Green Video Delivery in LTE-based Heterogeneous Cellular Networks

Apostolos Galanopoulos*, Georgios Iosifidis†, Antonios Argyriou*, and Leandros Tassiulas†
*Department of Electrical and Computer Engineering, University of Thessaly, Volos, 38221, Greece.
†Department of Electrical Engineering, and Institute for Networking Science, Yale University, New Haven, 06511 CT, USA.

***Abstract*—In this paper we present an optimization framework that formalizes the inherent trade-off between the user perceived quality of wireless video, and the energy consumption cost of the network. The former is formulated in the context of the emerging heterogeneous cellular networks (HCN) based on LTE. We also consider users that employ dynamic adaptive streaming over HTTP (DASH). Our framework quantifies this trade-off carefully, by delving into the details of DASH, the LTE network, and the HCN architecture. The result is a complex problem that is solved in two levels. The master problem is responsible for decisions regarding the the user association and the average power they are allocated. The solution of this problem also entails a decision about the encoding rate of the DASH video segments. The previous decision is used in order to perform resource allocation at a finer level by considering the technical details of LTE that allocates resource blocks and power simultaneously. Numerical results are presented with realistic parameters for the LTE network and the video traffic.**

***Index Terms*—Heterogeneous cellular networks, video delivery, video quality adaptation, resource allocation, network optimization**

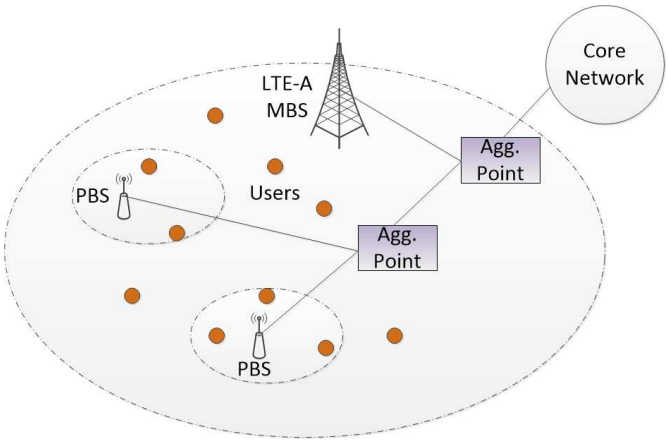

Fig. 1. An LTE-based HCN with a hierarchical backhaul topology. There are different types of base stations such as macrocellular (MBS) or pico-cell BSs (PBS). The aggregation points are switches that transfer the traffic of the BS to the core network .

I. Introduction

One of the most remarkable aspects of the mobile data tsunami that we are witnessing nowadays is that it is driven by an increasing volume of user demand for multiply-encoded and pre-stored video files [1], [2]. Mobile network operators (MNOs) aim to satisfy these requests with timely delivery of video content of high encoding quality, so as to increase the satisfaction of the users and hence their expected revenues. Nevertheless, this is a very challenging task that must be carefully addressed. On the one hand, it involves sophisticated network decisions as different users need to be served with different data rates due to their varying channel conditions. On the other hand, delivering high quality video files increases the load of the network and hence induces an often unbearable servicing cost for the MNOs.

One of the techniques that are currently gaining increasing popularity for addressing the first issue are the adaptive video streaming protocols. A prominent example is the dynamic adaptive streaming over HTTP (DASH) technique [3]. The main idea of DASH is that a video is stored as a sequence of short duration video segments with a typical time duration of 2 to 10 seconds. Each segment is available at different quality in terms of SNR encoding, spatial resolution, or even frame rate[1]. Based on the actual throughput that each user achieved during the delivery of the current segment, the protocol determines the quality (and hence the size) of the next segment to be delivered. The goal is to maximize the delivered video bit-rate while taking into account the actual network performance that each user experiences in practice.

However, it has been shown that allowing the users' devices to be fully responsible for requesting video segments (a pull-based content delivery system) results in inefficient usage of network resources, unfair allocation of network bandwidth (especially among flows sharing common resources), and even unnecessary variations of the bit rate [4], [5], [6], [7]. Therefore, it is imperative to employ a mechanism that will involve centralized video delivery decisions at the network side. Moreover, these decisions need to take into account the operating expenditures (OPEX) of the MNO. The latter are largely driven by the energy that is consumed by the active network components in the radio access network (RAN) such as the cellular base stations [8], [9], [10]. Clearly, from the perspective of the MNO, it would be ideal to minimize the energy costs while maximizing the user satisfaction by deliv-

[1]For example, in systems that use scalable video encoding (SVC), the quality of the delivered segment may refer to any combination of these different quality metrics.

ering high quality videos. However, these objectives are often conflicting. Additionally, they involve a variety of different network decisions such as the assignment of the users to the different base stations, and the allocation of the spectrum and transmission power of each base station to the users that it serves.

Devising such a servicing policy is a more intricate task nowadays where MNOs deploy small cells (micro-/pico-/femto-cells) that overlay the typical macrocellular base stations in order to increase their networks' capacity [11]. In these emerging heterogeneous cellular networks (HCNs), the users are often in range with multiple base stations (e.g., with a macro-BS and a pico-BS), possibly having different energy consumption profiles. Moreover, the smaller base stations are typically connected to the core network with long-range and low-capacity backhaul links. These links impose additional hard capacity constraints on the amount of traffic the respective base stations can serve, and at the same time consume an important amount of energy [12], [13]. Our goal in this work is to provide an optimization framework that takes video quality and servicing decisions for the mobile users served by an HCN, so as to improve its performance and reduce the overall energy costs.

A. Related Work and Contributions

The problems induced by DASH in cellular networks were recently studied in [14], which moreover proposed a rate allocation framework that improves the user-perceived network performance. References [15] and [16] followed a similar methodology. However, in these works (and in references therein) the authors do not consider the reduction of the energy consumption or any other type of operating cost of the cellular networks. On the contrary, [17] proposed a sophisticated mechanism for centralized (network-driven) decisions regarding the video quality, the transmission power and the servicing airtime for each user. The goal is to increase user satisfaction and reduce energy consumption. However, this analysis applies to typical cellular networks and does not take into consideration the implications of the HCNs architecture, such as the energy consumption costs and the capacity constraints of the backhaul links [18].

There is currently an increasing research interest for resource allocation problems in HCNs, though not in the context of the energy consumption incurred by the serviced requests. For example, [19] studies load balancing in HCNs, and [20], [21] propose joint user association and resource allocation mechanisms. More often than not, the goal is to maximize the aggregate data rate of the served users [22]. Regarding the works that study resource allocation for video delivery in HCNs the works are even more limited. A recent work reported in [23], studies rate allocation for DASH in HCNs that employ time-domain resource partitioning (TDRP), which is the dominant mechanism for minimizing interference in a HCN. This work optimizes TDRP in real-life HCNs in such a way that video quality is maximized. Despite their detailed models and rigorous analytical approaches, these works do not consider the important aspect of energy cost minimization. On the contrary, here we adopt as the main cost metric the total energy consumption that is induced in a typical HCN. This includes the energy consumption of the base stations, and the backhaul link of each base station.

In order to be able to quantify and control this metric, we need to consider a large set of network decisions including the users association, the spectrum, or, resource block (RB), allocation of each base station, and the power transmission in each RB for each user. Finding the optimal decisions is not only computationally challenging (typically involves NP-hard problems), but it also requires the solution of coupled problems in different time scales, often under limited information. For example, base station re-selection (and hence user association) in an HCN requires several seconds, while the RB and power allocation can be realized in milliseconds or whenever there is updated channel state information [24]. Additionally, the assignment of users to base stations has to be made using estimated values for their future expected channel.

To cope with these issues and be able to improve the network operation, we introduce a two-stage optimization framework. In the first stage, the optimization problem determines jointly the base station that each user will be associated with, the qualities of the video segments that will be delivered to each user[2], i.e., for a period of 2-10 seconds, and the total energy (for the base stations and the backhaul links) that will be consumed during this period. These decisions are taken at the network-wide level (e.g., for a set of BSs) and leverage average and expected values for the network and user-related parameters, the exact value of which is unknown at the time of their derivation, i.e., in the beginning of that period.

Accordingly in the second stage, each base station determines the resource block (spectrum) and power assignment to each user that is assigned to it, for each time frame that typically has a duration of tens of milliseconds. These decisions are taken by solving a proper scheduling problem and are updated in each frame[3], based on the feedback about the channel conditions of each user. Clearly, the scheduling decisions across the different frames are coupled with each other as they need to satisfy the respective decisions about the delivered video data (and hence the video quality), and the energy budget that were devised in the first stage.

The power and RB assignment problem is a well known NP-hard problem and several algorithms have been proposed for its heuristic or approximate solution [25]. Any of these well-known methods can be employed for our second-stage optimization problem in each frame, if modified properly so as to satisfy across the entire time period the data delivery and energy consumption overall constraints. On the other hand,

[2]Although this is not the typical operation of DASH, we consider a similar model where the file is divided in segments, and each segment is encoded in different qualities. Yet, the decision of the quality is made by the network.

[3]The framework is directly applicable to scenarios where users transmit their channel quality indicators every 2 or more frames. This time interval is a system-specific design choice, and does not affect our analysis.

by relaxing the discreteness of the delivered video qualities[4] the first-stage (long-period) decision problem can be solved using standard convex optimization techniques. For large set of available video qualities (e.g., as in the case of scalable video coding), this quantization has a relatively small impact.

To this end, the main technical contributions of this work can be summarized as follows:

- We introduce a comprehensive framework for video streaming with DASH in HCNs. Contrary to related work, e.g., [19], [14], [22], [23] this framework employs a detailed power cost model for the entire HCN (including the backhaul infrastructure), supports DASH video quality adaptation, and balances the quality of the delivered video (users' satisfaction) and the operator's cost.
- Resource allocation that corresponds to the actual operation of LTE-A (RB and power). Contrary to related work, [20], [14], [21], [23] the resource allocation problem affected by power consumption constraints set before, in a higher level of the problem.
- We propose a relaxed version of the scheduling optimization problem that allows the very fast derivation of the power and RB assignment solution which, albeit suboptimal, can be employed for real systems that have stringent time constraints.
- A detailed performance evaluation investigation is conducted. We found that a MNO can either decrease total energy consumption or provide the users with increased data rates resulting in improved video quality, by carefully adjusting the parameters of the proposed framework, or by adding more small cells to the topology.

The rest of this paper is organized as follows. Section II describes the basic features of a HCN, gives an insight of LTE's downlink resource grid and backhauling techniques to deliver video files. In Section III we introduce the two-stage optimization framework, discuss the complexity of the respective problems, and propose a solution approach. Section IV provides the performance evaluation of the framework, and Section V presents the conclusions and discusses open directions for future work.

II. Background and System Model

Radio Access Network Model. We consider a macrocell of an HCN, that is overlaid by a set of smaller base stations (SBSs) such as pico-cell BSs, and femto-cell BSs [11]. We denote with $\mathcal{I}$ the set of $I = |\mathcal{I}|$ BSs (including the macro-cellular BS, MBS) that can be overlapping. There exists a set $\mathcal{N}$ of $N = |\mathcal{N}|$ users within the macrocell with video content requests. We denote with $\mathcal{N}_i$ the set of the N_i users that are associated with each BS $i \in \mathcal{I}$. Also, we denote with $\mathcal{I}_n \subseteq \mathcal{I}$ the subset of BSs that are in range with each user $n \in \mathcal{N}$, and hence eligible to serve his request. We study the system for a time period of $T = 10$ seconds. Clearly, the different base stations induce different servicing cost to the network, as they have different energy consumption profiles [8], and involve different backhaul links. Also, the BSs have different servicing capacity as they are constraint by their backhaul links.

[4]For example, in the case of SVC, the combinations of the different (spatial, SNR, temporal) characteristics lead to a large set of qualities, hence reducing this quantization error.

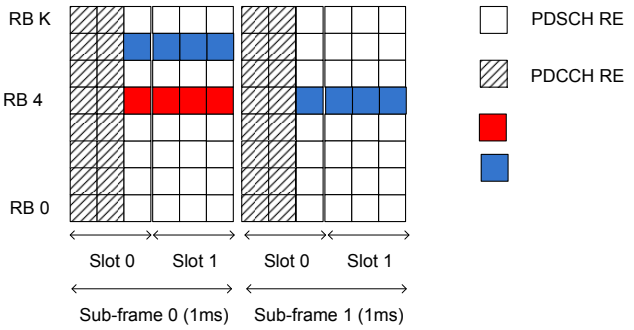

Fig. 2. LTE resource grid. Each small rectangle represents a Resource Element (RE). Each frame comprises 10 subframes, each subframe 2 slots and each slot 7 REs. PDSCH REs are used for users' data transmissions, while PDCCH REs constitute the control channel. RBs (1 slot in time and 12 subcarriers in frequency) are allocated to different UEs.

Backhaul Network. Namely, each BS is connected to the cellular evolved packet core (EPC) through the backhaul network. For the MBS this is a high-capacity direct link. However, the backhaul network for the smaller base stations typically involves more hops. Here, we adopt a tree topology for the backhaul network [12], as illustrated in Fig. 1. Namely, a switch is placed as the first hop of the backhaul path that immediately connects the MBS to the core network and forwards traffic destined to the small-cell users to a second switch that forwards it to the appropriate SBS. Each BS is connected to exactly one interface of the switch, so the rate that each BS can provide to the associated UEs is limited by the rate of that interface. If more SBSs than the total number of interfaces are deployed, a second switch is required to support the backhaul traffic. However, in this work we don't consider the scenario of dynamically increasing or decreasing the number of switches in the backhaul network. Following [12] and references therein, we assume that the basic parameters characterizing the backhaul network are the power consumed by one downlink interface of the aggregation switch denoted as $P_{dl} > 0$ (Watts), the number of downlink interfaces of the aggregation switch $max_{dl} \in \mathcal{Z}^+$, the maximum power consumption of the switch $P_{max,s} > 0$ (Watts), the maximum amount of traffic the switch can handle $C_{max} > 0$ (bps), and the data rate of a single downlink interface $C_{interface} > 0$ (bps). These parameters are considered constant and known.

User Model. Each user requests a video file that has to be delivered through the network that we described in the previous paragraphs. We assume that DASH is used for video delivery. Hence, for each user $n \in \mathcal{N}$, the network has to deliver a certain amount of data during a 10 second period that belong to a single DASH segment. The size of this segment obviously depends on the quality of the delivered

video description. The size of each one of the video quality descriptions for the video of user n is the set:

$$\mathcal{Q}_n = \{Q_{1n}, .., Q_{ln}, ..., \}\,. \tag{1}$$

In the general case, the quality of the delivered video is typically characterized as a convex function of the rate. In our optimization we define this quality as $\gamma_n \log \overline{D_n}$, where $\overline{D_n}$ is the average delivered rate to the user. The parameter γ_n for user n can be seen as the quality gradient of the requested video content, i.e., its quality versus the size in bits [26]. This formulation will allow our optimization to calculate the required $\overline{D_n}$, that must be bounded in practice by (1).

LTE-A PHY. In LTE a frame has a duration of 10ms and consists of 10 subframes (each consisting of 2 slots) as shown in Fig. 2. Consequently, the total number of available frames for the 10 second period is $T = 1000$ subframes. Each BS has a certain set $\mathcal{M}$ of M Resource Blocks (RB) that can allocate to each user during each frame[5]. We assume that each BS has access to the same number of RBs (i.e., we assume equal time/frequency resources are available across the different BSs) [24]. For each one of these RBs the BS can determine the transmission power [27]. We assume that RB allocation and power control decisions are taken in the beginning of a frame. The smallest unit of resource allocation in LTE is a RB that comprises 1 slot in the time domain and 12 OFDM sub-carriers in the frequency domain, while its total bandwidth is usually 180 KHz [28]. Depending on the available frequency band of the system (1.4 to 20 MHz), one can easily calculate the total number of RBs available for the duration of one LTE frame. For example, in a 2.5 MHz system, there exist 12 RBs in 1 slot and 240 in 1 frame. This RB organization is also illustrated in Fig. 2.

Channel Model. The BSs have a single omni-directional antenna that can be used in half-duplex mode for transmission and reception. We denote the average channel coefficient from the i-th BS to the n-th user as $\overline{h_{ni}}$. We assume that the fading coefficients are independent and $\overline{h_{ni}} \sim \mathcal{CN}(0,1)$, i.e. they are complex Gaussian random variables with zero mean and unit variance. All the channels are considered to be block-fading Rayleigh. The channel coefficients are quasi-stationary, that is they remain constant for the coherence period of the channel that is equal to the transmission length of the complete subframe (Fig. 2). We also consider the path loss and shadowing effects according to the LTE channel model [27]. Additive white Gaussian noise (AWGN) is assumed at every receiver with variance σ^2. We assume that users provide channel quality feedback (CQI) to the base stations once every frame [27].

III. Network Decisions and Optimization Framework

As it was explained in Section I, the goal of the operator is to maximize the video quality of the files that are delivered to the users, and, at the same time, minimize the respective servicing cost. These two objectives are often conflicting since increasing the transmission power or the spectrum on the one hand increases the network performance (in terms of video delivery) but may also increase significantly the respective operating expenditures. Clearly, it is important to provide a method for assessing this trade off in a systematic fashion.

A. User Association and Video Quality Selection

First we define the *master* optimization problem that is solved in the first stage. Its solution yields the decisions for the association of each UE to a BS, and the average allocated power from a BS to each UE for the time period T (10 seconds). The last decision also determines the average rate for each user, and therefore the delivered video quality. Hence, the master problem performs the *User Association and Video Quality Selection (UAVQS).*

The variables and constants used in the problem formulation are summarized in Table I. Specifically, the network determines for each user the associated BS and the transmission power that will be used to satisfy his request. We denote with $y_{ni} \in \{0,1\}$ the binary decision of whether user n will be associated with BS $i \in \mathcal{I}_n$, or not. The respective association and power allocation matrices are:

$$\boldsymbol{y} = \left(y_{ni} : n \in \mathcal{N}_i,\, i \in \mathcal{I}_n\right), \tag{2}$$

and,

$$\boldsymbol{P} = \left(P_{ni} \geq 0 : n \in \mathcal{N}_i, i \in \mathcal{I}_n\right). \tag{3}$$

The average rate for user n during the time period T is:

$$\overline{D_n} = \lfloor \frac{M}{N_i} \rfloor W_b \log(1 + \frac{\sum_i \overline{h_{ni}} P_{ni}}{\sigma^2}),\ \forall n \in \mathcal{N} \tag{4}$$

In the above $\lfloor \frac{M}{N_i} \rfloor$ assumes equal allocation of the M RBs to the $N_i = \sum_{n \in \mathcal{N}_i} y_{ni}$ users associated to BS i. Clearly, changing the number of RBs impacts the rate. We will discuss in detail this later. P_{ni} is the average power allocated to user n that is associated to BS i, during the interval T. Inside the logarithm we have the average SNR in order to obtain the ergodic Shannon channel capacity in bps.

Also, the total power allocated to the associated users of a BS i is:

$$P_i^{tx} = \lfloor \frac{M}{N_i} \rfloor \sum_{n \in N_i} P_{ni},\ \forall i \in \mathcal{I}_n\,. \tag{5}$$

This quantity is very important as it determines the energy consumption cost that each BS induces to the operator [8]. Besides, every BS $i \in \mathcal{I}$, based on its type, has a certain maximum transmission power $P_{max,i}$ [24] that must be respected by the power allocation decisions.

Similarly, there are constraints for the total backhaul power consumption. Following the analysis in [13], we assume that the power consumption consists of two components weighted by a variable c. The first one corresponds to the standard operating cost of the switch, while the second is proportional

[5]Without loss of generality, we assume here that each BS has at its disposal the same amount of spectrum.

Parameter	Description
$\overline{D_n}$	The average data rate enjoyed by user n
γ_n	Video quality gradient of user n
$\overline{h_{ni}}$	Average value of the channel between user n and base station i
P_i^{tx}	Radio power budget of base station i
P_i^{bh}	The power consumed at the backhaul switch due to the rate provided to base station i to serve the users associated to it.
$\mathcal{I}$	Set of base stations, macro and picos
I	Total number of base stations, i.e. $\lvert\mathcal{I}\rvert$
$P_{max,i}$	The maximum transmission power of BS i
W_b	Bandwidth of a Resource Block
$\mathcal{N}$	Set of users in the cell
$\mathcal{N}_i$	Set of users associated to BS i
M	The number of resource blocks available to the BS for allocation
y_{ni}	Takes the value of 1 if user n is associated to BS i and 0 otherwise
P_{ni}	The average power of each allocated RB to user n by base station i

TABLE I
MAIN OPTIMIZATION VARIABLES AND SYSTEM PARAMETERS.

to the rate provided to each BS's associated users. This analysis leads to the total power cost of the backhaul:

$$P^{bh} = \lceil \frac{I-1}{max_{dl}} + 1 \rceil c P_{max,s} + \sum_i P_i^{bh} \tag{6}$$

The first term in the above is fixed and does not affect the optimization once a switch has been deployed. At least one switch is required to support the MBS, and $\lceil \frac{I-1}{max_{dl}} \rceil$ for the remaining SBSs. Note that if the number of SBSs exceeds the number of available interfaces of a switch, one more has to be deployed. Hence, in the optimization only the second part is can be controlled since it depends on the traffic load. In [13] this is defined as:

$$P_i^{bh} = (1-c)\frac{\sum_{n\in\mathcal{N}_i}\overline{D_n}}{C_{max}} P_{max,s} + P_{dl},\ \forall i \in \mathcal{I} \tag{7}$$

Formally, the optimization problem that the operator needs to solve every T seconds so as to devise the user association and video quality decisions (*Problem UAVQS*), is:

$$\max_{\boldsymbol{y},\boldsymbol{P}} a\sum_{n=1}^{N}\gamma_n \log \overline{D_n} - b(\sum_{i=1}^{I} P_i^{tx} + P_i^{bh}) \tag{8}$$

$$\text{s.t.}, (4), (5), (7) \tag{9}$$

$$P_i^{tx} \le P_{max,i}\,, i \in \mathcal{I} \tag{10}$$

$$\sum_{i=1}^{I} y_{ni} = 1,\ \forall n \in \mathcal{N} \tag{11}$$

$$\sum_{n\in\mathcal{N}_i} \overline{D_n} \le C_{interface},\ \forall i \in \mathcal{I} \tag{12}$$

$$\overline{D_n}\cdot T \le \max \mathcal{Q} \tag{13}$$

$$P_{ni} \le y_{ni} P_{max,i},\ \forall\, n \in \mathcal{N}, i \in \mathcal{I}_n \tag{14}$$

$$P_{ni} \ge 0, y_{ni} \in \{0,1\}, \forall i \in \mathcal{I}, n \in \mathcal{N}\,. \tag{15}$$

Constraint set (10) ensures that the total transmission power of each BS $i \in \mathcal{I}$ will not exceed its maximum transmission

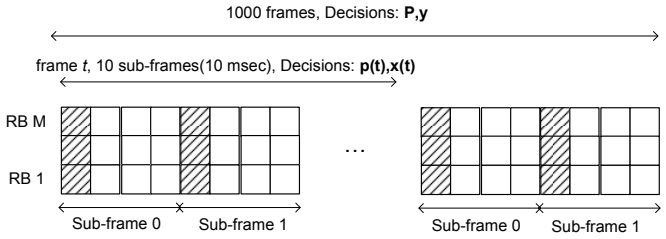

Fig. 3. Time-scale separation of the UAVQS (master problem), and LTERA problem. The former is solved in the beginning of each time period T while the latter in each frame t.

level. Also, the set of (11) ensures that each user will only be assigned to one BS exactly, i.e., there will not be users served by more than one BS, or users not been served by any BS. Eq. (12) imposes the backhaul capacity constraint for al the interfaces of the switches[6]. Also, it is clear that each user should not receive more service rate than it is necessary for the maximum available video quality level. This constraint is imposed by (13). Finally, (14) ensures that power is not allocated to a user that is not associated to a specific BS, and in any case does not exceed the maximum possible value.

The objective function in (8) is a weighted sum of the UE utilities plus the overall network power consumption. Note that both video quality and power can be translated to a financial cost expressed in $. Hence, in the above the balancing parameters a and b allow the MNO to precisely do that by quantifying the relative importance of the two parts of the objective. These parameters are expressed in $/byte and $/Watt, respectively.

The solution to (8) gives us the association decisions for the users, the power budget for each BS during the 10-second period (the vector $\boldsymbol{P}$), as well as the average data rate $\overline{D_n}$ of each user and the DASH video quality. This solution applies for a duration of 10 seconds and is illustrated in Fig. 3.

B. LTE Resource Allocation Problem

Now, we address the LTE resource allocation problem (LTERA) that is solved by each BS in every time frame $t = 1, 2, \ldots, T$ that has duration 10 milliseconds. The particular resources that are allocated to each user are the RBs and the power for each RB. The average power budget for each BS has already been decided by solving the previous problem. Now resources must be allocated to UEs so as to deliver the required data, while respecting the aforementioned solution.

The main motivation behind the formulation we will present next, is the observation that LTE provide CQI feedback from UEs for every frame (7-8 milliseconds). This information is critical since it allows the resource allocation algorithm to use this CQI for allocating optimally the available power and the discrete RBs. Hence, this problem is solved for every LTE frame. Also consistently with the concept of opportunistic

[6]We assume here that the backhaul links are dimensioned based on the capacity of the respective interfaces.

communication when multiple users experience different channel fades, the optimal course of action is to allocate more resources (power) to the user with the best channel. Another aspect of this formulation, is that because the solution provided by the master problem only limits the average power per BS and the average rate per user, we lift the assumption of an equal RB allocation to each user.

Let us now formally define the problem. The LTE resource allocation decisions of the base station for each frame t are (we drop the i notation from the decision vectors): (i) the RB assignment vector:

$$\boldsymbol{x}(t) = \left(x_{mn}(t) \in \{0,1\} : n \in \mathcal{N}_i,\, m \in \mathcal{M},\, t = 1, \ldots, T\right), \tag{16}$$

where $x_{mn}(t) \in \{0,1\}$ denotes whether the RB m is assigned to user n ($x_{mn}(t) = 1$) or not, during frame t, and, (ii) the RB power allocation vector:

$$\boldsymbol{p} = \left(p_{mn}(t) \geq 0 : n \in \mathcal{N}_i, m \in \mathcal{M},\, t = 1, \ldots, T\right), \tag{17}$$

where $p_{mn}(t)$ denotes the transmission power for RB m when allocated to user n, in frame t.

Regarding the total transmission power consumed by BS i, it is the sum of the power that it allocates to its RBs and it must be less or equal to the power budget limit set by the solution of the master problem:

$$\sum_{m \in \mathcal{M}} \sum_{n \in N_i} p_{mn}(t) \leq P_i^{tx} \tag{18}$$

Additionally, the power $p_{mn}(t)$ is constrained by the maximum transmission power of the BS, and is applicable only if the respective RB is indeed allocated to user n, i.e.,

$$0 \leq p_{mn} \leq x_{mn}(t) P_{max,i} \tag{19}$$

The above decisions yield a data rate $r_n(t)$ for user $n \in \mathcal{N}$ in frame t, that can be written as:

$$r_n(t) = \sum_{m \in \mathcal{M}} W_b \log(1 + \frac{\overline{h_n} p_{mn}(t)}{\sigma^2}) \tag{20}$$

In the above equation, we drop the index i from $\overline{h_{ni}}$ since the association of the user to BS is known from the UAVQS problem. Similarly with before, the served rate for all the users of a base station cannot exceed the respective capacity of its backhaul link:

$$\sum_{n \in N_i} r_n(t) \leq C_{interface},\ \forall i \in \mathcal{I} \tag{21}$$

In the above note that we assume that each backhaul link has the same capacity. Clearly this does not have to be the case in general. Also each RB for each BS is allocated only to one user:

$$\sum_{n \in N_i} x_{mn}(t) \leq 1,\ \forall\, m \in \mathcal{M}, \forall\, i \in \mathcal{I} \tag{22}$$

Formally, the RB allocation and power control problem that each BS $i \in \mathcal{I}$ solves, can be written as follows:

$$\max_{\boldsymbol{x}, \boldsymbol{p}} \sum_{n \in \mathcal{N}_i} u_n(t) \log r_n(t) \tag{23}$$
$$\text{s.t. } (18) - (22)$$

In the objective of LTERA we introduced a new parameter that is related to DASH. To explain the parameter $u_n(t)$, recall first that Q_n is the size of the 10-second DASH segment that will be delivered to a user n, which was decided after solving the UAVQS problem. Also let $D_n(t)$ be the total number of bits that have been delivered to user n until LTE frame t. We desire to track the progress of the LTERA solution with respect to the solution of the master problem. Hence, the ratio:

$$\frac{Q_n/10}{(\sum_{\tau=1}^{t} r_n(\tau))/t}, \tag{24}$$

serves as an indication of how the average data rate experienced by user n until frame t, deviates from the average rate $Q_n/10$ that the DASH video segment requires. The goal is to boost users that experience lower average rate compared to average segment rate up to frame t and slow them down in the opposite case. This is consistent with the works such as [16], [17] since it is effectively a way to track the contents of the playback buffer of each user and act accordingly. If a user has already received the entire segment before the end of the 10 second period, then u_n is set to 0 to prevent further allocation of unneeded resources. Formally we can write:

$$u_n(t) = \begin{cases} \frac{Q_n/10}{(\sum_{\tau=1}^{t} r_n(\tau))/t} & \text{if } D_n(t) \leq Q_n \\ 0 & \text{else} \end{cases},\ \forall n \in \mathcal{N}_i \tag{25}$$

Additional data pre-buffering can be allowed by minor enhancements of this parameter. For example we continue to allocate resources even if a complete DASH segment has been received.

C. Solution Approach and Suboptimal Problem Formulation

The above optimization problems clearly constitute non-convex mixed integer non-linear programs (MINLP). It is also known that even simpler instances of this power and RB allocation problems are NP-hard [25]. Hence, our approach for solving this problem is to adopt first a numerical approach. We used Matlab and specifically the Opti Toolbox [29] for solving these problems. More specifically, we used the BONMIN solver, which solves smooth twice differentiable MINLP. This approach can provide very fast solution for the master problem. However, the most challenging problem due to its larger size and the need to solve it within a few milliseconds, is clearly LTERA which is similarly a MINLP. For this problem, we define

1) Suboptimal Problem Formulation: Here, we define a convex alternative to LTERA that allows for its faster execution. Now the transmission power is constant and in particular it has the value that was calculated with the UAVQS problem. Setting a constant power definitely leads to a suboptimal solution, but the use of a continuous variable for RB allocation does not (after relaxing the discrete RB allocation variable). The practical implication is that the implementation of such a system should convert this continuous result into a discrete number of RBs. Consequently, the relaxed LTERA

optimization objective is:

$$\max_{\boldsymbol{x}} \sum_{n} u_n(t) \log r_n(t) \tag{26}$$

subject to: (21), (25) and

$$r_n(t) = x_n(t) M W_b \log(1 + \frac{\overline{h_n} P_n}{\sigma^2}),\ \forall n \in \mathcal{N} \tag{27}$$

$$\sum_{n \in \mathcal{N}_i} x_n(t) = 1 \tag{28}$$

$$0 \leq x_n(t) \leq 1,\ \forall n \in \mathcal{N} \tag{29}$$

The above problem is a convex one and can be solved in sufficiently small amount of time to allow its application in a real world system. The only variable is $x(t)$ and it is a continuous one and P_n is now constant and equal to the value calculated in the UAVQS problem. Thus the problem formulated is convex but also suboptimal.

IV. Performance Evaluation

In this section we present a performance evaluation of our proposed framework through simulations. We evaluate the performance of our framework with respect to the achieved data rate by the users, the average power consumption of the network, and the average delivered video quality. Several picocells and UEs are uniformly spread in an area of 3x3 Km. A single macro BS is located at the center of the area. The available bandwidth is 2.5 MHz [28] which means 240 RBs are distributed among the BSs. The bandwidth W_b of each RB is 180 KHz [28]. We assume a path loss propagation model in the HCN with the maximum transmission power of the BSs chosen according to 3GPP specifications. Different values for the weighting factors of the first problem objective function namely a and b are investigated through our simulation. Regarding the parameters of the backhaul switches they are set equal to P_{dl}=1W, max_{dl}=24, $P_{max,s}$=300 W, C_{max}=24 Gbps, $C_{interface}$=1 Gbps and were adopted from [12], [13]. Concerning video quality decisions, video traces from [30] were used and in particular a 352x288 resolution video segment with 5 different quality layers $Q_1 - Q_5$ with $Q_l > Q_{l-1},\ l \in \{1, ..., 5\}$.

Picocell density. The system performance for different picocell deployment densities are illustrated in Fig. 4. Here we wave set $a = 2$ and $b = 1$. One can see how the users' average data rate relates to the number of users served by the cell. The proposed framework allows the reduction of the data rate provided to the users as their number increases in order to save power. The total number of RBs is the same in case of 4 picocells as is in the case of 5, but we can see a slight increase in the average rate because 1 more picocell can provide more flexibility with respect to the BSs that can serve the users. Also this scenario with 5 BSs, higher spatial-reuse is achieved.

For the same experiment, we present the total power consumption in Fig. 4(b). This power consumption is the sum of the BS radio transmission power, and the power consumption of the backhaul. It increases as the number of users increase since the optimization is effectively trying to provide users with enough data rate while they share the total amount of RBs. A fifth picocell has the potential to lead to significant power savings that reach 16% when the number of users is equal to 12. The reason is that the same data rate that the same group of users require, can be served more efficiently by an additional BS. This fifth BS gives he ability to users to associate with it and enjoy communication of higher spectral efficiency.

Results for the delivered video quality are presented in Fig. 4(c) and are expressed in terms of the highest video quality layer that is delivered to a user. The results correspond to the same experiment. It is important to notice that the presence of 5 picocells increases the fraction of the users that receive high quality video. The value of our framework for a MNO is also evident again here. Our framework allows the precise quantification of different picocell densities on the quality of the delivered video.

From this first set of results we conclude that for constant a, b, and as the number of UEs in increased within a cell, to minimize the increase in the power cost, the delivered rate and video quality has to be reduced. It can be alleviated by introducing more small cells. Hence, our framework can allow the MNO to identify the optimal solution for delivering a certain video quality.

Balancing parameters. Our framework allows the operator to save even more power or offer a better video streaming service to the UEs by adopting different ratios of a/b. In this subsection the effect on data rate and power consumption when using different values for a/b is examined. The deployment in the following simulations is 4 picocells plus 1 macrocell. In Fig. 4(a) the average cell data rate in relation to the number of UEs is illustrated for different values of a/b. As a/b increases, more weight goes to the data rate of the users rather than the operator's power consumption. The impact of the increased data rate as a result of a bigger a/b value on power consumption is illustrated in Fig. 4(b). As UEs within a cell increase, the total power consumption increases in order to provide a certain average data rate to them. But as this level of service increases (a/b increases), the total power consumption increases too. Regarding the video quality we notice in Fig. 4(f) that the video with an increasing ratio a/b, the users receive higher video quality.

The choice of the number of picocells and the values of the balancing parameters define network's performance in terms of minimum QoE provision and maximum power consumption. A MNO can adjust those parameters according to their needs and succeed the desired performance.

Resource allocation and video segment delivery. Finally we evaluate the performance of the relaxed LTERA problem formulation. We consider its performance for the duration of an entire 10-second period. We measure the amount of bits that are received by the users during each one the 1000 frames of the 10-second period as a result of the relaxed LTERA solution of each frame. We also present an "ideal" system (dotted lines)

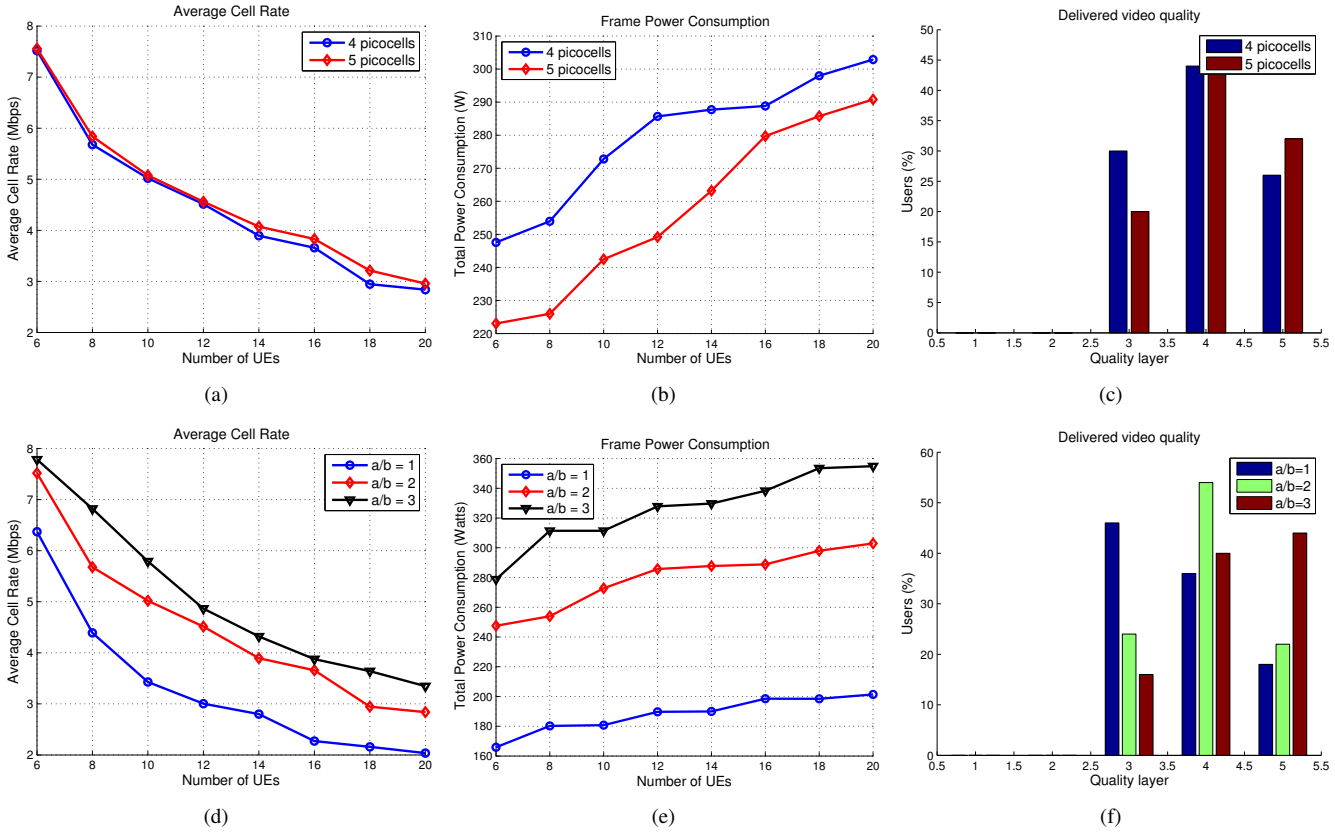

Fig. 4. Performance evaluation results. The effect of picocell density and balancing parameters on the average cell rate, network power consumption and user video quality. In (a)-(c) the ratio a/b is 2. In (d)-(f) The HetNet consists of 4 picocells.

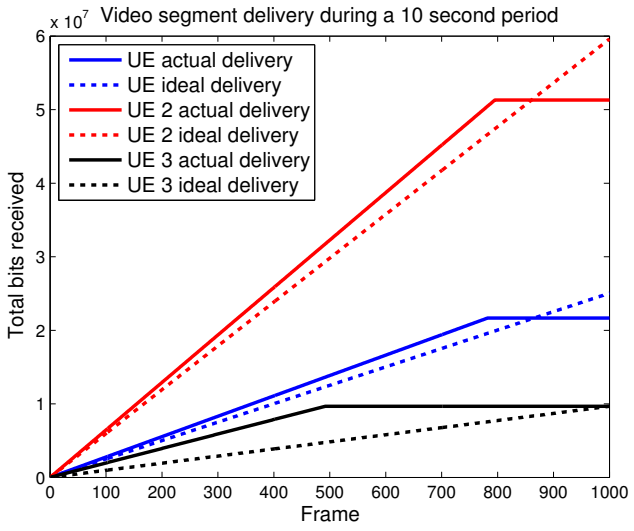

Fig. 5. LTE resource allocation results.

by assuming that the total number of bits is delivery equally among the 1000 frames. Fig. 5 displays the related results. The solid lines obtained with LTERA always stay above the dotted ones, for a certain user/quality, which serves as an indication that no buffer underrun occurs. This is expected, since the power level for each transmission from a user is based on UAVQS and the selected quality requires at most the bit rate D_n. This results in the early delivery of the segments and when they are finally received ($u_n(t)$ is set to 0), no more resources are allocated and thus the solid lines become flat.

V. Conclusions

In this paper, we presented an optimization framework that formalizes the inherent trade-off between the user perceived quality of wireless video, and the energy consumption cost of the network. The former is formulated in the context of the emerging HCNs based on LTE. We also considered users that employ DASH for video delivery. Our framework quantified this trade-off carefully, by delving into the details of DASH, the LTE network, and the HCN architecture. The result is a complex problem that was decomposed into a master problem, and several sub-problems that were solved sequentially. The detailed model presented and analyzed through simulation in this work is to be validated via real world experiments in our future work. Another promising idea is the extension of the model to support dynamically different number of available resources in the concept of the emerging LTE License Assisted Access (LAA). Numerical results indicate that performance improvements for the users can be achieved both by deploying more small cells or by increasing the allowed transmission

power (the servicing cost). Our framework allows the operator to configure the network depending on the desired trade-off between cost and user QoE.

ACKNOWLEDGMENTS

This research has been funded by the EU FP7 Project 318514 Convergence of wireless Optical Network and IT rEsources iN support of cloud services, (CONTENT).

REFERENCES

[1] Ericsson, "Mobility report: On the pulse of the networked society," June 2014.

[2] C. V. N. Index, "Global mobile data traffic forecast update, 2013–2018, cisco white paper, feb. 6, 2013," June 2014.

[3] T. Stockhammer, "Dynamic adaptive streaming over http –: Standards and design principles," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, ser. MMSys '11. New York, NY, USA: ACM, 2011, pp. 133–144. [Online]. Available: http://doi.acm.org/10.1145/1943552.1943572

[4] D. C. H. Nam, B. H. Kim, and H. G. Schulzrinne, "Mobile video is inefficient: A traffic analysis," Columbia University, Technical report, 2013.

[5] A. Mansy, M. Ammar, J. Chandrashekar, and A. Sheth, "Characterizing client behavior of commercial mobile video streaming services," in *Proceedings of Workshop on Mobile Video Delivery*, ser. MoViD'14. New York, NY, USA: ACM, 2013, pp. 8:1–8:6.

[6] S. Akhshabi, A. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http," in *MMSys*, 2011.

[7] S. Akhshabi, L. Anantakrishnan, A. Begen, and C. Dovrolis, "What happens when http adaptive streaming players compete for bandwidth," in *NOSSDAV*, 2012.

[8] O. Arnold, F. Richter, G.Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *ICT MobileSummit*, 2010.

[9] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green cellular networks: A survey, some research issues, and challnges," *IEEE Communication Surveys Tutorials*, vol. 13, no. 4, 2011.

[10] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Toward energy-efficient operation of base stations in cellular wireless networks," *Book chapter in Green Communications: Theoretical Fundamentals, Algorithms, and Applications*, 2012.

[11] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, 2013.

[12] P. Monti, S. Tombaz, L. Wosinska, and J. Zander, "Mobile backhaul in heterogeneous network deployments: Technology options and power consumption," in *Transparent Optical Networks (ICTON), 2012 14th International Conference on*, July 2012, pp. 1–7.

[13] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, Dec 2011, pp. 1–5.

[14] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, ser. MobiCom '13. New York, NY, USA: ACM, 2013, pp. 389–400. [Online]. Available: http://doi.acm.org/10.1145/2500423.2500433

[15] D. Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller, "Optimization of http adaptive streaming over mobile cellular networks," in *IEEE Infocom*, April 2013, pp. 898–997.

[16] V. Joseph and G. de Veciana, "Nova: Qoe-driven optimization of dash-based video delivery in networks," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 82–90.

[17] H. Abou-zeid, H. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, 2014.

[18] S. C. Forum, "Backhaul technologies for small cells," in *White Paper*, February 2013.

[19] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *CoRR*, vol. abs/1205.2833, 2012.

[20] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 1, pp. 248–257, January 2013.

[21] S. Deb, P. Monogioudis, J. Miernik, and J. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *Networking, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 137–150, Feb 2014.

[22] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *IEEE International Conference on Communications (ICC 2012)*, Ottawa, Canada, Jun. 2012, pp. 2485–2489.

[23] A. Argyriou, D. Kosmanos, and L. Tassiulas, "Joint time-domain resource partitioning, rate allocation, and video quality adaptation in heterogeneous cellular networks," *Multimedia, IEEE Transactions on*, vol. 17, no. 5, pp. 736–745, May 2015.

[24] A. Ghosh, J. Zhang, J. G. Andrews, and R.Muhamed, "Fundamentals of lte," *Wiley, Technology Series*, 2010.

[25] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user ofdma systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, 2014.

[26] A. Argyriou, "Compressed video streaming in cooperative wireless networks with interfering transmissions," *Communications, IEEE Transactions on*, vol. 60, no. 11, pp. 3407–3416, November 2012.

[27] "3GPP LTE-Advanced," http://www.3gpp.org/article/lte-advanced, 2010.

[28] J. Zyren, "White paper overview of the 3gpp long term evolution physical layer."

[29] J. Currie and D. I. Wilson, "OPTI: Lowering the Barrier Between Open Source Optimizers and the Industrial MATLAB User," in *Foundations of Computer-Aided Process Operations*, N. Sahinidis and J. Pinto, Eds., Savannah, Georgia, USA, 8–11 January 2012.

[30] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Communications Surveys and Tutorials, in print*, vol. 14, no. 4, pp. 1142–1165, 2012, Traces available at `trace.eas.asu.edu`.