16

MEASUREMENT-BASED CAC FOR VIDEO APPLICATIONS USING SBR SERVICE

F. Brichet and A. Simonian.
France Telecom - CNET/DAC/GTR
Address: 38-40 rue du Général Leclerc, F-92794 Issy Moulineaux Cedex 9, France
Tel: (+) 33 1 45 29 60 11, Fax: (+) 33 1 45 29 65 56
e-mail: {francois.brichet, alain.simonian}@cnet.francetelecom.fr

Abstract

This paper provides a conservative estimation of the required bandwidth for video sources multiplexed according to the bufferless scheme and using the Statistical Bit Rate transfer capability in ATM networks. For sources with sustainable bit rate significantly greater than the mean bit rate (as expected for video teleconference streams), we consider an estimation of the required bandwidth simply based on the measure of the link load. On the other hand, for video sources encoded using an algorithm such as that presented in [HRR97], we propose that this estimation be based on the measured variance of the total bit rate of already accepted sources. In both cases, the required bandwidth is calculated independently of any specific source model. This leads to a simple CAC policy which significantly increases the bandwidth saving with respect to the classical worst case framework.

Keywords

Connection Admission Control, real-time video applications, ressource allocation, equivalent bandwidth, measurement in ATM network, multiplexing gain.

1. INTRODUCTION

ATM Transfer Capabilities (ATC) have been defined by the ITU [ITU, Recommendation I.371] and the ATM Forum [ATMF96] in order to provide Quality of Service (QoS) guarantees to the different applications that the ATM based B-ISDN (Broadband Integrated Services Digital Network) has to handle. For each ATC, traffic management tools and efficient resource allocations have to be developed and implemented. The mapping between the different applications and the standardized ATCs is still an open issue. For video connections requiring high QoS, particularly in terms of network delay (e.g. for video-conferencing, video-teaching or live event video applications), it has been suggested that SBR

Performance and Management of Complex Communication Networks
T. Hasegawa, H. Takagi & Y. Takahashi (Eds.)

(Statistical Bit Rate) or VBR-rt (Variable Bit Rate-real time) is the most appropriate transfer capability [Rob95]. For such video connections, one of the main problems is the choice of the three parameters associated with the SBR capability, that is, the Peak Cell Rate (PCR), the Sustainable Cell Rate (SCR) and the Intrinsic Burst Tolerance (IBT). These parameters are supposed to be controlled by a leaky bucket, i.e., the Generic Cell Rate Algorithm [ITU, Recommendation I.371]. There are two main options for handling video connections using SBR capability.

According to the first option, video connections are encoded with an open-loop coder and the parameters are chosen in such a way that the proportion of lost (or tagged) cells at the GRCA is negligible. Video quality is consequently unaffected. It is known, however, that the set of parameters (PCR, SCR, IBT) which makes the GCRA transparent for a given video application is hard to predict. Even for applications like video-conferencing, SCR can only be considered as a very loose upper bound of the mean rate. In [RB95], it is shown that even for quite a large value of IBT (corresponding to a maximum burst size of the order of thousands cells), the sustainable rate has to be at least twice the real average rate.

On the other hand, for entertainement or news video applications, the extreme traffic variability over various time-scale [GW94, HL96] can lead to unacceptable cell loss at the GCRA. Such applications suggest to choose an alternative option making use of a closed-loop coder. Indeed, the MPEG-based rate control algorithm presented in [HRR97] ensures that the emitted video traffic conforms precisely to the declared parameters of the SBR service. In this case, the video quality can vary but the encoding algorithm guarantees that the incoming stream is not affected by the GCRA at the ingress of the network. Moreover, the real mean bit rate of the video stream is made equal to the sustainable bit rate.

For video sources using the SBR transfer capability, it is possible to control congestion by operating with either burst scale or cell scale congestion. In the case of burst scale congestion, large buffers in multiplexers are assumed to guarantee a low cell loss ratio when cell bursts are transmitted at peak cell rate. This is the so-called *multiplexing with buffer* or *Rate Sharing (RS) multiplexing* [COST242]. In the case where cell scale congestion only is considered, the buffer has just to absorb congestion due to coinciding cell arrivals from different streams. The only constraint is then to keep the sum of the rates of active connections less than the multiplexer output rate. This is the *bufferless multiplexing* or the *Rate Envelope Multiplexing (REM)* scheme [COST242]. Interactive or real-time video applications (like live events) which cannot tolerate delay have to be multiplexed according to the REM scheme. It is known that this multiplexing scheme leads to efficient network utilization only if the peak rate is a small fraction of the link rate.

In this paper, we consider video sources using the SBR transfer capability and multiplexed using the REM scheme. We address the problem of the definition of a suitable Connection Admission Control (CAC) for such sources. The PCR

and SCR are denoted by h and r, respectively. A "worst case" CAC is then derived by assuming that a source emits cells according to an on/off pattern, with the "on state" associated with transmission at peak cell rate and the "off state" associated with any silent period [ITU, Recommendation E.73x]. The IBT value is therefore unused. The stationary mean and variance of the source bit rate are then r and $r(h-r)$, respectively. Given the peak cell rate and sustainable cell rate, these are the maximum values for the mean and the variance. The above on/off model is conservative but it obviously corresponds to resource overallocation. This overallocation can be due to one of the following two factors.

- The SCR parameter is a loose upper bound of the mean bit rate or,
- the real traffic is less variant than an on/off model.

We here discriminate the above two types of overalloaction. In the first one, we propose to take a measurement of the mean bit rate of the already accepted sources into account, in order to improve the resource allocation estimation. For example, in the case of the video-conferencing application, we expect that a measurement of the mean rate will be significantly less that the sum of the SCR. The measurement of the mean rate of each traffic source would be unpractical and current signalling standards do not allow the communication of such parameters inside the network. It is rather more reasonable to consider that measurements of the stationary mean of the *total* bit rate on each network link can be performed regularly on time. Note that a measurement of the *instantaneous* total bit rate has been proposed in [GKK95] for applying the conservative on/off model to homogenous sources in an adaptative manner. In this paper, we show that measuring the *stationary* mean of the total bit rate for *heterogeneous* sources brings significant multiplexing gain. On the other hand, for video movies or news encoded according to the algorithm presented in [HRR97], the mean bit rate is equal to SCR. However, the bit rate can be less variant than that of the "worst case" model. For this type of traffic, we then suggest to improve the bandwidth gain by measuring the stationary variance of the total bit rate.

The rest of the paper is organised as follows. In Section 2 (resp. Section 3), we estimate the impact of the mean bit rate (resp. bit rate variance) on the resource allocation. In both cases, a simple conservative evaluation of the effective bandwidth per source is presented which is independent of any specific source model. In Section 4, the use of an effective bandwidth for the mixture of heterogeneous streams is justified by considering the linearity of the acceptance region. In Section 5, we develop a simple method for the resource allocation problem based *only* on the measurement of the mean (resp. the variance) of the total bit rate on a link and apply this method to the definition of a CAC mechanism for video connections using the SBR transfer capability. Final Section 6 is devoted to conclusive remarks.

2. IMPACT OF THE MEAN BIT RATE

2.1. General definitions and the worst case approach

We first introduce a general definition related to the performance evaluation of a multiplexer operating under the REM scheme. Denote by Λ_N the random variable describing the total stationary bit rate due to the superposition of N identical sources with bit rate distributed as variable λ. Let

$$\varphi_{m,v}(s) = E(e^{s\lambda}), \quad s \geq 0,$$

define the Laplace transform of the distribution of λ. In the sequel, the mean and the variance of λ are denoted by m and v, respectively. On a transmission link with capacity C, the saturation probability $Pr(\Lambda_N \geq C)$ associated with REM multiplexing can be evaluated by means of the Chernoff bound [Kel91], namely

$$Pr(\Lambda_N \geq C) \leq e^{-N \cdot I(C/N,m,v)} \tag{2.1}$$

with $I(C, m, v) = \sup_{s \geq 0} [sC - \log \varphi_{m,v}(s)]$.

This saturation probability can be seen as a relevant estimate for the actual Cell Loss Ratio (CLR), known as the standardized performance index for ATM connections. Now, consider SBR sources with declared parameters PCR $= h$ and SCR $= r$ multiplexed according to the REM scheme. The "worst case" model assumes that the bit rate λ of a source can have two states, namely, h with probability r/h and 0 with probability $1 - r/h$ [ITU, Recommendation I.371, GK94]. We then have $m = r$, $v_0 = r(h - r)$ and

$$\varphi_{r,v_0}(s) = 1 - \frac{r}{h} + \frac{r}{h} e^{sh}. \tag{2.2}$$

The resource allocation per source, also called effective bandwidth, can be evaluated as the ratio of C to the number N of identical sources which are admissible for a given saturation probability. A straigthforward calculation using bound (2.1) shows that, for the "worst case" model, the effective bandwidth e_0 is the minimum of h and the unique solution (greater than r) of equation

$$\frac{1}{h} \log \frac{e_0(h - r)}{r(h - e_0)} - \frac{1}{e_0} \log \frac{h - r}{h - e_0} = -\frac{\log \varepsilon}{C}, \tag{2.3}$$

where ε is the desired value of the saturation probability.

2.2. A conservative upper bound

We now assume that the value of the source mean rate (obtained by measurement) is not r but is equal to some value $m \leq r$. Given the peak rate and the mean rate, the worst case traffic is the on/off source described above, but with r replaced by m and corresponding variance $v = m(h - m)$. The associated Laplace transform of λ is then defined by

$$\varphi_{m,m(h-m)}(s) = 1 - \frac{m}{h} + \frac{m}{h} e^{sh} \tag{2.4}$$

and the corresponding effective bandwidth $e = e(m)$ is the minimum of h and the unique solution (greater than m) of equation

$$\frac{1}{h}\log\frac{e(h-m)}{m(h-e)} - \frac{1}{e}\log\frac{h-m}{h-e} = -\frac{\log\varepsilon}{C}. \tag{2.5}$$

When m tends to zero, the effective bandwidth $e(m)$ also tends to zero and, for $m = r$, $e(r) = e_0$ as provided by (2.3). It can be shown that the effective bandwith is a concave function of the mean (see Appendix 1). Due to this concavity, any tangent to the curve $e = e(m)$ provides a conservative estimation of the effective bandwidth. We thus suggest to linearise this function by the tangent at the point (r, e_0), as formulated in the following proposition.

Proposition 2.1 *Given a measure of the mean bit rate m of a SBR source with declared parameters PCR = h and SCR = r, a conservative estimation of the effective bandwidth is provided by*

$\bar{e}(m) = e_0 - (r-m)U(e_0)$ *with*

$$U(e_0) = \frac{e_0(e_0-r)}{r(h-r)\log\dfrac{h-r}{h-e_0}}.$$

The value of the tangent at r is obtained simply by differentiating (2.5) with respect to m. For numerical illustration, consider an OC-3 ($C = 155$ Mb/s) output link with the peak rate assumed to be $h = 2$ Mb/s, corresponding to a high quality video teleconference. We fix $\varepsilon = 10^{-9}$.

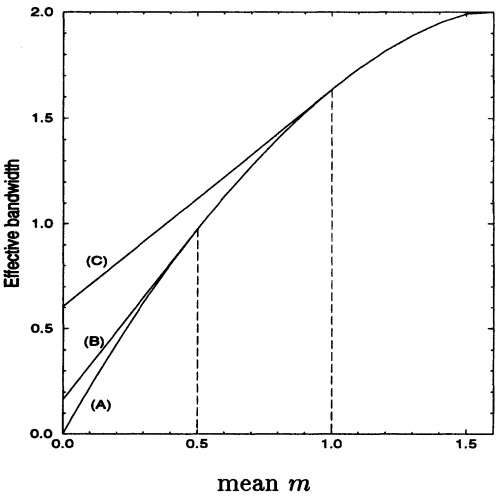

Figure 1. Effective bandwidth w.r.t. the mean bit rate for $C = 155$ Mbit/s and $h = 2$ Mbit/s.

Line (A) of **figure 1** gives the effective bandwidth with respect to the mean rate. Line (B) and line (C) correspond to the tangent at $m = r = 0.5$ Mb/s and $m = r = 1$ Mb/s, respectively.

3. IMPACT OF THE BIT RATE VARIANCE

In this section, we analyse the impact of the variance, assuming that the mean bit rate m is equal to the declared sustainable bit rate r as is the case, for instance, for video sources encoded according to [HRR97]. In this case, we can expect that the source profile delivered by the codec will not be an on/off pattern. We then suggest to measure the variance of the bit rate in order to reduce the bandwidth overallocation given by the on/off model. In this section, the Laplace transform $\varphi_{r,v}$ for fixed $m = r$ is simply denoted by φ_v.

In order to estimate the gain that we can expect from measuring the variance, we first consider an alternative profile source model with the same peak and mean bit rates, namely the three-state model.

3.1. A simple three-state model

Consider the source model where the bit rate can take three different states: the peak rate h, the sustainable rate r and 0, each state being taken with probability α, β and γ, respectively. This model, used for example in [EM91], corresponds to depicting a leaky bucket controlled source as follows. When the source is active and when there are tokens available in the leaky bucket, the source emits at rate PCR; when no tokens are available, the source bit rate is controlled by the leak rate; finally, the source generally has silent periods. v is the *unknown* stationary variance of the bit rate. From relations $r = \alpha h + \beta r$, $v + r^2 = \alpha h^2 + \beta r^2$ and $\alpha + \beta + \gamma = 1$, we readily have

$$\alpha = \frac{v}{h(h-r)}, \quad \beta = 1 - \frac{v}{r(h-r)}, \quad \gamma = \frac{v}{hr}. \tag{3.6}$$

For the three-state model, we have

$$\varphi_v(s) = \alpha e^{sh} + \beta e^{sr} + \gamma \tag{3.7}$$

with α, β and γ defined by (3.6). Note that a variance equal to zero implies that the source emits cells at a constant rate r and the resource allocation is then minimum. The effective bandwidth e is an increasing function of the variance and takes value between r (for a variance equal to zero) and the "worst case" allocation e_0 corresponding to variance $v = v_0$.

To illustrate the above discussion, we consider an OC-3 output link and a movie with peak rate $h = 6$ Mb/s, which can be seen as a typical peak rate value for MPEG movies [PZ95]. The mean rate is either 3, 2 or 1 Mb/s, corresponding to cases (A), (B) or (C). We fix $\varepsilon = 10^{-9}$. For such sources, **figure 2** gives the effective bandwidth $e = C/N$ as calculated from bound (2.1) for φ_v defined in (3.7), as a function of the ratio v/v_0. For cases (B) and (C), the "worst

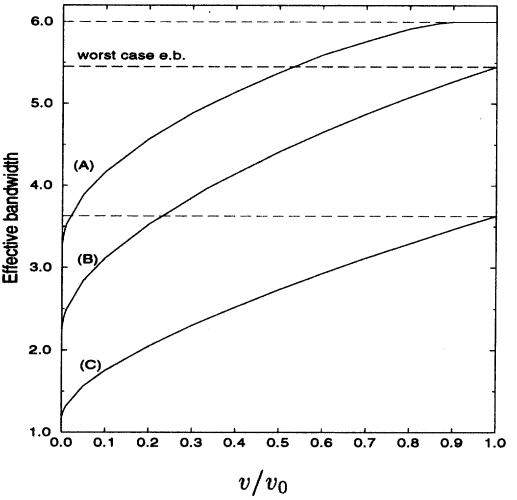

Figure 2. Effective bandwidth w.r.t. the stationary variance for the three-state model. $C = 155$ Mbit/s, $h = 6$ Mbit/s; (A): $r = 3$ Mbits/s, (B): $r = 2$ Mbits/s, (C) $r = 1$ Mbit/s

case" allocation e_0 obtained for $v = v_0$ is also indicated (dashed horizontal line). In case (A), i.e., $h = 6$ Mb/s and $r = 3$ Mb/s, e_0 is constrained by the value of the peak rate. For values of the ratio v/v_0 not too close to the origin, we observe a quasi-linear increase of the effective bandwidth as well as a significant gain with respect to the "worst case" allocation. This example gives some insight into the resource allocation economy which can be derived by the knowledge of the stationary variance of the rate. Recall that the saturation probability has been estimated here by considering a three-state model (with states h, m and 0). We must then note that

- given peak rate and mean rate, the worst case model is the on/off model. This has already been stated in [GK94];
- given peak rate, mean-rate and variance v, however, the "worst case" traffic is *not* the three-state model considered above (the three-state model corresponds to a realistic assumption only for a traffic effectively shaped by a leaky bucket [EM91]).

3.2. General property without specific model

In this section, we do not consider any a priori model, that is, no specific assumption is made on the distribution of random variable λ describing the bit rate of a source with peak rate h, mean rate r and variance v. In order to be able to use the Chernoff bound (2.1), our goal is to find a simple tight upper bound of the Laplace transform of the distribution of λ. This is provided by the following proposition.

Proposition 3.1 *The Laplace transform φ_v of λ verifies*

$\varphi_v(s) \leq \Phi_v(s) = \min(\Phi_v^{(1)}(s), \Phi_v^{(2)}(s))$ *for all* $s \geq 0$, *where*

$$\begin{cases} \Phi_v^{(1)}(s) = 1 + rs + \dfrac{v + r^2}{h^2}(e^{sh} - 1 - sh), \\ \Phi_v^{(2)}(s) = e^{rs}\left(1 + \dfrac{v}{\hat{h}^2}(e^{s\hat{h}} - s\hat{h} - 1)\right) \end{cases}$$

and $\hat{h} = \sup(h - r, r)$.

Proof: as $E(\lambda^k) \leq (v + r^2)h^{k-2}$ for $k \geq 2$, we have

$$\varphi_v(s) = 1 + rs + \sum_{k \geq 2} \frac{s^k E(\lambda^k)}{k!} \leq 1 + rs + \frac{v + r^2}{h^2} \sum_{k \geq 2} \frac{(sh)^k}{k!} = \Phi_v^{(1)}(s).$$

We verify that $\Phi_{v_0}^{(1)}(s) = 1 - r/h + e^{sh}r/h$, which corresponds to the Laplace transform (2.2) for the "worst case" model. For $v = 0$, however, $\Phi_v^{(1)}$ does not correspond to the Laplace transform associated with a source of constant rate m. To obtain another upper bound, consider the centered variable $\hat{\lambda} = \lambda - r$. We have $|\hat{\lambda}| \leq \hat{h} = \sup(h - r, r)$ and $E(\hat{\lambda}^2) = v$. Proceeding as above and noting that $E(\hat{\lambda}^k) \leq E(|\hat{\lambda}|^k) \leq v\hat{h}^{k-2}$ for $k \geq 2$, we derive

$$E(e^{s\hat{\lambda}}) \leq 1 + \frac{v}{\hat{h}^2}(e^{s\hat{h}} - s\hat{h} - 1).$$

We deduce that $E(e^{s\lambda}) \leq \Phi_v^{(2)}(s)$ and verify that $\Phi_0^{(2)}(s) = e^{rs}$, as expected. ■

Typically, $\Phi_v(s) = \Phi_v^{(1)}(s)$ when $v \to v_0$, and $\Phi_v(s) = \Phi_v^{(2)}(s)$ when $v \to 0$. For a given saturation probability ε, using proposition (3.1) and the Chernoff bound (2.1) associated with bound Φ_v, we then derive an upper bound for the effective bandwith of a source with given h, r and v which does not depend on the source profile.

Remark: in the evaluation of the Chernoff bound, we can also write

$$\log \Phi_v^{(2)}(s) \leq rs + \frac{v}{\hat{h}^2}(e^{s\hat{h}} - s\hat{h} - 1),$$

for all v, using the inequality $\log(1+u) \leq u$. We then obtain an easier evaluation of the saturation probability. This evaluation actually corresponds to the so-called *Bennett inequality* [SW86], namely

$$Pr(\Lambda_N \geq C) \leq \exp\left[-\frac{Nv}{h_0^2} \mathrm{H}_1\left(\frac{(C - Nr)h_0}{Nv}\right)\right] \tag{3.8}$$

where $\mathrm{H}_1(z) = (1 + z)\log(1 + z) - z$ is the large deviation function associated with a centered Poisson process with parameter 1.

As presented in Appendix 1, it can be shown that the effective bandwidth $e = e(v)$ is a concave function of the variance v. Moreover, we observe a quasi-linear increase in the part of the line where the estimation of the effective bandwidth is made using bound $\Phi_v^{(1)}$. We consequently suggest to linearise this curve so as to estimate the effective bandwidth $e(v)$ by the tangent at the point $v = v_0$. This estimation provides a conservative estimation of the effective bandwidth.

Proposition 3.2 *A conservative estimation $\hat{e}(v)$ of the effective bandwith is provided by the tangent at the point (v_0, e_0) of the curve $v \mapsto e(v)$, that is,*

$\hat{e}(v) = e_0 - (v_0 - v)T(e_0)$ *with*

$$T(e_0) = \frac{(h-e_0)e_0}{h^2(h-r)\log\left(\frac{h-r}{h-e_0}\right)}\left[\frac{h(e_0-r)}{r(h-e_0)} - \log\left(\frac{e_0(h-r)}{r(h-e_0)}\right)\right].$$

Details of the proof of this proposition have been deferred to Appendix 2.

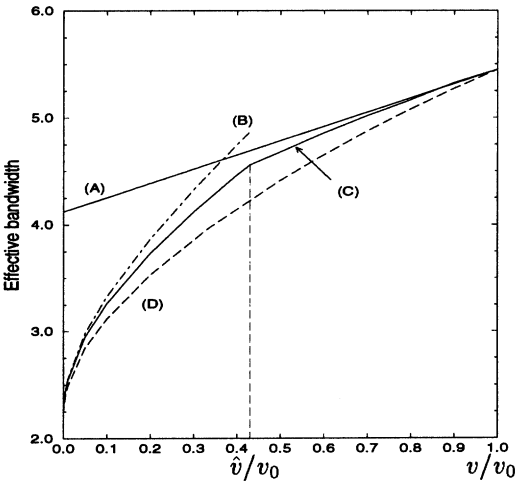

Figure 3. Effective bandwidth w.r.t. the variance for different models. $C = 155$ Mb/s, $h = 6$ Mb/s $r = 2$ Mb/s; (A): conservative evaluation by the tangent, (B): Bennett inequality (C): no a priori model, (D): three-states model.

For numerical illustration, line (C) in **figure 3** provides the effective bandwidth $e = C/N$ with respect to the variance, associated with bound Φ_v for $C = 155$ Mbit/s, $h = 6$ Mbit/s and $r = 2$ Mbit/s. The value of $v = \hat{v}$ above which we use bound $\Phi_v^{(1)}(s)$ instead of bound $\Phi_v^{(2)}(s)$ in the estimation of the saturation probability is also indicated. Line (D) gives the effective bandwidth obtained with the previous three-state model, providing an optimistic resource allocation evaluation compared to the case without a specific model. Line (B) corresponds to the effective bandwidth deduced from the Bennett inequality

(3.8). Line (A) depicts the conservative estimation of the effective bandwidth expressed by Proposition 3.2.

4. ACCEPTANCE REGION FOR HETEROGENEOUS SOURCES

So far, we have considered homogeneous sources, i.e., sources with the same parameter values (h, r, m) (in the case where the mean is measured) and (h, r, v) (in the case where the variance is measured). We now consider J classes of source. A source belonging to class j $(1 \leq j \leq J)$ is defined by the set of parameters (h_j, r_j, m_j) (or (h_j, r_j, v_j)) which is common to all sources belonging to this class. In this section, our purpose is to apply the notion of effective bandwith (calculated by assuming identical sources) to heterogeneous mixtures.

4.1. Case where the mean bit rate is estimated

For each class of parameters (h_j, r_j, m_j), we consider the on/off traffic model with Laplace transform as given in (2.4) where (h, m) is replaced by (h_j, m_j). We evaluate the saturation probability by simply extending the Chernoff bound to the heterogeneous case [Kel91]. We then consider the corresponding acceptance region $\mathcal{A}$, that is, the subset of vectors $(n_1, ..., n_J) \in \mathbb{N}^J$ such that the saturation probability is less than a desired value ε for a number n_j of class j-sources $(1 \leq j \leq J)$. For instance, **figure 4** represents the boundary of the acceptance region for a mixture of video teleconference sources (type 1) with peak rate $h_1 = 2$ Mb/s and $r_1 = 1$ Mb/s and video movie sources (type 2) with peak rate $h_2 = 6$ Mb/s and $r_2 = 2$ Mb/s. Line (A) correponds to the "worst case" allocation, line (B) correponds to the "worst case" allocation but for a "real" mean bit rate of $m_1 = 0.25$ Mb/s and line (C) corresponds to the acceptance region obtained simply by adding the effective bandwidth $\bar{e}(m_1)$ and the "worst case" allocation for the traffic of type 2. We observe the following points:

- the acceptance region boundaries are linear. This has also been verified for quite a large number of traffic mixtures;
- there is still a significant gain using the conservative estimation of the effective bandwidth given in Proposition 2.1, instead of the "worst case" allocation.

4.2. Case where the variance bit rate is estimated

For each class (h_j, r_j, v_j), we now consider the upper bound of the Laplace transform as given in Proposition 3.1 and we proceed as above for finding the acceptance region boundary $\mathcal{A}$. Note that, as for the homogeneous case, the acceptance region does not depend on any specific source model. We have drawn a large number of acceptance regions and we have observed that their boundary is always approximately linear. For example, the left hand side of **figure 5** represents different boundaries of the acceptance region for a particular video traffic mixture. We assume two source classes. Class 1 sources have a peak rate $h_1 = 6$ Mb/s and a sustainable rate $r_1 = 2$ Mb/s; class 2 sources have a peak rate $h_1 = 2$ Mb/s and a sustainable rate $r_1 = 1$ Mb/s. Line (A) corresponds to

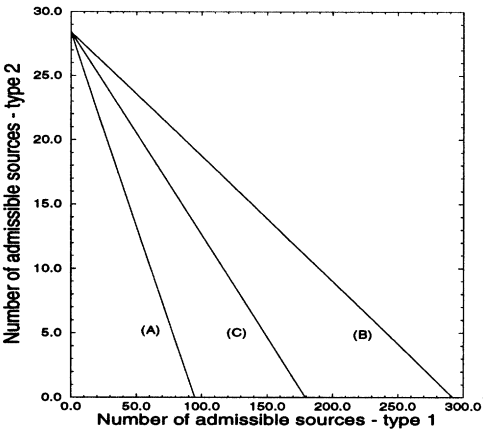

Figure 4. Boundaries of the acceptance region. $C = 155$ Mb/s, $h_1 = 2$ Mb/s, $r_1 = 1$ Mb/s, $m_1 = 0.25$ Mb/s, $h_2 = 6$ Mb/s, $r_2 = 2$ Mb/s.

the boundary for "worst case" traffics, i.e., $v_1 = 8(\text{Mb/s})^2$ and $v_2 = 1(\text{Mb/s})^2$. For line (B) (resp. line (C)), the variance of class 1 (resp. class 2) has been reduced to be equal to 10% of the worst case variance (i.e., $v_1 = 0.8(\text{Mb/s})^2$ for line (B) and $v_2 = 0.1(\text{Mb/s})^2$ for line (C)). Line (D0) corresponds to the case where variances of both classes have been reduced ($v_1 = 0.8(\text{Mb/s})^2$ and $v_2 = 0.1(\text{Mb/s})^2$). By drawing the chord (line (D1)), we have stressed the slight concavity of the acceptance region.

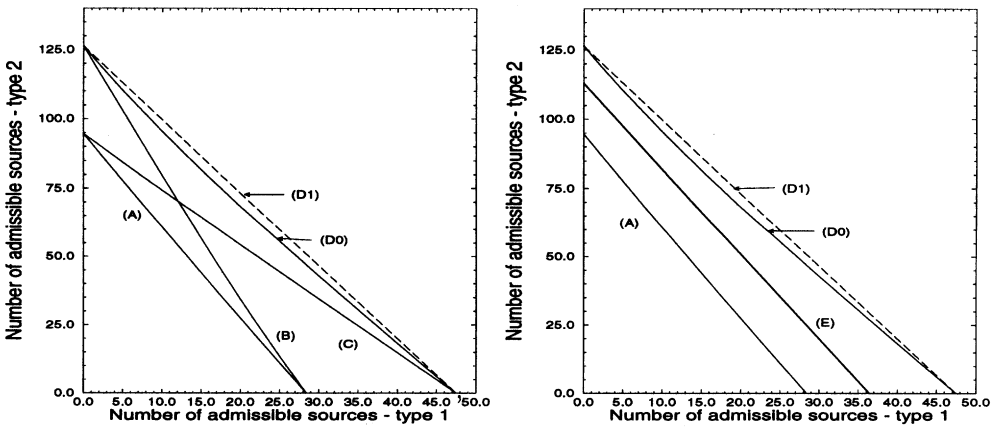

Figure 5. Boundaries of the acceptance region. $C = 155$ Mb/s, $h_1 = 6$ Mb/s, $r_1 = 2$ Mb/s, $h_2 = 2$ Mb/s, $r_2 = 1$ Mb/s.

On the right-hand side of **figure 5**, line (E) represents the linear boundary of the acceptance region deduced by simply adding the evaluations $\hat{e}_j(v_j)$ of the

effective bandwidth with the above values of v_1 and v_2 (curves (A), (D0) and (D1) remaining the same). The acceptance region is of course reduced, since $\hat{e}_j(v_j)$ is an upper bound of the effective bandwidth. We observe, however, the following two points.

- The acceptance region is well included in that delimited by (D0);
- there is still a significant gain with respect to the domain associated with the "worst case" allocation (represented by line (A)).

From an intensive numerical analysis of the acceptance region, we can then deduce the following heuristic proposition.

Proposition 4.1 ***(Heuristic)*** *A conservative estimation of the bandwidth required by a mixture of sources is given by the sum of estimations $\bar{e}_j(m_j)$ (resp. $\hat{e}_j(v_j)$) of the effective bandwidth as defined in Proposition 3.1 (resp. Proposition 3.2) for each source j composing the mixture.*

5. ESTIMATION OF THE REQUIRED BANDWIDTH FROM A GLOBAL MEASURE OF MEAN AND VARIANCE

5.1. Measurement of the total mean and variance

The previous estimations of the effective bandwidth asssume that the set of parameters (h, m, v) of each admitted source is known. In practice, for the SBR transfer capability, only parameters (h, r) are known by the network. Given a link in the network carrying N active sources with declared parameters (h_i, r_i) $(1 \leq i \leq N)$, a measurement of the mean $\mathcal{M}_N$ and variance $\mathcal{V}_N$ of the *total* bit rate offered to a given link can be performed on a link-by-link basis. As shown in Section 5.2 below, these measurements can enable us to use the results of the above sections for estimating the global bandwidth allocation on each link.

Now, address the way such global means and variances can be effectively measured. By stability, the mean input rate $\mathcal{M}_N$ is identical to the product $C \times \varrho$ where C is the link capacity and ϱ is the output link load. Measuring the latter can be performed by means of an "Exponentially Weighted Moving Average" (EWMA) algorithm which we recall below. Given $\mu \in]0, 1[$, define the sequence S_ℓ by $S_{\ell+1} = \mu S_\ell + (1 - \mu)R_\ell$, where $R_\ell = 1$ or 0 if a cell is transmitted in the ℓ-th slot (with duration the time transmission unit) or not. It is then known [CM65] that $S_\ell \to E(R) = \varrho$ as $\ell \to \infty$. It can be noted that the time scale for the convergence is of the order of $1/(1 - \mu)$.

The variance $\mathcal{V}_N$ can be classically estimated [DMM94] through the L-size sample $\Lambda(t_1)$, $\Lambda(t_2)$,..., $\Lambda(t_L)$ of the total bit rate at successsive instants t_1, t_2,...,t_L, by

$$\mathcal{V}_N = \frac{1}{L-1}\sum_{\ell=1}^{L}[\Lambda(t_\ell) - \mathcal{M}_N]^2 \quad \text{with} \quad \mathcal{M}_N = \frac{1}{L}\sum_{\ell=1}^{L}\Lambda(t_\ell).$$

The order of increment $t_\ell - t_{\ell-1}$ should be taken small with respect to the duration of a typical connection. The value of the number L of samples can be chosen so that the previous empirical mean comes close to the estimate provided by the EWMA algorithm recalled above. We can assume that the number of connections N remains unchanged during the time duration $t_L - t_1$ of the sample. However, in the case where quite a large number of video connections is carried on the link, we can simply update the value of the variance when its value changes significantly in time. Indeed, the departure of one connection has negligible impact on the variance of the global bit rate and it is not necessary to consider a measurement interval corresponds to a fixed number of active connections.

5.2. General framework for the estimation of the required bandwidth

5.2.1. Case where the total mean is measured

We here assume that the particular mean bit rate m_i due to any source i remains unknown but the global mean rate $\mathcal{M}_N$ is estimated as detailed previously. Thanks to Proposition 4.1 which enables us to simply add the effective bandwidth of the different sources, write that the bandwidth E_N required by the N sources on the link is such that

$$E_N \leq \max_{m_1,..,m_N}\left(\sum_{i=1}^{N}\bar{e}_i(m_i)\right) = \bar{E}_N \tag{5.9}$$

under constraints

$$\sum_{i=1}^{N} m_i = \mathcal{M}_N \quad \text{and} \quad m_i \leq r_i \text{ for } 1 \leq i \leq N, \tag{5.10}$$

where $\bar{e}_i(m_i)$ is the effective bandwidth associated with source i. Considering Proposition 3.1, we write $\bar{e}_i(m_i) = \alpha_i + \beta_i m_i$, where $\alpha_i = e_{0,i} - r_i U_i(e_{0,i})$, $\beta_i = U_i(e_{0,i})$ with the definition of U_i and $e_{0,i}$ being naturally extended from that of U and e_0. The system (5.9) and (5.10) can then be seen as a classical linear optimisation problem under linear constraints. The solution $\bar{E}_N$ can be actually obtained in an explicit manner [Tid97]. Let the N already admitted sources be numbered such that $\beta_1 \geq \beta_2 \geq ... \geq \beta_N$. Then, it can be easily shown that

$$\bar{E}_N = \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{K}\beta_i r_i + \beta_{K+1}\cdot(\mathcal{M}_N - \sum_{i=1}^{K} r_i) \tag{5.11}$$

where the index K is determined by $\sum_{i=1}^{K} r_i \leq \mathcal{M}_N \leq \sum_{i=1}^{K} r_i + r_{K+1}$.

Note that we have assumed that the measure of the global variance leads to an equality constraint. We can also introduce a confidence interval $\Delta\mathcal{M}_N$ associated with this measure and replace the equality constraint by the condition $\mathcal{M}_N - \Delta\mathcal{M}_N \leq \sum_{i=1}^{N} m_i \leq \mathcal{M}_N + \Delta\mathcal{M}_N$.

5.2.2. Case where the total variance is measured

The same framework can be applied if the mean rate of each source is assumed to be equal to its sustainable bit rate but the variance of the total bit rate is measured, the particular variance v_i of the bit rate of source i being unknown. The system is then

$$E_N \leq \max_{v_1,..,v_N} \left(\sum_{i=1}^{N} \hat{e}_i(v_i) \right) = \hat{E}_N \tag{5.12}$$

under constraints

$$\sum_{i=1}^{N} v_i = \mathcal{V}_N \quad \text{and} \quad v_i \leq r_i(h_i - r_i) \text{ for } 1 \leq i \leq N, \tag{5.13}$$

where $\hat{e}_i(v_i)$ is the effective bandwidth associated with source i. Considering Proposition 3.1, we write $\hat{e}_i(v_i) = \gamma_i + \delta_i v_i$, where $\gamma_i = e_{0,i} - r_i(h_i - r_i)T_i(e_{0,i})$ and $\delta_i = T_i(e_{0,i})$. Note that, exactly as in the case where the total mean is supposed to be measured, the value of $\hat{E}_N$ can be obtained in an explicit manner.

5.3. Numerical estimation of the gain

Let $p_N^{(1)}$ (resp. $p_N^{(2)}$) $\in [0,1]$ be the ratio between the measure of the global mean (resp. the variance) and the sum of the SCR (resp. "worst case" variance), that is,

$$p_N^{(1)} = \frac{\mathcal{M}_N}{\sum_{i=1}^{N} r_i}, \quad p_N^{(2)} = \frac{\mathcal{V}_N}{\sum_{i=1}^{N} r_i(h_i - r_i)}.$$

Denote also by $\mathcal{G}_N^{(1)}$ (and $\mathcal{G}_N^{(2)}$) the gain measured in % due to the estimation of the mean $\mathcal{M}_N$ (resp. the variance $\mathcal{V}_N$). The latter is defined as the complement to the ratio between the consumed bandwidth $\bar{E}_N$ (resp. $\hat{E}_N$), given this measure and the bandwidth without this knowledge, that is,

$$\mathcal{G}_N^{(1)} = 1 - \frac{\bar{E}_N}{\sum_{i=1}^{N} e_{0,i}}, \quad \mathcal{G}_N^{(2)} = 1 - \frac{\hat{E}_N}{\sum_{i=1}^{N} e_{0,i}}.$$

The left part of **figure 6** gives $\mathcal{G}_N^{(1)}$ with respect to $p_N^{(1)}$ for the following types of video teleconferencing sources: sources of type (I) and (II) have a peak rate equal to $h = 2$ Mb/s and a sustainable rate equal to $r = 1$ Mb/s and $r = 0.5$

Mb/s, respectively.

- Line (A) corresponds to the case where there are 159 sources of type (II) and no source of type (I);
- line (B) corresponds to the case where there are 47 sources of type (I) and 79 sources of type (II);
- line (C) corresponds to the case where there are 94 sources of type (I) and no source of type (II).

We have considered some limit cases where no source can be accepted whatever its type might be. The limit case when $p_N^{(1)} = 1$ corresponds to the "worst case" allocation (no measure or no information from it). The bend in the line (B) corresponds to a change in the solution of the linear optimisation problem (5.9) and (5.10). We observe that, even for a load around 70% the sum of the sustainable bit rate, the gain can be significant (around 20%).

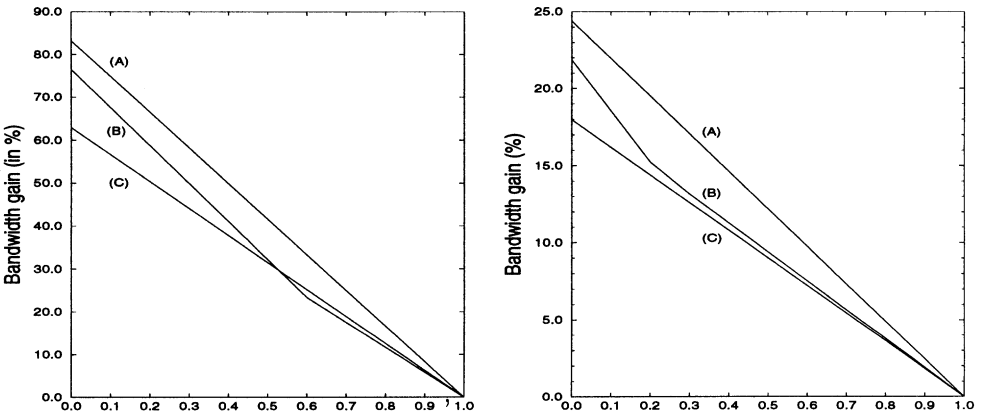

Figure 6. Bandwidth gain w.r.t. the measure of the normalised mean (left side) and variance (right side).

The right part of **figure 6** gives $\mathcal{G}_N^{(2)}$ with respect to $p_N^{(2)}$ for the types of sources already considered in section 4.2. Sources of type (I) and (II) have a peak rate equal to $h = 6$ Mb/s and $h = 2$ Mb/s, respectively, and a mean equal to $r = 2$ Mb/s and $r = 1$ Mb/s, respectively.

- Line (A) corresponds to the case where there are 28 sources of type (I) and no source of type (II);
- line (B) corresponds to the case where there are 17 sources of type (I) and 37 sources of type (II);
- line (C) corresponds to the case where there are 94 sources of type (II) and no source of type (I).

We have also considered some limit cases where no source can be accepted whatever its type might be. Off-line measurements of stationary variance for video applications indicate that one can expect the variance to be around 10 % or 20 % of the "worst case" variance [HR95]. In this domain, we observe significant bandwidth gain allowing acceptance of more connections.

Table 1 gives the bandwidth gain for a given measured normalised variance $p_N^{(2)} = 0.15$ and for homogeneous sources with varying peak rate h and burstiness h/m. The gain is particularly important for high peak rate and high burstiness. Entry ($\star$) corresponds to no multiplexing gain for "worst case" traffic, i.e., $e_0 = h$ and the tangent slope at (v_0, h) thus equals to zero (this is the same case as line (A) in figure 1)).

Table 1
Bandwidth gain with respect to the peak cell rate for a measured normalised variance of $p_N = 0.15$

$\mathcal{G}_N$	$h/r = 20$	$h/r = 10$	$h/r = 5$	$h/r = 3$
$h = 10$ Mb/s	50.6	40.8	24.0	$\star$
$h = 6$ Mb/s	45.4	40.6	31.8	20.7
$h = 2$ Mb/s	29.1	27.4	24.5	20.5

5.4. Application to the CAC for SBR capability

Assume that a new video connection with declared parameters (h_0, r_0) arrives for acceptance in the network. On each link k of the network path where this connection should be carried, we assume that a measure of the global mean bit rate and/or variance bit rate are performed. Note that in the case where there are only video teleconferencing connections carried in the network, it is not necessary to measure the variance. We then compute the two estimations of the bandwidth $\bar{E}^{(k)}$ and $\hat{E}^{(k)}$, as explained in Sections 5.2.1 and 5.2.2. The required bandwidth $E^{(k)}$ is taken as the minimum between these two values. The new connection is modelled by the corresponding "worst case" source, i.e., the source with parameters (h_0, r_0, v_0) for which we evaluate the effective bandwidth $e_0^{(k)}$ associated with link k. If, for each k, $E^{(k)} + e_0^{(k)}$ is less that the capacity $C^{(k)}$ of link k, the connection is accepted.

5.5. Comparison to the Gaussian model

Assuming N leaky bucket controlled sources with aggregate mean $\mathcal{M}_N$ and measured variance $\mathcal{V}_N$, it is natural to model the total bit rate by a Gaussian random variable $\mathcal{N}(\mathcal{M}_N, \mathcal{V}_N)$ (e.g. [Reg95, COST242]). The approximation of the bit rate distribution by a Gaussian variable is assumed to be accurate for *sufficiently*

large N. In this case, we write the saturation probability P_{sat} as

$$P_{sat} = \int_C^{+\infty} \frac{1}{\sqrt{2\pi \mathcal{V}_N}} \exp\left(-\frac{(x-\mathcal{M}_N)^2}{2\mathcal{V}_N}\right) dx.$$

For large N and for a desired $P_{sat} = \varepsilon$, we then deduce the bandwidth required E_N by N sources as being given by

$$E_N \approx \mathcal{M}_N + \sqrt{-2\mathcal{V}_N \log \varepsilon}. \tag{5.14}$$

When a new connection with declared parameters (h_0, r_0) arrives for acceptance on the link, we estimate E_{N+1} by using (5.14) with the following changes: $\mathcal{M}_N \to \mathcal{M}_{N+1} = \mathcal{M}_N + r_0$ and $\mathcal{V}_N \to \mathcal{V}_{N+1} = \mathcal{V}_N + r_0(h_0 - r_0)$. If the estimated value of E_{N+1} is less than the capacity, then the new connection is accepted. This is the basic CAC principle under this Gaussian assumption. This principle is simple but it has two main drawbacks:

- the convergence of the total bit rate to a Gaussian variable is not tightly controlled. The convergence speed in the case of homogeneous sources is indicated either by the Berry-Esseen theorem [Fel71, p.542] or [Shi84], or by using large deviations refinement through expressions depending of the third (and higher order) moments of the total bit rate [Fel71, p.342]. In both cases, considering the small target saturation probability, we cannot find satisfactory expressions for estimating the error made when we use the Gaussian model;
- in the case of heterogenous sources, the number of sources of each class has to be large: the Gaussian model, therefore, is not very appropriate for a heterogeneous aggregation of streams.

6. CONCLUSION

In this paper, we have analysed the impact of the stationary mean and variance of the bit rate on the effective bandwidth of video sources multiplexed using the REM scheme. More particularly, a conservative estimate of the effective bandwidth, independent of any other statistical properties of the source, has been derived. This estimate depends linearly on the mean (or on the variance) of the source bit rate. In the case where we can only measure the total mean (or variance) of the stationary bit rate on the link, we have derived a conservative estimation of the required resources on a specific link. This resource estimation leads to bandwidth gain when compared to the "worst case" resource allocation. The method gives important bandwidth economy for video sources either with a "real" mean significantly less than the sum of the sustainable bit rates or with a "real" variance significantly less than the "worst case" variance. Finally, the CAC mechanisms proposed in this paper rely on efficient measurement processes for mean and variance parameters. A few suggestions have been given for such processes ("Exponentially Weighted Moving Average" algorithm, empirical statistics) but the latter need to be further studied both in terms of physical im-

plementation and in terms of performance (convergence speed, precision). These questions are currently investigated.

Acknowledgement: we thank J. Roberts and L. Massoulié (FT-CNET) for useful discussions about this work.

BIOGRAPHY

• Alain Simonian received degrees from the Ecole Polytechnique, Paris, France in 1982 and from the Ecole Nationale Supérieure des Télécommunications in 1984. Since 1984, he has been working in the Centre National d'Etudes des Télécommunications, the France Telecom research center, in the area of Performance Evaluation of multiservice telecommunication networks. He received a Ph.D. Degree in Applied Mathematics from the University of Rennes in 1993. He is currently head of a research team dealing with the development of techniques from applied probability and queueing theory and their application to the modelling and Performance Analysis of broadband telecommunication networks.

• François Brichet received degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France in 1991. Since 1994, he has been working in the Centre National d'Etudes des Télécommunications in the area of Performance Evaluation of multiservice telecommunication networks. He is a member of the European Project COST 257 on "The impact of new services on the architecture and the performance of broadband network".

REFERENCES

[ATMF96] ATM Forum (1996), *Traffic management specification version 4.0*, Technical report contribution 95-0013R10, ATM Forum.

[COST242] J. Roberts, U. Mocci, J. Virtamo Editors (1996), *Broadband network teletraffic*, Final Report of Action COST 242, Springer Verlag, Lecture Notes in Computer Science 1155.

[CM65] D. R. Cox and H. D. Miller (1965), *The theory of stochastic processes*, Methuen, London.

[DMM94] Z. Dziong, O. Montanuy, L. G. Mason (1994), *Adaptive traffic admission in ATM networks - optimal estimation framework*, Proceedings ITC14, J.Labetoulle, J.Roberts ed., Elsevier.

[EM91] A. I. Elwalid, D. Mitra (1991), *Rate-based congestion control* Communication Systems D. Mitra and I. Mitrani Editors, Queueing Systems - theory and applications.

[Fel71] W. Feller (1971), *An introduction to probability theory and applications*, Volume II, J. Wiley editor, 1971.

[GW91] M. W. Garret and W. Willinger (1994), *Analysis, modeling and generation of self-similar VBR video traffic.* In Proceedings of SigComm 94, ACM.

[GKK95] R. J. Gibbens, F. P. Kelly, P. Key (1995), *A decision-theoretic*

approach to call admission control in ATM networks, IEEE Journal on Selected Areas in Communications, Special issue, 13(6): pp. 1101-1114.

[GK94] R. Griffiths, P. Key (1994), *Adaptive Call Admission Control in ATM Netwoks*, Proceedings of ITC 14, J. Labetoulle, J. Roberts ed., Elsevier, pp. 1089-1098.

[HRR97] M. Hamdi, J. W. Roberts, P. Rolin (1997), *Rate control for VBR video coders in broadband networks.* IEEE JSAC Vol.15, No 6, August 1997.

[HR95] M. Hamdi, P. Rolin (1995), *Resource requirements for VBR Mpeg traffic in interactive applications.* Technical Document RR-94017-RSM.

[HL96] D. P. Heyman, T.V Lakshman (1996), *Source models for VBR broadcast-video traffic*, IEEE/ACM Transactions on Networking, Vol.4, N0.1, February 1996.

[ITU, Recommendation E.73x] ITU-T Draft Recommendation E.73x (May 1996), *Methods for cell level traffic control in B-ISDN*, Geneva.

[ITU, Recommendation I.371] ITU-T Recommendation I.371 (May 1996), *Traffic control and congestion control in B-ISDN*, Geneva.

[Kel91] F.P. Kelly (1991), *Effective bandwidth at multi-class queues*, Queueing systems, Communication Systems, D. Mitra, I. Mitrani Editors.

[PZ95] P. Pancha and M. El Zarki (1995), *Leaky bucket access control for VBR MPEG video*, in Proceedings of the IEEE Infocom 95, Boston.

[Reg95] R. Rege (1995) *Equivalent bandwidth and related adimission criteria for ATM systems - a perforamce study*, International Journal of Communications Systems, Vol 7.

[RB95] A.R. Reibman, A.W. Berger (1995), *Traffic descriptors for VBR video teleconferencing over ATM networks*, IEEE/ACM Transactions on networking, Vol. 3, No. 3, June.

[Rob95] J.W. Roberts (1995), *What traffic handling capabilities for the B-ISDN?*, ITC Specialists Seminar, Leidschendam.

[Shi84] A. N. Shiryayev (1984), *Probability*, Springer Verlag.

[SW86] G. R. Shorak and J. A. Welner (1986), *Empirical processes with applications to statistics.*, J. Wiley editor.

[Tid97] S-E. Tidblom, *Improving the utilization of an ATM-link by meauring its load.*, COST 257 Technical Document 57.

APPENDIX

Appendix 1. Let $\Gamma = \{c \in \mathbb{R}^+ \mid c \leq e\}$ denote the domain under the curve $e = e(m)$ of section 2.1 (resp. $e = e(v)$ of section 3.2). From (2.1), we derive that $c \in \Gamma \Longleftrightarrow N \cdot I(c, m, v) \leq -\log \varepsilon$ with $c = C/N$, that is,

$$c \in \Gamma \Longleftrightarrow \forall s \geq 0,\ sc - \log \varphi_{m,v}(s) \leq ac, \tag{A.1}$$

where constant $a = -\log \varepsilon / C$ is fixed.

a) For varying m and "worst case" variance set to $v = m(h - m)$, we replace

$\varphi_{m,v}(s)$ by expression (2.4). By taking the exponential of each side of inequality $(A.1)$, we obtain

$$\frac{m}{h}(e^{sh}-1)+1 \geq e^{(s-a)c}. \tag{A.2}$$

Function $c \mapsto e^{(s-a)c}$ being convex, $(A.2)$ defines, for each $s \geq 0$, a convex domain Γ_s in the plane (c,m). The set Γ is therefore convex as the intersection of convex sets Γ_s.

b) Similarly, for fixed $m = r$ and variable variance $v \leq r(h-r)$, we replace $\varphi_{m,v}(s)$ by $\Phi_v^{(1)}(s)$. Condition $(A.1)$ can then be readily written as

$$a(s)v + b(s) \geq e^{(s-a)c} \tag{A.3}$$

with positive $a(s)$. For each $s \geq 0$, inequality $(A.3)$ defines a convex domain Γ'_s in the plane (c,v). The corresponding set Γ is therefore convex as the intersection of convex sets Γ'_s.

Appendix 2. We here detail the calculation for the slope $T(e_0)$ of the tangent at $v = v_0$ to the graph of function $v \mapsto e(v)$. The effective bandwidth with respect to v is defined by $e(v) = C/N(v)$. Differentiating $e(v)$ with respect to v, we then have $T(e_0) = \left(\frac{de}{dv}\right)_{v_0} = -\frac{e_0}{N_0}\left(\frac{dN}{dv}\right)_{v_0}$. To compute the latter derivative, recall from (2.1) that $N(v)$ satisfies the relation

$$N(v) \cdot I\left(\frac{C}{N(v)}, v\right) = \log \varepsilon. \tag{B.1}$$

for any variance v. Differentiating (B.1), we deduce

$$\left(\frac{dN}{dv}\right)_{v_0} = \frac{-N_0\left(\frac{\partial I}{\partial v}\right)_{v_0}(e_0,v_0)}{I(e_0,v_0) - e_0\left(\frac{\partial I}{\partial e}\right)_{e_0}(e_0,v_0)}. \tag{B.2}$$

The partial derivative of $I(e_0,.)$ at point (e_0,v_0) is calculated by setting $v = v_0 - \delta$ for small δ. We first derive the expansion of s, the value which achieves the maximum $I(e_0,v)$ in (2.1), to order $o(\delta)$. We then develop $I(e_0,v) = e_0 s - \log(\Phi_v^{(1)}(s))$ with respect to δ. The calculation gives

$$\left(\frac{\partial I}{\partial v}\right)_{v_0}(e_0,v_0) = \frac{e^{s_0 h} - 1 - h s_0}{h(h-m+me^{s_0 h})}. \tag{B.3}$$

From the definition of function $I(.,v_0)$, we readily have

$$\left(\frac{\partial I}{\partial e}\right)_{e_0}(e_0,v_0) = s_0. \tag{B.4}$$

Substituting (B.3) and (B.4) in (B.2), we obtain $T(e_0) = \left(\frac{de}{dv}\right)_{v_0}$ as given in Proposition 3.2.