffer occupancy resulting from the statistical multiplexing. In this manner, sources are accepted on a solid basis, i.e., a source is accepted whenever enough resources are available ($B_c = B_t$), then buffer occupancy is re-evaluated and updated to its true (effective) value, i.e., $B_c = B'_t$.

[1] Eqn. (4.8) is based on a well established property of the effective bandwidth stating that the latter is additive.

4.3.3 Model

The model used to merge input traffic streams together is shown in Fig. 4.2. Note that three sources are used to illustrate our work. The case of a greater number of sources is treated in a similar fashion and is straightforward.

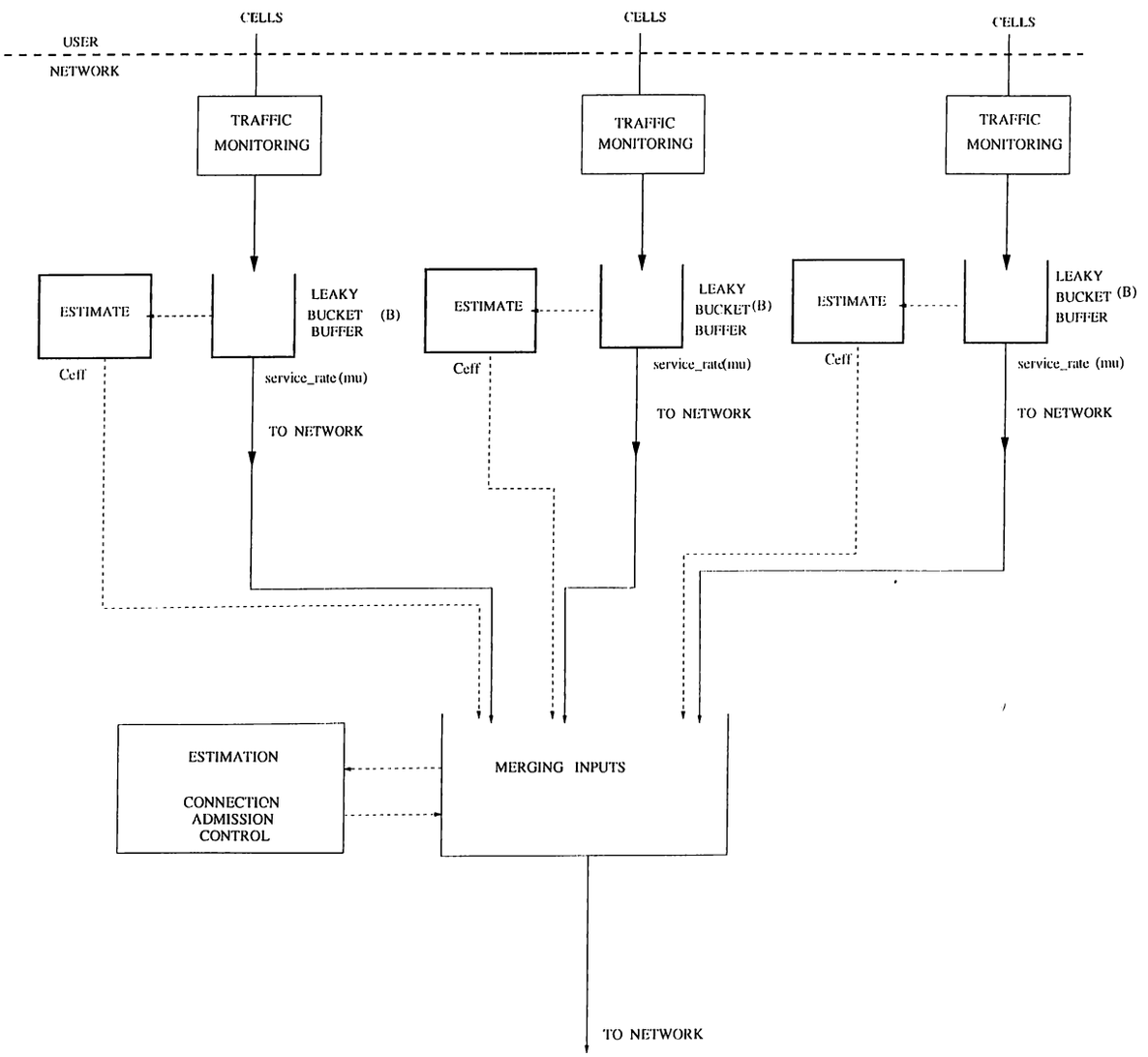

Figure 4.2: Model for Merging Input Streams and Performing CAC

4.3.4 Numerical Simulations

In these simulations, the units used are the following. Rates are given in cells per second and buffer sizes in cells.

Stage 1

Source 1-

The source has parameters : $\lambda_m = 20$ and $\lambda_p = 40$. It is to have a CWTD not exceeding $p = 10^{-4}$. An estimate for the CWTD as a function of μ is given in Fig. 4.3. We note that $CWTD = p$ for a value of $\mu = 30$. Hence, $\alpha_1 = 30$. A

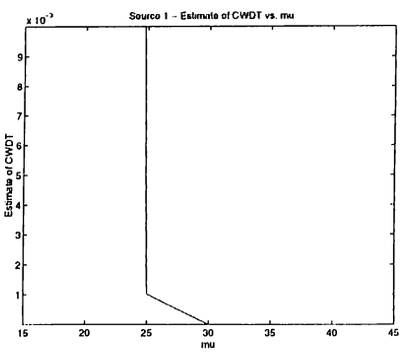

Figure 4.3: Source 1- Estimate of $CWTD$ vs. μ

plot for B_{cwtd} as a function of μ is given in Fig 4.4. For $\mu = 30$, $\beta_1 = 290$.

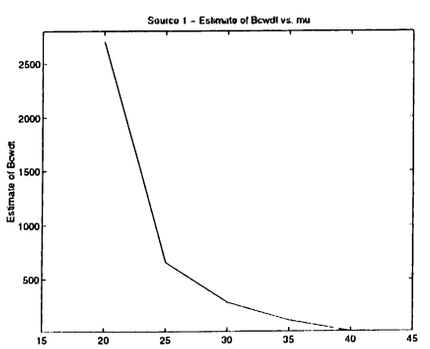

Figure 4.4: Source 1- Estimate of B_{cwtd} vs. μ

Source 2-

The source has parameters : $\lambda_m = 25$ and $\lambda_p = 50$. It is to have a CLP not exceeding $p = 10^{-4}$. An estimate for the CLP as a function of μ is given in Fig. 4.5.

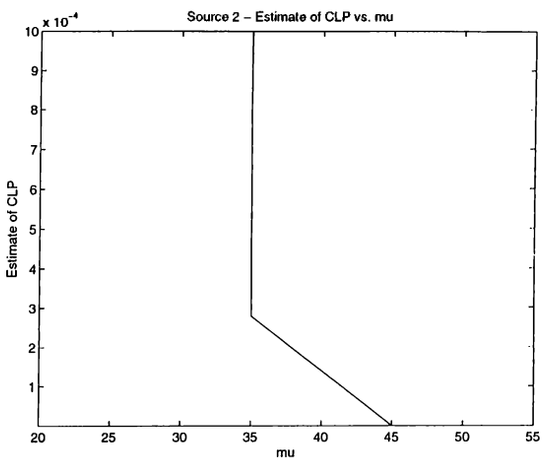

Figure 4.5: Source 2- Estimate of CLP vs. μ

We note that $CLP = p$ for a value of $\mu = 41.5$. Hence, $\alpha_2 = 41.5$. A plot for B_{clp} as a function of μ is given in Fig. 4.6.

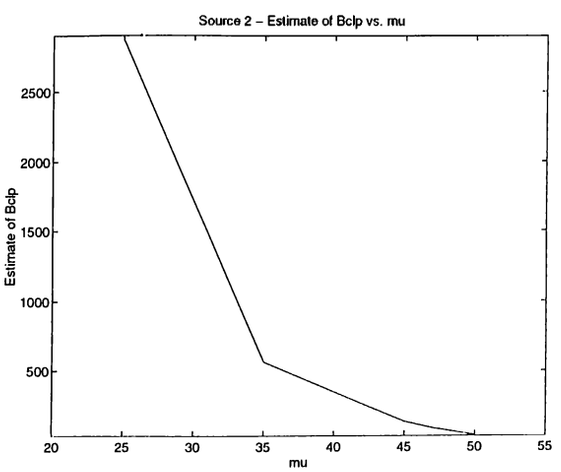

Figure 4.6: Source 2- Estimate of B_{clp} vs. μ

For $\mu = 41.5$, $\beta_2 = 135$.

Source 3-

The source has parameters : $\lambda_m = 30$ and $\lambda_p = 60$. It is to have both a CLP not exceeding $p = 10^{-4}$ and a CWTD not exceeding $p = 10^{-4}$. An estimate for the CLP as a function of μ is given in Fig. 4.7.

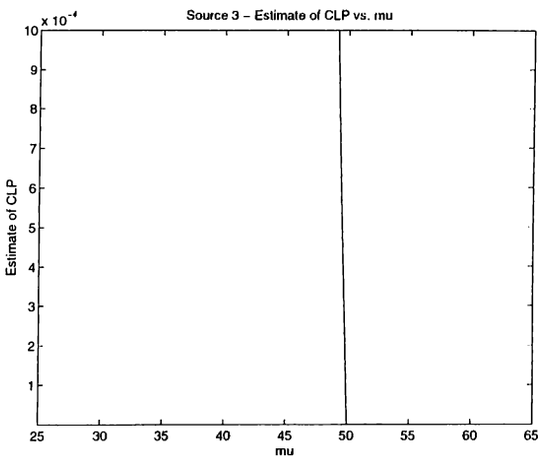

Figure 4.7: Source 3- Estimate of CLP vs. μ

We note that $CLP = p$ for a value of $\mu = 50$. Correspondingly, a plot for B_{clp} as a function of μ is given in Fig. 4.8.

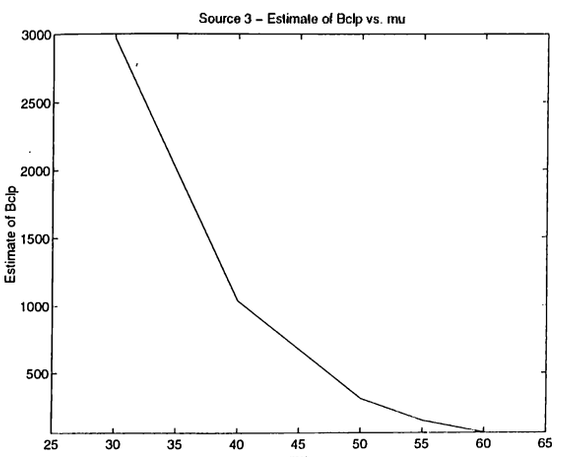

Figure 4.8: Source 3- Estimate of B_{clp} vs. μ

For $\mu = 41.5$, $B^*_{clp} = 280$.

An estimate for the CWTD as a function of μ is given in Fig. 4.9.

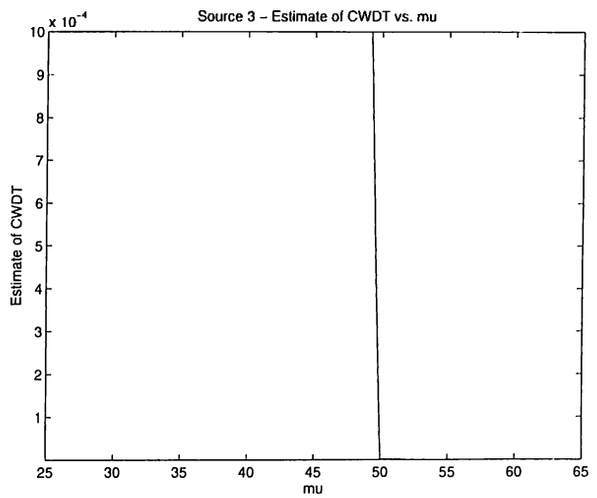

Figure 4.9: Source 3- Estimate of $CWTD$ vs. μ

We note that $CWTD = p$ for a value of $\mu = 50$. A plot for B_{cwtd} as a function of μ is given in Fig. 4.10.

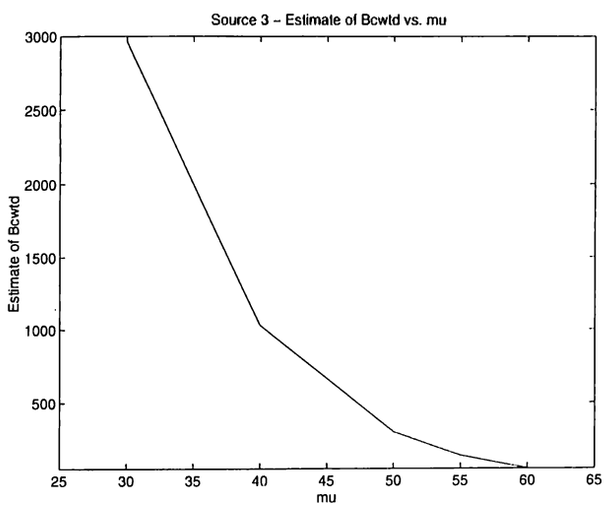

Figure 4.10: Source 3- Estimate of B_{cwtd} vs. μ

For $\mu = 50$, $B^*_{cwtd} = 270$.

Hence, $\alpha_3 = 50$ and $\beta_3 = \max(B^*_{cwtd}, B^*_{clp}) = 280$.

Stage 2

Coming to the second stage of our work, the sources are merged together and lead to a single leaky bucket with size B_t (given by Eqn. (4.7)) and service rate C_t (given by Eqn. (4.8)).

In our case, $C_t = \sum_{i=1}^{3} \alpha_i = 121.5$ and $B_t = \sum_{i=1}^{3} \beta_i = 705$.

With these values assigned to the buffer, the obtained results are the following.

```
results

service rate = 121.5
buffer size = 705
a  = 0.78
b  = 6.25
est_clp = 0.00000000
est_cwtd = 0.00000000
bclp = 174.1
bhsp = 149.8
```

We note that $p = 10^{-4}$ can be met for a smaller buffer size B_t', as expected from our analysis of the statistical multiplexing. Of a great importance and as in the case of individual sources, we have the values for the total effective buffer size that would meet the QoS for both CLP and CWTD specifications. For the case of achieving the desired CLP, we decrease the value of B_t from 705 to 174.1.

For $B_t = 174.1$, we get,

```
results

service rate = 121.5
buffer size = 174
```

```
a = 0.78
b = 6.58
est_clp = 0.00006284
est_cwtd = 0.00023426
```

and the desired CLP is achieved.

As for the case of achieving the desired CWTD, we decrease the value of B_t from 705 to 149.8.

For $B_t = 149.8$, we get,

```
results

service rate = 121.5
buffer size = 150

a = 0.78
b = 6.78
est_cwtd = 0.00069483
est_clp = 0.00017889
```

and the desired CWTD is achieved.

These values of B_t' are even smaller than the individual β's. This suggests that assigning C_t equal to the sum of the individual α's is enough to guarantee an overall performance that meets the individual QoS for a comparatively small buffer size.

Hence, CAC is performed as indicated by the algorithm of Section 4.3.2.

4.3.5 Comments

We have seen that a source may be characterized by an effective bandwidth, a measure that assigns a minimum bandwidth to meet a required QoS. Moreover, we introduced a new measure, the effective buffer size, a measure that assigns a minimum buffer space so as to satisfy a desired QoS.

Resource management would not be optimal unless statistical multiplexing gains are exploited. Indeed, in the case of merging input traffic streams together on a single link, we observed that for a total link capacity equal to the sum of the individual effective bandwidths of the sources, the required buffer size that would meet the desired QoS is much smaller than the sum of the individual effective buffer sizes.

In this line of thought, an application of the effective bandwidth theory, namely CAC, is applicable and is done by the algorithm described in Section 4.3.2.

4.4 Discussion of Results

In this chapter, we extended the work of Chapter 3 to cover the case of multi-input traffic. The virtual buffer method, quite efficient in traffic monitoring, is not very efficient in CAC. This is mainly due to the failure of this scheme to account for the asymptotic constant.

The effective bandwidth approximation is,

$$P(X > B) \approx ae^{-bB} \approx e^{-bB} \tag{4.14}$$

Actually, a is shown [9] to be asymptotically exponential itself, i.e., for n sources,

$$a_n \sim \sigma e^{-n\gamma} \tag{4.15}$$

where $\sigma > 0$, γ is positive for sources more bursty than Poisson and is negative for sources less bursty than Poisson. Hence, as the number of sources increases,

this approximation may perform very badly, as it leads to overestimation or underestimation of the link capacity, rendering the inclusion of the asymptotic constant inevitable. Nevertheless, this algorithm is undeniably simple and yields fairly good results in most of the practical settings.

As of the second scheme, an optimal pair, namely the effective bandwidth and the effective buffer size, could be derived, for QoS requirements expressed in terms of both CLP and CWTD. Multiplexing several input streams together on a single link provided valuable gains, as expected. Finally, based on this result, we could perform CAC, as given by the algorithm of Section 4.3.2.

Chapter 5

Conclusion

In this work, we investigated thoroughly the effective bandwidth concept, a measure of resource usage that adequately represents the trade-offs between sources of different types, taking proper account of their varying statistical characteristics and quality of service (QoS) requirements.

As ATM networks would support multi-media bursty traffic, often multiplexing a large number of input streams, an exact analysis and calculation of the cell loss probabilities (CLP) is intractable. In the presence of very small CLP values of the order of 10^{-9} and reasonably large buffer sizes, as in an ATM context, a large deviation approach is adopted and the effective bandwidth acts as an asymptotic approximation to the tail probabilities of loss.

Once bursty sources, transmitting traffic that can vary from image, data, voice, video or any combination of these, are assigned an effective bandwidth, network design and management tasks reduce to the simple cases of circuit-switched networks.

For the effective bandwidth and its underlying approximation, to be a viable tool, on-line, real-time estimation schemes are essential both in source characterization and in resource management applications.

However, from the theory and simulations of the effective bandwidth, it is

now clear that the concept has some underlying limitations. These are mainly due to some restrictions in the assumptions which fail to meet the real-life situations, to the uncertain characteristics of some sources and to the exclusion of the asymptotic constant in the effective bandwidth approximation.

The effective bandwidth approximation is validated through the large deviation principle and the underlying results are obtained in the asymptotic regime. The asymptotic regime is basically the assumption of infinite buffer size, i.e., $B \rightarrow \infty$ and very small CLP, i.e., $p \rightarrow 0$. However, the asymptotic regime significantly overestimates (case of sources that are more bursty than Poisson) or underestimates (case of sources less bursty than Poisson) the number of sources that can be multiplexed on a link.

Most of the work done on the effective bandwidth concept, assumed the sources to be stationary and ergodic. The stationarity assumption is known to fail in the case of long burst periods [23]. What would happen if sources are non-stationary, or even non-ergodic?

As for the quasi-static approximation, it is not solved yet. The approximation assumes that the system reaches steady-state. But what if a connection duration is not sufficiently long? The approximation would not hold. The question here is about the distribution of the transient process associated with a buffer servicing a number of connections whose effective bandwidth never exceeds the buffer capacity. But, will it in fact satisfy the expected QoS-constraint at each point in time. The proofs for the effective bandwidth are usually based on the transient arguments, that is, the system starts from an empty state which is not always true.

The effective bandwidth of a source depends on the statistical characteristics of the source. The source, however, may have difficulty providing such information. The uncertain characteristics of sources raise challenging practical and theoretical issues. Attempts to measure the effective bandwidth of a connection based on the very definition of the effective bandwidth may use an empirical averaging to replace the expectation operator. However, this may turn out to be

inaccurate. Suppose a user engages a connection on the basis of a peak rate demand, but after, a small quantity of traffic is produced. How would the network control schemes apply?

One of the most appealing features of ATM, is that it yields the way to statistical multiplexing. The effective bandwidth theory became very popular due to its simplicity in allocating the bandwidth independently of the number of sources being multiplexed. The tail distribution, $P(X > B)$, has been shown to be asymptotically exponential, i.e,

$$P(X > B) \approx ae^{-bB} \approx e^{-bB}. \tag{5.1}$$

However, the asymptotic constant a may differ from 1 as the number of superimposed sources increases. Actually, a is itself asymptotically exponential in the number of multiplexed sources, i.e., for n sources,

$$a_n \sim \sigma e^{-n\gamma}, \tag{5.2}$$

where $\sigma > 0$ and γ is positive for sources more bursty than Poisson and is negative for sources less bursty than Poisson. It has been suggested [10] that a refined asymptotic approximation, namely

$$P(X > B) \approx \sigma e^{-n\gamma} e^{-bB} \quad , \tag{5.3}$$

may be used.

Nevertheless, for very small CLP values, the effective bandwidth approximation holds well for values of a larger than 10^{-4} and smaller than 10^{4}, which is the case in a wide range of realistic problems.

With these limitations in mind, work is still developing in this area.

Appendix A

Large Deviation Principle

The large deviation theory is the theory of rare events, i.e., events which take place away from the mean, out in the tail distributions. In this sense, it studies the tails of the distributions.

For a tail probability of loss given by $P(X > B)$ to be held below a significantly small value p as B tends to ∞, the large deviation theory states that the tail probability has an exponential decaying function with rate I, i.e., $P(X > B) \approx e^{-IB}$.

Formally, the large deviation principle is stated as follows.

We use the following suggestive notation $P(M_n \approx) \asymp e^{-nI(x)}$ to indicate that the sequence $P(M_n > x)$ of probability distributions satisfies a large deviation principle with rate-function I, that is,

i. the function I is lower-semicontinuous

ii. for each real number a, the level set $x \in \mathbf{R} : I(x) \leq a$

iii. for each closed subset F of $\mathbf{R}$, $\lim_{n\to\infty} \sup \frac{1}{n} ln P(M_n \in F) \leq -\inf_{x\in F} I(x)$

iv. for each open subset G of $\mathbf{R}$, $\lim_{n\to\infty} \inf \frac{1}{n} ln P(M_n \in F) \geq -\inf_{x\in F} I(x)$

Appendix B

Gartner-Ellis Theorem

The Gartner-Ellis theorem provides the technical conditions under which the large deviation theory applies for a source.

We state the Gartner-Ellis theorem for a sequence M_n of real-valued random variables; the theorem can be extended to vector-valued random without too much difficulty. Define $\lambda_n(\theta) = \frac{1}{n} ln E e^{n\theta M_n}$ for $\theta \in \mathbf{R}$ and assume :

i. $\lambda_n(\theta)$ is finite for all θ.

ii. $\lambda(\theta) = \lim_{n\to\infty} \lambda_n(\theta)$ exists and is finite for all θ.

Then the upper bound iii. in the large deviation principle holds with rate-function $I(x) = \sup_{\theta\in\mathbf{R}} [x\theta - \lambda(\theta)]$.

If in addition $\lambda(\theta)$ is differentiable for all $\theta \in \mathbf{R}$, then the lower bound iv. in the large deviation principle holds.

References

[1] J. Abate, G. L. Choudhury, and D. M. Lucantoni "Asymptotic Analysis of Tail Probabilities Based on the Computation of Moments," *Annals of Applied Probability*, vol. 5, pp. 983–1007, 1995.

[2] D. Anick, D. Mitra, and M. M. Sondhi "Stochastic Theory of a Data-Handling System with Multiple Sources," *The Bell System Technical Journal*, vol. 61, pp. 1872–1894, 1982.

[3] D. D. Botvich and N. G. Duffield. "Large Deviations, the Shape of the Loss Curve, and Economies of Scale in Large Multiplexers,". Technical Report DIAS-APG-94-12, ., 1994. Submitted to Queueing Systems.

[4] C. S. Chang and J. A. Thomas "Effective Bandwidth in High Speed Digital Networks," *IEEE JSAC*, vol. 13, pp. 1091–1100, 1994.

[5] C. S. Chang. "Approximation of ATM Networks : Effective Bandwidths and Traffic Descriptors,". Technical Report RC 18954, IBM, May 1993.

[6] C. S. Chang "Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Trans. on Automatic Control*, vol. 39, pp. 913–931, 1994.

[7] C. S. Chang and L. S. Lou "Experiments of Effective Bandwidth for Markov Sources and Video Traces," *IEEE JSAC*, pp. 497–504, 1996.

[8] C. S. Chang, P. Heidelberger, and P. Shahabuddin "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Performance Evaluation*, vol. 20, pp. 45–65, 1994.

[9] G. L. Choudhury, D. M. Lucantoni, and W. Whitt "Squeezing the most out of ATM," *IEEE Transactions on Communications*, pp. 1872–1894, April 1994.

[10] G. L. Choudhury, D. M. Lucantoni, and W. Whitt "On the effectiveness of effective bandwidths for admission control in ATM networks," in J. Labetouille and J.W. Roberts, eds., *ITC*, volume 14, pp. 411–420. Elsevier Sciences B.V., 1994.

[11] G. L. Choudhury, D. M. Lucantoni, and W. Whitt "The BMAP/G/1 Queue: A Tutorial," in L. Donatiello and R. Nelson, eds., *Models and Techniques for Performance Evaluation of Computer and Communications Systems*, pp. 330–58. Springer Verlag, 1993.

[12] G. L. Choudhury, D. M. Lucantoni, and W. Whitt "Asymptotic Analysis of Tail Probabilities Based on the Computation of Moments," *Annals of Applied Probability*, vol. 5, pp. 719–740, 1993.

[13] C. Courcoubetis and J. Walrand. "Note on the Effective Bandwidth of ATM Traffic at a Buffer,". unpublished manuscript, 1991.

[14] C. Courcoubetis, G. Fouskas, and R. R. Weber. "An On-line Estimation Procedure for Cell Loss Probabilities in ATM Links,". to appear, 1995.

[15] C. Courcoubetis and R. R. Weber. "Effective Bandwidths for Stationary Sources,". to appear, 1995.

[16] C. Courcoubetis, G. Fouskas, and R. R. Weber "On the Performance of an Effective Bandwidths Formula," in Jacques Labetoulle and James W. Roberts, eds., *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress — ITC 14*, volume 1a of *Teletraffic Science and Engineering*, pp. 201–212. Elsevier Science B.V., June 1994. Antibes Juan-les-Pins.

[17] G. de Veciana, G. Kesidis, and J. Walrand "Resource Management in Wide-Area ATM Networks using Effective Bandwidths," *IEEE JSAC*, vol. 14, pp. 1080–1090, August 1995.

[18] G. de Veciana and J. Walrand. "Effective Bandwidths: Call Admission, Traffic Policing and Filtering for ATM networks,". to appear in *Queueing Systems*, 1994.

[19] K. M. Elsayed and H. G. Perros "On the Effective Bandwdith Theory of Arbitrary on/off Sources,". 1996.

[20] A. I. Elwalid and D. Mitra "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, 1993.

[21] R. Gibbens and P. Hunt "Effective Bandwidths for the Multi-type UAS channel," *Queueing Systems*, vol. 1, pp. 17–28, 1991.

[22] R. J. Gibbens. "Traffic Characterisation and Effective Bandwidths for Broadband Network Traces,". Research report. Submitted for publication. 1996-9, Statistical Laboratory, University of Cambridge, 1996.

[23] R. Guérin, H. Ahmadi, and M. Naghshineh "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE JSAC*, vol. 9, pp. 968–981, September 1991.

[24] L. Gun, V. G. Kulkarni, and P. F. Chimento "Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer," *IEEE JSAC*, vol. 13, pp. 1039–1047, August 1995.

[25] J.Y. Hui "Resource Allocation for Broadband Networks," *IEEE JSAC*, vol. SAC-6, pp. 1559–1608, 1988.

[26] J.Y. Hui. *Switching and Traffic Theory for Integrated Broadband Networks.* Kluwer Academic Publisher, Boston, 1990.

[27] F.P. Kelly "Effective Bandwidths at Multi-Type Queues," *Queuing Systems*, vol. 9, pp. 5–15, 1991.

[28] F.P. Kelly. *Stochastic Networks: Theory and Applications*, chapter Notes on the Effective Bandwidth. 1995.

[29] G. Kesidis, J. Walrand, and C. S. Chang "Effective Bandwidths for Multiclass Markov Fluids and other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 424–428, August 1993.

[30] G. Kesidis. *ATM Network Performance.* Kluwer Academic Publisher, Boston, 1996.

[31] J. T. Lewis and R. Russel. "An Introduction to Large Deviations,". Technical report, ., 1996. unpublished.

[32] B. L. Mark and G. Ramamurthy "Real-Time Estimation of UPC Parameters for Arbitrary Traffic Sources in ATM Networks,". San Francisco, March 1996. IEEE Infocom'96.

[33] A. K. Parekh and R. G. Gallager "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks : the Sngle Nde Case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–57, June 1993.

[34] N. Schroff and M. Schwartz "Improved Loss Calculations at an ATM Multiplexer,". San Fransisco CA, March 1996. IEEE INFOCOM 96.