Downlink OFDM Scheduling and Resource Allocation for Delay Constraint SVC Streaming

Xin Ji*, Jianwei Huang†, Mung Chiang‡, Francky Catthoor*
*Katholieke Universiteit Leuven & Interuniversity Micro-Electronics Center, Leuven, Belgium
†Department of Information Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
‡Department of Electrical Engineering, Princeton University, Princeton, NJ, USA

***Abstract*—Efficient delivery of multimedia contents over wireless network is essential for future communication networks. However, content distribution and network engineering are traditionally studied separately, which leads to suboptimal network performance. In this paper, we consider the problem of scheduling and resource allocation for multi-user video streaming over downlink OFDM channels. The video streams are precoded with the SVC coding scheme, which offers both quality and temporal scalabilities. The OFDM technology provides the maximum flexibility of resource allocation in terms of time, frequency, and power. We propose a gradient-based scheduling and resource allocation algorithm, which explicitly takes account of video contents, deadline requirements, and the previous transmission results when calculating users' priority weights. Simulation results show that our proposed algorithm always outperforms the content-blind and deadline-blind algorithms, with a performance gain as much as 6 dB in terms of average user PSNR improvement in a congested network.**

I. Introduction

The demand of video transmission over wireless networks exhibits an ever growing trend. However, content distribution and network engineering are typically studied and optimized separately, which leads to suboptimal network performance. This problem becomes more prominent in wireless networks, where the available network resource is highly dynamic and typically quite limited, which makes it challenging to support multiple high quality video streaming sessions. To overcome the challenges, we need to jointly design the video coding and content adaption together with efficient resource allocation to achieve the best video quality measured in terms of PSNR, delay guarantees, etc.

Among various wireless technologies, Orthogonal Frequency Division Multiplexing (OFDM) has been regarded as a promising option for future broadband wireless networks due to many of its advantages such as robustness against intersymbol interference and multipath fading, and no need for complex equalizations. It is the core technology for a number of wireless data systems, such as IEEE 802.16 (WiMAX), IEEE 802.11a/g (Wireless LANs), and IEEE 802.20 (Mobile Broadband Wireless Access) [1]. In particular, OFDM provides the network designer great flexibility in allocating wireless resources in time, frequency, and power. However, most of the previous work on OFDM has been focusing on data communications (e.g., [13]–[18]), where the main objective is to maximize the total system throughput or minimize the total transmission power. Due to the unique delay sensitive but error-tolerant features of the real-time video traffic, the previously proposed solutions are not suitable for supporting delay-constrained real-time video streaming applications.

In this paper, we will focus on the problem of video streaming over OFDM downlink channels. In particular, we will consider the case where video sources are pre-coded in SVC coding scheme [5]. Among various efficient coding compression and encoding schemes (e.g., [3]–[5]), SVC emerges as one of the most promising technologies to provide flexible solutions for transmission over heterogeneous networks and adaptation for various storage devices and terminals.

For multi-user video streaming over wireless networks, it has been shown that the system performance can be significantly improved by taking the video contents into explicit consideration (e.g., [6], [11], [12]). Reference [6] focuses on maximizing the number of admitted users by giving different priorities to different video subflows according to their importance. Power constraints and channel variations are not considered in [6]. In [11], video distortion is minimized by considering power and sub-carrier constraints in OFDM systems, without explicitly enforcing the delay constraint. In [12], a distortion-based gradient scheduling algorithm was proposed without considering the influence of video latency on resource allocation. In our work, we explicitly design a "delay function" to tackle the deadline approaching effect, thus greatly reduce the chances of deadline violation. Moreover, we consider a richer wireless model that captures channel variations, frequency diversity, and other practical system constraints.

The main contribution of this paper is to provide a framework for efficient multi-user SVC video streaming over OFDM wireless channels, with an objective of maximizing the average PSNR of all video users under a total downlink transmission power constraint. Here we fully utilize the temporal and quality scalabilities of the video coding and the time, frequency and multi-user diversities of the wireless system, with explicit consideration of the stringent delay constraint of each video frame. The core of the proposed algorithm is to dynamically adjust users' priority weights based on the current video contents, deadline requirements, as well as the previous transmission results, and allocate resource accordingly based on a gradient-based scheduling framework. Simulation study shows that the advantage of the proposed algorithm becomes significant (PSNR improvement of as high as 6 dB compared with the content and delay blind approaches) when the network is congested with many video streaming users.

The rest of the paper is organized as follows. Section II introduces the multi-user OFDM video streaming system, including specifications of SVC codings schemes and the OFDM wireless network model. Section III describes the problem formulation and the family of proposed algorithms. In Section IV, we examine the performance of our proposed algorithms through simulations. Concluding remarks are given in Section V.

II. System Description

A. Scalable Video Coding (SVC)

SVC is an extension of the H.264/MPEG4-AVC video coding standard [4] and provides three different scalabilities: spatial, temporal, and quality. An overview of the features and applications of SVC can be found in [5]. In this work, we focus on how to exploit the temporal and quality salabilities by adaptive scheduling and resource allocation.[1]

In SVC, the video frames are divided into groups, or called groups of pictures (GOPs). Typical SVC GOP structure is shown in Fig. 1. The video frames are encoded into different temporal and quality layers. In Fig. 1, the video data belonging to a specific temporal layer and a specific quality layer of a video frame is represented by a box. For the sake of the video distortion calculation, we regard these boxes as the smallest data unit and name them as *packets* hereafter.

The *quality scalability* refers to the fact that the receiver can reconstruct video sequences without receiving all quality layers. In Fig. 1, the dashed arrows depict the order of the enhancement layers for each video frame.

The *temporal scalability* is based on a temporal decomposition using hierarchical B pictures. In Fig. 1, the solid arrows depict the motion predictions for each frame. Notice that the temporal and quality salabilities are not totally independent. For example, packet L_{21} can only be decoded if the packets from its lower level quality layer (i.e., L_{20}) and previous temporal layer (i.e., L'_{41} and L_{41}) are all received.

In general, the quality and temporal scalabilities provide the possibility of adapting the video transmission to different network environments. Based on the previous discussions, we know that different packets in a GOP have different priorities. Some packets need to be received first in order to make other packets useful (i.e., decodable at the receiver), and this may not follow their own playback order. Also, the sizes of the packets at different quality and temporal layers are typically different. The aforementioned reasons result in the variable bit rate (VBR) nature of the compressed SVC video sequence. It is thus useful to calculate the required rate for delivering the video data with certain priority, and use that to facilitate the scheduling and resource allocation decisions.

Let us further assume the GOP size to be g. The total number of temporal levels within a GOP is $\log_2 g$ then. We use $P^{t,q,k}$ to denote a packet that is at temporal level t, quality layer q, and belongs to the kth frame in the current GOP. Here

[1]The spatial scalability is related to downsampling of the video frames, and its effect is difficult to measure in terms of PSNR. We will consider it in the future work.

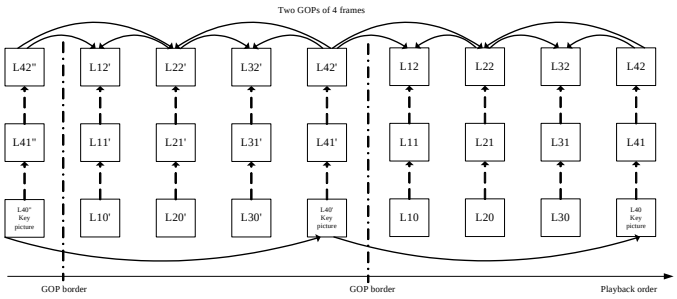

Fig. 1. GOP structure of SVC.

$1 \le k \le g$, $1 \le t \le \log_2 g$, and $0 \le q \le Q$ ($Q \le 3$ [5]). We group the packets with the same deadline as one *subflow*. Different from the subflow concept in [6], here packets from different quality layers will be grouped into different layers of one subflow since they have different priorities. This allows us to accurately capture the rate requirements of different packets within one GOP. For example, in Fig. 1, suppose all the packets necessary for decoding frame L_1 are grouped into one subflow. This subflow consists of packets L_{40}, L_{41}, L_{42} (and all the packets of former key pictures they depend on, i.e. L'_{40}, L'_{41}, L'_{42}; L''_{41}, L''_{41}, L''_{41} ... etc.), L_{20}, L_{21}, L_{22}, L_{10}, L_{11}, L_{12}. Among them, L_{40} (and the corresponding dependent packets from former GOPs), L_{20}, L_{10} form the base layer of this subflow. Other packets belong to the enhancement layers 1 and 2, respectively.

B. Wireless OFDM Network Model

We consider the downlink of a single cell OFDM system similar as the one in [7]. Time is divided into TDM time-slots, with each slot containing an integer number of OFDM symbols. The whole bandwidth is divided into a set of $\mathcal{N} = \{1, \ldots, N\}$ tones (frequency bands). During time slot t, the normalized channel gain (i.e., received signal-to-noise ratio (SNR) per unit transmission power) of user i on tone n is $e_{in,t}$. The power allocated to user i on tone n is $p_{in,t}$, which should satisfy the total power constraint at the base station: $\sum_{i,n} p_{in,t} \le P$. We also denote $x_{in,t}$ as the fraction that tone n is allocated to user i, and we have $x_{in,t} \in \{0, 1\}$ and $\sum_n x_{in,t} \le 1$, i.e., each tone can be allocated to at most one user. Note that the channel gains $e_{in,t}$'s will change from time slot to time slot, and thus the resource allocation decisions ($p_{in,t}$ and $x_{in,t}$) need to change accordingly.

User i's achievable rate on tone n in time slot t is:

$$r_{in,t} = B x_{in,t} \log(1 + \frac{p_{in,t}\tilde{e}_{in,t}}{x_{in,t} + a p_{in}\tilde{e}_{in,t}}) \qquad (1)$$

where B is the total bandwidth, $a < 1$ is the self-noise coefficient used to model the channel estimation error [20], and $\tilde{e}_{in,t} = e_{in,t}/(1+a)$.

The feasible rate region in time slot is denoted by $R(\boldsymbol{e}_t)$, where $\boldsymbol{e_t} = \{e_{in,t}, \forall i, n\}$ represents the normalized channel gains of all users on all tones in time slot t. Other vectors $\boldsymbol{x}_t$ and $\boldsymbol{p}_t$ are defined similarly. We also denote s_{in} as a maximum SNR constraint on tone n for user i, which does not change

with time. Then

$$\mathcal{R}(\boldsymbol{e_t}) = \left\{ \begin{array}{l} \boldsymbol{r}_t | r_{i,t} = \sum_n x_{in,t} B \log(1 + \frac{p_{in,t}\tilde{e}_{in,t}}{x_{in,t} + a p_{in}\tilde{e}_{in,t}}), \forall i \\ \quad \sum_{i,n} p_{in,t} \le P, \sum_i x_{in,t} \le 1, \forall n, (\boldsymbol{x}_t, \boldsymbol{p}_t) \in \chi_t. \end{array} \right\} \quad (2)$$

where

$$\chi_t := \left\{ (\boldsymbol{x}_t, \boldsymbol{p}_t) \ge 0 | x_{in,t} \in \{0,1\}, p_{in,t} \le \frac{x_{in,t} s_{in}}{e_{in,t}}, \forall i, n \right\}. \quad (3)$$

We also assume that there exist a feedback mechanism from users' receivers to the base station, such that the scheduler knows the an estimation of the joint channel state $\boldsymbol{e}_t$ at the beginning of time slot t. Such mechanism is available, for example, in the current WiMax 802.16e standards [21].

C. Video transmission over Wireless system

Let us consider a media server that is located in the backbone network and contains multiple video sequences. Once a video sequence is requested by a mobile user in a particular OFDM cell, it will be first transmitted through the backbone network to the base station of the corresponding base station. We assume that the backbone network is lossless and has high bandwidth, thus the transmission delay is negligible.[2] We assume that only one subflow of video sequence will be buffered at the base station for any user at any given time.[3] If the subflow cannot be fully transmitted to the user's receiver before its playback deadline, the frames within the partially received subflow may not be able to be decoded at the receiver. What the scheduler needs to decide is at each time slot t which users to transmit to and how much to transmit in order to achieve maximum network performance.

III. Resource Allocation for Scalable Video Streaming over OFDM Network

A. Problem Formulation

Consider a single cell OFDM system with I mobile users, each user has a utility of U_i representing the video quality received by user i, i.e., the average PSNR of the received video sequence. The value of U_i depends on the user's desired video content and its allocated transmission rate in each time slot. We want to solve the following problem:

$$\begin{array}{rl} \text{maximize} & \sum_i U_i(r_{i,1}, \ldots, r_{i,T}) \\ \text{subject to} & \boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t), t = 0, \ldots, T, \\ \text{variables} & \boldsymbol{r}_t, t = 0, \ldots, T. \end{array} \quad (4)$$

We want to maximize the total network utility within the time period $[0, T]$, subject to the constraint that the users' rate vector $\boldsymbol{r}_t$ at time t lies in the feasible rate region $\mathcal{R}(\boldsymbol{e}_t)$ (which is determined by users' channel gains and the total power constraint at the base station). This involves both *scheduling* (i.e., which users are allowed to transmit at each time slot) and *resource allocation* (i.e., how much resource is allocated to each active user in a given time slot). Problem (4) is difficult to solve since in general the utility function can not be written in close form. Moreover, it involves many integer constraints that are inherent for the OFDM system.

On the other hand, in our previous work [7] we have already obtained a thorough understanding of how to perform scheduling and resource allocation over OFDM networks for delay insensitive data traffic. The essential idea is to use a gradient-based algorithm proposed, i.e., solving the following problem in each time slot t, where $w_{i,t}$ is user i's priority weight in time slot t,

$$\begin{array}{rl} \text{maximize} & \sum_i w_{i,t} r_{i,t} \\ \text{subject to} & \boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t) \\ \text{variables} & \boldsymbol{r}_t. \end{array} \quad (5)$$

For the data transmission problem considered in [7], it is optimal to choose $w_{i,t}$ as the gradient of utility U_i at time t. This is not true for video streaming applications. *The key contribution of this paper is that we propose a family of methods to adaptively calculate the weights $w_{i,t}$'s in order to achieve the best overall long term video quality.*

B. Review: Scheduling and Resource Allocation with Fixed Priority Weights

Let us first review how to solve Problem (5) under *fixed* weights $w_{i,t}$'s as described in [7]. The key idea is to use Lagrangian dual relaxation to solve the problem. The corresponding the algorithm consists of two stages.

During the first stage, the integer channel allocation constraints (i.e., $x_{in,t} \in \{0,1\}$) are temporally ignored. We relax the total power constraint and total channel allocation constraints in the definition of $\mathcal{R}(\boldsymbol{e}_t)$ (i.e., (2)) with dual variables λ_t and $\mu_{n,t}$ for all n, respectively. Then for a fixed value of λ_t, we can analytically solve for the optimal values of $x^*_{in,t}(\lambda_t)$, $p^*_{in,t}(\lambda_t)$ and $\mu^*_{n,t}(\lambda_t)$. The computational complexity is $O(NK)$ since it involves searching for the maximum of K metrics (one from each user) on each of the N tones. The only remaining variable now is λ_t, which can be found using bi-section search and has a complexity that is independent of N and K[4].

The second stage involves enforcing integer channel allocation constraints based on $x^*_{in,t}(\lambda^*_t)$ obtained in the first stage. This requires breaking the ties over tones with fractional allocations and finding the maximum and minimum extreme points. The computational complexity is $O(NK)$. Finally we need to re-optimize the power allocation based on the integer channel allocation, which has a computational complexity

[2]When the delay is not negligible but upper-bounded, we can "shift" the users' deadlines accordingly and then the same analysis applies.

[3]If there is enough memory at the base station, we can buffer more than one subflow per user, which does not change the analysis.

[4]For example, the computational complexity of a bi-section search is $O(\log(1/\epsilon))$, where ϵ is the relative error bound target for the search.

independent of N and K. Details of the complete algorithm can be found in [7].

C. Priority Weights Calculation based on Video Distortions

In our work, the priority weights in (5) need to be calculated based on the required rates to deliver the current subflow and the corresponding distortion decrease. This is different from the delay insensitive date application considered in [7].

Let us denote the beginning of the current transmission time slot as t_c, the length (in bits) of user i's current unfinished subflow in the transmission queue at the scheduler as $l_{i_{t_c}}$, and the playback deadline of this subflow as t_i. In order to meet the deadline, the subflow needs to be transmitted at an average rate of

$$\hat{r}_i = \frac{l_{i_{t_c}}}{t_i - t_c} \tag{6}$$

Note that this may not be the actual rate that user i gets, which depends on the resource allocation decisions. If the subflow is delivered on time, the corresponding video distortion is D_{ic}.

Next we calculate the video distortion if the current subflow is not received on time. This enables us to calculate the distortion decrease for delivering the current subflow on time.

- If not all required base layer packets in the current subflow have been received by the users, then the receiver can use the last decodable frames to substitute the desirable frames, and the distortion is D_{il}. In other words, delivering the current subflow on time can lead to distortion decrease of

$$\Delta D_i = D_{il} - D_{ic}. \tag{7}$$

- If all required base layer packets have been received, but not all quality enhancement layers, then the receiver can construct the video frames based on the fully received quality layers. Assume that quality layers up to q_i level have been received, and the corresponding distortion is D_{iq_i}. In other words, delivering the current subflow on time can lead to distortion decrease of

$$\Delta D_i = D_{il} - D_{iq_i}. \tag{8}$$

Finally, the priority weight will be calculated based on the speed of distortion decrease in the current time slot as follows

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i} = \frac{\Delta D_i}{l_{i_{t_c}}}(t_i - t_c). \tag{9}$$

Another way to interpret (9) is the "derivative" of user i's utility function. By taking the users' video contents and deadlines into explicit consideration, we connect the distortion utility with the rate requirement of the video bitstreams.

A closer examination of (9) reveals that it is not very satisfactory. In particular, when the current time (t_c) approaches the deadline (t_i), the priority weight goes to zero. This is because for a given subflow, delivering it within a shorter time requires a larger rate, which leads to a reduced value of distortion decrease per unit rate. As a result, the scheduler tends to allocate a lot of resource to users in good channels and finish transmitting those subflows well before the deadlines. Then those users will have new subflows buffered at the scheduler, and those subflows have deadlines that are far away, thus again have high priority weights. The users in bad channels will seldom have chances to transmit. This is not satisfying and may lead to many deadline violations. Simulation results in Section IV also confirm this observation. To overcome such problem, we propose a framework in the next section to explicitly deal with the effect of approaching deadline, which can enforce the deadline to be satisfied with high probability while still achieving an overall good video quality.

D. Dealing with the Approaching Deadline Effect

The approach we take is to explicitly add a product term in the weight calculating, and this term increases when the deadline approaches. In this way, the system will allocate more resources to those "urgent" users and will have less deadline violations. This can avoid the case where the transmission priority always being given to those users with high rate-distortion slope at the beginning. With this assumption, the priority weight can be calculated as:

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i}\Gamma(t_i - t_c), \tag{10}$$

where the delay function Γ increases with an approaching deadline (i.e., when $t_i - t_c$ decreases). One choice that achieves the best overall performance in our simulation is to let

$$\Gamma(t_i - t_c) = \frac{1}{(t_i - t_c)^2}.$$

We will compare various choices of function Γ in Section IV.

E. Proposed Algorithms

The proposed joint scheduling and resource allocation algorithm for video streaming is given in Algorithm 1, which describes how the scheduling (i.e., which users to transmit) and resource allocation (how much each active is allowed to transmit) are done within each time slot t.

The computational complexity of the proposed algorithm comes from three parts: 1) Merging the remaining packets with the next subflow in Line 9. The worst case complexity of this step is $g(Q+1)$, where g is the GOP size and Q is the size of quality layers. 2) Calculating the priority weight $w_{i,t}$ according to(10) in Line 13, in particular, the calculation of the distortion decrease. For a video frame, the distortion of different quality layer can be pre-calculated before streaming. Only if the base layer of a video frame is not successfully received during the transmission, the distortion decrease need to be recalculated between the different frames. Since this rarely happens in practice (as verified by our simulations), the complexity comes from this part is negligible. 3) Solving the weighted rate maximization problem (5) in Line 15. This step is the most complicated part of the proposed algorithm, and detailed complexity analysis can be found in Section III.B. As this step is independent of the video content and is always achieved at the base station in a downlink OFDM system, it can be realized by dedicated hardware which can significantly decrease the realtime calculation time.

Algorithm 1 Joint Scheduling and Resource Allocation Algorithm for Multi-user Video Streaming

$t = 0$.
repeat
  $t = t + 1$.
  for all users $i = 1, \ldots, I$ **do**
    repeat
      check the deadline of the current subflow.
      if the deadline has already passed **then**
        discard packets not useful for decoding future packets.
        merge the remaining packets with the next subflow.
        let the next subflow be the current subflow.
      end if
    until the deadline of the current subflow has not passed
    calculate the priority weight $w_{i,t}$ according to (10).
  end for
  solve weighted rate maximization problem (5) using the algorithm described in Section III-B, and each user i is allocated transmission rate $r_{i,t}$.
  for all users $i = 1, \ldots, I$ **do**
    continue to transmit the current subflow with rate $r_{i,t}$.
    if the current subflow is transmitted successfully before the end of the time slot **then**
      obtain the next subflow from the media server.
      transmit with rate $r_{i,t}$.
    end if
  end for
until no more video to be streamed

IV. Simulation Study

A. Simulation Setup

We perform extensive simulations to show the performance gain of our proposed delay-aware scheduling and resource allocation algorithm with different delay function Γ.

The video sequences used in the experiments are encoded according in H.264 extended SVC standard (using JVT reference software, JSVM 8.12 [5]) at variable bit rates with an average PSNR of 35dB for each sequences. Four sequences ("Harbor", "City", "Foreman", "Mobile and calendar") are used to represent video with very different levels of motion activities. All the sequences are coded at CIF resolution (352×288, 4:2:0) and 30 frames per second. A GOP size of 8 is used. The first frame is encoded as I frame and all the key pictures of each GOP were encoded as P frames. Foreman sequence is encoded with an original rate 449.2 kbps and an average PSNR of 35.16dB; City sequence is encoded with an original rate 585.8 kbps and an average PSNR of 35.98dB; Harbor sequence is encoded with a rate of 1599.7 kbps and an average PSNR of 35.32dB; Mobile sequence is encoded with an original rate 2019 kbps and an average PSNR of 35.17dB.

For the wireless system, we perform simulation based on a realistic OFDM simulator with realistic industry measurements and assumptions commonly found in IEEE 802.16 standards [21]. We simulate a single OFDM cell with a total transmission power of $P = 6$W at the base station. The channel gains e_{ij} are the products of a fixed location-based term for each user i and a frequency-selective fast fading term. The location-based components were picked using an empirically obtained distribution for many users in a large system. The fast-fading term was generated using a block-fading model based upon the Doppler frequency (for the block-length in time) and a standard reference mobile delay-spread model (for variation in frequency). For a user's fast-fading term, each multi-path component was held fixed for $2m$sec (i.e., a fading block length), which corresponds to a 250MHz Doppler frequency. The delay-spread is 1μsec. The users' channel conditions are averaged over the applicable channelization scheme and fed back to the scheduler at the base station. All video users are randomly selected from the users with an average channel normalized SNR of at least 20dB. This makes sure that it is possible to support the minimum quality of the video streaming.

We considered a system bandwidth of 5MHz consisting of 512 OFDM tones, which are grouped into 64 subchannels (8 tones per subchannel). The symbol duration is 100μsec with a cyclic prefix of 10μsec. This roughly corresponds to 20 OFDM symbols per fading block (i.e., $2m$sec). This is one of the allowed configurations in the IEEE 802.16 standards [21]. The resource allocation is done once per fading block. For each video sequence, we report result that is averaged over 5 randomly generated channel realizations with a length of 10 seconds each (which corresponds to 10^5 OFDM symbols).

B. Different weight definitions

We simulate the algorithm with different delay functions Γ when calculating the weights $w_{i,t}$ in (10). In total, we simulate six algorithms. The first two algorithms are benchmark algorithms. To observe how the delay approaching effect and the video rate-distortion reciprocate with each other during the run-time, we simulate the last four algorithms according to our proposed method in section III.D with different level of emphasizes on deadline violation avoidance. We will show that algorithm $W_{\Gamma 2}$ achieves the best performance among all proposed ones.

- W_1: $w_{i,t} = 1$ for all i and t. This is the rate maximization algorithm, which is "content-blind" but widely used in data-oriented wireless communication systems (e.g., [7]).
- W_{rd}: $\Gamma(t_i - t_c) = 1$. This algorithm takes users' contents into consideration but does not explicit address the deadline approaching effect and thus is "deadline-blind".
- $W_{\Gamma 1}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)$.
- $W_{\Gamma 2}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)^2$.
- $W_{\Gamma 3}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)^3$.
- $W_{\Gamma 4}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)^4$.

Table I shows average PSNR achieved by four users request different four video clips with the same starting time. The initial playback deadline is set to be 200ms [9].

As we can see, the weighted gradient based scheduling reflects the rate-distortion properties of different video content. Under W_1 algorithm, the qualities of Mobile and Foreman are almost the same although they have very different rate-distortion properties. This is because W_1 simply maximizes

TABLE I
AVERAGE PSNR FOR 4 USERS WITH 200MS INITIAL PLAYBACK DEADLINE

Sequence	W_1	W_{rd}	$W_{\Gamma 1}$	$W_{\Gamma 2}$	$W_{\Gamma 3}$	$W_{\Gamma 4}$
Foreman	27.6218	25.8604	32.8488	33.3882	33.007	32.8352
City	34.1758	32.7642	34.0714	33.9814	33.6738	33.5146
Harbor	22.8458	18.1146	23.9022	26.1308	26.0102	25.8442
Mobile	27.191	15.2696	24.5064	27.6664	27.475	27.3092
Average	27.9588	23.0022	28.8324	30.2918	30.0416	29.8756

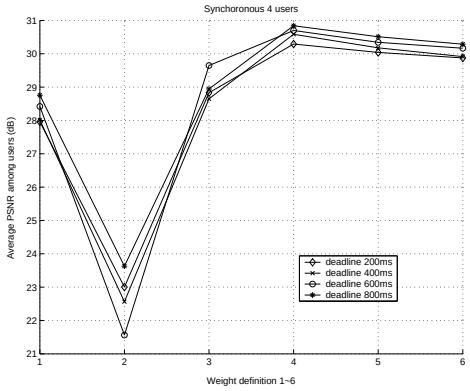

Fig. 2. Synchronous deadlines for 4 users. Horizontal axis represent different algorithms: 1 - W_1; 2 - W_{rd}; 3 - $W_{\Gamma 1}$; 4 - $W_{\Gamma 2}$; 5 - $W_{\Gamma 3}$; 6 - $W_{\Gamma 4}$;

the rate without considering the resulting video quality. Instead, by allocating network resource according to the users' video rate-distortion properties, the weighted scheduling and resource allocation schemes can dynamically adjust the resource allocation based on video contents.

Compared to the baseline W_1 algorithm, the W_{rd} algorithm actually decreases the average video quality among different users. This is due to the deadline approaching effect explained in Section III-C.

Once we take care of this effect by properly choosing Γ functions in $W_{\Gamma 1}$ to $W_{\Gamma 4}$, the average PSNR among users is improved over the simple total rate maximization scheme (W_1) by 0.9 dB to 2.3 dB. Results of $W_{\Gamma 2}$ achieves the best average PSNR value, while $W_{\Gamma 3}$ and $W_{\Gamma 4}$ tend to decrease the average PSNR value compared with $W_{\Gamma 2}$ since they put too much emphasis on not violating the deadlines.

C. Impact of different initial playback deadlines

Here we check the impact of different initial playback deadlines. The initial playback deadline, which refers to delay from the time user requires the video and the time the video starts to play at the receiver. According to the user satisfactory study in [9], we test the varied initial playback deadline between 200ms to 800ms. Four users still request the different video sequences from the server simultaneously. Other parameters are the same as before. The results are presented in Fig. 2. We can see that $W_{\Gamma 2}$ always achieve the highest average PSNR value under different deadlines. And typically the longer the initial deadline, the better the video quality.

D. Synchronous and Asynchronous requirements' influence

So far we have only considered the case of synchronously deadlines, i.e., all users start video streaming at the same time.

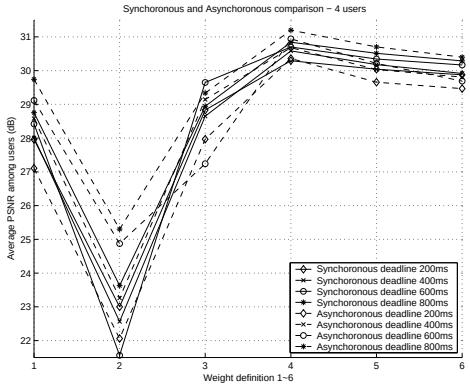

Fig. 3. Synchronous and asynchronous deadlines for 4 users. Horizontal axis represent different algorithms: 1 - W_1; 2 - W_{rd}; 3 - $W_{\Gamma 1}$; 4 - $W_{\Gamma 2}$; 5 - $W_{\Gamma 3}$; 6 - $W_{\Gamma 4}$;

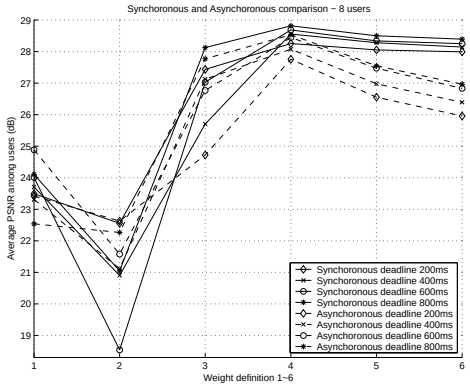

Fig. 4. Synchronous and Asynchronous deadlines for 8 users: 1 - W_1; 2 - W_{rd}; 3 - $W_{\Gamma 1}$; 4 - $W_{\Gamma 2}$; 5 - $W_{\Gamma 3}$; 6 - $W_{\Gamma 4}$;

In reality, it is more common that different users request video clips at different time, which we call asynchronous deadlines. In Fig. 3, we compare the results of both cases for four users. In the asynchronous deadline case, four users randomly start to request the different video sequences from the server within the first initial playback deadline. We also observe that the $W_{\Gamma 2}$ algorithm always performs the best.

E. Different user content and congestion range's influence

Fig. 4 shows the results of eight users requesting video sequences at the same time. Each of the 4 video sequences is required by 2 users. Synchronous and asynchronous cases are both shown here. For the asynchronous case, users still randomly request the video sequence within one playback deadline. The other setups are the same as section IV-B.

The effectiveness of our proposed algorithms is more obvious in this case, where the network is more congested. In particular, $W_{\Gamma 2}$ achieves as high as 6dB improvement in users' average PSNR compared with the W_1 rate maximization algorithm.

V. CONCLUSION

Traditionally the content distribution and network resource allocation are designed separately. Although working well

in certain wireline communication settings, this approach is certainly not optimal for wireless communication networks, where the available network resource changes rapidly in time. In this paper, we apply a joint design approach to solve the challenging problem of multi-user video streaming over wireless channels. We focused on the SVC coding schemes and the OFDM schemes, which are shown to be among the most promising technologies for video coding and wireless communications, respectively.

Building on the gradient-based scheduling framework designed in our previous work, we proposed an algorithm that explicitly calculate the users' priority weights based on the video contents, deadline requirements, and previous transmission results, and then optimize the resource allocation taking the current wireless channel conditions and various practical constraints into consideration. Simulation results show that our algorithm always outperforms the rate maximization (content-blind) scheme and the pure gradient-based (deadline-blind) scheme. The performance gain in terms of average user PSNR improvement is as much as 6 dB in a congested network. Finally, the performance of the algorithm is consistent under both synchronous and asynchronous deadlines.

REFERENCES

[1] "Orthogonal frequency-division multiplexing," Wikipedia, http://en.wikipedia.org/w/index.php?title=Orthogonal_frequency-division_multiplexing&oldid=153352313.

[2] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. of 2002 Allerton Conference*, 2002.

[3] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301 -317, March 2001.

[4] T. Wiegand,G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, July 2006.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable H.264/MPEG4-AVC extension," in *Proc. IEEE International Conference on Image Processing (ICIP'06)*, Atlanta, GA, USA, Oct. 2006.

[6] M. Van der Schaar, Y. Andreopoulos, and Z. Hu, "Optimized Scalable Video Streaming over IEEE 802.11 a/e HCCA Wireless Networks under Delay Constraints," *IEEE Trans. Mobile Computing*, vol. 5, pp. 755-768, June 2006.

[7] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink Scheduling and Resource Allocation for OFDM Systems," *Proc. of CISS*, Princeton Univeristy, 2006.

[8] *JSVM 8 Reference Software*, JVT-Q203, May 2007.

[9] J. Lou, H. Cai, and J. Li, "A Real-Time Interactive Multi-View Video System," *proc. of the 13th annual ACM international conference on Multimedia*, Singapore, 2005.

[10] P. Chou and Z. Miao, "Rate-Distortion Optimized Streaming of Packetized Media," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 390–404, 2006.

[11] G. M. Su, Z. Han, M. Wu, and K. J.R. Liu, "A Scalable Multiuser Framework for Video over OFDM Networks: Fairness and Efficiency," *IEEE Trans. on CSVT*, vol. 16, no. 10, pp. 1217–1231, 2006.

[12] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Content-Aware Resource Allocation and Packet Scheduling for Video Transmission over Wireless Networks," *IEEE J. Select. Areas Commun.*, vol. 25, no. 4, pp. 749–759, 2007.

[13] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms," *Communications, IEEE Transactions on*, vol. 52, no. 6, pp. 922–930, 2004.

[14] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun*, vol. 17, no. 10, pp. 1747–1758, 1999.

[15] J. Jang and K. Lee, "Transmit power adaptation for multiuser OFDM systems," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 2, pp. 171–178, 2003.

[16] T. Chee, C. Lim, and J. Choi, "Adaptive power allocation with user prioritization for downlink orthogonal frequency division multiple access systems," in *ICCS 2004*, pp. 210–214, 2004.

[17] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *IEEE Globecom*, 2000.

[18] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, July 2006.

[19] R. Jain, D. Chiu and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Systems." DEC Research Report TR-301, 1984.

[20] H. Jin, R. Laroia, and T. Richardson, "Superposition by position," preprint, 2006.

[21] "IEEE 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005," `http://www.ieee802.org/16/`.