# Mobile HTTP-based streaming using flexible LTE base station control

Izhak Rubin*, Stefania Colonnese†, Francesca Cuomo†, Federica Calanca† and Tommaso Melodia‡
*Electrical Engineering Dept., UCLA, Los Angeles, CA, USA
Email: rubin@ee.ucla.edu
†DIET, University of Rome Sapienza, Italy
Email: (francesca.cuomo, stefania.colonnese)@uniroma1.it
‡Dept. of Electrical and Computer Engineering Northeastern University, Boston, MA, USA
Email: melodia@ece.neu.edu

***Abstract*—This paper investigates the advantage of adopting a flexible resource control scheme when performing HTTP-based adaptive streaming across LTE systems. To guarantee video fluidity, mobile video streaming is known to require a large bandwidth overhead with respect to the net encoded video rate. The quality of a received video stream is impacted by variations in the size of the transmitted video packets (chunks), and by statistical fluctuations in the data rate at which the allocated downstream wireless channel operates. First, in considering an illustrative video scenario, we show that the chunk size distribution is heavy-tailed, and is well fit by a Gamma distribution. Second, we employ a HAS based proxy video manager and resource controller at the base station node. Based on the channel quality observed and reported by a mobile client, the manager selects the proper channel bandwidth and data rate levels at which to transmit the stream's chunks, in accordance with the selected encoded video rate and the configured Quality of Experience (QoE) level at which the user is targeted to receive the video stream.The communications data rate is also set to assure an acceptable low video reception stall probability. To illustrate the performance of such a dynamic bandwidth allocation scheme, we compare it with an operation that employs a stationary setting of the channel bandwidth, and we compute the gain achieved when such adaptations are performed at the base station node on a chunk by chunk basis. We show by analysis, and confirm by simulations, the improvements achieved in the system's performance behavior through the use of the adaptive resource allocation scheme.**

## I. Introduction

Video streaming applications are becoming very popular and there is an increased interest in providing them over 3G and 4G wireless platforms. Long Term Evolution (LTE) [1] systems, designed in accordance with advanced standard (LTE-A) structures and subsequent evolutions, will provide video streaming at high data rates and are expected to yield high throughput rates by using adaptations of the radio transmission process to variations embedded in the video traffic flows while also accounting for fluctuations in channel gain states [2].

The most popular streaming architecture, known as HTTP Adaptive Streaming (HAS), underlies the provision of several commercial streaming video services embedded with currently implemented video networking systems. The basic principle guiding the operation of HAS is based on storing, at the server side, multiple version of a video source, each encoded for reception at its targeted video quality level, and accordingly represented by its specific flow bit rate. Each encoded bit-stream is then parsed into video packets, usually referred to as "chunks", which encompass one or more GOPs (Group of Pictures), and are addressed by means of URLs available to the client through HTTP servers [3]. Every chunk $k$ represents a segment of the video that lasts for few seconds. The client, on the basis of the network conditions (i.e. the realized channel throughput rate) and/or the status of its playout buffer, sends requests of subsequent chunks to the server so as to receive them at throughput rate high enough (and at with sufficiently short temporal spacings) to avoid buffer starvation and video play stall events at the receiver.

In HAS, the video chunks can be selected from the video bitstream whose average rate best matches the current network bandwidth conditions. The actual video streaming rate is thus adapted to the end-to-end channel throughput rate experienced between the sender and the client [4]; nonetheless, it must be noticed that adaptation may come with a price of sudden video quality variations, that may affect the perceived quality [5].

A detailed presentation of video streaming issues that characterize (3G and 4G) wireless systems appears in [6]. The authors present both an analysis of Over-The-Top (e.g., You Tube, NetFlix) video content delivery and its impact on mobile networks. They also overview certain methods that may be employed to improve the performance of the video streaming delivery process. As for the first aspect, they observe that there are several key causes leading to video packet losses in the provision of HTTP-based mobile video streaming services. They identify methods that can be used to save the LTE downlink bandwidth levels.

To meet the desired Quality of Experience (QoE) level of the video stream received by a targeted mobile user, system designers often set the bandwidth of the downlink communications channel to yield a throughput data rate that is of the order of twice the average value of the video stream produced at the application layer at the configured compression level corresponding to the targeted QoE value [7] [8]. Such a setting

978-1-4799-8461-9/15/$31.00 2015 IEEE

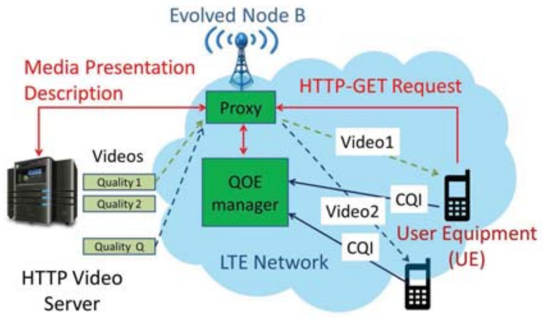


Fig. 1. Reference architecture

is typically based on a conservative estimate metric. As such, it may give rise to inefficient resource allocations. A related study involving QoE-capacity control tradeoffs is presented in [9]. Furthermore, apart the bandwidth allocation, a significant role in the QoE is played also by how the quality is varied. The paper [10] addresses the relationship between the bandwidth allocation and the quality selection and stresses the fact that users dislike significant quality drops. To this aim they propose DASH-qoe, a QoE-aware quality adaptation algorithm which basically insert intermediate quality levels during the quality switching.

In this framework we aim at designing a bandwidth allocation mechanism able to:

1) adapt the bandwidth allocated to a single user to the intrinsical fluctuations of the video data rate, while maintaining the user QoE;
2) adapt the bandwidth allocated to a single user to the channel quality fluctuations, by keeping user QoE;
3) efficiently accommodate multiple users with improved performance, in terms of bandwidth saving, with respect to the conventional HAS approach.

The considered architectural model is depicted in Figure 1. We assume that the Evolved Node B, denoted as Base Station (BS) in the following, acts as a proxy between the client and the server. This may be realized to different extents, either by partial pre-caching video data in the local server's proxy data base, which is stored at the BS node's facility, or by loading side information only, as the average rates of the video streams and the sizes of the next chunks to be transmitted.

As for the first objective, we will model and analyze an adaptive streaming operation at the BS site which, by acting as a video proxy, dynamically adapts the bandwidth allocated to a single user based on the chunk video rate and the channel quality fluctuations. Secondly, by collecting measures on the channel quality, the considered BS control mechanism is able to keep as constant as possible the video quality selected by the client application layer in order to avoid fluctuations of the perceived QoE.

As for the latter point, it is noted that in conventional HAS the QoE control is managed solely by the client. According to their reception quality, certain clients may require and consequently be allocated excess resources, as for being unintentionally recognized as privileged ones. To this aim, the use of a centralized controller has been proposed in the context of a HAS based operation, where QoE-aware resource managers are utilized to achieve a more fair resource allocations.

We notice that the use of proxy based adaptation mechanisms for video streaming on wireless systems have been also proposed in recent papers (like [11]) where the proxy helps in scheduling in an efficient and central way the access to the radio resources.

## II. Flexible Resource allocation for mobile streaming in LTE

We consider a system of $N$ User Equipments (UEs) served by a BS controller. The BS performs flexible resource allocation to efficiently provide the users with a constant QoE streaming services. To this aim, the BS implements the following stages: i) user profiling based on the user generated channel quality reports, denoted as $c$; ii) identification of the target quality level $q$ that will be provided to the user, and matching with the appropriate average video rate $R(q)$ in the set of video stored at the HAS-server side; iii) computation of the bandwidth $BW(k; q, c)$ to be allocated to the user for transmission of the $k$-th chunk to guarantee the aforementioned video quality $q$ having a channel quality $c$.

### A. Channel-quality based user profiling

The quality of signal reception at mobile stations residing in a given cell randomly fluctuates due to interference and mobility, and it is compactly described by the recorded Signal to Interference and Noise Ratio (SINR) levels. At a given instant of time, the SINR level $\sigma$ [dB] can vary over a range of values, as characterized by a probability density function $p_\Sigma(\sigma)$; practical values typically span the range 0 - 15 dB.

The BS is assumed to periodically attain and record the channel quality incurred at each client mobile, using the LTE Channel Quality Indicator (CQI) parameter signaled by each client based on the average measured SINR level. The CQI parameter, denoted as $c$, is an integer value signaling a SINR interval range, which is a member of a set of predefined intervals [12][1]. It spans the set $\{0, \ldots 15\}$. Accordingly, over a sliding period of time, the BS collects corresponding data and calculates a mass probability distribution function for the CQI value recorded for cell users.

Accordingly, the probability that a specific user requesting a streaming service over the cell will send a CQI level $c$ based on the experienced level of SINR is:

$$p_{CQI}[c] = \int_{I_c} p_\Sigma(\sigma) d\sigma, \quad c = c_{min} \ldots c_{max} \tag{1}$$

where $I_c$ is the SINR range corresponding to the $c$-th CQI level. Thereby, the channel quality estimated by the user due to mobility, fading, interference and low level phenomena is compactly summarized by the reported CQI index $c$.

### B. Identification of the target quality level $q$

The QoE value associated with the reception of a stream is related to the video stream's encoding rate and to the targeted maximum stall probability value, $\mathcal{P}_{stall}$ during the playout; in fact, since HAS relies on TCP, delay originated stalls are the only degradation introduced in the played video. From now on, we assume that the bandwidth, and henceforth the transmission rate, selected at the BS is sufficiently high to assure reception at the intended users in a manner that avoids stalls during the playout. With these settings, the target QoE is related to the video encoding average rate. Let us remark that the average rates $R_q,\ q = 1, \ldots Q$ actually available at the server site belong to a finite set that is signaled at the beginning of a session. The target QoE level $q$ belongs to the set $\{1, \ldots Q\}$, and it is associated to the actual video streams having net rate $R(q)$ [bps], as properly signaled at the beginning of the session, by the DASH manifest file, called the Media Presentation Description (MDP).

The BS node determines the QoE level (i.e. the rate of the stream ) to be provided to a mobile user that experience a certain CQI profile, as characterized by $p_{CQI}[c]$.

Let us notice that in conventional HAS the choice of the video quality is left to the client, that adaptively identifies the desired video quality level, among those provided by the server, based on the MDP that is sent at the beginning of the session; the client selection is directly related to the perceived throughput. On the contrary, in the proposed approach, the BS first identifies the quality level best suited to the channel quality user profile, and then drives the client by maintaining the throughput corresponding to the assigned quality level.

In order to select the appropriate QoE level for each user, the BS relies on pre-computed tables relating the average video rates $R(q),\ q = 1, \ldots Q$ of the bitstream candidate for transmission, with the occupied bandwidth, which in turns depends on the mass probability function $p_{CQI}[c]$ characterizing the user CQI profile. By doing this, the BS manages small channel quality fluctuations, which do not cause a quality switching at the application layer. On the other hand, the BS, by periodically updating the channel quality level profile, can drive client switching after larger channel quality fluctuations, e.g. due to user mobility.

Finally, to associate the CQI with a quality $q$, the BS can take into account small CQI fluctuations as well as the user willing-to-pay profile; the policy with which QoE and CQI are associated, $q = q(c)$, can be compactly characterized in terms of the joint probability $p_{CQI,QoE}(c, q)$, representing the probability that a user experiencing a SINR level corresponding to a CQI $c$ is provided a stream at QoE level that is equal to $q$.

Based on the LTE settings, when transmitting messages to a mobile user, reported CQI state univocally determines the underlying modulation/coding scheme (MCS) that is employed. The spectral efficiency, defined as the channel coding rate times the bits per symbol of the adopted modulation, varies accordingly. We denote the spectral efficiency level setting that corresponds to CQI $c$ as $\gamma(c)$ [bps/Hz].

The BS allocates to the user part of the totally available (net) LTE bandwidth, denoted as $BW_{LTE}$ [MHz]. We consider the Release-8 gross bandwidth set: $[1.4\ MHz, 3\ MHz, 5\ MHz, 10\ MHz, 15\ MHz, 20\ MHz]$, divided in so called LTE resource blocks (RB) occupying a bandwidth $BW_{RB} = 180\ kHz$ [13]; this corresponds to the net bandwidth $BW_{LTE}$ belonging to the set: $[1.08MHz, 2.7MHz, 4.5MHz, 9MHz, 13.5MHz, 18MHz]$. The bandwidth actually allocated by the BS to the user definitely determines the net wireless downlink channel (DLC) rate per user.

### C. Computation of the allocated bandwidth

The video stream at quality $q$ is conveyed in chunks, corresponding to a playout time of $\tau$ sec. We denote by $\lambda(k, q)$ the size (measured in bits) of chunk $k$. The chunk size is a random variable depending on the inherently random, time-variant nature of the video content as well as to the actual encoding settings. Then, the instantaneous level $r_k$ of the application layer rate used for the transmission of the $k$-chunk, $k = 1, 2, ...$, is given as:

$$r_k = \frac{\lambda(k, q)}{\tau}\ [bps]. \tag{2}$$

This can be translated to a wireless DLC rate $r_k^{(DLC)}$ [bps] that depends on the overhead that is added to the DLC to support the application rate $r_k$.

In accordance with the considered protocol stack, above the DLC, network and transports protocols implement their control policy so that the throughput at the application layer is only a fraction $\alpha$, with $\alpha < 1$, of the $r_k^{(DLC)}$. Specifically, we consider that the TCP throughput adapts to the current bottleneck that here is assumed to be the LTE bearer. Hence, a mobile user that requests a streaming service at QoE = $q$, requires a downlink wireless channel that transmits chunk $k$ at a data rate that is equal to:[1]:

$$r_k^{(DLC)} = r_k/\alpha\ [bps]. \tag{3}$$

The actual bandwidth required for the transmission of the chunk to the mobile user depends from the user's monitored CQI level, which ultimately determines the spectral efficiency level $\gamma(c)$ [bps/Hz].

This is exemplified in Table I. For example, for a user that reports CQI $c = 1$ the spectral efficiency is equal to $\gamma(1) = 0.1523\ bps/Hz$. Then, on the bandwidth of one RB, $BW_{RB} = 180\ kHz$, the rate available at the net of the channel encoding overhead equals to $\gamma(1) \times BW_{RB} = 0.1523\ bps/Hz \times 180\ kHz = 27.42\ kbps$. Remarkably, a fraction of the available time-frequency resources is dedicated to transmit channel control signaling as well as training information to be exploited at the receiver for synchronization

[1] In more details, $r_k^{(DLC)}$ is the value to be provided, on average, over the chunk $k$ download time.

TABLE I
DOWN LINK CQI $c$, NOMINAL SPECTRAL EFFICIENCY $\gamma(c)$ AND NOMINAL DOWNLINK (DL) THROUGHPUT $\theta_{RB}(c)$ (KBPS PER RB), NET SPECTRAL EFFICIENCY $\gamma_{net}(c)$ AND NET DOWNLINK (DL) THROUGHPUT $\theta_{RB}(c)$ (KBPS PER RB).

| CQI index $c$ | Modulation | Spectral efficiency $\gamma(c)$ [bps/Hz] | Nominal DL Throughput (kbps per RB) | Net spectral efficiency $\gamma_{net}(c)$ [bps/Hz] | Net DL Throughput $\theta_{RB}(c)$ (kbps per RB) |
|---|---|---|---|---|---|
| 0 | out of range | | | | |
| 1 | QPSK | 0.1523 | 27.42 | 0.1065 | 19.19 |
| 2 | QPSK | 0.2344 | 42.19 | 0.1640 | 29.53 |
| 3 | QPSK | 0.3770 | 67.85 | 0.2638 | 47.50 |
| 4 | QPSK | 0.6016 | 108.28 | 0.4211 | 75.80 |
| 5 | QPSK | 0.8770 | 157.85 | 0.6139 | 110.5 |
| 6 | QPSK | 1.1758 | 212.4 | 0.8222 | 148.15 |
| 7 | 16QAM | 1.4766 | 265.79 | 1.0333 | 186.05 |
| 8 | 16QAM | 1.9141 | 344.54 | 1.3389 | 241.17 |
| 9 | 16QAM | 2.4063 | 433.13 | 1.6833 | 303.19 |
| 10 | 64QAM | 2.7305 | 491.49 | 1.9111 | 344.04 |
| 11 | 64QAM | 3.3223 | 598.01 | 2.3222 | 418.60 |
| 12 | 64QAM | 3.9023 | 702.41 | 2.7278 | 491.68 |
| 13 | 64QAM | 4.5234 | 814.21 | 3.1611 | 569.94 |
| 14 | 64QAM | 5.1152 | 920.74 | 3.5778 | 644.51 |
| 15 | 64QAM | 5.5547 | 999.85 | 3.8833 | 699.89 |

TABLE II
APPLICATION LAYER $r_k$ AND DATA LINK LAYER $r_k^{(DLC)}$ RATE BOUNDS FOR DIFFERENT CQI VALUES OF THE LTE MOBILE USER $BW_{LTE} = 1.4\,MHz$.

| CQI | Maximum $r_k^{(DLC)}$ [kbps] | Maximum $r_k$ [kbps] |
|---|---|---|
| 4 | 450 | 288 |
| 5 | 660 | 422.4 |
| 6 | 888 | 568.32 |
| 7 | 1116 | 714.24 |
| 8 | 1446 | 925.44 |
| 9 | 1818 | 1163.5 |
| 10 | 2064 | 1321 |

TABLE III
MAXIMUM $r_k[kbps]$ IN THE DIFFERENT LTE BANDWIDTHS.

| | Max $r_k$ [kbps] | | |
|---|---|---|---|
| CQI | $1.4\,MHz$ | $3\,MHz$ | $5\,MHz$ |
| 4 | 288 | 720 | 1200 |
| 5 | 422.2 | 1056 | 1760 |
| 6 | 568.32 | 1420.8 | 2368 |
| 7 | 714.24 | 1785.6 | 2976 |
| 8 | 925.44 | 2313.6 | 3856 |
| 9 | 1163.5 | 2908.8 | 4848 |
| 10 | 1320.9 | 3302.4 | 5504 |

and channel estimation purposes[2]. In [13], the authors report the the net downlink (DL) throughput $\theta_{RB}(c)$ available for user data transmission within one RB for different CQIs $c$; these values are summarized in the right column of Table I. Hence, the net spectral efficiency can be defined as $\gamma_{net}(c) = \theta_{RB}(c)/BW_{RB}$ $[kbps/Hz]$; clearly, $\gamma_{net}(c)$ is always lower than $\gamma(c)$ and it can be used to estimate the bandwidth actually available for user data transmission.

On the basis of the number of RBs that can be assigned in a given LTE channel bandwidth (e.g., 6 RB for $BW_{LTE} = 1.4\,MHz$) in Table II we report the maximum application layer data rate in case of $BW_{LTE} = 1.4\,MHz$. Table III reports instead the maximum data rates in 3 different LTE channel bandwidths for the different CQI.

Consequently, a user experiencing CQI $c$ and provided with a stream at QoE level $q$ will require for the transmission of the $k$-th chunk a bandwidth level that is equal to

$$BW(k;q,c) = \frac{r_k^{(DLC)}}{\gamma_{net}(c)} = \frac{\lambda(k,q)}{\alpha\tau\gamma_{net}(c)}\ [Hz]. \qquad (4)$$

We note that the quality $q$ of the video stream to be sent to a mobile user that reports an observed CQI $c$ can be selected by employing a multitude of criteria and schemes. The consequent distribution of the required bandwidth $BW(k;q,c)$, given a recorded CQI level, follows then the distribution of the chunk size $\lambda(k,q)$, with parameters suitably scaled in accordance with the observed CQI $c$ level. The moments and percentile levels of the required bandwidth level are calculated accordingly, and are used to characterize the channel bandwidth resources that must be used by the BS which is acting as a proxy server that is placed between the user and the actual HAS video server. A qualitative behavior of the proposed scheme is reported in Figure 2. The user sends its CQI reports to the BS. The central manager at the BS dynamically adapts the bandwidth level allocated to the user on the basis of the reported CQI values, taking into consideration the lengths of the video chunks that are produced during each playout time of $\tau$ sec. In accordance with the effective available bandwidth level and the video quality requested by the mobile terminal,

[2] For instance, in a Time Division Duplexing System, every two RBs, up to 3 out of 14 OFDM symbols can be dedicated to the control channel, and one more symbol is devoted to channel estimation; besides, a selected subset of subcarriers, spanning the bandwidth of a few RBs, is selectively devoted to synchronization, namely twice every 11 RBs.

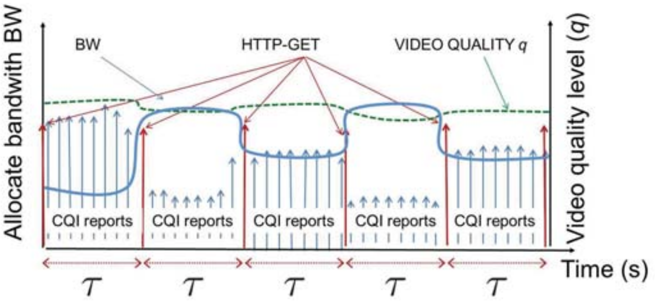


Fig. 2. Qualitative bandwidth allocation and QoE as a function of time.

the BS manager acts to keep the quality level $q$ to its prescribed level so that users experiences its desired QoE level.

In the following, we compute the number of users that can be accommodated by the described resource allocation scheme, under a targeted stall probability value, $\mathcal{P}_{stall}$ equal to zero[3]. The overall allocated downlink bandwidth level is equal to $BW$ $[Hz]$.

The analysis is carried out by assuming stationarity for the process characterizing the CQI value monitored at client mobile users. The bandwidth level used for transmitting a chunk to a user is then calculated based on the reported user's CQI level and the actual size of the chunk.

The expected value of the Dynamically Allocated (DA) bandwidth $BW_{DA}$ per user is evaluated as:

$$BW_{DA} = E\left\{\frac{\lambda(k,q)}{\alpha\tau\gamma_{net}(c)}\right\}[Hz] \quad (5)$$

where the expectation is carried out with respect to the CQI, the quality level and the random chunk size. Assuming a relationship $q = q(c)$ between the user's reported CQI level and the selected quality level, the expectation would then be carried out with respect to the distributions of the CQI and chunk size variables, yielding:

$$BW_{DA} = E_{CQI}\left\{E_{\Lambda|CQI}\left\{\lambda(k,q)/\gamma_{net}(c)\right\}\right\}/(\alpha\tau). \quad (6)$$

We set $\mu_q$ to denote the expected value of the chunk size under a given quality level $q = q(c)$. We then have:

$$\begin{aligned} BW_{DA} = \quad & E_{CQI}\left\{\mu_{q(c)}/\gamma_{net}(c)\right\}/(\alpha\tau) = \\ & \left(\textstyle\sum_{c_{min}}^{c_{max}} p_{CQI}[c]\mu_{q(c)}/\gamma_{net}(c)\right)/(\alpha\tau). \quad (7) \end{aligned}$$

The number of accommodated users is then calculated as:

$$N_{DA} = \lfloor BW_{LTE}/BW_{DA} \rfloor \quad (8)$$

For comparison purposes, we consider a Static Allocation (SA) scheme. The CQI reported by the user is received by the base station proxy manager and is used to determine the video quality level, as noted above. However, the bandwidth level allocated for the transmission of the stream's chunks is not varied on a chunk by chunk basis. This is in contrast with the operation performed by the dynamic scheme, whereby the length of each chunk is observed and the allocated bandwidth level is accordingly adapted, impacting the data rate and bandwidth levels occupied for the transmission of each chunk. Under the stationary scheme, the bandwidth level that is allocated for streaming to a user is calculated by averaging over the statistical values expected to characterize the variations of the chunk size variables (which do depend on the reported CQI and the corresponding selected quality level $q$). To avoid stall events (or, rather, reduce the probability of their occurrence), the data rate that is used for transmission of chunks across the downlink channel is set to a constant value that is equal to twice the average video rate $R(q)$ [7][14]. Such a setting is commonly employed, though it generally leads to inefficient utilization of spectral resources. We readily observe such a static allocation to require the following bandwidth level to be allocated for the transmission of a stream:

TABLE IV
EXAMPLE OF QUALITY ASSIGNMENT FOR DIFFERENT VIDEO ENCODER RATES: "HARRY POTTER" TRACES, [15].

| Quality $q$ | Quantization | Video Bit Rate $R(q)$ [kbps] |
|---|---|---|
| 1 | 20 | 3856.3 |
| 2 | 25 | 1192.11 |
| 3 | 30 | 527.7 |
| 4 | 35 | 273.65 |
| 5 | 40 | 150 |
| 6 | 45 | 81 |

$$BW_{SA} = E_{CQI}\left\{2R(q)/\gamma_{net}(c)\right\}/(\alpha). \quad (9)$$

We observe that the static allocation scheme requires the allocation of a larger bandwidth level. It assumes a value that is up to double the bandwidth used by the dynamic allocation scheme. The number of accommodated user is now given as:

$$N_{SA} = \lfloor BW_{LTE}/BW_{SA} \rfloor. \quad (10)$$

We note that the gap is reduced when the bandwidth allocation process involves the use of discrete bandwidth resource blocks. Also, one recognizes that the indicated advantage gained by the DA in bandwidth utilization is associated with the use of extra bandwidth and processing resources that are required for implementing the control and signaling processes.

## III. PERFORMANCE ANALYSIS

In this section, we present illustrative performance results for the streaming model described above. We carry out numerical analysis by using the performance equations presented above. We assume a H.265/HEVC 10.1 encoded video with a spatial resolution of 1920x1080, whose traces are publicly available at [15]. The traces correspond to the movie "Harry Potter", and are constructed by using 1799 encoded chunks. The encoding process operates at a rate of 24 frames per second, and the resulting GoP pattern is G24B7.

The video is segmented into chunks. Each chunk represents an encoding of a video segment of duration $\tau = 2s$ (48

[3] $\mathcal{P}_{stall} \leq p(r_k > r_{alloc})$, induced by random variations in the available and allocated physical channel rate and noting that network events may impose require an overhead ratio that is higher than the estimated $\alpha$ value.

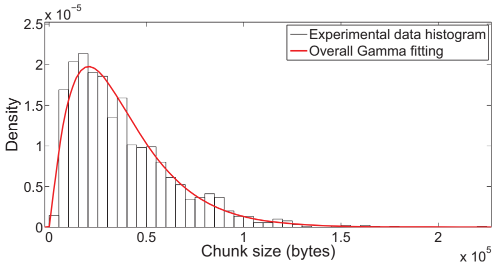


Fig. 4. Fitting of the QoE chunk size distributions in case of $q = 5$.

frames), corresponding to two GOPs. The quality of the video is further determined by the quantization level that is used, considering 6 different quantization levels. We assume that 6 different QoE levels are observed; each QoE level corresponds to an uninterrupted transmission of the stream at one out 6 encoded video quality levels. We identify the associated QoE level by the index $q$, which ranges from 1 to 6, with 1 designating the highest QoE level. We set the aggregate overhead efficiency of the involved layering process as $\alpha = 0.64$. Encoder settings and resulting video rates $R(q)$ for the 6 different qualities are summarized in Table IV.

### *A. Video characterization*

In Figs. 3 (a)-(c), we exhibit illustrative variations of chunk size $\lambda(k, q)$ sequences versus the temporal index $k$, as observed under different QoE quality values, namely $q = 1$, $q = 3$, $q = 5$. Large fluctuations in the chunk size are observed. These wide variations explain the large bandwidth margins that must be assigned to avoid the occurrence of reception stalls when a static bandwidth allocation scheme is employed. Using the measured chunk size sequences, we construct the corresponding chunk size distribution $p_\Lambda(\lambda)$ histograms. In Fig. 4, we display the chuck size histogram corresponding to $q = 5$. We show that a Gamma probability density function (pdf) can be used to provide a very good fit. We have shown a Gamma distribution function to provide excellent fit also when using other quality levels (though we do not report here the detail of such a fitting process due to space limitations). Consequently, in performing the following analyses we assume that the chunk size probability density function is well approximated by a Gamma pdf, whose parameters depend on the video statistics, i.e., $p_\Lambda(\lambda) \approx \Gamma(\lambda; \mu_q, \sigma_q^2)$.

### *B. Bandwidth allocation under constant CQI*

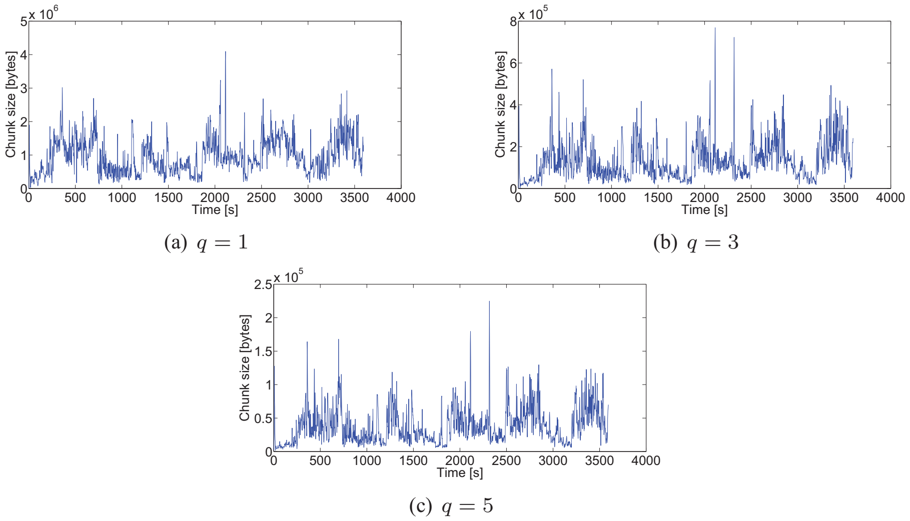
As discussed in Section II, given an overall bandwidth $BW_{LTE}$, a number $N_{RB}$ of 180 kHz wide bandwidth resource blocks (RBs) can be accommodated. The actual rate carried within each RB depends on the selected modulation/coding set (MCS), which is CQI dependent. Hence, the number of RBs allocated for the transmission of a stream to a mobile user, which is generated at a certain application-layer rate, depends on the CQI experienced by the user.

This is exemplified in Fig. 5, where the application layer rate and the corresponding DLC layer rate for different CQI levels are illustrated. The considered overall LTE channel bandwidth is $BW_{LTE} = 1.4\ MHz$. For each CQI level, a slotted bar represents the rate $\overline{r}^{(DLC)}$ allocated at the DLC level to carry data at an application layer rate $\overline{r} = 150kbps$ represented by the red line; this value equals to the average video rate at quality $q = 5$ (see Tab. IV). Within each bar, each colored slot represents the contribution of a single resource block to the overall transmit rate. In order to transmit at an application layer rate $\overline{r}$, the number of employed RBs decreases from 4 to 1 as the CQI increases from 4 to 10. Accordingly, the occupied bandwidth varies from $4 \times 180 = 720\ kHz$ to $180\ kHz$.

In Fig. 6, we exhibit the temporal variation of the number of RBs assigned for the transmission of the video stream analyzed in Fig. 3(c) and in Fig. 5. The bandwidth levels allocated to this stream are determined by the employed DA scheme.

In Fig. 7, we illustrate the bandwidth employed to transmit, in case of DA, to user stations recording different CQI levels, considering streams that are encoded at different quality values $q$. For each quality $q$, and CQI $c$, we evaluate the average rate and the bandwidth levels that are required to transmit it to a user station that experiences the indicated CQI value; let us remark that under these assumptions, the only randomness is related to the chunk size. Specifically, the average bandwidth is computed by first evaluating the bandwidth allocated to each chunk, quantized in terms of multiples of the RB rate as already shown in Fig. 5, and then averaging on the $\approx 1800$ chunks. As far as the standard deviation of the allocated bandwidth is concerned, the following comments are in order. Firstly, the deviation is caused by intrinsic fluctuation of the video chunk size, so that the ratio of the standard deviation and the average ranges in between 68% and 54% when quality $q$ changes from 6 to 1; specifically, it can be argued that video at higher quality, i.e., $q = 1$, shows larger absolute fluctuations, but the ratio of the standard deviation and the average is lower. Secondly, for each chunk, the actual allocated bandwidth depends on the CQI and it is quantized in a finite number of RBs as in the example in Fig. 5. In Tab. V, we report the standard deviation over average ratios of the allocated bandwidth as a function of the CQI $c$ and video quality $q$. We observe that in case of low CQI, small fluctuation in the chunk size may typically result into larger fluctuations the number of employed RBs. Thereby, for low quality video ($q = 6$), the standard deviation of the bandwidth ranges from a high value of $55\%$ of the average in case of low CQI down to a lower value of $15\%$ of the average in case of high CQI. On the other hand, for high quality video ($q = 1$), the standard deviation of the bandwidth is almost constant to $54\%$ of the average.

It is noted that, depending on the receiver's CQI level, it is not always possible to transmit streams at a sufficiently high quality level (bandwidth values above the dashed lines of Fig. 7 are not supported in the different $BW_{LTE}$). For comparison sake, we plot in Fig. 8, the corresponding levels of bandwidth that must be allocated to transmit a stream at a rate $R(q)$ under different incurred CQI levels, when a static allocation method is used. The distinct performance advantages attained through the use of the dynamic allocation



(a) $q = 1$

(b) $q = 3$

(c) $q = 5$

Fig. 3. Chunk size versus time for different quality levels

TABLE V
STANDARD DEVIATION OVER AVERAGE RATIOS OF THE ALLOCATED BANDWIDTH IN CASE OF DA AS A FUNCTION OF THE CQI $c$ AND VIDEO QUALITY $q$.

| $c$ / $q$ | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| 6 | 55% | 49% | 43% | 37% | 29% | 19% | 15% |
| 5 | 61% | 58% | 55% | 52% | 48% | 41% | 39% |
| 4 | 65% | 63% | 61% | 59% | 57% | 54% | 52% |
| 3 | 66% | 65% | 64% | 62% | 61% | 59% | 59% |
| 2 | 63% | 62% | 62% | 61% | 60% | 59% | 59% |
| 1 | 54% | 54% | 54% | 54% | 54% | 53% | 53% |

scheme are well demonstrated. In comparing with the static scheme, the dynamic bandwidth assignment scheme leads to significant reductions in the bandwidth level that is required for the streaming process, under each examined video quality and user recorded CQI levels. The use of a dynamic scheme also allows the system manager to set the stream's QoE to meet specific requirements by certain users. By observing again Table V and Fig. 8 we can then highlight that in the worst case, $q = 3$ and $c = 4$, the DA allows to save 35% of the allocated bandwidth with respect to the SA, while in the best case, $q = 6$ and $c = 10$, the bandwidth saving is around 85/

Obviously the bandwidth gains realized through the use of the DA reflects on the attained total number of users that can be supported over the specified $BW_{LTE}$.

### C. Bandwidth allocation under variable CQI

Let us now assume the system mobile users to be categorized into two classes. User members of the High Quality (HQ) class record CQI values that are governed by the probability distribution function $p_{CQI|HQ}(c)$. Users that belong to a Low Quality (LQ) class observe communications channels identified by the a CQI distribution that is given as $p_{CQI|LQ}(c)$. We assume that a percentage of $p_{HQ}$ users experience high channel quality, e.g. due to their distance from the BS, and $p_{LQ} = 1 - p_{HQ}$ experience low channel quality. The CQI probability distribution function of a randomly selected user is expressed as:

TABLE VI
QUALITY VS CQI IN THE DYNAMIC AND STATIC ALLOCATION MODES.

| CQI range | $q(c)$ (DA) | CQI range | $q(c)$ (SA) |
|---|---|---|---|
| 1,2,3 | 6 | 3,4 | 6 |
| 4,5 | 5 | 5 | 5 |
| 6,7 | 4 | 6,7,8 | 4 |
| 8,11 | 3 | 9,10,11 | 3 |
| 12,...,15 | 2 | 12,...,15 | 2 |

$$p_{CQI}(c) = p_{HQ} \cdot p_{CQI|HQ}(c) + p_{LQ} \cdot p_{CQI|LQ}(c). \quad (11)$$

We consider a mobile user that experiences and reports to the BS manager a CQI $c$. A stream of quality $q(c)$ is sent to this user, such that the resulting correspondingly required bandwidth value fits within the bandwidth capacity of 6 RBs ($BW_{LTE} = 1.4\ MHz$), with a small margin maintained to allow for transmission resiliency under small CQI level fluctuations. The association is shown in Table VI.

We assume that the user's CQI randomly changes according to a uniform distribution within the user's LQ/HQ class. this may model some interference, mobility, shadowing on the user access channel. As the CQI level observed at the mobile user changes, the above described system operations, under the DA and SA schemes, may not be able to preserve the same quality level. In Figure 9, we display results for the probability that the streamed quality remains unchanged under CQI variations. We notice that under the use of the DA scheme, a HQ user can be guaranteed to experience the same quality level with higher probability. In turn, when employing the SA scheme, we note that a LQ user keeps its initial quality level with a lower probability value. This behavior well illustrates the ability of the dynamic scheme to assure mobiles with higher persistence in maintaining the QoE level of the video streams.

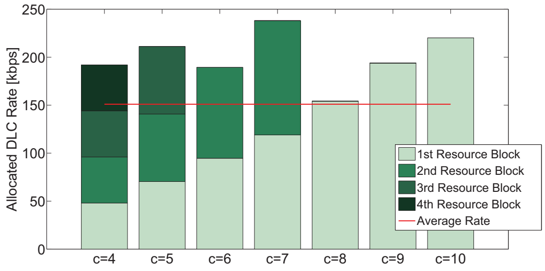


Fig. 5. Allocated DLC rate $\overline{r}^{(DLC)}$ as a function of the CQI $c$ for an application layer rate $\overline{r} = 150\ kbps$.

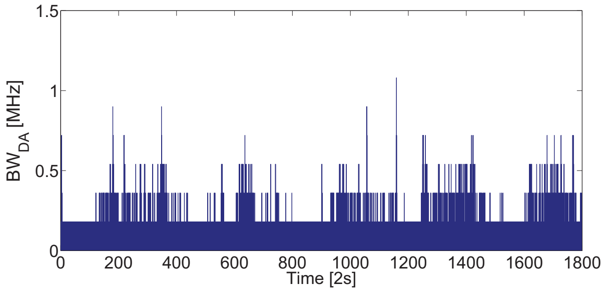


Fig. 6. Time evolution of the allocated bandwidth $BW_{DA}$ for the video streamed at quality $q = 5$.

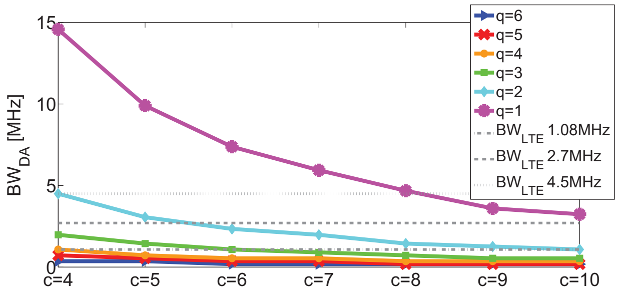


Fig. 7. Resource allocation with dynamic approach and comparison with three different LTE bandwidths

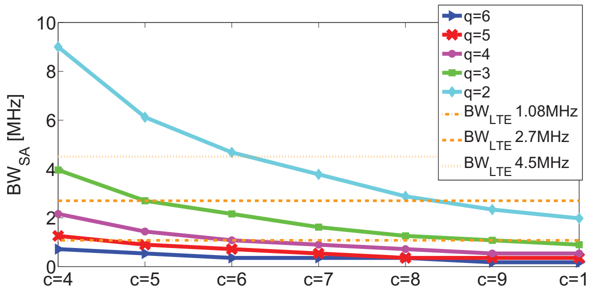


Fig. 8. Resource allocation with static approach and comparison with three different LTE bandwidths

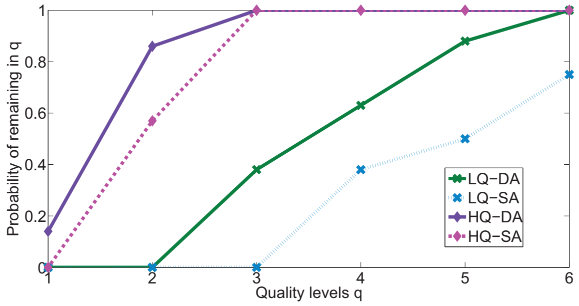


Fig. 9. Probability of remaining in a quality level $q$ for HQ and LQ users under DA and SA.

We have also analyzed the effectiveness of the dynamic allocation scheme in allocating bandwidth resources. For this purpose, we assume the system to support users that are divided into two classes; users that belong to the LQ class monitor CQI levels that are assumed to be uniformly distributed across the set $\{1, \cdots 8\}$, while users that belong to the HQ class observe CQI levels that are uniformly distributed across the set $\{9, \cdots 15\}$. A few performance indicators can be computed: the percentage of non accommodated users $p^{na}$, the average streamed video quality $E\{q\}$, where the expectation is evaluated wrt the CQI reported by the accommodated users only, the allocated bandwidth per user $BW$ [kHz], and the maximum number of users that can be accommodated by assuming an overall bandwidth $BW_{LTE} = 20MHz$. The ensuing results are summarized in Table VII. We observe the following. When the dynamic allocation scheme is used, the percentage of non accommodated users $p^{na}_{DA}$ is always equal to 0%. In contrast, we note that under the use of a static allocation scheme, no service is provided at low CQI levels, namely $c = 1, c = 2$. Under $p_{LQ}$ equal to 1, so that all users receive lower quality streams, if a SA scheme is used, the percentage of users that cannot be accommodated, $p^{na}_{SA}$, is equal to 25% of the cases. As the fraction of LQ users decreases, so that $p_{LQ}$ decreases from 1 to 0, the value of $p^{na}_{SA}$ decreases from 25% to 0%. The average streamed video quality is then evaluated by accounting only for the cases in which a video can be successfully provided and received. The average video quality of the DA scheme $E\{q\}$ (DA) overcomes the average video quality of the SA scheme $E\{q\}$ (SA) at any $p_{LQ}$. To further analyze the system's behavior we have evaluated, for the accommodated users, the allocated bandwidth levels $BW_{DA}$ and $BW_{SA}$, computed by using equations (7) and (9) for different values of $p_{LQ}$. The maximum number of users that can be accommodated, $N_{DA}$ and $N_{SA}$, is computed. We observe that, for any prescribed $p_{LQ}$ value, the dynamic allocation scheme systematically allows the accommodation of a larger number of users, at an equal or larger quality, through the use of much less bandwidth resources.

## IV. Discussion

As far as the complexity issue is concerned, two comments are in order. First, the http/video proxy functionalities at the BS do not require pre-fetching of the whole video data at the proxy site but just of selected data -namely the storing a description of the video bitstreams available at the server and

TABLE VII
OVERALL PERCENTAGE OF NOT ACCOMMODATED USERS, AVERAGE RECEIVED VIDEO QUALITY, ALLOCATED BANDWIDTH (IN KHZ) AND NUMBER OF ACCOMMODATED USERS USING DA AND SA FOR DIFFERENT VALUES OF $p_{LQ}$, $BW_{LTE} = 20\,MHz$.

| $p_{LQ}$ | 1 | 0.75 | 0.5 | 0.25 | 0 |
|---|---|---|---|---|---|
| $p_{DA}^{na}$ | 0% | 0% | 0% | 0% | 0% |
| $p_{SA}^{na}$ | 25% | 18.75% | 12.5% | 6.25% | 0% |
| $E\{q\}$ (DA) | 4.9 | 4.3 | 3.7 | 3.0 | 2.4 |
| $E\{q\}$ (SA) | 4.8 | 4.2 | 3.6 | 3.0 | 2.4 |
| $BW_{DA}$ | 525.11 | 492.33 | 459.56 | 426.78 | 394.0 |
| $BW_{SA}$ | 757.93 | 765.45 | 772.97 | 780.50 | 788.02 |
| $N_{DA}$ | 38 | 40 | 43 | 46 | 50 |
| $N_{SA}$ | 26 | 26 | 25 | 25 | 25 |

advertised to the client at the beginning of the session- and of the bitstream at the selected quality; this latter can be just partially pre-fetched and then continuously downloaded during transmission of the chunks on the wireless channel. Second, it is expected that the need for mobile streaming services will increase exponentially in the next few years, taking the lion share in the bandwidth consumption factors. Therefore, a reasonable improvement in the per user bandwidth efficiency could lead to high gains definitely motivating the adoption of more complex access network interfaces.

We note that by using the underlying models presented in this paper, one can also employ other dynamic allocation methods. For instance, the BS can encode a stream in a manner that depends on the user's experienced CQI level, so that users that experience higher SINR levels are granted video streams at higher QoE levels. As the user moves, its recorded SINR (and CQI) level may change, and then possibly inducing the reception of streams at varying QoE levels. For example, interior mobiles, which are located closer to the base-station (BS) node, tend to experience higher SINR levels. They can then be provided streams at higher QoE levels. Edge mobile users, which are located closer to the edge of the cell and are subjected to higher (inter-cell) interference signals, could then be provided streams at lower QoE levels. Under such an operation, the average required bandwidth per stream is expected to be much lower. A higher number of mobiles can then be provided support. We note that for multicast streaming over a cellular wireless downlink channel, the use of such variable QoE operations has been presented and analyzed in [16]. We note however that the latter uses no HAS and chunk by chunk based channel bandwidth and compression adaptations, and does not prescribe stall performance objectives, as a multicasting dissemination is performed.

For resource allocation management systems that do not wish to implement dynamic bandwidth variations, the average (or 95 percentile) required bandwidth level can be calculated and employed. Such an operation would however lead to stochastic stall occurrences. The corresponding stall probability will depend on application layer settings, such as the initial delay of the playout process, and the employed re-buffering technique. It is observed that the stall probability can be upper bounded by using the involved bandwidth related parameters: a maximum of 5% stall probability level is expected if the bandwidth level is sufficient for 95% of the chunks.

## V. CONCLUSIONS

In this paper, we present, model and evaluate a system that employs HAS streaming over a wireless access network, such as that implemented by using LTE technology, incorporating the use of a dynamic bandwidth allocation scheme. Bandwidth is allocated on a chunk by chunk basis, assuring the targeted mobile user to not experience stall events, based on the average channel quality experienced by the client. Such dynamic allocation has several advantages. In addition to performing more efficiently in terms of bandwidth allocation per user, it avoids or reduces the need for the application layer to dynamically adapt the quality of the video stream. These latter adaptations, as those envisaged by DASH, result in visual quality fluctuations that often degrade the user's experienced quality of video stream reception.

## REFERENCES

[1] M. Rinne and O. Tirkkonen, "LTE, the radio technology path towards 4G," *Computer Communications*, vol. 33, no. 16, pp. 1894 – 1906, 2010.

[2] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes, "The evolution to 4G cellular systems: LTE-Advanced," *Physical Communication*, vol. 3, no. 4, pp. 217 – 244, 2010.

[3] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[4] S. Colonnese, P. Frossard, S. Rinauro, L. Rossi, and G. Scarano, "Joint source and sending rate modeling in adaptive video streaming," *Signal Processing: Image Communication*, vol. 28, no. 5, pp. 403 – 416, 2013.

[5] C. Mueller, S. Lederer, and C. Timmerer, "A proxy effect analyis and fair adatpation algorithm for multiple competing Dynamic Adaptive Streaming over HTTP clients," in *IEEE Visual Communications and Image Processing, VCIP*, 2012.

[6] H. Nam, K. H. Kim, B. H. Kim, D. Calin, and H. Schulzrinne, "Towards dynamic QoS-aware over-the-top video streaming," in *IEEE WoWMoM 2014*, June 2014, pp. 1–9.

[7] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," in *Proceedings of the second annual ACM conference on Multimedia systems*, ser. MMSys '11, 2011, pp. 157–168.

[8] B. Wang, J. Kurose, P. Shenoy, and D. Towsley, "Multimedia streaming via TCP: An analytic performance study," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 4, no. 2, 2008.

[9] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. Andrews, "Video capacity and QoE enhancements over LTE," in *Communications (ICC), 2012 IEEE International Conference on*, June 2012, pp. 7071–7076.

[10] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "Qdash: A qoe-aware dash system," in *Proceedings of the 3rd Multimedia Systems Conference*, ser. MMSys '12, 2012, pp. 11–22.

[11] W. Pu, Z. Zou, and C. W. Chen, "Video adaptation proxy for wireless dynamic adaptive streaming over http," in *Packet Video Workshop (PV), 2012 19th International*, May 2012, pp. 65–70.

[12] "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 8.8.0 Release 8)," 2009.

[13] A. Ghosh and R. Ratasuk, "Essentials of LTE and LTE-A," *Cambridge Wireless Essentials Series*, 2011.

[14] S. Colonnese, F. Cuomo, R. Guida, and T. Melodia, "Performance evaluation of sender-assisted http-based video streaming in wireless ad hoc networks," *Ad Hoc Networks*, vol. 24, no. Part B, pp. 74 – 84, 2015.

[15] "Video Trace Files and Statistics," 2014. [Online]. Available: http://trace.eas.asu.edu/videotraces2/h265/

[16] H.-B. Chang, I. Rubin, and O. Hadar, "Scalable Video Downlink Multicasting in Multi-cell Cellular Wireless Networks," in *IEEE Globecom Workshop 2014*, December 2014.