10

Scheduling and Resource Allocation for SVC Streaming over OFDM Downlink Systems

Xin Ji[1], Jianwei Huang[2], Mung Chiang[3], Gauthier Lafruit[4] and Francky Catthoor[5]

[1,5]*IMEC/University of Leuven*

[2]*the Department of Information Engineering,The Chinese University of Hong Kong*

[3]*the Department of Electrical Engineering, Princeton University*

[4]*IMEC*

[1,4,5]*Belgium*

[2]*Hong Kong*

[3]*NJ, USA*

1. Introduction

The demand of video transmission over wireless 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, where the available network resource is highly dynamic and typically limited in terms of supporting high quality multimedia applications. This makes it challenging to achieve efficient multi-user video streaming over wireless channels.

In this chapter, we consider the problem of multi-user video streaming over Orthogonal Frequency Division Multiplexing (OFDM) networks, where videos are coded in Scalable Video Coding (SVC) format. OFDM is a promising technology for future broadband wireless networks, due to many of its advantages such as robustness against intersymbol interference and the usage of lower complexity equalization at the receiver. It is suitable for supporting high spectrum efficiency communications, and thus is chosen as 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) (34). The resource allocation in OFDM is done in the dimension of power, frequency, and time, and thus is very flexible. SVC, on the other hand, is one of the most promising technologies to enable high coding performance and flexibility (29). It has the attractive capabilities of reconstructing lower resolution or lower quality signals from partially received bitstreams, and hence provides flexible solutions for transmission over heterogeneous networks and allows easy adaptation to various storage devices and terminals. In this chapter, we focus on designing efficient multi-user video streaming protocols that fully exploit the resource allocation flexibility in OFDM and performance scalabilities in SVC.

Most of the previous work on downlink resource allocation in OFDM system focused on elastic data transmissions, where users do not have stringent deadline constraints (e.g., (5; 8; 19; 35; 41)). In (35), the goal was to minimize the total transmit power given users'

target bit rates. In (41), the authors investigated the downlink throughput maximization problem with dynamic sub-carrier allocation and fixed power allocation. (19) also considered maximizing the sum-rate without any minimum bit-rate target. In (8), the authors proposed Best Sub-carrier Allocation (BSA) for voice and data users that utilizes the feedback of the radio channel quality and sorts users to choose sub-carrier based on their radio channel feedback. Moreover, (5; 19; 41) considered suboptimal heuristics that use a constant power per sub-carrier. However, due to the deadline requirement feature of real-time video applications, these solutions may not be optimal for delivering multi-user, delay-constrained, real-time streaming video applications.

Video transmission over OFDM channels has been studied recently (15; 36). However, neither of these results considered power allocation, which is critical to wireless multimedia data transmission. 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. In (10), sub-carriers and power are allocated based on rate-distortion model. In (31), video distortion is minimized by considering power and sub-carrier constraints in OFDM systems. Neither (10) nor (31) explicitly took the delay constraint into account.

SVC standard brings various scalabilities (e.g. temporal, spatial, and quality) through adaptation of the bit stream, thus is particularly relevant in heterogeneous network contexts. One niche area of the application of SVC is the transmission over wireless networks. There have been several research results reporting SVC transmission over wireless networks. Most of them focused on exploiting the scalable feature of SVC to provide QoS quarantee for the end users ((6; 28), and the references therein). In (11), the layered bitstream of SVC is exploited in conjunction with a specific congestion control algorithm for distributing video to subscriber stations of an 802.16 system. In (9), the rate distortion model proposed for H.264/AVC is extended to include the effect of random packet loss on the scalable video layers of SVC and the resulting overall video distortion. Reference (32) focused on maximizing the number of admitted users in the communication system by giving different priorities to different video subflows according to their importance. None of the aforementioned solutions for SVC transmission over wireless networks considered power control. An unequal power allocation scheme was proposed in (3) for the transmission of SVC packets over WiMAX communications channels. In (26), 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 chapter is to provide a framework for efficient multi-user 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 framework presented in ((2; 16; 30)). In previous work (13), an efficient downlink OFDM resource allocation algorithm for *elastic data* traffic has been successfully designed, which is provably optimal for long term utility maximization subject to stochastic channel variations of wireless networks. In this chapter, we generalize such framework to real-time video streaming by further considering dynamically adjusted priority weights based on the current video contents, deadline requirements, and the previous transmission results. The following steps are involved in the proposed joint optimization:

1. Unlike conventional wireless streaming approach, where video data is transmitted indifferently with the achievable rate, we divide the video data into subflows based on the contribution of distortion decrease and the delay requirements of individual video frames.

As discussed in Section 3, this allows the most important video data get transmitted with more priorities and avoid the waste of the network resources.

2. Based on the existing gradient related approach, the rate-distortion weighted transmission scheduling strategy is established in Section 4.3. Our proposed solution involves calculating the weights of the current subflows according to their rate-distortion properties, playback deadline requirements and the previous transmission results.
3. The inherent prioritization brought from the aforementioned weight definition is however conflict with the so-called deadline approaching effect. In Section 4.4, we proposed to deliberately add a product term to the weight calculation which increases when the deadline approaches. This allows the weights of the subflows with low rate-distortion ratio being gracefully increased when their playback deadline approach. We propose a family of algorithms and identify the best tradeoff between meeting deadlines and maximizing the overall video quality.

The resulting algorithms not only fully utilize the temporal and quality scalabilities of the SVC scheme, but also thoroughly explore the time, frequency and multi-user 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.
The remainder of this chapter is organized as follows. Section 2 introduces the OFDM network model. Section 3 describes the SVC scheme. Section 4 describes the problem formulation and the proposed algorithms. In Section 5, we examine the performance of our proposed solutions through simulations. Concluding remarks are given in Section 6.

2. OFDM model of the wireless transmission

The OFDM network model considered here is similar as in (13). Different video bitstreams are transmitted from the base station to a set $\mathcal{I} = \{1, \ldots, I\}$ of mobile users in an OFDM cell. Time is divided into TDM time-slots that contain 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 $\boldsymbol{r}_t = (r_{1,t}, \ldots, r_{I,t})$ from the current feasible rate region $\mathcal{R}(\boldsymbol{e}_t) \subseteq \mathbb{R}_+^K$, where $\boldsymbol{e}_t$ indicates the time-varying channel state information available at the scheduler at time t. For presentation simplicity, we omit the time index t in the following.
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. We denote the power allocated to user i on tone j as p_{ij} and the fraction of time that tone allocated to user i as x_{ij}. The total power allocation must satisfy $\sum_{i,j} p_{ij} \leq P$, i.e., the total downlink power constraint at the base station. The total allocation for each tone j must satisfy $\sum_i x_{ij} \leq 1$. For a given allocation with perfect channel estimation, user i's feasible rate on tone j is $r_{ij} = x_{ij} B \log(1 + \frac{p_{ij} e_{ij}}{x_{ij}})$, which corresponds to the Shannon capacity of a Gaussian noise channel with bandwidth $x_{ij}B$ and received SNR $p_{ij}e_{ij}/x_{ij}$.[1] 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 set bandwidth $B = 1$ in the following analysis.
In practical OFDM networks, imperfect carrier synchronization and channel estimation may result in "self-noise" (e.g. (20; 22)). With self-noise, user i's feasible rate on tone j becomes

[1] To better model the achievable rates in a practical system we can re-normalize e_{ij} by γe_{ij}, where $\gamma \in [0, 1]$ represents the system's "gap" from capacity.

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

where $\beta << 1$ is the self-noise coefficient. Under these assumptions, we have

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

where $\mathcal{X} := \prod_{j=1}^{N} \mathcal{X}_j$, and for all $j \in \mathcal{J}$,

$$\mathcal{X}_j := \left\{ (\boldsymbol{x}^j, \boldsymbol{p}^j) \geq \boldsymbol{0} : x_{ij} \leq 1, p_{ij} \leq \frac{x_{ij}\tilde{s}_{ij}}{\tilde{e}_{ij}}, \; \forall i \in \mathcal{I}, \forall j \in \mathcal{J} \right\}, \tag{2}$$

with $\boldsymbol{x}^j := (x_{ij}, \forall i \in \mathcal{I})$ and $\boldsymbol{p}^j := (p_{ij}, \forall i \in \mathcal{I})$. Here, $\tilde{s}_{ij} = \frac{\Gamma_{ij}}{1 - \Gamma_{ij}\beta}$, where $\Gamma_{ij} < 1/\beta$ is a maximum SNR constraint on tone j for user i, e.g., to model a constraint on the maximum rate per tone due to a limitation on the available modulation and coding schemes. [2]
We assume that $\tilde{e}_{ij}$ is known by the scheduler for all i and j as is β (equivalently, the estimation error variance). In a frequency division duplex (FDD) system, this knowledge can be acquired by having the base station transmit pilot signals, from which the users can estimate their channel gains and feedback to the base station. In a time division duplex (TDD) system, these gains can also be acquired by having the users transmit uplink pilots; the base station can then exploit reciprocity to measure the channel gains. In both cases, this feedback information would need to be provided within the channel's coherence time.

3. SVC Scheme of video coding

SVC is an extension of the H.264/MPEG4-AVC video coding standard (33) and provides three different scalabilities: spatial, temporal, and quality. An overview of the features and applications of SVC can be found in (29). In this chapter, we focus on how to exploit the temporal and quality salabilities by adaptive scheduling and resource allocation.[3]
In SVC, the video frames are usually divided into groups, or called groups of pictures (GOPs). The typical SVC GOP structure is shown in Fig. 1, where we assume that one GOP consists of 4 frames. The video frames are further encoded into different temporal and quality layers. One box in Fig. 1 represents the data belonging to one specific temporal layer and one specific quality layer. For the purpose of video distortion calculation, we regard a box as the smallest decodable data unit and call it a "packet". All the packets in one column represent one frame. For example, frame L_1 consists of three packets: L_{10}, L_{11}, and L_{12}.
The packets at the same horizontal level belong to the same quality layer. The *quality scalability* refers to the fact that a video decoder can reconstruct video sequences without receiving all quality layers. After receiving the base 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 required playback deadline of the corresponding video frames. In Fig. 1, the dashed arrows depict the enhancement layers order for each video frame.

[2] Another important practical constraint is that each subchannel can be allocated to at most one user, i.e., $x_{ij} \in [0, 1]$. For simplicity, we do not consider such constraint in this chapter. Interested readers are referred to (13) for related detailed discussions.

[3] 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.

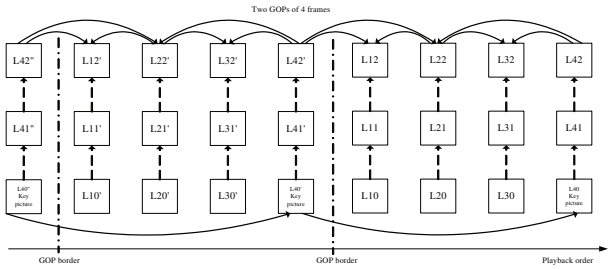

Fig. 1. GOP structure of SVC.

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 hierarchical B pictures scheme. In Fig. 1, the solid arrows depict the motion predictions for each frame. 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} becomes decodable 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.
The quality and temporal scalabilities provide the possibility of adapting the video transmission to different network environments. 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 sizes of the packets at different quality and temporal layers are typically different. Because 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 video data with same priority, and use that to facilitate the scheduling and resource allocation decisions.
Let's assume the GOP size is g. The total number of temporal levels within a GOP is $\log_2 g$ then. Also we use $P^{t,q,k}$ to denote the packet that belongs to frame k, quality layer q, and temporal level t in the current GOP. Here $1 \leq k \leq g$, $1 \leq t \leq \log_2 g$, and $0 \leq q \leq Q$. Normally we have $Q \leq 3$ (29). We group the packets with the same deadline as one *subflow* in a way similar as that proposed in (32). For example, in Figure 1, suppose all the packets that are necessary for decoding frame L_1 to be one subflow. This subflow consists of packet 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''_{40}, L''_{41}, L''_{42} ... etc.), L_{20}, L_{21}, L_{22}, L_{10}, L_{11}, L_{12}. Different from the subflow concept in (32), 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 former GOPs), L_{20}, L_{10} belong to the base layer of the current subflow. Other packets belong to the enhancement layers 1 and 2, respectively. This allows us to accurately capture the rate requirements of different packets within one GOP.

4. Scheduling and resource allocation algorithms

4.1 Gradient-based scheduling framework

Consider a media server that is connected to the base station through a high bandwidth backbone network. Each of the K 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 has high bandwidth, thus 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.[4] If the subflow 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 overall network streaming quality in the long run, under time varying channel conditions and variable rate video contents.

Our starting point is the stochastic gradient-based scheduling framework presented in (2; 16; 30). 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 $\boldsymbol{r}_t$ in $\mathcal{R}(\boldsymbol{e}_t)$ 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 obtain

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_{i \in \mathcal{I}} \frac{\partial U_i(W_{i,t})}{\partial W_{i,t}} r_{i,t}. \tag{3}$$

The above policy has been shown to yield utility maximizing solutions under time-varying rate region (2; 16; 30), i.e., maximizing $\sum_{i \in \mathcal{I}} U_i(W_{i,t})$. 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 (3) needs to be solved for each time slot.

In (13), we proposed an efficient algorithm to solve Problem (3) for an OFDM downlink system with elastic data transmission. Next in Section 4.2 we will briefly review the proposed algorithm in (13). Then in Section 4.3 we will explain the special challenges introduced by the real-time streaming applications and discuss how the algorithm in (13) can be generalized to our case.

4.2 Weighted rate maximization algorithm under fixed weights

Consider a given time slot t, where we define $w_{i,t} = \partial U_i(W_{i,t})/\partial W_{i,t}$. According to (1), Problem (3) can be stated as follows,

$$\begin{aligned} \max_{(\boldsymbol{x},\boldsymbol{p}) \in \mathcal{X}} \; & V(\boldsymbol{x},\boldsymbol{p}) := \sum_i w_i \sum_j x_{ij} \log\left(1 + \tfrac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right) \\ \text{subject to:} \quad & \sum_{i,j} p_{ij} \le P, \text{ and } \sum_i x_{ij} \le 1, \, \forall j \in \mathcal{N}. \end{aligned} \tag{4}$$

Here we omit time index t for simplicity. We can solve this problem via a dual decomposition method (4) with complexity $O(NK)$.

First consider the Lagrangian,

$$L(\boldsymbol{x},\boldsymbol{p},\lambda,\boldsymbol{\mu}) := \lambda P + \sum_{j=1}^{N} L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j), \tag{5}$$

where

$$L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j) := \mu_j + \sum_{i=1}^{K} w_i x_{ij} \log\left(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right) - \mu_j \sum_{i=1}^{K} x_{ij} - \lambda \sum_{i=1}^{K} p_{ij}, \tag{6}$$

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

and $\boldsymbol{\mu} = (\mu_j)_{j=1}^N$. The corresponding dual function is

$$L(\lambda, \boldsymbol{\mu}) := \max_{(\boldsymbol{p},\boldsymbol{x})\in\mathcal{X}} L(\boldsymbol{x}, \boldsymbol{p}, \lambda, \boldsymbol{\mu}) = \lambda P + \sum_{j=1}^{N} \max_{(\boldsymbol{p}^j,\boldsymbol{x}^j)\in\mathcal{X}_j} L_j(\boldsymbol{x}^j, \boldsymbol{p}^j, \lambda, \mu_j).$$

Since Problem (4) is convex and satisfies Slater's condition, there is no duality gap and so $V^* := \min_{\lambda\geq 0, \boldsymbol{\mu}\geq \mathbf{0}} L(\lambda, \boldsymbol{\mu})$ is the optimal objective value (4).
First, we show that the dual function can be calculated in closed form. Define[5]

$$q(\beta, z) := \begin{cases} z, & \text{if } \beta = 0, \\ \left(\frac{2\beta+1}{2\beta(\beta+1)}\right)\left(\sqrt{1+\frac{4\beta(\beta+1)}{(2\beta+1)^2}z} - 1\right), & \text{if } \beta > 0, \end{cases}$$

$$h\left(\beta, \omega, \tilde{s}_{ij}\right) := \log\left(1 + \frac{q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}}{1+\beta\left(q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}\right)}\right) - \frac{1}{\omega}\left(q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}\right).$$

and $\mu_{ij}(\lambda) := w_i h\left(\beta, \frac{w_i\tilde{e}_{ij}}{\lambda}, \tilde{s}_{ij}\right)$. Then the dual function is

$$L(\lambda, \boldsymbol{\mu}) := \lambda P + \sum_{j=1}^{N} L_j(\lambda, \mu_j), \tag{7}$$

where $L_j(\lambda, \mu_j) := L_j(\boldsymbol{x}^{j,*}, \boldsymbol{p}^{j,*}, \lambda, \mu_j) = \sum_i \left(\mu_{ij}(\lambda) - \mu_j\right)^+ + \mu_j$.
Second, we can further simplify the dual function by optimizing over $\boldsymbol{\mu}$, i.e.,

$$L(\lambda) := \min_{\boldsymbol{\mu}\geq\mathbf{0}} L(\lambda, \boldsymbol{\mu}) = \lambda P + \sum_j \mu_j^*(\lambda), \tag{8}$$

where for every tone j, the minimizing value of μ_j^* is achieved by $\mu_j^*(\lambda) = \max_i \mu_{ij}(\lambda)$.
Since $L(\lambda)$ is the minimum of a convex function over a convex set, it is a convex function of λ and can be solved numerically. The overall dual-based algorithm involves evaluating $L(\lambda)$ for a fixed value of λ as an inner loop, and a one-dimensional search over λ as an outer loop. The outer loop has a constant complexity that is independent of J and I[6]. The inner loop has a complexity of $O(JI)$ due to searching for the maximum of I metrics on each of the J tones. Thus the total complexity of this stage is $O(JI)$. Details of the algorithm can be found in (13).

4.3 Dynamic weight calculation for streaming applications

The algorithm presented in Section 4.2 solves the weighted rate maximization problem under fixed weights. For elastic data applications, the weights are calculated as the gradients of the utility functions. This weight calculation method, however, is not suitable for real-time video streaming application since the stringent delay constraints are not explicitly considered. This motivates us to design a different weight calculation method in this chapter, which will be based on the required 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, its length is l_i bits and the playback deadline is $t_i > 0$.

[5] Here $(x)^+ = \max(x, 0)$ and $x \wedge y = \min(x, y)$.
[6] The computational complexity of a bi-section search is $O(\log(1/\epsilon))$, where ϵ is the relative error bound target for the search.

In order to meet the deadline, the subflow needs to be transmitted at an average rate of

$$\hat{r}_i = \frac{l_i}{t_i}. \tag{9}$$

Note that this may not be the actual rate that user i gets, which depends on the resource allocation decisions.

Denote the distortion of the corresponding frame is D_{ic} if the current subflow can be 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 $t = 0$. This can be calculated as follows:

1. If some of the base layer packets within the current subflow have not been received by the users at time $t = 0$, then the receiver will use the last decodable frames to substitute the desired frames and achieves distortion $D_{il} (> D_{ic})$ at time $t = 0$. In this case, successfully delivering the current subflow on time can lead to distortion decrease of

$$\Delta D_i = D_{il} - D_{ic}. \tag{10}$$

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 layer 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 can lead to distortion decrease of

$$\Delta D_i = D_{il} - D_{iq}. \tag{11}$$

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

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i} = \frac{\Delta D_i}{l_i} t_i. \tag{12}$$

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.

Nevertheless, using the weight definition of (12) and solving Problem (3) 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 (12) 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 approaching deadline will have higher priority. As a result, weighted rate maximization based on (12) will give users in good channels extra advantage.

For users with the same weight, a user in good channel condition requires less resource to achieve the same transmission rate and thus is favorable. 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 5 also confirm this problem.

To tackle this problem, we next propose a framework to explicitly consider the effect of approaching deadline, which can enforce the deadline to be satisfied with high probability while still achieving an overall good video quality.

4.4 Mitigating the approaching deadline effect

We propose to explicitly add 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 enforces the system to allocate more resources to "urgent" users and reduce deadline violations. The new priority weight can be calculated as:

$$w_{i,t} = \frac{\Delta D_i}{\hat{r}_i} \Gamma(t_i). \tag{13}$$

where the delay function Γ decreases with t_i. One choice that achieves the best overall performance in our simulation is to have

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

We will give more examples of function Γ in Section 5.

4.5 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 rate each active user gets). For each time slot t, there are three key steps in the algorithm:

1. The priority weight of each user is calculated based on its previous transmission results and the deadline of the current subflow.
2. The base station performs the scheduling and resource allocation based on users' priority weights using the algorithm in Section 4.2.
3. Each user transmits the packets based on the allocated resource.

According to the way that the subflow is defined in Section 3, 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 are not decodable due to lack of necessary dependent packets.[7] It is clear that all three steps converge, thus we know that Algorithm 1 converges.

The computational complexity of the proposed algorithm comes from three parts:

1. Merging the remaining packets with the next subflow. The worst case complexity of this step is $O(g(Q+1))$, where 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 $O(Ig(Q+1))$, where I is the total number of users.
2. Calculating the priority weight $w_{i,t}$ according to(13). 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 distortion decrease needs

[7] 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.

```
initialization t = 0;
repeat
    t = t + 1;
    forall the user i do
        repeat
            check the deadline of the current subflow;
            if the deadline has passed then
                discard those packets not useful for decoding future packets;
                merge the remaining packets with the next subflow, which becomes
                the "current" subflow;
            end
            Calculate the priority weight w_{i,t} according to (13)
        until the deadline of the current subflow has not passed;
    end
    Solve weighted rate maximization problem (4) using the algorithm described in
    Section 4.2, and each user i is allocated transmission rate r_{i,t};
    forall the user 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
    end
until no more video to be streamed;
```

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

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 (4), which has complexity $O(IJ)$, where J is the total number of subchannels.

The overall complexity of the algorithm for each time slot t is then $O(I(J + g(Q+1))$.

5. Simulation study

5.1 Simulation setup

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

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 dramatically different levels of motion activities. The rate and the quality of the different sequences are shown in Table 1. 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.

Sequence	Bitrate	Average PSNR
Mobile	2019 kbps	35.17 dB
Foreman	449.2 kbps	35.16 dB
City	585.8 kbps	35.98 dB
Harbour	1599.7 kbps	35.32 dB

Table 1. Encoding rates and average PSNRs of different sequences

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 (17). 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 (17). The resource allocation is done once per fading block. For each video sequence, we report results that are averaged over 5 randomly generated channel realizations with a length of 10 seconds each (which corresponds to 10^5 OFDM symbols).

5.2 Different weight definitions

We simulate the algorithm with different counter-deadline approaching effect functions Γ when calculating the weights $w_{i,t}$ in (13). To illustrate the effectiveness of our proposed algorithm, we also compared with rate maximization algorithm and the algorithm proposed in (26). In total, we simulate seven algorithms. The first two algorithms are benchmark algorithms, and the last five algorithms are our proposed ones with different levels of emphasis on deadline violation avoidance. We will show that algorithm $W_{\Gamma 2}$ achieves the best performance among all proposed ones.

- 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., (13)). On top of this, we use the packet dropping policy for SVC proposed in (24).
- W_2 (benchmark 2: deadline-blind approach): the weights in this approach are defined according to (26). Instead of grouping packets into subflows, the schedular will transmit every packet following the order of Method II proposed in (27), which has been proven to achieve similar results as the optimal one. Though special care has been taken to

force every new GOP data to be buffered either after the current GOP's deadline is expired or until all the current GOPs of each user have been transmitted, compared to our subflow scheme (which explicitly consider the deadline of each frame), it is considered as a "deadline-blind" benchmark.

- 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 2 shows average PSNR achieved by four users requesting four different video clips with the same starting time. The initial playback deadline is set to be 200ms (25).

Sequence	W_1	W_2	W_{rd}	$W_{\Gamma 1}$	$W_{\Gamma 2}$	$W_{\Gamma 3}$	$W_{\Gamma 4}$
Mobile	28.5316	26.7014	18.6482	20.6136	28.0960	27.6642	27.4646
Foreman	29.0880	30.7430	27.2240	30.6424	33.5992	33.2444	33.0476
City	34.2552	31.0290	33.5274	34.1902	34.0882	33.8188	33.6754
Harbour	23.5310	26.9150	20.1732	21.6224	26.1610	26.0774	25.9670
Average	28.8514	28.8470	24.8932	26.7672	30.4862	30.2012	30.0388

Table 2. Average PSNR for 4 users with 200ms initial playback deadline

As we can see, the weighted gradient based scheduling reflects the rate-distortion properties of different video contents. Under W_1 algorithm, 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. 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. Since the benchmark algorithm W_2 does not dynamically organize the video packet into different subflow or change the weights according to the run-time transmission results, it achieves inferior results compared to our proposed algorithms ($W_{\Gamma 2}$ to $W_{\Gamma 4}$).

Compared to the benchmark W_1 and W_2 algorithms, the W_{rd} algorithm actually decreases the average video quality among different users. This is due to the deadline approaching effect explained in Section 4.3. Once we take care of this effect by properly chosen 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 1.1 dB to 1.6 dB. Results of $W_{\Gamma 2}$ reaches 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.

Figures 2, 3, 4 and 5 show the PSNR values of the first 200 frames achieved by four users requesting different four video clips concurrently under a particular channel realization. The initial playback deadline is set to be 400ms. In these figures, the results of weight definition W_1, W_2 and our best approach $W_{\Gamma 2}$ are compared.

From the figures, we see that compared to algorithm $W_{\Gamma 2}$, algorithm W_1 only considers rate maximization and hence user 2 and user 4's video qualities are sacrificed. Some of the frames' PSNR value of user 1 and user 3 may be higher than those of our proposed algorithm, however without significant performance improvement compared to the video quality of the proposed ones. This proves that our proposed rate-distortion related gradient based scheme is more efficient.

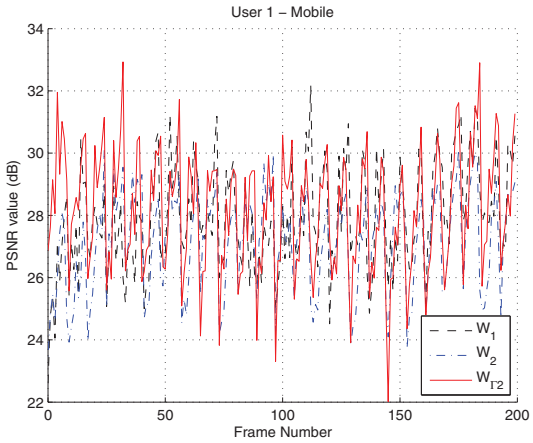

Fig. 2. Frame PSNR of User 1 - Mobile.

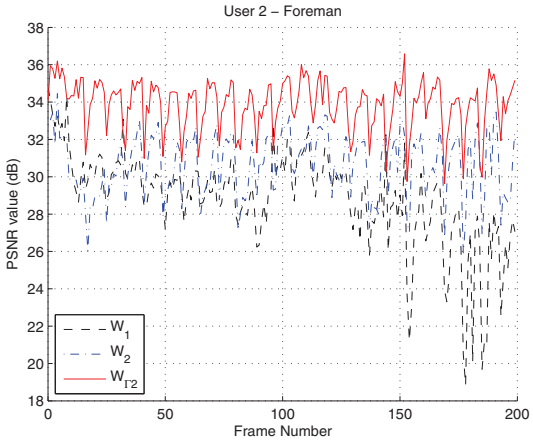

Fig. 3. Frame PSNR of User 2 - Foreman.

5.3 Effect of different initial playback deadlines

We now study the impact of different initial playback deadlines in Figure 6. The initial playback deadline means the delay between the time when the user requests the video and the time when the video starts to play at the receiver. According to the user satisfactory study in (25), we test various initial playback deadlines between 200ms to 800ms. Four users request the different video sequences from the server simultaneously. Other parameters are the same as in Section 5.1. We can see that $W_{\Gamma 2}$ always reaches the highest average PSNR value under different deadlines.

5.4 Synchronous and asynchronous requirements' influence

So far we have only considered the cases of synchronously deadlines, i.e., all users start requesting the video streaming applications at the same time. In reality, it is more common

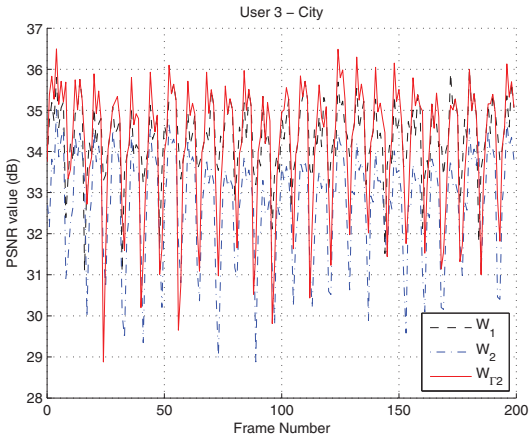

Fig. 4. Frame PSNR of User 3 - City.

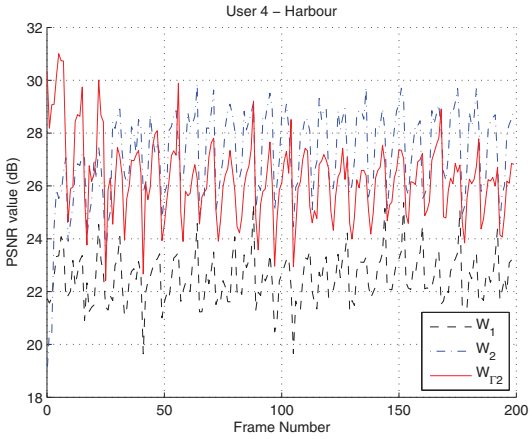

Fig. 5. Frame PSNR of User 4 - Harbour.

that different users request video clips at different time, which we call *a*synchronous deadlines cases. In Figure 7, we compare the results of these cases for four users. In the asynchronous deadline cases, four users randomly start to request the different video sequences from the server within the first initial playback deadline. We again observe that the $W_{\Gamma 2}$ algorithm always performs the best.

5.5 Different user content and congestion range's influence

Figure 8 shows the results of eight users requesting video sequences concurrently. Each of the 4 video sequences is requested by 2 users. Synchronous and asynchronous cases are both shown here. For the asynchronous cases, users randomly request the video sequence within one playback deadline. The other setups are the same as in Section 5.2.

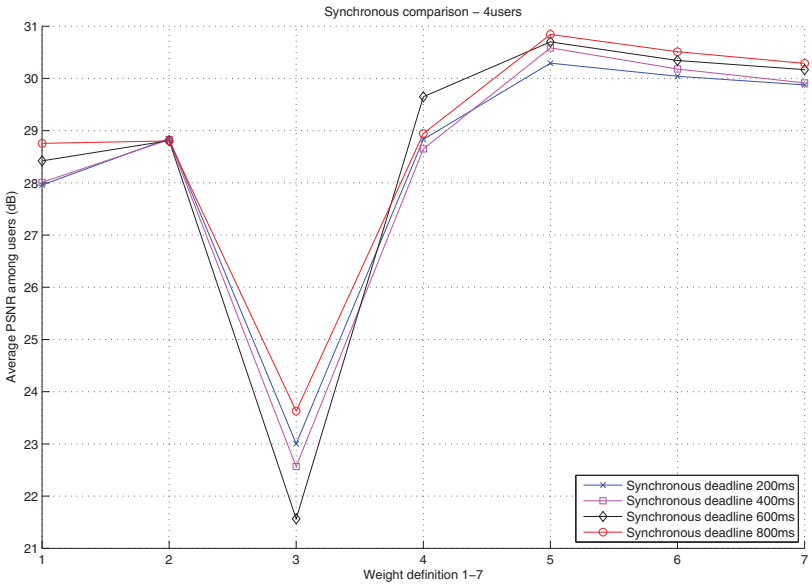

Fig. 6. Synchronous deadlines for 4 users. Horizontal axis represent different algorithms: 1 - W_1; 2 - W_2; 3 - W_{rd}; 4 - $W_{\Gamma 1}$; 5 - $W_{\Gamma 2}$; 6 - $W_{\Gamma 3}$; 7 - $W_{\Gamma 4}$;

The effectiveness of our proposed algorithms is more obvious compared to the rate maximization algorithm W_1 in heavily congested network case. For asynchronous cases with playback deadline of 800ms, algorithm $W_{\Gamma 2}$ achieves as high as 6dB improvement in users' average PSNR value. In the asynchronous cases, the advantage of porposed alogirthm 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.

5.6 Fairness analysis

Motived by the Jain's fairness index (18), we propose the following index to evaluate the fairness of video qualities achieved by different algorithms:

$$VideoQualityFairness = \frac{(\sum_i PSNR_i)^2}{n \sum_i (PSNR_i)^2} \tag{14}$$

The fairness index ranges from 1/n (worst case) to 1 (best case). For each algorithm, we show the fairness index of different simulation settings in Tables 3, 4, 5 and 6. In all cases, Algorithm W_2 always achieves the highest fairness index. However, Table 2 shows that it achieves so by sacrificing the video quality. All of our five proposed algorithms achieve a fairness index of more than 0.98 most of the time. We also find that W_{rd} always has the worst fairness property, which means this algorithm does not consider the fairness but only emphasizes on the rate-distortion property. In fact, both considering the deadline approaching effect and

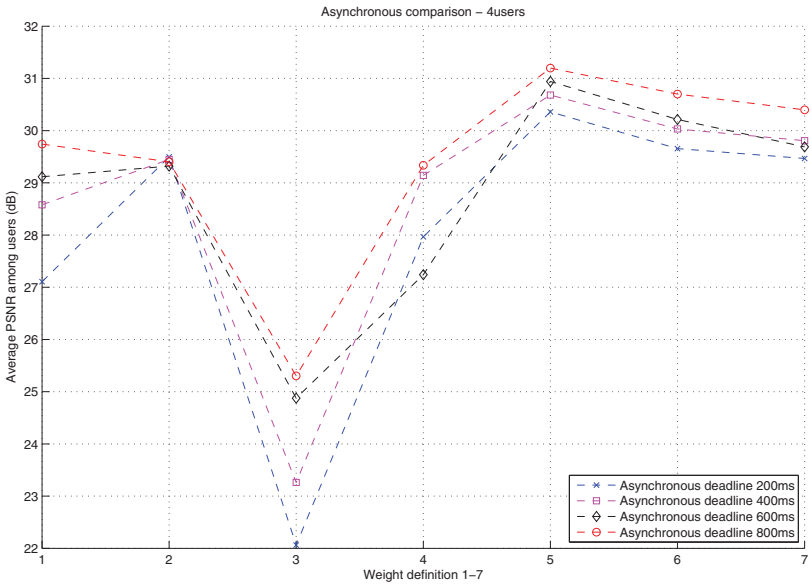

Fig. 7. Asynchronous deadlines for 4 users. Horizontal axis represent different algorithms: 1 - W_1; 2 - W_2; 3 - W_{rd}; 4 - $W_{\Gamma 1}$; 5 - $W_{\Gamma 2}$; 6 - $W_{\Gamma 3}$; 7 - $W_{\Gamma 4}$;

using "GOP control" can improve fairness. The "GOP control" benchmark algorithm (W_2) pursues absolute fairness, thus decreases the overall video quality.

Weight	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
W_1	0.9848	0.9354	0.986	0.941	0.9426
W_2	0.995	0.9953	0.9962	0.9898	0.9942
W_{rd}	0.9423	0.9328	0.8281	0.9314	0.9341
$W_{\Gamma 1}$	0.9799	0.9162	0.9287	0.9544	0.9335
$W_{\Gamma 2}$	0.9856	0.9845	0.9868	0.9806	0.9818
$W_{\Gamma 3}$	0.9873	0.9855	0.9877	0.9797	0.982
$W_{\Gamma 4}$	0.9869	0.9848	0.988	0.9794	0.9821

Table 3. 4 users with synchronous initial playback deadline of 200ms

6. Conclusion

Traditionally the content distribution and network resource allocation are designed separately. Although working well in the wireline communication settings, this approach could be far from optimal for wireless communication networks, where the available network resource changes rapidly in time. In this chapter, 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 among the most promising technologies for video coding and wireless communications, respectively.

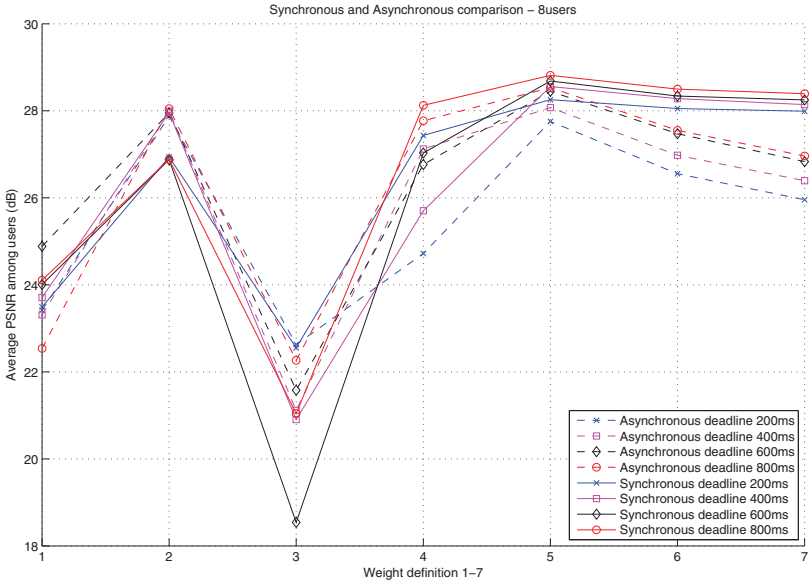

Fig. 8. Synchronous and Asynchronous deadlines for 8 users: 1 - W_1; 2 - W_2; 3 - W_{rd}; 4 - $W_{\Gamma 1}$; 5 - $W_{\Gamma 2}$; 6 - $W_{\Gamma 3}$; 7 - $W_{\Gamma 4}$;

Weight	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
W_1	0.9656	0.8883	0.9139	0.9342	0.9088
W_2	0.9938	0.9949	0.9888	0.9931	0.9951
W_{rd}	0.971	0.8965	0.9292	0.9373	0.947
$W_{\Gamma 1}$	0.9806	0.9804	0.9747	0.9751	0.9816
$W_{\Gamma 2}$	0.9839	0.9832	0.9773	0.9777	0.9861
$W_{\Gamma 3}$	0.9836	0.9836	0.9774	0.978	0.9859
$W_{\Gamma 4}$	0.9841	0.9829	0.9767	0.9779	0.9845

Table 4. 8 users with synchronous initial playback deadline of 200ms

Weight	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
W_1	0.9851	0.9172	0.9805	0.8884	0.9428
W_2	0.9945	0.9947	0.9963	0.9793	0.9936
W_{rd}	0.8701	0.9534	0.8264	0.8208	0.941
$W_{\Gamma 1}$	0.9725	0.9817	0.95	0.8494	0.9754
$W_{\Gamma 2}$	0.9851	0.9846	0.9869	0.9813	0.9824
$W_{\Gamma 3}$	0.9831	0.9805	0.9861	0.9804	0.9823
$W_{\Gamma 4}$	0.9834	0.9778	0.9867	0.9811	0.9823

Table 5. 4 users with asynchronous initial playback deadline of 200ms

Weight	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
W_1	0.9863	0.9414	0.9799	0.9417	0.9429
W_1	0.9949	0.9956	0.9956	0.9897	0.9936
W_{rd}	0.9437	0.9247	0.9436	0.9329	0.8332
$W_{\Gamma 1}$	0.9803	0.8958	0.9826	0.9763	0.9254
$W_{\Gamma 2}$	0.9853	0.9834	0.9893	0.982	0.9832
$W_{\Gamma 2}$	0.9861	0.9852	0.9893	0.9812	0.9823
$W_{\Gamma 2}$	0.9865	0.9847	0.9893	0.9818	0.9843

Table 6. 4 users with synchronous initial playback deadline of 800ms

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. We first divide the video data into subflows based on their contribution of distortion decrease and the delay requirements of individual video frames. Then we propose to calculate the weights of the current subflows according to their rate-distortion properties, playback deadline requirements and the previous transmission results. To tackle the deadline approaching effect, we also propose to explicitly add to the weight calculation a product term which increases when the deadline approaches.

Simulation results show that our algorithms always outperform the rate maximization (content-blind) scheme and the pure gradient-based (deadline-blind) scheme. Besides improving the average video quality, the proposed algorithms also lead to a fair allocation. Finally, the performance of the algorithms are consistent under both synchronous or asynchronous deadlines.

7. References

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

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

[3] Z. Ahmad, S. Worrall and A. Kondoz, "Unequal power allocation for scalable video transmission over WiMAX", *IEEE International Conference on Multimedia and Expo, ICME'08*, pp. 517-520, Hannover, Germany, 2008.

[4] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, Massachusetts: Athena Scientific, 1999.

[5] T. Chee, C. C. Lim, and J. Choi, "Adaptive Power Allocation with User Prioritization for Downlink Orthogonal Frequency Division Multiple Access Systems," in *Proc. of 9th IEEE International Conf. on Communication Systems*, pp. 210-214, Sept. 2004.

[6] Hsing-Lung Chen, Po-Ching Lee and Shu-Hua Hu, "Improving Scalable Video Transmission over IEEE 802.11e through a Cross-Layer Architecture," *The Fourth International Conference on Wireless and Mobile Communications*, pp. 241-246, 2008.

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

[8] N. Damji, T. Le-Ngoc, "Dynamic Downlink OFDM Resource Allocation with Interference Mitigation and Macro Diversity for Multimedia Services in Wireless Cellular Systems", WCNC 2005

[9] 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", *Circuits and Systems for Video Technology, IEEE Transactions on*, Volume: 18, Issue: 7, pp. 875-887, July 2008.

[10] Hojin Ha, Changhoon Yim and Young Yong Kim, "Distortion Management Scheme for Multiuser Video Transmission in OFDM Systems", *5th IEEE Consumer Communications and Networking Conference, CCNC'08*, pp. 795-799, Jan. 2008.

[11] O. Hillestad, A. Perkis, V. Genc, S. Murphy and J. Murphy, "Adaptive H.264/MPEG-4 SVC video over IEEE 802.16 broadband wireless networks," *Packet Video 2007*, pp. 26-35, Lausanne, Switzerland, Nov. 2007.

[12] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, "Multiuser Transmit Optimization for Multicarrier Broadcast Channels: Asymptotic FDMA Capacity Region and Algorithms," in *IEEE Trans. on Communications*, vol. 52, no. 6, pp. 922-930, June 2004.

[13] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink Scheduling and Resource Allocation for OFDM Systems," *IEEE Trans. on Wireless Commun., Accepted, 2008*.

[14] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," in *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1150-1158, Nov. 2003.

[15] Seoshin Kwack, Hanbyeol Seo and Byeong Gi Lee, "Suitability-Based Subcarrier Allocation for Multicast Services Employing Layered Video Coding in Wireless OFDM Systems", *IEEE 66th Vehicular Technology Conference, VTC-2007 Fall*, pp. 1752 - 1756, Sept. 30 2007-Oct. 3 2007.

[16] H. Kushner and P. Whiting, "Asymptotic properties of proportional-fair sharing algorithms," in *Proc. 40th Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2002.

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

[18] 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.

[19] J. Jang and K. B. Lee, "Transmit Power Adaptation for Multiuser OFDM System," in *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171-178, Feb. 2003.

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

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

[22] J. Lee, H. Lou and D. Toumpakaris, "Analysis of Phase Noise Effects on Time-Direction Differential OFDM Receivers," *IEEE GLOBECOM*, 2005

[23] 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.

[24] 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." ICME, 2006

[25] 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.

[26] 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.

[27] 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," *IEEE ICASSP'07*, Honolulu, USA, Apr. 2007.
[28] T. Schierl, T. Stockhammer and T. Wiegand, "Mobile Video Transmission using Scalable Video Coding (SVC)," *IEEE Trans. on Circuits and Systems for Video Technology, Special issue on Scalable Video Coding*, June 2007.
[29] 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.
[30] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," Operations Research, vol. 53, No. 1, pp. 12–25, 2005.
[31] 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 Circuits and Systems for Video Technology*, vol. 16, no. 10, pp. 1217–1231, 2006.
[32] 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 Transactions on Mobile Computing*, vol. 5, pp. 755-768, June 2006.
[33] 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.
[34] "Orthogonal frequency-division multiplexing," Wikipedia, http://en.wikipedia.org/w/index.php?title=Orthogonal_frequency-division_multiplexing&oldid=153352313.
[35] C. Y. Wong, R. S. Cheng, K. B. Letaief and R. D. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit and Power Allocation," in *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, Oct. 1999.
[36] J. Xu, X. Shen, J. W. Mark, and J, Cai, "Quasi-Optimal Channel Assignment for Real-Time Video in OFDM Wireless Systems," *Wireless Communications, IEEE Transactions on*, Volume 7, Issue 4, pp. 1417 - 1427, April 2008.
[37] Y. Yi and M. Chiang, "Stochastic network utility maximization: A tribute to Kelly's paper published in this journal a decade ago," *European Transactions on Telecommunications*, vol. 19, no. 4, pp. 421-442, June 2008.
[38] H. Yin and H. Liu, "An Efficient Multiuser Loading Algorithm for OFDM-based Broadband Wireless Systems," in *Proc. of IEEE Globecom*, pp. 103-107, Nov. 2000.
[39] 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.
[40] Y. J. Zhang and K. B. Letaief, "Adaptive Resource Allocation and Scheduling for Multiuser Packet-based OFDM Networks," in *Proc. of IEEE ICC*, pp. 2949-2953, June 2004.
[41] Y. J. Zhang and K. B. Letaief, "Multiuser Adaptive Subcarrier-and-Bit Allocation With Adaptive Cell Selection for OFDM Systems," in *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, Sept. 2004.