An overview of effective bandwidth methods

Thijs Harleman
thijs.harleman@nokia.com

Presentation date: Oct 18, 1999

Abstract

This report gives an overview of the effective bandwidth approach, which is one of the proposed methods to perform CAC in B-ISDN networks. The essential idea behind this approach is to estimate the required bandwidth (i.e. service rate) for a set of traffic sources from a queue model, given the source characteristics, the buffer size B and QoS constraints. If we can find an approximately linear bound on the maximum number of sources of each source type, the capacity allocated to each source of each type is understood as the effective bandwidth. We first treat some common concepts in the effective bandwidth approach. Then we give an overview of the major streams for estimating the effective bandwidth found in literature, in which we mainly focus on the results and the differences between the various methods.

NOKIA

CONTENTS

1. INTRODUCTION 2
2. BASIC CONCEPTS FOR EFFECTIVE BANDWIDTH 4
2.1 The notion of effective bandwidth 4
2.2 General approach for estimating the effective bandwidth 4
2.3 Key properties of effective bandwidth methods 4
2.3.1 Arrival distribution 5
2.3.2 Service time distribution 5
2.3.3 Homogeneous versus heterogeneous traffic sources 5
2.3.4 Number of traffic sources 5
2.3.5 Source rate distribution 5
2.3.6 Queue size 6
2.3.7 Buffering discipline 6
2.3.8 Detail level of the model 6
3. OVERVIEW OF SOME EFFECTIVE BANDWIDTH METHODS 8
3.1 Fluid flow approximations 8
3.2 Markovian queue models 10
3.3 Large deviations approximations 10
4. GLOSSARY 13
5. REFERENCES 14

1. INTRODUCTION

Generally, B-ISDN networks are expected to integrate a large number of traffic streams with a wide variety of traffic characteristics, while still providing some guaranteed Quality of Service (such as allowed cell or packet loss rate and allowable queueing delay). The traffic sources generating these streams can transmit data at different rates, where the rate may vary between 0 and some peak rate. Hence, we can achieve some *statistical multiplexing gain* by allowing the cumulative peak rate of a set of different traffic streams to exceed the available link capacity, thereby increasing the utilization of the networks resources. This is illustrated in Figure 1.

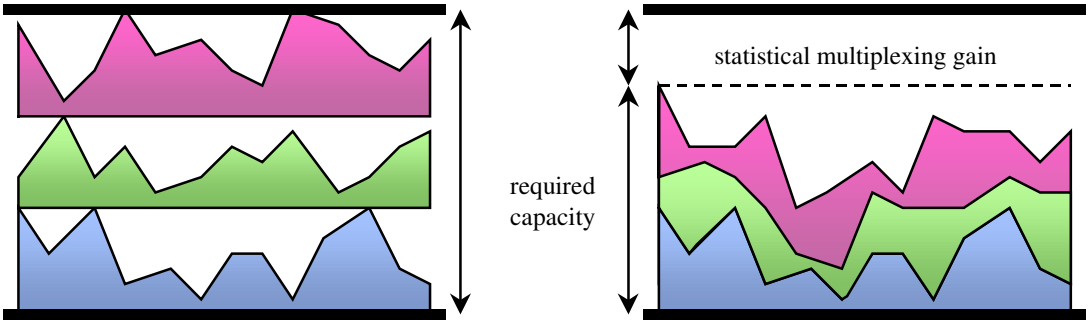

Figure 1 Statistical multiplexing of traffic streams

This report gives an overview of the *effective bandwidth* approach, which is one of the proposed methods to perform CAC in B-ISDN networks. The essential idea behind this approach is to estimate the required bandwidth (i.e. service rate) for a set of traffic sources from a queue model, given the source characteristics, the buffer size B and QoS constraints (e.g. the allowable packet loss probability ε, which typically is chosen in the order of 10^{-5} - 10^{-9}). This estimate for the effective bandwidth will vary between

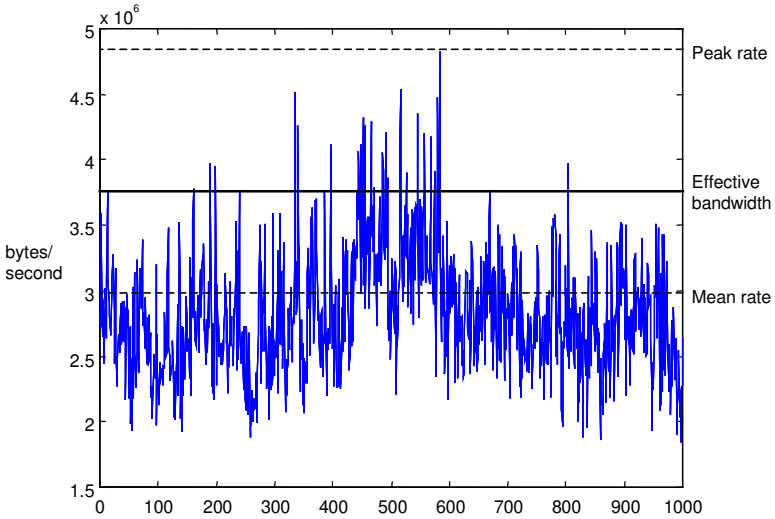

Figure 2 Example of rate bounds for a real traffic sample

the mean and peak rate of the cumulative traffic; the burstier the traffic is, the closer the effective bandwidth will be to the peak rate. Figure 2 illustrates these bounds for a real measured traffic sample (taken from [6]).

Chapter 2 describes some common concepts used in all effective bandwidth methods. Then, Chapter 3 summarizes some well-known methods and discusses the differences in model assumptions, applicability and results for these methods.

2. BASIC CONCEPTS FOR EFFECTIVE BANDWIDTH

Before we take a look at some effective bandwidth methods from literature, we first treat some general concepts used in these methods. This chapter describes the notion of effective bandwidth, explains the high-level general approach for estimating and then discusses what the relevant properties are for comparing the different effective bandwidth methods.

2.1 The notion of effective bandwidth

For predicting the performance of multiplexers loaded with heterogeneous traffic streams, we require a wide range of traffic characteristics. However, to simply the CAC in B-ISDN networks, it is more convenient to use the notion of effective bandwidth, as suggested by many authors in literature. To decide if a new connection can be accepted, we then simply compare the sum of effective bandwidths for the traffic streams with the available link capacity.

In COST 242 [6], two essential characteristics of effective bandwidth are given:

- the effective bandwidth of a traffic stream is independent of the other streams with which it is mixed;
- the sum of the effective bandwidths of two independent traffic streams is to be equal to the effective bandwidth of their superposition; this is the additivity property.

It must also be mentioned that, in order to be practically feasible, the effective bandwidth calculations should be possible in real-time, so the approximation method cannot be too complex.

2.2 General approach for estimating the effective bandwidth

The effective bandwidth of a traffic stream is chosen as the smallest capacity (server rate) C that solves the admission criterion. In most cases, the admission criterion is taken as the allowed buffer overflow probability ε. . For infinite buffers, the buffer overflow probability can be found from the complementary distribution function $G(x)$ of the buffer occupancy, i.e. the probability that the queue length X exceeds the buffer size B. The admission criterion then becomes:

$$G(B) = \Pr\{X > B\} < \varepsilon \qquad (1)$$

in which $G(B)$ is a function of the variable C. We could also choose the mean waiting time as the admission criterion (e.g. see De Veciana and Walrand [10]), but this approach is less common.

2.3 Key properties of effective bandwidth methods

The accuracy and applicability of the various effective bandwidth methods described in the literature greatly depends on the assumptions made in modelling the queue model (with takes into account both source and queue characteristics). As usually simplifying assumptions need to be made to keep the analysis tractable, these assumptions may lead to smaller or larger inaccuracies of the results.

In this section we focus on the key properties by which we can categorize the different methods in literature.

In the general queueing case traffic generated from a number of sources is multiplexed on a single link. Traffic that cannot immediately be served is enqueued in a buffer. The following subsections briefly describe the different variables of interest used in literature when building a queueing model.

2.3.1 Arrival distribution

The arrival distribution describes the rate at which source events occur. Depending on the model under study, this arrival rate can have a different interpretation. For example, in the case of Markov-modulated sources an arrival can indicate a change in the source rate, whereas in the case of a M/G/1 queue it can indicate the arrival of a packet. Most methods are based on Markovian or Markov-modulated sources, because they reduce the complexity of the analysis. However, it must be noticed that the memoryless assumption (which is typical for all Markov models) does not necessarily agree with realistic traffic (which can exhibit self-similar properties).

2.3.2 Service time distribution

The service time distribution describes the amount of time a packet remains in the server. For many queueing models this is the same as the burst length distribution. The M/M/1 queue is an example of a queue with Markovian service times (i.e. exponentially distributed).

2.3.3 Homogeneous versus heterogeneous traffic sources

As earlier described in Chapter 1, broadband networks need to accommodate a wide variety of traffic types. To take this feature into account in queueing modelling, it is preferable if the model can deal with sources with different stochastic behavior (heterogeneous sources). Many authors use a model with homogeneous traffic sources as a starting point of their analysis, which they then generalize to case of heterogeneous traffic. It must be noticed that in most models with heterogeneous traffic the sources have the same underlying model (e.g. Poisson) but with different parameters.

2.3.4 Number of traffic sources

Mostly, a single source in isolation is only studied in literature as a starting point (because the analysis becomes easier in that case). Then, the results are generalized to the case of multiple sources (which can than be either a homogeneous of heterogeneous mix).

2.3.5 Source rate distribution

Some more elementary methods for calculating the effective bandwidth are based on ON/OFF models, in which each source either is silent or transmits at its peak rate (typical for data traffic). Obviously, the more general methods in which the source transmission rate is chosen from a finite set of values have wider applications (e.g. they can also be used for modelling VBR traffic).

2.3.6 Queue size

Generally, three different cases for the queue size can be considered. The most elementary case is the *bufferless* case (buffer size 0), which means that traffic is not enqueued at all but simple blocked. Thus, in this case the focus is on the packet blocking probability instead of buffer overflow probability. The next case is the case of the *infinite buffer*, which by definition is an approximation of any real case. The most general case is the case of a *finite buffer*.

When the infinite buffer case is used as an approximation of a real situation (finite buffer), the overflow probability is necessarily overestimated. This is because the buffer occupancy distribution for a finite queue will demonstrate a (slight) shift in probability mass from the tail to the head of the queue with respect to the case of an infinite buffer. Thus, the buffer overflow probability for an infinite buffer acts as an upperbound for the overflow probability of finite buffers. Obviously, the approximation error of the infinite buffer approximation becomes larger for smaller buffers.

2.3.7 Buffering discipline

Although in queueing analysis we can encounter various buffering disciplines, for practical purposes mostly the FIFO (also referred to as FCFS) discipline is considered. It should be mentioned that when the network should support multiple QoS classes, for instance in ATM networks, priority buffering is also a relevant buffering discipline.

2.3.8 Detail level of the model

In general, the behavior of a traffic source can be described at different levels ([4], [7]) (e.g. in the case of ATM traffic: on the connection, burst and cell level, which is illustrated in Figure 3). Depending on the time scale we are interested in, we can choose a model that focuses on either the cell level (or packet if applicable) or burst level traffic. For instance, when we focus on the burst level behavior of traffic, we might even forget about the individual cells or packets and focus our analysis entirely on the traffic intensity. The connection level is of less interest in recent traffic modelling.

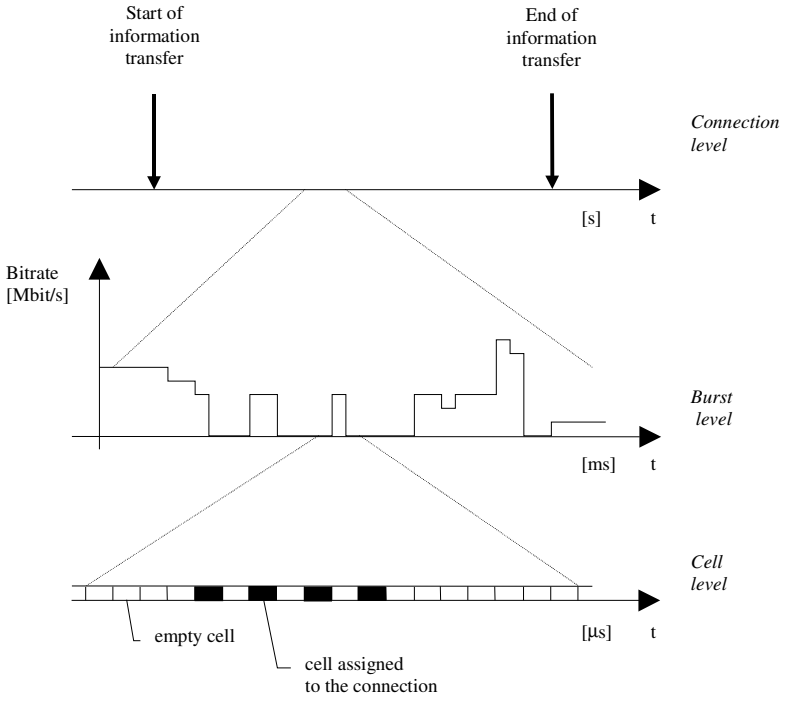

Figure 3 Three-level view of ATM traffic

3. OVERVIEW OF SOME EFFECTIVE BANDWIDTH METHODS

This chapter introduces some different methods to estimate the effective bandwidth for a connection or set of connections. Instead of delving into details, we will try to list the key applications, assumptions, approximations and results for each of the described methods. This will enable the reader to select the method(s) of his interest, for which he can find a more detailed treatment in the references given.

3.1 Fluid flow approximations

In general, the resources (links, buffers) in high-speed networks carry discrete units of data (i.e. packets or cells), so the traffic and buffer models should also be discrete. However, when we are dealing with large buffer sizes the model granularity becomes so small that we may approximate the discrete traffic flow by a continuous one (referring again to Figure 3, it means that we forget entirely about the cell level and only analyze the burst level). The analysis then becomes similar to a reservoir in which a fluid flows in at a variable rate and leaks out a constant rate C whenever the reservoir is non-empty, hence the name fluid flow approximation.

Gibbens and Hunt [2] consider a buffered channel loaded with traffic from a heterogeneous set of ON/OFF sources with exponentially distributed ON and OFF times. As both the traffic rates and the queue length are chosen from a continuous domain, their model is equivalent a to a Markov-modulated fluid flow approximation.

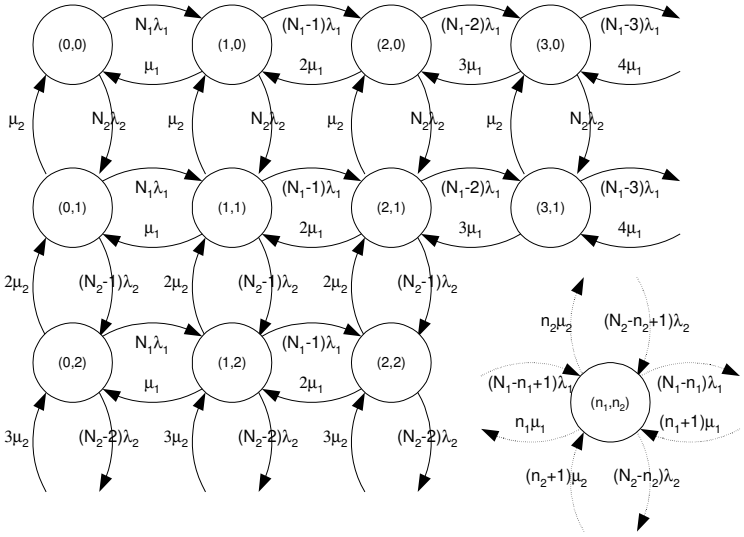

Figure 4 State transition diagram of Markov modulator for two source types

The system state for this model is governed by two equations: a matrix difference equation that describes the evolution of the joint source state (which is a Markov process) and a differential equation that describes the evolution of the buffer occupancy in time. Figure 4 shows the state diagram of a simple modulating Markov chain for two source types (The state (n_1, n_2) indicates the number of active sources of each type, N_i is the total number of sources for each source type and $1/\lambda_i$ and $1/\mu_i$ are the average OFF and ON times, respectively). The formal solution for the steady-state distribution

of this model can be found in a quite straightforward as it is a well known eigenvalue problem. The solution for the steady-state probability vectors then look like:

$$\boldsymbol{\pi}(x) = \sum_i a_i \eta_i e^{\eta_i x} \mathbf{e_i} \tag{2}$$

$$\boldsymbol{\psi} = \sum_i a_i \mathbf{e_i} \tag{3}$$

in which $\boldsymbol{\pi}(x) = (\pi_1(\mathbf{n}, x), \pi_2(\mathbf{n}, x), \ldots)$ is the joint pmf-pdf vector for the system state $\mathbf{n}$ and the queue length (for $x > 0$), $\boldsymbol{\psi}$ is the probability mass vector for $x = 0$ and a_i and $\mathbf{e_i}$ are the eigenvalues and eigenvectors of the eigenvalue problem. We omit here the details of the eigenvalue problem (such as the matrixes and boundary conditions), but we only focus on the structure of the results given in (2) and (3). We must mention that this solution depends on the rate c at which the queue is emptied.

Interestingly, the steady-state distribution of the queue length (which can be found from (2) by multiplying the state probability vector $\boldsymbol{\pi}(x)$ with the unit vector) is a sum of exponential functions. Of course, in order to arrive at a stable system, all eigenvalues must have negative real parts. Now, Elawid and Mitra analyze the same model in [1] and they state that for all practical situations there exists a *dominant eigenvalue*, which is real and is larger than the real part of the other eigenvalues. We can then see from (2) that the other exponentials decay faster than the exponential for this dominant eigenvalue. Thus, for large buffers we can approximate the tail distribution of the queue as being purely exponential:

$$\Pr\{X > x\} \approx e^{-|\eta_0(c)|x} \tag{4}$$

in which $\eta_0(c)$ is the dominant eigenvalue, which here is depicted as a function of the capacity c. Using the overflow constraint (1) we discussed earlier, the effective bandwidth $\hat{c}$ for the aggregate traffic stream is the solution for c in

$$\eta_0(c) = -\log \varepsilon / B \tag{5}$$

Gibbens and Hunt show that in their model the constraint (1) can be approximated for a large buffer by the constraint:

$$\sum_i \alpha_i(\zeta) N_i \le C \tag{6}$$

This is exactly the result we wanted to arrive at: at linear constraint for different source types. The $\alpha_i(\zeta)$, found as:

$$\alpha_i(\zeta) = \frac{\zeta\gamma_i + \mu_i + \lambda_i - \sqrt{(\zeta\gamma_i + \mu_i - \lambda_i)^2 + 4\lambda_i\mu_i}}{2\zeta} \tag{7}$$

can thus be used as an estimate for the effective bandwidth ($\zeta = \log \varepsilon / B$, γ_i gamma is the peak rate of each source type). It must be noticed that, although the solution

given by (2) and (3) looks complete, it is not practical feasible to solve it for dimensions (i.e. number of source types) larger than 1.

Guérin et all. [3] use a fluid flow approximation which is a simple case of the Markov-modulated fluid process. They find for the effective bandwidth of single ON/OFF sources in a finite FIFO buffer of length x the following effective bandwidth formula:

$$\hat{c} = \frac{R_{\text{peak}}}{2} - \frac{\lambda+\mu}{2\theta} + \sqrt{\left[\frac{R_{\text{peak}}}{2} - \frac{\lambda+\mu}{2\theta}\right]^2 + \frac{\lambda R_{\text{peak}}}{\theta}} \tag{8}$$

where $\theta = \frac{1}{x}\ln\left(\frac{1}{\varepsilon}\right)$, R_{peak} is the peak rate of the source in the ON state and λ and μ are the arrival and service rates, respectively. Following a similar matrix approach for the case of multiple ON/OFF sources of the same type, they find better estimates for the effective bandwidth of each source.

3.2 Markovian queue models

The steady-state probability distributions for M/M/1 and M/G/1 queues are described in many books on queueing theory, so we shall not delve into details here.

The approach taken in [5] and [10] is that for a set of heterogeneous sources with exponentially distributed interarrival times and generally distributed service times (burst lengths), an estimate for the effective bandwidth is found using both a time constraint (the average delay seen by packets) and a loss constraint (as given in (1)). Then it is shown that these constraints can be approximated by a linear constraint similar to the one given in (6). De Veciana and Walrand also discuss a variant with two service priorities. For more details the reader is referred to the articles mentioned.

3.3 Large deviations approximations

Instead of finding an explicit expression for the buffer occupancy distribution, Hui [4] and Kelly [5] find some effective bandwidth estimations from large deviations theory. For the bufferless case, they show that an upperbound on the link overload probability can be found based on the moment generating functions of the sources' rate distributions. The moment generating function is well known in probability theory and it is equivalent to the mirrored Laplace transform of a random variable's probability density function:

$$M(s) = \mathrm{E}\left[e^{sX}\right] = \int_{-\infty}^{\infty} f(x)e^{sx}dx = \hat{F}(-s) \tag{9}$$

where $M(s)$ is the moment generation function, $f(x)$ the pdf and $\hat{F}(s)$ the Laplace transformed pdf.

As large deviations theory is too abstract to discuss extensively here, we limit ourselves to only giving the key equations and results. The upperbound on the link overflow probability is given as

$$\Pr\{S \geq C\} \leq \varepsilon = 10^{-\gamma} \tag{10}$$

where $S = \sum_{j=1}^{N} \sum_{i=1}^{n_j} X_{ji}$ is the sum of the random source rates X_{ji} of n_j sources out of N different source type and C is the capacity of the link. From this upperbound and the Chernoff bound:

$$\frac{1}{n} \log \Pr\{X_1 + X_2 + \ldots + X_n \geq 0\} \leq \inf_s \log M(s) \tag{11}$$

it can be shown that (10) holds for all combination of sources in the source vector $\underline{n} = (n_1, n_2, \ldots, n_N)$ satisfying the condition:

$$\sum_{j=1}^{N} \alpha_j^* n_j + \frac{\gamma}{s^*} \leq C \tag{12}$$

where $\alpha_j^* = \frac{\log M_j(s^*)}{s^*}$ and s^* is a value for which s takes the infinimum in $\inf_s \sum_{j=1}^{N} n_j M_j(s) - sC$

The most interesting part of equation (12) is its interpretation: when a bufferless link is loaded with a heterogeneous set of N source types (for each of which the rate distribution can be freely chosen), the upperbound on the link overload probability as given in equation (10) results in a *linearly constrained admission region* in N-dimensional space. The factors α_j in (12) can thus be used as the effective bandwidth of each connection of type *j*.

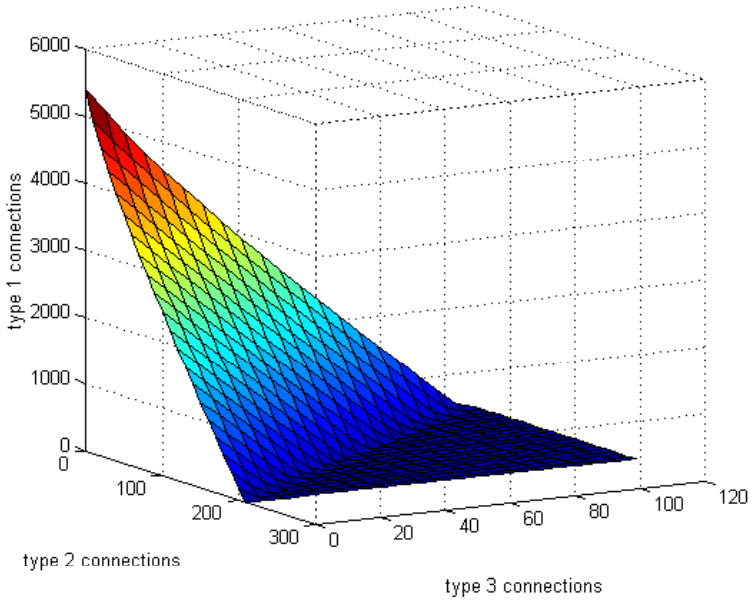

Figure 5. Example of an admission region for ε=10^{-9}

To illustrate this linearity of the admission region boundary, Figure 5 shows an example of this boundary in 3-dimensional space (taken from [8]). The traffic consists of heterogeneous mixes of three different types of ON/OFF sources. The parameters used for calculating the admission region are shown in Table 1. The actual admission region is the volume between the origin, the x-, y-, and z-axes and the triangular admission region boundary.

Table 1. A realistic traffic mix

link rate c=150 Mbps		
service	peak rate h	burst probability p
1	64 kbps	0.4
2	2 Mbps	0.2
3	2 Mbps	0.4

4. GLOSSARY

ATM	Asynchronous Transfer Mode
B-ISDN	Broadband Integrated Services Digital Network
CAC	Connection Admission Control
FCFS	First Come First Served
FIFO	First In First Out
pdf	probability density function
pmf	probability mass function
QoS	Quality of Service

5. REFERENCES

[1] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks", *IEEE/ACM TNET*, Vol. 1, No. 3, pp. 329-343, June 1993

[2] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel", *Queueing Systems*, Vol. 9, pp. 17-28, 1991

[3] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", *IEEE JSAC*, Vol. 9, No. 7, pp. 968-981 , September 1991

[4] J.Y. Hui, "Resource allocation for broadband networks", *IEEE JSAC*, Vol 6, No. 9, pp. 1598-1608, December 1988

[5] F.P. Kelly, "Effective bandwidths at multi-class queues", *Queueing Systems*, Vol. 9, pp. 5-16, 1991

[6] A. Petäjistö, *An Empirical Time-Series Analysis of Internet Protocol Datatraffic. Special study for the Systems Analysis Laboratory.* Espoo, Finland: Helsinki University of Technology, Systems Analysis Laboratory, 1997

[7] F.C. Schoute, "Simple decision rules for acceptance of mixed traffic streams", *Proceedings of 12th International Teletraffic Congress, Italy*, paper 4.2 A.5, 1998

[8] J.W. Roberts (Ed.), *Performance Evaluation and Design of Multiservice Networks. Final Report COST 224,* Luxembourg: Commission of European Communities, Information Technologies and Sciences, 1992

[9] J. Roberts, U. Mocci, and J. Virtamo (Eds.), *Performance evaluation and design of broadband multiservice networks, Final report of action COST242*, Berlin: Springer Verlag, 1996

[10] G. de Veciana and J. Walrand, "Effective bandwidths: Call admission, traffic policing and filtering for ATM networks", *Queueing Systems*, Vol. 20, pp. 37-59, 1995