Optimal Smoothing for Guaranteed Service

Jean-Yves Le Boudec and Olivier Verscheure
EPFL
Institute for Computer Communications and Applications (ICA)
1015-Lausanne, Switzerland

Abstract

We consider a scenario where multimedia data is sent over a network offering a guaranteed service such as ATM VBR or the guaranteed service of the IETF. A smoothing device writes the stream into a networking device for transmission, possibly with some pre-fetching; at the destination, the decoder waits for an initial *playback delay* and reads the stream from the receive buffer. We consider the problem of whether there exists a smoothing which minimizes the playback delay and the receive buffer size over all possible strategies, given that we know a service curve property for the flow in the network. We show that there does exist such an optimal smoothing. It can be expressed using the deconvolution operator of min-plus algebra. We obtain the smallest playback delay which can be achieved by smoothing, provided that the information about the network is reduced to its service curve β. We also give a constructive expression for the deconvolution operator, using a time inversion transform, introduced in the paper. We illustrate on some examples the difference with optimal shaping, a smoothing strategy which aims at minimizing buffer and delay on the sender side but does not allow pre-fetching. We apply the theory to the determination of the minimum T-SPEC required to support a given flow with admissible playback delay or decoding buffer size constraints.

1 Introduction

We consider scenarios of transmission of *variable bit rate* (VBR) video over a guaranteed network [14, 10]. One approach, called *rate control*, consists in modifying the encoder output so that it becomes compliant with the negotiated arrival curve [13, 17, 9]. Rate control is considered as an important issue in video coding since it significantly affects video quality. An alternative approach is to smooth the video stream. Many researches have already focused on video bandwidth smoothing techniques. In [16, 12, 11], the aim is to reduce the burstiness of VBR stream by pre-fetching data *at a series of fixed rates*. Using a series of fixed rates simplifies the allocation of resources in video servers and the communication network. Performance evaluation of these techniques is given in [15].

In this paper, we consider the transmission of a multimedia VBR stream over one VBR channel (e.g., ATM VBR or the guaranteed service of the IETF). A smoothing device writes the stream into a networking device for transmission, possibly with some pre-fetching; at the destination, the decoder waits for an initial *playback delay* and reads the stream from the receive buffer. We consider the problem of whether there exists a smoothing which minimizes the playback delay and the receive buffer size over all possible strategies, given that we know a service curve property for the flow in the network. We show that there does exist such an optimal smoothing.

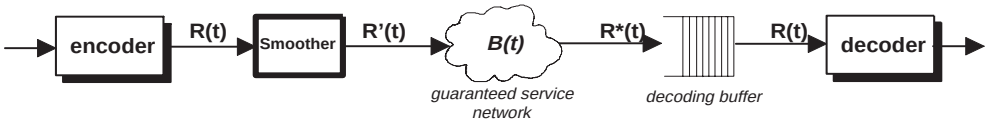


Figure 1: *The prefetching scenario.*

The scenario is illustrated on Figure 1. A multimedia stream is encoded, and then input into a smoothing device. The smoothing device writes the stream into a networking device for transmission; its output R' is constrained by a specified arrival curve σ. For example, a flow conforming to the IETF specification for integrated service [5], with maximum packet size M, peak rate P, sustainable rate r and burst tolerance b, has an arrival curve defined by $\sigma(t) = \min(M + Pt, b + rt)$ for $t > 0$ and $\sigma(t) = 0$ if $t \leq 0$. A similar definition holds for ATM variable bit rate services. We call $R(t)$ the cumulative number of bits observed on the encoded flow, starting from an arbitrary point in time, and $R'(t)$ the output of the smoother. The smoothing constraint is expressed as $R'(t+u) - R'(t) \leq \sigma(u)$ for all $u \geq 0$. It can also be written as $R' \leq R' \otimes \sigma$, where $\otimes$ is the min-plus convolution operation, defined by $(R' \otimes \sigma)(t) = \inf_u (R'(t-u) + \sigma(u))$ [3, 1].

We assume that the data is carried by a network offering the guaranteed service of the IETF. This implies that the transformation imposed on R' by the network can be decomposed into a fixed delay, and a variable delay. The variable delay can be assumed to satisfy a constraint expressed by the service curve concept [1, 5].

At the destination, the receiving device stores incoming bits into a buffer before passing them to the decoder. The decoder starts reading from the decoding buffer after a delay D, and then reads at a rate imposed by $R(t)$. The delay D is called the *playback delay* of the receiver.

Note that we allow some prefetching, namely, we do not require that $R'(t) \leq R(t)$. Prefetching is commonly used with pre-recorded streams for which the smoother S is composed of both a disk server and a scheduler.

A popular smoothing device is the shaper [3, 1]. A shaper is a device which outputs bits whenever doing so does not violate the arrival curve constraint; otherwise it stores the bits in a buffer. An optimal shaper is one that maximizes the number of bits output on any time interval. For an optimal shaper, the output R' is given by $R'(t) = (R \otimes \sigma)(t)$. The formula is true under the assumption that σ is sub-additive (namely $\sigma(s+t) \leq \sigma(s) + \sigma(t)$) and $\sigma(t) = 0$ for $t \leq 0$. It is known that these technical conditions on σ are not a restriction, since any arrival curve can be replaced by one which satisfies them. The arrival curves defined for Internet integrated services or for ATM and mentioned above do satisfy these assumptions, as do any concave arrival curves [1].

It can be shown that an optimal shaper minimizes the buffer requirement and the delay experienced in the smoother. However, an optimal shaper is optimal only at the sender side. In this paper we consider another problem, namely, we would like to minimize the playback delay D and the buffer size at the receiver. Also note that the causality restriction $R'(t) \leq R(t)$ is true for an optimal shaper, but is not required from our smoother.

In order to simplify the arguments, we consider first an intermediate, simplified problem, where the network part is not considered (Problem 1). With Problem 1, the issue is to find a smoothed output $R'(t)$ which satisfies the following constraints.

Definition 1.1 (Constraints for Problem 1) *For Problem 1, the smoother must produce an output R' such that*

1. *(Smoothing Constraint):* $R' \leq R' \otimes \sigma$
2. *(Real-Time Constraint):* $R'(t) \geq R(t - D)$

A smoothing solution is a couple (R', D), where R' is the output of the smoother, with $R'(t) = 0$ for $t < 0$, and where D is the playback delay. With Problem 1, we try to find a smoothing solution which minimizes the playback delay D and the buffer constraint at the receiver. We show that there exists indeed such a solution. It is given by applying the deconvolution operator, defined later in the paper, to the initial input $R(t)$, and then shifting in time in order to consider only non-negative times. Among all solutions satisfying the constraints above, the corresponding playout delay, which is the horizontal deviation $h(R, \sigma)$ between the functions R and σ, is then minimal, and so is the buffer requirement at the receiver.

Then we consider the general problem with a non-null network (Problem 2). Here the network is assumed to provide a variable plus a fixed delay. First, we can reduce to the case where the fixed delay is zero, since it does not impact the smoothing method. Second, the variable part of the delay can be assumed to satisfy a service curve constraint β. This means that the relationship between R' and R^* cannot be known exactly by the sending side, however, we can assume that [1]

$$R^*(t) \geq (R' \otimes \beta)(t)$$

For example, the Internet guaranteed service assumes that every node offers a service curve of the form $\beta(t) = \rho(t - L)^+$ for some latency L and rate ρ. It is further assumed that the latency parameter L depends on the rate ρ according to $L = \frac{C_0}{\rho} + E_0$ for some constants C_0 and E_0. The values of C_0 and E_0 are computed during reservation setup, with a protocol such as the Resource Reservation Protocol (RSVP) [1, 2].

With Problem 2, we consider service strategies that ignore the details of the network, but do know the service curve β. Thus, with Problem 2, the constraints for a smoother are the following.

Definition 1.2 (Constraints for Problem 2) *For Problem 2, the smoother must produce an output R' such that*

1. *(Smoothing Constraint):* $R' \leq R' \otimes \sigma$
2. *(Real-Time Constraint after traversing a network offering a service curve β):*$(R' \otimes \beta)(t) \geq R(t - D)$

As with Problem 1, we show that there indeed exists an optimal solution, namely a solution which minimizes playback delay and buffer size at the receiver. It is also expressed using the deconvolution operator. We also show that the minimal playback delay is the horizontal deviation $h(R, \sigma \otimes \beta)$ between the functions R and $\sigma \otimes \beta$.

The paper continues as follows. In Section 2, we give the theoretical results on deconvolution which form the basis of our results. We also introduce the time inversion transform, and use it to give a constructive definition of deconvolution. In Section 3 we give an application to Problems 1 and 2

and show the existence of optimal smoothing strategies. In Section 4, we give some applications to corollary problems, such as determining the minimum T-SPEC required to support a given flow with a maximum admissible playback delay D. In Section 5, we give numerical applications using multimedia streams encoded with MPEG-2.

2 Deconvolution as a Smoothing Operation

In this section we introduce some new network calculus concepts which support our study of Problems 1 and 2.

2.1 Previous Results on Deconvolution

Call $\mathcal{F}$ the set of wide-sense increasing functions of time with values in $[0, +\infty]$, which are equal to 0 for very negative values. More precisely, a function $t \to S(t)$ is in $\mathcal{F}$ if it is wide-sense increasing, if $S(t) \geq 0$ and if there exists some T_0 such that $S(t) = 0$ if $t \leq T_0$. It is traditional to consider $T_0 = 0$; in other words, to consider only non-negative times. However, in this paper, it is more convenient to allow some negative times. Functions in $\mathcal{F}$ are used to represent the cumulative number of bits observed on a flow, starting from an arbitrary point in time.

The key operation for prefetching is deconvolution, which we recall now.

Definition 2.1 (Deconvolution [6]) *For two functions of time f and g, the deconvolution $f \ominus g$ is defined by*

$$(f \ominus g)(t) = \sup_{u \in \mathbb{R}} \{f(t+u) - g(u)\}$$

Note that $f \ominus g$ may be non-zero for negative times even if this is not the case for f and g. Also note that if $g(0) = 0$ then $(f \ominus g)(t) \geq f(t)$. Figure 2.1 shows the value of $S \ominus \sigma$ when S is an impulse function.

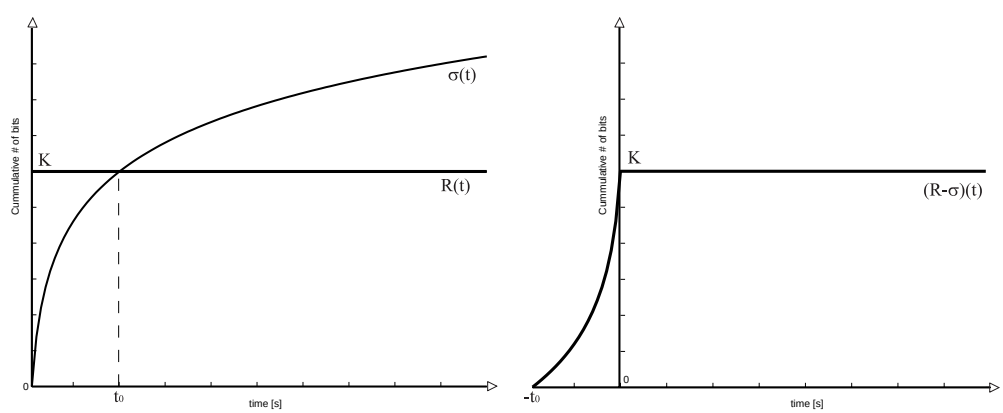


Figure 2: *$R(t)$, $\sigma(t)$ and the deconvolution $(R \ominus \sigma)(t)$.*

The deconvolution operator has the following properties [6, 4] :

Theorem 2.1 (Properties of Deconvolution) *For any three functions of time f, g and h:*

1. $(f \ominus g) \leq h$ *if and only if* $f \leq (g \otimes h)$
2. $(f \ominus g)$ *is the minimum solution to the problem* $f \leq (g \otimes x)$, *where* $t \to x(t) \in \mathbb{R}$ *is the unknown function.*
3. $(f \ominus g) \ominus h = f \ominus (g \otimes h)$

Proof: We give the proof of the last item, which to our knowledge is unpublished. The other proofs are similar.

For a fixed value of t, let $A = [(f \ominus g) \ominus h](t)$ and $B = [f \ominus (g \otimes h)](t)$. We show first that $A \geq B$. For all $s \geq 0$ and all $0 \leq u \leq s$ we have

$$A \geq \gamma(t+u) - h(u)$$

where $\gamma = f \ominus g$. Similarly, by definition of $\ominus$:

$$\gamma(t+u) \geq f(t+u+s-u) - g(s-u)$$

Putting the two formulas together gives

$$g(s-u) + h(u) \geq f(t+s) - A$$

Since this is true for all $0 \leq u \leq s$, we have

$$(g \otimes h)(s) \geq f(t+s) - A$$

The above is true for all s, which shows that $B \leq A$.

Conversely, for all $\epsilon > 0$ there is some v such that

$$A \leq \gamma(t+v) - h(v) + \epsilon$$

There is also some u such that

$$\gamma(t+v) \leq f(t+v+u) - g(u) + \epsilon$$

Thus

$$A \leq f(t+v+u) - g(u) - h(v) + 2\epsilon \leq f(t+v+u) - (g \otimes h)(u+v) + 2\epsilon$$

Thus $A \leq B + 2\epsilon$ for all $\epsilon > 0$, thus $A \leq B$. $\square$

2.2 An Application for both Problems 1 and 2

We can apply the properties of deconvolution to the following theorem; it is the the central result for both Problems 1 and 2. It shows deconvolution as a smoothing operator.

Theorem 2.2 *Consider some function $S \in \mathcal{F}$, and a wide-sense increasing function σ. Assume that σ is sub-additive and $\sigma(t) = 0$ for $t \leq 0$. Then, among all real valued functions $t \to y(t)$ satisfying*

$$\begin{cases} y(t) \geq S(t) \\ y \text{ is } \sigma\text{-smooth} \end{cases} \tag{1}$$

there exists one function y^* which lower bounds all others. This function is given by $y^* = S \ominus \sigma$.

The theorem says that (i) $S \ominus \sigma$ satisfies the constraints in (1) and (ii) for any other function y satisfying the constraints, we have $y(t) \geq y^*(t)$ for all $t \in \mathbb{R}$.

Proof: We start by showing that $y^* = S \ominus \sigma$ satisfies the constraints in (1). Firstly, it follows from $\sigma(0) = 0$ that $y^* \geq S$. Secondly, we have from the third item in Theorem 2.1

$$y^* \ominus \sigma = (S \ominus \sigma) \ominus \sigma = S \ominus (\sigma \otimes \sigma)$$

now $\sigma \otimes \sigma = \sigma$ because $\sigma(0) = 0$ and σ is sub-additive. Thus $y^* \ominus \sigma = y^*$. It follows trivially that

$$y^* \ominus \sigma \leq y^*$$

which means that y^* is σ-smooth.

We procede now with showing the minimality of y^*. Let y be some solution to the constraints in (1). The second constraint can be rewritten as

$$y \ominus \sigma \leq y$$

Now $S \leq y$ from the first constraint; it follows from the last two formulas that $S \ominus \sigma \leq y \ominus \sigma \leq y$ thus $y^* \leq y$. □

The $S \to S \ominus \sigma$ transformation The previous theorem introduces the transformation $S \to S \ominus \sigma$. There is some similarity with the transformation associated with an optimal shaper. Indeed, for a shaper with service curve σ (with σ sub-additive and $\sigma(0) = 0$), the output is equal to $S \otimes \sigma$ [3, 1], if S is the input. The transformation $S \to S \ominus \sigma$ is also a smoothing operation, and like the other one, it is idempotent, namely, $(S \ominus \sigma) \ominus \sigma = S \ominus \sigma$. However, it can be shown using the construction in the following section that, unlike the other one, it is anti-causal, namely $(S \ominus \sigma)(t)$ depends only on $S(s)$ for $s \geq t$.

2.3 Computational Representation of Deconvolution

We will use the deconvolution operator in the next sections to show the existence of, and define, optimal prefetching strategies. The definition of deconvolution does not lend itself well to implementation, since it requires computing a supremum over the whole function, for every value of time. This is similar to the problem of computing the output $S \otimes \sigma$ of an optimal shaper, since computing the min-plus convolution is in general also complex. However, there are many cases where this can be simplified [3], in particular if σ is the minimum of a finite number of affine functions (for example, $\sigma(t) = \min(Pt + B, rt + m)$), as is defined for VBR traffic with the Internet integrated services or with ATM [5]). In such cases, the output of a shaper can be realized by implementing a multiple leaky bucket, which is considerably simpler from an implementation point of view; indeed, if time is discrete, computing the output at time $t + 1$ for a shaper defined by n leaky buckets requires the knowledge of only the n bucket levels, not the complete history $S(s)_{\{s \leq t\}}$.

We provide now a computational representation of deconvolution which lends itself to easy implementations in the case where σ is the minimum of a finite number of affine functions. The idea is use a time inversion; this transforms the deconvolution into convolution.

First we introduce some notation. We consider $\mathcal{F}_0$ defined as the subset of functions in $\mathcal{F}$ with a finite lifetime, namely:

Definition 2.2 (Set of functions with a finite lifetime)

$$\mathcal{F}_0 = \{S \in \mathcal{F} \ : \ \textit{there exists a finite } T \textit{ such that } S(t) = S(T) \textit{ for } t \geq T\}$$

Considering functions in $\mathcal{F}_0$ instead of $\mathcal{F}$ is not a restriction in practice; however, it is a convenient restriction for the rest of this section.

For a function S in $\mathcal{F}$, we use the notation $S(+\infty)$ as a shorthand for $sup_{t \in \mathbb{R}} S(t) = \lim_{t \to +\infty} S(t)$.

The following lemma will be also be used.

Lemma 2.1 *For any wide sense increasing function σ such that $\sigma(t) = 0$ for $t \leq 0$ and $\lim_{t \to +\infty} \sigma(t) = +\infty$, we have, for all $S \in \mathcal{F}$:*

1. *If $S \in \mathcal{F}_0$ then $S \otimes \sigma \in \mathcal{F}_0$*
2. *$(S \otimes \sigma)(+\infty) = S(+\infty)$*

Proof: Define $L = S(+\infty)$ and call T a number such that $S(t) = L$ for $t \geq T$.

The assumption that $\sigma(0) = 0$ implies that $S \otimes \sigma \leq S$. Thus

$$(S \otimes \sigma)(t) \leq L \text{ for } t \geq T \tag{2}$$

Now since $\lim_{t \to +\infty} \sigma(t) = +\infty$, there exists some $T_1 > T$ such that $\sigma(t) \geq L$ for all $t > T_1$. Now let $t > 2T_1$ and consider that

$$(S \otimes \sigma)(t) = \inf_{u \geq 0} \{\sigma(u) + S(t - u)\}$$

If $u > T_1$, then $\sigma(u) \geq L$. Otherwise, $u \leq T_1$ thus $t - u \geq t - T_1 > T_1$ thus $S(t-u) \geq L$. Thus in all cases $\sigma(u) + S(t-u) \geq L$. Thus we have shown that

$$(S \otimes \sigma)(t) \geq L \text{ for } t > 2T_1 \tag{3}$$

Combinining (2) and (3) shows the theorem. □

Note that the second item in the lemma is true even if S is not in $\mathcal{F}_0$.

Definition 2.3 (Time Inversion) *For a fixed $T \in [0, +\infty[$, the inversion operator $\mathcal{I}_T$ is defined on $\mathcal{F}_0$ by:*

$$\mathcal{I}_T(S)(t) = S(+\infty) - S(T-t)$$

This operation is a time inversion. Graphically, it can be obtained by a rotation of 180° around the point $(\frac{T}{2}, \frac{S(+\infty)}{2})$.

It is simple to check that $\mathcal{I}_T(S)$ is in $\mathcal{F}$ (because S has a finite lifetime) and that $\mathcal{I}_T(S)$ is in $\mathcal{F}_0$ (because by definition of $\mathcal{F}$, S is zero for very negative values).

Proposition 2.1 (Properties of time inversion)

1. *(Symmetry): for all $S \in \mathcal{F}_0$, we have : $\mathcal{I}_T(\mathcal{I}_T(S)) = S$*
2. *(Preservation of Total Value): $\mathcal{I}_T(S)(+\infty) = S(+\infty)$*
3. *(Preservation of Smoothness): For any fixed σ and T, $S \in \mathcal{F}_0$ is σ-smooth if and only if $\mathcal{I}_T(S)$ is σ-smooth*

Proof: The first and second items are straightforward. We now prove the third one.

Consider some $S \in \mathcal{F}_0$ and call $\widehat{S} = \mathcal{I}_T(S)$. Assume that S is σ-smooth. This means that $S(t) - S(s) \leq \sigma(t-s)$ for all $s \leq t$. We have

$$\widehat{S}(t) - \widehat{S}(s) = S(+\infty) - S(T-t) - S(+\infty) + S(T-s)) = S(T-s) - S(T-t) \leq \sigma(t-s)$$

Conversely, if $\widehat{S}$ is σ-smooth, then $S = \mathcal{I}_T(\widehat{S})$ and apply the previous statement to $\widehat{S}$. □

Theorem 2.3 (Representation of Deconvolution by Time Inversion) *Let $S \in \mathcal{F}_0$ be a function with finite lifetime, and let T be such that $S(T) = S(+\infty)$. Let σ be a wide-sense increasing function, with $\sigma(t) = 0$ for $t \leq 0$ and $\lim_{t \to +\infty} \sigma(t) = +\infty$. Then*

$$S \ominus \sigma = \mathcal{I}_T(\mathcal{I}_T(S) \otimes \sigma) \tag{4}$$

The theorem says that $S \ominus \sigma$ can be computed by first inverting time, then smoothing as with an optimal shaper, then inverting time again. Figure 3 shows a graphical illustration. It is easy to understand now why deconvolution is anticausal. The assumption that $\lim_{t \to +\infty} \sigma(t) = +\infty$ means that the smoothing does not put a limit on the total number of bits that are output.

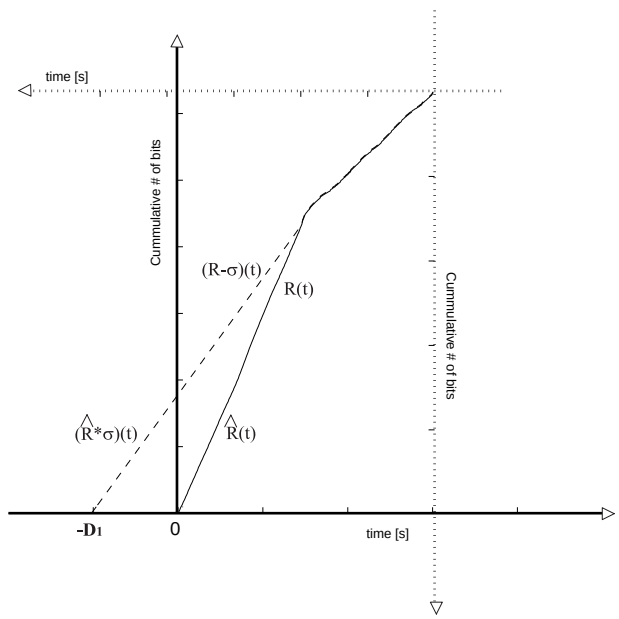


Figure 3: Graphical illustration of the construction of deconvolution.

Proof: The proof consists in computing the right handside in Equation (4). Call $\widehat{S} = \mathcal{I}_T(S)$. We have, by definition of the inversion

$$\mathcal{I}_T(\mathcal{I}_T(S) \otimes \sigma) = \mathcal{I}_T(\widehat{S} \otimes \sigma) = (\widehat{S} \otimes \sigma)(+\infty) - (\widehat{S} \otimes \sigma)(T - t)$$

Now from Lemma 2.1 and Proposition 2.1, item 2 :

$$(\widehat{S} \otimes \sigma)(+\infty) = \widehat{S}(+\infty) = S(+\infty)$$

Thus, the right-handside in Equation (4) is equal to

$$S(+\infty) - (\widehat{S} \otimes \sigma)(T - t) = S(+\infty) - \inf_{u \geq 0}\{\widehat{S}(T - t - u) + \sigma(u)\}$$

Again by definition of the inversion, it is equal to

$$S(+\infty) - \inf_{u \geq 0}\{S(+\infty) - S(t + u) + \sigma(u)\} = \sup_{u \geq 0}\{S(t + u) - \sigma(u)\}$$

□

3 Solutions to Problems 1 and 2

We now apply the results of the previous section to Problems 1 and 2

3.1 Solution to Problem 1

We now define a function which will appear to be the optimal solution to Problem 1, in a sense which will be explained later. First, it is useful to introduce the definition of horizontal deviation between two wide sense increasing functions:

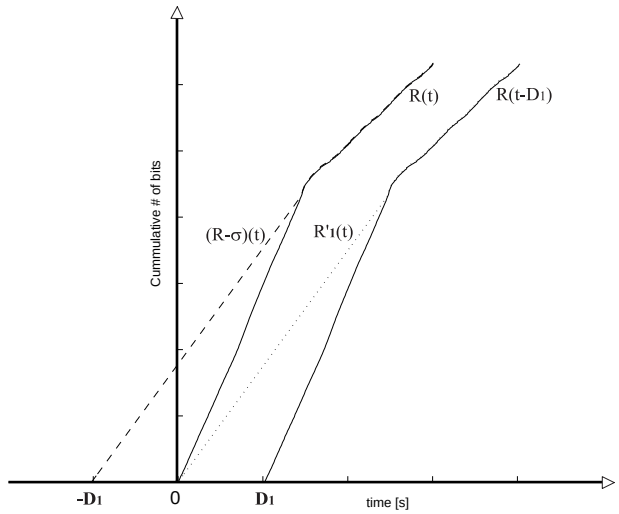


Figure 4: Definition of the Optimal Solution R_1'

Definition 3.1 (Horizontal Deviation h [1]) *For two wide-sense increasing functions α and β, define the horizontal deviation $h(\alpha, \beta)$ by*

$$h(\alpha, \beta) = \sup_{s \geq 0} \left(\inf \{ T : T \geq 0 \text{ and } \alpha(s) \leq \beta(s+T) \} \right) \tag{5}$$

The horizontal deviation has a simple intuitive meaning: if R is the input function (cumulative arrival function) into some arbitrary system, and R^* is the output function, then $h(R, R^*)$ is the maximum virtual delay through the system. The virtual delay is defined as the delay that would experience a bit of information if the system would be first in, first out.

Definition 3.2 (Definition of R_1') *Consider a given input function $R \in \mathcal{F}_0$ and a smoothing curve σ. Define R_1' by*

$$R_1'(t) = (R \ominus \sigma)(t - D_1)$$

with $D_1 = h(R, \sigma)$.

Figure 4 illustrates the definition. Note that D_1 is also given by

$$-D_1 = \sup\{t : (R \ominus \sigma)(t) \leq 0\} \tag{6}$$

To see why this is true, simply consider the set of values T such that $(R \ominus \sigma)(-T) \leq 0$.

For a smoothing solution (R', D), we call $X_{(R',D)}$ the buffer requirement at the receiver. We have

$$X_{(R',D)} = \sup_{t \geq 0} \{ R'(t) - R(t - D) \}$$

We now come to our main result for Problem 1.

Theorem 3.1 (Optimality of R_1') *Assume that $\sigma(t) = 0$ for $t \leq 0$ and σ is sub-additive. Then*

1. *(R_1', D_1) is a smoothing solution for Problem 1.*
2. *Conversely, for any other smoothing solution (R', D) for Problem 1 we have : $D \geq D_1$ and $X_{(R',D)} \geq X_{(R_1',D_1)}$.*

The theorem says that R_1' achieves the best buffer and playback delays over all possible solutions to Problem 1. It also implies that D_1 is the smallest playback delay that can be achieved. The theorem also says that there exists a single solution which achieves both bounds.

Proof: It follows from the definition of D_1 that $R_1'(t) = 0$ for $t < 0$. From Theorem 2.2, $R \ominus \sigma$ is σ-smooth and thus so is R_1'. Also, $R \ominus \sigma \geq R$ and thus $R_1'(t) \geq R(t - D_1)$. This shows that R_1' is a smoothing solution for Problem 1.

Conversely, let (R', D) be another solution. We have $R'(t) \geq R(t - D)$ and thus

$$R'(t + D) \geq R(t)$$

It follows from Theorem 2.2 that

$$R'(t + D) \geq (R \ominus \sigma)(t) \tag{7}$$

Since $R'(t) = 0$ for all $t < 0$ it follows that $(R \ominus \sigma)(u) = 0$ for all $u < -D$, by the definition of D_1, it comes that $D_1 \leq D$.

Now we can rewrite Equation (7) as

$$R'(t) \geq R_1'(t - D + D_1) \tag{8}$$

This shows that $X_{(R',D)} \geq X_{(R_1',D_1)}$. □

The proof of the theorem provides a stronger statement in Equation (8). It says that if we time-shift the optimal solution R_1' so that it has the same playback delay as another solution R', then R_1' is, at every time instant, no earlier than R'. In other words, the optimality is not only for the playback delay and the buffer requirement, it is all along the solution (Figure 5).

The optimal smoothing solution can be computed in practice using the representation of $R \ominus \sigma$ given in Theorem 2.3. First, R is inverted in time, yielding function $\widehat{R}$. This requires storing $R(t)$ in a complete array; then $\widehat{R} \otimes \sigma$ is computed, usually by simulating the operation of one or two leaky bucket controller. Then the result is time-inverted again.

3.2 Solution to Problem 2

The application to Problem 2 is easy to understand if we analyze how we could map the results in Section 2 to Problem 1. The idea is to consider the problem shifted in time, with time origin at the end of the playback delay. We first give the line of reasoning that leads to the existence and the value of an optimal solution.

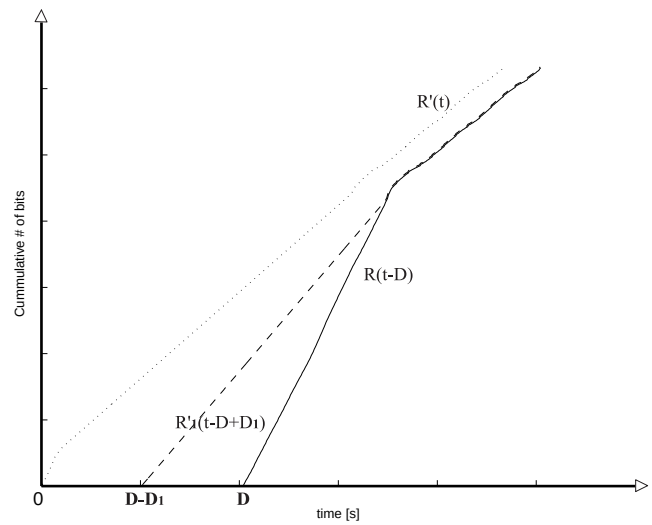


Figure 5: Comparison of a solution R' with the optimal solution R'_1.

Let us apply this reasoning to Problem 2; consider a smoothing solution (R', D) and take $t = 0$ at the end of the playback delay D. Call S the time-shifted version of R', namely, $S(t) = R'(t - D)$. We must have

$$(S \otimes \beta)(t) \geq R(t)$$

by definition of Problem 2 and because the convolution by β is time-invariant. From Theorem 2.1, item 1, it follows that

$$\begin{cases} S \geq R \ominus \beta \\ S \text{ is } \sigma\text{-smooth} \end{cases}$$

thus, from Theorem 2.2 applied to $R \ominus \beta$, we must have

$$S \geq (R \ominus \beta) \ominus \sigma \tag{9}$$

Following the same reasoning as for Problem 1, the right-handside in Equation (9) should be the optimal solution to Problem 2, after an appropriate time-shift. We can now proceed formally as in the previous section and start by defining what will prove to be the optimal solution to Problem 2:

Definition 3.3 (Definition of R'_2) *Consider a given input function $R \in \mathcal{F}_0$ and a smoothing curve σ. Define R'_2 by*

$$R'_2(t) = ((R \ominus \beta) \ominus \sigma)(t - D_2) = (R \ominus (\beta \otimes \sigma)(t - D_2)$$

where D_2 is the horizontal deviation between R and $\beta \otimes \sigma$:

$$D_2 = h(R, \beta \otimes \sigma)$$

Note that the equivalence between the two definitions of R_2' comes from Theorem 2.1. Also note that, as with Problem 1, D_2 can be defined by

$$-D_2 = \sup\{t : (R \ominus (\beta \otimes \sigma))(t) = 0\}$$

.

We can now come to our main result for Problem 2.

Theorem 3.2 (Optimality of R_2') *Assume that $\sigma(t) = 0$ for $t \leq 0$ and σ is sub-additive. Then*

1. *(R_2', D_2) is a smoothing solution for Problem 2.*
2. *Conversely, for any other smoothing solution (R', D) for Problem 2 we have : $D \geq D_2$ and $X_{(R',D)} \geq X_{(R_2',D_2)}$.*

Proof: The proof is similar to that of Theorem 3.1.

In practice, we also apply the representation given in Theorem 2.3 in order to compute R_2' from R. The first step is to time invert R, yielding function $\widehat{R}$.

The second step is to compute $\widehat{R} \otimes \beta \otimes \sigma$. However, we cannot simply use leaky buckets for computing $\widehat{R} \otimes \beta \otimes \sigma$, because, unlike σ, $\beta \otimes \sigma$ is not the minimum of n affine functions. However, a simple decomposition of $\beta \otimes \sigma$ brings the solution in the (common) case where $\beta(t) = \rho(t - L)^+$. Assume that σ is defined as a minimum of affine functions, for example $\sigma(t) = \min(Pt + M, rt + b)$ for $t > 0$, 0 otherwise. Then, using easy rules found for example in [1] we find

$$\beta \otimes \sigma = \delta_L \otimes \alpha$$

where

- δ_L is the impulse function defined by $\delta_L(t) = 0$ if $0 \leq t \leq L$ and $\delta_L(t) = +\infty$ if $t > L$
- $\alpha(t) = \min(\sigma(t), \rho t) = \min(Pt + M, rt + b, \rho t)$

Note that for any function S, $(S \otimes \delta_L)(t) = S(t - L)$. Thus $\widehat{R} \otimes \beta \otimes \sigma = (\widehat{R} \otimes \alpha) \otimes \delta_L$ and the second step consists in computing $\widehat{R} \otimes \alpha$, using $n + 1$ leaky bucket controllers if σ is defined with n leaky buckets; then shift in time by L. Then the third step is to apply time inversion as previously.

Comments on Theorem 3.2 Note that D_2 in the theorem is the smallest playback delay which can be achieved by smoothing, provided that the information about the network is reduced to its service curve β.

In particular, we compare the scenario of Problem 2 to the following (delay equalization): assume that at the receiver, we first introduce a buffer in order to equalize the variable delay imposed by the network with service curve β; then we use a playback buffer to compensate for fluctuations due to pre-fetching. In this new scenario, the pure playback delay is the D_1 of Problem 1, since now the composition of the network and the delay equalization buffer have a null variable delay. However, we need to dimension the equalization buffer for a delay $g = h(\sigma, \beta)$ which is the worst case delay

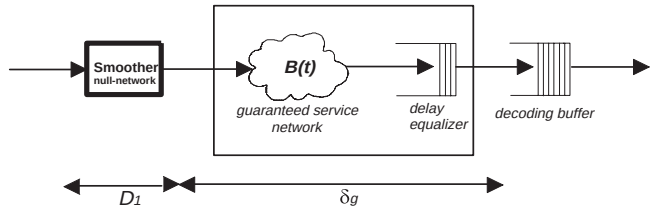


Figure 6: *Comparing global smoothing versus delay equalization at receiver*

through the original network. Call $D^* = D_1 + g$ the total delay that needs to be provisioned for that scenario. We know that $D_2 \leq D^*$. This scenario is illustrated by Figure 6.

The figure also illustrates that, in general, equality cannot be assumed. In other words, separate delay equalization and pre-fethcing are suboptimal.

4 Application to some Corollary Problems

From the fundamental results obtained in the previous section, we derive the solutions to some corollary problems. In Section 3, we have demonstrated that the optimal smoother based on the deconvolution operator required the minimal playback delay and decoding buffer size under a given smoothing curve σ (Problem 1). We have also shown that this result still holds when a non-null network is considered (Problem 2). In this section, we provide deterministic bounds on the smoothing curve $\sigma(t)$ given either a decoding buffer size B or the maximum admissible playback delay D, in the case of a network characterized by a service curve $\beta(t)$ (corollaries to Problem 2). It is to be noted that $\beta(t)$ is intimately related to the desired smoothing curve $\sigma(t)$. Related numerical applications will be given in Section 5.

Corollary 4.1 (Delay Constraint) *Consider a given input function $R \in \mathcal{F}_0$ and a playback delay D. There exists a solution (R', D) to Problem 2 if and only if, for all t,*

$$(\sigma \otimes \beta)(t) \geq R(t - D) \tag{10}$$

The corollary says that, if the maximum admissible playback delay is D, then it is necessary that $(\sigma \otimes \beta)(t)$ is lower-bounded by $R(t - D)$, and conversely; if Eq. 10 is true, then it is possible to attain the target playback delay D.

Proof: We know that, among all possible solutions to Problem 2, the playback delay, D_2, induced by optimal smoothing is the smallest playback delay. Also, the playback delay is strictly non-increasing when $(\sigma \otimes \beta)(t)$ increases. Thus, for a given playback delay D, all other possible solutions to Problem 2 lead to a greater $(\sigma \otimes \beta)(t)$ curve.

The playback delay D_2 is defined as

$$(R \ominus (\sigma \otimes \beta))(-D_2) \leq 0$$

It follows from definition 2.1 that

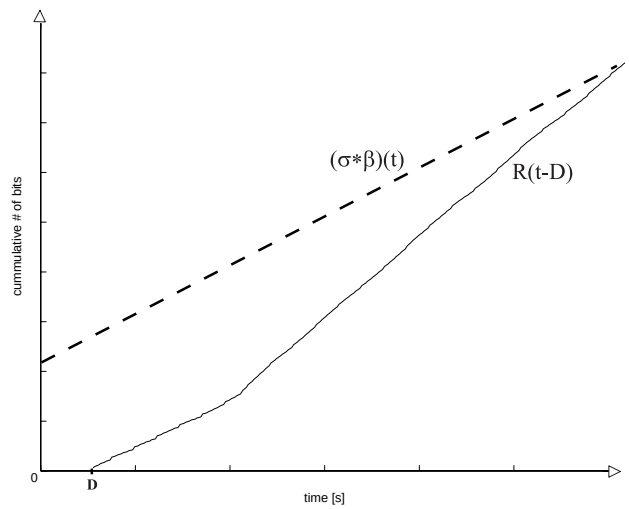


Figure 7: Representation of $R(t-D)$ and an admissible $(\sigma \otimes \beta)(t)$.

$$sup_{u \geq 0}\{R(-D_2 + u) - (\sigma \otimes \beta)(u)\} \leq 0$$

Since this is true for all $u \geq 0$, we have

$$(\sigma \otimes \beta)(t) \geq (R \otimes \delta_{D_2})(t)$$

and, since D_2 is the smallest playback delay, we have, for all smoothing solutions,

$$(\sigma \otimes \beta)(t) \geq (R \otimes \delta_D)(t)$$

□

The corollary is illustrated by figure 7.

Corollary 4.2 (Buffer Constraint) *Consider a given input function $R \in \mathcal{F}_0$ and a decoding buffer of size B. There exists a solution (R', B) to Problem 2 if and only if*

$$(\sigma \otimes \beta)(t) \geq (R \ominus R)(t) - B$$

Proof: The smallest decoding buffer size B_2 induced by optimal smoothing is defined as:

$$(R \ominus (\sigma \otimes \beta))(t) - R(t) \leq B_2$$

Proof: We know that, among all possible solutions to Problem 2, the required decoding buffer size, B_2, induced by optimal smoothing is the smallest buffer size. Also, the buffer size is strictly non-increasing when $(\sigma \otimes \beta)(t)$ increases. Thus, for a given decoding buffer size B, all other possible solutions to Problem 2 lead to a greater $(\sigma \otimes \beta)(t)$ curve.

In order to avoid the decoding buffer to overflow, $(\sigma \otimes \beta)$ must, at every time instant, verify

$$(R \ominus (\sigma \otimes \beta))(t) - R(t) \leq B_2$$

It follows from the definition of the convolution and deconvolution operators that

$$(\sigma \otimes \beta)(t) \geq (R \ominus R)(t) - B_2$$

and, since B_2 is the smallest decoding buffer size, we have for all smoothing solutions

$$(\sigma \otimes \beta)(t) \geq (R \ominus R)(t) - B$$

□

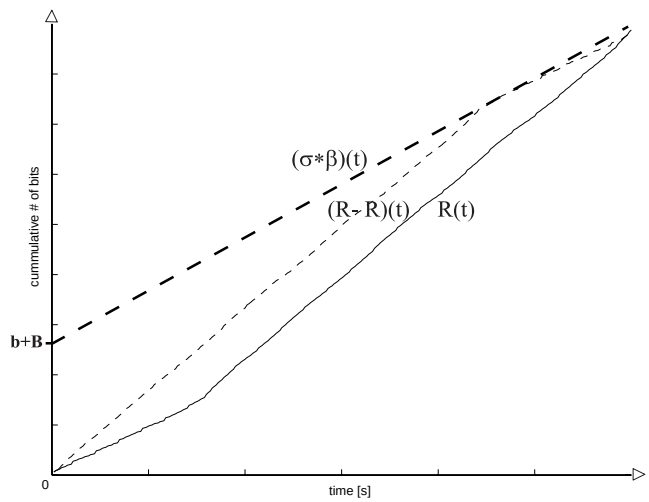


Figure 8: Representation of $R(t)$, $(R \ominus R)(t)$ and an admissible $(\sigma \otimes \beta)(t)$.

The corollary is illustrated by figure 8.

From Corollaries 4.1 and 4.2, we see that the condition on $(\sigma \otimes \beta)(t)$ to induce a zero playback delay does not necessarily imply a null decoding buffer size. Indeed, since $R(0) = 0$, $(R \ominus R)(t) \geq R(t)$. Therefore, $\sup_{s \geq 0}\{(R \ominus R)(s) - R(s)\}$ corresponds to the required buffer size under the constraint $(\sigma \otimes \beta)(t) = R(t)$. However, $(\sigma \otimes \beta)(t) \geq (R \ominus R)(t)$ implies $D = 0$.

In the special case of Problem 1 (null network), we can have some more explicit results, as shown below.

Case Study: null network ($\beta(t) = \delta_0(t)$): Assume the input flow $R(t)$ and the playback delay D are given. We further assume that the service curve $\beta(t)$ is of the form $\delta_0(t)$ (null network). This implies that $(\sigma \otimes \beta)(t) = \sigma(t)$. We consider the practical case for which the output of the optimal smoother is constrained by two leaky buckets $\sigma_1(t) = M + Pt$ and $\sigma_2(t) = b + rt$. Therefore, $\sigma(t) = \min\{\sigma_1(t), \sigma_2(t)\}$. We propose to find both the minimal peak rate p_{min} and the minimal sustainable rate r_{min} that strictly insure the playback delay D.

¿From Corollary 4.1, it follows:

$$\sigma_i(t) \geq (R \otimes \delta_D)(t) \text{ for } i = \{1, 2\}$$

Thus, the peak rate P and the couple (sustainable rate r, bucket size b) must, at every time instant, verify:

$$\begin{cases} M + Pt & \geq & R(t - D) \\ b + rt & \geq & R(t - D) \end{cases}$$

Since this is true for all $t \geq 0$, we derive:

$$\begin{cases} P & \geq & sup_{t\geq 0}\left\{\frac{R(t-D)-M}{t}\right\} \\ r & \geq & sup_{t\geq 0}\left\{\frac{R(t-D)-b}{t}\right\} \end{cases}$$

The sustainable rate r and the bucket size b are obviously related to each other. However, assume $b = K$ is given, both the minimal peak rate and the minimal sustainable rate are given by

$$\begin{cases} P_{min} & = & sup_{t\geq 0}\left\{\frac{R(t-D)-M}{t}\right\} \\ r_{min} & = & sup_{t\geq 0}\left\{\frac{R(t-D)-K}{t}\right\} \end{cases}$$

We have assumed so far that there was no relationship between the choice of arrival curve σ and the service curve β. In the rest of this section, we assume to have some knowledge about the relation between the smoothing curve σ and the service curve β.

Case Study: Guaranteed Service of the IETF : The Internet guaranteed service assumes that every node offers a service of the form $\beta(t) = \rho(t-L)^+$ for some latency L and rate ρ. The latency parameter L depends on the rate ρ according to $L = \frac{C_0}{\rho} + E_0$. Using the IETF jargon, ρ is contained in the list of R-SPEC parameters. The constants C_0 and E_0 depends on the route taken by the flow throughout the network. They are both determined during the advertisment phase (in the PATH messages, assuming routing does not change with the traffic parameters). The rate ρ, provided by the network, is not know a priori by a source, it is discovered during the advertisement phase using PATH messages, and accumulated in the AdSpec. With the guaranteed service, a source advertizes an arrival curve σ of the form $\sigma(t) = min(M + Pt, b + rt)$, and destinations choose a target admissible network delay T_0. The choice of a specific service curve $\beta(t) = \rho(t-L)^+$ (or equivalently, of a rate parameter ρ) is done during the reservation phase and cannot be know exactly in advance.

We consider the following problem. Assume that an input flow and a fixed maximum playback delay D are given. Assume that source and destination are able to agree on what reservation should be done, by some out-of-band mechanism. The question is: which choices of $\sigma(t) = min(M+Pt, b+rt)$ and of T_0 are admissible in order to guarantee that the reservation that will subsequently be performed ensures a playback delay not exceeding D.

For a given (but unknown) (σ, T_0), the reservation rate is not known. Let us call by $\mathcal{D}(\sigma, T_0)$, the set of rates that may be allocated by the network. A rate ρ is in $\mathcal{D}(\sigma, T_0)$ if and only if

$$\begin{cases} h(\sigma, \beta(\rho)) & \leq T_0 & \text{(a)} \\ \rho & \geq r & \text{(b)} \end{cases} \tag{11}$$

where $\beta(\rho)(t) = \rho(t-L)^+$ and $\sigma(t) = min(M + Pt, b + rt)$.

Call $\rho_1(\sigma, T_0)$ the solution to Equation 11(a) and define $\rho_{min}(\sigma, T_0) = \max(\rho_1(\sigma, T_0), r)$. It is easy to see that

$$\mathcal{D}(\sigma, T_0) = [\rho_{min}(\sigma, T_0), +\infty[$$

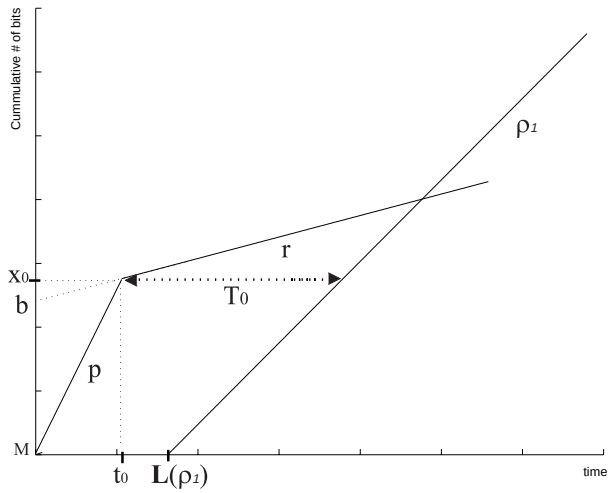


Figure 9: *T-SPEC* (P, M, r, b, T_0) *and R-SPEC* $(\rho_1, L(\rho_1))$.

We now determine the solution ρ_1 to Equation 11(a) in terms of the $\sigma(t)$ parameters (P, M, r, b). A graphical illustration is given by Figure 9.

We easily derive $\rho_1 = \frac{rt_0+b}{t_0+T_0-L(\rho_1)}$ where $L(\rho_1)$ is the delay parameter contained in the R-SPEC and is equal to $\frac{C_0}{\rho_1} + E_0$.

Finally, we obtain

$$\rho_1 = \frac{rt_0 + b + C_0}{t_0 + T_0 - E_0} \tag{12}$$

with $t_0 = \frac{b-M}{p-r}$.

Note that that all variables $(r, t_0, b, C_0, T_0, E_0)$ have non-negative values. Thus, Equation 12 has an admissible solution if and only if $T_0 \geq E_0 - \frac{b-M}{p-r}$. Moreover, the condition $\rho_1(\sigma, T_0) \geq r$ is equivalent to $r \leq \frac{b+C_0}{T_0-E_0}$.

Now we proceed with analyzing the conditions on (σ, T_0). Every rate $\rho \in \mathcal{D}(\sigma, T_0)$ corresponds to a rate that the network may potentially reserve (return in its R-SPEC). Thus, we require that every rate $\rho \in \mathcal{D}(\sigma, T_0)$ must necessarily verify the constraint on the playback delay:

$$h(R, \sigma \otimes \beta(\rho)) \leq D \tag{13}$$

One must notice that $h(R, \sigma \otimes \beta(\rho))$ decreases when ρ increases. Indeed, for a given time t, $(\sigma \otimes \beta(\rho))(t)$ increases with the rate ρ. Therefore, we can conclude that it is necessary and sufficient that

$$h(R, \sigma \otimes \beta(\rho_{min}(\sigma, T_0))) \leq D \tag{14}$$

From Corollary 4.1, Equation 14 may be rewritten as:

$$(\sigma \otimes \beta(\rho_{min}))(t) \geq R(t-D) \quad \text{for} \quad i = \{1,2\} \tag{15}$$

where $(\sigma \otimes \beta(\rho_{min}))(t)$ is the shifted version of $(\sigma \otimes \alpha)(t) = \min\{M + Pt, b + rt, \rho_{min}t\}$ by the amount of time $L(\rho_{min})$.

Therefore, we obtain that the following must be true for all t:

$$\begin{cases} b + rt & \geq & R(t - D + L(\rho_{min})) & \text{(a)} \\ \rho_{min}t & \geq & R(t - D + L(\rho_{min})) & \text{(b)} \end{cases} \tag{16}$$

with $L(\rho_{min}) = \frac{C_0}{\rho_{min}} + E_0$.

First, we analyze Equation 16(b).

Let $u = t - D + \frac{C_0}{\rho_{min}} + E_0$; Equation 16(b) may be rewritten as:

$$\rho_{min}(D - E_0) - C_0 \geq \sup_{u \geq 0}\{R(u) - \rho_{min}u\}$$

which is equivalent to

$$\rho_{min}(D - E_0) - C_0 \geq -\inf_u\{\rho_{min}u - R(u)\}$$

Call $\overline{\overline{R}}$ the concave conjugate of R, namely $\overline{\overline{R}}\,(\rho_{min}) = \inf_u\{\rho_{min}u - R(u)\}$. Equation 16(b) is equivalent to

$$\overline{\overline{R}}\,(\rho_{min}) + \rho_{min}(D - E_0) - C_0 \geq 0 \tag{17}$$

From the concavity of $\overline{\overline{R}}$ we can conclude that there exists one ρ_2 (independent of (σ, T_0)) such that Equation 16(b) is equivalent to

$$\rho_{min} \geq \rho_2 \tag{18}$$

The value of ρ_2 is the only positive solution of

$$\overline{\overline{R}}\,(\rho_2) + \rho_2(D - E_0) - C_0 = 0 \tag{19}$$

The graphical solution to Equation 17 is represented on Figure 10. It illustrates that there exists a solution to Equation 17 if and only if the following conditions are met:

$$\begin{cases} D \geq E_0 \quad \text{and,} \\ C_0 = 0 \quad \text{if} \quad D = E_0 \end{cases}$$

Similarly, Equation 16(a) is equivalent to

$$\overline{\overline{R}}\,(r) + r(D - E_0) + b - r\frac{C_0}{\rho_{min}} \geq 0 \tag{20}$$

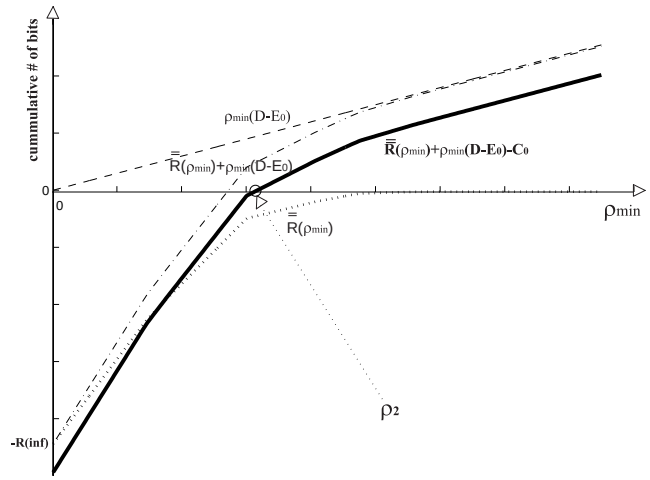


Figure 10: *Graphical solution to Equation 17: given the flow $R(t)$, the parameters C_0, E_0 accumulated in the PATH messages and the maximum admissible playback delay D, the figure shows how to compute ρ_2.*

It is not clear whether Equation (20) can be simplified. In summary so far, the conditions on (σ, T_0) are that both Equations (18) and (20) are satisfied.

Now, if it turns out from the values of (σ, T_0), that $\rho_1 \leq r$, then $\rho_{min} = r$ and Equation (20) is redundant. The only condition is thus $r \leq \rho_2$ in that case.

In summary, the procedure to test whether a choice of parameters (σ, T_0) is compatible with a playback delay is as follows. Assume that $T_0 < D$, that $D > E_0$ and that $T_0 \geq E_0 - \frac{b-M}{p-r}$, otherwise there is no solution. Then:

- Compute $\rho_1(\sigma, T_0)$ using Equation (12).
- If $r \geq \frac{b+C_0}{T_0-E_0}$ then check whether $r \geq \rho_2$. If it is true, then (σ, T_0) is admissible, otherwise not.
- Else (namely if $r < \frac{b+C_0}{T_0-E_0}$, check $\rho_1(\sigma, T_0) \geq \rho_2$ and Equation (20). If both are true, then (σ, T_0) is admissible, otherwise not.

A numerical illustration is given in Section 5.3.

5 Numerical Examples

In this section, we present some experimental results. We consider the scenario illustrated by figure 1, where the codec is MPEG-2 [1] and the network offers the guaranteed service of the IETF (for example using a resource reservation protocol such as RSVP).

[1] MPEG-2 is the most appropriate compression standard for video broadcast applications.

The RTP/UDP/IP protocols stack has now been widely accepted for the delivery of delay- and loss-sensitive services over packet networks. In such a scenario, every single packet contains 40 bytes of pure header information (assuming no header compression technique is used). RFC 2250 defines the packet format for MPEG-1 and MPEG-2 audio and video [7]. It specifies payload identifier and encapsulation schemes for the different packet formats (i.e., transport or program streams). For MPEG-2 transport streams, the RTP payload must contain an integral number of transport stream (TS) packets. According to the MPEG-2 standard, a TS packet is a 188-byte length packet, which encapsulates both video and system information.

In our simulations, we use a 300 frame-long sequence conforming to the ITU-R 601 format (720*576, 25 fps). The sequence is composed of 3 video scenes that differ in terms of spatial and temporal complexities. It has been encoded in an open-loop VBR (OL-VBR) mode, as interlaced video, with a structure of 11 images between each pair of I-pictures and 2 B-pictures between every reference picture. For this purpose, the widely accepted TM5 video encoder [8] has been utilized. Also, we consider the encapsulation of 2 MPEG-2 TSs per RTP packet. Therefore, every packet sent throughout the IP network contains $2 * 188 + 40 = 416$ bytes.

Figure 11 shows how the number of RTP packets vary with the frame number (or, equivalently, with time as each frame corresponds to a time-slot of 40 ms).

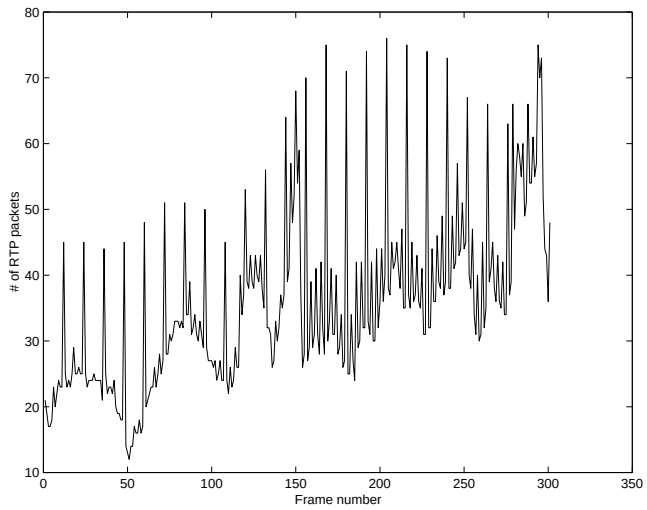


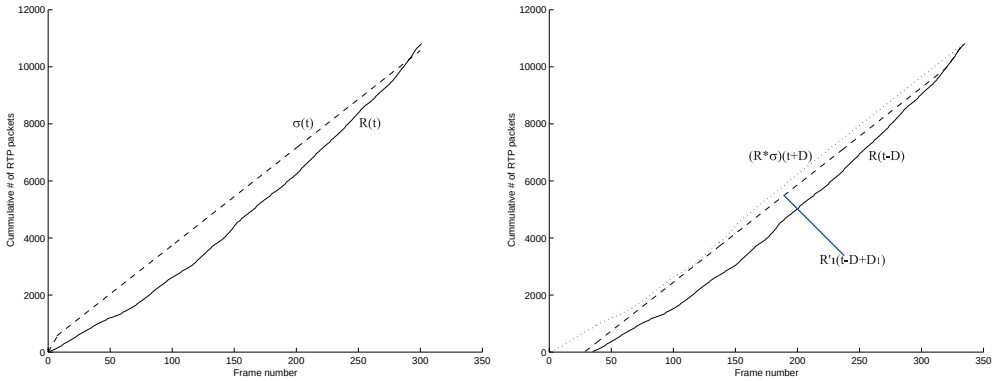
Figure 11: *MPEG-2 Trace: number of RTP packets per frame.*

5.1 Application to Problem 1

Figure 7(a) represents the cummulative RTP packets rate, $R(t)$, and the arrival curve, $\sigma(t)$. The arrival curve conforms to the IETF specification for integrated service [5]. The maximum packet size M is equal to 416 bytes. The peak and sustainable rates are, respectively, the peak and the average rates of the MPEG-2 stream ($P = 5.8$ Mbits/s and $r = 2.7$ Mbits/s). Finally, the bucket can absorb up to 332 MPEG-2 RTP packets (close to 1 Mbits).

Figure 7(b) shows the optimal solution to Problem (1), $R_1'(t)$. We find that the minimum playback delay is $D_1 = 0.24s$ (6 frames) and the minimum decoding buffer size is 3.35 Mbits. It further

illustrates that if we smooth with the optimal shaper $(R \otimes \sigma)(t)$ instead of the optimal smoother, then we do find a non-optimal result. Indeed, the minimum playback delay is now $D = 1.36s$ (34 frames) which corresponds to a minimum decoding buffer size of 4.26 Mbits. The delay D is defined such as the real-time constraint is met.



(a) (b)

Figure 12: *(a) Cummulative RTP packets rate, $R(t)$, and arrival curve, $\sigma(t)$. (b) $R(t)$, optimal smoothing, $R'_1(t)$, and optimal shaping, $(R \otimes \sigma)(t)$*

5.2 Application to Problem 2

We now consider the transmission of our MPEG-2 stream over an IP network characterized by a service curve $\beta(t) = \rho(t - L)^+$. Assume that the delay L and the rate ρ equals to, respectively, $1s$ (25 frames) and 3.4 Mbits/s (slightly greater than the average bit rate but lower than the peak bit rate).

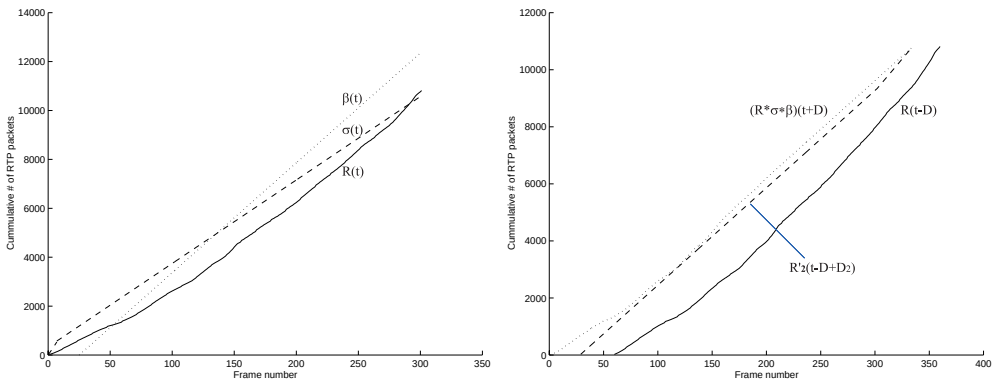
Figure 8(a) shows the cummulative rate, $R(t)$, the arrival curve, $\sigma(t)$ and the service curve offered by the network, $\beta(t)$.

The optimal solution to Problem 2, $R'_2(t)$, is represented on figure 8(b). The delay D_2 is of $1.24s$ (31 frames) due to the maximal delay the network may introduce. The decoding buffer must be at least of 5.92 Mbits. It is to be noted that $R'_2(t)$ is not the shifted version of $R'_1(t)$. Indeed, $\sigma(t)$ is not equal to $(\beta \otimes \sigma)(t)$. Figure 5.2(b) also shows that, once again, the optimal shaper $(R \otimes \sigma \otimes \beta)(t)$ leads to a sub-optimal solution. Indeed, the playback delay $D = 2.36s$ is greater than D_2.

5.3 Admissible (σ, T_0) under maximum playback delay D and IntServ network assumptions

We now give a numerical illustration of the last case study proposed in Section 4. We use the MPEG-2 trace presented hereabove.

Assume we know the input flow $R(t)$, the smoothing curve $\sigma(t) = \min(Pt + M, b + rt)$ with $(P, M, r, b) = (5, 3.328 \times 10^{-3}, 2.5, 1)$ expressed in MBits and, a target maximum network delay $T_0 = 1$s.



(a) (b)

Figure 13: *(a) Cummulative RTP packets rate, $R(t)$, arrival curve, $\sigma(t)$, and service curve, $\beta(t)$. (b) $R(t)$, optimal smoothing, $R'_2(t)$, and optimal shaping, $(R \otimes \sigma \otimes \beta)(t)$*

We determine whether or not the specified (σ, T_0) is an admissible solution under a maximum playback delay $D = 2.0$s and IntServ network assumptions.

We further assume to know the constant C_0 and E_0 respectively equal to 0.5 MBits and 0.2s.

We follow the test procedure mentioned in the case study. We first compute $\frac{b+C_0}{T_0-E_0} = 1.875$ Mbits/s. The specified sustainable rate $r = 2$ Mbits/s is greater then 1.875 Mbits/s. Therefore, the specified (σ, T_0) is admissible if $r \geq \rho_2$. Using Eq. 19, we find $\rho_2 = 0.28$ MBits/s. Therefore, (σ, T_0) is indeed an admissible solution.

6 Conclusion

We have shown that there exists an optimal smoothing which minimizes playback delay and receive buffer size. We have expressed it using the deconvolution operator of min-plus algebra. The existence of an optimal smoothing means that we can compute the minimum playback delay and buffer sizes that are required at a receiver, no matter what the scheduling strategy is at the sender. The minimum playback delay is the horizontal deviation between the functions R and $\sigma \otimes \beta$. We have also proposed practical applications such as finding the minimal T-SPEC that insures a desired maximum playback delay.

The results in this paper are also a contribution to the theory called network calculus. In addition to the optimal shaper based on min-plus convolution, described for example in [3, 1], we have defined another smoothing operator, based on deconvolution. We have given a representation of the deconvolution in terms of time inversion and convolution, and shown on some example how this can be used in practice to compute optimal smoothing.

The results in this paper provide absolute bounds. However their computation requires a knowledge of the complete sequence $R(t)$. The bounds could be used to evaluate sub-optimal smoothing strategies.

References

[1] J.-Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Transactions on Information Theory*, 44:1087–1096, August 1998.

[2] Bob Braden, David Clark, and Scott Shenker. Integrated services in the internet architecture: an overview, June 1994. RFC 1633, IETF.

[3] C.S. Chang. On deterministic traffic regulation and service guarantee: A systematic approach by filtering. In *Proc of Infocom 1997*, 1997.

[4] R. L. Cruz. Sced+: Efficient management of quality of service guarantees. In *IEEE Infocom'98, San Francisco*, March 1998.

[5] R. Guérin and V. Peris. Quality-of-service in packet networks - basic mechanisms and directions. *Computer Networks and ISDN, Special issue on multimedia communications over packet-based networks*, 1998.

[6] Le Boudec Jean-Yves. Network calculus made easy, 1996. Technical report EPFL-DI 96/218 `http://lrcwww.epfl.ch/PS_files/d4paper.ps`.

[7] D. Hoffman, G. Fernando, V. Goyal and R. Civanlar. RTP Payload Format for MPEG1/MPEG2 Video. IETF RFC-2250, January 1998.

[8] C. Fogg. mpeg2encode/mpeg2decode. By the *MPEG Software Simulation Group*, 1996.

[9] Chi-Yuan Hsu, Antonio Ortega and Amy R. Reibman. Joint Selection of Source and Channel Rate for VBR Video Transmission under ATM Policing Constraints. *IEEE Journal on Selected Areas in Communications*, 1997. accepted for publication.

[10] D. Reininger, et al. Variable bitrate mpeg video : Characteristics, modeling and multiplexing. In *Proceedings of the ITC-14*, pages 295–306. Elsevier, 1994.

[11] J. McManus and K. Ross. Video on Demand over ATM: Constant-Rate Transmission and Transport. In *IEEE INFOCOM*, Mar. 1996.

[12] J. Salehi, Z. Zhang, J. Kurose, and D. Towsley. Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements Through Optimal Smoothing. In *ACM SIGMETRICS*, May 1996.

[13] M. Hamdi and J. W. Roberts. QoS Guaranty for Shaped Bit Rate Video Connections in Broadband Networks. In *IEEE Computer Society Press*, pages 153–162, september 1995.

[14] T. V. Lakhsman, A. Ortega and A. R. Reibman. VBR video: Trade-offs and potentials. In *Proceedings of the IEEE*, Jul. 1997.

[15] W. Feng and J. Rexford. A Comparison of Bandwidth Smoothing Techniques for the Transmission of Prerecorded Compressed Video. In *IEEE INFOCOM*, 1997.

[16] W. Feng, F. Jahanian, and S. Sechrest. Optimal Buffering for the Delivery of Compressed Prerecorded Video. In *IASTED/ISMM Int'l Conference on Networks*, Montreal, Canada, Jan. 1995.

[17] Wei Ding and Bede Liu. Joint Encoder and Channel Rate Control of VBR Video over ATM Networks. In *SPIE Electronic Imaging - Digital Video Compression*, volume 2668, January 1996.