Scheduling and Resource Allocation for SVC Streaming over OFDM Downlink Systems

Xin Ji, Jianwei Huang, Mung Chiang, Gauthier Lafruit, and Francky Catthoor

***Abstract*—We consider the problem of scheduling and resource allocation for multiuser video streaming over downlink orthogonal frequency division multiplexing (OFDM) channels. The video streams are precoded using the scalable video coding (SVC) scheme that offers both quality and temporal scalabilities. The OFDM technology provides the flexibility of resource allocation in terms of time, frequency, and power. We propose a gradient-based scheduling and resource allocation algorithm, which prioritizes the transmissions of different users by considering video contents, deadline requirements, and transmission history. Simulation results show that the proposed algorithm outperforms the content-blind and deadline-blind algorithms with a gain of as much as 6 dB in terms of average PSNR when the network is congested.**

***Index Terms*—Multiuser, OFDM, resource allocation, SVC, video streaming.**

I. Introduction

The demand of high-quality video over communication networks exhibits an ever growing trend. However, content distribution and resource allocation are typically studied and optimized separately, which leads to suboptimal network performance. This problem becomes more prominent in wireless networks, since the typically time-varying and limited network resource makes efficient multiuser video streaming particularly challenging.

In this letter, we consider the problem of multiuser video streaming over orthogonal frequency division multiplexing (OFDM) networks, where videos are coded in a scalable video coding (SVC) scheme. OFDM is the core technology for a number of wireless data systems (e.g., WiMAX and wireless LANs). The resource allocation in OFDM can be flexibly performed over power, frequency, and time. SVC, on the other hand, allows reconstructing lower quality signals from partially received bitstreams. It provides flexible solutions for transmission over heterogeneous networks and allows easy adaptation to various storage devices and terminals. Our focus in this letter is to design efficient streaming protocols that fully exploit various flexibilities in both OFDM and SVC.

There is a growing literature on SVC-based video transmission over wireless networks. Most of them focused on exploiting the scalable feature of SVC to provide QoS quarantee for the end users (e.g., [12], [13]). In [15], the layered bitstream of SVC is exploited together with congestion control algorithm for distributing video to WiMax subscriber stations. In [16], the rate distortion model proposed for H.264/AVC was extended to include the effect of random packet loss on SVC. Reference [2] focused on maximizing the number of admitted users by giving different priorities to different video subflows according to their importance. None of the above results considered power control. An unequal power allocation scheme was proposed in [17] for the transmission of SVC packets in a WiMax system. In [5], a distortion-based gradient scheduling algorithm was proposed. However, they did not consider the influence of video latency on resource allocation.

The main contribution of this letter is to provide a framework for efficient multiuser SVC video streaming over OFDM wireless channels. The objective is to maximize the average PSNR of all video users under a total downlink transmission power constraint. The basis of our approach is the stochastic subgradient-based scheduling (e.g., [10]) and our previous work [3] that considers downlink OFDM resource allocation for *elastic data* traffic.

In this letter, we generalize the framework in [3] to real-time video streaming by further considering dynamically adjusted priority weights based on the current video contents, deadline requirements, and the transmission history. The resulting algorithms not only fully utilize the temporal and quality scalabilities of the SVC scheme, but also thoroughly explore the time, frequency, and multiuser diversities of the OFDM system. Simulations show that the proposed algorithms are better than the content-blind and delay-blind approaches, and the improvement becomes quite significant (e.g., PSNR improvement of as high as 6 dB) in a congested network.

Manuscript received December 08, 2008; revised on August 17, 2009. First version published July 7, 2009; current version published September 30, 2009. This work is supported in part by ONR Grant N00014-07-1-0864, NSF Grant CNS-0720570, by the Competitive Earmarked Research Grants (Project Number 412308) established under the University Grant Committee of the Hong Kong Special Administrative Region, China, the Direct Grant (Project Number C001-2050398) of The Chinese University of Hong Kong, and the National Key Technology Research and Development Program (Project Number 2007BAH17B04) established by the Ministry of Science and Technology of the People's Republic of China. This paper was recommended by Associate Editor T. Inguen.

X. Ji, G. Lafruit, and F. Catthoor are with the Interuniversity Micro Electronics Center, University of Leuven, 3001 Leuven, Belgium (e-mail: jixinxin@gmail.com; lafruit@imec.be; catthoor@imec.be).

J. Huang is with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong, China (e-mail: jwhuang@ie.cuhk.edu.hk).

M. Chiang is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: chiangm@princeton.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2026812

The remainder of this letter is organized as follows. Sections II and III introduce the OFDM and SVC models. Section IV describes the problem formulation and the proposed algorithms. In Section V, we examine the performance of the proposed solutions through simulations. Concluding remarks are given in Section VI.

II. OFDM Wireless Network Model

The OFDM network model considered here is similar to that in [3]. Different video bitstreams are transmitted from the base station to a set $\mathcal{I} = \{1, \ldots, I\}$ of mobile users in a single OFDM cell. Time is divided into equal length slots, and each slot contains an integer number of OFDM symbols. The entire frequency band is divided into a set $\mathcal{J} = \{1, \ldots, J\}$ of tones (carriers). The rate achieved by user i at time t, $r_{i,t}$, depends on the resource (tone and power) allocation and the channel gains. In each time-slot, the scheduling and resource allocation decision can be viewed as selecting a rate vector $\mathbf{r}_t = (r_{i,t} \forall i \in \mathcal{I})$ from the current feasible rate region $\mathcal{R}(\mathbf{e}_t) \subseteq \mathbb{R}_+^K$, where $\mathbf{e}_t$ indicates the time-varying channel state information available at the scheduler at time t. For presentation simplicity, we omit the time index t when we only focus on the resource allocation problem in one time slot.

For each tone $j \in \mathcal{J}$ and user $i \in \mathcal{I}$, let e_{ij} be the received signal-to-noise ratio (SNR) per unit power. The power allocated to user i on tone j is p_{ij}, and the fraction of time that tone allocated to user i is x_{ij}. The total power allocation satisfies $\sum_{i,j} p_{ij} \leq P$, where P is the total downlink power constraint at the base station. The allocation for each tone j satisfies $\sum_i x_{ij} \leq 1$. With perfect channel estimation, user i's feasible rate on tone j is $r_{ij} = x_{ij} B \log(1 + (p_{ij} e_{ij}/x_{ij}))$, which corresponds to the Shannon capacity of a Gaussian noise channel with bandwidth $x_{ij} B$ and received SNR $p_{ij} e_{ij}/x_{ij}$. This SNR arises since the active transmission power that user i transmits on tone j is p_{ij}/x_{ij} when only a fraction x_{ij} of the tone is allocated. Without loss of generality we let $B = 1$ in the analysis.

In practical OFDM networks, imperfect carrier synchronization and channel estimation may result in "self-noise" (e.g., [7], [11]). With self-noise, user i's feasible rate on tone j becomes

$$r_{ij} = x_{ij} \log\left(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right), \tag{1}$$

where $\tilde{e}_{ij}$ is the estimated value of e_{ij} and $\beta << 1$ is the self-noise coefficient. Details of how (1) is derived can be found in [3]. The feasible rate region is then

$$\mathcal{R}(\mathbf{e}) = \left\{ \begin{array}{l} \mathbf{r} : r_i = \sum_j x_{ij} \log\left(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right) \forall i \in \mathcal{I}, \\ \sum_{i,j} p_{ij} \leq P, \sum_i x_{ij} \leq 1 \; \forall j \in \mathcal{J}, (\mathbf{x}, \mathbf{p}) \in \mathcal{X} \end{array} \right\}. \tag{2}$$

Here $\mathcal{X} := \prod_{j=1}^N \mathcal{X}_j$ is the feasible region for $(\mathbf{x}, \mathbf{p})$, where

$$\mathcal{X}_j := \left\{ (\mathbf{x^j}, \mathbf{p^j}) \geq \mathbf{0} : \mathbf{x_{ij}} \leq \mathbf{1}, \mathbf{p_{ij}} \leq \frac{\mathbf{x_{ij}\tilde{s}_{ij}}}{\tilde{\mathbf{e}}_{\mathbf{ij}}} \; \forall \mathbf{i} \in \mathcal{I} \; \forall \mathbf{j} \in \mathcal{J} \right\} \tag{3}$$

with $\mathbf{x^j} := (\mathbf{x_{ij}} \forall \mathbf{i} \in \mathcal{I})$ and $\mathbf{p^j} := (\mathbf{p_{ij}} \forall \mathbf{i} \in \mathcal{I})$. Here, $\tilde{s}_{ij} = \Gamma_{ij}/(1 - \Gamma_{ij}\beta)$ is the normalized maximum SNR constraint, where $\Gamma_{ij} < 1/\beta$ is a maximum SNR constraint on tone j for user i. This models the limitation on the available modulation and coding schemes.

We assume that $\tilde{e}_{ij}$s (for all i and j) and β (e.g., the estimation error variance) are known by the scheduler. This is possible in both frequency division duplex (FDD) and time division duplex (TDD) systems [3]. In both cases, this feedback information would need to be provided within the channel's coherence time.

III. SVC Scheme of Video Coding

SVC is an extension of the H.264/MPEG4-AVC video coding standard and provides three different scalabilities: spatial, temporal, and quality. An overview of SVC can be found in [1]. In this letter, 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 groups of pictures (GOPs). The typical SVC GOP structure is shown in Fig. 1, where for illustration purposes we assume that one GOP consists of four frames. The video frames are further encoded into different temporal and quality layers. One box in Fig. 1 represents the data belonging to a combination of one specific temporal layer and one specific quality layer. For the purpose of video distortion calculation, we regard one box as the smallest decodable data unit and call it a "packet." All the packets in one column represent one video frame. For example, frame L_1 consists of two packets: L_{10} and L_{11}.

The packets at the same horizontal level belong to the same quality layer. Due to quality scalability, a video decoder can reconstruct *video* sequences without receiving all quality layers. After receiving the base layer (the lowest layer), the decoder can already provide a video with some reasonable quality. The video quality can be improved if one or more enhancement quality layers are received before the playback deadline of the corresponding video frames.

The packets at the same vertical level (i.e., in the same frame) belong to the same temporal layer, and different frames may belong to the same temporal layer. The *temporal scalability* is based on a temporal decomposition using a hierarchical B pictures scheme. For example, only after receiving packets L'_{40} and L_{40} (together with all the base layer of video frames they depend on), packet L_{20} can be decoded at the receiver. Notice that the temporal and quality salabilities are not 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.

It is clear 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

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

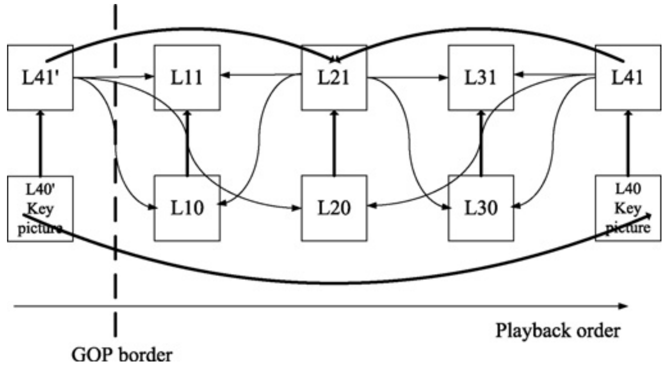

Fig. 1. GOP structure of SVC.

sizes of the packets at different quality and temporal layers are typically different. Because of this, the compressed SVC video bitstream exhibits a variable bit rate (VBR) nature. It is thus useful to calculate the required rate for delivering the packets with the same priority, and use that to facilitate the scheduling and resource allocation decisions.

To facilitate the analysis, we divide the video flow (bitstream) into *subflows*. The packets with the same deadline, i.e., all the video bitstream data that is necessary to receive before correctly decoding one video frame, are grouped into one subflow. Let us assume the GOP size is g. The total number of temporal levels within a GOP is then $\log_2 g$. Also we use $P^{t,q,k}$ to denote the packet that belongs to temporal level t, quality layer q, and frame k in the current GOP. Here $1 \leq t \leq \log_2 g$, $0 \leq q \leq Q$, and $1 \leq k \leq g$. Normally, we have $Q \leq 3$ [1]. For the example in Fig. 1, suppose all the packets that are necessary for decoding frame L_1 belong to one subflow. This subflow consists of packets L_{40} and L_{41} (and all the packets of the former key pictures they depend on, i.e., L'_{40}, L'_{41}, ... etc.), L_{20}, L_{21}, L_{10}, and L_{11}. Different from the subflow concept in [2], here we also differentiate different quality layers within the same subflow. Among the packets inside this subflow, L_{40} (and the corresponding dependent packets from the former GOPs), L_{20}, and L_{10} belong to the base layer of the current subflow. Other packets belong to the enhancement layer. This allows us to accurately capture the rate requirements of different packets within one GOP.

IV. Scheduling and Resource Allocation Algorithms

A. Gradient-based Scheduling Framework

Consider a media server that is connected to the base station through a high bandwidth backbone network. Each of the I mobile users in the OFDM cell requests a separate *video* sequence to be streamed from the media server. We assume that the backbone network is lossless and the transmission delay from the media server to the OFDM base station is negligible. For each user, only one GOP of the requested sequence will be buffered at the base station at any given time.[2] If a subflow of the GOP cannot be fully received by the mobile user before its playback deadline, the frames within the partially received subflow may not be able to be decoded at the receiver. Our objective is to design a scheduling and resource allocation algorithm that achieves the maximum long-term average overall network streaming quality, under time varying channel conditions and variable rate video contents.

Our starting point is the stochastic gradient-based scheduling framework presented in, e.g., [10]. In this framework, each user i is assigned a utility function $U_i(W_{i,t})$ depending on their average throughput $W_{i,t}$ up to time t, which is used to quantify fairness between users. During each scheduling epoch t, the system objective is to choose a rate vector $\mathbf{r}_t$ in $\mathcal{R}(\mathbf{e_t})$ that maximizes a (dynamic) weighted sum of the users' rates, where the weights are determined by the gradient of the sum utility across all users. Hence, the scheduling and resource allocation decision is to

$$\max_{\mathbf{r_t} \in \mathcal{R}(\mathbf{e_t})} \sum_{i \in \mathcal{I}} \frac{\partial U_i(W_{i,t})}{\partial W_{i,t}} r_{i,t}. \qquad (4)$$

The above policy has been shown to yield utility maximizing solutions under a time-varying rate region, i.e., maximizing $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i\in\mathcal{I}} U_i(W_{i,t})$.

The main advantage of this policy is its greedy nature, *i.e.*, the optimization at time t does not require any rate region information of other time slots (past or future). We notice that Problem (4) needs to be solved for each time slot.

Based on this, in [3] we proposed an efficient algorithm to solve Problem (4) for an OFDM downlink system with elastic data transmission. The algorithm applies to any weighted rate maximization with fixed weights, and will be used as a module of our algorithms.

B. Dynamic Weight Calculation For Streaming Applications

Weight calculation based on utility gradient is not suitable for real-time video streaming application, since the stringent delay constraints are not explicitly considered. This motivates the design of a different weight calculation method, which involves calculation of the rates to deliver the current subflow and the corresponding distortion decrease.

Without loss of generality, assume that the current time slot starts at $t = 0$. For user i's current unfinished subflow at the base station, denote its length as l_i bits and the playback deadline as $t_i > 0$. To meet the deadline, one needs to transmit the subflow at a rate of

$$\hat{r}_i = \frac{l_i}{t_i}. \qquad (5)$$

Note that this is the desired rate instead of the actual allocated rate.

Denote the corresponding distortion of the corresponding frame as D_{ic} if the current subflow is successfully received before the required playback deadline. Video distortion can be regarded as the negative function of user's utility. The *distortion decrease* depends on how much distortion is at time $t = 0$ and can be calculated as follows.

1) If some of the base layer packets within the current subflow have not been received at time $t = 0$, then the receiver will use the last decodable frames to substitute

[2] If there is enough memory at the base station, we can buffer more than one GOP per user and the analysis remains essentially unchanged.

the desired frames and achieve distortion $D_{il}\,(>D_{ic})$ at time $t = 0$. This means successfully delivering the current subflow on time leads to a reduction in distortion by

$$\Delta D_i = D_{il} - D_{ic}. \tag{6}$$

2) If up to q quality layer packets within the current subflow have been fully received at time $t = 0$, where q is less than the maximum number of quality layers available, then the receiver can construct the video frames based on the received quality layers and achieves a distortion D_{iq} $(>D_{ic})$. In this case, successfully delivering the current subflow on time leads to a reduction in distortion by

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

Similar to the utility gradient for elastic data traffic, here we can calculate the speed of reduction in distortion (i.e., priority weight) in the current time slot as follows:

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i} = \frac{\Delta D_i}{l_i} t_i. \tag{8}$$

By taking the users' video contents and deadlines into explicit consideration, we connect the distortion (i.e., utility) with the rate requirement of the video bitstreams.

On the other hand, using the weight definition of (8) only to solve Problem (4) may not lead to good overall video quality. This is due to the "approaching deadline effect." Assume user i's unfinished subflow length l_i is fixed, and so is the possible distortion decrease ΔD_i. If the deadline is approaching, i.e., t_i becomes smaller, priority weight calculated based on (8) actually decreases. This is because for a given amount of data, delivering it within a shorter amount of time requires a larger transmission rate, which leads to a smaller distortion decrease per unit rate. This is counter-intuitive, however, since we would expect that a user with an approaching deadline will have higher priority.

For users with the same weight, a user in good channel condition requires less resource to achieve the same transmission rate and thus is favored under (8). Once a user's current subflow is transmitted completely, the next new subflow has a longer deadline, i.e., a larger t_i, which leads to a higher priority weight and more resource allocation. This means that users in worse channels will seldom have chances to transmit and will face a lot of deadline violations. Simulation results in Section V confirm this problem.

To tackle this problem, next we explicitly enforce the deadlines to be satisfied with high probability while still achieving an overall good video quality.

C. Mitigating the Approaching Deadline Effect

We propose adding a product term to the weight calculation. This term is a decreasing function of t_i, i.e., it increases when the deadline approaches. This solution shares some similarity with the method proposed in [14], which can enforce the system to allocate more resources to "urgent" users and reduce deadline violations. The new priority weight can be calculated as

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i} \Gamma(t_i) \tag{9}$$

where the delay function $\Gamma(t_i)$ decreases with t_i. One choice that achieves the best overall performance in our simulation is

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

More examples of function Γ will be given in Section V.

Algorithm 1. Joint Scheduling and Resource Allocation Algorithm for Multiuser Video Streaming.

```
initialization $t = 0$ repeat
    $t = t + 1$;
    forall user $i$ do
        repeat
            check the deadline of the current subflow;
            if the deadline has passed then
                discard those packets (at the base station)
                not useful for decoding future packets;
                fetch the next subflow from the media
                server and merge with the left useful
                packets;
            end
        until the current subflow deadline has not passed
        ;
        Calculate the priority weight $w_{i,t}$ as in (9)
    end
until no more video to be streamed ;
Solve weighted rate maximization problem using the
algorithm described in [3] with fixed weights such that
each user $i$ transmits with rate $r_{i,t} \geq 0$;
forall user $i$ do
    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
end
```

D. Proposed Algorithms

The proposed joint scheduling and resource allocation algorithm is given in Algorithm 1, which describes which users to transmit and how much rate each active user gets at any given time slot.

According to the way that the subflow is defined in Section III, each user transmits the packets in the base quality layer first (from all temporal layers), and then the packets from enhancement quality layers. The video quality degradation is mainly due to two reasons: (i) some packets are discarded at the scheduler before transmission since their deadlines have already passed, or (ii) some packets are discarded at the receiver because they cannot be decoded due to lack of necessary dependent packets.[3] Both have been taken into consideration in Algorithm 1.

The overall worst-case complexity of Algorithm 1 for each time slot is $O(I(J + gQ)$ based on the analysis of three parts.

1) Merging the remaining packets with the next subflow, with a complexity of $O(gQ)$. Here g is the GOP size and Q is the maximum number of the quality layers. Since this needs to be done by each user, the overall complexity is $O(IgQ)$, where I is the total number of users.
2) Calculating the priority weight $w_{i,t}$ as in (9). For a video frame, the distortion of different quality layers can be pre-calculated before streaming. Only if the base layer of a subflow is not successfully received during the transmission, the reduction in distortion needs to be re-calculated between the different frames. Since this rarely happens in practice (as verified by our simulations), the complexity coming from this part is negligible.
3) Solving the weighted rate maximization problem as in [3], with a complexity $O(IJ)$, where J is the total number of subchannels.

V. Simulation Studies

A. Simulation Setup

We perform extensive simulations to show the performance gain of our proposed algorithm with different delay functions.

The *video* sequences used in the experiments are encoded according to H.264 extended SVC standard (using JVT reference software, JSVM 8.12 [5]) at variable bit rates with an average PSNR of 35 dB for each sequence. Four sequences (*Harbor*, *City*, *Foreman*, and *Mobile*) 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 at an original rate of 449.2 kbps and an average PSNR of 35.16 dB; *City* sequence is encoded at an original rate of 585.8 kbps and an average PSNR of 35.98 dB; *Harbor* sequence is encoded at a rate of 1599.7 kbps and an average PSNR of 35.32 dB; *Mobile* sequence is encoded at an original rate 2019 kbps and an average PSNR of 35.17 dB.

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 [9]. We simulate a single OFDM cell with a total transmission power of P = 6W at the base station. Other wireless system parameters are the same as that in Section IV of [3], except that all video users are randomly selected from the users with an average channel normalized SNR of at least 20 dB. This makes sure that it is possible to support the minimum quality of the video streaming.

TABLE I

Average PSNR For Four Users With 200 ms Initial Playback Deadline

Content	W_1	W_2	W_{rd}	$W_{\Gamma 1}$	$W_{\Gamma 2}$	$W_{\Gamma 3}$
Mobile	28.5316	26.7014	18.6482	20.6136	28.0960	27.6642
Foreman	29.0880	30.7430	27.2240	30.6424	33.5992	33.2444
City	34.2552	31.0290	33.5274	34.1902	34.0882	33.8188
Harbor	23.5310	26.9150	20.1732	21.6224	26.1610	26.0774
Average	28.8514	28.8470	24.8932	26.7672	30.4862	30.2012

B. Different Weight Definitions

We simulate the algorithm with different functions Γ that mitigate the approaching deadline effect when calculating the weights $w_{i,t}$ in (9). The effectiveness of the proposed algorithm is illustrated by comparison with the rate maximization algorithm and the algorithm proposed in [5]. We simulate a total of six algorithms. The first two algorithms are benchmark algorithms, and the last four algorithms are our proposed ones with different levels of emphases on deadline violation avoidance. We will show that algorithm $W_{\Gamma 2}$ achieves the best performance among all proposed ones.

1) W_1 (benchmark 1: content-blind approach): $w_{i,t} = 1$ for all i and t. This is the rate maximization algorithm, which is "content-blind" but widely accepted in data-oriented wireless communication systems (e.g., [3]). On top of this, we use the packet dropping policy for SVC proposed in [8].
2) W_2 (benchmark 2: deadline-blind approach): the weights in this approach are defined as in [5]. Instead of grouping packets into subflows, the schedular will transmit every packet following the order of Method II proposed in [6], which has been proven to achieve similar results as the optimal one. It is considered as a "deadline-blind" benchmark.
3) 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 also "deadline-blind."
4) $W_{\Gamma 1}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)$.
5) $W_{\Gamma 2}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)^2$.
6) $W_{\Gamma 3}$: $\Gamma(t_i - t_c) = 1/(t_i - t_c)^3$.

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

Under algorithm W_1, the qualities of *Mobile* and *Foreman* are similar although they have very different rate-distortion properties. This is because W_1 simply maximizes the rate without considering the resulting video quality. Since the benchmark algorithm W_2 does not dynamically organize the video packets into different subflows or change the weights according to the run-time transmission results, it achieves inferior results compared with our proposed algorithms ($W_{\Gamma 2}$ to $W_{\Gamma 4}$).

Compared to the benchmark algorithms W_1 and W_2, algorithm W_{rd} actually decreases the average video quality among different users. This is due to the deadline approaching effect explained in Section IV-B. Once we take care of this effect by properly chosen Γ functions in $W_{\Gamma 1}$ to $W_{\Gamma 3}$, the average

[3] We assume that the transmitter chooses the appropriate modulation and coding schemes to match the channel conditions of each user such that there is no data corruption during the transmission.

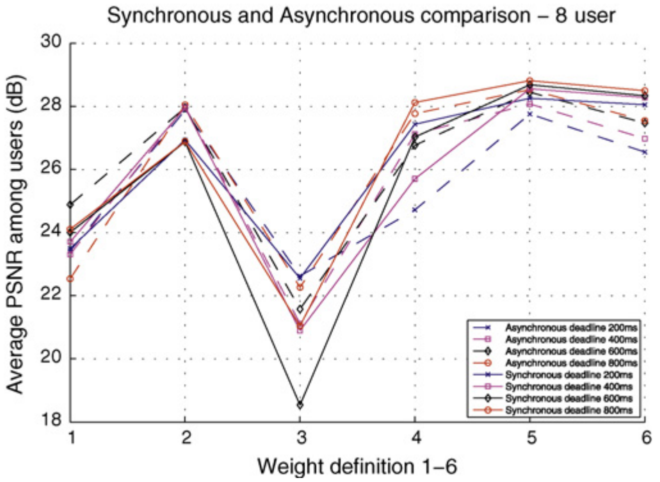

Fig. 2. Synchronous and asynchronous deadlines for eight users: 1, W_1; 2, W_2; 3, W_{rd}; 4, $W_{\Gamma 1}$; 5, $W_{\Gamma 2}$; 6, $W_{\Gamma 3}$; 7, $W_{\Gamma 4}$.

PSNR among users is improved over the simple total rate maximization scheme (W_1) by 1.1 dB to 1.6 dB.

C. Different Initial Playback Deadlines, User Contents and Congestion Range's Influence

Fig. 2 shows the results of eight users requesting video sequences concurrently. Each of the four video sequences is requested by two users. In this figure, the impacts of different initial playback deadlines, user contents, and congestion range are checked. We test various initial playback deadlines between 200 ms to 800 ms.

We first test the cases of synchronous deadlines, i.e., all users start requesting the video streaming applications at the same time. In reality, it is more common that different users request video clips at different time, which we call asynchronous deadlines cases. In the figure, we compare the results of these cases by letting the users randomly start to request the different video sequences from the server within the first initial playback deadline. We observe that the $W_{\Gamma 2}$ algorithm always performs the best.

The effectiveness of our proposed algorithms is more noticeable compared to the rate maximization algorithm W_1 in the heavily congested network case. For asynchronous cases with playback deadline of 800 ms, algorithm $W_{\Gamma 2}$ achieves as high as 6 dB improvement in users' average PSNR value. In the asynchronous cases, the advantage of proposed algorithm is not so obvious as compared to algorithm W_2. This is because the congestion of network is so heavy that "GOP control" is almost as effective as the deadline approaching control. Besides, little can be exploited by dynamically adapting weights according to the video rate-distortion properties.

VI. Conclusion

In this letter, we apply a cross-layer design approach to solve the challenging problem of multiuser 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 in our previous work, we proposed a family of algorithms 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 various wireless practical constraints into consideration. Simulation results show that our algorithms always outperform the rate maximization (content-blind) scheme and the pure gradient-based (deadline-blind) scheme. The performance of the algorithms is consistent under both synchronous or asynchronous deadlines.

As part of the future work, we plan to extend the proposed algorithms to the case of uplink OFDM systems and multicell downlink OFDM systems, where efficient elastic data transmission schemes have been already proposed [18], [19].

References

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[2] 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 Comput.*, vol. 5, no. 6, pp. 755–768, Jun. 2006.

[3] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288–296, Jan. 2009.

[4] J. Reichel, H. Schwarz, and M. Wien, *Joint Scalable Video Model 8 (JSVM 8)*, Joint Video Team, Doc. JVT-X202.

[5] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Content-aware resource allocation and packet scheduling for video transmission over wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 749–759, May 2007.

[6] P. Pahalawatta, T. N. Pappas, R. Berry, T. Pappas, and A. Katsaggelos, "Content-aware resource allocation for scalable video transmission on multiple users over a wireless network," in *Proc. IEEE Acoust. Speech Signal Process. (ICASSP 2007)*, Honolulu, HI, Apr. 2007, pp. I-853–I-856.

[7] H. Jin, R. Laroia, and T. Richardson, "Superposition by position," in *Proc. IEEE Inf. Theory Workshop 2006 (ITW 2006), Preprint*, pp. 222–226.

[8] G. Liebl, T. Schierl, T. Wiegand, and T. Stockhammer, "Advanced wireless multiuser video streaming using the scalable video coding extensions of H.264/MPEG4-AVC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Toronto, ON, Canada, 2006, pp. 625–628.

[9] IEEE Std 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks–Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 2004.

[10] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operat. Res.*, vol. 53, no. 1, pp. 12–25, 2005.

[11] J. Lee, H. Lou, and D. Toumpakaris, "Analysis of phase noise effects on time-direction differential OFDM receivers," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 5, no. 2. Dec. 2005, p. 2679.

[12] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile video transmission using scalable video coding (SVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1204–1217, Sep. 2007.

[13] H. Chen, P. Lee, and S. Hu, "Improving scalable video transmission over IEEE 802.11e through a cross-layer architecture," in *Proc. Int. Conf. Wireless Mobile Commun.*, 2008, pp. 241–246.

[14] C. De Vleeschouwer, J. Chakareski, and P. Frossard, "The virtue of patience in low-complexity scheduling of packetized media with feedback," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 348–365, Feb. 2007.

[15] O. Hillestad, A. Perkis, V. Genc, S. Murphy, and J. Murphy, "Adaptive H.264/MPEG-4 SVC video over IEEE 802.16 broadband wireless net-

works," in *Proc. Packet Video 2007*, Lausanne, Switzerland, Nov. 2007, pp. 26–35.

[16] Y. P. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. M. Alnuweiri, "A link adaptation scheme for efficient transmission of H.264 scalable video over multirate WLANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 875–887, Jul. 2008.

[17] Z. Ahmad, S. Worrall, and A. Kondoz, "Unequal power allocation for scalable video transmission over WiMAX," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hannover, Germany, 2008, pp. 517–520.

[18] J. Huang, V. Subramanian, R. Berry, and R. Agrawal, "Scheduling and resource allocation in OFDMA wireless communication systems," in *Orthogonal Frequency Division Multiple Access*, to be published by Auerbach Publications, CRC Press, Taylor and Francis Group.

[19] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 2, pp. 226–234, Feb. 2009.