Video-Aware Time-Domain Resource Partitioning in Heterogeneous Cellular Networks

Antonios Argyriou*, Dimitrios Kosmanos*, Leandros Tassiulas†, Yanwei Liu‡, Song Ci.§

*Department of Electrical and Computer Engineering, University of Thessaly, Volos, 38221, Greece.
†Yale Institute for Network Science, Yale University, New Haven, 06511 CT, USA.
‡High Performance Network Lab, Chinese Academy of Sciences, Beijing 100190, China.
§University of Nebraska-Lincoln, Omaha 68046, USA.

Abstract—Heterogenous cellular networks (HCN) consist of macrocells and small cells that are overlaid in the same geographical area. Hence, critical is that the high power macrocell shuts off its transmissions for a fraction of the time to allow the low power small cells to transmit without interference. This is the time-domain resource partitioning (TDRP) mechanism. In this paper we investigate video communication in HCNs when TDRP is employed. More specifically we consider the problem of maximizing the average video quality of all users, by jointly optimizing the rate allocated to each specific video stream and the quality that it is streamed. The resulting mixed integer linear program (MILP) formulation is solved numerically. Simulation results indicate clearly that as the small cells and the users are increased the proposed system can improve significantly the video quality.

Index Terms—Heterogeneous cellular networks, small cells, intra-cell interference, video streaming, video distribution, rate allocation, resource allocation.

I. INTRODUCTION

Wireless bandwidth has always been a valuable resource. The lack of wireless bandwidth in cellular systems today is addressed with the very promising solution of heterogenous cellular networks (HCNs) that require the deployment of low power small cell base stations (SCBSs) [1]. HCNs are primarily deployed to address the explosive demand for high quality video in mobile devices [2], [3]. Thus, the reality that the mobile network operators (MNOs) have to face is that small cells are responsible for transmitting high volumes of video data (and will be for the foreseeable future as indicated in [2], [3]). To respond to this huge increase of video content delivered through HCNs, the MNOs have to seek ways to optimize the utilization of the HCN resources. One way to further optimize the use of resources in such a video delivery system is through caching [4], [5]. However, caching is the first step and only minimizes data transfers through the backhaul link. Once the video is cached and is available at a SCBS, there are still steps that need to be taken especially when the video is delivered through streaming to the user. HCNs do not simply consist of the deployment of SCBSs and their associated backhaul links. There are intricate algorithmic details that are unique to HCNs that need further exploration.

In this paper we depart from the study of the architectural implications of HCNs on video delivery, and we delve into the new physical layer (PHY) transmission schemes. The reason we must delve more into the details of HCNs is that they bring with them new challenges and eventually new algorithms and protocols. One of the key problems unique to HCNs is that of intra-cell or cross-tier interference. Intra-cell interference is caused between the macro base station (MBS) and the low power base stations (BS). Fig. 1 illustrates this case where the MBS interferes with the picocell BS (PBS) and femto BS (FBS) transmissions. It is evident even from this simple example that intra-cell interference must be minimized. One strategy for handling this type of interference in HCNs, is time domain resource partitioning (TDRP) according to which the MBS shuts off its transmission for a subset of the available resources. This technique seems very promising and it was recently standardized through the introduction of almost blank subframes (ABS) in 3GPP LTE under the more general scheme of enhanced ICIC (eICIC) [6]. With TDRP users associated to the picocells and femtocells can achieve higher data rates in the ABSs since interference from the MBS is limited to the bare minimum [6]. Nevertheless, time-domain resources must still be allocated to the macrocell to ensure umbrella coverage for the complete network.

Interference coordination and TDRP with ABS for HCNs is a topic investigated only recently because ABS was also very recently standardized in 3GPP. Deb et al. in [7] derived the optimal fraction from the available ABSs and regular subframes, that each picocell should use. The authors assumed
a constant fraction of ABS $\eta$ that are configured by the HCN. In a more recent work by Sing and Andrews [8], the authors investigated the joint optimization of time-domain resource partitioning expressed through the parameter $\eta$, together with user association (for traffic offloading). However, the authors consider an equal rate allocation to the associated users. There are also numerous studies for video delivery optimization and resource allocation in different types of networks but not for the TDRP-based systems that we focus in this paper. The authors in [9] investigated scheduling and resource allocation for a downlink LTE cellular system that employs discrete decisions for the video streaming rate. The same topic, but for scalable encoded video, was considered in [10], and for cooperative cellular networks in [11]. To the best of our knowledge there is no work that addresses TDRP in the context of wireless video distribution and streaming for HCNs.

The system we consider can be described more effectively with the help of Fig. 1 and Fig. 2. Fig. 1 indicates the users and how they are associated to each BS, while the fraction of the rate allocated to these users is depicted in Fig. 2. During the fraction of the ABS resources $\eta$, the transmission rate is different when compared to the RS that occupy a fraction $1 - \eta$ of the total resources. In this example user 6 can only be allocated resources during the RS. For this system we investigate the optimal allocation of the available rate to the associated users and the optimal video quality for each user given a TDRP $\eta$ that the HCN operator has configured.

II. SYSTEM MODEL AND ASSUMPTIONS

In Fig. 1 we present the network topology that we study in this paper and it includes a single macrocell with a MBS, the PBSs, and the users. Each base station $j$ in the set $\mathcal{J}$ communicates with the set of users $\mathcal{N}_j$. The MBS shuts off its transmissions for a fraction of the resources that is denoted with $\eta$. During these resources all the picocells transmit and interfere with every active user in the network. Thus, we consider resource reuse across BSs of the same tier (PBSs in our case) which is one of the main benefits of small cells since it allows spatial reuse. The aggregate average interference power that a user $i$ receives is denoted as $I_{\text{ABS},i}$. During the fraction of the non-blanked resources, or regular subframes $1 - \eta$, both the MBS and PBSs transmit and the aggregate interference power that user $i$ receives is denoted as $I_{\text{RS},i}$. The time period that contains ABS and regular subframes is called the ABS period [6].

We consider the unicast streaming of a set of pre-compressed and packetized video streams from each BS $j$ to its associated users in the set $\mathcal{N}_j$. The users associate to the proper BS by using an signal-to-interference plus noise ratio (SINR) biasing rule [8]. This means that a user is associated to the small cell $j$, and not the MBS, if the following is true:

$$SNR_{\text{PBS},j} + \text{Bias} \geq SNR_{\text{MBS}}$$

This ensures that user are offloaded to the small cells. If this is not enforced, then users will associate to the MBS that offers higher SINR and so the benefit of using small cells cannot be observed. This is a critical parameter for HCNs studied in [8].

Regarding the PHY details, every node has a single omnidirectional antenna that can be used in half-duplex mode for transmission and reception. The transmission power that the PBS and MBS use is $P_{\text{PBS}}$, and $P_{\text{MBS}}$ respectively. We denote the channel from the $j$-th BS to the $i$-th user as $h_{j,i}$. We assume that the fading coefficients are independent and $h_{j,i} \sim \mathcal{CN}(0,1)$, 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 PHY block. We also consider the path loss and shadowing effects according to the LTE channel model [6]. Additive white Gaussian noise (AWGN) is assumed at every receiver with variance $\sigma^2$. A modulation and coding scheme (MCS) with $m$ bits/symbol is used by each BS while its value is determined by each PBS independently and optimally as we will later explain. The set of available MCSs is $\mathcal{MCS} = \{1, ..., 7\}$, i.e., we assume that the most spectral efficient MCS is 128-QAM. We also assume that users provide only average channel quality feedback (CQI) to the base stations.

Now each user $i$ may either request a fixed video quality or the best possible video quality. Without loosing generality we assume that all the PBSs are assumed to have cached the video file for all the users [4], [5].\(^1\) Now if BS $j$ transmits to user $i$ the $r$-th video description, the average bitrate that must be sustained is

$$R_{ir} = \frac{S_{ir}}{T_i + B_{ir}} \text{ bits/sec},$$

where $S_{ir}$ is the size of the $r$-th video description, $T_i$ is the total playback time of the file and $B_{ir}$ is the buffering delay.

We also assume static conditions for the users, i.e., we do not consider user churn. The reason is that we are interested to optimize the system operation within several minutes (the duration of the video viewing). Typically the user population per small cell and the content they receive varies slowly throughout the day as shown in [12].

A. Utility Function

In this paper we define a utility modeling framework that can support compressed non-scalable and scalable video. In particular for non-scalable video we assume that we have available the rate-distortion (RD) information associated with video frame $n$ denoted as $\Delta D(n)$ which is the total decrease

\(^1\)Our system can easily accommodate the case that the video stream originates from a server at cost of higher delay.
in the mean square error (MSE) distortion that will affect the video stream if the frame is delivered to the client by its prescribed deadline [13]. These values can be obtained fairly easily during the encoding of the video sequence. Hence, the utility function for the video of user $i$ when it receives the video quality description indexed by $r$, is defined as the reduction of the reconstruction distortion of the video:

$$ U_{ir} = \frac{\sum \Delta D(n)}{S_{ir}} \text{ MSE utility/bit} \quad (2) $$

The final result of the previous discussion is that a single video sequence for user $i$ will be available at the following discrete set of utilities:

$$ \mathcal{U}_i = \{U_{i1}, \ldots, U_{ir}, \ldots\} \text{ with } r \in \mathcal{R}_i \quad (3) $$

### B. Throughput with Adaptive MCS

We consider that the BSs optimize independently PHY parameters of the point-to-point links, as it is typically done in every communication system [14]. To estimate the average communication rate that each user $i$ achieves when it is associated to small cell $j$ we proceed as follows. During the ABS, a user receives the aggregate interference from all the simultaneously transmitting PBSs since there is resource reuse as we explained. Thus, the average SINR between the PBS and user $i$ is:

$$ E[\gamma_{i\text{ABS}}^2] = \frac{P_{\text{PBS}} E[|h_{\text{PBS},i}|^2]}{I_{\text{ABS},i} + \sigma^2} \quad (4) $$

Now, during the regular subframes, each user associated to a small cell has to suffer higher interference because the MBS is also active and so the SINR of user $i$ is:

$$ E[\gamma_{i\text{RS}}^2] = \frac{P_{\text{MBS}} E[|h_{\text{MBS},i}|^2]}{I_{\text{RS},i} + \sigma^2} \quad (5) $$

Also for the users associated to the MBS it will be:

$$ E[\gamma_{i\text{RS}}^2] = \frac{P_{\text{MBS}} E[|h_{\text{MBS},i}|^2]}{I_{\text{RS},i} + \sigma^2} \quad (6) $$

The average SINR expressions allow us to calculate the resulting average data rate for user $i$ associated to PBS $j$ under MCS $m$ as:

$$ C_{jim} = m \cdot \text{eff} \cdot S \cdot (1 - P_s)^{L/m} \text{ bits/sec}, \quad (7) $$

where $S$ is the symbol rate, eff is the efficiency of the MCS, and the probability of symbol error $P_s$ under $2^m$-QAM is [15]:

$$ P_s = 4(1 - 2^{-m/2})Q\left(\sqrt{\frac{3}{2^m - 1}E[\gamma_i]}\right) \quad (8) $$

In our system the PHY and link-layer system at each BS selects optimally for the average SINR the MCS that ensures the highest point-to-point rate [16]. Formally this is:

$$ C_{ji} = \max_{m \in \text{MCS}} C_{jim} $$

### III. Problem Formulation and Solution

Now we are ready to describe quantitatively and also define formally the problem we address in this paper. For each user $i$ associated to BS $j$ the MNO that manages the HCN must select the highest quality video description $r$, and the rate allocated to it. This is the Rate Allocation and Video Quality Selection (RAVQS) problem. Let $x_{jir}^{\text{ABS}}, x_{jir}^{\text{RS}} \in \{0, 1\}$ indicate that user $i$ that is associated to BS $j$, is served with video description $r$ in an ABS and RS respectively. Let also $x_{jir}^{\text{ABS}} \in \{0, 1\}$ denote the fraction of the ABS resources that the PBS $j$ allocates to user $i$ for streaming the $r$-th video description. Similarly for the RS, i.e., $x_{jir}^{\text{RS}} \in \{0, 1\}$. Hence, the decisions of each base station $j$ are: (a) the video quality selection vector for all the associated users, i.e., $x_{jir}^{\text{ABS}} \in \{0, 1\}$ (b) the resource allocation vector for all users $x_{jir}^{\text{ABS}} \in \{0, 1\}$ and similarly for the resource and video quality selection vectors for the regular slots. To minimize the notation later in our solution, we also define different concatenations of the variable vectors as follows: $z_j = (x_{jir}^{\text{ABS}})_{r \in \mathcal{R}_i}, j \in \mathcal{J}, i \in \mathcal{N}_j$. Formally, the objective is:

$$ \sum_{j \in \mathcal{J}\setminus\{0\}} \sum_{i \in \mathcal{N}_j} \sum_{r \in \mathcal{R}_i} (x_{jir}^{\text{ABS}} + x_{jir}^{\text{RS}}) U_{ir} + \sum_{i \in \mathcal{N}_0} \sum_{r \in \mathcal{R}_i} x_{jir}^{\text{RS}} U_{ir} \quad (7) $$

In the above recall that $U_{ir}$ is the average utility/bit of the video flow transmitted to user $i$ and at $r$-th quality level. Also, $j=0$ indexes the MBS that cannot transmit during an ABS. Thus, the objective in (7) is expressed in terms of video quality delivered to the complete HCN. For the first set of constraints we have to recall that the total fraction of the blank ABS resources that are available at the PBSs (there is resource reuse across the PBSs) is $\eta$. This leads to:

$$ \sum_{i \in \mathcal{N}_j} \sum_{r \in \mathcal{R}_i} x_{jir}^{\text{ABS}} \leq \eta, \forall j \in \mathcal{J}\setminus\{0\} \quad (8) $$

$$ \sum_{i \in \mathcal{N}_j} \sum_{r \in \mathcal{R}_i} x_{jir}^{\text{RS}} \leq 1 - \eta, \forall j \in \mathcal{J} \quad (9) $$

When a particular description $r$ is selected, the average bitrate in bits/sec (see (1)) that must be sustained by a user $i$, if it is streaming the description $r$, is less than the rate that can be achieved during both the ABS and RS. Also the resources allocated during ABS and RS will determine the average rate. The above can be formally written as:

$$ x_{jir}^{\text{ABS}} R_{jir} \leq x_{jir}^{\text{ABS}} C_{jir}^{\text{ABS}} + x_{jir}^{\text{RS}} C_{jir}^{\text{RS}}, \forall r \in \mathcal{R}_i, i \in \mathcal{N}_j, j \in \mathcal{J} \quad (10) $$

We also have the constraint that resources cannot be allocated to a video description $r$ if it is not actually selected:

$$ z_{jir}^{\text{ABS}} \leq x_{jir}^{\text{ABS}}, \forall r \in \mathcal{R}_i, \forall i \in \mathcal{N}_j, j \in \mathcal{J}\setminus\{0\} \quad (11) $$

$$ z_{jir}^{\text{RS}} \leq x_{jir}^{\text{RS}}, \forall r \in \mathcal{R}_i, \forall i \in \mathcal{N}_j, j \in \mathcal{J} \quad (12) $$

We also need the integer constraints according to which only one video description $r$ can be used for each user. Thus:

$$ \sum_{r \in \mathcal{R}_i} x_{jir}^{\text{ABS}} \leq 1, \forall i \in \mathcal{N}_j, j \in \mathcal{J}\setminus\{0\} \quad (13) $$

$$ \sum_{r \in \mathcal{R}_i} x_{jir}^{\text{RS}} \leq 1, \forall i \in \mathcal{N}_j, j \in \mathcal{J} \quad (14) $$
During the regular slots the PBSs can also transmit together with the MBS, albeit with lower spectral efficiency. In this case the rate will be lower. However, we must ensure that across ABS and RS the same video description is streamed:

\[ x_{jir}^{ABS} = x_{jir}^{RS}, \forall r \in R_i, \forall i \in N_j, j \in J\backslash\{0\} \] (15)

All the previous discussion leads to the definition of the optimization problem we solve in this paper:

\[
\max_{x,r} \sum_{j \in J(0)} \sum_{i \in N_j} (x_{jir}^{ABS} + x_{jir}^{RS})U_{ir} + \sum_{i \in N_0} \sum_{r \in R_i} x_{jir}^{RS}U_{ir}
\]

subject to (8) – (15)

A. Solution

The problem we defined in the previous subsection, together with the associated constraints constitutes a mixed integer linear program (MILP). The MILP formulation can be solved with polynomial algorithms through LP relaxation techniques. For solving the MILP formulation we will present numerical results in the next section obtained with CPLEX Optimization Studio V12.5.0. Note that in this paper we evaluate the initial potential of our approach on system-level through numerical evaluation for constant user population and infrastructure deployment. Our future work will be focused on the development of a practical protocol that extracts this gain. Regarding the parameters of the MILP formulation that correspond to the channels between the BSs and the users, they can be available in the form of CQI measurements. This information is needed by the a central controller (CC) in order to solve the MILP in a centralized fashion. The variances of the AWGN can similarly be available at the CC since they are device specific. Also note that the problem can be solved offline since it only requires average statistics.

IV. Performance Evaluation

In this section, we examine the performance of our framework through numerical evaluation. The parameter settings are as follows. Downlink MBS and PBS transmit power are equal to 46dBm and 30dBm respectively [8]. Distance-dependent path loss is given by \( L(d) = 128.1 + 37.6 \log_{10}(d) \), where \( d \) is the distance in Kms [6], and the shadowing standard deviation is 8 dB. The user speed is 3 kmp/h (quasi-static as we already stated). The macrocell area is set to be a circle with radius equal to 1 Km. The wireless channel parameters include a channel bandwidth of \( W=20 \) MHz, noise power spectral density of \( \sigma^2=10^{-6} \) Watt/Hz. The traffic model is that of an infinite full buffer for every user (i.e. available video content for all of them). The user distribution and picocell locations are random and uniform within the complete macrocell. In our experiments we set the biasing threshold to 0 dB for all the systems in order to obtain \( N_j \), and then run our optimization. Our tested systems include the proposed scheme RAVQS, a system that performs video/utility-aware rate allocation (RA) but for a fixed randomly chosen video quality for each user, and a system that is oblivious to the video aware rate allocation but instead it uses the currently state-of-the-art approach for proportional fair allocation (PFRA). The presented results are averaged over 100 randomly generated topologies. Also for the given user population in each simulation run we assume that 75% of the associated users receive video, based on recent traffic reports [3]. The video content used in the experiments consists of 26 CIF (352x288), and high definition (1920x1080) sequences that were encoded with SVC H.264 [17] as a single layer [18]. The videos are compressed at 30 fps and different rates ranging from 128 Kbps and reaching values<7 Mbps. The frame-type patterns were G16B1,G16B3,G16B7,G16B15, i.e., there are different numbers of B frames between every two P frames and a GOP size is always equal to 16 frames. In all the figures we present the average video quality delivered (our problem objective) to the picocell users versus the average quality in the macrocell associated users (only from the MBS to its associated users) for different constant values of \( \eta \). Regarding the presentation of the results we show the average video quality (in terms of the representation \( r \)) that is delivered.
to the picocell and macrocell users. For example one data point that has the value 3.2 indicates that on average the users received the quality representation 3.2. Hence, higher values indicate that the clients received on average higher video quality representations. The data points in these figures correspond to different values of \( \eta \).

The results for all systems can be seen in Fig. 3. RAVQS is superior when compared to PFRA and RA for high user density and low PBS density in Fig. 3(a). As the number of the PBSs is increased in Fig. 3(b), all the systems can achieve higher performance. The reason is that fewer users are associated to each picocell and so a higher rate is available for each user under any rate allocation scheme. So more picocells lead as expected to better results due to the higher available rate per user. Another important result is that for constant PBS density we have better system performance as the user population grows. As the user density increases this leads to higher importance for the rate allocation algorithm, since the rate of a single PBS is shared across more users. Also note that in the left part of the \( x \) axis, where all the resources are essentially allocated to the picocells, we have the maximum possible utility performance in the network (\( \eta \approx 1 \)). In this regime, the performance gap between RAVQS and the other scheme is increased as the number of picocells and users is increased.

V. CONCLUSIONS

In this paper we presented a framework for improving the quality of video streaming in a HCN that employs TDRP. TDRP is essential for the efficient operation of HCNs and when high quality video distribution enters the game, efficiency becomes even more critical. Our framework focused on this challenge, i.e., it ensures optimal and video-aware rate allocation and quality adaptation in HCNs that apply TDRP. We formulated this problem as MILP that was solved numerically. Our results indicate the benefits of our system for increasing user and picocell densities. Our future work will be focused on the design of a distributed solution algorithm, and the co-design of our system with practical streaming protocols like dynamic adaptive streaming over HTTP (DASH). Also evaluating other system parameters of HCNs is a potential avenue for investigation.

REFERENCES