DOI:
10.1109/GLOCOM.2015.7417260

Document Version
Publisher's PDF, also known as Version of record

Delay-constrained Video Transmission: A Power-efficient Resource Allocation Approach for Guaranteed Perceptual Quality

Seyed Ehsan Ghoreishi, Adnan Aijaz, and A. Hamid Aghvami
Centre for Telecommunications Research, King's College London, London WC2R 2LS, U.K.
E-mail: {seyed_ehsan.ghoreishi, adnan.aijaz, hamid.aghvami}@kcl.ac.uk

***Abstract*—In this paper we investigate power-efficient resource allocation for transmission of perceptual quality guaranteed video over LTE downlink under delay quality of service (QoS) constraints. We formulate the resource allocation problem as the minimization of sum power in the downlink under user-perceived quality and statistical delay QoS provisioning. We solve the problem using dual decomposition and employ the ellipsoid method to update dual variables. Experimental results have shown significant performance enhancement of the proposed system in terms of power efficiency while satisfying the statistical delay-bounded QoS and perceptual quality requirements.**

***Index Terms*—Perceptual quality, OFDMA, QoS, Scalable Video Coding, resource allocation.**

I. Introduction

The extensive growth in the adoption of smartphones and tablets has led to a continuous increase in video traffic to mobile devices. By 2019, mobile video will represent 72 percent of global mobile data traffic [1]. With the explosive increase of video traffic over wireless networks, it has become necessary to support more simultaneous video streams while guaranteeing a certain level of quality for individual users. Real-time video transmission requires maintaining stringent delay bounds and monitoring perceptual video quality to ensure a good user experience.

Deterministic delay bounds are hard to guarantee over wireless networks due to time varying nature of the wireless channel [2]. Moreover, for real time applications, the key QoS requirement is the bounded delay instead of average delay (which is only a marginal statistics). Therefore, statistical delay bound provisioning techniques are considered as a design guideline by defining constraints in terms of the delay-bound violation probability.

Achieving higher perceived video quality through maximizing source bit rate [3] results in an increase in throughput and power requirements. Higher system throughput does not necessarily mean higher video quality at the receiver. With the increase of system throughput, the average packet loss rate will also increase which could degrade user-perceived video quality [4]. In a multi-cell wireless networks, this will also generate higher interference to the other cells.

In order to support transmission of real-time applications over wireless channels using Orthogonal Frequency Division Multiple Access (OFDMA), cross-layer techniques have been adopted in literature for dynamic resource allocation. However, the common practice is to maximize quality [3], throughput [5], [6] or energy efficiency [6], [7] or minimize expected distortion [8] or power consumption [9]–[11].

To the best of our knowledge, the problem of perceptual quality-aware power-efficient resource allocation under statistical delay-bounded QoS guarantees in OFDMA systems has not been investigated before and hence is the main focus of this paper. We derive a resource allocation policy that minimizes power for a target user-perceived video quality such that all users in the network can achieve their target statistical delay bound.

The main contributions of this paper are as follows. (i) We provide an empirical mapping between perceptual video quality and source bit rate. (ii) We model the statistical delay QoS requirements in terms of queue length decaying rate which can be jointly determined by the effective bandwidth [2] of the arrival traffic and the effective capacity [12] of the wireless channel. (iii) We model the video specific resource allocation problem as the minimization of sum power in the downlink subject to perceptual quality guarantees, statistical QoS provisioning as well as the power and resource block (RB) allocation constraints of OFDMA. (iv) We employ a duality-based algorithm where dual variables are updated using the efficient ellipsoid method.

The rest of this paper is organized as follows. The system model and problem formulation are presented in Section II. Section III describes the proposed resource allocation algorithm. Section IV conducts numerical analyses of the model. The conclusion is presented in Section V.

II. System Model and Problem Formulation

We focus on the downlink of LTE networks and consider a single cell, multi-user scenario. The system consists of K mobile users (video streams), which are indexed by the set $\mathcal{K} \triangleq \{1, ..., k, ..., K\}$. Each pre-encoded H.264/SVC video stream k requires a bounded delay of $d_k^{\max}$, a delay violation probability of Γ_k and a source bit rate of $A_k^{\min}$ bits per second which guarantees a target perceptual video quality Q_k. We assume that the total number of available RBs are indexed by the set $\mathcal{L} \triangleq \{1, ..., l, ..., L\}$.

Using the source bit rate adaptation module (Fig. 1a), for a user k, we encode the video sequence at a target bit rate $A_k^{\min}$ which satisfies a certain quality perception Q_k. We then

978-1-4799-5952-5/15/$31.00 ©2015 IEEE

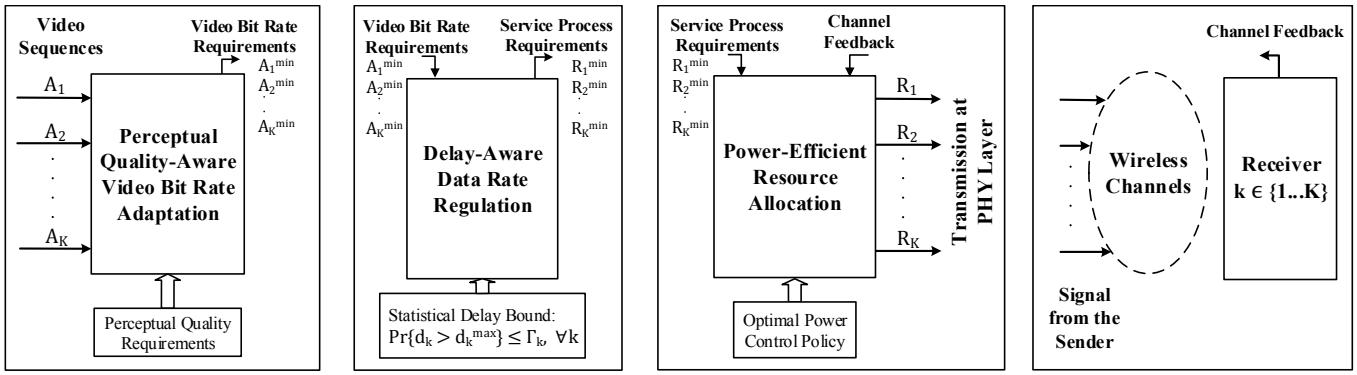


Fig. 1: The system modeling framework for video transmission over wireless network: (a) Bit rate adaptation module. (b) Data rate adaptation module. (c) Resource allocation module. (d) Receiver.

determine θ_k and the minimum required data rate, $R_k^{\min}$, in order to guarantee a specific delay QoS requirement given by $(d_k^{\max}, \Gamma_k)$, using the delay-aware data rate adaptation module (Fig. 1b). Lastly, the resource allocation module (Fig. 1c) allocates resources integrating $R_k^{\min}, \theta_k$ and the optimal power control policy $\mu_{k,l}$ presented in Theorem 1.

In the following subsections, we concisely explain the modular framework and formulate the resource allocation problem mathematically.

A. Perceptual Quality-Aware Source Bit Rate Adaptation

Peak signal-to-noise-ratio (PSNR) is widely used as a measure of the quality degradation of digitally encoded video. It is calculated as the error between the original and the reconstructed pictures. For a video sequence, PSNR can be derived as $\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{256^2}{\varepsilon^2(i)}\right)$, where $\varepsilon^2(i)$ is the pixel luminance mean-squared error between corresponding frame i in the reference and compressed videos, and N is the number of frames in the degraded video.

Perceptual quality is receiving considerable interest as a method to quantify the multimedia experience of mobile users. A survey of objective and subjective quality evaluation of scalable video coding (SVC) can be found in [13]. [14] relates perceptual quality to PSNR as

$$q_k = \frac{1}{1+e^{b_1^{(k)}(\text{PSNR}_k - b_2^{(k)})}}, \tag{1}$$

where $b_1^{(k)}$ and $b_2^{(k)}$ are parameters that depend on the video characteristics. In (1), $q_k = 0$ indicates the best quality and $q_k = 1$ indicates the worst quality. A perceptual Quality of Experience (QoE) metric derived in [15], based on the metric in [14], is expressed as

$$Q_k = q_{\max}^{(k)}\left(1-\frac{1}{1+e^{b_1^{(k)}(\text{PSNR}_k - b_2^{(k)})}}\right)\frac{1-e^{-b_3^{(k)}\frac{f^{(k)}}{f_{\max}^{(k)}}}}{1-e^{-b_3^{(k)}}}, \tag{2}$$

where $b_1^{(k)}$, $b_2^{(k)}$ and $b_3^{(k)}$ are parameters that depend on the video characteristics, $q_{\max}^{(k)}$ is a constant corresponding to maximum quality, $f^{(k)}$, is the frame rate at which the video is displayed and $f_{\max}^{(k)}$ is the maximum frame rate.

We encode source video sequences with H.264 SVC at multiple bit rates to produce compressed videos with different levels of quality for each video content. We use the error concealment method proposed in [16] in order to make (2) suitable to assess transmission over wireless systems and maintain the same frame rate after error concealment ($f^{(k)} = f_{\max}^{(k)}$). Therefore, the target perceptual video quality of the k^{th} stream, (2), can be simplified to

$$Q_k = q_{\max}^{(k)}\left(1-\frac{1}{1+e^{b_1^{(k)}(\text{PSNR}_k - b_2^{(k)})}}\right), \tag{3}$$

where $q_{\max}^{(k)} = 100$, thus displaying perceptual quality on a scale from 0 to 100.

For each user k, we determine the required minimum video source rate that satisfies the user's perceptual quality requirements. Therefore, having encoded the source streams at multiple bit rates, we measure the PSNR variation with the bit rate. For each encoded bit rate, we calculate the corresponding user-perceived quality using (3). Repeating over all sequences, we provide an empirical mapping between perceptual video quality and source bit rate. Therefore, for each user k, we find the minimum bit rate $A_k^{\min}$ that satisfies the user perceptual quality requirements.

B. Optimal Data Rate Adaptation for Statistical Delay QoS Guarantees

Statistical QoS guarantees have been extensively investigated in literature in the context of effective bandwidth $\mathcal{B}_E$ and effective capacity $\mathcal{C}_E$ functions [2], [12]. The effective bandwidth is defined as the minimum constant service rate required by a given arrival process for which a statistical QoS requirement specified by θ_k is fulfilled. θ_k characterizes the queue length decaying rate. Inspired by the effective bandwidth, [2] proposed effective capacity. The effective capacity

is defined as the maximum constant arrival rate that a given service process can support in order to guarantee statistical delay-QoS requirements specified by θ_k. Specifically, for a dynamic queuing system, under sufficient conditions, the queue length process, $\mathcal{Q}(t)$, converges in distribution to a random variable $\mathcal{Q}(\infty)$ such that [17]

$$-\lim_{z_t \to 0} \frac{\ln(Pr\{\mathcal{Q}(\infty)\})}{z_t} = \theta_k. \tag{4}$$

Considering a discrete-time arrival process $\{A[i], i = 1, 2, \cdots\}$ and the time-accumulated arrival process $S_B[t] \triangleq \sum_{i=1}^{t} A[i]$, effective bandwidth can be expressed as

$$\mathcal{B}_E(\theta_k) = -\lim_{t \to \infty} \frac{1}{t\theta_k} \log\left(\mathbb{E}\left\{e^{-\theta_k S_B[t]}\right\}\right), \tag{5}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation. Moreover, for effective bandwidth, the probability of delay-bound violation can be approximated as [12]

$$Pr\{d_k > d_k^{max}\} \approx e^{-\theta_k \mathcal{B}_E(\theta_k) d_k^{max}} \leq \Gamma_k, \tag{6}$$

where $d_k^{\max}$ and Γ_k are the delay-bound and delay violation probability thresholds for a user k. Likewise, given a discrete-time, stationary and ergodic stochastic service process $\{R[i], i = 1, 2, \cdots\}$ and the time-accumulated service process $S_C[t] \triangleq \sum_{i=1}^{t} R[i]$, effective capacity is given by [2], [12]

$$\mathcal{C}_E(\theta_k) = -\lim_{t \to \infty} \frac{1}{t\theta_k} \log\left(\mathbb{E}\left\{e^{-\theta_k S_C[t]}\right\}\right). \tag{7}$$

We drop the time-frame index [i] for the corresponding variables to simplify notations. Since the service rate $R_k^{\min}$ is an uncorrelated process, the effective capacity formulation simplifies to [18]

$$\mathcal{C}_E(\theta_k) = -\frac{1}{\theta_k} \log\left(\mathbb{E}\left\{e^{-\theta_k R_k^{\min}}\right\}\right). \tag{8}$$

We model the statistical delay guarantees in terms of QoS exponent, effective bandwidth/capacity, and delay-bound violation probability as in [18]. For a given arrival process $A_k^{\min}$ determined in section II-A, we get the corresponding effective bandwidth using (5). We then apply (6) to calculate the solution QoS exponent θ_k^* that guarantees a specific delay QoS requirement given by $(d_k^{\max}, \Gamma_k)$. Having found θ_k^* and $\mathcal{B}_E(\theta_k^*)$, we design the corresponding data rate, R_k^{min}, such that $\mathcal{C}_E(\theta_k^*) \geq \mathcal{B}_E(\theta_k^*)$ is satisfied

$$-\frac{1}{\theta_k^*} \log\left(\mathbb{E}\left\{e^{-\theta_k^* R_k^{\min}}\right\}\right) \geq \mathcal{B}_E(\theta_k^*). \tag{9}$$

1) Rate Requirements: The signal-to-noise ratio (SNR) for the k^{th} user over the l^{th} RB is given by $\gamma_{k,l} = \frac{P_{k,l}|h_{k,l}|^2}{\sigma^2}$, where $P_{k,l}$ is the transmission power of the k^{th} user over the l^{th} RB, $h_{k,l}$ is the channel fading coefficient, and σ^2 denotes the power of additive white Gaussian noise (AWGN). The item $\frac{|h_{k,l}|^2}{\sigma^2}$ is called channel-to-noise ratio (CNR), which fully reflects the quality of each wireless channel. This paper assumes perfect channel state information (CSI) at both base station (BS) and each user, which enables BS to dynamically allocate power and rate on each tone according to channel conditions.

Using Shannons capacity formula, the upper bound on the achievable service rate for the k^{th} user over the l^{th} RB, denoted by $R_{k,l}$ can be expressed as

$$R_{k,l} = B \log_2\left(1 + \frac{\mu_{k,l} P_{k,l} |h_{k,l}|^2}{\sigma^2}\right), \tag{10}$$

where B is the bandwidth of each RB and $\mu_{k,l}(\theta_k, \gamma_{k,l})$ denotes the optimal power control policy to be discussed later. Applying the power adaptation, the instantaneous transmit power becomes

$$\mathcal{P}_{k,l} = \mu_{k,l}(\theta_k, \gamma_{k,l}) P_{k,l} \qquad \forall l \in \mathcal{L}, \forall k \in \mathcal{K}. \tag{11}$$

2) Power Control Policy: The power control policy, denoted by $\mu_{k,l}(\theta_k, \gamma_{k,l})$, gives the relationship between R_k , θ_k and allocated power. Conventionally, the power control policy is expressed as a function of SNR only. However in our case, it is a function of both SNR and QoS exponent.

Theorem 1. *The optimal power control policy [5] for the k^{th} user over the l^{th} RB, denoted by $\mu_{k,l}(\theta_k, \gamma_{k,l})$ can be expressed as*

$$\mu_{k,l}(\theta_k, \gamma_{k,l}) = \frac{1}{\gamma_{k,l}} \left[\left(\frac{\gamma_{0_{k,l}}}{\gamma_{k,l}}\right)^{\frac{1}{q_k - 1}} - 1\right]^+, \tag{12}$$

where $[x]^+ = \max(0, x)$, $q_k = -\frac{\theta_k B}{\ln 2}$ is defined as the normalized QoS exponent and $\gamma_{0_{k,l}}$ is the cutoff SNR.

Proof. The proof directly extends from [5]. □

C. Problem Formulation

The resource allocation problem is mathematically formulated as

$$\min_{P_{k,l}, x_{k,l}} \; P_s = \sum_{k=1}^{K} \sum_{l=1}^{L} \mathcal{P}_{k,l} x_{k,l} \tag{13}$$

subject to:

$$\sum_{l=1}^{L} B \log_2(1 + \frac{\mathcal{P}_{k,l}|h_{k,l}|^2}{\sigma^2}) x_{k,l} \geq R_k^{\min} \quad \forall\, k \in \mathcal{K} \tag{13a}$$

$$\sum_{k=1}^{K} \sum_{l=1}^{L} \mathcal{P}_{k,l} x_{k,l} \leq P^{\max} \tag{13b}$$

$$\sum_{k=1}^{K} x_{k,l} \leq 1 \qquad \forall\, l \in \mathcal{L} \tag{13c}$$

$$x_{k,l} \in \{0, 1\},\ \mathcal{P}_{k,l} \geq 0 \qquad \forall k, \forall l. \tag{13d}$$

The objective of the optimization problem in (13) is power and RB allocation for different users in order to minimize the cumulative transmit power, P_s, in the downlink. It is subject to different constraints of OFDMA along with satisfying the statistical delay-bound, maximum transmission power and data rate requirements. R_k^{min} in (13a) is the minimum required data

rate for the delay constrained video services of receiver k, specified in Section II-B. The value of $P^{\max}$ in (13b) puts an upper limit on the power radiated by the transmitter. (13c) and (13d) indicate that each RB can be allocated to one receiver exclusively. We use binary variables $x_{k,l} \in \{0,1\}$ to represent the RB assignment in multi-user systems, where $x_{k,l} = 1$ indicates RB l is used to serve user k and $x_{k,l} = 0$ otherwise.

III. Duality-Based Resource Allocation

In this section we derive some desirable properties of the optimal solution. We propose to solve (13) using dual decomposition. The Lagrangian of problem (13) is given by

$$\begin{aligned}\boldsymbol{\mathcal{L}}(\mathbf{X},\mathbf{P},\lambda,\boldsymbol{\nu}) =& \sum_{k=1}^{K}\sum_{l=1}^{L}\mathcal{P}_{k,l}x_{k,l} + \lambda\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\mathcal{P}_{k,l}x_{k,l}\right.\\ &\left. - P^{\max}\right) + \sum_{k=1}^{K}\nu_k\left(R_k^{min} - \sum_{l=1}^{L}R_{k,l}x_{k,l}\right)\\ =& \sum_{l=1}^{L}\left[\sum_{k=1}^{K}(1+\lambda)\,\mathcal{P}_{k,l}x_{k,l} - \sum_{k,l}^{K}\nu_k R_{k,l}x_{k,l}\right]\\ &+ \sum_{k=1}^{K}\nu_k R_k^{min} - \lambda P^{\max}, \end{aligned} \tag{14}$$

where $\mathbf{X}$ and $\mathbf{P}$ are both $K \times L$ matrices with elements $x_{k,l}$ and $P_{k,l}$ respectively. λ is the dual variable for the power constraint and $\boldsymbol{\nu} = [\nu_1, ..., \nu_k, ..., \nu_K]$ is the dual vector for the data rate constraint. We define the Lagrangian dual function $g(\lambda, \boldsymbol{\nu})$ as

$$g(\lambda,\boldsymbol{\nu}) = \begin{cases} \min\limits_{\boldsymbol{X},\boldsymbol{P}} \ \boldsymbol{\mathcal{L}}(\mathbf{X},\mathbf{P},\lambda,\boldsymbol{\nu}) \\ \text{s.t.} \\ \sum\limits_{k=1}^{K} x_{k,l} \le 1 \qquad\qquad \forall\ l \in \mathcal{L} \\ x_{k,l} \in \{0,1\},\ \mathcal{P}_{k,l} \ge 0 \quad \forall\ k, \forall\ l, \end{cases} \tag{15}$$

and the dual problem is

$$G = \max_{\lambda \ge 0, \boldsymbol{\nu} \succeq 0} g(\lambda, \boldsymbol{\nu}). \tag{16}$$

In general, there is a non-zero duality gap in presence of integer constraints. However, when time-sharing condition is satisfied, we have an asymptotically zero duality gap as L goes to infinity, and for practical systems with finite L, the duality gap is still nearly zero [19]. Via Lagrangian relaxation (14), we have removed the coupling among RBs. Thus, $g(\lambda, \boldsymbol{\nu})$ is decomposed into L sub-problems which can be independently solved at each RB, given $(\lambda, \boldsymbol{\nu})$. The sub-problem at RB l is

$$\begin{aligned}\min_{\boldsymbol{X}_l,\boldsymbol{P}_l} \boldsymbol{\mathcal{L}}_l(\mathbf{X}_l,\mathbf{P}_l) =& \sum_{k=1}^{K}\mathcal{P}_{k,l}x_{k,l} + \sum_{k=1}^{K}\lambda\mathcal{P}_{k,l}x_{k,l}\\ &- \sum_{k,l}^{K}\nu_k R_{k,l}x_{k,l} \end{aligned} \tag{17}$$

subject to:

$$\sum_{k=1}^{K} x_{k,l} \le 1,\ x_{k,l} \in \{0,1\},\ \mathcal{P}_{k,l} \ge 0 \qquad \forall\ l,$$

where $\mathbf{X}_l$ and $\boldsymbol{\mathcal{P}}_l$ are vectors of $x_{k,l}$ and $\mathcal{P}_{k,l}$ at RB l. By visiting the constraints in (17), we note that $\mathbf{X}_l$ is an all-zero vector except for one binary non-zero entry. Hence, we can first calculate the optimal value of

$$\mathcal{F}_{k,l} = \begin{cases} \min\limits_{\mathbf{P}_l} \ (1+\lambda)\mathcal{P}_{k,l} - \nu_k R_{k,l} \\ \text{s.t.} \\ \mathcal{P}_{k,l} \ge 0 \qquad\qquad \forall\ k, \end{cases} \tag{18}$$

at each l and then find optimality for sub-problem l within the vector $\boldsymbol{\mathcal{F}}_l = [\mathcal{F}_{1,l}, \mathcal{F}_{2,l}, \cdots, \mathcal{F}_{K,l}]$. Therefore, we achieve the scheduling vector $\boldsymbol{X}_l$ for RB l as

$$x_{k,l} = \begin{cases} 1 & k = k^* = \arg\min\limits_{k} \boldsymbol{\mathcal{F}}_l, \mathcal{P}^*_{k,l} \ne 0 \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Substituting (11), (10) and (12) into (18), we have

$$\mathcal{F}_{k,l} = \begin{cases} \min\limits_{\mathbf{P}_l} \ (1+\lambda)\,\mu_{k,l}P_{k,l} \\ -\nu_k B \log_2\left(1 + \frac{\mu_{k,l}P_{k,l}|h_{k,l}|^2}{\sigma^2}\right) \\ \text{s.t.} \\ \mathcal{P}_{k,l} \ge 0 \qquad\qquad \forall\ k. \end{cases} \tag{20}$$

By taking derivative with respect to $P_{k,l}$, we obtain the optimal $P_{k,l}$ allocation on RB as l

$$P^*_{k,l} = \begin{cases} \gamma_{0_{k,l}}\left(\frac{\sigma^2}{|h_{k,l}|^2}\right)\left(\frac{\nu_k}{(1+\lambda)}\frac{B}{\ln 2}\frac{|h_{k,l}|^2}{\sigma^2} + 1\right)^{1-q_k} & k = k^* \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

By updating the dual vector $(\lambda, \boldsymbol{\nu})$ at each iteration, the Ellipsoid Method [20] can efficiently solve dual problem (16) and achieve dual optimality $(\lambda^*, \boldsymbol{\nu}^*)$. The subgradient is required by ellipsoid method at each iteration. The subgradient at the n^{th} iteration is derived in the following proposition.

Proposition 1. *For the optimization problem* (13) *with dual defined in* (16)*, a subgradient for* $g(\lambda, \boldsymbol{\nu})$*) is*

$$\begin{aligned} d(\lambda_k(n)) =& \sum_{k=1}^{K}\sum_{l=1}^{L}\mathcal{P}^*_{k,l}(n)x_{k,l}(n) - P^{\max}\\ d(\nu_k(n)) =& R_k^{min} - \sum_{l=1}^{L}R^*_{k,l}(n)x_{k,l}(n), \end{aligned} \tag{22}$$

where $\mathcal{P}_{k,l}(n)^* = \mu_{k,l}(\theta_k, \gamma^*_{k,l})P^*_{k,l}$ *and* $R^*_{k,l}(n) = B\log_2(1+\frac{\mu_{k,l}(\theta_k,\gamma^*_{k,l})P^*_{k,l}|h_{k,l}|^2}{\sigma^2})$. $P^*_{k,l}$ *minimizes* (15) *at* λ *and* $\boldsymbol{\nu}$.

Proof. By definition of $g(\lambda, \boldsymbol{\nu})$ in (15)

$$g(\lambda', \boldsymbol{\nu}') \leq \sum_{k=1}^{K}\sum_{l=1}^{L} \mathcal{P}^*_{k,l} x_{k,l} + \lambda' \Bigg(\sum_{k=1}^{K}\sum_{l=1}^{L} \mathcal{P}^*_{k,l} x_{k,l} - P^{\max} \Bigg) + \sum_{k=1}^{K} \nu'_k \left(R_k^{min} - \sum_{l=1}^{L} R^*_k x_{k,l} \right)$$
$$= g(\lambda, \boldsymbol{\nu}) + (\lambda' - \lambda) \Bigg(\sum_{k=1}^{K}\sum_{l=1}^{L} \mathcal{P}^*_{k,l} x_{k,l} - P^{\max} \Bigg) + \sum_{k=1}^{K} (\nu'_k - \nu_k) \left(R_k^{min} - \sum_{l=1}^{L} R^*_k x_{k,l} \right). \quad (23)$$

Thus, proposition 1 is proven using subgradient definition. □

Lemma 1. *The optimal dual variables* $(\lambda^*, \boldsymbol{\nu}^*)$ *must satisfy*

$$0 \leq \nu_k^* \leq \nu_k^{\max} = \frac{\ln 2}{B} \mu_\alpha P_{k,l} \left(1 + \lambda^{\max}\right) \quad \forall\, k \in \mathcal{K}, \quad (24)$$

$$0 \leq \lambda^* \leq \lambda^{\max} = \frac{B}{\ln 2} \frac{\nu^*}{\mu_\beta P_{k,l}}, \quad (25)$$

where μ_α *and* μ_β *are the total channel inversion [21], [22] and water-filling [21], [23] power-control policies respectively.*

Proof. The dual variables $(\lambda^*, \boldsymbol{\nu}^*)$ must satisfy the Karush-Kuhn-Tucker (KKT) conditions in order to be optimal. Taking the partial derivative of (17) at RB l with respect to $P_{k,l}$, we obtain (26) and (27).

$\mu_{k,l}$ is upper-bounded by the channel inversion scheme denoted by μ_α and lower-bounded by the water-filling power adoption denoted by μ_β [5]. Therefore, we obtain the upper bound $\nu_k^{\max}$ by letting $\mu_{k,l} = \mu_\alpha$ and $P_{k,l} = P^{\max}$ in (26) and the upper bound $\lambda^{\max}$ by substituting $\nu_k^{\max}$ and $\mu_{k,l} = \mu_\beta$ into (27). □

Using Lemma 1, one may choose an initial ellipsoid $\mathbf{A}(0)$ with a center $\mathbf{z}(0)$ in which the optimal $(\lambda^*, \boldsymbol{\nu}^*)$ reside. The details, e.g. the update algorithm and stopping criterion can be found in [20].

A summary of our proposed algorithm is provided in Algorithm 1.

Algorithm 1: Power efficient resource allocation

initialize $(\lambda(0), \boldsymbol{\nu}(0))$ and the initial ellipsoid, $\boldsymbol{A}(0)$;
repeat
  initialize $P_{k,l}$;
  for $l = 1$ *to* L **do**
    for $k = 1$ *to* K **do**
      Obtain the optimal $P_{k,l}$ through (21);
      Calculate vector $\mathcal{F}_{k,l}$ in (18) with optimal $P_{k,l}$;
      Get the optimal assignment for RB l by (19);
    end
  end
  Update $(\lambda, \boldsymbol{\nu})$ and $\boldsymbol{A}$ via the ellipsoid method with the subgradients in (22);
until *$(\lambda, \boldsymbol{\nu})$ convergence*;

IV. Numerical and Simulation Results

In our simulations, we consider the downlink of a single-cell OFDMA system. The system bandwidth is 10 MHz. Therefore, 50 usable RBs are available per transmission time interval (TTI). The channel model accounts for small scale Rayleigh fading, large scale path loss, and shadowing (log-normally distributed). We consider 8 uniformly distributed users in the coverage area with a minimum distance of 50 m from the eNodeB.

In this paper, video coding is performed by JSVM9.19.15. The CIF (352×288) video sequences City and Foreman are used in the simulations. The parameters $(b_1^{(k)}, b_2^{(k)})$ are set to (0.34, 29.09) for Foreman and (0.34, 26.3) for City as in [15]. The sequences are encoded at different bit rates with 30 fps to produce different levels of PSNRs (see Fig. 2a). Using (3), Fig. 2b presents an empirical mapping between PSNR and perceptual quality.

TABLE I: Simulation Configuration Parameters

Parameter	Value
Cell radius	1 km
Path loss	$128.1 + 37.6 \log_{10}(r), r$ in km
Standard deviation of shadowing	8 dB (90% cell edge coverage)
Quality requirements for different scenarios	
Scenario 1	Foreman sequence, $d_1^{max} = 150$ ms, $\Gamma_1 = 10^{-2}$, $Q_1 \approx 70$
Scenario 2	Foreman sequence, $d_2^{max} = 100$ ms, $\Gamma_2 = 10^{-3}$, $Q_2 \approx 80$
Scenario 3	City sequence, $d_3^{max} = 70$ ms, $\Gamma_3 = 10^{-4}$, $Q_3 \approx 90$

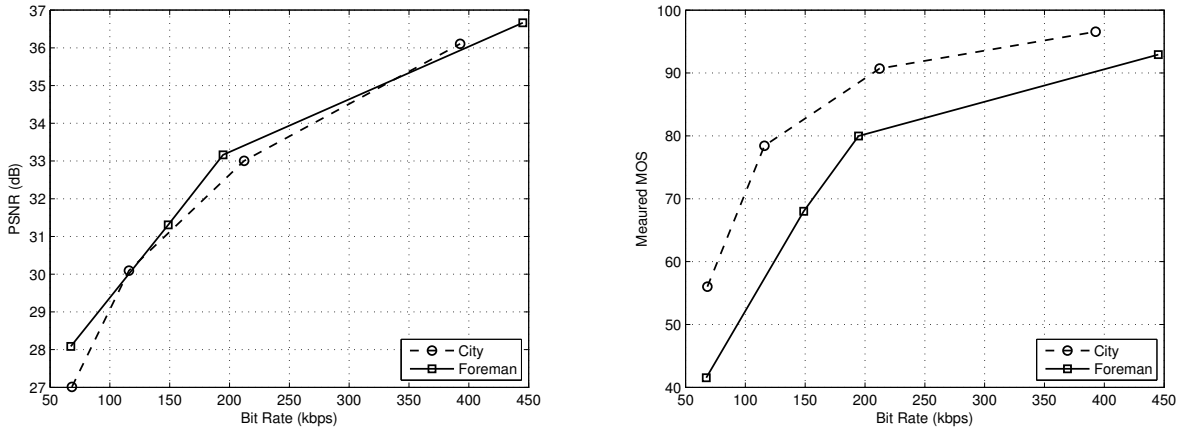
In order to analyze our proposed approach, we consider three scenarios in each of which, users have different quality and delay requirements as shown in TABLE I. *Scenario 1* has the highest and *Scenario 3* has the lowest quality requirements. The performance of the proposed algorithm is evaluated on these scenarios.

We compare our algorithm with WSPmin scheme in [9] and VAWS method in [24]. WSPmin minimizes the total transmission power with a minimum rate constraint on each user. In VAWS, subcarriers are assigned to satisfy minimum rate constraint with the assumption of equal power allocation per subcarrier. It then refines the initial uniform power allocation given the subcarrier assignment in the last stage to ensure that minimum rate requirements are met. It lastly repeats the previous phases to refine power allocation. Nevertheless, WSPmin and VAWS do not provide statistical delay QoS guarantees. The data rate requirements for the users served by WSPmin and VAWS are randomly varying from 100 kbps 400 kbps as multiples of 50 kbps.

Fig. 3 illustrates θ_k for a user k in *Scenario 2* for different delay QoS requirements $(d_k^{\max}, \Gamma_k)$. The θ_k for the target delay

$$\begin{aligned}\nu_k^* =& \frac{\ln 2}{B}\left(\frac{|h_{k,l}|^2}{\sigma^2}\right)^{\frac{q_k}{1-q_k}}\left(\frac{P_{k,l}}{\gamma_{0_{k,l}}}\right)^{\frac{1}{1-q_k}}(1+\lambda^*) - \frac{\ln 2}{B}\frac{\sigma^2}{|h_{k,l}|^2}(1+\lambda^*) \\ =& \frac{\ln 2}{B}\frac{1}{\gamma_{k,l}}\left[\left(\frac{\gamma_{k,l}}{\gamma_{0_{k,l}}}\right)^{\frac{1}{1-q_k}} - 1\right]P_{k,l}(1+\lambda^*) = \frac{\ln 2}{B}\mu_{k,l}P_{k,l}(1+\lambda^*)\end{aligned} \tag{26}$$

$$\lambda^* = \frac{B}{\ln 2}\frac{\nu_k^*}{\frac{1}{\gamma_{k,l}}\left[\left(\frac{\gamma_{k,l}}{\gamma_{0_{k,l}}}\right)^{\frac{1}{1-q}} - 1\right]P_{k,l}} = \frac{B}{\ln 2}\frac{\nu_k^*}{\mu_{k,l}P_{k,l}}. \tag{27}$$



(a) (b)

Fig. 2: Perceptual quality-rate mapping: (a) PSNR vs. bit rate. (b) MOS vs. bit rate.

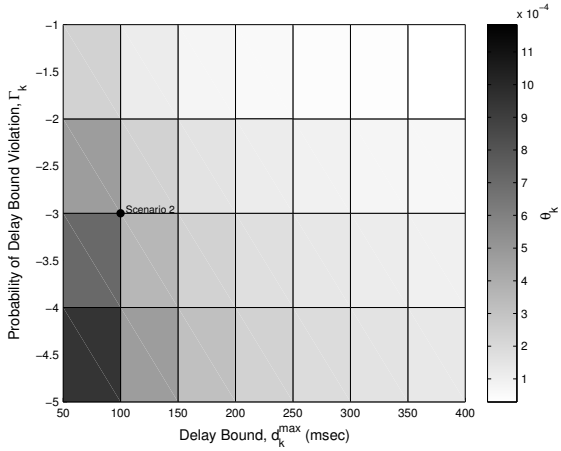


Fig. 3: Probability of delay violation of a user k in *Scenario 2* (y-axis in logarithmic scale).

bound and target delay bound violation probability in *Scenario 2* is highlighted on Fig. 3. Similarly, we can also plot the heatmaps of θ_k for the other scenarios, which are omitted for lack of space.

Fig. 4 plots the sum power for different scenarios and resource allocation schemes under different average cell border CNRs. As in [7], the noise variance σ^2 is set to ensure average cell border CNR ρ_0. We see that in terms of power efficiency, the proposed method performs considerably better than WSPmin and VAWS . For instance, in *Scenario 1*, with an average cell border CNR of 4 dB, our proposed approach can improve power efficiency by 52.9% compared with WSPmin and 72.2% compared with VAWS.

We also note that the power efficiency reduces as the video quality requirements increases. This is due to the fact that more power is allocated per RB in order to provide a more stringent delay QoS guarantee and satisfy the higher perceptual quality requirements. For instance, in *Scenario 3* which has higher quality requirements, with the same average cell border CNR of 4 dB, power efficiency is improved by 30.2% and 55.3% compared with WSPmin and VAWS respectively.

Fig. 5 demonstrates the Cumulative Distribution Function (CDF) of sum power for different scenarios and resource allocation algorithms generated over 100 iterations. The noise power is set to -174 dBm/Hz. We note that in *Scenario 1*, our proposed scheme outperforms both WSPmin and VAWS algorithms in terms of power efficiency by performing 16.21%

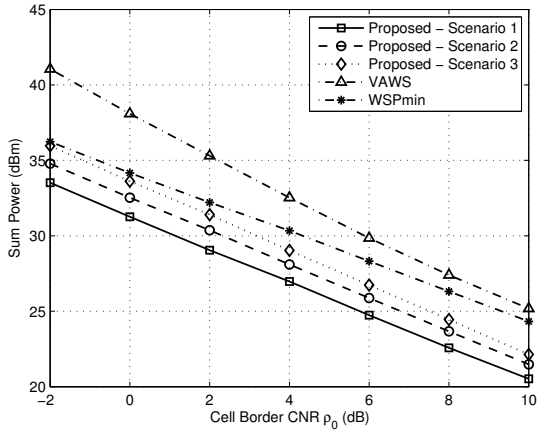


Fig. 4: Sum power versus border CNR, ρ_0.

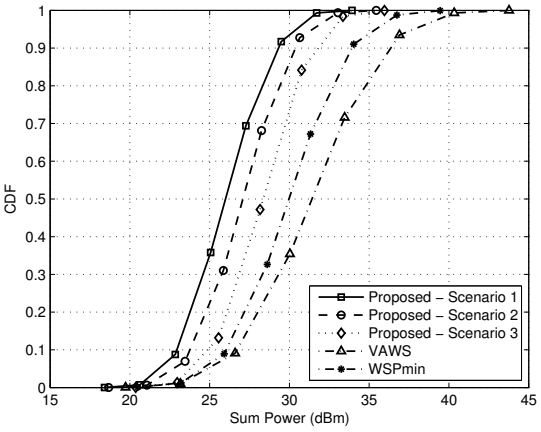


Fig. 5: CDF of sum power.

better than the former and 22.3% better than the latter in 90% of the times. Likewise, in *Scenario 3*, compared with WSPmin and VAWS, the proposed approach improves power efficiency by 10.25% and 18.58% in 90% of the times respectively.

Lastly, we discuss the complexity of the proposed scheme. The complexity to solve all sub-problems in (17) is $\mathcal{O}(KL)$. Therefore, the complexity of ellipsoid method with $(K+1)$ dual variables is $\mathcal{O}(KL(K+1)^2)$ [9]. The overall complexity can be estimated by $\mathcal{O}(KL(K+1)^2\log_2(\frac{1}{\epsilon}))$ where ϵ is the required accuracy (polynomial complexity).

V. Conclusion

In this paper we have investigated power efficient resource allocation for the downlink of LTE networks under user-perceived quality and statistical delay QoS constraints. The resource allocation problem is solved using duality-based approach. Numerical results have shown that the proposed resource allocation algorithm not only outperforms classical algorithms in terms of power efficiency but also satisfies the QoS requirements of different users for the target perceptual qualities.

References

[1] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*, Cisco White Paper, February 2015.

[2] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

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

[4] H. Luo, S. Ci, D. Wu, J. Wu, and H. Tang, "Quality-driven cross-layer optimized video delivery over lte," *IEEE Commun. Mag.*, vol. 48, no. 2, pp. 102–109, February 2010.

[5] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, August 2007.

[6] B. Li, S. Li, C. Xing, Z. Fei, and J. Kuang, "A QoE-based OFDM resource allocation scheme for energy efficiency and quality guarantee in multiuser-multiservice system," in *Globecom Workshops, 2012 IEEE*, Anaheim, California, December 2012, pp. 1293–1297.

[7] X. Xiao, X. Tao, and J. Lu, "QoS-aware energy-efficient radio resource scheduling in multi-user OFDMA systems," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 75–78, 2013.

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

[9] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *IEEE International Symposium on Information Theory*, Seattle, WA, USA, July 2006, pp. 1394–1398.

[10] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, June 2008.

[11] C. Isheden and G. Fettweis, "Energy-efficient link adaptation with shadow fading," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, Budapest, May 2011, pp. 1–5.

[12] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer Science & Business Media, 2000.

[13] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, "Quality assessment of multidimensional video scalability," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 38–46, April 2012.

[14] Video Quality Expert Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," www.vqeg.org, 2000.

[15] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286–298, March 2011.

[16] E. Yaacoub, F. Filali, and A. Abu-Dayya, "QoE enhancement of SVC video streaming over vehicular networks using cooperative LTE/802.11p communications," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 37–49, Feb 2015.

[17] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[18] Q. Du and X. Zhang, "Statistical qos provisionings for wireless unicast/multicast of multi-layer video streams," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 420–433, Apr. 2010.

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

[20] S. Boyd, "Ee364b course note," Stanford University, 2015. [Online]. Available: www.stanford.edu/class/ee364b

[21] A. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, October 1997.

[22] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, November 1997.

[23] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2006.

[24] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct 2014.