

Anticipatory Radio Resource Management for Mobile Video Streaming with Linear Programming

Dimitrios Tsilimantos, Amaya Nogales-Gómez, and Stefan Valentin

Mathematical and Algorithmic Sciences Lab,

FRC, Huawei Technologies

{dimitrios.tsilimantos, amaya.nogales.gomez, stefan.valentin}@huawei.com

Abstract—In anticipatory networking, channel prediction is used to improve communication performance. This paper describes a new approach for allocating resources to video streaming traffic while accounting for quality of service. The proposed method is based on integrating a model of the user’s local play-out buffer into the radio access network. The linearity of this model allows to formulate a Linear Programming problem that optimizes the trade-off between the allocated resources and the stalling time of the media stream. Our simulation results demonstrate the full power of anticipatory optimization in a simple, yet representative, scenario. Compared to instantaneous adaptation, our anticipatory solution shows impressive gains in spectral efficiency and stalling duration at feasible computation time while being robust against prediction errors.

I. INTRODUCTION

Video streaming generated 45% of all mobile data traffic in 2014 and is predicted to increase to 62% by 2019 [1]. Although much effort has been spent to increase the capacity of mobile networks, it is still a major challenge for operators to assure sufficient streaming quality for mobile users. The technical difficulty here is to provide high resolution and fluency while reaching high Spectral Efficiency (SE) with a time-variant wireless channel. Keeping the user’s video play-out buffer filled at high rate, is often not possible or inefficient in difficult coverage situations, with interference, or at high speed. Once the play-out buffer runs empty, the video stalls, and the user’s experience is heavily reduced. There is now a wide consensus that stalls are a major cause of dissatisfaction for the users of mobile streaming services [2]–[4].

A. Idea and Contributions

In this paper, we address this problem by Anticipatory Radio Resource Management (ARRM). Based on the knowledge of future channel states, the Base Station (BS) allocates wireless channel resources over upcoming time slots in order to fill the user’s play-out buffer before a poorly covered area is reached. While moving through this area, the user perceives a fluent video stream from the preemptively filled buffer without using wireless channel resources. At the same time, these resources can be allocated to users with higher channel state. This increases overall spectral efficiency by reaching multi-user diversity gains, while satisfying the minimum bitrate requirement for streaming from the user’s local memory but not from the channel.

Following this concept of ARRM, this paper contributes a model to predict the state of the user’s play-out buffer at the BS. The linearity of this model allows to formulate two Linear Programming (LP) problems for ARRM. The first formulation maximizes spectral efficiency while avoiding stalls but becomes infeasible if the wireless channel state is too low to prevent a buffer under-run. Our second formulation avoids this problem by trading off stalling duration and SE. A detailed performance study shows that this formulation achieves outstanding SE gains at high QoS. These gains are reached at feasible computational time and decrease only slightly if channel prediction errors are taken into account.

The focus of this paper is entirely on the RRM for the final hop in mobile streaming. Aspects related to bottlenecks in the backbone and to the optimization of content storage (e.g., caching, CDN replication) are not considered in this work. Nonetheless, our model covers adaptive streaming, e.g., [5], [6], by including a time-variant traffic rate in the optimization.

B. Related Work

The anticipatory, or proactive, allocation of wireless channel resources is typically used to compensate for delayed channel state information for a small number of upcoming transmission times [7], [8]. Operating close to the coherence time, these schemes predict small-scale fading in the millisecond regime. Such short-term prediction is inapplicable for video streaming. This is a consequence of the relatively large segment size of common HTTP Adaptive Streaming Protocols, e.g., [6]. To transfer a single segment, current cellular networks [9] may require hundreds of milliseconds transmission time. Thus, ARRM operates at a time scale where small-scale fading averages out and propagation loss dominates, which becomes time-variant with user mobility. As a result, channel prediction at this time scale is often based on combining the prediction of user trajectory with coverage data [10], [11]. Based on analyzing coverage maps, a linear model for the prediction error was presented in [12], which will be used in Section III.

Based on such long-term prediction of the wireless channel, only few authors have studied the anticipatory resource allocation for media streaming so far. This paper extends [13] by including initial buffer states, avoiding infeasible cases by trading off SE with QoS, and by studying robustness to prediction errors. In [14] an upper bound for ARRM was presented but only for the single-user case and without

an operational formulation of the problem. Our paper goes beyond this work by presenting two tractable formulations for the multiple user case together with a concise study. Recently, a RRM framework was proposed in [15] which maximizes a non-linear Utility function by a greedy algorithm. Besides providing a linear formulation, our work exploits anticipation for the allocation of resources.

C. Paper Structure

The remainder of the paper is organized as follows. In Section II, we describe the resource allocation model for the multi-user case and discuss the proposed optimization problems. We present our simulation results in Section III and conclude with our remarks in Section IV.

II. A MATHEMATICAL MODEL FOR ANTICIPATORY RRM

In this paper we consider the downlink of a multi-user Orthogonal Frequency Division Multiple Access (OFDMA) system when the average channel gain is predicted for the video users over the next T time slots, the prediction horizon. The OFDMA system is widely used in different standards such as in Long Term Evolution (LTE), [9]. In such a system, the bandwidth is divided into N Physical Resource Blocks (PRBs), each one with a bandwidth B which can be assigned to the different users. Let us define the set of users, time slots in the prediction horizon and BSs as $\mathcal{K} = \{1, \dots, K\}$, $\mathcal{T} = \{1, \dots, T\}$ and $\mathcal{M} = \{1, \dots, M\}$ respectively.

A. Parameters and Variables

We consider the following input parameters and variables:

- Achievable data rate per PRB with $k \in \mathcal{K}, t \in \mathcal{T}$:

$$S_{k,t} = B \log_2 \left(1 + \frac{P|\hat{h}_{k,t}|^2}{\Gamma(\sigma^2 + I)} \right),$$

where P is the transmit power, σ^2 and I is the noise and the interference respectively, $|\hat{h}_{k,t}|^2$ is the predicted channel gain and Γ is the Signal to Interference plus Noise Ratio (SINR) gap that accounts for the bit error rate in practical modulations.

- Required play-out video bit rate $V_{k,t}$ with $k \in \mathcal{K}, t \in \mathcal{T}$. Note that the time index covers adaptive streaming, e.g., [5], [6]. We assume this value to be obtained by cross-layer signaling or traffic profiling.
- Base station assignment parameter with $k \in \mathcal{K}, t \in \mathcal{T}$:

$$a_{k,t}^m = \begin{cases} 1 & \text{if user } k \text{ is at time slot } t \text{ in BS } m \\ 0 & \text{otherwise} \end{cases}$$

In the following, we use the upper index d for $V_{k,t}$ and $S_{k,t}$ when we want to refer to bits instead of rates, i.e. $V_{k,t}^d = V_{k,t}T_d$ and $S_{k,t}^d = S_{k,t}T_d$ respectively, where T_d is the duration of one time slot. Moreover, we consider the following variables that characterize the problem:

- Fraction of assigned resources $\omega_{k,t} \in \mathbb{R}^+$ with $k \in \mathcal{K}, t \in \mathcal{T}$ that represent the proportion of overall PRBs assigned to user k at time slot t .

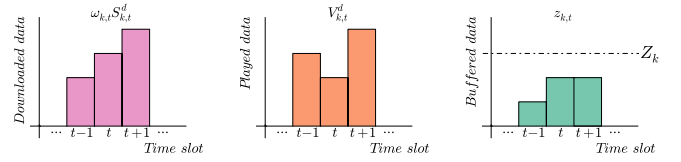


Fig. 1. Model for the user's play-out buffer, illustrated for various time slots. The buffer state at the end of time slot t is the result of the remaining buffered data from previous time slot (right) plus the downloaded data (left) minus the played data (center).

- Buffer state with $k \in \mathcal{K}, t \in \{0\} \cup \mathcal{T}$:

$$z_{k,0} = \zeta_k, \quad 0 \leq z_{k,t} \leq Z_k,$$

where $z_{k,t}$ is the remaining data for user k at the end of time slot t , Z_k is the buffer size and ζ_k is the initial buffered data.

- Stalling time $\ell_{k,t}$ with $k \in \mathcal{K}, t \in \mathcal{T}$, which represents the fraction of time slot t for which user k did not receive the required play-out data.

B. LP Formulation Without Stalls

The proposed model is illustrated in Fig. 1, which shows the evolution of the buffer state over the time slots. In general, each user k requires $V_{k,t}^d$ bits per time slot t in the buffer to play a video. At any time slot for a given user, there are $\omega_{k,t}S_{k,t}^d + z_{k,t-1}$ bits in the buffer, and a stall will take place if this amount is less than $V_{k,t}^d$. On the contrary, if $\omega_{k,t}S_{k,t}^d + z_{k,t-1} \geq V_{k,t}^d$, the unused bits are carried over to the next time slot. To avoid stalls, we formulate the following constraints for the proposed buffer model:

$$\begin{aligned} z_{k,1} &= \omega_{k,1}S_{k,1}^d + \zeta_k - V_{k,1}^d \\ z_{k,2} &= \omega_{k,2}S_{k,2}^d + z_{k,1} - V_{k,2}^d = \sum_{t=1}^2 (\omega_{k,t}S_{k,t}^d - V_{k,t}^d) + \zeta_k \\ &\vdots \\ z_{k,T} &= \sum_{t=1}^T (\omega_{k,t}S_{k,t}^d - V_{k,t}^d) + \zeta_k, \end{aligned}$$

which can be reduced to the constraint $\forall k \in \mathcal{K}, \forall t \in \mathcal{T}$:

$$z_{k,t} = \sum_{i=1}^t (\omega_{k,i}S_{k,i}^d - V_{k,i}^d) + \zeta_k, \quad z_{k,t}, \omega_{k,t} \in \mathbb{R}^+. \quad (1)$$

Based on this linear buffer model, we can formulate the previously-stated resource allocation problem for multiple video users as

$$\min_{\omega} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} \omega_{k,t} \quad (2a)$$

$$\text{s.t.} \quad 0 \leq \sum_{i=1}^t (\omega_{k,i}S_{k,i}^d - V_{k,i}^d) + \zeta_k \leq Z_k \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \quad (2b)$$

$$\sum_{k \in \mathcal{K}} \omega_{k,t} a_{k,t}^m \leq N_m \quad \forall t \in \mathcal{T}, \forall m \in \mathcal{M}, \quad (2c)$$

$$\omega_{k,t} \in \mathbb{R}^+ \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \quad (2d)$$

where the objective function is chosen to minimize the total allocated resources, subject to the constraints for the buffer

level and the available resources per BS. Since both the objective function and the constraints are linear, our formulation represents an LP, which can be solved with conventional optimization software [16] in polynomial time on the average.

C. LP Formulation With Stalls

The previous optimization problem takes advantage of channel state prediction in order to maximize spectral efficiency while perfectly avoiding stalls. However, with consistently poor radio coverage, stalls will be unavoidable and problem (2a)-(2d) will become infeasible. In this section, we propose a variant of the model described in Section II-B where stalls are allowed, but stalling time is minimized. First, we formulate the following constraints for the proposed buffer model:

$$z_{k,1} = V_{k,1}^d \ell_{k,1} + \omega_{k,1} S_{k,1}^d + \zeta_k - V_{k,1}^d \quad (3a)$$

$$z_{k,2} = V_{k,2}^d \ell_{k,2} + \omega_{k,2} S_{k,2}^d + z_{k,1} - V_{k,2}^d \quad (3b)$$

$$\vdots \quad (3c)$$

$$z_{k,T} = \sum_{t=1}^T (V_{k,t}^d \ell_{k,t} + \omega_{k,t} S_{k,t}^d - V_{k,t}^d) + \zeta_k, \quad (3c)$$

which can be reduced to the constraint $\forall k \in \mathcal{K}, \forall t \in \mathcal{T}$:

$$z_{k,t} = \sum_{i=1}^t (V_{k,i}^d \ell_{k,i} + \omega_{k,i} S_{k,i}^d - V_{k,i}^d) + \zeta_k, \\ z_{k,t}, \omega_{k,t}, \ell_{k,t} \in \mathbb{R}^+.$$

In the previous model (2a)-(2d), $\omega_{k,t} S_{k,t}^d + z_{k,t-1} < V_{k,t}^d$ implies that the model is infeasible since it leads to $z_{k,t} < 0$. Under the same circumstances, the new buffer model, however, gives $z_{k,t} = 0$ and, thus, leads to a feasible solution. In addition, it yields a positive value for the stalling time $\ell_{k,t}$. On the contrary, in the case that $\omega_{k,t} S_{k,t}^d + z_{k,t-1} > V_{k,t}^d$, the buffer model (3a)-(3c) could lead to unrealistic solutions where $z_{k,t} > 0$ and $\ell_{k,t} > 0$. We can avoid such cases by minimizing the stalling time in the objective function, which leads to the following solutions $\forall k \in \mathcal{K}, \forall t \in \mathcal{T}$:

$$z_{k,t} = \max(\omega_{k,t} S_{k,t}^d + z_{k,t-1} - V_{k,t}^d, 0), \\ \ell_{k,t} = \frac{1}{V_{k,t}^d} \max(-\omega_{k,t} S_{k,t}^d - z_{k,t-1} + V_{k,t}^d, 0).$$

Finally, we choose the objective function of this new model in order to minimize a trade-off between the total allocated resources and the stalling time. Controlling this trade-off requires to introduce the free parameter $\gamma \in \mathbb{R}^+$, where higher values of γ prioritize stalling time minimization. Thus, we can formulate the previously stated resource allocation problem for multiple video users as the LP problem:

$$\min_{\omega, z, \ell} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} (\omega_{k,t} + \gamma \ell_{k,t}) \quad (4a)$$

s.t.

$$\sum_{i=1}^t (V_{k,i}^d \ell_{k,i} + \omega_{k,i} S_{k,i}^d - V_{k,i}^d) + \zeta_k \geq 0 \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \quad (4b)$$

$$\sum_{i=1}^t (V_{k,i}^d \ell_{k,i} + \omega_{k,i} S_{k,i}^d - V_{k,i}^d) + \zeta_k \leq Z_k \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \quad (4c)$$

$$\sum_{k \in \mathcal{K}} \omega_{k,t} a_{k,t}^m \leq N_m \quad \forall t \in \mathcal{T}, \forall m \in \mathcal{M}, \quad (4d)$$

$$\omega_{k,t}, \ell_{k,t} \in \mathbb{R}^+ \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}. \quad (4e)$$

III. NUMERICAL RESULTS

We begin this section by discussing the performance metrics and simulation assumptions for the study of different resource allocation schemes. Then, we provide numerical results for the LP formulation (4a)-(4e) for multiple users with video streaming traffic.

A. Performance Metrics

We focus on the following three performance metrics:

- 1) *Cell spectral efficiency* (bits/s/Hz/cell): This critical measure for wireless networks is commonly defined as the data rate that the BS transmits over a given bandwidth, divided by the number of cells. According to our mathematical model, SE is given by

$$SE = \frac{\sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}_S} \omega_{k,t} S_{k,t}^d}{MB \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}_S} \omega_{k,t}}, \quad (5)$$

where $\mathcal{T}_S = \{1, \dots, T_S\}$ is the set of T_S time slots until K users are served.

- 2) *Stalling duration* (s): Since user mobility leads to time-variant channel gains, the allocated data rate may not always support the traffic rate of the video stream. If, in this case, the user's play-out buffer runs empty, the video stream stalls. Since stalling duration is recognized as a main factor for QoS [2]–[4], we assign higher priority to the stalling duration than to the allocated resources. In mathematical terms, the parameter γ in (4a) is chosen high enough to ensure that the solution of the optimization gives the minimum resources for the minimum feasible stalling time.
- 3) *Computational time* (s): Since RRM has to adapt to the state of channel and traffic in real time, the optimization problem has to be solved sufficiently fast. However, it is worth mentioning that the optimization does not have to be performed for each Transmission Time Interval (TTI). This results from the fact that HTTP-based streaming protocols separates the video stream into segments, each containing several seconds of video time. As only complete segments can be played, which requires typically hundreds of milliseconds transmit time per segment, it is sufficient to choose a slot duration T_d at this time scale.

B. Simulation Scenario and Parameters

A simple two-cell scenario is assumed throughout the remainder of the paper. Video users arrive at the system following a Poisson arrival process with rate $\lambda = \frac{K}{T_N T_d}$, where $T_N \leq T_S$ is the number of time slots that the user stays connected. Each user moves with constant speed in a straight line from the first BS to the second one and requests video streaming, approximating the situation for vehicular users in a highway. The chosen inter-site distance corresponds to an typical LTE deployment in urban areas. Note that this simple scenario already captures the main idea of anticipatory resource allocation, since the channel gain in the cell edge

between the two BSs can likely lead to stalls if multiple users are sharing the BS resources at the same time. By predicting the channel gain, the user's play-out buffer can be filled in advance under better channel conditions and be consumed in the cell edge, where the channel gain is low.

To account for the channel, we adopt the 3GPP path-loss model $PL = 128.1 + 37.6 \log_{10} d + L_s$ [17], where d (km) is the distance between the user and the serving BS, which we consider to be the nearest one, and L_s is the shadowing factor. For the sake of simplicity, we do not explicitly calculate interference and only a margin that includes both intra- and inter-cell interference is introduced. The SINR gap between the achieved spectral efficiency and the Shannon channel capacity is modeled as a simple function of Bit Error Rate (BER), as in [18]. Fast fading is not taken into account, as we assume that it is averaged out over hundreds of milliseconds of a time slot. A new optimization is performed either when a new user arrives at the system or after T_c time slots, where $T_c \leq T$ is a fixed optimization step. This way, we keep the most recent results $\omega_{k,t}$ for each time slot and user, an approach which proves to have an important impact on the results for a specific range of values of T .

Monte-Carlo simulations are performed and the average values are evaluated over 1000 iterations. A summary of the main simulation parameters are presented in Table I, where the slot duration is $T_d = 167$ ms given the respective values of user speed, inter-site distance and number of time slots T_N . Finally, the video traffic rates $V = 1.5$ Mbits/s and $V = 6$ Mbits/s are chosen, which are average rates for 720p and 2k HD streaming. The rest of the values of V are used as intermediate levels. These values may seem high but are realistic for LTE metro-cells with a small number of video users. Our selection was confirmed by field measurements in two larger European cities and comes at no loss in generality. For simplicity, a common value V is used for all users and time slots, although our formulation can handle varying traffic bitrates as well.

C. Model of Channel Prediction Errors

Perfect channel prediction was assumed to be available so far in order to formulate the optimization problem in the previous sections. However, this is not the case in practical systems even with the most advanced predictors. Here we introduce a simple model for the channel prediction error $\epsilon_{k,t}$ that allows us to examine the robustness of our proposed algorithm. The predicted channel gain is then given by $|h_{k,t}|^2 = |h_{k,t}|^2 + \epsilon_{k,t}$ in dB scale, where $\epsilon_{k,t}$ is assumed to follow for all users a Normal distribution with zero mean and standard deviation that is linear to the prediction horizon, i.e. $\epsilon_{k,t} \sim N(0, \sigma_t^2)$, where $\sigma_t = \frac{t}{T} \sigma$ and σ^2 is the variance of the error for a prediction horizon of T . Although simple, this model incorporates the fact that the error increases with the prediction horizon and approximates well the linear gradient found in [12]. We leave more sophisticated analysis of prediction errors for our future work.

TABLE I
MAIN SIMULATION PARAMETERS.

Parameter	Value
Total BS Tx power P	46 dBm
BS antenna gain	18 dBi
Available PRBs in a BS	50
PRB bandwidth B	180 kHz
Noise spectral density	-174 dBm/Hz
Receiver noise figure	10 dB
Interference margin	6 dB
Shadow fading margin	10 dB
SINR gap Γ	$-\ln(5\text{BER})/1.5$ [18]
BS inter-site distance	500 m
Total number of time slots T_N	100
User speed	30 m/s
BS antenna height	35 m
Path-loss model	3GPP empirical [17]
Number of users K	[1, 30]
Prediction horizon T	[1, T_N] time slots
Maximum buffer size Z_k	20 Mbits
Video play-out rate V	{1.5, 2.5, 4, 6} Mbits/s
Optimization step T_c	[1, T] time slots
Trade-off parameter γ	[1, 10^4]

D. Performance Study

We start by studying the effect of prediction horizon T on the SE, under the assumption of perfect channel prediction. Fig. 2 shows the SE of a single user for a set of different video play-out rates V , as a function of the prediction horizon T . Two cases are examined, the first one when $T_c = T$, i.e. a new optimization is performed every T slots, and the second one when $T_c = 1$ slot.

We observe that as the encoding rate V increases, lower SE is achieved above a certain value of T . This is due to the constraint of the maximum buffer size that allows for lower V a longer video duration to be buffered close to the BS, where the channel conditions are better. Moreover, it is interesting to notice that as T increases, the curves of the SE oscillate for $T_c = T$ until they reach a maximum constant value. Thus, in some cases, a higher prediction horizon leads to worse results in terms of SE. This counter-intuitive effect depends on the set of time slots for which the optimization is performed. For example, for $T = 40$ slots, the second optimization is performed close to the cell edge where the resources are expensive in terms of bandwidth. On the other hand, for $T = 30$ slots, the second optimization occurs early enough to anticipate for the cell edge and the buffer is then filled at a lower price. When a new optimization is realized every time slot for $T_c = 1$ slot, then the curves are monotonic and the full potential of the prediction is exploited. This result shows that, even without prediction error, there is already a trade-off between computational effort and SE. We conclude that a small value for T_c should be used in order to provide near to optimal results for all the different parameter values.

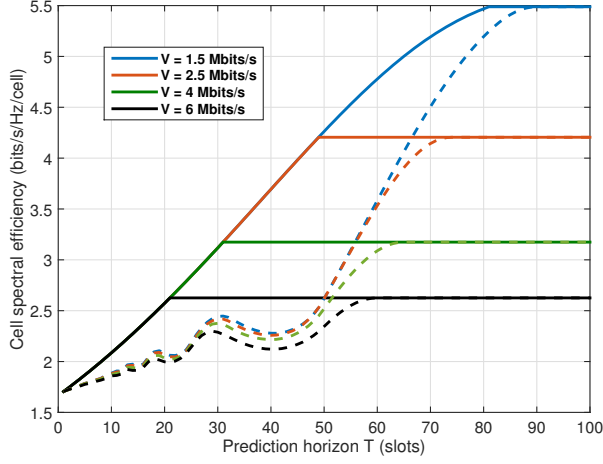


Fig. 2. Single user SE as a function of T and V ; solid lines for $T_c = 1$ slot and dashed lines for $T_c = T$.

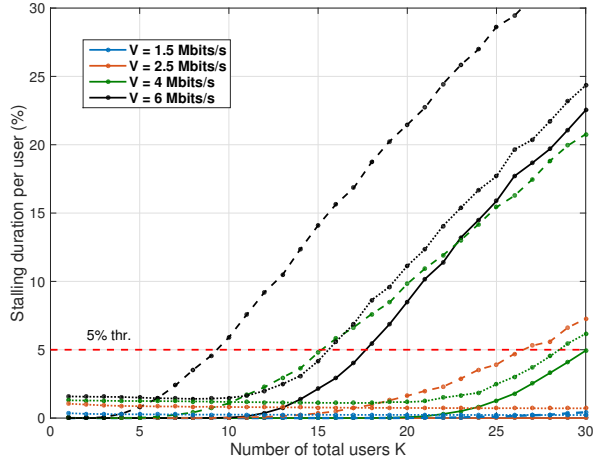


Fig. 3. Stalling duration as a function of users K ; solid lines for ARRM with perfect channel prediction, dotted lines for ARRM with $\sigma = 10$ dB and dashed lines for baseline.

Now that we have illustrated the relation between basic parameters of our model, we proceed by studying the QoS in terms of stalling duration, with and without prediction errors. In the following figures, we denote by ‘ARRM’ the case with $T = 100$ and $T_c = 20$ time slots respectively and by ‘baseline’ the approach without prediction, where the BS instantaneously allocates the necessary resources to satisfy the given play-out data rate. Fig. 3 presents the stalling duration per user as a fraction of the total time the user spends in the system, for different values of K, V . As expected, ARRM (solid lines) provides a clear reduction of the stalling duration compared to the baseline (dashed lines). For example, by limiting the stalling duration to 5%, we can see that the maximum number of users, served under this QoS constraint, is almost doubled for $V = \{4, 6\}$ Mbits/s. For lower V , the gains cannot be defined exactly in this example, since the QoS of ARRM is so high that more than the simulated number of $K = 30$

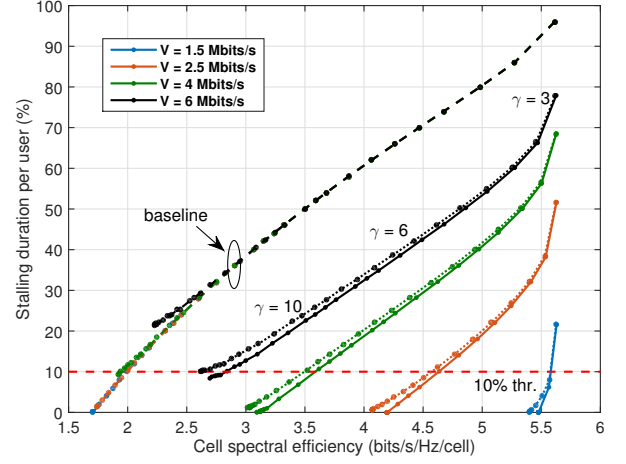


Fig. 4. Cell SE vs. stalling duration for $K = 20$; solid lines for ARRM with perfect channel prediction, dotted lines for ARRM with $\sigma = 10$ dB and dashed lines for baseline.

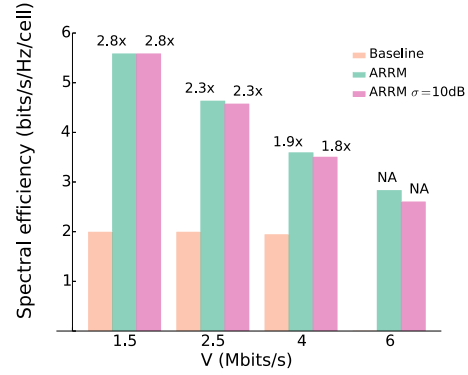


Fig. 5. Cell SE for 10% stalling duration and gains compared to baseline

users is supported. The effect of the channel prediction error is marginal as we can see for the case where ARRM with $\sigma = 10$ dB (dotted lines). Here, the gains for $V = \{4, 6\}$ Mbits/s are only slightly reduced compared to ARRM with perfect channel prediction.

Our ARRM formulation (3a)-(3c) trades off SE and stalling duration. In the previous results, a large value of γ was assumed in order to prioritize the minimization of the stalling time. If $\gamma \leq \max \frac{V_{k,t}}{S_{k,t}}, \forall k, t$, then less resources can be allocated at the cost of higher stalling. The above expression of the γ threshold is easily found by setting $\ell'_{k,t} = \ell_{k,t} + \delta$ and $\omega'_{k,t} = \omega_{k,t} - \frac{\delta V_{k,t}}{S_{k,t}}$, with $\delta > 0$, that satisfy all the constraints of (4a)-(4e) and then by verifying when this solution leads to a better objective value. Fig. 4 illustrates the set of the optimal solutions for the complete range of γ values from Table I for $K = 20$ users. We can see that ARRM achieves a better trade-off for all the curves, i.e. higher cell SE is obtained for a given value of stalling duration. For $V = 6$ Mbits/s, we can also notice that the curves do not reach the x-axis, which means that although ARRM reduced stalling, stalls cannot be entirely avoided under such high load. The effect of the prediction

TABLE II
MEDIAN COMPUTATIONAL TIME OF ONE OPTIMIZATION.

K'	Time (ms)			
	Baseline	ARRM		
		$T = 20$	$T = 50$	$T = 100$
1	0.06	0.23	0.44	0.93
10	0.18	2.44	6.17	10.01
20	0.24	6.17	15.67	23.26
30	0.31	11.48	28.61	42.74

error with $\sigma = 10$ dB on the trade-off curves is higher as γ increases, but remains marginal for all the studied cases. For a better illustration of the achieved SE gains, Fig. 5 shows the SE reached at 10% stalling duration. We can see in this figure that ARRM with $Z = 20$ Mbits achieves an impressive increase of cell SE up to 2.8 times, while satisfying the above QoS-constraint. For $V = 6$ Mbits/s, we can also notice that the baseline does not satisfy the QoS-constraint. Note that these high gains are robust to channel prediction errors.

E. Computational Time

We now study the computational time for the solution of the proposed LP formulation over different sets of parameters. Let us represent the number of simultaneously active users by the variable K' , with $K' \leq K$. We measure the computation time for one single optimization for different values of K' and prediction horizon T on a typical server processor (i.e., the Intel Xeon CPU running at 3.3 GHz) using the optimization engine CPLEX v12.6, [16].

The variables K' and T are the two key factors that define the size of the linear system in (4a)-(4e) to be $2TK'$ variables and $T(2K' + M)$ constraints. Thus, K' and T have a large effect on the computational time. We have also verified that V only insignificantly affects the runtime and assume, thus, a constant $V = 1.5$ Mbits/s when studying this metric. Table II provides the median computational time of one optimization. As we can see from this table, the computational time increases with the number of active users and with the prediction horizon. However, even for $T = 100$ slots, the corresponding time strongly indicates that the proposed optimization problem can provide anticipatory resource allocation sufficiently fast for practical systems.

IV. CONCLUSIONS

In this paper, we studied Anticipatory Radio Resource Management (ARRM) for mobile video streaming based on channel state prediction and knowledge of the traffic rate. An LP formulation was proposed that provides the optimal solution in a computationally efficient manner.

Our numerical results for a representative scenario with multiple users and two base stations provide high insight. We verified that ARRM highly increases the QoS compared to the baseline scheme without anticipation. Further studying spectral efficiency and stalling time, displays the proper choice

of the trade-off parameter and reveals an impressive spectral efficiency gain at high QoS. This gain is only slightly reduced if a model for channel prediction is introduced into the study. This shows how robust our ARRM formulation is against the side effects of practical channel prediction. Further practicality is demonstrated by a low computational time, which supports real-time solutions even for large instances of the problem.

In our future work, we will further study the robustness of ARRM under practical assumptions. This requires to study various error models for channel prediction and traffic rate estimation. We aim to include an error term into the ARRM formulation for further robustness. Finally, simulations of larger topologies, QoS, channel and traffic models are required, in order to verify the exceptional quality and efficiency in further scenarios.

REFERENCES

- [1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2014–2019," *white paper*, Feb. 2015.
- [2] Huawei, "Mobile video service performance study," *white paper*, Jun. 2015.
- [3] T.-Y. Huang *et al.*, "A Buffer-based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service," in *SIGCOMM*, Aug. 2014, pp. 187–198.
- [4] M. Seufert *et al.*, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 2015.
- [5] R. Pantos and W. May, "HTTP live streaming," IETF, Informational Internet-Draft 2582, Sep. 2011.
- [6] ISO/IEC, "Dynamic adaptive streaming over HTTP (DASH)," ISO/IEC, International Standard DIS 23009-1.2, 2012.
- [7] A. Aguiar, A. Wolisz, H. Lederer, and H. Karl, "Channel-adaptive schedulers with state-of-the-art channel predictors," in *European Wireless*, Apr. 2005, pp. 1–6.
- [8] A. Chiumento *et al.*, "Adaptive CSI and feedback estimation in LTE and beyond: a Gaussian process regression approach," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, 2015.
- [9] D. Astély *et al.*, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, 2009.
- [10] J. Yao, S. S. Kanhere, and M. Hassan, "Improving QoS in High-Speed Mobility Using Bandwidth Maps," *IEEE Trans. Mob. Comput.*, vol. 11, no. 4, pp. 603–617, 2012.
- [11] H. Riiser *et al.*, "Video streaming using a location-based bandwidth-lookup service for bitrate planning," *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, vol. 8, no. 3, p. 24, 2012.
- [12] Q. Liao, S. Valentin, and S. Stanczak, "Channel gain prediction in wireless networks based on spatial-temporal correlation," in *SPAWC*, Stockholm, Sweden, June 2015, pp. 400–404.
- [13] S. Sadr and S. Valentin, "Anticipatory buffer control and resource allocation for wireless video streaming," *CoRR*, vol. abs/1304.3056, 2013. [Online]. Available: <http://arxiv.org/abs/1304.3056>
- [14] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *INFOCOM, 2013 Proceedings IEEE*, Turin, Italy, April 2013, pp. 2706–2714.
- [15] A. El Essaili *et al.*, "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 988–1001, Jun. 2015.
- [16] IBM-Cplex, v. 12.6, <http://www-01.ibm.com/software/integration/optimization/cplex/>.
- [17] 3GPP, "Further advancements for E-UTRA physical layer aspects," *3GPP Technical report, TR 36.814 V9.0.0*, Mar. 2010.
- [18] A. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct 1997.