Cross-Layer Resource Allocation for Scalable Video Over OFDMA Wireless Networks: Tradeoff Between Quality Fairness and Efficiency

Kuan Lin and Sorina Dumitrescu, *Senior Member, IEEE*

***Abstract*—This work addresses the tradeoff between quality fairness and system efficiency for scalable video delivery to multiple users over OFDMA wireless networks. We consider a cross-layer optimization framework seeking to maximize the sum-PSNR corresponding to average user rates, subject to relaxed PSNR-fair constraints. More specifically, a pure quality-fairness (PF) problem is solved first to determine the maximum PSNR value obtained by imposing the same PSNR level to all users. Next the constraints in the PF problem are relaxed by allowing the relative difference between the PSNR of each video and the PF PSNR value to be within some range $[0, \sigma]$. Thus, the parameter σ controls the tradeoff between quality fairness and system efficiency. The PF problem is equivalent to the quality fairness problem proposed by Cicalò and Tralli, which was solved using a vertical decomposition approach. Further, we convert the optimization problem with the relaxed fairness constraints into a convex problem and solve it using established techniques. Our simulation results show that by varying the value of σ, a wide range, densely populated, of trade-off points between quality fairness and efficiency can be achieved. Additionally, a subjective quality assessment reveals that while the maximum efficiency scheme (ME), i.e., when $\sigma = \infty$, may compromise the quality of the high demanding videos, the PF scheme may sacrifice the quality of the low demanding videos. On the other hand, by providing a trade-off between PF and ME, the proposed scheme has the potential of finding a middle ground where all users are satisfied.**

***Index Terms*—Cross-layer, orthogonal frequency division multiple access (OFDMA), resource allocation, scalable video coding, quality fairness/efficiency trade-off.**

I. Introduction

With the advancement of video compression technology and the rapid development and deployment of network infrastructure, recent years have witnessed an unprecedented growth in demand for video services. According to recent forecasts [1], video will represent 72% of the total mobile data traffic by 2019, compared to 55% in 2014. When the broadcast operators deliver different compressed video programs to multiple users sharing a resource-limited wireless network, the design and optimization of the video communication system should consider two essential service objectives, namely, fairness and efficiency. To achieve fairness, the system should provide fair service, typically in terms of video quality, to all users subscribing to video services with the same quality level. The second objective, efficiency, is to attain the highest overall video quality with constraints on the system resources. To attain the highest efficiency while providing fairness whenever needed, cross-layer optimization is one of the approaches that can be exploited [2], [3]. In this paper, we will limit our discussion of cross-layer approaches to the medium access control (MAC) and the application (APP) layers. We address the above issue and present a MAC-centric cross-layer optimization framework. That is, according to quality fairness requirements which are specified by the utilities and constraints defined at the APP layer, an adaptive resource allocator (ARA) at the MAC layer optimally distributes the system resources among users so that the overall video quality is maximized.

Orthogonal frequency division multiple access (OFDMA) is one of the key physical layer techniques for the current wireless standards such as IEEE 802.16e [4] and 3GPP-Long Term Evolution (LTE) [5]. It has become the workhorse for wireless broadband applications due to its ability to provide high-rate wireless connectivity. In order to fully exploit the temporal, frequency and multiuser diversities of a OFDMA system, a highly adaptive resource allocation scheme should be adopted to jointly allocate the system resources, e.g., subcarriers and transmission power. By exploiting the channel statistics of different users, the opportunistic resource allocation scheme [6] assigns the system resources in favor of the users with better channel conditions, and thus maximizes the spectral efficiency. However, such an opportunistic allocation scheme often sacrifices the transmissions of the cell-edge users experiencing poor channel quality, thereby resulting in an unfair video quality among users. Moreover, in a system transmitting videos, the goal is to fully utilize the system resources to maximize the system efficiency in terms of the overall video quality rather than spectral efficiency. Therefore, a resource allocation scheme which maximizes the system efficiency and maintains quality fairness among users is desired.

To transmit video streams over wireless networks, it is required that the video source rates can be adapted to meet different user requirements under the time-varying channel conditions. This can be achieved by the use of an APP layer source rate adaptation entity which also improves transmission

Manuscript received January 5, 2016; revised July 26, 2016 and February 2, 2017; accepted February 19, 2017. Date of publication March 3, 2017; date of current version June 15, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Honggang Wang.

The authors are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: link24@mcmaster.ca; sorina@mail.ece.mcmaster.ca).

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

Digital Object Identifier 10.1109/TMM.2017.2678198

stability, avoids buffer overflow, etc. Scalable Video Coding (SVC) [7] is a highly attractive tool to achieve source rate adaptation. With SVC, a video sequence is encoded into a single multi-layer stream. The source rate adaptation is achieved by sequentially dropping layers until the target rate is achieved. Moreover, the rate-distortion (R-D) information can be predicted during the encoding procedure without re-encoding the video and provides a one-to-one mapping between rate and video quality.

The cross-layer optimization for multi-video delivery over wireless networks has been a very active research area. The authors of [8] presented a cross-layer multiuser resource allocation algorithm for video transmission in downlink OFDM networks where the objective is to minimize the overall video distortion. The algorithm consists of subcarrier assignment and power allocation relying on the R-D function. The R-D function considers a temporal error propagation effect and relative importance (RI) imposing constraints on individual user rate and quality of service. However, since the R-D relationships can be significantly different among different videos, while RI imposes constraints on the required rate rather than the required video quality, even a fair service, i.e., with the same RI value for all users, can lead to large quality variation among users. To ensure fairness in terms of video quality, the authors in [9] proposed a content-aware distortion-fair video delivery scheme for multihop video communications. Instead of providing bandwidth fairness, it assures max-min distortion-fair sharing among users. The cross-layer resource allocation is guided by exploiting the temporal prediction structure of the video sequences and a frame drop distortion metric based on the frame importance. The main drawback of the scheme is that the source rate adaptation is based on a coarse distinction of the data importance at frame level and could lead to a waste of bandwidth if the thresholds of dropping frames are not carefully selected. In [10] a scheduling strategy relying on the concept of Nash equilibrium, for scalable video transmission to multiple users over OFDMA systems, is devised. It is based on a metric named frame significance throughput (FST) considering the temporal dependencies among frames in a video sequence. The FST is incorporated into a payoff metric, which is exploited by the scheduler at the MAC layer to guide the resource allocation procedure in order to maximize the Nash product of the received video quality of each user and achieve quality fairness among users. More recent works have been proposed for cross-layer video transmission optimization with the goal of maximizing the minimum video quality across users and thus providing max-min quality fairness [11], [12], or of maximizing (respectively, minimizing) the overall received video quality (distortion) without addressing fairness [13], [14].

The authors of [3] proposed a distortion-fair cross-layer optimization framework for scalable video delivery to multiple users over OFDMA wireless networks. The optimization seeks to maximize the sum of the ergodic (average) rate assigned to users while minimizing the distortion difference among the received videos. The optimization problem is "vertically" decomposed into two sub-problems at the MAC and APP layers, respectively. An iterative local approximation (ILA) algorithm is proposed to obtain the global solution. The globally optimal solution under the distortion-fairness constraint aims to attain zero distortion difference between any two users' received videos. Thus, the framework proposed in [3] for achieving quality fairness is also based on the max-min criterion.

However, under such purely fair schemes ([3], [11], [12]), a majority of the available resources must be allocated to the users having poor channel conditions or requiring high-complexity videos so that they can achieve the same quality level as other users who are likely to achieve much higher quality improvement if assigned the same amount of resources. In other words, achieving pure fairness among users usually comes at the cost of sacrificing the video quality of a set of users and decreasing the system efficiency in terms of overall received video quality. Therefore, there is an inherent conflict between fairness and efficiency.

The authors of [15] proposed a cross-layer framework for sending multiple scalable videos over OFDM networks where trade-offs between quality fairness and system efficiency can be achieved. The video streams are transmitted across J transmission intervals. The optimization problem is broken down into J sequential problems, each of which is solved during a transmission interval to either ensure fairness or improve efficiency. To ensure fairness, the problem is formulated as minimizing the maximal end-to-end distortion received among all users. To improve efficiency, the problem is formulated as minimizing the overall end-to-end distortion among all users. Due to the NP-hard nature of the fairness and efficiency problems, two suboptimal algorithms were proposed to solve them. Further, the authors applied the fairness algorithm for the first x transmission intervals to ensure the baseline fairness, and the efficiency algorithm for the remaining $J - x$ intervals to improve the overall efficiency. In this way a desired trade-off between fairness and efficiency can be achieved by varying the value of x. However, such a transmission interval-based optimization, without considering the ergodic rate, does not allow to fully exploit the temporal diversity. In addition, the framework supports only $J + 1$ trade-off points, and thus it is inadequate when a set of denser trade-off points is required.

II. Contribution

As pointed out earlier, quality fairness and system efficiency are conflicting requirements. Focusing entirely only on one of them might not guarantee satisfaction of all users. For instance, when the system efficiency is maximized the set of users with good channel conditions and/or low demanding videos (referred to as set A of videos) will be assigned more resources at the expense of sacrificing, possibly drastically, the quality of the users with poor channel conditions and/or high demanding videos (set B). At the other extreme is the pure quality-fairness scenario, which enforces the same level of (objective) quality to all users. In this scenario, the quality of videos in set B is increased while the quality of videos in set A is decreased until they reach the same level. However, it could happen that, while the perceived visual quality of videos in set B increases, that of some videos in set A is degraded below a pleasing/acceptable level. Thus,

TABLE I
MOST USED NOTATIONS AND ACRONYMS

Notation	Definition	Notation	Definition
$K, \mathcal{K}$	Number, resp. set, of users	$\boldsymbol{\tau}, \boldsymbol{p}$	PHY layer time, resp. power, allocation policies
$M, \mathcal{M}$	Number, resp. set, of subcarriers	$\mathcal{S}$	Set of all possible PHY layer allocation policies
$\boldsymbol{F}$	Source rate vector	$\mathcal{A}$	Set of feasible PHY layer allocation policies
$\boldsymbol{R}$	PHY layer rate vector	$\mathcal{R}$	PHY layer rate region
Q	PSNR	$\mathcal{R}_f$	Set of rate vectors satisfying the PF constraints
q^*	Optimal PSNR value for pb. (5)	σ	Parameter controlling the constraints in problem (10)
$\mathcal{R}_\sigma$	Set of rate vectors satisfying (11b)	$\boldsymbol{f}_{\min}, \boldsymbol{f}_{\max}$	Lower and upper limits on source rates obtained
bd $\mathcal{R}$	Boundary of $\mathcal{R}$		from constraints (10b) - (10d)
Abbr.	Full name	Abbr.	Full name
APP	Application	ARA	Adaptive Resource Allocator
BS	Base Station	ERA	Equal-rate Adaptation
GOP	Group of Pictures	IDR	Instantaneous Decoding Refresh
ILA	Iterative Local Algorithm	MAC	Medium Access Control
ME	Maximum Efficiency	MANE	Media-aware Network Element
MS	Multimedia Server	MGS	Medium-grain Quality Scalable Coding
MSE	Mean Square Error	NALU	Network Abstraction Layer Unit
PHY	Physical	OFDMA	Orthogonal Frequency Division Multiple Access
PF	Pure Quality-fairness	PSNR	Peak Signal-to-noise Ratio
R-D	Rate-Distortion	RTP	Real-time Transport Protocol
RS	Reed-Solomon	SNR	Signal-to-noise Ratio
SVC	Scalable Video Coding	UXP	Unequal Erasure Protection
TB	Transmission Block		

both the maximum efficiency scenario on one side, and the pure fairness scenario on the other side, may lead to unsatisfied users. A natural solution to this problem is to trade off between the two extremes. Therefore, it is important to investigate such trade-off schemes.

In this paper, we propose a new cross-layer optimization framework for the transmission of scalable videos to multiple users in OFDMA wireless networks addressing the trade-off between quality fairness and system efficiency. The quality is measured using the PSNR corresponding to average user rates under average power constraints.

The main idea is to relax the fairness constraints so that to allow the PSNR of each video to be some distance away from the common PSNR level corresponding to the pure quality-fairness (PF) problem. This distance is controlled by using a parameter σ which bounds from above the relative difference between the PSNR of each video and the PF PSNR value. More specifically, the problem is formulated as maximizing the sum of PSNRs under the relaxed constraints mentioned above. We will show that, by gradually increasing σ (starting from 0), a wide-range and dense set of trade-off points can be achieved, thus overcoming the limitation of the trade-off framework proposed in [15]. The two extremes of this range of points correspond to the PF scenario, when $\sigma = 0$, respectively, the maximum efficiency scenario (ME), when $\sigma = \infty$.

In order to solve the optimization problem we proceed in two steps. First the PF PSNR value is determined by maximizing the common PSNR level enforced to all videos. Note that this optimization criterion is essentially equivalent to the max-min quality fairness criterion used in prior work [3], [11], [12], [15]. We solve the PF problem using the same vertical decomposition approach employed in [3], but with a faster algorithm for the source adaptation problem at the APP layer. Further, once the PF PSNR value is found, the problem with relaxed PSNR-fair constraints is converted to a convex optimization problem by transforming the constraints on PSNR into linear constraints in terms of rate. The latter problem is a general utility-based resource allocation problem for which a low-complexity algorithm has been proposed to obtain an almost surely optimal solution [16].

Finally, our simulation results validate the fact that, by gradually increasing σ, a densely populated set of points trading off between PF and ME, can be obtained. Additionally, with an appropriate choice of the parameter σ, the disadvantages of both extreme schemes - PF, respectively, ME - could be mitigated, leading to the satisfaction of all users.

The remainder of the paper is organized as follows. Section III describes the system architecture and the models for transmitting scalable video in OFDMA networks. In Section IV the PF problem and its solution are discussed. In the following section the problem with relaxed quality-fairness constraints is introduced. The problem is further converted to a convex optimization problem and its solution based on [16] is reviewed. The practical performance of the proposed optimization framework is evaluated in Section VI, followed by a discussion in Section VII. Finally, Section VIII concludes the paper.

Notation: Vectors and sets are denoted by bold and calligraphic letters, respectively. $\boldsymbol{x}^T$ denotes the transpose and $\|\boldsymbol{x}\|_p$ the p-norm of $\boldsymbol{x}$. $\mathbf{0}$ is the all-zero vector and $\mathbf{1}$ is the all-one vector. The symbol $\vee$ means "OR", $\wedge$ means "AND", $[x]^+ \triangleq \max(x, 0)$, $[x]^+_\epsilon \triangleq \max(x, \epsilon)$ and $\mathbb{E}_{\boldsymbol{\gamma}}[\cdot]$ denotes the expectation with respect to the random process $\boldsymbol{\gamma}$. Given two vectors $\boldsymbol{x} = [x_1, \cdots, x_K]^T$ and $\boldsymbol{x'} = [x'_1, \cdots, x'_K]^T$, $\boldsymbol{x} \preceq \boldsymbol{x'}$ means that $x_k \leq x'_k$, for all $1 \leq k \leq K$, while $\boldsymbol{x} \preccurlyeq \boldsymbol{x'}$ means that $\boldsymbol{x} \preceq \boldsymbol{x'}$ and $\boldsymbol{x} \neq \boldsymbol{x'}$. Table I contains the notations and abbreviations that are used most often in the paper.

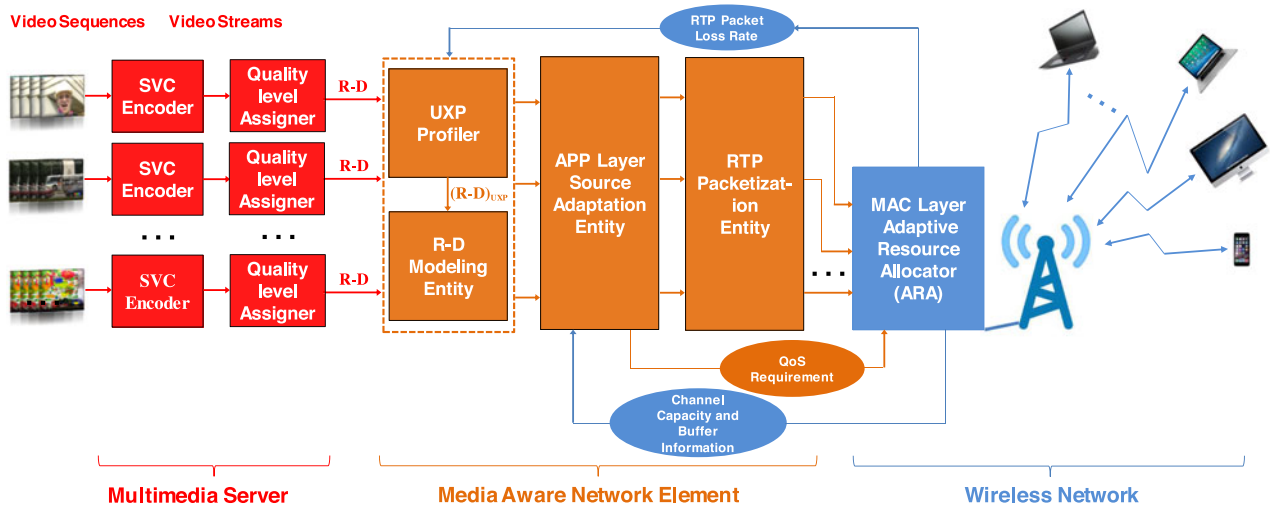

Fig. 1. Architecture and components of the multi-user video delivery system.

III. System Description

In this section we first discuss the architecture and functionality of a general multi-user video delivery system. Next we review a continuous semi-analytical R-D model for quality scalable video streams. Then we present the physical layer model for a OFDMA wireless network.

A. Multi-user Video Delivery System

We consider a general multi-user video delivery system shown in Fig. 1 where the system components are arranged in three groups, namely, the multimedia server (MS), the media-aware network element (MANE), and the wireless network. The MS encodes a set of video sequences, each of which is requested by a user, to fully support quality scalability. Each encoded video stream is then organized into network abstraction layer units (NALUs), each of which is a packet of an integer number of bytes. Then, the quality level assigner evaluates the priority level of each NALU according to its contribution to the quality of the reconstructed video. Such priority level information is embedded into the header of the NALU and will be exploited by the source adaptation entity. It should be pointed out that the encoding and priority level assessment are carried out off-line. The pre-encoded video streams are stored in databases at the MS, whereas the R-D information will be forwarded to the MANE.

The unequal erasure protection (UXP) profiler periodically collects the R-D information from the MS, and the estimated real-time transport protocol (RTP) packet loss rate from the base station (BS). Then it computes, according to a predefined protection policy and the available information, the rates and the expected reconstructed video qualities of the videos after erasure protection. This information is fed to the R-D modeling entity and is used for the R-D modeling of the UXP protected video streams. The source adaptation entity removes, according to the results of the source adaptation algorithm, the needless NALUs from each original video stream to form a valid substream intended for a user. It should be pointed out that the source adaptation algorithm requires as inputs the estimated channel capacity and buffer status information from the BS, and the information about the R-D models at the MANE. Each outcoming substream is forwarded to the real-time transport protocol (RTP) packetization entity where the substream will be protected by a UXP scheme based on Reed-Solomon (RS) codes. The resulting RS codewords, containing both data and parity symbols, are arranged into a transmission block (TB) and interleaved over a number of RTP packets. Finally, the RTP packets will be sent to the MAC/PHY layers through the UDP/IP protocol stack. We point out that in this work we use an adaptation of the algorithm of [17] to compute the optimal UXP redundancy allocation in R-D sense. For details we refer the reader to [18].

The adaptive resource allocator (ARA) at the BS adaptively allocates the system resources among users with the aim of maximizing the overall average rates while satisfying the quality requirement provided by the APP layer. The MANE and the BS exchange information about the channel capacity, buffer status, RTP packet-loss rate and quality requirement in a cross-layer style at regular intervals called *application periods* (in the order of seconds). It is worth noting that whereas the processes of R-D modeling, UXP, source adaptation and RTP packetization at the MANE are executed per *application period*, the resource allocation process at the BS is carried out every time slot (in the order of milliseconds).

B. Rate Distortion Models for the Error-Protected Quality Scalable Video Streams

The SVC standard provides three common scalable modes, namely, temporal, spatial and quality scalability, which enable to adapt a video stream in terms of frame rate, frame size and frame fidelity, respectively. In this paper, we focus on quality scalable videos with fixed frame rate and size. The SVC standard provides two ways to achieve quality scalability, namely coarse-grain quality scalable coding (CGS) and medium-grain quality scalable coding (MGS). With CGS, the provision of a video with different qualities is enabled by dropping quality layers one by one until the target bit rate is achieved. However, the number of

supported bit rates is limited to the number (up to eight) of CGS quality layers [19]. In comparison to CGS, MGS allows more extractable rate points (up to 128) by dividing each quality layer into up to 16 MGS layers, each of which can be dropped for the purpose of rate adaptation. In this paper, we will focus on MGS scalability.

Scalable video sequences are commonly arranged into sets of frames named groups of pictures (GOPs). Each GOP begins with an intra-coded (I-frame) or inter-coded (P-frame), which is followed by a fixed number of B-frames. The frame interval between any two consecutive I-frames is called an instantaneous decoding refresh (IDR) period. In this paper, we will focus on an IDR-based video transmission. We assume that at the beginning of each *application period* I_k successive frames of the video sequence intended for user k, are encoded to generate an MGS video stream.

The operational R-D point for the corresponding video stream is obtained by averaging the rate and distortion (i.e., mean squared error - MSE), respectively, over all I_k frames. In this work we convert the distortion to PSNR in order to measure the video quality, using the relation $PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right)$. Thus, from each operational R-D point an operational rate-PSNR point can be computed. The aforementioned R-D information of the compressed video stream is further used by the UXP profile to obtain a rate-PSNR point corresponding to the error protected video stream (i.e., the TB). We point out that the PSNR is actually the expected PSNR over all possible packet loss scenarios for the RTP packets. The set of operational points is discrete since different rates correspond to different numbers of transmitted packets, while the size of each RTP packet remains fixed.

In [20] the authors proposed a general continuous semi-analytical R-D model, which has been verified for SVC quality scalable videos in [21] and [22], to estimate the relationship between the rate and distortion at the encoder side. According to this model the rate of the video for user k is a parametric function $F_k(Q)$ of the PSNR Q, as follows:

$$F_k(Q) = \frac{\theta_k}{255^2 10^{-Q/10} + \alpha_k} + \beta_k,\ Q \in [Q_{k,\min}, Q_{K,\max}] \tag{1}$$

where $Q_{k,\min}$ and $Q_{K,\max}$ are the minimum and maximum PSNR values corresponding to the minimum rate $F_{k,\min}$ and maximum rate $F_{K,\max}$, respectively. We emphasize that the three parameters θ_k, α_k and β_k are dependent on the video content, encoder and RTP packet loss rate. They can be estimated using curve-fitting methods over a number of empirical rate-PSNR points. According to extensive simulations, a general curve-fitting algorithm needs at least six empirical R-D points and a certain number of iterations and function evaluations to guarantee high accuracy for most of the video sequences [22].

C. Physical Layer Model for OFDMA Wireless Networks

We consider the downlink of a single-cell OFDMA wireless network with K users and M orthogonal subcarriers indexed by the sets $\mathcal{K} = \{1, 2, \cdots, K\}$ and $\mathcal{M} = \{1, 2, \cdots, M\}$, respectively. We assume a subcarrier bandwidth B and total average power $\bar{P}$. The channel gain between BS and user k, on subcarrier m at the nth time slot, is denoted by $h_{k,m}[n]$, and modeled as a stationary and ergodic complex Gaussian random process (Rayleigh fading). Therefore, the distribution of $h_{k,m}[n]$ is independent of the time slot index n. In the subsequent discussion, we drop the time slot index n when the context is clear, for notational brevity.

The normalized signal-to-noise ratio (SNR), i.e., the SNR corresponding to unit transmission power, of user k on subcarrier m, is given as $\gamma_{k,m} = |h_{k,m}|^2/\sigma^2$, where σ^2 is the variance of the zero-mean additive white Gaussian noise (AWGN) at the receiver. We let $\boldsymbol{\gamma} = \{\gamma_{k,m}, \forall\, k, m\}$ denote the set of the SK realizations of the normalized SNR random processes. Throughout the paper, we assume that the BS has perfect knowledge of $\boldsymbol{\gamma}$, and that $\boldsymbol{\gamma}$ is fixed per time slot, but varies across time slots. Based on $\boldsymbol{\gamma}$, the ARA at the BS optimally allocates the available power and subcarriers to all users per time slot.

Consider for now that per time slot a subcarrier can be shared by multiple users over nonoverlapping time fractions of a time slot duration t_{slot}. Let $\tau_{k,m} \geq 0$ and $p_{k,m} \geq 0$ denote the non-negative time fraction and the average power, respectively, allocated for transmission to user k on subcarrier m. Since the transmission to user k is only activated for a fraction of the time slot, the transmission power allocated to user k, during the active time fraction, is $p_{k,m}/\tau_{k,m}$. Taking into account the adaptive modulation and coding (AMC) scheme adopted by the PHY layer, the maximum achievable rate of user k on subcarrier m is given by

$$r_{k,m}(\tau_{k,m}, p_{k,m}) = \begin{cases} B\tau_{k,m} R\left(\frac{\gamma_{k,m} p_{k,m}}{\tau_{k,m}}\right) & \tau_{k,m} > 0 \\ 0 & \tau_{k,m} = 0 \end{cases} \tag{2}$$

where $R(x) = a_1 \log_2(1 + x/a_2)$, and a_1 and a_2 are two parameters named *rate adjustment* and *SNR gap* that are introduced to account for the particular AMC scheme in use [23]. Since the rate in (2) is a function of $\tau_{k,m}$ and $p_{k,m}$, the ARA seeks to specify the set of allocation policies $\boldsymbol{\tau}(\boldsymbol{\gamma}) = \{\tau_{k,m}(\boldsymbol{\gamma}), \forall\, k, m\}$ and $\boldsymbol{p}(\boldsymbol{\gamma}) = \{p_{k,m}(\boldsymbol{\gamma}), \forall\, k, m\}$ per channel realization $\boldsymbol{\gamma}$. If the optimal $\boldsymbol{\tau}^*$ and $\boldsymbol{p}^*$ are found, the corresponding optimal rates, following from (2), will be $\boldsymbol{r}^*(\boldsymbol{\gamma}) = \{r_{k,m}(\tau^*_{k,m}(\boldsymbol{\gamma}), p^*_{k,m}(\boldsymbol{\gamma})), \forall\, k, m\}$.

Consider a sufficiently long *application period* t_{ap} over which it is reasonable to approximate the time-averaged rate, through ergodicity, by its ensemble average with respect to the random process $\boldsymbol{\gamma}$. Then, the maximum achievable rate for user k, averaged over an *application period*, is given by

$$\begin{aligned} R_k(\boldsymbol{\tau}, \boldsymbol{p}) &= \frac{1}{N_{\text{slot}}} \sum_{n=1}^{N_{\text{slot}}} \left[\sum_{m \in \mathcal{M}} r_{k,m}(\tau_{k,m}[n], p_{k,m}[n]) \right] \\ &\simeq \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{m \in \mathcal{M}} r_{k,m}\left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right) \right] \end{aligned}$$

where $N_{\text{slot}} = \left\lfloor \frac{t_{ap}}{t_{\text{slot}}} \right\rfloor \gg 1$, is the number of time slots within an *application period*. Supposing without loss of generality that the overhead introduced by different network layers is unity, the average PHY rate is equal to the average source rate, i.e.,

$R_k(\boldsymbol{\tau},\boldsymbol{p}) = F_k(Q)$. According to (1), the relationship between the PSNR of the video transmitted to user k and the average PHY rate can be described by $Q_k = F_k^{-1}(R_k(\boldsymbol{\tau},\boldsymbol{p}))$.

Let us denote by $\mathcal{S}$ the set of all possible allocation polices $\boldsymbol{\tau}(\boldsymbol{\gamma})$ and $\boldsymbol{p}(\boldsymbol{\gamma})$, i.e.,

$$\mathcal{S} \triangleq \{(\boldsymbol{\tau},\boldsymbol{p}) \mid \tau_{k,m}(\boldsymbol{\gamma}) \geq 0, p_{k,m}(\boldsymbol{\gamma}) \geq 0, \forall\, k,m, \sum_{k=1}^{K} \tau_{k,m}(\boldsymbol{\gamma}) \leq 1, \forall\, m\}$$

and $\mathcal{A} \triangleq \{(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{S} \mid \mathbb{E}_{\boldsymbol{\gamma}}[\sum_{k\in\mathcal{K}} \sum_{m\in\mathcal{M}} p_{k,m}(\boldsymbol{\gamma})] \leq \bar{P}\}$. Let us, further, denote $\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \triangleq [R_1(\boldsymbol{\tau},\boldsymbol{p}), \cdots, R_K(\boldsymbol{\tau},\boldsymbol{p})]^T$ the maximum achievable ergodic rate vector and $\boldsymbol{R} = [R_1, \cdots, R_K]^T$ an ergodic rate vector. The ergodic rate region of the OFDMA downlink channel can be defined as

$$\mathcal{R} \triangleq \bigcup_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} \{\boldsymbol{R} \mid \boldsymbol{0} \preceq \boldsymbol{R} \preceq \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})\}. \tag{3}$$

Since the rate $r_{k,m}(\tau_{k,m}, p_{k,m})$ in (2) is a jointly concave function of $\tau_{k,m}$ and $p_{k,m}$, the ergodic rate region $\mathcal{R}$ in (3) is a convex set of the rate vectors [16].

IV. Pure Quality-Fairness Problem

In order to present our proposed framework for the trade-off between quality fairness and system efficiency, we need first to discuss the pure quality-fairness problem. From a pure quality-fairness perspective, we are interested in all users obtaining the same quality level represented by some target PSNR value q. However, since the attainable PSNR range for different users may be different, the target value q may not be included in this range for some users. Therefore, the value assigned to users in the latter category will be the attainable PSNR which is closest to q. In order to model this requirement we define the function $\hat{Q}_k(q)$ which maps every $q \in \mathbb{R}^+$ to the closest achievable PSNR of the kth video, namely

$$\hat{Q}_k(q) \triangleq \begin{cases} Q_{k,\min} & q \leq Q_{k,\min} \\ q & Q_{k,\min} < q < Q_{K,\max} \\ Q_{K,\max} & q \geq Q_{K,\max} \end{cases} \tag{4}$$

for $1 \leq k \leq K$ and $q \in \mathbb{R}^+$.

Further, define the set $\mathcal{R}_f$ of rate vectors that satisfy the pure quality-fairness constraints as

$$\mathcal{R}_f \triangleq \{\boldsymbol{R} \mid \exists q \geq 0 \text{ such that } R_k = F_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}\}.$$

Then we formulate the pure quality-fairness problem as the following constrained PSNR maximization:

$$\max_{q\in[Q^{\text{all}}_{\min}, Q^{\text{all}}_{\max}]} \quad q \tag{5a}$$

$$\text{s.t.} \quad (\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{A}, \tag{5b}$$

$$\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{R}_f \tag{5c}$$

where $Q^{\text{all}}_{\min}$ denotes the minimum of $Q_{k,\min}$ over all k, and $Q^{\text{all}}_{\max}$denotes the maximum of $Q_{k,\min}$ over all k. We will show that problem (5) is equivalent to the distortion-fair sum-rate maximization problem in [3]. The authors of [3] define a distortion difference $\Delta(D_i, D_j)$ and formulate the problem as

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \quad \|\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})\|_1 \tag{6a}$$

$$\text{s.t.} \quad \triangle(D_i, D_j) = 0, \forall i,j \in \mathcal{K} \tag{6b}$$

$$\boldsymbol{F}_{\min} \preceq \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \preceq \boldsymbol{F}_{\max} \tag{6c}$$

$$\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{R} \tag{6d}$$

where $\boldsymbol{F}_{\max} = [F_{1,\max}, \cdots, F_{K,\max}]^T$ and $\boldsymbol{F}_{\min} = [F_{1,\,\min}, \cdots, F_{K,\,\min}]^T$, for $1 \leq k \leq K$. Additionally, the distortion difference $\triangle(D_i, D_j)$ is defined as

$$\triangle(D_i, D_j) \triangleq \begin{cases} 0 & (D_i, D_j) \in \mathbb{D} \vee (D_j, D_i) \in \mathbb{D} \\ |D_i - D_j| & \text{otherwise} \end{cases}$$

where $\mathbb{D} \triangleq \{(D_i, D_j) \mid (D_i = D_{i,\,\max} < D_j) \vee (D_i = D_{i,\,\min} > D_j)\}$. The PSNR difference $\triangle(Q_i, Q_j)$ can be defined in a similar manner. Notice that the fairness constraints in (6b) restrict the feasible solutions to the set of rate vectors

$$\mathcal{R}_f^c \triangleq \{\boldsymbol{R} \mid \triangle\left(F_i^{-1}(R_i), F_j^{-1}(R_j)\right) = 0, \forall i,j \in \mathcal{K}\}$$

where $F_k^{-1} : [F_{k,\min}, F_{k,\max}] \to [Q_{k,\min}, Q_{k,\max}]$ is the inverse mapping of $F_k(\cdot)$ defined in (1), for $k \in \mathcal{K}$. Then the following equivalence result holds, whose proof is deferred to the appendix.

Lemma 1: One has $\mathcal{R}_f = \mathcal{R}_f^c$. Then it follows that problem (5) is equivalent to problem (6).

It was proved in [3] that the optimal solution $\boldsymbol{R}^*$ to problem (6) is the unique point in $\mathcal{R}_f^c \cap \mathbf{bd}\,\mathcal{R}$, where $\mathbf{bd}\,\mathcal{R} \triangleq \{\boldsymbol{R} \in \mathcal{R} \mid \nexists \boldsymbol{r} \in \mathcal{R} \text{ with } \boldsymbol{R} \preccurlyeq \boldsymbol{r}\}$. The authors of [3] "vertically" decomposed the optimization problem into a resource allocation problem at the MAC layer and a source adaptation problem at the APP layer. The optimal solution was further obtained through the ILA algorithm, which is an iterative procedure built between the MAC and APP layers. In virtue of Lemma 1 we can use the same decomposition approach to solve problem (5). Next we briefly review the MAC layer and APP layer problems as formulated in [3] and propose a faster solution for the APP layer problem.

Algorithm 1: ILA algorithm

1: **Initialize:** $i = 0$; give a directional vector $\tilde{\boldsymbol{\phi}}^{(0)} \succcurlyeq \boldsymbol{0}$ and tolerance $\epsilon > 0$; set $e^{(0)} = 10\epsilon$;
2: Solve problem (8) to obtain $\tilde{\boldsymbol{R}}^{(0)}$ and $\tilde{\boldsymbol{w}}^{(0)}$
3: **while** $e^{(i)} > \epsilon$ **do**
4: $\quad i = i + 1$
5: $\quad$ Find $\tilde{\boldsymbol{F}}^{(i)}$ such that:
6: $\quad \tilde{\boldsymbol{F}}^{(i)} \in \boldsymbol{R}_f \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}^{(i-1)}, \tilde{\boldsymbol{w}}^{(i-1)})$
7: $\quad \tilde{\boldsymbol{\phi}}^{(i)} = \tilde{\boldsymbol{F}}^{(i)} / \|\tilde{\boldsymbol{F}}^{(i)}\|_1$
8: $\quad$ Solve problem (8) to obtain $\tilde{\boldsymbol{R}}^{(i)}$ and $\tilde{\boldsymbol{w}}^{(i)}$
9: $\quad e^{(i)} = \|\tilde{\boldsymbol{R}}^{(i)} - \tilde{\boldsymbol{F}}^{(i)}\|_1$
10: **end while**

The problem at the MAC layer is

$$\max_{r\geq 0,(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} r \tag{8a}$$

$$\text{s.t.} \qquad \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \succeq \boldsymbol{\phi} r \tag{8b}$$

where $\boldsymbol{\phi}=[\phi_1,\cdots,\phi_K]^T \succeq \mathbf{0}$ defines the direction of the line connecting the origin and the point obtained at the application layer. Problem (8) is a well-investigated resource allocation problem [24] that can be solved efficiently given the information of the directional vector $\boldsymbol{\phi}$. It is shown in [24] that even the solution of (8) can be obtained through solving a weighted sum of average rate (WSAR) problem

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} \boldsymbol{w}^T\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})$$

and the solution resides on the boundary $\mathbf{bd}\,\mathcal{R}$. However, differently from the WSAR problem, the weight vector $\boldsymbol{w}$ is not predefined, but rather, it is evaluated in the dual domain and constrained by $\boldsymbol{\phi}$. Let us denote by $\tilde{\boldsymbol{R}}$ the optimal rate and $\tilde{\boldsymbol{w}}=[w_1,\cdots,w_K]^T \succcurlyeq \mathbf{0}$ that are obtained after solving the problem at the MAC layer. The tangent space to the rate region $\mathcal{R}$ at the point $\tilde{\boldsymbol{R}}$ can then be identified with the null space of $\tilde{\boldsymbol{w}}$ defined as follows:

$$\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}},\tilde{\boldsymbol{w}}) \triangleq \{\boldsymbol{R} \mid \tilde{\boldsymbol{w}}^T(\boldsymbol{R}-\tilde{\boldsymbol{R}})=0\}.$$

Then the problem solved at the APP layer is the problem of determining the intersection

$$\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}},\tilde{\boldsymbol{w}}) \cap \mathcal{R}_f. \tag{9}$$

Finally, the ILA algorithm solves iteratively problems (8) and (9). Its pseudocode is presented in Algorithm 1.

In virtue of Lemma 1 and of the optimality result of [3, Lemma 1] we conclude that, assuming that $\boldsymbol{F}_{\min}\in\mathcal{R}$ and $\boldsymbol{F}_{\max}\notin\mathcal{R}$, and starting from an initial $\boldsymbol{R}\succeq\mathbf{0}$, the ILA algorithm converges to the unique solution $\boldsymbol{R}^*\in\boldsymbol{R}_f\cap\mathbf{bd}\,\mathcal{R}$, of problem (5), i.e., $\lim_{i\to\infty}\tilde{\boldsymbol{F}}^{(i)}=\boldsymbol{R}^*$.

Next we discuss the solution to problem (9). We point out that the algorithm to solve the APP layer problem proposed in [3] requires $K(K-1)/2$ iterations in the worst case. Each iteration consists of numerically solving an equation to obtain a candidate distortion value $\tilde{D}$ to be assigned to a set of videos, followed by at most K rate evaluations. Since the number of terms needed to evaluate the equation is $O(K)$, it follows that the time complexity of each iteration is $O(KI)$, where I denotes the number of inner iterations needed to determine numerically the value of $\tilde{D}$. This leads to a worst-case time complexity of $O(K^3I)$ for the whole algorithm. However, the authors of [3] point out that in their extensive simulations with practical data, K outer iterations were sufficient, which translates to $O(K^2I)$ operations to solve the APP layer problem.

On the other hand, our formulation of the pure quality-fairness problem directly shows that $\mathcal{R}_f$ is a curve parameterized by q, therefore the intersection in (9) can be found by means of a bisection search over q. For this define the function

$$\Gamma(\boldsymbol{F},\tilde{\boldsymbol{R}},\tilde{\boldsymbol{w}}) \triangleq \sum_{k\in\mathcal{K}} \tilde{w}_k(F_k-\tilde{R}_k).$$

Algorithm 2: Fast algorithm to solve problem (9)

```
if Γ(F_min, R̃, w̃) > 0 then
    report infeasibility
else if Γ(F_max, R̃, w̃) ≤ 0 then
    report infeasibility and set F* = F_max
else
    Initialize: low = Q^all_min; high = Q^all_max; set tolerance e_bs;
    while (high − low)/2 > e_bs do
        q* = (high + low)/2;
        for all k ∈ K do
            if q* ≤ Q_{k,min} then
                Q*_k = Q_{k,min}; F*_k = F_{k,min};
            else if q* ≥ Q_{K,max} then
                Q*_k = Q_{K,max}; F*_k = F_{K,max};
            else
                Q*_k = q*; F*_k = F_k(Q*_k), based on model (1);
            end if
        end for
        if Γ(F*(q*), R̃, w̃) < 0 then
            low = q*;
        else if Γ(F*(q*), R̃, w̃) > 0 then
            high = q*;
        else
            break
        end if
    end while
end if
```

According to (1) and (4), the rate $F_k(\hat{Q}_k(q)), \forall k\in\mathcal{K}$, is a nondecreasing function of q. Using further the fact that $\tilde{w}_k\geq 0$ for all k, it follows that the function $\Gamma(\boldsymbol{F}(q),\tilde{\boldsymbol{R}},\tilde{\boldsymbol{w}})=\sum_{k\in\mathcal{K}}\tilde{w}_k(F_k(\hat{Q}_k(q))-\tilde{R}_k)$ is also a nondecreasing function of q. Therefore, we can apply the bisection search method to find q^* such that $\Gamma(\boldsymbol{F}(q^*),\tilde{\boldsymbol{R}},\tilde{\boldsymbol{w}})=0$ and obtain the solution $\boldsymbol{F}^*(q^*)$ to (9). We summarize the pseudocode of the bisection search-based source adaptation algorithm in Algorithm 2.

It is easy to see that each iteration takes $O(K)$ time, while the total number of iterations is $O(\log\frac{Q^{\text{all}}_{\max}-Q^{\text{all}}_{\min}}{e_{bs}})$, where e_{bs} is the tolerance value for the optimal q^*. We conclude that the time complexity of the proposed solution to problem (9) is $O(K\log\frac{Q^{\text{all}}_{\max}-Q^{\text{all}}_{\min}}{e_{bs}})$. Further, assuming that the values I and $\log\frac{Q^{\text{all}}_{\max}-Q^{\text{all}}_{\min}}{e_{bs}}$ are comparable, it follows that the proposed algorithm is faster by a factor of $O(K^2)$ in the worst case, and by a factor of $O(K)$ in the average case, based on the experimental average running time reported in [3].

V. Problem With Relaxed Quality-Fairness Constraints

This section presents the proposed framework for the trade-off between system efficiency and quality fairness, which is the main contribution of this work. Our strategy to achieve this trade-off is to relax the fairness constraints in the PF problem

by allowing the PSNRs of different users to be at some distance away from the common PSNR value q^* corresponding to the optimal PF point [i.e., the solution of problem (5)]. The larger this distance is, the looser the fairness constraints are and the higher the potential efficiency is. Therefore, we use a parameter σ to denote the upper bound imposed on the ratio between this distance and q^*. Consequently, this parameter will measure the trade-off between fairness and efficiency.

Thus, we formulate the problem with relaxed quality-fairness constraints as follows:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} \quad \sum_{k\in\mathcal{K}} Q_k(R_k) \tag{10a}$$

$$\text{s.t.} \quad |Q_k(R_k) - q^*| \le q^*\sigma, \forall k \in \mathcal{K}_0 \tag{10b}$$

$$Q_k(R_k) = Q_{k,\min}, \forall k \in \mathcal{K}_1 \tag{10c}$$

$$Q_k(R_k) = Q_{K,\max}, \forall k \in \mathcal{K}_2 \tag{10d}$$

$$R_k = R_k(\boldsymbol{\tau}, \boldsymbol{p}) \tag{10e}$$

where $\mathcal{K}_1 = \{k \in \mathcal{K} \mid q^*(1+\sigma) < Q_{k,\min}\}$, $\mathcal{K}_2 = \{k \in \mathcal{K} \mid q^*(1-\sigma) > Q_{K,\max}\}$ and $\mathcal{K}_0 = \mathcal{K} \setminus (\mathcal{K}_1 \cup \mathcal{K}_2)$. The equality constraints in (10c) and (10d) are motivated by the following considerations. If the target interval $[q^*(1-\sigma), q^*(1+\sigma)]$ and the attainable PSNR range $[Q_{k,\min}, Q_{K,\max}]$ for user k are disjoint, then the achievable value closest to q^* is assigned as the PSNR of user k.

Notice that the optimal solution to problem (10) guarantees that the PSNR difference between any two users in the set $\mathcal{K}_0$ is within $2\sigma q^*$, while all remaining users either have smaller PSNRs than users in $\mathcal{K}_0$, but achieve their individual maximum quality, or they have higher PSNRs than all users in $\mathcal{K}_0$, but achieve their individual minimum quality. Additionally, when $\sigma = 0$ problem (10) is equivalent to problem (5).

Clearly, problem (10) is not convex, however it can be converted to a convex one by writing the constraints (10b)- (10d) in terms of rate. This task is simplified by the fact that the PSNR functions are strictly monotone in the rate. For this denote for all $k \in \mathcal{K}_0$

$$\begin{cases} f_{k,\min} \triangleq \max\{F_{k,\min}, F_k(q^*(1-\sigma))\} \\ f_{K,\max} \triangleq \min\{F_{K,\max}, F_k(q^*(1+\sigma))\} \end{cases}.$$

Further, for all $k \in \mathcal{K}_1$ let $f_{k,\min} = f_{K,\max} \triangleq F_{k,\min}$, and for all $k \in \mathcal{K}_2$ let $f_{k,\min} = f_{K,\max} \triangleq F_{K,\max}$. Finally, denote $\boldsymbol{f}_{\min} \triangleq [f_{1,\min}, \cdots, f_{k,\min}]^T$ and $\boldsymbol{f}_{\max} \triangleq [f_{1,\max}, \cdots, f_{K,\max}]^T$. Thus, problem (10) can be cast as

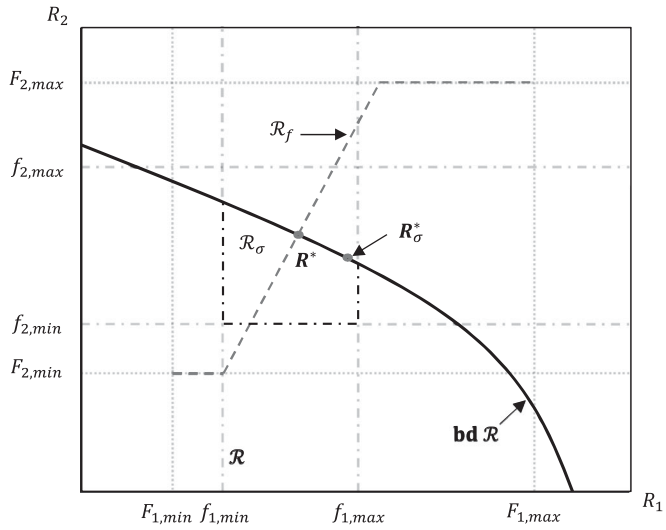

Fig. 2. Illustration of the relevant rate regions for problems (5) and (11) for an example with two users. $\boldsymbol{R}^*$ is the optimal solution to problem (5), whereas $\boldsymbol{R}^*_\sigma$ is the optimal solution to problem (11).

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} \quad \sum_{k\in\mathcal{K}} Q_k(R_k) \tag{11a}$$

$$\text{s.t.} \quad \boldsymbol{f}_{\min} \preceq \boldsymbol{R} \preceq \boldsymbol{f}_{\max} \tag{11b}$$

$$\boldsymbol{R} \in \mathcal{R} \tag{11c}$$

According to constraints (11b) and (11c), any feasible solution to (11) belongs to the set $\mathcal{R}_\sigma = \{\boldsymbol{R} \in \mathcal{R} \mid \boldsymbol{f}_{\min} \preceq \boldsymbol{R} \preceq \boldsymbol{f}_{\max}\}$. Note that $\mathcal{R}_\sigma$ is not empty if and only if $\boldsymbol{f}_{\min} \in \mathcal{R}$. Moreover, the problem has a trivial solution if $\boldsymbol{f}_{\max} \in \mathcal{R}$. Therefore, we will assume that $\boldsymbol{f}_{\min} \in \mathcal{R}$ and $\boldsymbol{f}_{\max} \notin \mathcal{R}$. Since the objective (11) is concave [25] and increasing, the optimal solution $\boldsymbol{R}^*_\sigma$ must be on the boundary $\mathbf{bd}\,\mathcal{R}$. Fig. 2 illustrates the relevant regions and the solution for problems (5) and (11) for an example with two users.

Problem (11) can be reformulated as

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}, \boldsymbol{f}_{\min}\preceq\boldsymbol{R}\preceq\boldsymbol{f}_{\max}} \quad \sum_{k\in\mathcal{K}} Q_k(R_k)$$

$$\text{s.t.} \quad R_k \le \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{m\in\mathcal{M}} r_{k,m}\left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right)\right], \forall k \in \mathcal{K}$$

$$\mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} p_{k,m}(\boldsymbol{\gamma})\right] \le \bar{P}. \tag{12}$$

$$L(\boldsymbol{\tau},\boldsymbol{p},\boldsymbol{R},\lambda,\boldsymbol{\mu}) = \sum_{k\in\mathcal{K}} Q_k(R_k) + \lambda\left\{\bar{P} - \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} p_{k,m}(\boldsymbol{\gamma})\right]\right\} + \sum_{k\in\mathcal{K}}\mu_k\left\{\mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{m\in\mathcal{M}} r_{k,m}\left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right)\right] - R_k\right\}$$

$$= \sum_{k\in\mathcal{K}} Q_k(R_k) + \boldsymbol{\mu}^T\boldsymbol{R} + \lambda\bar{P} + \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} c_{k,m}\left(\lambda, \boldsymbol{\mu}, r_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right)\right]. \tag{13}$$

Note that problem (12) is a strictly feasible convex optimization problem because of the concavity of the objective function and of the function $r_{k,m}(\cdot)$. The solution to (12) can be found using the Lagrangian dual method, as in [16]. We briefly review here the relevant results from [16].

Let $\boldsymbol{\mu}$ be the Lagrangian multiplier vector related to the constraints on the rates and let λ be the Lagrangian multiplier related to the average power constraint. Then, the Lagrangian associated with (12) is given in (13), shown at the bottom of the previous page, where $c_{k,m}(\lambda, \boldsymbol{\mu}, \tau_{k,m}, p_{k,m}) \triangleq \mu_k r_{k,m}(\tau_{k,m}, p_{k,m}) - \lambda p_{k,m}$. Further, the related Lagrangian dual function is

$$\Theta(\lambda, \boldsymbol{\mu}) = \max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}, \boldsymbol{f}_{\min}\preceq\boldsymbol{R}\preceq\boldsymbol{f}_{\max}} L(\boldsymbol{\tau}, \boldsymbol{p}, \boldsymbol{R}, \lambda, \boldsymbol{\mu})$$

and the dual problem is $\min_{\lambda>0,\boldsymbol{\mu}\succeq\mathbf{0}} \Theta(\lambda, \boldsymbol{\mu})$. For given λ and $\boldsymbol{\mu}$, $\Theta(\lambda, \boldsymbol{\mu})$ can be derived by solving two decoupled subproblems across $\boldsymbol{R}$ and $(\lambda, \boldsymbol{\mu})$, respectively. The first subproblem is associated with $\boldsymbol{R}$, i.e.,

$$\max_{\boldsymbol{f}_{\min}\preceq\boldsymbol{R}\preceq\boldsymbol{f}_{\max}} Q_k(R_k) + \boldsymbol{\mu}^T\boldsymbol{R} \tag{14}$$

which is a convex optimization problem, for which efficient algorithms to find the solution $\boldsymbol{R}^*(\boldsymbol{\mu})$ are available.

The second subproblem is related to $(\boldsymbol{\tau}, \boldsymbol{p})$ and is given as

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \lambda\bar{P} + \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} c_{k,m}(\lambda, \boldsymbol{\mu}, r_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma}))\right].$$

Given λ and $\boldsymbol{\mu}$, the unique solution to (14) is attained when each subcarrier is exclusively assigned to a single user per time slot and the power is allocated per user across subcarriers following a water-filling principle, i.e.,

$$\tau^*_{k,m}(\boldsymbol{\gamma}) = \begin{cases} 1 & k = k^*_m \\ 0 & \forall k \neq k^*_m \end{cases} \qquad \forall\, m$$

where $k^*_m = \arg\max_{k\in K} [\mu_k \log_2(1 + \lambda_{k,m}\tilde{p}^*_{k,m}) - \lambda\tilde{p}^*_{k,m}]$ and $\tilde{p}^*_{k,m}(\boldsymbol{\gamma}) = [\frac{a_1 B\mu_k}{\lambda \ln 2} - \frac{a_2}{\gamma_{k,m}}]^+$. The corresponding optimal power allocation is

$$p^*_{k,m}(\boldsymbol{\gamma}) = \begin{cases} \tilde{p}^*_{k,m}(\boldsymbol{\gamma}) & k = k^*_m \\ 0 & \forall k \neq k^*_m \end{cases} \qquad \forall\, m.$$

Since (12) is convex and Slater's condition holds, the duality gap between the primal and dual problems is zero. Therefore, replacing λ and $\boldsymbol{\mu}$ with the optimal dual variables λ^* and $\boldsymbol{\mu}^*$ provides the almost surely optimal resource allocation policy $\boldsymbol{\tau}^*(\lambda^*, \boldsymbol{\mu}^*, \boldsymbol{\gamma})$ and $\boldsymbol{p}^*(\lambda^*, \boldsymbol{\mu}^*, \boldsymbol{\gamma})$ and the corresponding optimal rate vector $\boldsymbol{R}^*(\boldsymbol{\mu}^*)$, which is on the boundary of the rate region $\mathcal{R}$. The optimal λ^* and $\boldsymbol{\mu}^*$ can be obtained through the method of stochastic subgradient iterations, i.e.,

$$\begin{cases} \lambda[n+1] = \lambda[n] + \delta(\sum_{m\in\mathcal{M}} p^*_{k,m}(\boldsymbol{\gamma}[n]) - \bar{P}) \\ \mu_k[n+1] = \mu_k[n] + \delta(R^*_k(\boldsymbol{\mu}[n]) - \sum_{m\in\mathcal{M}} r^*_{k,m}(\boldsymbol{\gamma}[n])). \end{cases} \tag{15}$$

Starting from any initial $\lambda > 0$ and $\boldsymbol{\mu} \succ \mathbf{0}$, the iterations in (15) converge to the optimal λ^* and $\boldsymbol{\mu}^*$.

Finally, it should be pointed out that the optimal solution $\boldsymbol{R}^*_\sigma = \boldsymbol{R}^*(\boldsymbol{\mu}^*)$ may not be achievable since the available SVC encoding schemes support only a discrete set of rate values. Following the common practice, the optimal discrete solution is obtained by extracting the largest achievable rate which is smaller than $\boldsymbol{R}^*_\sigma$.

VI. Numerical Results

In this section we assess the practical performance of the proposed optimization framework. We consider a OFDMA wireless network with $K = 6$ users and $M = 144$ subcarriers, unless otherwise stated, a time slot duration $t_{\text{slot}} = 0.5$ ms and a total average power $\bar{P} = 1$ W. The bandwidth of each subcarrier is 15 kHz. The Rayleigh fading channels between the BS and each user are simulated using the ITU Vehicular Channel A model [26] which has a root mean square delay spread $\tau_{rms} = 0.37$ μs and 50% coherence bandwidth of $B_c = 1/(5\tau_{rms}) \approx 540$ kHz. The average normalized SNRs for all users are assumed to be 25 dBW. The modulation and coding scheme adopted at the PHY layer are characterized by a *rate adjustment* $a_1 = 0.905$ and an *SNR gap* $a_2 = 1.34$ [23].

We encode six 160-frame videos, one for each user, with different spatial-temporal complexities, i.e., Foreman, Ice, Soccer, Crew, Football and Mobile,[1] in CIF resolution with a frame-rate of 30 frames per second. Each sequence is encoded IDR-period-by-IDR-period by the JSVM reference software [27] with the GOP size and IDR period set to 8 and 16 frames, respectively. The encoded stream consists of one base layer and two enhancement layers, and the basis quantization parameters for encoding the three layers are set to 40, 34 and 28, respectively. Each enhancement layer is further split into five MGS layers with MGS vector [3 2 4 2 5]. Then, the post-processing priority level assignment is carried out. The estimate of the three parameters of model (1) is performed every IDR period. The duration of the *application period* is set to an IDR period, which leads to an *application period* window $N_{\text{slot}} = 1066$.

Moreover, we set the size of an RTP packet to 600 bytes and simulate an RTP packet loss rate r_{rtp} of 5% as in [21] and [3]. The maximum number of bytes per RS codeword in the UXP scheme (i.e., the maximum number of RTP packets) is set to 255. The minimum number of packets is dependent on the video content and on r_{rtp} since it has to ensure that the base layer is transmitted.

To assess the individual received video quality, we use the PSNR calculated using the luminance MSE, averaged over all 160 frames, if not specified otherwise, i.e.,

$$PSNR = 10\log_{10}\left(\frac{255^2}{aveMSE}\right). \tag{16}$$

To measure the system efficiency, we average the PSNRs for all user received videos, i.e.,

$$avePSNR = (1/K)\sum_{k\in\mathcal{K}} PSNR_k.$$

[1]The video sequences were downloaded from ftp://ftp.tnt.uni-hannover.de/pub/svc/testsequences/.

The higher avePSNR is, the higher system efficiency we have. On the other hand, the quality fairness is evaluated using the standard deviation of the PSNRs, i.e.,

$$stdPSNR = \sqrt{(1/K)\sum_{k\in\mathcal{K}}(PSNR_k - avePSNR)^2}.$$

Lower stdPSNR corresponds to fairer service.

We will use the acronym σ-F to refer to the resource allocation obtained by solving the problem with relaxed fairness constraints (10). Recall that the pure-quality fairness scheme (PF) and the maximum efficiency scheme (ME) are the two extreme cases of σ-F, corresponding to $\sigma = 0$, respectively, $\sigma = \infty$. We will compare the performance of σ-F, for intermediate values of σ, with the extremes PF[2] and ME. We also include the comparison with an equal-rate adaptation scheme, referred to as ERA, which provides fairness among users in terms of allocated video rate without violating the maximum and minimum rate constraints. The ERA problem can be formulated from problem (5) by replacing the fairness constraints in (5c) with new rate-fair constraints, i.e., $R_k(\boldsymbol{\tau}, \boldsymbol{p}) = \hat{F}_k(f), \forall k \in \mathcal{K}$, with the objective of maximizing f, where $f \geq 0$ and $\hat{F}_k(f)$ is defined similarly to (4). The solution to this problem can be obtained by using the ILA algorithm where the APP layer algorithm aims to find an optimal rate-fair solution rather than a quality-fair solution.

We emphasize that in our experiments the value of the expected PSNR (i.e., accounting for all packet loss scenarios) was generally very close to the PSNR of the transmitted video (i.e., the value achieved when all packets are received). This fact suggests that the UXP scheme generally ensures that the whole transmitted video stream can be correctly recovered from packet erasures with very high probability. Therefore, we use the PSNR value of the transmitted video sequence in our assessment in the sequel.

Fig. 3 illustrates the performance of the σ-F scheme, in terms of fairness and system efficiency, for various values of σ, in comparison with ERA. The values of σ range from 0 to 0.3 in increments of 0.01. Additionally, the values $0.32, 0.34, 0.36, 0.38, \infty$, are considered too. As σ increases, both avePSNR and stdPSNR increase until reaching the ME point. These results show that the fairness is traded off gracefully against the system efficiency as σ increases, as expected, due to the increasingly looser fairness constraints in problem (10). In our setting the ME point is already obtained when σ is about 0.28, meaning that the relaxed fairness inequalities do not constrain the solution anymore. It is important to note that as σ increases from 0 in sufficiently small increments, a dense set of trade-off points can be obtained covering the whole range between the PF and ME points. Additionally, it is worth pointing out the poor performance of ERA in terms of quality fairness compared to σ-F, since the latter can achieve (with $\sigma = 0.28$) the same avePSNR, but with about 0.8 dB lower standard deviation in PSNR than ERA.

It is interesting to compare our trade-off framework with that proposed in [15] in terms of flexibility. Examining the results

[2]It is worth pointing out that the results obtained with PF are expected to be very close to those obtained using the algorithm of [3].

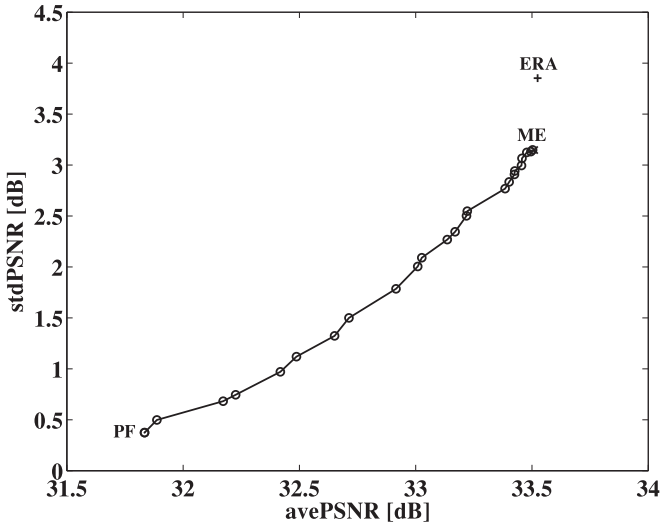

Fig. 3. stdPSNR versus avePSNR obtained with σ-F for various values of σ. The ERA point is also plotted.

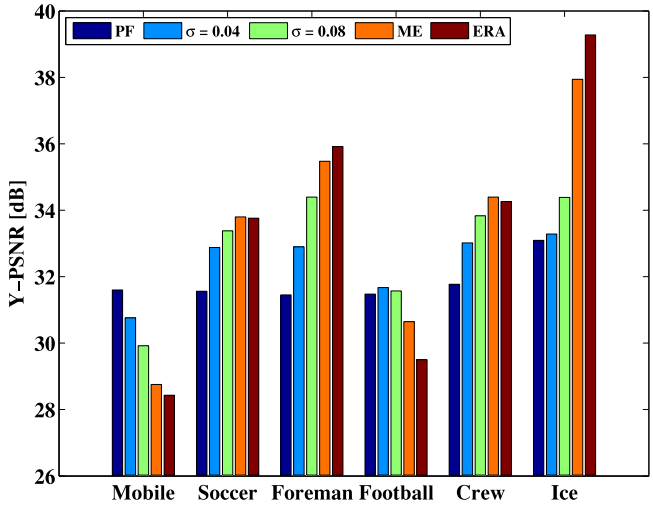

Fig. 4. PSNR of each video obtained with σ-F, for $\sigma = 0.04, 0.08$, PF, ME, and ERA.

reported in [15, Fig. 7] we see that there is a big gap between the pure quality-fairness point and the next trade-off point (about 2 dB in stdPSNR). Thus, their framework can achieve only points at some distance away from the pure quality-fairness point. On the other hand, our σ-F scheme overcomes this limitation, since it can achieve trade-off points close to PF. More specifically, our framework ensures trade-off points densely covering the whole range from PF to ME.

Fig. 4 shows the PSNR of each video corresponding to the σ-F allocation scheme for $\sigma = 0.04$ and $\sigma = 0.08$ in comparison with PF, ME and ERA. We see that in terms of PSNR difference the results are as expected for the σ-F scheme including its two extreme cases PF and ME. In other words, the absolute PSNR difference increases gradually with σ from the lowest value, achieved with PF, to the highest value, achieved with ME. Interestingly, for Soccer, Foreman, Crew and Ice, the PSNR value increases with σ, while for Mobile and Football it decreases, in the case of Football the decrease occuring after an initial

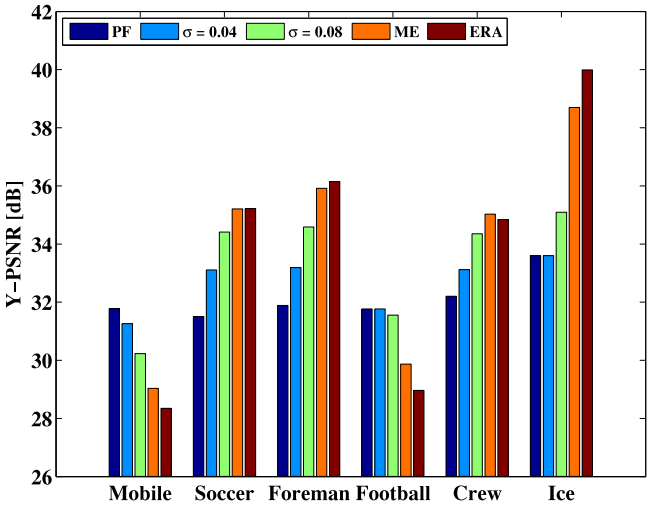

Fig. 5. PSNR of the third IDR for each video obtained with σ-F, for $\sigma = 0.04, 0.08$, PF, ME, and ERA.

(almost) stationary phase. Further, we observe that the PSNR values for ME and ERA are very close for all videos except for Football and Ice where we see an about 1 dB difference. Finally, these results suggest that, while the subjective quality achieved with ME and ERA for Soccer, Foreman, Crew and Ice may be high, the quality of Mobile and Football may be too low since the PSNR values for the latter group are much lower than for the former group.

We point out that the σ-F allocation scheme can also be applied to the scenario without RTP packet loss, and thus without the UXP module. Simulation results reported in [18] for the no packet loss case are similar in spirit to those reported here.

Next we perform the visual comparison between the proposed σ-F scheme ($\sigma = 0.08$) and PF, ME and ERA for decoded frames in the third IDR. For reference we also include a bar-graph showing the PSNR values achieved for the third IDR for the above mentioned five scenarios (Fig. 5). Fig. 6 compares reconstructed frames obtained with ME and ERA for Mobile and Football. We observe that the reconstructions are very similar. While differences exist they are hardly noticeable. The same observation holds for Soccer, Foreman, Crew and Ice. The corresponding reconstructions for Soccer, Foreman, Crew and Ice can be found in [28, Figs. 1 and 2].

Figs. 7 and 8 contain the frame reconstructions for ME, PF and σ-F, with $\sigma = 0.08$, for Soccer and Football (Fig. 7), respectively, Foreman and Mobile (Fig. 8). We observe that ME/ERA ensures a good reconstruction quality for Soccer and Foreman, but unsatisfactory for Football and Mobile, which contain blurred regions (the football field and the two players in the center in Football; the numbers in the calendar and the picture above them in Mobile). The PF scheme sharply improves the reconstruction for Football and Mobile, but it visibly deteriorates the reconstruction of Soccer and Foreman. For instance, in Soccer the number on the soccer player's shirt and the person with the dog in the background become blurred. In Foreman, the details of the face become unclear. On the other hand, we see that the σ-F scheme ensures a pleasing reconstruction for all four videos by finding a middle ground between PF and ME/ERA. A clear degradation in quality from ME to PF can also be observed in Crew and Ice, while σ-F visibly improves the quality in comparison with PF. The reconstructions for Crew and Ice are available in [28]. Thus, our results show that, while both PF and ME have the ability of providing a high visual quality to some users, this may come at the expense of drastically worsening the quality for other users. The σ-F scheme, on the other hand, gradually decreases the difference in visual quality between users until possibly a balancing point is reached where each user has a high enough perceived quality.

In order to understand this behavior it is instructive to analyze the rate-PSNR curves of the videos, which are plotted in Fig. 9. We see from the plots that for Football and Mobile the PSNR increases much more slowly with the rate than for the other four videos. This explains why the ERA scheme, which allocates the same rate to all users leads to high discrepancies in quality between the reconstructions. A similar effect is achieved with the ME scheme since it allocates more rate to videos that ensure higher increase of the PSNR. On the other hand, the discrepancies in performance exhibited by the PF scheme, which enforces very close PSNR values for all videos, may be due to the fact that the ranges of achievable PSNR values are different. Thus the PSNR values corresponding to maximum quality (i.e., for the highest rate) in the case of Soccer, Foreman, Crew and Ice are much higher than for Mobile and Footbal. This suggests that the correspondence between the PSNR value and the subjective quality level is different for the two groups of videos. Interestingly, lower achievable PSNR values correspond to the more demanding videos, i.e., Mobile and Football (referred to as group B), while higher values can be achieved by the videos in the other group (group A). This suggests that for the same level of subjective quality, videos in group A should have higher PSNR values than those in group B. This explains why the PF scheme cannot ensure the same subjective quality level to all videos.

The σ-F scheme, on the other hand, allows for the PSNR values to be selected from a wider region around the common PF PSNR value. Since the objective of the optimization is to increase the sum-PSNR within this region, the videos in group A will likely acquire PSNR values near the upper bound, while videos in group B will likely acquire PSNR values smaller than the PF value, but certainly not smaller than the lower bound. The parameter σ constraints this region ensuring that the lower bound guarantees a good enough quality for the videos in group B. On the other hand, the videos in group A exhibit an increase in quality versus the PF level. Thus, the σ-F scheme is more attractive than PF. The σ-F scheme is also more fair than PF in terms of perceived quality since it decreases the discrepancies in visual quality between users. Additionally, by gradually increasing σ it may be possible to reach a balancing point where the degradation of videos in group B is still small while the increase in quality of videos in group A is high enough so that all users are fully satisfied. In our experiments such a value of σ exists, but this might not happen all the time. Clearly, for such a balancing point to exist the available sum-rate must be high

Fig. 6. Visual comparison between ME (left) and ERA (right) for Mobile (top) and Football (bottom).

enough, at least higher than the sum of the minimum individual video rates needed for a good enough visual quality.

VII. Discussion

While in our experiments the σ-F scheme has a beneficial effect in terms of improving the visual quality, a natural question is whether this effect will persist in more general scenarios. We think that this behavior can be generalized as explained next. Consider a scenario where neither PF, nor ME ensure sufficient visual quality to all users, but the set of videos can be partitioned into the following groups:

1) group $\mathcal{A}$, containing videos for which ME ensures a very high visual quality, but the quality under PF is too low;
2) group $\mathcal{B}$, containing videos for which PF ensures a very high visual quality, but the quality under ME is too low;
3) group $\mathcal{C}$, containing videos for which both ME and PF ensure very good perceived quality.

Since videos in groups $\mathcal{B}$ and $\mathcal{C}$ have very high quality in the PF regime, they can afford to lose some rate and still maintain a pleasing quality. Thus, the σ-F scheme guarantees good enough quality for groups $\mathcal{B}$ and $\mathcal{C}$ if σ is sufficiently small, say $\sigma \leq \sigma_1$ for some σ_1.

Assume now that the PSNRs of videos in group $\mathcal{A}$ increase in the σ-F case versus the PF case, and let σ_2 be the smallest σ which brings the perceived quality of these videos at a high enough level. Thus, if $\sigma_2 \leq \sigma \leq \sigma_1$, all three groups of users are satisfied in the σ-F case. On the other hand, if $\sigma_2 > \sigma_1$ such a balancing σ value does not exist. However, for $\sigma < \sigma_1$, the σ-F scheme still brings some improvement in quality for videos in group $\mathcal{A}$, thus lowering the users' dissatisfaction versus the PF case. In the same time, users in groups $\mathcal{B}$ and $\mathcal{C}$ remain fully satisfied. We conclude that the σ-F scheme may bring some benefit versus the PF scenario even if σ is small.

The above conclusions rely on the assumption that the PSNRs of videos in group $\mathcal{A}$ are guaranteed to increase in the σ-F case versus the PF case. Next we present some theoretical results to support this claim.

Let us denote by $F_{k,\sigma}$ the rate assigned to user k in the σ-F scenario, for $0 \leq \sigma \leq \infty$, where $\sigma = 0$ corresponds to PF, and $\sigma = \infty$ to ME. Further, let $q'_{k,\sigma} \triangleq Q'_k(F_{k,\sigma})$, where $Q'_k(F_k)$ denotes the derivative of $Q_k(F_k)$. As discussed previously, the ME scheme is equivalent to σ_M-F for some large enough σ_M. For simplicity, let us assume that $Q_k(F_{k,\infty}) \in (q^*(1-\sigma_M), q^*(1+\sigma_M)) \subseteq (Q_{k,\min}, Q_{k,\max})$. Then the interval constraining the rate for user k in problem (11) is $[F_k(q^*(1-\sigma)), F_k(q^*(1+\sigma))]$, for all $0 < \sigma < \sigma_M$. Additionally, denote for each k and $\sigma \geq 0$, $q'_{k,\sigma+} \triangleq Q'_k(F_k(q^*(1-\sigma)))$ and $q'_{k,\sigma-} \triangleq Q'_k(F_k(q^*(1+\sigma)))$. Since the functions $Q_k(F_k)$ are strictly

Fig. 7. Visual comparison between ME (first row), PF (second row), and σ-F with $\sigma = 0.08$ (third row) for Soccer (left) and Football (right).

concave, it follows that $q'_{k,\sigma+} > q'_{k,0} > q'_{k,\sigma-}$ for all k and $\sigma > 0$. The following lemma, proved in the appendix, will be used in the sequel.

Lemma 2: For each $0 < \sigma < \sigma_M$ there is some value q'_σ such that the following hold for each $1 \le k \le K$.

a) If $F_{k,\sigma} \in (F_k(q^*(1-\sigma), F_k(q^*(1+\sigma)))$ then $q'_{k,\sigma} = q'_\sigma$.
b) If $F_{k,\sigma} = F_k(q^*(1-\sigma))$ then $q'_{k,\sigma} \le q'_\sigma$.
c) If $F_{k,\sigma} = F_k(q^*(1+\sigma))$ then $q'_{k,\sigma} \ge q'_\sigma$.

Applying Lemma 2 for σ_M we obtain that $q'_{k,\infty} = q'_{k,\sigma_M} = q'_{\sigma_M} = q'_\infty$ for all k. We may assume that $q'_{1,0} \le q'_{2,0} \le \cdots \le q'_{K,0}$. Further, for each $\sigma \ge 0$ let $k_1(\sigma)$ be the largest k_1 for which $q'_{k,\sigma+} < q'_\infty$ for all $1 \le k \le k_1$, and let $k_2(\sigma)$ denote the smallest integer such that $q'_{k_2(\sigma)} > q'_\infty$ and

$$\sum_{j=k_2(\sigma)}^{K} \frac{1}{q'_{j,\sigma-}} < \sum_{i=1}^{k_1(\sigma)} \frac{1}{q'_{i,\sigma+}}. \tag{17}$$

Fig. 8. Visual comparison between ME (first row), PF (second row), and σ-F with $\sigma = 0.08$ (third row) for Foreman (left) and Mobile (right).

Additionally, let $\mathcal{K}(\sigma)$ denote the set of integers k such that $q'_{k,\sigma-} > q'_{k_2(\sigma)}$. Clearly, $\mathcal{K}(\sigma) \subseteq \{k_2(\sigma)+1, \cdots, K\}$. The following result is proved in the appendix.

Proposition 1: Let $0 < \sigma < \sigma_M$. Then the following hold: $q'_\sigma < q'_{k_2(\sigma)}$, $Q_k(F_{k,\sigma}) > q^*$ for all $k \geq k_2(\sigma)$, and $Q_k(F_{k,\sigma}) = q^*(1+\sigma)$ for all $k \in \mathcal{K}(\sigma)$.

Remark 1: It is likely that indexes k with high ME PSNR value also have high $q'_{k,0}$. Thus, the set $\{k_2(\sigma), \cdots, K\}$ is likely the set of the N videos with highest PSNR in the ME regime, for $N = K - k_2(\sigma) + 1$. This set can be regarded as the set $\mathcal{A}$. Thus, according to Proposition 1, it is guaranteed that all videos in $\mathcal{A}$ have the σ-F PSNR higher than in the PF case. A moment of thought reveals that as σ decreases towards 0, N is nondecreasing, approaching the value $K - k_2(0) + 1$. Additionally, $K - k_2(0) + 1 \geq k_1(0)$ and the set of indexes $\{1, \cdots, k_1(0)\}$ corresponds to videos for which the ME PSNR is smaller than the PF PSNR. Thus, Proposition 1 guarantees that, for sufficiently small σ, the number of videos whose PSNR increases

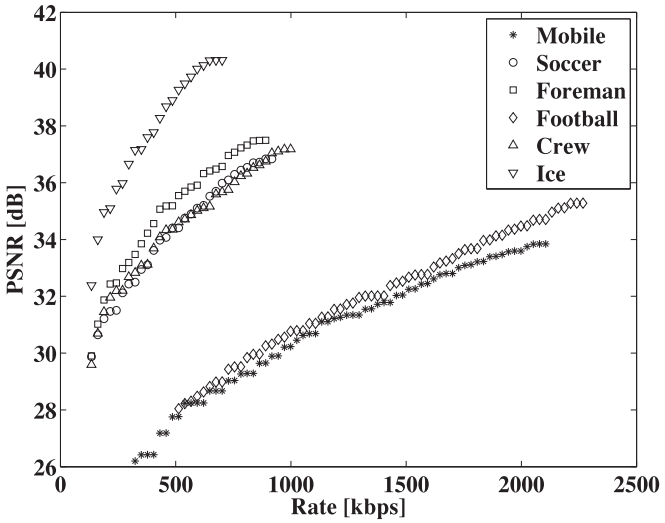

Fig. 9. Rate-PNSR curves for the third IDR for all six video sequences.

versus the PF case is at least equal to the number of videos whose ME PSNR is smaller than in the PF case.

VIII. Conclusion

In this paper, we tackle the problem of sending scalable videos to multiple users over OFDMA wireless networks considering the trade-off between quality fairness and system efficiency. We propose a cross-layer optimization framework for maximizing the sum of the PSNR values corresponding to average user rates subject to relaxed PSNR fair constraints. The optimization problem is solved in two steps. In the first step, we solve a pure quality-fairness (PF) problem which maximizes the common PSNR value enforced to all users. In the second step, we relax the constraints on quality fairness by allowing individual PSNRs to be some distance away from the PF value. This relative distance is controlled by a parameter σ, which thus governs the trade-off between quality fairness and efficiency. Our simulation results demonstrate that, by slowly increasing σ, a rich set of points gracefully trading off quality fairness for system efficiency, can be obtained. Furthermore, our experiments show that the maximum efficiency allocation may degrade the quality of the high demanding videos, while the PF scheme may drastically reduce the subjective quality of the low complexity videos. The proposed scheme is able to overcome the above disadvantages with an appropriate choice of σ, leading to a high subjective quality for all videos.

Appendix A

Proof of Lemma 1 First we will show that $\mathcal{R}_f \subseteq \mathcal{R}_f^c$. For this let $q \geq 0$. Then we have to show that

$$\Delta(\hat{Q}_i(q), \hat{Q}_j(q)) = 0 \tag{18}$$

for all $i \neq j \in \mathcal{K}$. Let us fix some $i \neq j \in \mathcal{K}$. Next we have to distinguish between the following cases: 1) $Q_{i,\min} \leq q \leq Q_{i,\max}$ and $Q_{j,\min} \leq q \leq Q_{j,\max}$; 2) $Q_{j,\min} \leq q < Q_{i,\min}$ or $Q_{i,\min} \leq q < Q_{j,\min}$; 3) $q < Q_{i,\min}$ and $q < Q_{j,\min}$; 4) $Q_{j,\max} \leq q < Q_{i,\max}$ or $Q_{i,\max} \leq q < Q_{j,\max}$; 5) $q > Q_{i,\max}$ and $q > Q_{j,\max}$. In case 1) we have $\hat{Q}_i(q) = q = \hat{Q}_j(q)$, thus (18) trivially holds. In case 2) if inequalities $Q_{j,\min} \leq q < Q_{i,\min}$ hold, then we have $\hat{Q}_i(q) = Q_{i,\min} > q = \hat{Q}_j(q)$, which implies that (18) is true. The other subcase can be treated similarly. Consider now case 3). Then one has $\hat{Q}_i(q) = Q_{i,\min}$ and $\hat{Q}_j(q) = Q_{j,\min}$ and (18) holds again. The other cases can be treated analogously.

Let us prove now that $\mathcal{R}_f^c \subseteq \mathcal{R}_f$. Let $\boldsymbol{R} \in \mathcal{R}_f^c$ and $Q_k = F_k^{-1}(R_k)$ for each $k \in \mathcal{K}$. We have to show that there is some $q \geq 0$ such that $\hat{Q}_k(q) = Q_k$ for all $k \in \mathcal{K}$. To this end, consider first the following sets of indexes: $\mathcal{I}_0 \triangleq \{k \in \mathcal{K} | Q_{k,\min} < Q_k < Q_{k,\max}\}$, $\mathcal{I}_1 \triangleq \{k \in \mathcal{K} | Q_{k,\min} = Q_k\}$ and $\mathcal{I}_2 \triangleq \{k \in \mathcal{K} | Q_{k,\max} = Q_k\}$.

Consider now the case when $\mathcal{I}_0$ is non-empty. The fact that $\Delta(Q_i, Q_j) = 0$ for all i, j implies that the value Q_k for all $k \in \mathcal{I}_0$ is the same. Let us choose q as the common value of Q_k for $k \in \mathcal{I}_0$. It follows that $\hat{Q}_k(q) = Q_k k \in \mathcal{I}_0$. Further, for $i \in \mathcal{I}_1$, the fact that $\Delta(Q_i, Q_k) = 0$ for any $k \in \mathcal{I}_0$, implies that $q \leq Q_{i,\min}$, leading to the conclusion that $\hat{Q}_i(q) = Q_{i,\min} = Q_i$. Likewise, for $j \in \mathcal{I}_2$, the fact that $\Delta(Q_j, Q_k) = 0$ for any $k \in \mathcal{I}_0$, implies that $q \geq Q_{j,\max}$, leading to the conclusion that $\hat{Q}_j(q) = Q_{j,\max} = Q_j$.

Finally, we have to consider the case when $\mathcal{I}_0$ is empty. If $\mathcal{I}_1 \neq \emptyset$ we choose $q \triangleq \min\{Q_{k,\min} | k \in \mathcal{I}_1\}$. Then clearly, $\hat{Q}_k(q) = Q_{k,\min} = Q_k$ for any $k \in \mathcal{I}_1$. For $j \in \mathcal{I}_2$, the fact that $\Delta(Q_j, Q_k) = 0$ for any $k \in \mathcal{I}_1$, implies that $q \geq Q_{j,\max}$, leading to the conclusion that $\hat{Q}_j(q) = Q_{j,\max} = Q_j$. On the other hand, in the case when $\mathcal{I}_1 = \emptyset$ and $\mathcal{I}_2 \neq \emptyset$ we choose $q \triangleq \max\{Q_k | k \in \mathcal{I}_2\}$ and the conclusion follows. With this observation the proof is complete.

Proof of Lemma 2: Let us consider two distinct users i and j such that $F_i(q^*(1-\sigma)) \leq F_{i,\sigma} < F_i(q^*(1+\sigma))$ and $F_j(q^*(1-\sigma)) < F_{j,\sigma} \leq F_i(q^*(1+\sigma))$. Let $\delta > 0$ be such that $F_{i,\sigma} + \delta \leq F_i(q^*(1+\sigma))$ and $F_j(q^*(1-\sigma) \leq F_{j,\sigma} - \delta$. If the optimal σ-F rate assignment is changed such that an amount of δ rate is moved from user j to user i, the sum-PSNR cannot increase. Therefore

$$Q_i(F_{i,\sigma} + \delta) + Q_j(F_{j,\sigma} - \delta) \leq Q_i(F_{i,\sigma}) + Q_j(F_{j,\sigma})$$

which implies that

$$\frac{Q_i(F_{i,\sigma} + \delta) - Q_i(F_{i,\sigma})}{\delta} \leq \frac{Q_j(F_{j,\sigma}) - Q_j(F_{j,\sigma} - \delta)}{\delta}.$$

Applying the limit as $\delta \to 0$, we obtain that $q'_{i,\sigma} \leq q'_{j,\sigma}$. If additionally, $F_i(q^*(1-\sigma)) < F_{i,\sigma}$ and $F_{j,\sigma} < F_i(q^*(1+\sigma))$, then we have $q'_{i,\sigma} \geq q'_{j,\sigma}$, leading to $q'_{i,\sigma} = q'_{j,\sigma}$. Thus, q'_σ is the common value of $q'_{k,\sigma}$ for users k falling in Case a). The rest of the claim follows based on similar arguments. ■

Proof of Proposition 1: Assume first that $q'_\sigma < q'_{k_2(\sigma)}$ holds. Then $q'_\sigma < q'_k$ for all $k \geq k_2(\sigma)$, and, in virtue of Lemma 2 and of the concavity of $Q_k(F_k)$ it follows that $Q_k(F_{k,\sigma}) > q^*$, for all $k \geq k_2(\sigma)$. Moreover, we have $q'_{k,\sigma-} > q'_\sigma$ for all $k \in \mathcal{K}(\sigma)$. Thus, Lemma 2 implies that $F_{k,\sigma} = F_k(q^*(1+\sigma))$, for $k \in \mathcal{K}(\sigma)$. Thus, it only remains to prove that $q'_\sigma < q'_{k_2(\sigma)}$ holds.

Assume, for the sake of contradiction, that $q'_\sigma \geq q'_{k_2(\sigma)}$. Then $q'_\sigma \geq q'_{i,\sigma+}$ for all $i \leq k_1(\sigma)$, and $q'_\sigma > q'_j$ for all $j < k_2(\sigma)$. Using Lemma 2 it follows that $F_{i,\sigma} = F_i(q^*(1-\sigma))$ for all $i \leq k_1(\sigma)$ and $F_{j,\sigma} < F_j(q^*)$ for all $j < k_2(\sigma)$. Since the sum-rate under PF and σ-F is the same, it follows that

$$\sum_{i=1}^{k_1(\sigma)} (F_i(q^*) - F_i(q^*(1-\sigma)) \leq \sum_{k=k_2(\sigma)}^{K} (F_k(q^*(1+\sigma)) - F_k(q^*). \tag{19}$$

Since $Q_k(F_k)$ is concave it follows that

$$q'_{k,\sigma-} \leq \frac{q^*(1+\sigma) - q^*}{F_k(q^*(1+\sigma)) - F_k(q^*)} \leq q'_k$$

which implies that $F_k(q^*(1+\sigma)) - F_k(q^*) \leq \frac{q^*\sigma}{q'_{k,\sigma-}}$. Similarly, we obtain that $F_k(q^*) - F_k(q^*(1-\sigma)) \geq \frac{q^*\sigma}{q'_{k,\sigma+}}$. Combining these relations with (19) it follows that

$$\sum_{i=1}^{k_1(\sigma)} \frac{q^*\sigma}{q'_{i,\sigma+}} \leq \sum_{k=k_2(\sigma)}^{K} \frac{q^*\sigma}{q'_{k,\sigma-}} \tag{20}$$

which contradicts relation (17), completing the proof. ■

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019," Cisco Inc., San Jose, CA, USA, Feb. 2015.

[2] M. van Der Schaar and S. N. Sai, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Commun.*, vol. 12, no. 4, pp. 50–58, Aug. 2005.

[3] S. Cicalò and V. Tralli, "Distortion-fair cross-layer resource allocation for scalable video transmission in OFDMA wireless networks," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 848–863, Jan. 2014.

[4] *Air Interface for Fixed and Mobile Broadband Wireless Access Systems*, IEEE Standard 802.16e-2005, 2006.

[5] *3rd Generation Partnership Project, Technical Specification Group Radio Access Network; Physical Layer Aspects for Evolved Universalterrestrial Radio Access (UTRA)*, 3GPP Standard TR 25.814 v.7.0.0, Sep. 2006.

[6] D. Munaretto and M. Zorzi, "Robust opportunistic broadcast scheduling for scalable video streaming," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2012, pp. 2134–2139.

[7] 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.

[8] H. Ha, C. Yim, and Y. Y. Kim, "Cross-layer multiuser resource allocation for video communication over OFDM networks," *Comput. Commun.*, vol. 31, no. 15, pp. 3553–3563, Sep. 2008.

[9] Y. Li, Z. Li, M. Chiang, and A. R. Calderbank, "Content-aware distortion-fair video streaming in congested networks," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1182–1193, Oct. 2009.

[10] N. Khan, M. G. Martini, and Z. Bharucha, "Quality-aware fair downlink scheduling for scalable video transmission over LTE systems," in *Proc. IEEE 13th Int. Workshop Signal Process. Adv. Wireless Commun.*, Jun. 2012, pp. 334–338.

[11] Z. Chen, M. Li, and Y.-P. Tan, "Perception-aware multiple scalable video streaming over WLANS," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 675–678, Jul. 2010.

[12] A. A. Khalek, C. Caramanis, and R. W. Heath, "Delay-constrained video transmission: Quality-driven resource allocation and scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 60–75, Feb. 2015.

[13] E. Maani, P. V. Pahalawatta, R. Berry, T. N. Pappas, and A. K. Katsaggelos, "Resource allocation for downlink multiuser video transmission over wireless lossy networks," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1663–1671, Sep. 2008.

[14] L. He and G. Liu, "Quality-driven cross-layer design for h.264/avc video transmission over OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6768–6782, Dec. 2014.

[15] G.-M. Su, Z. Han, M. Wu, and K. Liu, "A scalable multiuser framework for video over ofdm networks: Fairness and efficiency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 10, pp. 1217–1231, Oct. 2006.

[16] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4359–4372, Jul. 2011.

[17] S. Dumitrescu, X. Wu, and Z. Wang, "Efficient algorithms for optimal uneven protection of single and multiple scalable code streams against packet erasures," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1466–1474, Nov. 2007.

[18] K. Lin, "Quality fairness-oriented cross-layer resource allocation for scalable video delivery over OFDMA wireless networks," M.S. thesis, Dept. Elec. Comp. Eng., McMaster Univ., Hamilton, ON, USA, 2015.

[19] J. De Cock, S. Notebaert, P. Lambert, and R. Van De Walle, "Architectures for fast transcoding of h.264/avc to quality-scalable SVC streams," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1209–1224, Nov. 2009.

[20] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 1012–1032, Jun. 2000.

[21] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Channel aware multiuser scalable video streaming over lossy under-provisioned channels: Modeling and analysis," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1366–1381, Nov. 2008.

[22] S. Cicalò, A. Haseeb, and V. Tralli, "Fairness-oriented multi-stream rate adaptation using scalable video coding," *Signal Process., Image Commun.*, vol. 27, no. 8, pp. 800–813, Sep. 2012.

[23] M. Mazzotti, S. Moretti, and M. Chiani, "Multiuser resource allocation with adaptive modulation and LDPC coding for heterogeneous traffic in OFDMA downlink," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 2915–2925, Oct. 2012.

[24] I. C. Wong and B. L. Evans, "Adaptive downlink ofdma resource allocation," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, Oct. 2008, pp. 2203–2207.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] E. SMG, "Universal mobile telecommunications system (UMTS); selection procedures for the choice of radio transmission technologies of the UMTS," ETSI, Sophia Antipolis, France, Tech. Rep. 101 112, Nov. 1997.

[27] J. Reichel, H. Schwarz, and M. Wien, "Joint scalable video model 11 (JSVM 11)," Joint Video Team, Geneva, Switzerland, Doc. JVT-X202, Jul. 2007.

[28] K. Lin and S. Dumitrescu, "Cross-layer resource allocation for scalable video over OFDMA wireless networks: Trade-off between quality fairness and efficiency. Supplementary figures," 2017. [Online]. Available: http://www.ece.mcmaster.ca/~sorina/papers/TMM2017SupplFigs.pdf.

Authors' photographs and biographies not available at the time of publication.