Throughput and Delay Analysis in Video Streaming over Block-Fading Channels

G. Cocco[‡], D. Gündüz[*] and C. Ibars[†]

[‡] German Aerospace Center (DLR), Weßling, Germany

[*] Imperial College, London, United Kingdom

[†] Intel Corporation, Santa Clara, CA, USA

giuseppe.cocco@dlr.de, d.gunduz@imperial.ac.uk, christian.ibars.casas@intel.com

Abstract—We study video streaming over a slow fading wireless channel. In a streaming application video packets are required to be decoded and displayed in the order they are transmitted as the transmission goes on. This results in per-packet delay constraints, and the resulting channel can be modeled as a physically degraded fading broadcast channel with as many virtual users as the number of packets. In this paper we study two important quality of user experience (QoE) metrics, namely *throughput* and *inter-decoding delay*. We introduce several transmission schemes, and compare their throughput and maximum inter-decoding delay performances. We also introduce a genie-aided scheme, which provides theoretical bounds on the achievable performance. We observe that adapting the transmission rate at the packet level, i.e., periodically dropping a subset of the packets, leads to a good tradeoff between the throughput and the maximum inter-decoding delay. We also show that an approach based on initial buffering leads to an asymptotically vanishing packet loss rate at the expense of a relatively large initial delay. For this scheme we derive a condition on the buffering time that leads to throughput maximization.

I. INTRODUCTION

Video traffic constitutes a large portion of today's Internet data flow, and it is foreseen to exceed 70% of the total IP traffic within the next five years [1]. A significant portion of the video traffic is generated by streaming applications, such as YouTube and Netflix. This, together with the increasing utilization of mobile terminals for streaming high-definition video content, poses growing challenges to mobile network operators in terms of bandwidth availability and quality of user experience (QoE).

Mobile wireless channels are often modelled with block fading, where the channel gain stays constant during the channel coherence time, and changes independently across channel blocks according to a certain probability distribution [2]. From the extensive literature on fading channels (see, e.g., [3]-[9]), it emerges that a pivotal role for reliable communications is played by the delay constraint, which is a critical design parameter in streaming applications.

In [10] and [11] the broadcast strategy proposed in [12] is used to improve the end-to-end quality in multimedia transmission. However, the broadcast strategy requires encoding bits into multiple superposed messages of increasing rates, and this level of fine adaptation is not possible in practical multimedia communication systems, in which the encoding rate is fixed by a higher layer application[1][13]. Moreover, practical network architectures are strictly layered, and the channel encoder is typically oblivious to the video coding scheme used by the application layer; and therefore, rate adaptation is usually not possible at the code level. Video packets received by the channel encoder are already video-encoded at a fixed rate, which cannot be changed. On the other hand, the channel encoder can choose to drop some of the video packets, and achieve rate adaptation at the packet level at the expense of *inter-decoding delay* at the receiver.

In the Moving Picture Experts Group (MPEG) standard, the video encoder output units are called group of pictures (GOP). Each GOP consists of an I- frame and a number of P- and B-frames [14]. A GOP can be decoded and displayed independently of the previous and following GOPs. We assume that a whole GOP (or an integer number of GOPs) forms one video packet, and the coding rate is normalized such that the display time of a GOP (or an integer number of GOPs) is equal to the channel coherence time[2].

We consider streaming over a Gaussian block fading channel, in which the transmitter has no channel state information (CSIT), which is the case for networks with large round trip delay (like satellite networks), or wireless broadcast networks with a large number of users[3]. Due to the lack of CSIT, the transmitter uses a fixed transmission rate. In order to minimize the probability of packet loss over the channel, the transmission rate is kept at the minimum value that allows no freezing in the display process at the receiver provided no packet is lost. This implies that the transmission time of

---

[1]Some streaming protocols, such as HTTP Live Streaming, allow rate adaption among only a limited number of available rates.

[2]With this we implicitly assume a slow varying channel, for example, a mobile terminal moving at pedestrian speed.

[3]In the downlink channel with many receiving terminals, acquisition of CSIT is not viable, since this requires the transmission of an extensive amount of information which may result in the *feedback implosion* problem [15].

a packet is equal to its display time (assuming that the time needed to process the packet at the receiver is negligible), which is assumed to be constant for all packets. In the streaming scenario, this imposes a different decoding deadline for each video packet, i.e., the first packet needs to be received by the end of the first channel block, the second packet by the end of the second block, and so on. Modeling the decoder at each channel block as a distinct virtual receiver, this channel can be seen as a physically degraded fading broadcast channel with as many virtual users as the number of channel blocks.

The loss of a data packet implies the loss of the corresponding GOP; and hence, an interruption in the playback of the video at the end user, which lasts until the next packet is received. In [16] the quality degradation due to GOP losses as perceived by the end user has been assessed by streaming pre-recorded videos while introducing video segment losses in a controlled fashion. The results illustrate that users are more tolerant to long freezes with respect to choppy playback, that is, few long freezing events are on average preferred to many short freezing events. However, this is no longer true if the transmission is for a live event, such as a sport event or news video. In this case, the loss of a large chunk of video content, which may lead to loss of important information, is much worse than choppy playback quality. In this paper we target the latter kind of video content, and consider the interdecoding delay as a performance measure.

The effect of GOP loss in video streaming has been studied in [17], [18] and [19]. In the video streaming literature, the problem is usually tackled at the network level, focusing on the effect of packet loss rate, delay and jitter [20]. However, these parameters are usually assumed to be given as fixed values to the system designer, or studied from a networking perspective, where packet losses are mainly due to buffer overflow, while jitter is due to the congestion level of the network, link failures and dynamic routing. The problem of radio resource allocation in wireless multimedia transmission over frequency selective channels is studied in [21] and [22].

We study the interaction between the physical layer and the display process of the received video data. In particular, we study different communication strategies, each of which adopts a different policy to select the subset of messages to be transmitted, as well as the amount of resources (in terms of transmission time) dedicated to each message, which has an impact on the successful decoding probability. The performance of these strategies is evaluated based on two figures of merit: average throughput and maximum inter-decoding delay [23]. The interaction between the display process and the lower layers is of fundamental importance for streaming services such as Dynamic Adaptive Streaming over HTTP (DASH), that need an estimation of the link quality in order to provide an adequate QoE to the end users. In its current implementation DASH uses the information about the link status at each user in order to optimize the QoE that can be provided with the available resources [24]. However, DASH systems require a feedback link that instructs the transmitter on the highest bit-rate that can be received in the current channel condition, whereas we assume no information on the current channel state at the transmitter, and thus the optimisation of the transmission strategy at the transmitter has to be done independently of the current channel condition.

While there is an extensive literature on the higher layer analysis of video streaming applications [25], research on the physical layer aspects of streaming focus mostly on code construction [26], [27], [28]. The diversity-multiplexing trade-off for a streaming system is studied in [29]. The channel model we study here is the dual of the streaming transmitter model studied in [30], [31], where the data packets, rather than being available at the transmitter in advance and having a per-packet delay constraint, arrive gradually over time, and have a global delay constraint.

We propose four different transmission schemes based on time-sharing[4]. More elaborate transmission techniques have been previously studied in literature such as in [10]. In [33] the problem of still images transmitting over slow fading channel using a FEC-based multiple description encoder over an OFDM modulation was studied. Unlike in such previous works, we exclusively focus on time-sharing transmission because of its applicability in practical systems, as it leads to lower complexity decoding schemes with respect to, for example, successive interference cancellation, which is required in the case of superposition transmission. Moreover, the throughput and delay analysis is not completely understood even for this relatively simpler transmission scheme. In particular, we consider *memoryless transmission (MT)*, *equal time-sharing (eTS)*, *pre-buffering (PB)* and *windowed time-sharing (wTS)* schemes. We also consider an informed transmitter (IT) bound on the achievable throughput and delay performances, assuming perfect CSIT. We compare these achievable schemes and the informed transmitter bound in terms of both through-put and maximum inter-decoding delay. Our results provide fundamental performance bounds as well as an insight for the design of practical video streaming systems over wireless fading channels.

The rest of the paper is organized as follows. In Section II we present the system model. In Section III we derive informed transmitter bounds on throughput and average maximum delay. In Section IV we presents four different transmission schemes and, for each of them, we analyze throughput and delay. Section V contains the numerical results, while the conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider a video streaming system over a block fading channel. The channel is constant for a block of n channel uses and changes in an independent and identically distributed (i.i.d.) manner from one block to the next. We assume that the file to be streamed to the receiver consists of M independent packets denoted by $W_1, \ldots, W_M$, all available at the transmitter at the very beginning. The receiver wants to decode these packets gradually as the transmitter continues its transmission.

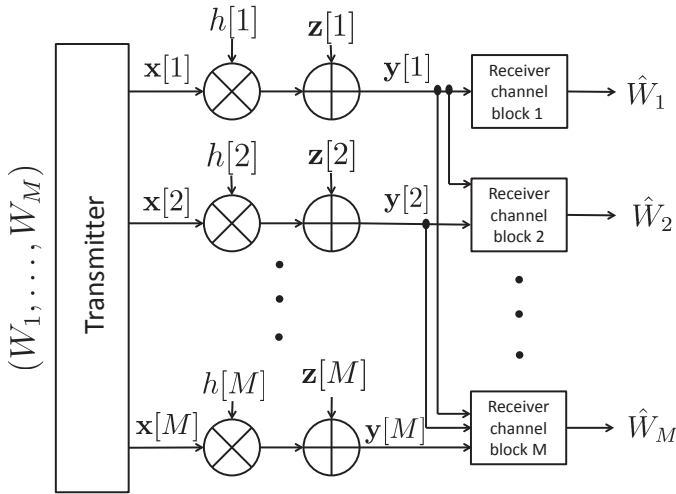[4]Part of the present work has been presented in [32].

Fig. 1. Equivalent channel model for streaming a video file composed of M packets over M blocks of the fading channel to a single receiver with a per packet delay constraint.

We assume that the packet W_t needs to be decoded by the end of channel block t, $t = 1, \ldots, M$, otherwise it becomes useless. The data packets all have the same size; and it is assumed that each packet is generated at rate R bits per channel use (bpcu), which is fixed by the application layer, i.e., W_t is chosen randomly with uniform distribution from the set $\mathcal{W}_t = \{1, \ldots, 2^{nR}\}$ [34]. The channel in block t is given by

$$\mathbf{y}[t] = h[t]\mathbf{x}[t] + \mathbf{z}[t],$$

where $h[t]$ is the channel state, $\mathbf{x}[t]$ is the length-n channel input vector, $\mathbf{z}[t]$ is a vector of i.i.d. zero mean unit-variance Gaussian noise, and $\mathbf{y}[t]$ is the length-n channel output vector at the receiver. Instantaneous channel states are known only at the receiver, while the transmitter has only statistical channel knowledge, i.e., it knows the probability density function (pdf) of $h(t)$. We have a short-term average power constraint of P, i.e., $\mathrm{E}[\mathbf{x}[t]\mathbf{x}[t]^\dagger] \leq nP$ for $t = 1, \ldots, M$, where $\mathbf{x}[t]^\dagger$ represents the Hermitian transpose of $\mathbf{x}[t]$.

The channel from the source to the receiver can be seen as a physically degraded broadcast channel, such that the decoder at each channel block acts as a virtual receiver trying to decode the packet corresponding to that channel block. See Fig. 1 for an illustration of this channel model. We denote the instantaneous channel capacity over channel block t by C_t:

$$C_t \triangleq \log_2(1 + \phi[t]P), \tag{1}$$

where $\phi[t] = |h[t]|^2$ is a random variable distributed according to a zero-mean pdf $f_\Phi(\phi)$. We define $\overline{C} \triangleq E\{C_t\}$, $E\{x\}$ being the mean value of x.

We define the average throughput, $\overline{T}$, as the average decoded rate at the end of M channel blocks:

$$\overline{T} \triangleq \frac{R}{M} \sum_{m=1}^{M} m \cdot \eta(m), \tag{2}$$

where $\eta(m)$ is the probability of decoding exactly m messages out of M.

In addition to the average throughput, we also study the *frame delay*, which is defined as the maximum number of consecutive channel blocks in which the corresponding message is not decoded, denoted by $D^{\max}$. When a video packet over a channel block is not decoded at the receiver, video playback at the receiver's device stalls, and the user continues to see the same video frame until a new GOP is successfully received. Since $D^{\max}$ is also a random variable whose realization depends on the channel, we consider the *average maximum delay* $\overline{D}^{\max}$ as our performance measure. We have:

$$\overline{D}^{\max} \triangleq \sum_{d=1}^{M} d \cdot Pr\{D^{\max} = d\} = \sum_{d=1}^{M} Pr\{D^{\max} \geq d\}. \tag{3}$$

In the next section, we first study an informed transmitter bound on the system performance, assuming perfect CSIT about all the future channel realizations.

III. INFORMED TRANSMITTER BOUND

An upper bound on the achievable average throughput and a lower bound on the average maximum inter-decoding delay can be obtained by assuming that the transmitter is informed about the exact channel realization over all the M channel blocks non-causally. This allows the transmitter to optimally allocate the available resources among the messages. In particular, knowing the channels *a priori* the transmitter can choose the optimal subset S_{opt} of messages to be transmitted that maximizes $\overline{T}$ and minimizes $\overline{D}^{\max}$. Note that power allocation across channel blocks is not possible due to short-term power constraint. In order to find the set of messages S_{opt} that minimizes the average maximum delay, we first find the maximum number of decodable messages for the given channel realizations. It follows from the physically degraded broadcast channel model depicted in Fig. 1 that the total number of messages that can be decoded up to channel block t, denoted by $\Psi(t)$, $t = 1, \ldots, M$, is bounded as:

$$\Psi(t) \leq \min\left\{t, \left\lfloor \frac{I^{\mathrm{tot}}(t)}{R} \right\rfloor\right\}, \tag{4}$$

where $I^{\mathrm{tot}}(t) \triangleq \sum_{i=1}^{t} C_i$, is the total mutual information (MI) accumulated up to and including channel block t, while $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. At each channel block t, we check whether we can decode packet W_t in addition to the packets that have already been decoded. Note that there is no gain in decoding a packet prior to its decoding deadline. Let $v(t) \in \{0, 1\}$ denote the decoding event for W_t, i.e., $v(t) = 1$, if W_t is decoded, and $v(t) = 0$ if not. We have $\Psi(t) = v(1) + \cdots + v(t)$, and

$$v(t+1) = \begin{cases} 1 & \text{if } I^{\mathrm{tot}}(t+1) \geq (\Psi(t) + 1)R, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This recursion returns the M-length binary vector $\mathbf{V} = [v(1) \cdots v(M)]$, which corresponds to a transmission scheme
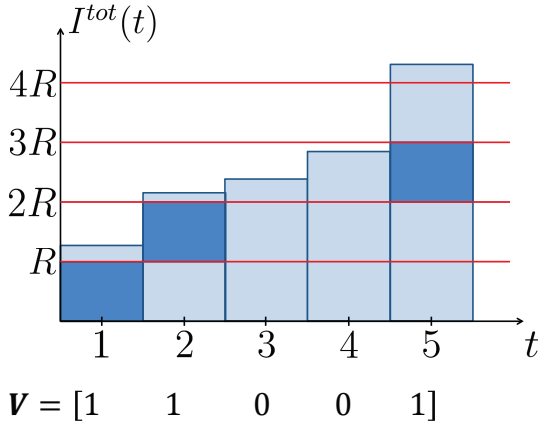
Fig. 2. $I^{tot}(t)$ plotted against t, and the corresponding vector $\mathbf{V}$ in case of throughput-optimal transmission. The light blue bars represent the amount of MI accumulated in each of the 5 channel blocks considered, while the dark blue rectangles indicate a decoding event and represent the amount of MI that is used to decode a message.
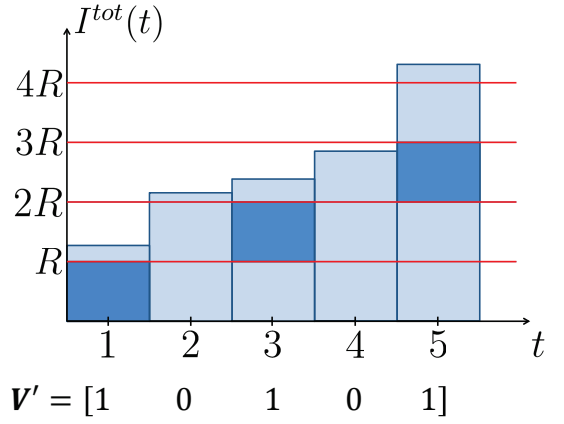

Fig. 3. $I^{tot}(t)$ plotted against t, and the corresponding vector $\mathbf{V}$ in case of throughput- and delay-optimal transmission. The light blue bars represent the amount of MI accumulated in each of the 5 channel blocks considered, while the dark blue rectangles indicate a decoding event and represent the amount of MI that is used to decode a message.

that maximizes the throughput. Although $\mathbf{V}$ represents an optimal solution in terms of $\overline{T}$, it may be suboptimal in terms of $\overline{D}^{\max}$. From the maximum delay perspective it may be a better choice not to transmit some of the packets even if enough mutual information could be accumulated by their deadlines, and instead to transmit packets that are further in the sequence. This is equivalent to shifting rightwards some of the 1's in $\mathbf{V}$ so that the number of consecutive 0's in the vector is minimized. Note that this process leaves the throughput unchanged.

Let us consider the example shown in Fig. 2, where the mutual information accumulated by the receiver at the end of channel block t, $I^{tot}(t)$ is plotted against the channel block number. The lines $I^{tot}(t) = jR$, $j = 1, \ldots, 4$, indicate the threshold values of $I^{tot}(t)$ after which a new message can be decoded. The vector $\mathbf{V}$ has entries equal to 1 in correspondence to decoding events (shadowed areas) and zero in correspondence to channel blocks in which the receiver does not decode the corresponding message.

With reference to Fig. 2, the iterative process described by Eqn. (5) returns the sequence $\mathbf{V} = [11001]$. This allocation achieves a throughput of $3/5$ and a maximum delay of 2. However, a better choice for the transmitter is to transmit message W_3 instead of W_2, as shown in Fig. 3. This gives the new allocation $\mathbf{V}' = [10101]$, which has the same throughput as $\mathbf{V}$ but a maximum delay of $D^{max} = 1$ instead of 2.

In order to minimize the maximum delay, the transmitter can choose to drop a message even if it could be decoded with high probability. In other words, the resources are allocated to a message with a higher index, which, if decoded, would lead to a lower maximum delay. Note that the maximum delay is optimized without decreasing the average throughput. Next we provide the necessary definitions and results to introduce the algorithm `Min_Del_Max_Rate`, which optimizes both $\overline{T}$ and $\overline{D}^{\max}$.

Definition 3.1: Let $\mathbf{V}_{\text{lb},D}$ denote the binary string of length M with maximum number of consecutive zeros equal to D, which has the smallest number of 1's and the smallest decimal representation.

If $M > D$, $\mathbf{V}_{\text{lb},D}$ can be constructed by taking a sequence of M zeros and starting from the $(D+1)$-th most significant bit (i.e., the leftmost one), substituting a 0 with a 1, every D bits. If $M = D$, $\mathbf{V}_{\text{lb},D}$ is the all-zero string of length M.

Let us clarify the definition considering an example with $M = 5$. To each value of D in the set $\{0, 1, 2, 3, 4, 5\}$ corresponds a different vector $\mathbf{V}_{\text{lb},D}$: $\mathbf{V}_{\text{lb},0} = [11111]$, $\mathbf{V}_{\text{lb},1} = [01010]$, $\mathbf{V}_{\text{lb},2} = [00100]$, $\mathbf{V}_{\text{lb},3} = [00010]$, $\mathbf{V}_{\text{lb},4} = [00001]$ and $\mathbf{V}_{\text{lb},5} = [00000]$.

Definition 3.2: We define $\Psi(t) = \sum_{n=1}^{t} v(n)$ and $\Psi_{\text{lb},D}(t) = \sum_{n=1}^{t} v_{\text{lb},D}(n)$, where $v(n)$ and $v_{\text{lb},D}(n)$ are the n-th bits, starting from the most significant ones, of $\mathbf{V}$ (tentative allocation vector returned by recursion (5)) and $\mathbf{V}_{\text{lb},D}$ (see Definition 1), respectively. In other words, $\Psi(t)$ and $\Psi_{\text{lb},D}(t)$ are the cumulative sum, from left, of the vectors $\mathbf{V}$ and $\mathbf{V}_{\text{lb},D}$, respectively, up to the t-th coordinate.

With reference to the example in Fig. 2, we have $\Psi(1), \ldots, \Psi(5) = 1, 2, 2, 2, 3$. For $D = 2$, we have $\mathbf{V}_{\text{lb},2} = [00100]$, and $\Psi_{\text{lb},2}(1), \ldots, \Psi_{\text{lb},2}(5) = 0, 0, 1, 1, 1$.

Theorem 1 Given the allocation vector $\mathbf{V}$ returned by recursion (5), a maximum delay less than or equal to D^* is achievable if the following holds: $\Psi(t) \geq \Psi_{\text{lb},D^*}(t)$, $\forall t \in \{1, \ldots, M\}$.

Proof We recall that $\Psi_{\text{lb},D}(t)$ is the total number of 1's among the leftmost t bits of the sequence $\mathbf{V}_{\text{lb},D}$ (see Definition 1), while $\Psi(t)$ is the total number of 1's among the leftmost t bits of the sequence $\mathbf{V}$. $\Psi(t) \geq \Psi_{\text{lb},D}(t)$, $\forall t \in \{1, \ldots, M\}$, implies that $\mathbf{V}$ has at least as many 1's

as $\mathbf{V}_{\text{lb},D}$ among the leftmost t positions, $\forall t \in \{1, \ldots, M\}$, which, in turn, implies that $\mathbf{V}$ achieves a maximum delay that is no greater than D^*, which concludes the proof.

In order to find the minimum possible maximum delay starting from a given sequence $\mathbf{V}$, one can start with a delay $D^* = 0$ and check if the condition of Theorem 1 is satisfied. If not, the maximum delay is increased by 1, and so on.

Algorithm 1 `Min_Del_Max_Rate(`$\mathbf{V}$`)`

$M = \text{length}(\mathbf{V})$
if $\mathbf{V}$ $==$ $[0, \ldots, 0]$ **then**
  // if no packet can be decoded return the all zero sequence
    $\mathbf{S}_{\text{opt}} = [0, \ldots, 0]$
    return $\mathbf{S}_{\text{opt}}$
end if
$D, k = 0$
while found $== 0$ **do**
    found $= 1$
    $\mathbf{V}_{\text{lb},D} = [0, \ldots, 0]$ // vector of M zeros
    for $i = 1$ to $\left\lfloor \frac{M}{D+1} \right\rfloor$ **do**
        $\mathbf{V}_{\text{lb},D}[i(D + 1)]$ $=$ 1
  // assign 1 to the $i(D+1)$-th component
    end for
    cumsum_d $= 0$
    cumsum_lb $= 0$
    for $j = 1$ to M **do**
        cumsum_d $=$ cumsum_d$+\mathbf{V}[j]$ // calculate $\Psi(j)$
        cumsum_lb $=$ cumsum_lb$+\mathbf{V}_{\text{lb}}[j]$
  // calculate $\Psi_{\text{lb},D}(j)$
        if cumsum_d $<$ cumsum_lb **then**
  // if cumulative sum is lower, start again increasing delay
            found $= 0$
            exit for
        end if
    end for
    if found $== 1$ **then**
        $D_{\text{IT}}^{\max} = D$
        exit while
    end if
    $D = D + 1$
end while
$\mathbf{S}_{\text{opt}} = \mathbf{V}_{\text{lb},D_{\text{IT}}^{\max}}$
excess_0 $= \text{sum}(\mathbf{V}) - \text{sum}(\mathbf{V}_{\text{lb},D})$
while k $<$ excess_0 **do**
  // assign 1 to the rightmost excess _0 zeros of $\mathbf{V}_{\text{lb},D_{\text{IT}}^{\max}}$
    if $\mathbf{S}_{\text{opt}}[M - k] == 0$ **then**
        $\mathbf{S}_{\text{opt}}[M - k] = 1$
        $k = k + 1$
    end if
end while
return $\mathbf{S}_{\text{opt}}$

Using Theorem 1, the `Min_Del_Max_Rate` algorithm (Algorithm 1) has been obtained. The algorithm takes as input the vector $\mathbf{V}$, which is obtained using the recursion in Eqn. (5). First the algorithm calculates the minimum achievable maximum delay $D_{\text{IT}}^{\max}$ (see Theorem 1 and the following note) and derives the vector $\mathbf{V}_{\text{lb},D_{\text{IT}}^{\max}}$. Then it calculates the difference in the number of ones between $\mathbf{V}$ and $\mathbf{V}_{\text{lb},D_{\text{IT}}^{\max}}$ (excess_0 in the algorithm). By definition of $D_{\text{IT}}^{\max}$, excess_0 is greater

than or equal to zero. Using $\mathbf{V}_{\text{lb},D_{\text{IT}}^{\max}}$ as an initialization allocation vector, the vector $\mathbf{S}_{\text{opt}}$ is then constructed by simply substituting the rightmost excess_0 zeros with ones. The output of the algorithm is the set of messages $\mathbf{S}_{\text{opt}}$ (containing a 1 or a 0 in position t if message W_t is to be transmitted, or not) that constitutes the optimal transmission choice in terms of both throughput and maximum delay. It can be easily shown that Algorithm 1 has a complexity which is quadratic in M.

In order to clarify the procedure just described, let us consider again the example in Fig. 2. The recursion in Eqn. (5) returns the vector $\mathbf{V} = [11001]$, which corresponds to $\Psi = [12223]$. The algorithm starts with a tentative delay $D_{\text{IT}}^{\max} = 0$, and generates the corresponding sequence $\mathbf{V}_{\text{lb},0} = [11111]$, with $\Psi_{\text{lb},0} = [12345]$. Since the condition of Theorem 1 is not satisfied ($\Psi(3) < \Psi_{\text{lb},0}(3)$), a minimum maximum delay $D_{\text{IT}}^{\max} = 0$ cannot be achieved, and the tentative delay is increased by 1, i.e., $D_{\text{IT}}^{\max} = 1$. The corresponding sequences $\mathbf{V}_{\text{lb},1} = [01010]$ and $\Psi_{\text{lb},1} = [01122]$ are then calculated. The cumulative function $\Psi_{\text{lb},1}$ satisfies the condition of Theorem 1, which implies that the minimum achievable maximum delay is $D_{\text{IT}}^{\max} = 1$. At this point the algorithm calculates the optimal allocation vector. First, the difference in the number of ones between vector $\mathbf{V}_{\text{lb},1}$ and vector $\mathbf{V}$ (excess_0) is computed, which in the example is equal to excess_0=1. Finally, the rightmost excess_0 zeros in $\mathbf{V}_{\text{lb},1}$ are set to 1, which leads to the allocation sequence $\mathbf{S}_{\text{opt}} = [01011]$.

IV. TRANSMISSION SCHEMES

In this section we introduce four different transmission schemes based on time-sharing. Each channel block is divided among the messages for which the deadline has not yet expired. Thus, while the first channel block is divided among all the messages $W_1, \ldots, W_M$, the second channel block is divided among messages $W_2, \ldots, W_M$, as the deadline of message W_1 expires at the end of the first block. In general the encoder divides channel block t into $M - t + 1$ portions $\alpha_{tt}, \ldots, \alpha_{Mt}$, such that $\alpha_{mt} \geq 0$ and $\sum_{m=t}^{M} \alpha_{mt} = 1$. In channel block t, $\alpha_{mt}n$ channel uses are allocated for the transmission of message W_m. We assume that Gaussian codebooks are used in each portion for each message, and the corresponding codelengths are sufficient to achieve the instantaneous capacity. Then the total amount of received mutual information relative to message W_m is:

$$I_m^{\text{tot}} \triangleq \sum_{t=1}^{m} \alpha_{mt} C_t. \tag{6}$$

The proposed schemes differ in the way the channel uses are allocated among the messages for which the deadline has not yet expired. Different time allocations lead to different average throughput and average maximum delay performances.

A. Memoryless Transmission (MT)

In *memoryless transmission (MT)* each message is transmitted only within the channel block just before its expiration, that is, message W_t is transmitted over channel block t. Equivalently we have $\alpha_{mt} = 1$, if $t = m$, and $\alpha_{mt} = 0$,

otherwise. In MT message W_t can be decoded if and only if $C_t \geq R$. Due to the i.i.d. nature of the channel state over blocks, the successful decoding probability $p \triangleq Pr\{C_t \geq R\}$ is constant over messages. The probability that exactly m messages are decoded is given by:

$$\eta(m) \triangleq \binom{M}{m} p^m (1-p)^{M-m}. \tag{7}$$

The average number of decoded messages for the MT scheme is $\overline{T}_{MT} = Mp$.

Next we derive the exact expression for the average maximum delay for MT, denoted by $\overline{D}_{MT}^{\max}$. The term $Pr\{D^{\max} \geq d\}$ in the summation in Eqn. (3) is the probability that a sequence of M Bernoulli random variables with parameter p contains at least d consecutive zeros. This probability can be evaluated by modeling the number of consecutive zeros as a Markov chain, and finding the probability of reaching the final absorbing state of d consecutive zeros. This probability is given in the following theorem:

Theorem 2: Let $x_1, \cdots, x_M$ be a sequence of i.i.d. Bernoulli random variables with parameter $p = \mathrm{E}[x_i]$. The probability of having at least d consecutive zeros in the sequence is given by:

$$Pr\{D^{\max} \geq d\} = \sum_{i=0}^{k} \sum_{r_i=1}^{s_i} a_{d,r_i} \binom{M + r_i - 1}{r_i - 1} \left(\frac{1}{\varphi_{di}}\right)^M, \tag{8}$$

where $k \in \{0, \ldots, M\}$, $k \leq d+1$ is the number of distinct zeros of the polynomial $(1-z)q_d(z)$ where:

$$q_d(z) = 1 - p \sum_{j=1}^{d} z^j (1-p)^{j-1}, \tag{9}$$

φ_{di}, $i \in \{0, \ldots, k\}$, are the zeros of $(1-z)q_d(z)$ with multiplicity s_i, a_{d,r_i}, $r_i \in \{1, \ldots, s_i\}$, are constants derived from the partial fraction expansion of

$$\frac{(zp)^d}{(1-z)q_d(z)}. \tag{10}$$

Proof: See Appendix.

Finally, by plugging (8) into (3) we find:

$$\overline{D}_{MT}^{\max} = \sum_{d=1}^{M} \left[\sum_{i=0}^{k} \sum_{r_i=1}^{s_i} a_{d,r_i} \binom{M + r_i - 1}{r_i - 1} \left(\frac{1}{\varphi_{di}}\right)^M \right]. \tag{11}$$

B. Equal Time-Sharing (eTS) Transmission

In the equal time-sharing (eTS) transmission scheme each channel block is equally divided among all the messages whose deadline has not expired yet, that is, for $m = 1, \ldots, M$, we have $\alpha_{mt} = \frac{1}{M-t+1}$ for $t = 1, \ldots, m$, and $\alpha_{mt} = 0$, for $t = m+1, \ldots, M$.

In eTS, messages whose deadlines are later in time are allocated more resources; and hence, are more likely to be decoded. We have $I_i^{\mathrm{tot}} < I_j^{\mathrm{tot}}$ for $1 \leq i < j \leq M$. Hence, the probability of decoding exactly m messages is:

$$\eta(m) \triangleq Pr\{I_m^{\mathrm{tot}} \geq R \geq I_{m-1}^{\mathrm{tot}}\}, \tag{12}$$

for $m = 0, 1, \ldots, M$, where we define $I_0^{\mathrm{tot}} = 0$ and $I_{M+1}^{\mathrm{tot}} = \infty$. Since the decoded messages in eTS are always the last ones, we can express the average maximum delay of eTS, $\overline{D}_{\mathrm{eTS}}^{\max}$, as a function of its average throughput $\overline{T}_{\mathrm{eTS}}$ as follows:

$$\overline{D}_{\mathrm{eTS}}^{\max} \triangleq \sum_{m=0}^{M} (M-m) \cdot \eta(m)$$

$$= \sum_{m=0}^{M} M \cdot \eta(m) - \sum_{m=0}^{M} m \cdot \eta(m)$$

$$= M \left(1 - \frac{\overline{T}_{\mathrm{eTS}}}{R}\right). \tag{13}$$

The numerical analysis of eTS, together with other schemes is presented in Section V.

C. Pre-Buffering (PB) Transmission

In most practical streaming systems the receiver first accumulates GOPs in the playout buffer and then starts displaying them at a constant frame rate once a sufficient portion of the video has been received, in order to compensate for the delay jitter of arriving packets [35]. We consider a slightly different version of this type of streaming transmission in which only the last B messages are transmitted while the first packets are not transmitted at all. The first $M - B + 1$ channel blocks are used to convey information relative to the last B packets as explained in the following. We call this method *pre-buffering (PB)* transmission.

The initial buffering phase introduces a start-up delay of $M - B$ channel blocks. On the other hand, if a sufficiently large buffering period is chosen, all the transmitted messages can be received correctly, achieving an average throughput of $R\frac{B}{M}$. Transmitted messages are encoded with equal time allocation over the first $M - B + 1$ blocks. Due to the delay constraint, message W_{M-B+1} is transmitted up to channel block $M - B + 1$. Hence, in block $M - B + 2$ the last $B - 1$ messages are transmitted with equal time allocation. The process continues up until channel block M, in which only message W_M is transmitted. Next we indicate with $\overline{T}_{\mathrm{PB}}(B)$ and $\overline{D}_{\mathrm{PB}}^{\max}(B)$ the average throughput and the average maximum delay achieved by the scheme using a buffering period of B channel blocks, respectively. The number B_{opt} of messages to be transmitted is chosen so that

$$B_{\mathrm{opt}} = \underset{B \in \{1, \ldots, M\}}{\arg\min} \left\{\overline{D}^{\max}(B)\right\}. \tag{14}$$

Next we show that the B_{opt}, as defined in Eqn. (14), also maximizes the average throughput. The average throughput

when transmitting only the last B messages is given by:

$$\overline{T}_{\text{PB}}(B) = \frac{R}{M} \sum_{m=1}^{B} Pr\,\{\text{decode at least } m \text{ messages}\}$$

$$= \frac{R}{M} \sum_{m=1}^{B} Pr\,\left\{ I_{M-m+1}^{\text{tot}} \geq R \right\}, \qquad (15)$$

where the mutual information accumulated by the receiver for message W_m, for $m = M - B + 1, M - B + 2, \ldots, M$, is given by:

$$I_m^{\text{tot}} = \frac{1}{B} \sum_{t=1}^{M-B+1} C_t + \sum_{t=M-B+2}^{m} \frac{C_t}{M-t+1}. \qquad (16)$$

From Eqn. (15) we have:

$$\overline{T}_{\text{PB}}(B) = \frac{R}{M}\left[B - \sum_{m=1}^{B} Pr\,\left\{ I_{M-m+1}^{\text{tot}} < R \right\} \right]$$

$$= \frac{R}{M}\left[B - \sum_{m=1}^{B} Pr\,\left\{ D^{\text{max}} \geq M - m + 1 \right\} \right] (17)$$

The average maximum delay when only the last B messages are transmitted is:

$$\overline{D}_{\text{PB}}^{\text{max}}(B) = M - B + \sum_{d=1}^{B} Pr\,\left\{ D^{\text{max}} \geq M - B + d \right\} (18)$$

From (17) and (18) we find

$$\overline{T}_{\text{PB}}(B) = R\left(1 - \frac{\overline{D}^{\text{max}}(B)}{M} \right),$$

and finally

$$\underset{B \in \{1, \cdots, M\}}{\arg\min} \left\{ \overline{D}_{\text{PB}}^{\text{max}}(B) \right\} = \underset{B \in \{1, \cdots, M\}}{\arg\max} \left\{ \overline{T}_{\text{PB}}(B) \right\}. \qquad (19)$$

This proves that the average throughput and the maximum delay can be optimized simultaneously. It is not straightforward to come up with an analytical expression for the optimal value of B in the PB scheme for the general case. In the following theorem we derive the optimal fraction of messages $\alpha_{\text{opt}} = \frac{B_{\text{opt}}}{M}$, such that almost all of the transmitted messages can be decoded with probability that approaches 1 asymptotically as M goes to infinity, if a fraction $\alpha' < \alpha_{\text{opt}}$ of the messages is transmitted, while a fraction smaller than α_{opt} of the messages can be decoded if $\alpha' > \alpha_{\text{opt}}$.

Theorem 3 Average throughput of αR can be achieved in the limit of infinite M by transmitting $\alpha M + o(M)$ messages as long as

$$\alpha < \alpha_{\text{opt}} \triangleq \frac{1}{\frac{R}{\overline{C}} + 1}.$$

If $\alpha > \alpha_{\text{opt}}$, the achieved average throughput is smaller than $\alpha_{\text{opt}} R$.

Proof Assume that the last B messages, i.e., $W_{M-B+1}, \ldots, W_M$, are transmitted, with $B = M\alpha + o(M)$, $\alpha \leq 1$. Message W_{M-B+1}, for which the deadline expires

first, is the one that accumulates the least amount of mutual information, that is:

$$I_{M-B+1} = \frac{1}{B} \sum_{t=1}^{M-B+1} C_t. \qquad (20)$$

The probability of decoding all the transmitted messages is then:

$$Pr\,\{I_{M-B+1} \geq R\} = Pr\,\left\{ \frac{1}{B} \sum_{t=1}^{M-B+1} C_t \geq R \right\}$$

$$= Pr\,\left\{ \sum_{t=1}^{M-B+1} \frac{C_t}{M-B+1} - \overline{C} \geq \frac{B}{M-B+1} R - \overline{C} \right\}$$

$$= Pr\,\left\{ S_{M-B+1} - \overline{C} \geq \frac{B}{M-B+1} R - \overline{C} \right\}, \qquad (21)$$

where $S_{M-B+1} \triangleq \sum_{t=1}^{M-B+1} \frac{C_t}{M-B+1}$, is the sample mean of the instantaneous channel capacity over the first $M - B + 1$ channel blocks. From the law of large numbers it follows that:

$$\lim_{M \to \infty} Pr\,\left\{ \left| S_{M\left(1-\alpha-\frac{o(M)}{M}\right)} - \overline{C} \right| > \delta \right\} = 0, \ \forall \delta > 0. \qquad (22)$$

Using equations (21) and (22) we find:

$$\lim_{M \to \infty} Pr\,\{I_{M-B+1} \geq R\} = \begin{cases} 1, & \text{if } \lim_{M \to \infty} \frac{B}{M-B+1} R < \overline{C} \\ 0, & \text{if } \lim_{M \to \infty} \frac{B}{M-B+1} R > \overline{C}. \end{cases} \qquad (23)$$

We can write:

$$\lim_{M \to \infty} \frac{B}{M-B+1} R = \lim_{M \to \infty} \frac{M\alpha + o(M)}{M - M\alpha + o(M)} R$$

$$= \frac{\alpha}{1-\alpha} R. \qquad (24)$$

Finally, using Eqn. (24) in Eqn. (23) we find:

$$\lim_{M \to \infty} Pr\,\{I_{M-B+1} \geq R\} = \begin{cases} 1, & \text{if } \alpha < \alpha_{\text{opt}} \\ 0, & \text{if } \alpha > \alpha_{\text{opt}}. \end{cases} \qquad (25)$$

Eqn. (25) implies that if a fraction of messages α' larger than α_{opt} is transmitted, then the average throughput is less than $\alpha_{\text{opt}} R$, which concludes the proof.

In Section V, we provide a numerical optimization of the PB scheme, and compare it with the other proposed transmission strategies and the upper bound. As we will see from the numerical results, this buffering approach can improve the average throughput significantly as it provides rate adaptation at the packet level by eliminating some of the packets, thus increasing the correct decoding probability of the remaining packets.

D. Windowed Time Sharing (wTS)

We have seen in the PB scheme that transmitting only a subset of the messages can improve the system throughput by allowing rate adaptation at the packet level. However, in the PB scheme only the last B packets are transmitted leading to a minimum delay of $M - B$ channel blocks. In the next scheme, called the windowed time-sharing (wTS) scheme, $\lceil M/B \rceil$ messages are transmitted, where $\lceil x \rceil$ is the smallest integer greater than or equal to x; however, unlike in PB,

the transmitted messages are distributed among the whole set of available messages, that is, only one from B consecutive packets is transmitted over B consecutive channel blocks. So, for instance, if $B = 3$, the first message to be transmitted is W_3, which is repeated over channel blocks 1, 2 and 3, followed by message W_6, which is transmitted in the next three channel blocks, and so on.

The parameter B can be optimized according to two different criteria, namely to maximize the average throughput or to minimize the delay, which leads to the two variants of the wTS scheme, which we call *throughput-wTS (T-wTS)* and *delay-wTS (D-wTS)*, respectively. In wTS a message is decoded with probability p_B given below:

$$p_B = Pr\left\{I_{kB} \geq R\right\} = Pr\left\{\sum_{t=kB-W+1}^{\min\{kB,M\}} C_k \geq R\right\}, \quad (26)$$

for $k \in \{1, \ldots, \lceil \frac{M}{B} \rceil\}$. A lower bound on $\overline{D}_{wTS}^{\max}$ can be found by substituting $\lfloor \frac{M}{B} \rfloor$ for M in Eqn. (11), p_B for p in equations (9) and (10) and multiplying Eqn. (11) with B. An upper bound can be found in a similar way by using $\lceil \frac{M}{B} \rceil$ instead of $\lfloor \frac{M}{B} \rfloor$. Similarly, an upper and a lower bound on $\overline{T}_{wTS}$ are given by $\lceil \frac{M}{B} \rceil \cdot p_B$ and $\lfloor \frac{M}{B} \rfloor \cdot p_B$, respectively. Analytical optimization of parameter B in both the T-wTS and D-wTS schemes is elusive and we resort to the numerical analysis presented in the next section.

V. NUMERICAL RESULTS

In this section we compare the average throughput and the average maximum delay of the proposed schemes numerically. The channel model used in the simulations is a Rayleigh block fading channel, in which the channel gain $\phi[t]$ in block number t, $t = 1, \ldots, M$ (see Eqn. 1) is a unit-mean exponential random variable that changes in an i.i.d. fashion at the beginning of each channel block and stays constant until the beginning of the next one. Fig. 4 and Fig. 5 show the average throughput and the average maximum delay for the proposed schemes, respectively, for $R = 1$ and SNR $= -5$ dB. Both variants of the wTS scheme perform close to the informed transmitter lower bound in terms of the maximum delay, while the PB scheme is the one with the highest average throughput, followed by T-wTS and D-wTS. The eTS scheme shows quite poor performance in terms of both the delay and the throughput. From the plots it emerges that wTS in its two variants T-wTS and D-wTS, can help to reduce the inter-decoding delay while achieving a relatively good average throughput in the low SNR regime. The transmitter can choose between the two schemes based on its preference between higher throughput and lower inter-decoding delay. While PB provides the highest throughput among the proposed schemes, its inter-decoding delay is significantly high, due to the initial buffering time. PB might be a particularly attractive choice for video streams of long duration, for which the users would be willing to have a larger startup delay to enjoy a higher throughput for the rest of the video.
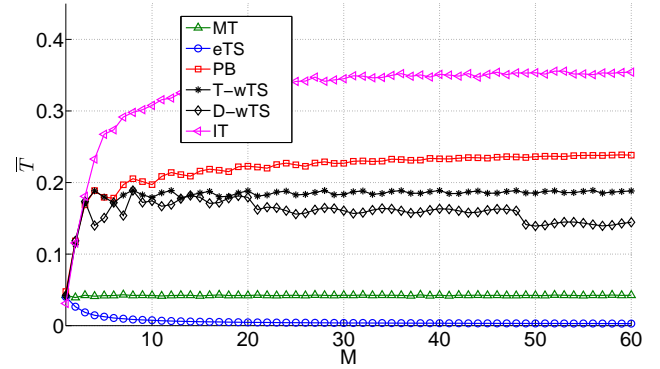

Fig. 4. Average throughput $\overline{T}$ plotted against the number of messages transmitted for $SNR = -5$ dB and $R = 1$ bpcu.
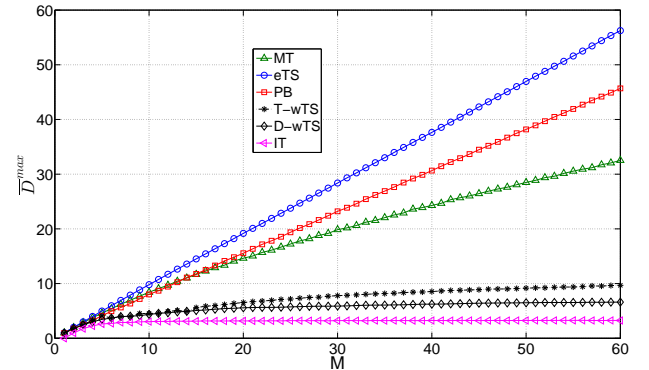

Fig. 5. Average maximum delay $\overline{D}^{\max}$ plotted against the number of transmitted messages for $SNR = -5$ dB and $R = 1$ bpcu.

Fig. 6 and Fig. 7 show the average throughput and the average maximum delay, respectively, for the proposed schemes for $R = 1$ and SNR $= 5$ dB. Also for this SNR level the two variants of the wTS scheme perform close to the informed transmitter lower bound in terms of maximum delay. The highest average throughput is achieved by the T-wTS scheme together with the MT scheme, followed by the PB, D-wTS and eTS schemes. From Fig. 6 and Fig. 7 we see that, when the SNR is high, the MT scheme, together with the T-wTS scheme, achieves the best performances in terms of both delay and average throughput. This suggests that a simple memoryless approach is sufficient when the channel SNR is sufficiently high, while at low SNR more complex encoding techniques can help to significantly improve the performance. The D-wTS scheme shows a sudden decrease in the average throughput, which, with reference to Fig. 6, also corresponds to a decrease in the slope of the curve at points corresponding to $M = 7$ and $M = 48$. This is due to the optimization of the window size B. We recall that in D-wTS the window size represents the number of channel blocks dedicated to a message, and is optimized so as to achieve the minimum average maximum delay. While a large B leads to a high decoding probability, it implies a small
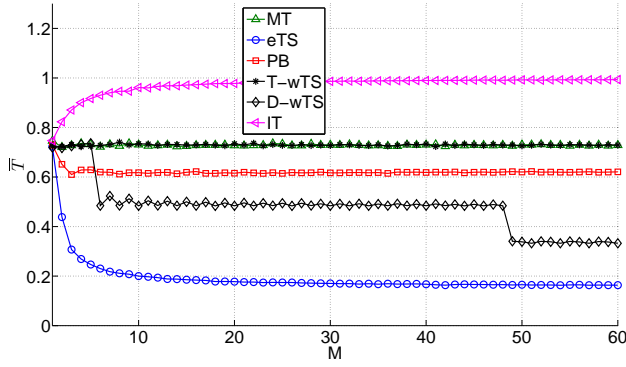
Fig. 6. Average throughput $\overline{T}$ plotted against the number of messages transmitted for $SNR = 5$ dB and $R = 1$ bpcu.
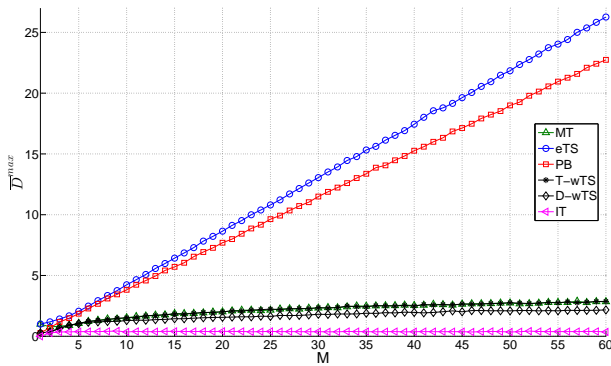

Fig. 8. Optimal window size (B) for the T-wTS scheme plotted versus the total number of messages (M) for $SNR = 5$ dB.

Fig. 7. Average maximum delay $\overline{D}^{max}$ plotted against the number of transmitted messages for $SNR = 5$ dB and $R = 1$ bpcu.

Fig. 9. Average throughput $\overline{T}$ plotted against the SNR for $M = 40$ packets and $R = 1$ bpcu.

number of transmitted messages, which bounds from below the minimum delay by B. As a matter of fact, only $\lceil \frac{M}{B} \rceil$ messages are transmitted in the wTS scheme, which implies that the maximum delay, in a given realization, is a multiple of B. If, for instance, $B = 2$ and $m = 3$ consecutive messages are lost, the corresponding delay is $m \cdot B = 6$. Formally, given a window size B^* there is a certain probability $p_{B^*}^l$ of not decoding a message. For any fixed $m \in \{0, \dots, M\}$, using Eqn. (8) it can be easily shown that the probability of losing at least m consecutive messages increases with M. Thus a value B^* which is optimal for a certain M, may not be the optimal for a larger number of messages, as the probability that more than one consecutive messages get lost increases with M. The optimal choice may be to increase B, so that the probability of losing consecutive messages is decreased. This is confirmed by Fig. 8, where the optimal window size, obtained numerically, is plotted against the total number of messages. An increase in B implies a decrease in the slope of the average number of decoded messages, since a smaller fraction of messages is transmitted, as shown in the plots. The T-wTS scheme, in which B is optimized so as to achieve the maximum average throughput, shows a good tradeoff between the average throughput, which, unlike D-
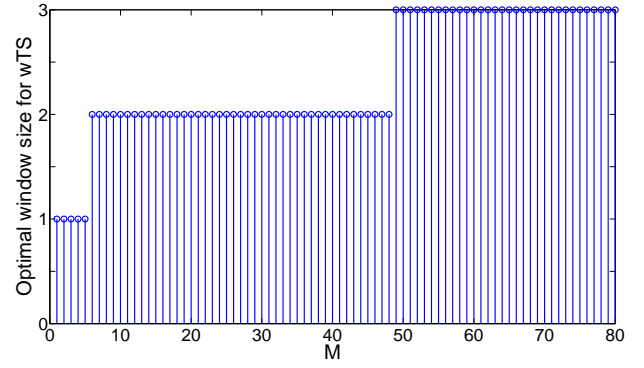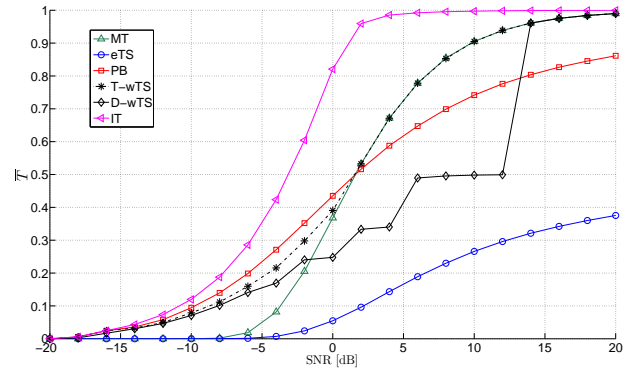
wTS, is almost independent of the number of messages, and the average maximum delay, performing close to the D-wTS scheme.

In Figures 9 and 10 the average throughput and the average maximum delay, respectively, are plotted against average SNR. The plots were obtained for $M = 40$ packets and
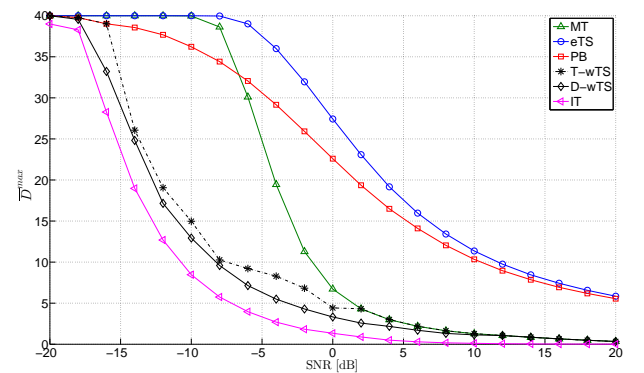

Fig. 10. Average maximum delay $\overline{D}^{max}$ plotted against the SNR for $M = 40$ packets and $R = 1$ bpcu.

$R = 1$ bpcu. As observed in Figures 4 and 6, for $M = 40$, the PB scheme outperforms all other schemes in terms of throughput at low SNR (lower than 2 dB), while T-wTS and MT achieve almost the same performance, and outperform the PB scheme at higher SNRs. From the figures we observe that the PB scheme is the most robust one against packet losses at low SNR, while at higher SNR it is outperformed by all the schemes but the trivial MT. In terms of maximum delay, PB shows relatively poor performance for most of the considered SNR range, which is due to the initial buffering phase. Note that, if, unlike assumed in this paper, the loss of large consecutive chunks of the content were not an issue, and choppy playback were to be avoided, the PB scheme would be the best among the considered schemes since it guarantees that, once the buffering phase is finished, no additional packet is lost, as proven in Theorem 3 for the asymptotic case.

VI. CONCLUSIONS

We have studied the streaming of stored video content over slow fading channels with per-packet delay constraints. In addition to the classical throughput metric, we have also considered the inter-decoding delay, i.e., the number of consecutive video GOPs that cannot be decoded successfully, as a performance measure. We have proposed four different transmission schemes based on time-sharing. We have carried out theoretical as well as numerical analysis for the average throughput and maximum delay performances. We have also derived bounds on both the average throughput and maximum inter-decoding delay by introducing an informed transmitter bound, in which the transmitter is assumed to know the channel states in advance. We have seen that the wTS scheme can provide a good trade-off between the average throughput and the maximum inter-decoding delay by deciding on the proportion of transmitted video packets. In practice this corresponds to reducing the coding rate of the video at the packet level. We have also proved that in the PB scheme almost all transmitted messages can be decoded with a probability that goes to 1 as M goes to infinity if only a fraction of the messages smaller than a threshold value, which depends on the transmission rate and the average channel capacity, are transmitted.

APPENDIX

Proof of Theorem 1

The probability of having a run of at least d, $d \in \{0, \ldots, M\}$, consecutive zeros in the sequence is equivalent to finding the probability of state d after M steps in the Markov chain depicted in Fig. 11. The state d is an absorbing state, i.e., once the process reaches that state, it remains there with probability 1. Let $\mathbf{p}_t$ be a d-length probability mass function, where $\mathbf{p}_t(i)$, $i = 0, \ldots, d$, denotes the probability of being in state i at step t. The vector $\mathbf{p}_t$ of state occupancy at step t for the Markov chain in Fig. 11 can be obtained as:

$$\mathbf{p}_t = \mathbf{p}_{t-1}\mathbf{H} = \mathbf{p}_0\mathbf{H}^t, \tag{27}$$
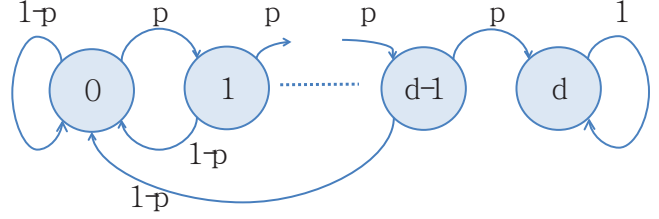

Fig. 11. Markov chain for the calculation of the average maximum delay in memoryless transmission.

where $\mathbf{p}_0 = [1\ 0\ \cdots\ 0]$ and $\mathbf{H}$ is the $(d+1)\times(d+1)$ transition matrix of the chain which has the following structure:

$$\mathbf{H} = \begin{pmatrix} 1-p & p & 0 & 0 & \cdots & 0 & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & & & \vdots \\ 1-p & 0 & 0 & 0 & \cdots & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{28}$$

The probability of being in state d after M steps, $\mathbf{p}_M(d)$, can be found from Eqn. (27). Since $\mathbf{p}_0 = [1\ 0\ \cdots\ 0]$ we have:

$$\mathbf{p}_M(d) = \mathbf{H}^M(1, d+1). \tag{29}$$

In order to evaluate $\mathbf{H}^M(1, d+1)$, we apply the *Z-transform* to Eqn. (27), taking into account that the recursive formula is defined only for $t \geq 1$. The Z-transform $\mathcal{P}(z)$ of a discrete vector function $\mathbf{p}_t$ is defined as [36]:

$$\mathcal{P}_z \triangleq \mathcal{Z}(\mathbf{p}_t) = \sum_{t=0}^{+\infty} \mathbf{p}_t z^t. \tag{30}$$

To account for the fact that $t \geq 1$ in Eqn. (27) we can write:

$$\sum_{t=1}^{+\infty} \mathbf{p}_t z^t = \sum_{t=0}^{+\infty} \mathbf{p}_t z^t - \mathbf{p}_0 = \mathcal{P}_z - \mathbf{p}_0, \tag{31}$$

and

$$\begin{aligned} \sum_{t=1}^{+\infty} \mathbf{p}_{t-1}\mathbf{H} z^t &= z\sum_{t=1}^{+\infty} \mathbf{p}_{t-1}\mathbf{H} z^{t-1} \\ &= z\sum_{t=0}^{+\infty} \mathbf{p}_t\mathbf{H} z^t \\ &= z\mathcal{P}_z\mathbf{H}. \end{aligned} \tag{32}$$

Plugging Eqn. (31) and Eqn. (32) into Eqn. (27) we find:

$$\mathcal{P}_z = \mathbf{p}_0\left(\mathbf{I} - z\mathbf{H}\right)^{-1}, \tag{33}$$

where $\mathbf{I}$ is the $(d+1) \times (d+1)$ identity matrix.

The Z-transform $\mathcal{C}_z$ of a matrix $\mathbf{C}_t$ of functions in the discrete variable t is defined as:

$$\mathcal{C}_z \triangleq \mathcal{Z}(\mathbf{C}_t) = \sum_{t=0}^{+\infty} \mathbf{C}_t z^t. \tag{34}$$

Note that in Eqn. (34) the term z^t is a scalar function of z and t which is multiplied to each of the elements of matrix $\mathbf{C}_t$. By comparing Eqn. (33) with Eqn. (27), we see that $(\mathbf{I} - z\mathbf{H})^{-1}$

is the Z-transform of the matrix $\mathbf{H}^t$ having functions in the discrete variable t as elements. We have:

$$\mathbf{I} - z\mathbf{H} = \begin{pmatrix} 1-z(1-p) & -zp & 0 & 0 & \cdots & 0 & 0 \\ -z(1-p) & 1 & -zp & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & & \vdots \\ -z(1-p) & 0 & 0 & 0 & \cdots & 1 & -zp \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-z \end{pmatrix} \tag{35}$$

Once $(\mathbf{I} - z\mathbf{H})^{-1}$ is known, it is sufficient to inversely transform it and get $\mathbf{H}^t$. We find the inverse of matrix (35) for a generic d using Gauss-Jordan elimination. As we only need the element $\mathbf{H}^M(1, d+1)$, we only report the first row of $(\mathbf{I} - z\mathbf{H})^{-1}$ in Eqn. (36) at the top of the next page, where

$$q_d(z) \triangleq 1 - p\sum_{j=1}^{d} z^j (1-p)^{j-1}. \tag{37}$$

The probability of being in state d at step M is the inverse Z-transform of element $(1, d+1)$ of matrix $(\mathbf{I} - z\mathbf{H})^{-1}$, i.e.:

$$\mathbf{p}_M(d+1) = \mathcal{Z}^{-1}\left\{ \frac{(zp)^d}{(1-z)q_d(z)} \right\}_{t=M}, \tag{38}$$

where $\mathcal{Z}^{-1}\{\mathcal{P}_z\}$ is the inverse Z-transform of $\mathcal{P}_z$ defined as [36]:

$$\mathcal{Z}^{-1}\{\mathcal{P}_z\} = \frac{-1}{2\pi j} \oint_\gamma \mathcal{P}_z z^{-t-1} dz = \mathbf{p}_t, \tag{39}$$

γ being a counterclockwise-oriented circle around the origin of the complex plane. An easier way to solve Eqn. (38) is to decompose the Z-transform using partial fraction decomposition, i.e., rewriting $\mathcal{P}_z$ as:

$$\mathcal{P}_z = \frac{(zp)^d}{(1-z)q_d(z)} = \sum_{i=0}^{k}\sum_{r_i=1}^{s_i} a_{d,r_i}\left(\frac{1}{1-\frac{z}{\varphi_{d,i}}}\right)^{r_i}, \tag{40}$$

where $\varphi_{d,i}$, $i \in \{0, \ldots, k\}$, are the $k \le d+1$ distinct zeros with degree $d+1$ and multiplicity s_i of the polynomial $(1-z)q_d(z)$, while a_{d,r_i}, $r_i \in \{1, \ldots, s_i\}$, are constants deriving from the partial fraction expansion of $\mathcal{P}_z$. Once in the form of Eqn. (40), $\mathcal{P}_z$ can be inversely transformed using the linearity of the inverse Z-transform and the fact that:

$$\mathcal{Z}^{-1}\left\{\left(\frac{1}{1-\frac{z}{\varphi_{d,i}}}\right)^{r_i}\right\} = \binom{t+r_i-1}{r_i-1}\left(\frac{1}{\varphi_{d,i}}\right)^t. \tag{41}$$

Eqn. (41) follows from the fact that:

$$\mathcal{Z}\left\{\left(\frac{1}{\varphi}\right)^t\right\} \triangleq \sum_{t=0}^{\infty}\left(\frac{1}{\varphi}\right)^t z^t$$
$$= \sum_{t=0}^{\infty}\left(\frac{z}{\varphi}\right)^t$$
$$= \frac{1}{1-z/\varphi}, \tag{42}$$

for $|z| < \varphi$, and from the fact that the Z-transform of the convolution of sequences is the product of the Z-transform of the individual sequences (see [36, Appendix 1] for further details). Finally, using Eqn. (42) and Eqn. (40) and putting $t = M$, we find:

$$Pr\{D^{\max} \ge d\} = \mathbf{p}_M(d+1)$$
$$= \sum_{i=0}^{k}\sum_{r_i=1}^{s_i} a_{d,r_i}\binom{M+r_i-1}{r_i-1}\left(\frac{1}{\varphi_{di}}\right)^M. \tag{43}$$

REFERENCES

[1] Cisco Sysems Inc, "Cisco visual networking index," White paper, Cisco Sysems, Inc, http://www.cisco.com/, Feb. 2014.

[2] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, May 1994.

[3] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. on Info. Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[4] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Trans. on Info. Theory*, vol. 45, no. 5, pp. 1468–1489, July 1999.

[5] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. on Info. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[6] H. Zhang, Y. Zheng, M. A. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 344–353, Apr. 2010.

[7] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels. II. Delay-limited capacities," *IEEE Trans. on Info. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[8] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[9] C. Gong and X. Wang, "Adaptive transmission for delay-constrained wireless video," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 1, pp. 49–61, Jan. 2014.

[10] C. T .K. Ng, D. Gündüz, A. J. Goldsmith, and E. Erkip, "Distortion minimization in Gaussian layered broadcast coding with successive refinement," *IEEE Trans. on Info. Theory*, vol. 55, no. 11, pp. 5074–5086, Nov. 2009.

[11] D. Gündüz and E. Erkip, "Joint source-channel codes for MIMO block-fading channels," *IEEE Trans. on Info. Theory*, vol. 54, no. 1, pp. 116–134, Jan. 2008.

[12] S. Shamai, "A broadcast strategy for the Gaussian slowly fading channel," in *IEEE Int. Symp. Info. Theory*, Ulm, Germany, June-July 1997.

[13] S. C. Draper and M.D. Trott, "Costs and benefits of fading for streaming media over wireless," *IEEE Network*, vol. 20, no. 2, pp. 28–33, Mar. 2006.

[14] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communications*, Prentice Hall, 2002.

[15] A. Sali, G. Acar, B. Evans, and G. Giambene, "Feedback implosion suppression algorithm for reliable multicast data transmission over geostationary satellite networks," in *Int. Workshop on Satellite and Space Commun.*, Salzburg, Sep. 2007.

[16] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[17] T.-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. C. Cosman, and A.R. Reibman, "A versatile model for packet loss visibility and its application to packet prioritization," *IEEE Trans. on Image Processing*, vol. 19, no. 3, pp. 722–735, Mar. 2010.

[18] Y. J. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames," *IEEE Trans. on Circuits and Syst. for Video Technology*, vol. 18, no. 7, pp. 861–874, July 2008.

$$(\mathbf{I} - z\mathbf{H})^{-1}_{[1,:]} = \frac{1}{(1-z)q_d(z)} \begin{bmatrix} (1-z) & (1-z)(zp) & (1-z)(zp)^2 & \cdots & (1-z)(zp)^{d-1} & (zp)^d \end{bmatrix}. \tag{36}$$

[19] A. Huszak and S. Imre, "Analysing GOP structure and packet loss effects on error propagation in MPEG-4 video streams," in *IEEE Int. Symp. on Commun., Control and Signal Processing (ISCCSP)*, Limassol, Cyprus, Mar. 2010.

[20] International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), "ITU-T recommendation Y.l541, network performance objectives for IP-based services," http://www.itu.int/, Dec. 2011.

[21] L. Toni, P. C. Cosman, and L. B. Milstein, "Channel coding optimization based on slice visibility for transmission of compressed video over OFDM channels," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1172–1183, Aug. 2012.

[22] S. Zhao, Y. Zhang, and L. Gui, "Optimal resource allocation for video delivery over MIMO OFDM wireless systems," in *IEEE Global Telecommun. Conference (GLOBECOM)*, Miami, FL, U.S.A., Dec. 2010.

[23] W. Zeng, C. T. K. Ng, and M. Medard, "Joint coding and scheduling optimization in wireless systems with varying delay sensitivities," in *IEEE Commun. Society Conf. on Sensor, Mesh and Ad Hoc Commun. and Networks (SECON)*, Seoul, Korea, June 2012.

[24] R. Aparicio-Pardo, K. Pires, A. Blanc, and G. Simon, "Transcoding live adaptive video streams at a massive scale in the cloud," in *ACM Multimedia Systems Conference*, Portland, Oregon, U.S.A., Mar. 2015.

[25] M. van der Schaar, S. Krishnamachari, C. Sunghyun, and X. Xiaofeng, "Adaptive cross-layer protection strategies for robust scalable video transmission over 802.11 WLANs," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1752–1763, Dec. 2003.

[26] M. C. O. Bogino, P. Cataldi, M. Grangetto, E. Magli, and G. Olmo, "Sliding-window digital fountain codes for streaming multimedia contents," in *Int. Symp. Circuits Systems*, New Orleans, LA, U.S.A., May 2007.

[27] A. Badr, A. Khisti, and E. Martinian, "Diversity embedded streaming erasure codes (DE-SCo): Constructions and optimalty," in *IEEE Global Telecommun. Conf.*, Miami, FL, U.S.A., Dec. 2010.

[28] D. Leong and T. Ho, "Erasure coding for real-time streaming," in *IEEE Int. Symp. Info. Theory (ISIT)*, Boston, MA, U.S.A., July 2012.

[29] A. Khisti and S.C. Draper, "Streaming data over fading wireless channels: The diversity-multiplexing tradeoff," in *IEEE Int. Symp. Info. Theory*, St. Petersburg, Russia, Aug. 2011.

[30] G. Cocco, D. Gündüz, and C. Ibars, "Real-time broadcasting over block-fading channels," in *IEEE Int. Symp. Wireless Commun. Syst.*, Aachen, Germany, Nov. 2011.

[31] G. Cocco, D. Gündüz, and C. Ibars, "Streaming transmission over block fading channels with delay constraint," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 9, pp. 4315–4327, Aug. 13.

[32] G. Cocco, D. Gündüz, and C. Ibars, "Throughput and delay analysis in video streaming over block-fading channels," in *IEEE Int. Conf. on Commun. (ICC)*, Budapest, Hungary, June 2013.

[33] L. Toni, Y.S. Chan, P.C. Cosman, and L.B. Milstein, "Channel coding for progressive images in a 2-D time-frequency OFDM block with channel estimation errors," *IEEE Trans. on Image Processing*, vol. 18, no. 11, pp. 2476–2490, Nov 2009.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, second edition edition, 2006.

[35] T. H. Luan, L. X. Cai, and S. Xuemin, "Impact of network dynamics on user's video quality: Analytical framework and QoS provision," *IEEE Trans. on Multimedia*, vol. 12, no. 1, pp. 64–78, Jan. 2010.

[36] L. Kleinrock, *Queueing Systems*, vol. I, John Wiley and Sons, 1975.