Network calculus applied to optimal multimedia smoothing

Patrick Thiran[†‡], Jean-Yves Le Boudec[‡] and Frédéric Worm[‡]
†Sprint ATL, 1 Adrian Court, Burlingame, CA 94010, USA
‡ICA-DSC (Dept. of Communication Systems)
EPFL: Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland
thiran,leboudec @ica.epfl.ch, fworm@eurecom.fr

Abstract—**We consider a scenario where multimedia data are sent over a network offering a guaranteed service. A smoothing device writes the stream into a transmitting device with limited input buffer; at destination, the decoder waits for an initial playback delay and reads the stream from the receiver buffer. We assume that some limited look-ahead is possible at the source, and that the playback buffer size is limited. First, we consider the case where the stream is delivered from the source directly to the destination buffer. We obtain closed-form expressions of the minimal required values in the case of a CBR smoothing Then we consider the case where the stream is first transmitted through a backbone network to an intermediate server, which relays the multimedia data to the final destination via an access network. We compute the requirements on playback delay, buffer sizes and amount of look-ahead.**

Keywords—**Network calculus, min-plus algebra, video smoothing, ATM, RSVP/IP.**

I. INTRODUCTION

One approach to stream video is to act on the quantization levels at the encoder output: this is called rate control, see e.g. [5]. Another approach is to smooth the video stream, using a smoother fed by the encoder, see e.g. [11], [14], [10]. This paper focuses on this second approach.

Let us start by describing the simpler problem where the video stream is directly delivered from the server to the client, through a single network, as shown on the top of Figure 1. At the sender side, a smoothing device reads the encoded video stream $R(t)$ and sends a stream $x(t)$ that must conform to a traffic envelope or *arrival curve* σ [3], [4], that is, such that over any time window of any length τ, the traffic sent to the network may not exceed $\sigma(\tau)$. For lack of space, we will only consider the simplest (and most important in practice) case of a constant rate r shaping curve: $\sigma(t) = rt$, for $t \geq 0$. In this paper, we will refer to this function as a *constant rate curve* and denote it by $\lambda_r(t)$: $\lambda_r(t) = rt$ for $t \geq 0$ and $\lambda_r(t) = 0$ for $t < 0$. The methodology extends however directly to more general smoothing curves.

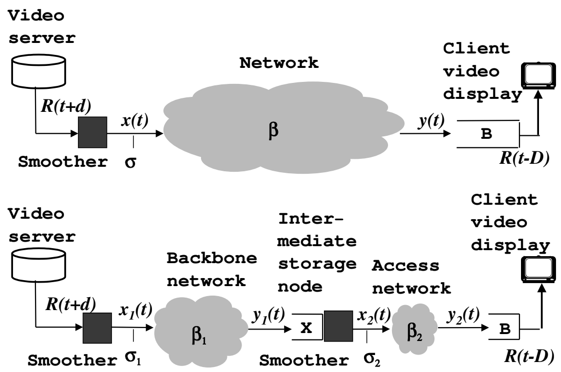

Fig. 1. The two smoothing scenarios over a single network (top) and over two networks with a local caching node (bottom).

We take the transmission start as origin of time: this implies that $x(t) = 0$ for $t \leq 0$. The video stream R will be played back at the receiver after D units of times, the *playback delay*: the output of the decoding buffer B must therefore be $R(t-D)$.

The network offers a guaranteed service to the flow x. If y denotes the output flow, it is not possible, in general, to express y as a function of x. However we assume that the service guarantee can be expressed by a service curve β [3], [4], meaning that the cumulative output at time t, $y(t)$, is at least equal to the cumulative input up to some time $s \leq t$, $x(s)$, augmented of $\beta(t-s)$ for the remaining time interval $(t-s)$. For example, the IETF assumes that RSVP routers offer a service curve of the form $\beta(t) = C[t-L]^+ = \max\{0, C(t-L)\}$. We will refer to this service curve as a *rate-latency curve* (C is the rate, L the latency), and denote it by $\beta_{L,C}(t)$. Another example is a network which is completely transparent to the flow (i.e. which does not incur any jitter to the flow nor rate limitation). We speak of a *null network*. It offers a service curve $\beta(t) = \delta_0(t)$, where δ_τ is defined for $\tau \geq 0$ as

$$\delta_\tau(t) = \begin{cases} \infty & \text{if} \quad t > \tau \\ 0 & \text{if} \quad t \leq \tau. \end{cases} \tag{1}$$

We assume in this paper that the encoding buffer size is large enough to contain the full data stream. This restriction allows us to simplify the mathematics for this paper, but our method can also cope with a finite size encoding buffer, the results are left for a forthcoming paper. On the other hand, the receiver (decoding) buffer, denoted by B, is a much more scarce resource.

As the stream is pre-recorded and stored in the video server, it allows the smoother to prefetch some of the data before schedule. We suppose that the smoother is able to prefetch data for up to d time units ahead. This *look-ahead delay* can take values ranging from zero (in the most restrictive case where no prefetching is possible) up to the length of the full stream. The sum of the look-ahead delay and playback delay is called the *total delay*, and is denoted by T: $T = D + d$.

These constraints are described more mathematically in Section II. In Section III, we apply a result of min-plus algebra to compute (i) the requirements on the playback delay D, on the look-ahead delay d, and on the client buffer size B guaranteeing a lossless transmission for given smoothing and service curves β and σ, and (ii) a scheduling strategy at the smoother that will achieve transmission for the above computed values. In the CBR case ($\sigma = \lambda_r$), we will give closed-form expressions for these quantities. We will also compute what is the typical amount of look-ahead delay d necessary to achieve this value on some real video traces in Section IV. At high rates, the playback delay is zero, but increases rapidly when the rate is too small. Between these two sets of rates, there is a small range of quasi-optimal rates which bring the playback delay to a very small value thanks to a small amount of look-ahead on the trace stored in the server.

We then consider in Section V the more complex setting where two networks separate the video server from the client: the first one is the backbone network, offering a service curve β_1 to the flow, and the second one is a local access network, offering a service curve β_2 to the flow, as shown on the bottom of Figure 1. This scenario models intelligent, dynamic caching often done at local network head-ends. We will compute the requirements on D, d, B and on the buffer X of this intermediate node in Section VI. Moreover, we will see in Section VII that for constant rate shaping curves and rate-latency service curves, the size of the client buffer B can be reduced by implementing a particular smoothing strategy instead of FIFO scheduling at the intermediate node. If both networks are null networks, Rexford and Towsley [11] have proven that the optimal resource allocation places all buffer requirements at the client node, and none at the intermediate node ($X = 0$). Using a different approach based on network calculus, we show that it is no longer the case when the backbone network between the two server nodes may introduce some bounded amount of jitter, and that a non-zero buffer X is needed.

II. Requirements for lossless video smoothing over a single network

Let us now formulate the constraints implied by the simple scenario of the top of Figure 1.

Let us introduce some notation first: R, x, y are the cumulative values of the flows (e.g., $R(t)$ denotes therefore the number of bits of the video flow in $[0, t]$). They belong to $\mathcal{F}_0$, the set of wide-sense increasing sequences or functions f taking values in $\mathbb{R}^+ \cup \{\infty\}$ and that are zero for non positive times, namely, for which $0 \leq f(t_1) \leq f(t_2)$ for any $0 < t_1 < t_2$ and $f(t) = 0$ for any $t \leq 0$. $f \wedge g$ denotes $\min\{f, g\}$ and $f \vee g$ denotes $\max\{f, g\}$. Finally, two useful operators from min-plus algebra [2] that can be used for deriving the results are the min-plus convolution of two functions or sequences $f, g \in \mathcal{F}_0$

$$(f \otimes g)(t) = \inf_{s \,:\, 0 \leq s \leq t} \{f(t-s) + g(s)\} \tag{2}$$

and the min-plus deconvolution of two functions or sequences $f, g \in \mathcal{F}_0$

$$(f \oslash g)(t) = \sup_{u \,:\, u \geq 0} \{f(t+u) - g(u)\}. \tag{3}$$

Note that $(f \otimes g) \in \mathcal{F}_0$ but that $(f \oslash g)$ is not necessarily zero for non positive times, and hence does not belong to $\mathcal{F}_0$ in general. A useful property of these two operators that if $f, g, h \in \mathcal{F}_0$ then

$$f \leq g \otimes h \quad \text{if and only if} \quad f \oslash g \leq h. \tag{4}$$

We can now formalize the constraints that completely define the first problem (illustrated on top of Figure 1), and that will also apply to the second problem (illustrated on bottom of Figure 1).

• **Causal flow** x: As mentioned above, the chosen origin of time is such that $x(t) = 0$ for $t \leq 0$, or equivalently

$$x(t) \leq \delta_0(t). \tag{5}$$

• **Smoothness constraint**: Flow x is constrained by an arrival curve $\sigma(\cdot)$. This means that for all $t \geq 0$ and for all $s \in [0,t]$, $x(t) - x(s) \leq \sigma(t-s)$ or equivalently, with the use of the convolution operator (2), that for all $t \geq 0$

$$x(t) \leq (x \otimes \sigma)(t). \tag{6}$$

• **Playback delay constraint (no playback buffer underflow)**: The data is read out from the playback buffer after D unit of times at a rate given by $R(t-D)$. This implies that $y(t) \geq R(t-D)$. However we do not know the exact expression of y as a function of x. All we know is that the network guarantees a service curve β, namely that for all $t \geq 0$ there exists some $s \in [0,t]$ so that $y(t) \geq x(s) + \beta(t-s)$. This can be recast, using (2), as $y(t) \geq (x \otimes \beta)(t)$. The output flow may therefore be as low as $(x \otimes \beta)(t)$, and hence we can replace y in the previous inequality to obtain $(x \otimes \beta)(t) \geq R(t-D)$. Using (4) we can recast this latter inequality as

$$x(t) \geq (R \oslash \beta)(t-D) \tag{7}$$

for all $t \geq 0$.

• **Playback buffer constraint (no playback buffer overflow)**: The size of the playback buffer is limited to B, and to prevent any overflow of the buffer, we must impose that $y(t) - R(t-D) \leq B$ for all $t \geq 0$. Again, we do not know the exact value of y, but we know that it can be as high as x, but not higher, because the network is a causal system. Therefore the constraint becomes, for all $t \geq 0$,

$$x(t) \leq R(t-D) + B. \tag{8}$$

• **Look-ahead delay constraint**: We suppose that the encoder can prefetch data from the server up to d time units ahead, which translates in the following inequality:

$$x(t) \leq R(t+d). \tag{9}$$

III. Constraints on the delays and buffer sizes for a single network

In this section we compute the solution of the smoothing problem with a single network. Inequalities (5) to (9) can be recast as two sets of inequalities as follows:

$$\begin{aligned} x(t) &\leq \delta_0(t) \wedge R(t+d) \wedge \{R(t-D)+B\} \\ &\quad \wedge (\sigma \otimes x)(t) \qquad (10) \\ x(t) &\geq (R \oslash \beta)(t-D). \qquad (11) \end{aligned}$$

There is a solution x to the smoothing problem if and only if it simultaneously verifies (10) and (11). This is equivalent to requiring that the maximal solution of (10) is larger than the right hand side of (11) for all t.

Let us first compute the maximal solution of (10). Inequality (10) has the form

$$x(t) \leq a(t) \wedge (\sigma \otimes x)(t) \tag{12}$$

where

$$a(t) = \delta_0(t) \wedge R(t+d) \wedge \{R(t-D)+B\}. \tag{13}$$

The existence of a maximal solution to (12) is a consequence of Theorem 4.70, item 6 from [2]. The application of this theorem, as made in [8], gives its explicit formulation, which is

$$x_{\max}(t) = (\overline{\sigma} \otimes a)(t) \tag{14}$$

where $\overline{\sigma}$ is the sub-additive closure of σ defined as

$$\overline{\sigma} = \inf\{\delta_0, \sigma, \sigma \otimes \sigma, \dots\} = \inf_{n \in \mathbb{N}} \{\sigma^{(n)}\} \tag{15}$$

and where $\sigma^{(n)}$ is the nth self-convolution of σ, defined recursively by $\sigma^{(n+1)} = \sigma^{(n)} \otimes \sigma$ and $\sigma^{(0)} = \delta_0$. Here $\sigma(t) = rt$ is an additive function, so $\overline{\sigma}(t) = \sigma(t) = rt$. The latter also holds for more general functions σ that are only sub-additive (i.e. such that $\sigma(t) + \sigma(t') \geq \sigma(t+t')$ for all $t, t' \geq 0$).

Replacing a by its expression in (13), we compute that the maximal solution of (10) is

$$\begin{aligned} x_{\max}(t) &= \sigma(t) \wedge \{(\sigma \otimes R)(t+d)\} \\ &\quad \wedge \{(\sigma \otimes R)(t-D)+B\}. \qquad (16) \end{aligned}$$

We are now able to compute the smallest values of the playback delay D, of the total delay T and of the playback buffer B ensuring the existence of a solution to the smoothing problem, thanks to following theorem. The requirement on d for reaching the smallest value of D is therefore $d = T - D$.

Theorem 1: The smallest values of D, T and B ensuring a lossless smoothing to a (sub-additive) curve σ through a network offering a service curve β are

$$D_{\min} = \inf\{t \geq 0 : (R \oslash (\beta \otimes \sigma))(-t) \leq 0\} \quad (17)$$

$$T_{\min} = \inf\{t \geq 0 : ((R \oslash R) \oslash (\beta \otimes \sigma))(-t) \leq 0\} \quad (18)$$

$$B_{\min} = ((R \oslash R) \oslash (\beta \otimes \sigma))(0). \quad (19)$$

Proof: The set of inequalities (10) and (11) has a solution if, and only if, the maximal solution of (10) is larger or equal to the right hand side of (11) at all times. This amounts to impose that for all $t \in \mathbb{R}$

$$\begin{aligned}(R \oslash \beta)(t - D) - \sigma(t) &\leq 0\\ (R \oslash \beta)(t - D) - (\sigma \otimes R)(t + d) &\leq 0\\ (R \oslash \beta)(t - D) - (\sigma \otimes R)(t - D) &\leq B.\end{aligned}$$

Using the deconvolution operator and its properties, the latter three inequalities can be recast as

$$\begin{aligned}(R \oslash (\beta \otimes \sigma))(-D) &\leq 0\\ ((R \oslash R) \oslash (\beta \otimes \sigma))(-T) &\leq 0\\ ((R \oslash R) \oslash (\beta \otimes \sigma))(0) &\leq B.\end{aligned}$$

The minimal values of D, T and B satisfying these three inequalities are given by (17), (18) and (19). These three inequalities are therefore the necessary and sufficient conditions ensuring the existence of a solution to the smoothing problem. ■

Now that we have computed the optimal parameters, we will discuss the smoothing strategy that realizes a possible solution $x(t)$ with this setting.

Let us first mention that a dual approach (using max-plus algebra instead of min-plus algebra) to the previous one is to replace the set of inequalities (10) and (11) by the equivalent set

$$x(t) \leq \delta_0(t) \wedge R(t + d) \wedge \{R(t - D) + B\} \quad (20)$$

$$x(t) \geq (R \oslash \beta)(t - D) \vee (x \oslash \sigma)(t). \quad (21)$$

One can then compute the minimal solution of (21), which is

$$x_{\min}(t) = (R \oslash (\beta \otimes \sigma))(t - D), \quad (22)$$

and compute the constraints on d, D and B ensuring that it verifies (20): one would get the very same values of $D_{\min}$, $T_{\min}$ and $B_{\min}$ given by (17) (18) and (19). This solution is the one proposed in [9]. It does achieve the values of $D_{\min}$ and $B_{\min}$ given by (17) and (19) but requires however the evaluation at time t of a supremum over all values of R up to the end of the trace.

On the contrary, the maximal solution (16) requires only the evaluation of an infimum at time t over the past values of R and over the future values of R up to time $t + d_{\min}$, with $d_{\min} = T_{\min} - D_{\min}$. Of course, we still need the complete traffic trace to dimension $D_{\min}$, $d_{\min}$ and $B_{\min}$. However, once we have these values, we do not need the full stream for the computation of the smoothed input to the network.

If N denotes the number of frames in the stream, the complexity in computing the bounds on $D_{\min}$ is $O(N)$, and on $d_{\min}$ and $B_{\min}$ is $O(N^2)$, which can be reduced to $O(N)$ by using the time-inversion technique in [9]. The implementation of the smoothing strategy as (16) has however a very low complexity. Indeed, in the CBR discrete time setting ($\sigma = \lambda_r$, $t \in \mathbb{N}$), we can rewrite (16) as

$$x_{\max}(t) = \{x_{\max}(t-1)+r\} \wedge R(t+d) \wedge \{R(t-D)+B\},$$

and therefore only necessitates two comparisons at each time $t \in \mathbb{N}_0$.

The top of Figure 2 shows the maximal solution (16) for a CBR smoothing curve $\sigma(t) = \lambda_r(t) = rt$ and a service curve $\beta(t) = \delta_0(t)$, whereas the bottom shows the minimal solution (22). Any non-decreasing function $x(t)$ such that $x_{\min}(t) \leq x(t) \leq x_{\max}(t)$ and $x(t) \leq (x \otimes \lambda_r)(t)$ is a solution to the smoothing problem, for the same optimal values of the playback delay, look-ahead delay and client buffer size. This gives the set of all solutions. A particular solution among these can be selected to further minimize another metric, such as the ones discussed in [6], e.g. number of rate changes or rate variability. For example, for the latter metric, we could apply the algorithm devised in [14] to select a solution minimizing the playback delay and buffer together with the rate variability by replacing the envelopes $B(t)$ and $D(t)$ defined in [14] by $x_{\max}(t)$ and $x_{\min}(t)$, respectively. The difference is that contrary to $B(t)$ and $D(t)$, $x_{\max}(t)$ and $x_{\min}(t)$ always satisfy the constant rate r constraint, no matter how low it is as long as Theorem 1 is satisfied.

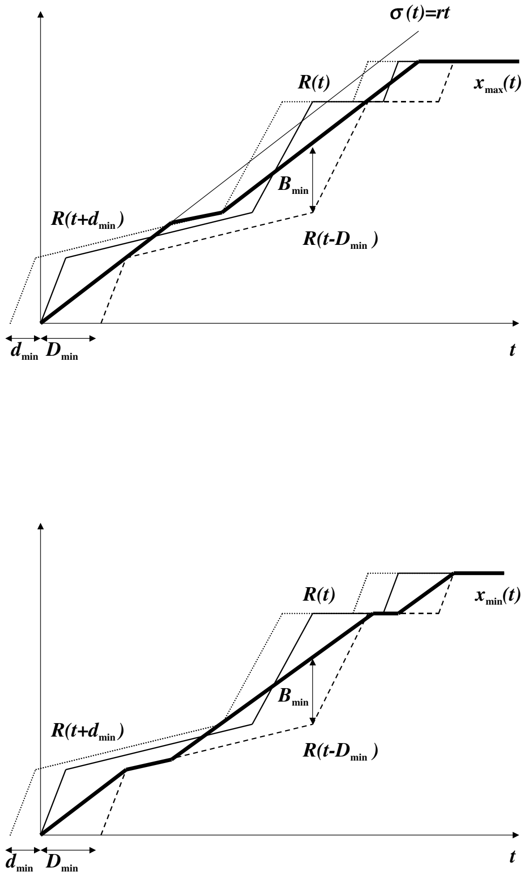

Fig. 2. In bold, the maximal solution (top figure) and minimal solution (bottom figure) to the CBR smoothing problem with a null network.

IV. Optimal constant rate smoothing with a rate-latency service curve

Let us compute the above values in the case of a constant rate (CBR) smoothing curve $\sigma(t) = \lambda_r(t) = rt$ (with $t \geq 0$) and a rate-latency service curve of the network $\beta(t) = \beta_{L,C}(t) = C[t-L]^+$. We assume that $r \leq C$, the less interesting case where $r > C$ being handled similarly. We will often use the decomposition of a rate-latency function as the min-plus convolution of a pure delay function, as defined by (1), with a constant rate function:

$$\beta_{L,C} = \delta_L \otimes \lambda_C. \tag{23}$$

In this particular case, the optimal values (17), (18) and (19) become the following ones.

Theorem 2: If $\sigma = \lambda_r$ and $\beta = \beta_{L,C}$ with $r < C$, the smallest values of D, of T and of B are

$$\begin{aligned}
D_{\min} &= L + \frac{1}{r}(R \oslash \lambda_r)(0) && (24)\\
&= L + \frac{1}{r}\sup_{t\geq 0}\{R(t) - rt\}\\
T_{\min} &= L + \frac{1}{r}((R \oslash R) \oslash \lambda_r)(0) && (25)\\
&= L + \frac{1}{r}\sup_{t,u\geq 0}\{R(t+u) - R(u) - rt\}\\
B_{\min} &= ((R \oslash R) \oslash \lambda_r))(L) \leq rT_{\min}. && (26)
\end{aligned}$$

Proof: To establish (24) and (25), we note that R and $(R \oslash R) \in \mathcal{F}_0$. Because of (23) and as $r < C$,

$$\begin{aligned}
\beta \otimes \sigma &= \beta_{L,C} \otimes \lambda_r = \delta_L \otimes \lambda_C \otimes \lambda_r\\
&= \delta_L \otimes \lambda_r = \beta_{L,r}
\end{aligned}$$

so that we can apply Lemma 5 of the Appendix with $f = R$ and $f = (R \oslash R)$, respectively.

To establish (26), we develop (19) as follows

$$\begin{aligned}
&((R \oslash R) \oslash (\beta \otimes \sigma))(0)\\
&= ((R \oslash R) \oslash (\delta_L \otimes \lambda_r))(0)\\
&= \sup_{u\geq 0}\{(R \oslash R)(u) - \lambda_r(u-L)\}\\
&= ((R \oslash R) \oslash \lambda_r)(L)\\
&= \sup_{u\geq L}\{(R \oslash R)(u) - \lambda_r(u-L)\}\\
&= \sup_{u\geq L}\{(R \oslash R)(u) - \lambda_r(u)\} + rL\\
&\leq \sup_{u\geq 0}\{(R \oslash R)(u) - \lambda_r(u)\} + rL\\
&= ((R \oslash R) \oslash \lambda_r)(0) + rL = rT_{\min}.
\end{aligned}$$

■

In the particular case where $L = 0$ and $r < C$ the network is completely transparent to the flow, and can be considered as a null network meaning that we can replace $\beta(t)$ by $\delta_0(t)$. The values (24), (25) and (26) become, respectively,

$$\begin{aligned}
D_{\min} &= \frac{1}{r}(R \oslash \lambda_r)(0) && (27)\\
T_{\min} &= \frac{1}{r}((R \oslash R) \oslash \lambda_r)(0) && (28)\\
B_{\min} &= ((R \oslash R) \oslash \lambda_r))(0) = rT_{\min}. && (29)
\end{aligned}$$

We have computed these values for different traces. Three of them are displayed on Figure 3, in bytes par frame: a video conference trace of 5000 frames and a Jurassic Park movie of 40000 frames, which are traces of MPEG-1 encoded streams available from [12], [13] (for lisibility, only 1000 frames of each of these two traces are displayed in Figure 3), and an MPEG-2 video sequence of 400 frames from [9]. The encoding generates 25 frames per second.

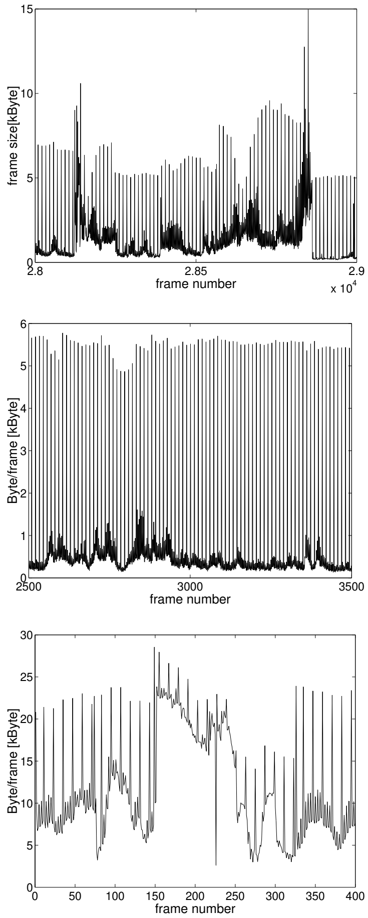

Fig. 3. Three traces: Jurassic Park (top), video conference (middle) and MPEG2 video-sequence (bottom) in bytes per frame.

Since $B_{\min}$ is directly proportional to $T_{\min}$ because of (29), we show only the graphs of the values of $D_{\min}$ and $d_{\min} = T_{\min} - D_{\min}$ for different traces on Figures 4, 5 and 6 as a function of the CBR smoothing rate r. We observe three qualitative ranges of rates: (i) the very low ones where the playback delay is very large, and where look-ahead does not help in reducing it; (ii) the quasi-optimal middle range where the playback delay is negligible, thanks to the use of look-ahead and (iii) the high rates above the peak rate of the stream, which do not require any playback nor lookahead of the stream. The boundaries between these three ranges differ from one stream to another, but are clearly present in all traces.

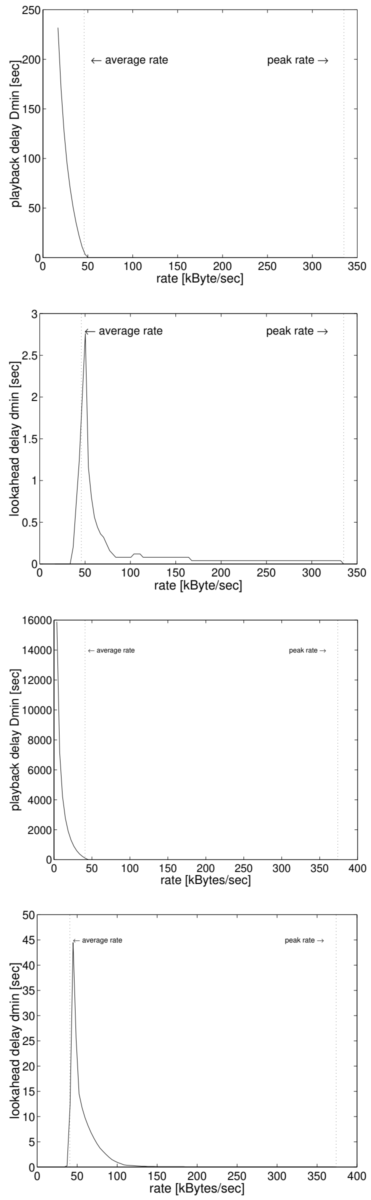

Fig. 4. Minimum playback delay $D_{\min}$ and corresponding look-ahead delay $d_{\min}$ for a constant rate smoothing of the Jurassic park trace, for different lengths: the first 3300 frames (top two graphs) and the full 40000 frames (bottom two graphs)

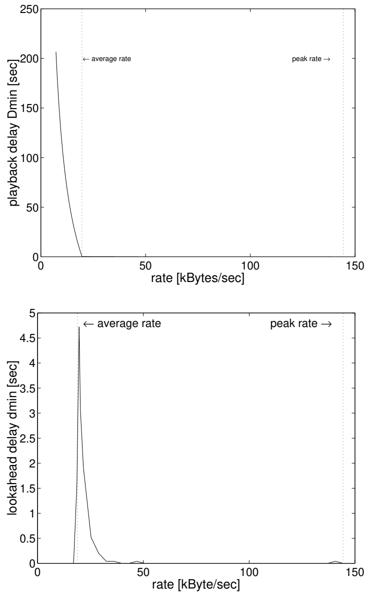

Fig. 5. Minimum playback delay $D_{\min}$ and corresponding look-ahead delay $d_{\min}$ for a constant rate smoothing of the video conference trace (5000 frames).

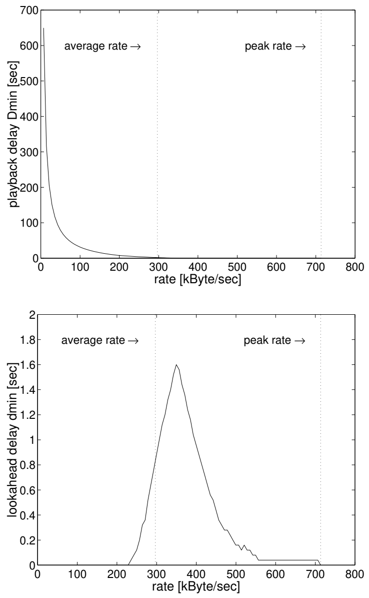

Fig. 6. Minimum playback delay $D_{\min}$ and corresponding look-ahead delay $d_{\min}$ for a constant rate smoothing of the MPEG-2 video trace shown on bottom of Figure 3 (400 frames).

V. Requirements for lossless video smoothing over two networks

We now move to the smoothing problem over two networks. Here two flows need to be computed: the first one $x_1(t)$ at the input of the backbone network, and the second one $x_2(t)$ at the input of the local access network, as shown on the bottom of Figure 1.

The constraints on both flows are now as follows:

- **Causal flow** x_1: This constraint is the same as (5), but with x replaced by x_1:

$$x_1(t) \leq \delta_0(t), \tag{30}$$

- **Smoothness constraint**: Both flows x_1 and x_2 are constrained by two arrival curves σ_1 and σ_2:

$$x_1(t) \leq (x_1 \otimes \sigma_1)(t) \tag{31}$$
$$x_2(t) \leq (x_2 \otimes \sigma_2)(t). \tag{32}$$

- **No playback and intermediate server buffers underflow**: The data is read out from the playback buffer after D unit of times at a rate given by $R(t-D)$, which implies that $y_2(t) \geq R(t-D)$. On the other hand, the data is retrieved from the intermediate server at a rate given by $x_2(t)$, which implies that $y_1(t) \geq x_2(t)$. As we do not know the expressions of the outputs of each network, but only a service curve β_1 and β_2 for each of them, we can replace y_1 by $x_1 \otimes \beta_1$ and y_2 by $x_2 \otimes \beta_2$, and reformulate these two constraints by

$$x_2(t) \leq (x_1 \otimes \beta_1)(t) \tag{33}$$
$$x_2(t) \geq (R \oslash \beta_2)(t-D). \tag{34}$$

- **No playback and intermediate server buffers overflow**: The size of the playback and cache buffers are limited to B and X, respectively, and to prevent any overflow of the buffer, we must impose that $y_1(t) - x_2(t) \leq X$ and $y_2(t) - R(t-D) \leq B$ for all $t \geq 0$. Again, we do not know the exact value of y_1 and y_2, but we know that they are bounded by x_1 and x_2, respectively, so that the constraints becomes, for all $t \geq 0$,

$$x_1(t) \leq x_2(t) + X \tag{35}$$
$$x_2(t) \leq R(t-D) + B. \tag{36}$$

- **Look-ahead delay constraint**: this constraint is the same as in the single network case:

$$x_1(t) \leq R(t+d). \tag{37}$$

VI. Constraints on the delays and buffer sizes for two networks

Inequalities (30) to (37) can be recast as three sets of inequalities as follows:

$$\begin{aligned} x_1(t) &\leq \delta_0(t) \wedge R(t+d) \wedge (\sigma_1 \otimes x_1)(t) \\ &\qquad \wedge (x_2(t) + X) \qquad (38) \\ x_2(t) &\leq \{R(t-D) + B\} \wedge (\beta_1 \otimes x_1)(t) \\ &\qquad \wedge (\sigma_2 \otimes x_2)(t) \qquad (39) \\ x_2(t) &\geq (R \oslash \beta_2)(t-D). \qquad (40) \end{aligned}$$

With T denoting transposition, let us introduce the following notations: for $t \geq 0$

$$\begin{aligned} \vec{x}(t) &= [x_1(t) \quad x_2(t)]^T \\ \vec{a}(t) &= [\delta_0(t) \wedge R(t+d) \quad R(t-D) + B]^T \\ \vec{b}(t) &= [0 \quad (R \oslash \beta_2)(t-D)]^T \end{aligned}$$

and for $0 \leq s \leq t$

$$H(t-s) = \begin{bmatrix} \sigma_1(t-s) & \delta_0(t-s) + X \\ \beta_1(t-s) & \sigma_2(t-s) \end{bmatrix}.$$

We extend $\wedge$ and $\otimes$ to vectors and matrices in the usual way [1]. For example, with $\vec{f}, \vec{g} \in \mathcal{F}_0^2$ and $A, B \in \mathcal{F}_0^2 \times \mathcal{F}_0^2$ we define $\vec{f} \wedge \vec{g}$ as the coordinate-wise minimum of $\vec{f}$ and $\vec{g}$, and $A \wedge B$ as the entry-wise minimum of A and B. We write $\vec{f} \leq \vec{g}$ with the meaning that $\vec{f}_j \leq \vec{g}_j$ for $j = 1, 2$. For a constant K, we note $\vec{f} + K$ the vector defined by adding K to all elements of $\vec{f}$. $A \otimes \vec{f}$ has entries $\min_{j=1,2}(A_{ij} \otimes \vec{f}_j)$ for $i = 1, 2$ and $A \otimes B$ has entries $\min_{j=1,2}(A_{ij} \otimes B_{jk})$ for $i, k = 1, 2$.

With these notations, the set of inequalities (38), (39) and (40) can therefore be recast as

$$\vec{x}(t) \leq \vec{a}(t) \wedge (H \otimes \vec{x})(t) \qquad (41)$$

$$\vec{x}(t) \geq \vec{b}(t). \qquad (42)$$

We will follow the same approach as in Section III: we first compute the maximal solution of (41) and then derive the constraints on D, T (and hence d), X and B ensuring the existence of this solution. The application of Theorem 4.70, item 6, from [2] in the two-dimensional case gives an explicit formulation of the maximal solution of (41), which is

$$\vec{x}_{\max}(t) = (\overline{H} \otimes \vec{a})(t). \qquad (43)$$

$\overline{H}$ is the sub-additive closure of H defined as

$$\overline{H} = \inf_{n \in \mathbb{N}} \{H^{(n)}\} \qquad (44)$$

where $I_\delta = H^{(0)}$ is the 2×2 matrix whose diagonal entries are δ_0 and whose non-diagonal entries are ∞, and $H^{(n)}$ denotes the nth self-convolution of H. Application of (44) to matrix H is straightforward, but involves a few manipulations which are skipped here for lack of space. Denoting

$$\begin{aligned} \alpha &= \sigma_1 \otimes \sigma_2 \otimes \inf_{n \in \mathbb{N}} \left\{ \beta_1^{(n+1)} + nX \right\} \qquad (45) \\ &= \sigma_1 \otimes \sigma_2 \otimes \beta_1 \otimes \overline{\beta_1 + X} \end{aligned}$$

we find that

$$\overline{H} = \begin{bmatrix} \sigma_1 \wedge (\alpha + X) & (\sigma_1 \otimes \sigma_2 + X) \wedge (\alpha + 2X) \\ \alpha & \sigma_2 \wedge (\alpha + X) \end{bmatrix}$$

and therefore the two coordinates of the maximal solution of (41) are

$$\begin{aligned} x_{1\max}(t) &= \sigma(t) \wedge \{\alpha(t) + X\} \wedge \\ &\wedge \quad (\sigma_1 \otimes R)(t+d) \\ &\wedge \quad \{(\alpha \otimes R)(t+d) + X\} \\ &\wedge \quad \{(\sigma_1 \otimes \sigma_2 \otimes R)(t-D) + B + X\} \\ &\wedge \quad \{(\alpha \otimes R)(t-D) + B + 2X\} \qquad (46) \\ x_{2\max}(t) &= \alpha(t) \wedge (\alpha \otimes R)(t+d) \\ &\wedge \quad \{(\sigma_2 \otimes R)(t-D) + B\} \\ &\wedge \quad \{(\alpha \otimes R)(t-D) + B + X\}. \qquad (47) \end{aligned}$$

We can now express the constraints on X, B, D and d that will ensure that a solution exists by requiring that (47) be larger than (40). The result is stated in the following theorem, whose proof is similar to that of Theorem 1.

Theorem 3: The lossless smoothing of a flow to (sub-additive) curves σ_1 and σ_2, respectively, over two networks offering service curves β_1 and β_2 has a solution if and only if the D, T, X and B verify the following set of inequalities, with α defined by (45):

$$\begin{aligned} (R \oslash (\alpha \otimes \beta_2)(-D) &\leq 0 \qquad (48) \\ ((R \oslash R) \oslash (\alpha \otimes \beta_2))(-T) &\leq 0 \qquad (49) \\ ((R \oslash R) \oslash (\sigma_2 \otimes \beta_2))(0) &\leq B \qquad (50) \\ ((R \oslash R) \oslash (\alpha \otimes \beta_2))(0) &\leq B + X. \qquad (51) \end{aligned}$$

VII. Optimal constant rate smoothing over two networks

Let us compute the above values in the case of two constant rate (CBR) smoothing curves $\sigma_1 = \lambda_{r_1}$ and $\sigma_2 = \lambda_{r_2}$. We assume that each network offers a rate-latency service curve $\beta_i = \beta_{L_i,C_i}, i = 1,2$. We assume that $r_i \leq C_i$ In this case the optimal values of D, T and B become the following ones, depending on the value of X.

Theorem 4: Let $r = r_1 \wedge r_2$. Then we have the following three cases depending on X:

(i) If $X \geq rL_1$, then $D_{\min}, T_{\min}$ and $B_{\min}$ are given by

$$\begin{aligned} D_{\min} &= L_1 + L_2 + \frac{1}{r}(R \oslash \lambda_r)(0) && (52)\\ T_{\min} &= L_1 + L_2 + \frac{1}{r}((R \oslash R) \oslash \lambda_r)(0) && (53)\\ B_{\min} &= ((R \oslash R) \oslash \lambda_{r_2})(L_2) \\ &\quad \vee \{((R \oslash R) \oslash \lambda_r)(L_1 + L_2) - X\} \\ &\leq ((R \oslash R) \oslash \lambda_r)(L_2). && (54) \end{aligned}$$

(ii) If $0 < X < rL_1$ then $D_{\min}$, $T_{\min}$ and $B_{\min}$ are bounded by

$$\begin{aligned} &\frac{X}{r} + L_2 + \frac{L_1}{X}(R \oslash \lambda_{\frac{X}{L_1}})(0) \leq D_{\min} \\ &\quad \leq L_1 + L_2 + \frac{L_1}{X}(R \oslash \lambda_{\frac{X}{L_1}})(0) && (55)\\ &\frac{X}{r} + L_2 + \frac{L_1}{X}((R \oslash R) \oslash \lambda_{\frac{X}{L_1}})(0) \leq T_{\min} \\ &\quad \leq L_1 + L_2 + \frac{L_1}{X}((R \oslash R) \oslash \lambda_{\frac{X}{L_1}})(0) && (56)\\ &((R \oslash R) \oslash \lambda_{\frac{X}{L_1}})(L_1 + L_2) - r_2 L_1 \leq B_{\min} \\ &\quad \leq ((R \oslash R) \oslash \lambda_{\frac{X}{L_1}})(L_2) && (57) \end{aligned}$$

(iii) Let K be duration of the stream. If $X = 0 < rL_1$ then $D_{\min} = K$.

Proof: One easily verifies that $\delta_{L_1}^{(n+1)} = \delta_{(n+1)L_1}$ and that $\lambda_{C_1}^{(n+1)} = \lambda_{C_1}$. Since $\beta_1 = \beta_{L_1,C_1} = \delta_{L_1} \otimes \lambda_{C_1}$ because of (23), and since $r = r_1 \wedge r_2 \leq C_1$, (45) becomes

$$\begin{aligned} \alpha &= \lambda_r \otimes \inf_{n \in \mathbb{N}} \left\{\delta_{(n+1)L_1} \otimes \lambda_{C_1} + nX\right\} \\ &= \delta_{L_1} \otimes \inf_{n \in \mathbb{N}} \left\{\delta_{nL_1} \otimes \lambda_r + nX\right\}. && (58) \end{aligned}$$

(i) If $X \geq rL_1$, then one checks that $(\delta_{nL_1} \otimes \lambda_r)(t) + nX \geq rt = \lambda_r(t)$ for all $t \geq 0$, which implies that the infimum in (58) is reached for $n = 0$ and hence that $\alpha = \delta_{L_1} \otimes \lambda_r$. It follows that

$$\alpha \otimes \beta_2 = \delta_{L_1+L_2} \otimes \lambda_r. \qquad (59)$$

Inserting this last relation in (48), and using Theorem 5 we establish (52), (53) and the equality in (54). The inequality in (54) is obtained by noticing that $r_2 \geq r$ and that

$$\begin{aligned} &((R \oslash R) \oslash \lambda_r)(L_1 + L_2) - X \\ &= \sup_{u \geq 0}\{(R \oslash R)(u + L_1 + L_2) - ru\} - X \\ &= \sup_{v \geq L_1}\{(R \oslash R)(v + L_2) - r(v - L_1)\} - X \\ &\leq \sup_{v \geq 0}\{(R \oslash R)(v + L_2) - rv\} + (rL_1 - X) \\ &\leq ((R \oslash R) \oslash \lambda_r)(L_2). \end{aligned}$$

(ii) If $0 < X < rL_1$, the computation of α does not provide a rate-latency curve anymore, but a function that can be bounded below and above by the two following rate-latency curves: $\beta_{L_1,X/L_1} \leq \alpha \leq \beta_{X/r,X/L_1}$. Therefore, replacing (59) by

$$\delta_{L_1+L_2} \otimes \lambda_{\frac{X}{L_1}} \leq \alpha \otimes \beta_2 \leq \delta_{\frac{X}{r}+L_2} \otimes \lambda_{\frac{X}{L_1}}$$

and applying Theorem 5 for both bounding rate-latency curves, we get respectively the lower and upper bounds.

(iii) If $X = 0$ and $rL_1 > 0$ then (58) yields that $\alpha(t) = 0$ for all $t \geq 0$. In this case (48) becomes $\sup_{u \geq 0}\{R(u - D)\} \leq 0$. This is possible only if D is equal to the duration of the stream. ■

It is interesting to examine these results for the particular values of X.

The first one is $X = \infty$. If the intermediate server is a greedy shaper [7], [3] whose output is $x_2(t) = (\sigma_2 \otimes y_1)(t)$, one could have applied Theorem 2 with $\sigma_2 = \lambda_r$ and $\beta = \beta_1 \otimes \sigma_2 \otimes \beta_2 = \delta_{L_1+L_2} \otimes \lambda_{r_2} = \beta_{L_1+L_2,r_2}$ to find out that D and T are still given by (52) and (53) but that $B = ((R \oslash R) \oslash \lambda_r)(L_1 + L_2)$ is larger than (54). Using the caching scheduling (47) instead of a greedy shaping one allows therefore to decrease the playback buffer size, but not the delays. The intermediate buffer node does not need to be infinite, but can be limited to rL_1.

The second one is $X = 0$. Then whatever the rate $r > 0$, if $L_1 > 0$ the playback delay is the length of the stream, which makes streaming impossible in practice.

When $L_1 = L_2 = 0$ however (in which case we have two null networks) $X = rL_1 = 0$ is the optimal intermediate node buffer allocation, as shown in [11] using other arguments. When $L_1 > 0$, this is no longer the case.

VIII. CONCLUSION

Network calculus enabled us to compute the minimal requirements on the playback delay, buffer size and look-ahead delay to ensure lossless video smoothing over one or two networks offering guaranteed service. The same approach could be extended for internetworking with more than two networks. We provided closed-form expressions for the most important practical case of a constant rate smoothing over a rate-latency service curve. We showed that look-ahead can prove useful over a quasi-optimal range of rates by keeping the playback delay to a very small value.

APPENDIX

Theorem 5: Let $f \in \mathcal{F}_0$ and $\beta_{L,C}$ be a rate-latency curve. Define

$$W = \inf\{t \geq 0 \;:\; (f \oslash \beta_{L,C})(-t) \leq 0\}.$$

Then

$$W = L + \frac{1}{C}(f \oslash \lambda_C)(0). \tag{60}$$

Proof: As $f(t) = 0$ for $t \leq 0$ and as $\beta_{L,C} = \delta_L \otimes \lambda_C$, we can write

$$\begin{aligned}
&(f \oslash \beta_{L,C})(-W) \\
&= \sup_{u \geq 0}\{f(u - W) - (\delta_L \otimes \lambda_C)(u)\} \\
&= \sup_{u \geq 0}\{f(u - W) - \lambda_C(u - L)\} \\
&= \sup_{v \geq -W}\{f(v) - \lambda_C(v + W - L)\} \\
&= \sup_{v \geq 0}\{f(v) - \lambda_C(v + W - L)\} \\
&= \sup_{v \geq 0}\{f(v) - \lambda_C(v)\} - C(W - L) \\
&= (f \oslash \lambda_C)(0) - CW + CL,
\end{aligned}$$

from which we deduce the smallest value of W making the left-hand side of this equation non-positive is given by (60). ∎

REFERENCES

[1] R. Agrawal, F. Baccelli, R. Rajan 'An Algebra for queueing networks with time varying service and its application to the analysis of integrated service networks', Report RR-3435, INRIA, Sophia-Antipolis, France, May 1998.

[2] F. Baccelli, G. Cohen, G. J. Olsder and J.-P. Quadrat. *Synchronization and Linearity, An Algebra for Discrete Event Systems*, John Wiley and Sons, August 1992.

[3] C.S. Chang, *Performance Guarantees in Communication Networks*, Springer-Verlag, 2000.

[4] R.L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE JSAC*, **vol. 13**(6), pp. 1048–1056, Aug. 1995.

[5] N. G. Duffield, K. K. Ramakrishan and A. R. Reibman, 'SAVE: An algorithm for smoothed adaptative video over explicit rate networks', *IEEE/ACM Transactions on Networking*, **vol. 6**, pp. 717–728, Dec. 1998.

[6] W.-C. Feng and J. Rexford, 'Performance evaluation of smoothing algorithms for transmitting Variable-Bit-Rate video', *IEEE Transactions on Multimedia*, **vol. 1**, pp. 302–312, Sept. 1999.

[7] J.-Y. Le Boudec. 'Application of Network Calculus to Guaranteed Service Networks', *IEEE Transactions on Information Theory*, **vol. 44**, pp. 1087–1096, May 1998.

[8] J.-Y. Le Boudec and P. Thiran. 'Network calculus using Min-plus System Theory', Chapter 9 of *High-Performance Networks for Multimedia Applications*, A. Danthine et al editors, Kluwer, Boston, 1999.

[9] J.-Y. Le Boudec and O. Verscheure. 'Optimal smoothing for guaranteed service', Technical report DSC2000/014, EPFL, March 2000.

[10] J. M. McManus and K.W. Ross, 'Video-on-Demand over ATM: Constant-rate transmission and transport,' *IEEE JSAC*, **vol. 14**(6), pages 1087–1098, August 1996.

[11] J. Rexford and D. Towsley, 'Smoothing Variable-Bit-Rate Video in an Internetwork', *IEEE/ACM Transactions on Networking*, **vol. 7**, pp. 202–215, April 1999.

[12] `ftp-info3.informatik.uni-wuerzburg.de`

[13] O. Rose, 'Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems', Proc. 20th Annual Conf. on Local Computer Networks, Minneapolis, pp. 397–406, 1995.

[14] J. D. Salehi, Z-L Zhang, J. F. Kurose and D. Towsley, 'Supporting stored video: Reducing rate variability and end-to-end resource requirements through optimal smoothing', *IEEE/ACM Transactions on Networking*, **vol. 6**, pp. 397–410, Dec. 1998.