Cross-Layer Approach for HTTP-Based Low-Delay Adaptive Streaming in Mobile Networks

Hieu Le$^{\pm}$, Konstantin Miller$^{\pm}$, Arash Behboodi$^{\mp}$ and Adam Wolisz$^{\pm}$
($\pm$) Technische Universität Berlin, Germany, ($\mp$) RWTH Aachen University, Germany

Abstract—**The increasing number of mobile devices with high processing power and high-resolution screens had led to an enormous growth of mobile video traffic. Mobile network operators face the requirement to efficiently support large numbers of concurrent unicast streaming sessions. In the present work, the long-term quality of experience perceived by the user, the fairness, and the overall system efficiency are addressed simultaneously from the cross-layer perspective by jointly optimizing the video adaptation and the wireless resource allocation. One fundamental challenge of the cross layer design is that the time scale of video adaptation—seconds—differs by several orders of magnitude from the one of resource allocation—milliseconds. We focus on the low-delay live streaming, which is particularly sensitive to the throughput fluctuation. We consider the streaming both in the downlink and in the uplink, explicitly taking into account the imperfect synchronization in the uplink. Our proposed solution consists of two components. First, we formulate the problem of video adaptation as a quality of experience based max-min optimization problem that leverages the link rate estimate in the lower network layers. Second, we propose a dynamic resource allocation scheme that takes into account the demands of the streaming clients. These two components together aim at a fair and efficient cross-layer streaming system. An accurate estimation of link rate on the time scale of seconds, required for this problem, is particularly difficult in mobile networks. As a separate contribution, several link rate estimation approaches are evaluated. The prediction algorithms assuming the static resource allocation, although computationally less complex, may lead to inaccurate prediction results in comparison to the one when dynamic resource allocation scheme is used. In this work, the spectral efficiency gain by dynamic resource allocation can be approximated and used to improve throughput predictions. We evaluate the proposed approach against state-of-the-art baselines. The results reveal significant improvements of quality of experience in all studied use cases.**

Keywords*—*Adaptive streaming, mobile networks, dynamic resource allocation, throughput prediction, cross layer optimization

I. Introduction

The video industry has broadly adopted adaptive streaming technologies, such as Dynamic Adaptive Streaming over HTTP (DASH) [1], to provide a smooth viewing experience in dynamic environments. However, there are several challenges to develop robust streaming solutions in mobile networks. First, although most of the streamed content nowadays are Video on Demand (VoD), the amount of live streaming is growing rapidly [2]. While current live streaming services can exhibit a latency of several tens of seconds, a *low-delay* streaming targets a particularly low latency of a few seconds or less. Using the *de-facto* standard DASH, which has been primarily developed to replace progressive download for VoD, to efficiently stream low-delay contents, especially in mobile networks, is still an open challenge.

Second, typical streaming clients operate entirely on the application layer, while treating the network as a black box. This is particularly true for browser-based applications (e.g. YouTube and Netflix[1]). They often adapt the target video bit rate to the playback buffer and a—potentially implicit—estimation of the future throughput (by, for instance, extrapolating previous application-layer throughputs). Meanwhile, the scheduling strategies of modern networks like Long Term Evolution (LTE) and Worldwide Interoperability for Microwave Access (WiMAX) are commonly designed based on the proportional fairness approach to efficiently allocate radio resources to single-rate videos [3], and do not incorporate efficient frameworks for adaptive videos [4]. The separation significantly facilitates the deployment and reduces the complexity. Consequently, this design is broadly used despite the potential performance gain by cross-layer and client coordination techniques [5], [6]. Recently, however, in order to cope with the explosive traffic growth, mobile network operators are increasingly forced to look for new ways to more efficiently serve the video traffic, creating a strong incentive to consider cross-layer approaches.

There are two concrete aspects of adaptive streaming over mobile networks where cross-layer techniques can help boosting the performance. First, operating purely on the application layer, it is extremely challenging to almost impossible to compute a good estimation of the future throughput. This leads to inaccurate predictions, reducing the video performance. In this case, the link state information is beneficial to improve the prediction accuracy [7]. Second, the information about video characteristics can help Base Station (BS) to efficiently allocate radio resources to users that can benefit most from them. Furthermore, a joint consideration of multiple streams can assure a certain level of fairness, while avoiding overloading the Radio Access Network (RAN) [8].

Motivated by the above considerations, we propose a cross-layer approach to jointly optimize the video adaptation (VA) and the wireless resource allocation (RA) for the low-delay live streaming. The approach consists of two components. The first component—the video quality selector—selects appropriate video qualities, which maximize the long-term Quality of Experience (QoE) of multiple users subject to the capacity constraints of the wireless resource. We consider two use cases: hard latency constraints (LSH), and soft latency constraints (LSS). While client skips segments that miss their playback

978-1-5386-2723-5/17/$31.00 ©2017 IEEE

[1] http://youtube.com, http://netflix.com

deadlines to meet the given strict upper bound on the live latency, the client in the latter case play out the content without gaps. It means, in the LSS case, whenever a segment cannot be delivered by its playback deadline, the playback is halted until the segment download is completed, effectively increasing the latency. We study this use case with a configurable initial latency equal to a small multiple of the segment duration. Our proposed approach takes into account not only the dynamics of network conditions, but also—in the LSS case—users individual buffer levels. In addition, the proposed solution takes advantage of the more accurate throughput predictions that are based on the link state information.

Given the quality selection as the output of the first component, the second component, a dynamic resource allocation (DRA) scheme, strives to deliver the requested video qualities. In this context, one particular challenge arises as the time scale of VA—in seconds— is much larger than the one of RA—in milliseconds. To deal with that challenge, a sequential process of adapting RA to the instantaneous wireless channel states is performed to gradually match the users throughput with the demand. By leveraging DRA, we achieve valuable throughput gains by exploiting frequency and multiuser diversities to efficiently combat the varying channel. We consider both the downlink and uplink cases. For the uplink, we explicitly consider the imperfect synchronization among users by incorporating the mitigation of Multiple Access Interference (MAI) as deriving DRA schemes.

Finally, the problem of predicting the available throughput in mobile networks, which can strongly impact the video performance, is non-trivial even given the availability of link state information [9]. To select the efficient candidates, we evaluate several state-of-the-art prediction approaches, which are established based on the concept of ergodic capacity, through Monte Carlo simulations using a realistic non-stationary channel model. For each method, we compare the throughputs achieved as using either static resource allocation (SRA) or DRA. We observe SRA tends to give inaccurate predictions and the low bound (i.e. for a safety margin) of ergodic capacity should be most appropriate for SRA. Furthermore, a properly designed scheme of DRA significantly increases the users throughput by up to 150% as compared with the one of SRA. Based on this result, we propose to integrate the past throughput gains by DRA into the problem of predicting the future throughput, which is then used as the input for the video quality selection.

We evaluate the proposed approach using extensive Monte Carlo simulations developed on the simulation framework OMNeT++ [10], leveraging the advanced channel simulation tool QuaDRiGa [11]. The numerical results show significant performance improvements with regard to several performance metrics. For example, our approach increases the users' Quality Index (QID), defined by the mean quality minus the standard deviation (in terms of the Peak Signal to Noise Ratio (PSNR)), by more than 10dB for downlink and 20dB for uplink (in the LSH case). Moreover, it significantly reduces the number and the duration of video stalls by up to several hundred times.

II. Related Work

There is a large body of literature on the stand-alone adaptive streaming application exploiting various approaches such as the control theory [12], the Markov decision process [13], the machine learning [14], and heuristics selecting the video quality based on the playback buffer and/or the average throughput [15]. While most of them target VoD, some address live streaming [16]. Alternatively, our approach involves a tight cooperation between the lower network layers and the application layer to increase the spectral efficiency, and it involves the client coordination to increase the fairness.

Cross-layer approaches for the single-user video streaming typically focus on the packet scheduling, the error protection and the video adaptation as maximizing the perceived quality (e.g. [12]). However, several measurement campaigns have shown that robust adaptive streaming algorithms must look beyond the single-user view to account for the overall performance of multiple adaptive streams that compete for a bottleneck link like the RAN [8]. Several studies then propose to coordinate clients sharing a bottleneck link. In [17] the authors consider a simplified network model assuming a constant total capacity that can be arbitrarily distributed among the clients. Alternatively, other works adopt the ergodic capacity for each link rather than the static capacity of shared channels [16], [18], [19]. In contrast to these works, we assume a non-stationary channel, leverage cross-layer based throughput predictions, and apply a detailed network model explicitly including DRA and MAI in uplink. Moreover, we consider the low-delay streaming with latencies of just a few seconds.

A systematic framework for optimizing multiple streams over the Time Division Multiple Access (TDMA) based wireless networks is introduced in [20]. The work in [7] presents a solution for multiple adaptive video streams over LTE, where users can monitor RA at BS as estimating future capacity. To the best of our knowledge, however, no studies have been explicitly considered the natural tendency of imperfect synchronization in uplink.

Approaches that deploy application-layer throughput predictions or the path probing for video streaming include [12], [16]. Often, however, such studies consider prediction time scales that are more appropriate for VoD. Recently several studies have reported the great prediction enhancement by considering cross-layer link state information [7], [9], [21].

A few works attempt to exploit information about wireless link to generate relatively good prediction of link rate, which is then fed to VA operations (e.g. [18], [19]). Suitable prediction methods for cross layer approaches are typically based on the concept of ergodic capacity. Particularly, there might be three candidates in literature. The first one leverages the channel estimation to reconstruct the Probability Density Function (PDF) of the fading channel, and to generate sufficient samples off-line for the prediction [22]. The second one applies a low bound (i.e. safety margin) on the ergodic capacity, which is formulated as an exponential integral function of the average channel gain as used in [19]. The third one aims at providing tight bounds on the ergodic capacity by using results from multivariate statistics [23].

III. System Model

This section describes (i) the streaming model and (ii) the network model in our study. Table I summarizes the notation.

A. Streaming model

In this work we adopt the streaming model similar to the standard MPEG-DASH [24]. We assume each video content is available in several independent representations targeting different perceived qualities and different video bit rates. Each representation is split into segments, such that switching the representation is feasible on each segment boundary. The client issues requests to download the segments, selecting the most appropriate representation based on the observed network conditions. After a segment is downloaded, it is stored in the playback buffer until its playback deadline is reached.

Before playing-back, the client pre-buffers one or multiple segments to reach a target buffer level. The goal is to mitigate the negative impact of throughput fluctuations. Later, in the course of the streaming session, if a segment is not fully downloaded by its playback deadline, the playback is halted. The response actions depend on the use case. In the case of LSS, the client waits until the late segment is completely downloaded, and then resumes the playback. This is often termed re-buffering. In the LSH case, the download of the late segment is terminated, its playback is skipped, and the client proceeds with downloading and playing subsequent segments.

We assume the video data is dowloaded using a lossless transport protocol like Transmission Control Protocol (TCP). Then the performance goal is to pursuit a high QoE by adapting the video bit rate to the network throughput. Among the main factors influencing the QoE are (1) the pre-buffering duration, (2) the duration of playback interruptions (due to rebuffering in the LSS case or skipped segments in the LSH case), (3) the average video quality and (4) the quality fluctuation during the streaming session.

We adopt the following notation. $M = \{0, \dots, |M| - 1\}$ denotes the set of Mobile Stations (MSs) (i.e. video clients), indexed by $m \in M$. $L = \{0, \dots, |L| - 1\}$ denotes the set of available video representations, indexed by $l \in L$. We use index $s \in \mathbb{Z}$ to indicate the individual segments in the stream. Then we denote the number of segments played by a client by N_{seg}. Each segment shall contain T_{seg} seconds of video; we call T_{seg} the segment duration. To improve readability, we assume all streams have the same number of representations, the same segment duration, and each user plays N_{seg} segments.

Since we consider the live streaming, the segments are not initially available for download. Rather, they can be downloaded after they have been published to the server. We assume that segment s becomes available for download at time $s \cdot T_{\text{seg}}$. This provides a natural way to divide the time into slot of length T_{seg}, indexed by the segment index $s \in \mathbb{Z}$. We assume, at the start of time slot s, the client selects the most appropriate representation for segment s, based on its performance objectives, and issues a download request.

We assume the playback starts at the beginning of time slot $s = 0$, and the prebuffering is completed by that moment. We denote the buffer level, measured in seconds of playback time, at the beginning of time slot $s \geq 0$ by $B_m(s)$. We denote the number of segments downloaded prior to starting the playback by $N_{\text{pre}} \geq 1$. That is, $B_m(0) = N_{\text{pre}} \cdot T_{\text{seg}}$. Note that since segment s only becomes available at the beginning of time slot s, it holds $0 \leq B_m(s) \leq B_m(0)$. For the LSH case, we require $N_{\text{pre}} = 1$, setting the live latency to roughly $2T_{\text{seg}}$ (T_{seg} seconds to record the segment, and up to T_{seg} seconds to download it, plus server-side and the client-side processing overhead).

TABLE I. NOTATION OVERVIEW

$Q_m(s)$	Selected video quality for user m in slot s
$R_m(s)$	Selected video bit rate for user m in slot s
$B_m(s)$	Buffer level at the start of time slot s
$Q_{l,s,m}$	Quality of representation l for segment s in stream m
$R_{l,s,m}$	Video bit rates of representation l associated with $Q_{l,s,m}$
$\hat{C}_m(s)$	Estimated link rate of stream m in slot s
$\hat{G}_m(s)$	Targeted resource share for stream m in slot s
$\hat{A}_m(s)$	Expected throughput for m in slot s
$C_m(s)$	Empirical link rate of user m in slot s
$G_m(s)$	Actual resource share for user m in slot s
$A_m(s)$	Actual throughput for user m in slot s
$p_m(i,t)$	Transmission power on sub-channel i in t of user m
$h_m(i,t)$	Channel gain on sub-channel i in t of user m
$\iota_m(i,t)$	MAI on sub-channel i in t of user m
$\gamma_m(i,t)$	SINR on sub-channel i in OFDMA frame $t$of user m
$a_m(t)$	No. of data bits sent in OFDMA frame t for user m

Note that in the LSS case, a client may download more than one segment during a time slot, in order to raise its buffer level to the maximum value $N_{\text{pre}}T_{\text{seg}}$. In this case, the start and end of the download is not necessarily aligned with time slot boundaries, as it is in the LSH use case. This is in contrast to several the related studies assuming that exactly one segment can be downloaded per time slot, thus defeating the purpose of adaptive streaming by preventing the client from dynamically adjusting its buffer level by adapting the video bit rate.

We denote the perceived quality of segment s in representation l in stream m by $Q_{l,s,m}$. We denote its Mean Media Bit Rate (MMBR), which is the size of the segment in bits divided by T_{seg}, by $R_{l,s,m}$. After the representation for segment s is selected, we denote the resulting quality and MMBR by $Q_m(s)$ and $R_m(s)$, respectively.

B. Wireless network model

We consider a single cell, where OFDMA is used both in the downlink and the uplink, and Time Division Duplexing (TDD) is employed as the multiplexing method. In the cell, N_{sc} subchannels are shared between $|M|$ MSs. In the time domain, one OFDMA frame consists of a fixed number of OFDMA symbols. Consequently, one resource block spans over one subchannel in frequency and one OFDMA frame in time. In general, the size of resource blocks is chosen to be smaller than the coherence time and the coherence bandwidth of the underlying fading channel to efficiently exploit channel diversities while minimizing the overhead for resource addresses. As a result, one can assume a flat fading channel for each individual resource block, and the fading processes vary independently between different resource blocks. We use index t to index OFDMA frames. The video time slot s contains OFDMA frames $t \in \{sN_{\text{fis}}, sN_{\text{fis}} + 1, \dots, sN_{\text{fis}} + (N_{\text{fis}} - 1)\}$, where N_{fis} is the number of frames per time slot.

In this work, we explicitly consider the realistic aspect that users' signals tend to be imperfectly synchronized in the uplink, resulting in MAI, which can strongly degrade user throughput [25]. Fortunately, MAI can be mitigated by a proper RA [26]. Let $\iota_m(i,t)$ be the MAI experienced by user m in frame t subchannel i. The detailed derivation of MAI can be found in [27]. Then the Signal to Noise plus Interference Ratio (SINR) on subchannel i, frame t for user m is given by

$$\gamma_m(i,t) = \frac{p_m(i,t)h_m(i,t)}{\sigma^2_{\text{noise}} + \iota_m(i,t)} \tag{1}$$

where σ^2_{noise} denotes the thermal noise, the transmission power $p_m(i,t)$ is assumed to be the same for all users, and channel gain $h_m(i,t)$ includes path loss, shadowing and flat fading.

As most of modern wireless systems, the link adaptation implemented by Adaptive Coding and Modulation (ACM) is used to adapt Modulation and Coding Schemes (MCSs) to the instantaneous SINR subject to a predefined tolerable error rate. To reflect that, a function $F(\gamma_m(i,t))$ returns the number of bits that can be sent according to the selected MCS for the given SINR $\gamma_m(i,t)$. The function $F(.)$ is basically an increasing step function. The total amount of bits transmitted in the frame t is given by:

$$a_m(t) = \sum_{i=0}^{N_{sc}} x_m(i,t) F(\gamma_m(i,t)) \quad , \tag{2}$$

where the binary variable $x_m(i,t)$ indicates whether the resource block (i,t) is allocated to user m or not. Finally, the resource share assigned to user m in video time slot s is given by $G_m(s)$, and the corresponding throughput achieved by user m during the video time slot s is given by $A_m(s)$. We have $A_m(s) = 1/T_{\text{seg}} \sum_t a_m(t)$.

IV. Joint Quality Selection and Resource Allocation

In this section, we first describe the general idea of the proposed solution. A flow chart of the overall system is depicted in Figure 1. We then proceed to describe two key components that comprise our proposed solution: the video quality selection, and the wireless RA. Finally, we propose a system architecture that may be adopted to integrate the solution in existing mobile networks.

A. Cross-layer approach for low-delay live streaming

In this work, we strive to develop an approach that pursues several goals. First, it explicitly considers the QoE perceived by the users as its main performance metric. Concretely, it maximizes the long-term average video quality, while reducing the quality fluctuation to provide a smooth playback experience. Second, it aims at maintaining a high level of fairness by maximizing the minimum QoE among users. Finally, it avoids playback interruptions by considering users' demands and buffer levels while allocating radio resource blocks.

At the beginning of time slot s, for each user m, we select a representation for the newest available segment s. Following [18], we define the user's QID after selecting the representation for segment s as $U_m(s) = \text{mean}(\mathcal{A}_m(s)) - \text{std}(\mathcal{A}_m(s))$. Here $\mathcal{A}_m(s) = \{Q_m(s'), 0 \leq s' \leq s\}$ is the adaptation trajectory for segments $0,\ldots,s$; std is the standard deviation. Note that only $Q_m(s)$ is unknown as computing $U_m(s)$, and in the LSH case, which requires skipping late segments, the quality of a skipped segment s' is set to $Q_m(s') = 0$.

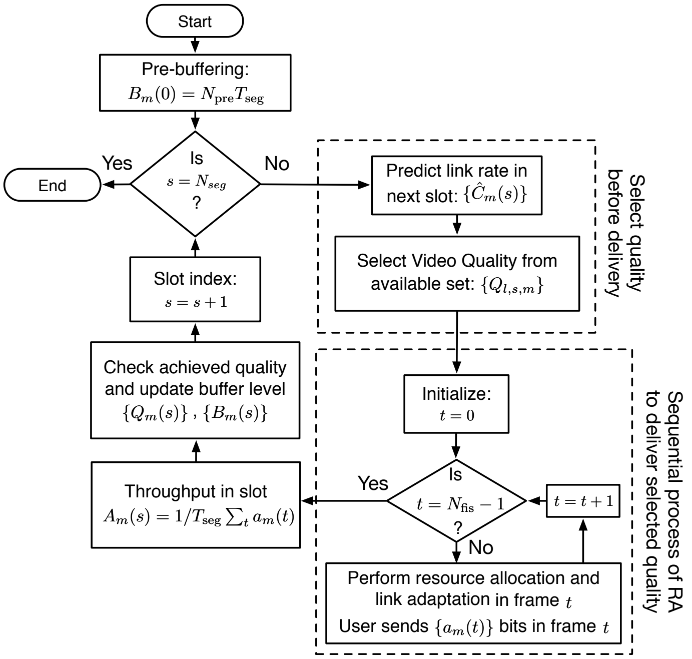
Start
Pre-buffering:
$B_m(0) = N_{\text{pre}}T_{\text{seg}}$
Yes
Is
$s = N_{seg}$
?
No
End
Predict link rate in
next slot: $\{\hat{C}_m(s)\}$
Select Video Quality from
available set: $\{Q_{l,s,m}\}$
Select quality
before delivery
Slot index:
$s = s+1$
Check achieved quality
and update buffer level
$\{Q_m(s)\}$, $\{B_m(s)\}$
Initialize:
$t=0$
Sequential process of RA
to deliver selected quality
Yes
Is
$t = N_{\text{fis}} - 1$
?
No
$t=t+1$
Throughput in slot
$A_m(s) = 1/T_{\text{seg}} \sum_t a_m(t)$
Perform resource allocation and
link adaptation in frame t
User sends $\{a_m(t)\}$ bits in frame t

Fig. 1. Illustration of the proposed cross-layer adaptive streaming approach

We then maximize the minimum QID $U_m(s)$ across all users. This maximization is constrained by the estimated link rates $\{\hat{C}_m(t), m \in M\}$ of the clients. In this section, we assume that link rate estimations are given. We will study the problem of accurately estimating the link rate in Section V.

At the same time, in each scheduling time slot corresponding to one OFDMA frame t with $sN_{\text{fis}} \leq t \leq sN_{\text{fis}} + (N_{\text{fis}} - 1)$, we allocate each client a link rate with the objective to minimize the maximum queue size at the transmitter for this client. To achieve this, we solve an optimization problem to adapt the RA and the ACM to the instantaneous channel quality, in order to gradually meet the throughput demand for the selected video quality. By doing that, the frequency and multiuser diversities can be exploited to provide valuable throughput gain.

Note that while in the LSH case, the transmission queue may contain at most one video segment for each user, in the LSS case, it may contain up to N_{pre} segments as the user tries to ramp up its buffer level after experiencing a buffer underrun.

B. Video Adaptation

We formulate here the problem of video quality adaptation performed at the beginning of the time slot s.

1) Live streaming with hard latency constraints: We first formulate the video quality selection problem for the LSH case. In this work, we target a particularly low delay of two times the segment duration (neglecting the server-side and client-side processing). At the beginning of the streaming session, the client loads one segment into the playback buffer and immediately starts the playback. That is, $B_m(0) = T_{\text{seg}}$. Such a low delay forces the scheduler in the RAN to allocate sufficient resources to each client so that the average throughput in each video adaptation time slot is greater than the MMBR of the segment in the selected representation. If segment s cannot be delivered by its deadline, which is at the end of the time slot s, it is discarded. By targeting this extreme case, we aim to push the limit of the DRA approach.

We introduce the binary optimization variable $z_{l,s,m} \in \{0,1\}$, which takes the value 1 if user m chooses representation l $(0 \leq l < |L|)$ in slot s and 0 otherwise. Then the optimization problem for video quality selection in slot s with hard latency constraints yields the following form.

$$\begin{aligned}
\max_{m \in M} \min \quad & U_m(s) && \text{(LSH)}\\
\text{s.t.} \quad & Q_m(s) = \sum_{l \in L} z_{l,s,m} Q_{l,s,m}\,, \quad \forall m \in M && \text{(C1)}\\
& R_m(s) = \sum_{l \in L} z_{l,s,m} R_{l,s,m}\,, \quad \forall m \in M && \text{(C2)}\\
& \sum_{l \in L} z_{l,s,m} \leq 1\,, \qquad \forall m \in M && \text{(C3)}\\
& R_m(s) \leq \hat{G}_m(s)\hat{C}_m(s)\,, \qquad \forall m \in M && \text{(C4.H)}\\
& \sum_{m \in M} \hat{G}_m(s) \leq 1 && \text{(C5)}
\end{aligned}$$

Here, constraints (C1) and (C2) express the quality and MMBR of the segment s in the selected representation. Constraint (C3) ensures that each user m selects exactly one representation for segment s. Finally, constraints (C4.H) and (C5) represent the capacity constraints. More precisely, constraint (C4.H) ensures that the MMBR of segment s in the selected representation does not exceed the throughput expected in time slot s, given by the link rate estimate $\hat{C}_m(s)$ multiplies by the user's link share $\hat{G}_m(s)$. Constraint (C5) ensures that the total allocated resource blocks does not exceed the total available limit.

It is important to note that, due to the discrete set of perceived qualities and MMBRs of available representations, the problem of video adaptation in general is non-linear, non-convex and, hence, NP-hard [28]), thus there is no efficient approaches available to derive the global optimum of such problems. The formulation in (LSH) alleviates the challenge by exploit the piecewise linearization method to provide a linear programming problem for obtaining an approximated global optimal solution.

2) Live streaming with soft latency constraints: In this case, a playback buffer can be exploited to absorb short-term throughput degradations. Concretely, we allow the MMBR to exceed the estimated throughput when playback buffer level is high enough, so that the segment can still be downloaded prior to its playback deadline. Consequently, constraint (C4.H) in (LSH) transforms to

$$R_m(s) \leq \hat{G}_m(s)\hat{C}_m(s)\Big(1 + \lambda \frac{B_m(s-1)}{T_{seg}}\Big)\,, \forall m \in M\,, \qquad \text{(C4.S)}$$

where $\lambda \in (0,1)$ represents a safety margin to reduce the risk of a buffer underrun.

C. Dynamic Resource Allocation

Next, we exploit DRA to proactively match the long-term user throughputs with the selected video bit rates. Particularly, we consider a series of RA decisions, each of which takes place in one OFDMA frame. We formulate an optimization problem which takes into account instantaneous SINR, achieved throughput in video time slot, and the selected representation. At the output, optimal RA and associated MCSs are given. Especially, we separately consider the problems in the downlink and the uplink. The key difference between them relies in the mitigation of MAI in the uplink. We adopt the max-min formulation to balance between efficiency and fairness.

1) DRA for the downlink: We consider the problem of DRA for the OFDMA frame t, with $(0 \leq t \leq N_{\text{fis}} - 1)$, in the time video time slot s. Let $\tilde{a}_m(t) = \sum_{i=0}^{t-1} a_m(i)$ be the sum of video data (in bits) that user m has sent over all previous OFDMA frames in the time slot s. Then the amount of data delivered in time t is given in (2) is the result of the RA decision. To pursuit the fairness w.r.t. video quality, the achieved throughput sum is further normalized by the amount of video data that has been buffered at transmitter and not yet delivered to receivers. The normalization weight is given by $W_m(s) = \sum_{i=\epsilon_m}^{s} R_m(i) T_{\text{seg}}$, where ϵ_m is the index of the segment will be played-back next. The final optimization problem at time t is given as:

$$\begin{aligned}
\max_{m \in M} \min \quad & \frac{1}{W_m(s)}\Big(\tilde{a}_m(t) + \sum_i x_m(i,t) F(\gamma_m(i,t))\Big) && \text{(DL)}\\
\text{s.t.} \quad & \sum_m x_m(i,t) \leq 1\,, \ \forall i \in \{0, \ldots, N_{\text{sc}} - 1\}\,, && \text{(C6)}
\end{aligned}$$

The constraint (C6) essentially ensures each frequency subchannel can be assigned to at most one user.

2) DRA for the uplink: Unlike the downlink, efficient scheduling algorithm needs to consider the negative impact of MAI on the throughput performance in the uplink. Note that MAI is determined by not only the offsets in time and frequency between imperfectly synchronized users signals, but also the RA schemes. The optimization problem for the uplink is based on our previous work in [26]. The ACM function $F(.)$ is modeled by a piece-wise linear function. K SINR thresholds of MCSs is fixed and denoted by Λ_k for $k = 1, \ldots, K$. As SINR passing higher thresholds, ACM can provide higher throughputs. If SINR $\gamma_m(i,t)$ belongs to the interval $(\Lambda_k, \Lambda_{k+1}]$, then ACM uses MCS that can transmit b_k number of bits, i.e., $F(\gamma_m(i,t)) = b_k$. Furthermore, the ACM selection constraint can be further simplified as follows:

$$\Lambda_k \leq \gamma_m(i,t) \Leftrightarrow \Lambda'_{i,m,k} \geq \frac{\iota_m(i,t)}{\sigma^2} \qquad (3)$$

where $\Lambda'_{i,m,k} = 10^{(p_m(i,t)h_m(i,t)/\sigma^2_{\text{noise}} - \Lambda_k)/10} - 1$. Let $y_m(i,t)$ be the number of bits user m sends on subchannel i, $z_{i,m,k}$ is an indicator variable for the MCS k on subchannel i of user m. Write $\iota_m(i,t) = \sum_{j \neq i} \sum_{m' \neq m} \iota_{i,m}^{j,m'}$ where $\iota_{i,m}^{j,m'}$ is the MAI caused by the subchannel j of user m' on subchannel i of user m. Using these new variables, the problem of DRA in

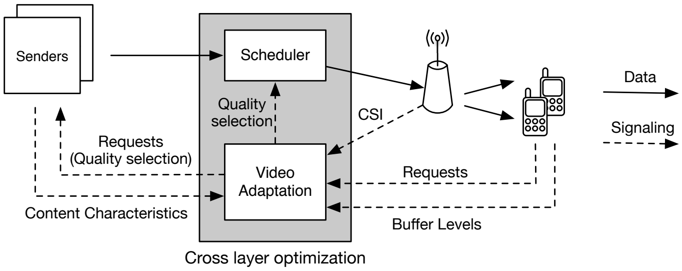

Fig. 2. System architecture for the adaptive streaming in downlink

the uplink frame t can be written as:

$$\max_{m\in M}\min \quad \frac{1}{W_m(s)}\Big(\tilde{a}_m(t)+\sum_i x_m(i,t)y_m(i,t)\Big) \tag{UL}$$

$$\text{s.t.} \quad \sum_m x_m(i,t)\le 1\,,\ \ \forall i\in\{0,\ldots,N_{\text{sc}}-1\}\,, \tag{C6}$$

$$\forall i,m, \sum_k z_{i,m,k}\Lambda'_{i,m,k} \ge \sum_j\sum_{m'}\frac{\iota^{j,m}_{i,m}}{\sigma^2}x_{m'}(j,t) \tag{C7}$$

$$\sum_k z_{i,m,k}\le 1 \tag{C8}$$

$$y_m(i,t)\le \sum_k z_{i,m,k}b_k \tag{C9}$$

Similar to problem (LSH), the problem (UL) is a linear optimization problem thanks to the piece wise linear method. Note that $a_m(t)$ can be written as $a_m(t) = \sum_{i=0}^{N_{\text{sc}}-1} x_m(i,t)y_m(i,t)$ followed by constraints (C7)-(C9). The above formulation is obtained by replacing $a_m(t)$ by this expression in (DL).

D. Proposed system architecture

The presented cross-layer approach can be integrated into existing mobile network technologies via a software upgrade of the schedulers. Currently, state-of-the-art schedulers incorporate RA techniques based on the proportional-fairness (regarding channel quality) for non-adaptive single-rate video streams only [4]. In order to efficiently schedule adaptive multi-rate streams, the proposed solutions assumes that information about the receivers' playback buffer levels and video content characteristics is available at the controller.

1) Streaming in Downlink: The proposed architecture for the downlink is illustrated in Figure 2. First, the network controller, which can be co-located with the scheduler at the BS, collects the buffer levels (in the LSS use case) and users requests in addition to Channel State Information (CSI) measured at the BS or fed back from the MSs. Next it performs a sequence of decisions, which includes (1) predicting the link rate in the next time slot and (2) selecting video qualities subject to constraints on the available resources. Requests for selected video qualities are sent to the scheduler and on behalf of the users to the remote servers. The task of the scheduler is then to adapt RA and ACM to instantaneous channel state within each Orthogonal Frequency-Division Multiple Access (OFDMA) frame in order to satisfy the resulting demands of the users in the cell.

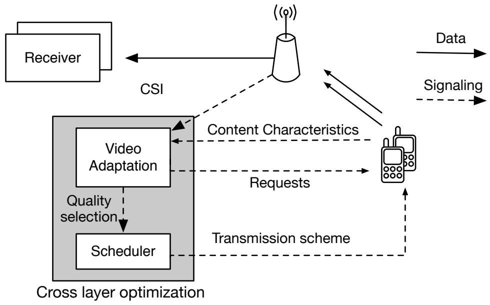

Fig. 3. System architecture for adaptive streaming in uplink

2) Streaming in Uplink: In the uplink, the considered wireless cell is at the sender side. A realization of this use case which is still compatible with the DASH streaming paradigm may use the ServerPush feature of HTML5 [29] to actively push the video segments selecting for each of them the appropriate representation. The proposed architecture is illustrated in Figure 3. The buffer level of the receiver may be reported using the MPEG-DASH reporting functionality.

V. LINK RATE ESTIMATINON

Accurate link rate estimation is crucial to choose appropriate video qualities but very challenging due to, among others, the dynamics of wireless channel. While the underestimation results in lower video quality, the overestimation causes recurrent video re-buffering or skipping video segments. One way to significantly enhance the prediction is to exploit information about wireless link and to estimate the link rate based on the ergodic capacity as used in, for instance, [18], [19], [30]. However, the ergodic capacity expression implicitly assumes that the resources are not adapted to the particular channel state and therefore the throughput is averaged out over all possible channel realizations. This assumption does not hold for DRA. In this work, we show that the ergodic capacity expression can be modified slightly to provide predictions in this case.

A. Ergodic Capacity-based Throughput Prediction

As mentioned above, fading processes on radio resource blocks are independent. Therefore, transmission on each resource block amounts to transmission on an independent channel. The total number of parallel transmission is equal to number of used resource blocks in a given time frame. With sufficiently large number of transmission and no CSI,the average throughput can be approximated well by the ergodic capacity [31]. Let $C^{\text{erg}}_m(s)$ (in [bps]) denote the ergodic capacity that user m achieves with source blocks spanning over N_{sc} subchannels and N_{fis} OFDMA frames within slot s. Without DRA, the SINR within slot s can be considered as a random variable independently realized over each resource block. Denote the SINR random variable by $\gamma_m(s)$. We have

$$\frac{1}{N_{\text{fis}}N_{\text{sc}}}\left\{\sum_{t=0}^{N_{\text{fis}}-1}\sum_{i=0}^{N_{\text{sc}}-1}\log_2(1+\gamma_m(i,t)\right\}\longrightarrow \\ \mathbb{E}_\gamma\Big[\log_2(1+\gamma_m(s))\Big]=C^{\text{erg}}_m(s). \tag{4}$$

The expected value is computed with respect to channel gains. Note that since the number of OFDMA resource blocks within one time slot is very large, roughly 10^5 blocks in 500ms time slot with total bandwidth of 18MHz, the convergence to ergodic capacity occurs even when users do not take all resource blocks. If a user receives a resource share of $\hat{G}_m(s)$ during time slot s, and if the approximated average throughput is equal to $\hat{C}_m(s)$, the throughput can be approximated by $\hat{A}_m(s) = \hat{G}_m(s)\hat{C}_m(s)$. This approach can be adapted to both transmission directions (i.e. downlink without MAI and uplink with MAI). Three methods are considered for evaluating the ergodic capacity or approximating it. First one is Statistical Generation (SG) method. It assumes that the PDF of fading process is available by analyzing channel gains in the previous time slot [22]. Knowing the underlying PDF, ergodic capacity can be computed offline by averaging over sufficient number of channel coefficients. The output of this method is denoted by $C_m^{\mathrm{SG}}(s)$. If the PDF is unknown, one can use running averages over the instantaneous fading realizations. Second method, Low Bound Prediction (LBP) method, used in [19] provides a lower bound on the ergodic capacity. If the average SNR in the last time slot is given by $\bar{\gamma_m}(s-1)$, the LBP method approximates the ergodic capacity by:

$$C_m^{\mathrm{LBP}}(s) = e^{1/\bar{\gamma}_m(s-1)} E_i\left(1, \frac{1}{\bar{\gamma}_m(s-1)}\right), \tag{5}$$

where $E_i(1,x)$ is the exponential integral function defined as $E_i(1,x) = \int_x^\infty \frac{e^{-\tau}}{t} d\tau$. Tight Bound Prediction (TBP), introduced in [23], approximates the ergodic capacity by using bounds on ergodic capacity. The output of this method, C_m^{TBP}, is given by ($\rho \approx 0.57721566$ the Euler's constant):

$$C_m^{\mathrm{TBP}}(s) = \log_2(1 + \bar{\gamma}_m(s-1)(e^{-\rho})). \tag{6}$$

B. Estimation with Dynamic Resource Allocation

As discussed above, the ergodic capacity in (4) is based on static resource allocation assumption. However, DRA allocates resource blocks to users subject to the long-term video quality goal and therefore the throughput on each resource block will be intricately dependent on the CSI of that block. This process increases spectral efficiency but deviates from assumptions behind ergodic capacity. In this section, the previously obtained throughput prediction is adjusted to DRA assumption. By considering the throughput gain achieved by dynamic resource allocation in the last time slot, the reference value to derive the capacity prediction for the next time slot is updated.

The prediction consists of the following steps: 1. The ergodic capacity $C_m^{\mathrm{erg}}(s+1)$ is computed based on the channel states in the previous time slot, namely $\{h_m(i,t)\}$ the channel coefficients in time slot s. The ergodic capacity, however, is an underestimation for average throughput of DRA. It will be modified using the DRA gain of the previous block to improve the prediction.

2. The spectral efficiency gain $\rho_m(s)$ (in b/s/Hz) achieved by DRA is computed using the achieved throughput compared to the predicted value in the last time slot s. We have

$$\rho_m(s) = \frac{A_m(s)}{G_m(s)} - \hat{C}_m(s) \tag{7}$$

where $A_m(s)$ is the actual accumulated throughput for user m using $G_m(s)$ resource blocks in time slot s.

3. The spectral efficiency in the next time slot is predicted as:

$$\hat{C}_m(s+1) = C_m^{\mathrm{erg}}(s+1) + \beta\rho_m(s) \tag{8}$$

where β is a coefficient to scale the impact of previous gain $\rho_m(s)$ on the expectation.

Based on the predicted $\hat{C}_m(s+1)$, the accumulated throughput is obtained as:

$$\hat{A}_m(s+1) = \hat{C}_m(s+1)\hat{G}_m(s+1) \tag{9}$$

Note that $\hat{G}_m(s+1)$ is found by solving (LSH).

VI. Evaluation

We evaluate the proposed algorithms through Monte Carlo simulations based on the network simulator OMNeT++ with realistic models of wireless link and a video trace file.

A. Simulation setup

1) Wireless Link: We consider a cell, where 4 users share 16 sub-channels. In time domain, one OFDMA frame spans over 47 OFDM symbols. Users move in cell at the speed of 50km/h following the Manhattan mobility grid model. The propagation channel consists of either Light of Sight (LOS) or non-LOS (NLOS), with the distance following the probability density function $\mathbb{P}_{\mathrm{NLOS}}(d) = 0.9(1-(1.24-0.6\log(d)^3)^{1/3})$. Path loss models for LOS and NLOS cases are based on COST-231 channel model. We model shadowing loss that includes spatial correlation based on the MOSAIC model in [32]. Finally, unlike most of other work, we explicitly consider the fact that the fading process within a long duration of video chunk is most likely non-stationary. To do that, we adopt the model of piece-wise stationary fading processes, each of which is reasonably assumed to last for 100 ms [33]. The time varying statistics of these stationary channels are derived from the highly detailed QuaDRIGA simulator [11]. Moreover the imperfect synchronization between users signals are considered in uplink transmission. Specifically, users signals yield a small residual frequency offsets, which are equally distributed from 0 to 30Hz (due to, e.g. oscillator inaccuracy and the Doppler effect), while perfectly synchronized in time. In addition, the specification of MCSs in LTE are simulated for the link adaptation.

2) Link Rate Prediction: We evaluate the prediction accuracy of three approaches, which are SG, LBP and TBP, in downlink and uplink with realistic assumptions including non-stationary channels and imperfect synchronization in uplink. In this section, each user is assumed to demand a throughput equal to$\hat{A}_m(s) = \hat{C}_m(s)/M$ in the next slot. To achieve that goals, we first adopt the SRA approach, where users take equal resource share (i.e. $G_m(s) = 1/M, \forall m$). Afterward, DRA is chosen, where the link rate is set as the target rate, i.e., we set $W_m(s) = \hat{C}_m(s)$ in (DL) for downlink and (UL) for uplink. The expected throughput at the beginning of slot $\{\hat{A}_m(s)\}$ is compared with the throughput achieved by using either SRA or DRA at the end of slot $A_m(s)$. In this work, the performance metrics is the relative prediction errors defined as $100(A_m(s) - \hat{A}_m(s))/\hat{A}_m(s)$ (in percentage). We perform total 30 simulation runs of 4 users in 60 seconds.

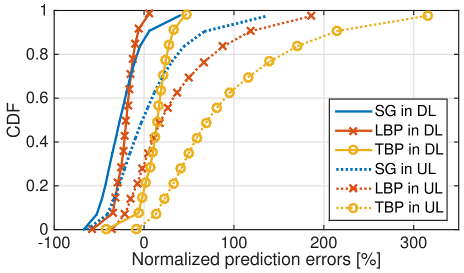

Fig. 4. Performance of prediction methods

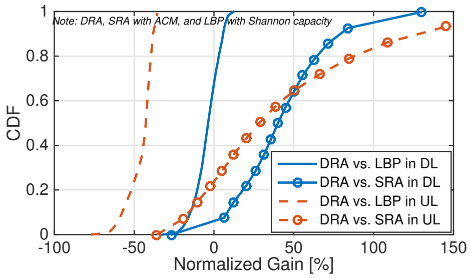

Fig. 5. Throughput performance gain by DRA

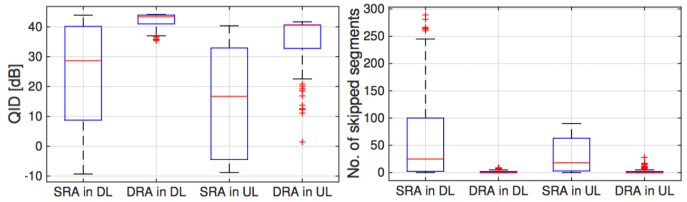

Fig. 6. Video performance in case of LSH

3) Video Traffic: We use the trace file of the movie The silence of the lambs provided by Arizona State University [34] to feed the simulation. In the chosen trace, video content is encoded 9 times separately; the encoding scheme is MPEG-4 single layer (non-scalable) with the format G16B15. For convenience, size of video chunk is set to equal one GOP of one second. It means there are total 200 OFDMA frames within each video chunk. To take into account the heterogeneity among users video contents, users video sequences are extracted from the common trace but from different starting points. In case of LSH, video streams have the same size of 300s in downlink and 90s in uplink, and for LSS simulation terminates only when a minimum number of video chunks are played back, which are 300s in downlink and 90s in uplink.

B. Evaluation of Link Rate Prediction

We show the numerical result for SRA cases in Figure 4, where the solid lines correspond to downlink and dash lines to uplink. Note that continuous Shannon rate is used here instead of discrete ACM rate for computing user throughput. The achieved throughput is this comparable to the prediction model. As it is shown, in general, TBP provides mostly overestimated values. LBP is the most conservative one since the expected values rarely overshoot the channel capacity. Finally SG tends to give both under- and over-estimation. In addition, all methods lead to overestimation in uplink due to the lack of consideration of MAI. Based on this result, the LBP seems to be more appropriate choice for the SRA approaches in order to avoid video stalls.

Next, the results for DRA approaches are shown in comparison with the predicted throughput using LBP method and achieved throughput through SRA. Note that the actual user throughput is calculated using a realistic model of link adaptation for both SRA and DRA, as opposed the prediction step where the continuous Shannon rate is considered. The results shown in Figure 5 illustrate the large improvement of DRA compared to SRA. From this figure, one can also see that throughput achievement by DRA tends to fluctuate around the predicted value in downlink and exceed 50% of that in uplink, respectively.

C. Evaluation of Video Performance

We then evaluate the video performance of the proposed algorithms and the base line, which uses SRA and no exploitation of DRA and buffer levels. First, for the LSH, consistent with the results from the previous section, link rates of incoming time slot are estimated equal 60% and 40% of the lower bound of ergodic capacity (i.e. $\hat{C}_m^{LBP}$) for downlink and uplink, respectively. In addition, after some preliminary investigation, the coefficient β in (8), used to scale the impact of DRA gain, is chosen as $\beta = 0.2$. We present the QoE index (QID), which is computed by function $U_m()$ for all segments of streamsand number of skipped segments in Figure 6. As it can be seen, DRA can greatly improve the QoE by not only increasing the median of QID, by 10dB in downlink and more than 20dB in uplink, but also strongly mitigates number of video stalls for both downlink and uplink cases.

Regarding the LSS cases, predicted link rates are computed equal 60% of $\hat{C}_m^{LBP}$ for SRA in downlink and uplink; these ratios are 80% and 60% used for DRA in downlink and uplink. The parameters of proposed algorithms are selected as follows $\beta = 0.2$, $\lambda = 0.1$ for downlink, and $\beta = 0.2$, $\lambda = 0.1$ for uplink. We show the overall number and duration of video stalls in cell in addition to QID in Figure 7. It can be seen that the performance gains by DRA are less obvious due to the deployment of buffer to absorb the throughput fluctuation. Particularly, DRA increases the median QID by more than 2dB and 0.5dB in downlink and uplink, respectively. Meanwhile, the advantage of DRA can be easily noticed as the proposed approaches can efficiently avoid the video stalls w.r.t. both number of events as well as duration.

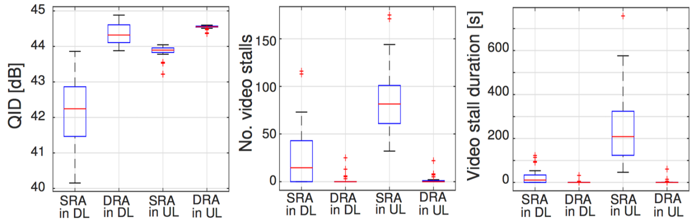

Fig. 7. Video performance in case of LSS

VII. Conclusion

We propose novel cross-layer approaches to enhance the QoE of multiple adaptive live streams in a mobile cell. The proposed solutions consist of two key components: video quality selection, and dynamic resource allocation. We address two use cases: low-delay streaming with hard latency constraints and with soft latency constraints. We separately consider downlink and uplink transmissions, where imperfect synchronization in uplink distinguishes these two cases. Through DRA, the multiuser and frequency diversities can be efficiently exploited and, at the same time, multiple access interference in uplink can be mitigated in order to provide valuable throughput gains, and resources can be allocated more efficiently to users who benefits most from them. As a separate contribution, we evaluate several link rate estimation methods revealing significant throughput gains due to DRA. Simulation results demonstrate significant QoE gains for all considered use cases.

References

[1] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.

[2] P. Sweeting, "Video in 2014: Going Live and Over the Top," Gigaom, Tech. Rep., 2014.

[3] M. Andrews, *A Survey of Scheduling Theory in Wireless Data Networks*. New York, NY, USA: Springer New York, 2007, pp. 1–17.

[4] "Improved support for dynamic adaptive streaming over http in 3gpp," 3GPP, Tech. Rep. TR 26.938 v1.6.0, 2014.

[5] M. Van Der Schaar and S. Shankar N, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, Aug. 2005.

[6] O. Oyman, J. Foerster, Y.-j. Tcha, and S.-c. Lee, "Toward enhanced mobile video services over WiMAX and LTE [WiMAX/LTE Update]," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 68–76, Aug. 2010.

[7] X. Xie, X. Zhang, S. Kumar, and L. E. Li, "pistream: Physical layer informed adaptive video streaming over lte," in *Proc. 21st Annu. Int. Conf. on Mobile Computing and Networking*. New York, NY, USA: ACM, 2015, pp. 413–425.

[8] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proc. 19th Annu. Int. Conf. on Mobile Computing and Networking (MobiCom '13)*. New York, NY, USA: ACM, 2013, pp. 389–400.

[9] X. K. Zou, J. Erman, V. Gopalakrishnan, E. Halepovic, R. Jana, X. Jin, J. Rexford, and R. K. Sinha, "Can Accurate Predictions Improve Video Streaming in Cellular Networks?" in *Proc. 16th Int. Work. on Mobile Computing Systtem and Application (HotMobile '15)*. New York, NY, USA: ACM Press, Feb. 2015, pp. 57–62.

[10] A. Varga and R. Hornig, "An overview of the omnet++ simulation environment," in *Proc. 1st Int. Conf. on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops*, ser. Simutools '08. Brussels, Belgium: ICST, 2008, pp. 60:1–60:10.

[11] K. Börner, J. Dommel, S. Jaeckel, and L. Thiele, "On the requirements for quasi-deterministic radio channel models for heterogeneous networks," in *Proc. 2012 International Symposium on Signals, Systems, and Electronics (ISSSE)*, Oct. 2012, pp. 1–5.

[12] X. Yin, V. Sekar, and B. Sinopoli, "Toward a Principled Framework to Design Dynamic Adaptive Streaming Algorithms over HTTP," in *Proc. 13th ACM Work. on Hot Topics in Networks (HotNets)*, 2014.

[13] A. Bokani, M. Hassan, and S. Kanhere, "HTTP-Based Adaptive Streaming for Mobile Clients using Markov Decision Process," in *Proc. 20th Int. Packet Video Workshop*. IEEE, Dec. 2013, pp. 1–8.

[14] M. Claeys, S. Latre, J. Famaey, and F. D. Turck, "Design and evaluation of a self-learning HTTP adaptive video streaming client," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 716–719, 2014.

[15] K. Miller, E. Quacchio, G. Gennari, and A. Wolisz, "Adaptation Algorithm for Adaptive Streaming over HTTP," in *Proc. 19th Intl. Packet Video Workshop*, 2012, pp. 173–178.

[16] K. Miller, A.-K. Al-Tamimi, and A. Wolisz, "QoE-Based Low-Delay Live Streaming Using Throughput Predictions," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, pp. 1–24, 2016.

[17] G. Cofano, L. De Cicco, T. Zinner, A. Nguyen-Ngoc, P. Tran-Gia, and S. Mascolo, "Design and experimental evaluation of network-assisted strategies for http adaptive streaming," in *Proc. 7th Int. Conf. on Multimedia Systems (MMSys '16)*. New York, NY, USA: ACM, 2016.

[18] V. Joseph and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks," in *IEEE INFOCOM 2014 - IEEE Conf. Comput. Commun.* IEEE, Apr. 2014, pp. 82–90.

[19] D. Bethanabhotla, G. Caire, and M. Neely, "Adaptive Video Streaming for Wireless Networks with Multiple Users and Helpers," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 1–1, 2014.

[20] F. Fu and M. Der Schaar, "A systematic framework for dynamically optimizing multi-user wireless video transmission," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 308–320, Apr. 2010.

[21] Q. Xu, S. Mehrotra, Z. Mao, and J. Li, "PROTEUS: Network Performance Forecast for Real-Time, Interactive Mobile Applications," in *Proc. 11th Annu. Int. Conf. Mobile Systems, Applications, and Services (MobiSys '13)*, 2013, p. 347.

[22] E. Yaacoub and Z. Dawy, "A Survey on Uplink Resource Allocation in OFDMA Wireless Networks," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 2, pp. 322–337, 2012.

[23] O. Oyman, R. Nabar, H. Bolcskei, and A. Paulraj, "Tight lower bounds on the ergodic capacity of rayleigh fading mimo channels," in *Proc. IEEE Global Telecommunication Conference (GLOBECOM '02)*, vol. 2, Nov. 2002, pp. 1172–1176.

[24] "MPEG-DASH (ISO/IEC 23009-1)," Moving Picture Experts Group, Tech. Rep., 2012.

[25] H. Le, A. Behboodi, and A. Wolisz, "Dynamic resource allocation in ofdma uplink for mai mitigation and throughput improvement," in *Proc. 80th Vehicular Technology Conference (VTC2014-Fall)*, Sep. 2014, pp. 1–5.

[26] ——, "Quality Driven Resource Allocation for Adaptive Video Streaming in OFDMA Uplink," in *Proc. 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 1277–1282.

[27] A. M. Tonello, N. Laurenti, and S. Pupolin, "Analysis of the uplink of an asynchronous multi-user DMT OFDMA system impaired by time offsets, frequency offsets, and multi-path fading," in *Proc. 52nd Vehicular Technology Conference (VTC2000-Fall)*, vol. 3. IEEE, 2000, pp. 1094–1099.

[28] A. Seetharam, P. Dutta, V. Arya, J. Kurose, M. Chetlur, and S. Kalyanaraman, "On Managing Quality of Experience of Multiple Video Streams in Wireless Networks," *IEEE Trans. Mob. Comput.*, vol. 14, no. 3, pp. 619–631, Mar. 2015.

[29] W3C, "Server-Sent Events," W3C, Tech. Rep., 2015.

[30] A. El Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-Based Traffic and Resource Management for Adaptive HTTP Video Delivery in LTE," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 6, pp. 988–1001, Jun. 2015.

[31] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[32] D. Kitchener, M. Naden, W. Tong, and P. Zhu, "Correlated Lognormal Shadowing Model," IEEE, Tech. Rep. IEEE C802.16j-06/059, 2006.

[33] A. Duel-Hallen, Shengquan Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 62–75, May. 2000.

[34] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 4, pp. 1142–1165, 2012.