Multiuser Video Streaming Rate Adaptation: A Physical Layer Resource-Aware Deep Reinforcement Learning Approach

Kexin Tang†, Nuwen Kan‡, Junni Zou∗, Xiao Fu†, Mingyi Hong† and Hongkai Xiong∗
†School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, {txk1994-china, kannw_1230, xionghongkai}@sjtu.edu.cn, zou-jn@cs.sjtu.edu.cn
‡Department of Electrical and Engineering and Computer Science, Oregon State University, USA., xiao_fu@oregonstate.edu
∗School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, {txk1994-china, kannw_1230, xionghongkai}@sjtu.edu.cn, zou-jn@cs.sjtu.edu.cn

Abstract—In this paper, we propose a cross-layer decision framework for multiuser adaptive video delivery over time-varying and mutually interfering wireless cellular network. The key idea is to synthetically design the physical-layer optimization-based beamforming scheme (performed at the base stations) and the application-layer deep reinforcement learning (DRL)-based rate adaptation scheme (performed at the user terminals), so that a very complex multi-user overall fair long-term quality of experience (QoE) maximization problem can be decomposed to two layers and solved effectively. Extensive simulations show that the proposed cross-layer design is effective and promising.

Index Terms—Wireless video streaming, beamforming, rate adaptation, deep reinforcement learning, cross-layer design.

I. INTRODUCTION

Mobile video streaming has been a major drive for the exponential growth of the global mobile data traffic, which has already accounted for 60% of the total mobile data traffic in 2016, and this number is projected to be increased to 78% by 2021 [1]. This trend imposes significant challenges to the task of real-time delivery of resource-demanding video streams over wireless networks.

To ensure high quality of experience (QoE) for mobile users, while coping with ever growing user heterogeneity (in terms of demands for video content, display devices and available network resources) and dynamic changing network conditions, flexible video delivery techniques such as dynamic adaptive streaming over HTTP (DASH) have been developed. In DASH [2], each video is encoded at different bitrates and/or resolutions to generate several representations. These representations are further divided into small chunks containing a few seconds of video content (typically 2 seconds). In this way, users are able to dynamically select chunk representation to fit their network conditions. A variety of user-side rate adaptation techniques [3], [4] have been proposed to improve the user-perceived QoE based on locally observed throughput/buffer status. However, these algorithms have inferior performance (e.g., lower video quality, frequent quality switches, QoE unfairness) in the presence of multiple competing video flows [5].

To deal with these issues, some cross-layer approaches [6], [7] have been proposed, which jointly optimize the physical layer transmission rates of all users with QoE fairness objective, and then overwrite the requested video bitrates by a network-side centralized proxy server to match the optimization result. Note that these methods are less appealing as centralized operations are usually undesired for upper layers due to privacy/security reasons and the asynchronous nature of user requests. More importantly, these algorithms merely focus on short-term QoE maximization problem (within a time slot). In practice, however, it is much more preferable to offer high and fair QoE to users over a long period of time, since the event of video watching can easily span hours.

To this end, in this paper, we propose a novel cross-layer decision framework for multiuser adaptive streaming over wireless cellular network. Specifically, in the physical layer, we formulate a proportional fairness maximization problem in terms of user’s average video quality over a period time. A quality-driven dynamic resource allocation (QDDRA) algorithm is then proposed, which is capable of determining the transmission rates for users to achieve fair resource allocation. In the application layer, we model the rate adaptation logic into a reinforcement learning (RL) task. Each user is regarded as an independent agent that learns the desired chunk representation by interacting with changing wireless network environment, so that her/his long-term QoE is maximized. Particularly, we leverage a state-of-the-art deep reinforcement learning (DRL) algorithm called asynchronous advantage actor-critic (A3C) to train two neural networks, where the actor network is used to generate a rate adaptation policy, together with a critic network to evaluate the learned policy. Extensive simulations are conducted to showcase the effectiveness of our approach.

II. FRAMEWORK AND SYSTEM MODELS

A. Framework

As illustrated in Fig. 1(a), we consider a wireless adaptive streaming system consisting of a single DASH server and multiple mobile users located in a cellular network. Suppose that the server stores $F$ video files denoted as $\mathcal{F} = \{1, 2, \ldots, F\}$, each of which is fragmented into small video chunks comprising $T_{\text{chunk}}$ seconds of video. Each chunk is independently encoded into $L$ different representations with $\mathcal{A}_{f,m}$ being...
the available bitrate set for $m$-th chunk of the video file $f$. Furthermore, assume that $I$ users request video playback from the server and compete for the wireless resource. The channel states between users and base stations could change frequently due to user mobility and channel fading. We consider a time-slotted system with each time slot being $T_{\text{slot}}$ seconds, and channel states are stable within the duration of a time slot.

Our proposed cross-layer decision framework is shown in Fig. 1(b). According to the cross-cell channel state information at each time slot, the transmission rate allocated to each user is determined by the QDDRA algorithm. The transmit beamforming vectors for all base station-user pairs. On this basis, each user locally learns the optimal chunk representation using DRL technique to maximize her/his long-term QoE. Hence, there is no centralized computation or synchronization required on the application layer. This is very different from the existing short-term cross-layer QoE optimization works in [6], [7]. Arguably, our framework better fits the user-dependent nature of video streaming services.

B. Wireless Network Model

We consider a cellular network with $K$ cells, modeled by MIMO interfering broadcast channel. There is a single base station $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ within cell $k \in \mathcal{K}$, and it is equipped with $N_a^k$ transmit antennas and sends video data to $I_k$ users located in different areas of the cell $k$. Let $I_k$ denote the $i$-th user in cell $k$ who has $N_a^k$ receive antennas, and the set of all users is represented as $I = \bigcup_{k \in \mathcal{K}} I_k$ where $I_k = \{1_k, 2_k, \ldots, I_k\}$. The set of users located in cell $k$. Denote by $\mathbf{v}_i^k \in \mathbb{C}^{N_a^k}$ and $\mathbf{H}_i^k \in \mathbb{C}^{N_a^k \times N_a^k}$ the transmit beamformer and the downlink channel gain from the base station $k$ to user $i$ at time slot $t$. Taking into account a linear channel model and additive white Gaussian noise (AWGN) with probability distribution $CN(0, \sigma_i^2 I_{N_a^k})$, the signal-to-interference-plus-noise ratio (SINR) at the receiver $i$ is

$$\text{SINR}_i^k = \mathbf{H}_i^k \mathbf{v}_i^k \mathbf{H}_i^k \mathbf{H}_i^k \mathbf{H}_i^k \Theta_i^k, \quad \forall i_k \in I$$

with $\Theta_i^k = \sum_{j \neq i} \mathbf{H}_i^k \mathbf{H}_i^k \mathbf{v}_j^k \mathbf{v}_j^k \mathbf{H}_i^k \mathbf{H}_i^k + \sigma_i^2 I_{N_a^k}$. Accordingly, the achievable data transmission rate for user $i_k$ is

$$R_{i_k} = B \log_2 \det \left( I_{N_a^k} + \frac{\text{SINR}_i^k}{\Gamma} \right), \quad \forall i_k \in I,$$

where $B$ represents the channel bandwidth and $\Gamma$ is the SNR gap depending on the modulation scheme.

C. QoE Model

Suppose that user $i_k \in I$ requests the playback for the video file $f \in \mathcal{F}$ and downloads the video chunks with the desired representation in turn. We denote by $\tau_m \in \mathcal{A}_{f,m}$ the selected bitrate for the $m$-th chunk, and the video quality $q_m$ perceived by the user about the chunk can be expressed as:

$$q_m = g(a_m, z_m),$$

where $z_m$ is a content-dependent parameter vector that indicates the complexity of the chunk $m$. The parametric rate-quality function $g(\cdot): \mathbb{R}^+ \rightarrow \mathbb{R}^+$ maps the encoding bitrate to some quality metrics (e.g., PSNR, SSIM).

Assume that the video player starts to download chunk $m$ at time $t_m$, then the experienced average throughput for downloading the chunk can be expressed as:

$$C_m = \frac{1}{t_{m+1} - t_m} \int_{t_m}^{t_{m+1}} R_{i_k}^t dt,$$

where $R_i^t$ is transmission rate for the user defined in Eq. (2). Here, $t$ denotes a continuous time index, and $R_i^t$ can be treated as a step function which changes every $T_{\text{slot}}$ seconds. Thus, the download time of the chunk $m$ can be derived as $d_m = \tau_m(a_m)/C_m$, where $\tau_m(a_m)$ denotes the size of chunk $m$ encoded at bitrate $a_m$. The downloaded video chunks are stored in the user’s playout buffer. Let $b_m \in [0, b_{\text{max}}]$ denote the buffer occupancy (measured in seconds) when the video player attempts to request chunk $m$. A rebuffering event occurs when the buffer becomes empty, and the corresponding rebuffering time can be formulated as:

$$\phi_m = (d_m - b_m)_{+},$$

where the notation $(x)_{+} = \max\{x, 0\}$.

To achieve high efficiency of the adaptive streaming system, the user should watch the highest possible video quality based on its available wireless resource. Meanwhile, frequent quality switches and rebuffering events should be avoided to guarantee smooth and stall-free playback. Therefore, we can define the user-perceived QoE of the chunk $m$ as

$$QoE_m = q_m - \lambda(q_m - q_m-1) - \rho \phi_m,$$

where $\lambda$ and $\rho$ are non-negative parameters used to trade off the instantaneous video quality, quality fluctuations and rebuffering events in the QoE evaluation.

III. QUALITY-DRIVEN DYNAMIC RESOURCE ALLOCATION

From the QoE model defined in last section, it can be seen that the physical-layer resource allocation (i.e., transmission rate allocated to each user) has a significant influence on the ultimate performance of adaptive streaming system. Therefore, we first exploit the rate-quality function to map the transmission rate into video quality, that is,

$$q_{i_k}^t = g_i(R_{i_k}^t, z_{i_k}^t).$$
The parameter vectors $z^t_i$ is the complexity of video viewed by the user $i_k$, and its transmission rate $R_{ik}'$ is determined by the beamforming vectors $v^t_i$ as shown in Eqs. (1)(2). Then, we formulate a proportional fairness maximization problem in terms of average video quality of each user as:

$$\max \sum_{v^t_i} \log Q_{ik}'$$

$$\text{s.t.} \sum_{i_k \in I_k} |w_{i_k}^t|^2 \leq P_{ik}, \quad \forall k \in \mathcal{K},$$

where $P_{ik}$ is the power budget of base station $k$. $Q_{ik}'$ represents the average quality of user $i_k$ up to time slot $t$, i.e.,

$$Q_{ik}' = \beta q_{ik}' + (1 - \beta)Q_{ik}^{-1}.$$  

Here, $\beta \in (0, 1]$ is used to control the impact of average video quality obtained in the previous time slots.

In practice, proportional fairness maximization problem can be approximately implemented using a weighted sum maximization problem [8]. Thus, Eq. (8) can be converted to the following weighted sum-quality maximization problem:

$$\max \sum_{v^t_i} \alpha^t_i q_{ik}'$$

$$\text{s.t.} \sum_{i_k \in I_k} |w_{i_k}^t|^2 \leq P_{ik}, \quad \forall k \in \mathcal{K}$$

with $\alpha^t_i = 1/Q_{ik}^{-1}$. Eq. (10) can be easily solved by extending the popular WMMSE algorithm [9]. Thus, the quality-driven dynamic resource allocation (QDDRA) algorithm is proposed in Algorithm 1 ($u^t_i$ and $w^t_i$ are auxiliary variables).

**Algorithm 1 QDDRA algorithm.**

1. **Input:** $\alpha^t_i = 1/|Q_{ik}^{-1}|$, $z^t_i$, $H^t_i$, $\forall i \in \mathcal{K}, i_k \in I$;
2. **Initialize** $v^t_i$ such that $|w_{i_k}^t|^2 = \frac{P_{ik}}{\alpha^t_i}$, $\forall i_k \in I$;
3. **Repeat** for each user $i_k$:
   5. $u^t_{ik} \leftarrow \left( \sum_{j \in I_k} H^t_{ik} v_j^t \right) / \|v_j^t\|^2_{H^t_{ik}}$;
   6. $w^t_{ik} \leftarrow \left( \sum_{j \in I_k} (H^t_{ik} H^t_{ik})^{-1} u^t_{ik} v_j^t \right) / \|v_j^t\|^2_{H^t_{ik}}$;
   7. $q^t_{ik} \leftarrow \alpha^t_i \left( \sum_{j \in I_k} H^t_{ik} v^t_{ik} \right) / \|v^t_{ik}\|^2_{H^t_{ik}}$;
4. **Until** some stopping criteria is met.
5. **Compute** $R^t_i$, $q^t_i$, $Q^t_i$ based on Eq. (2)(7)(9) respectively.
6. **Output:** $R^t_i$, $q^t_i$, $Q^t_i$, $\forall i_k \in I$;
7. **Update** $t \leftarrow t + 1$.

IV. DRL-BASED RATE ADAPTATION

According to the dynamically allocated wireless resource, each user should adapt the chunk representation to underlying network conditions such that its long-term QoE is maximized. We model the rate adaptation process as a RL task where an agent (e.g., the mobile DASH user) learns the best action (i.e., the best bitrate of chunk to be downloaded) to achieve the anticipated goal (i.e., maximizing the long-term QoE) from the interaction with the environment. The learning process is independently executed in the application layer of each user.

To be specific, we define the state $s_m$, action $a_m$ and reward $r_m$ as follows:

1. **State:** The system state at the time when the video player starts to request the download of the chunk $m$ is defined as $s_m = (C_m, \bar{d}_m, \tau_m, b_m, \omega_m, \delta_m)$, which is characterized by the measured network throughput experienced by the previous $n$ video chunks, $C_m = \{C_{m-n}, \ldots, C_{m-1}\}$, which depend on the transmission rates determined by the QDDRA algorithm; the download time of the past $n$ video chunks, $\bar{d}_m = \{d_{m-n}, \ldots, d_{m-1}\}$; the complexity of the chunk $m$, $\omega_m$; the available sizes of the chunk $m$, $\tau_m$; the current buffer occupancy, $b_m$; the number of remaining chunks in the video, $\delta_m$; and the video quality of the last downloaded chunk, $\delta_m = \delta_m - 1$.

2. **Action:** The action $a_m$ corresponds to the selected bitrate for the chunk $m$.

3. **Reward:** The scalar reward is an available immediate feedback from the environment when the agent takes an action, here we consider the QoE of chunk $m$ as reward, that is, $r_m = QoE_m$.

We employ A3C algorithm to train actor neural network and critic neural network that approximate the policy function and value function respectively. In order to derive the optimal policy $\pi_{\theta}$, we update the policy parameters $\theta$ by performing gradient ascent on the expected total reward. The gradient of the expected total cumulative reward is

$$\nabla_{\theta} \mathbb{E}_{\pi_\theta} \left[ \sum_{m=0}^{\infty} \gamma^m r_m \right] = \mathbb{E}_{\pi_\theta} \left[ \nabla_{\theta} \log \pi_\theta (a_m | s_m; \theta) A^m (s_m, a_m) \right],$$

where $A^m (s_m, a_m) = Q^m (s_m, a_m) - V^m (s_m)$ is the advantage function, which depicts the difference between the expected return when deterministically selecting an action $a_m$ in state $s_m$ and the expected return for actions drawn from policy $\pi_\theta$.

We estimate the advantage function by the temporal difference method with n-step bootstrapping, and train the critic network parameters $\theta_v$ by the following update rule:

$$\theta_v \leftarrow \theta_v - \mu' \sum_m \nabla_{\theta_v} (A(s_m, a_m, \theta_v))^2,$$

where $\mu'$ is the learning rate of the critic network. Further, we can update the policy parameters $\theta$ of the actor network by

$$\theta \leftarrow \theta + \mu \sum_m \nabla_{\theta} \log \pi_\theta (a_m | s_m; \theta) A(s_m, a_m) + \varphi \nabla_{\theta} h(\pi_\theta (| s_m))$$

with $\mu$ being the learning rate of the actor network. Here, $h(\cdot)$ is the entropy of the policy used to encourage exploration and the hyperparameter $\varphi$ controls the strength of exploration. The pseudocode for the A3C algorithm can be found in [10].

V. EXPERIMENTS

In this section, we evaluate the performance of the proposed cross-layer decision framework for multiuser adaptive streaming in wireless cellular networks. We encode three test video sequences (BigBuckBunny, SitaSingstheBlues and DucksTakeOff) into 6 different representations with encoding bitrate set as {0.3, 0.75, 1.2, 1.85, 2.85, 3.2} Mbps. Then they
are further divided into multiple 2-second chunks and stored at the DASH server. We adopt PSNR as the video quality metric and set the penalty weights in QoE metric as $\lambda = 0.5$, $\rho = 4$. Besides, four users are assumed to watch the videos and compete the wireless resource.

We take the weighted sum mean-square error minimization (WMMSE) algorithm [9] as baseline for wireless resource allocation. For rate adaptation part, we adopt rate-based (RB) [3] and buffer-based (BB) [4] schemes for comparison. We illustrate the effectiveness of our proposed cross-layer combination of QDDRA algorithm and DRL-based rate adaptation logic (called QDDRA_DRL) over the following five schemes: QDDRA_BB, QDDRA_RB, WMMSE_DRL, WMMSE_BB and WMMSE_RB, which combine different methods in physical and application layer respectively.

We carry out the QDDRA and WMMSE algorithms in Matlab to generate network throughput traces for training and testing respectively. The actor and critic neural network have the same architecture (a convolution layer and a full connection layer) except that the actor network has a softmax output while a linear output for the critic network. The learning rate of the actor and critic neural network are set as $10^{-5}$ and $10^{-3}$, respectively. The discount factor $\gamma = 0.99$ and entropy factor $\varphi = 0.5$. We report the average QoE of all chunks when the users watch videos over test traces.

Fig. 2a shows the overall average normalized QoE of all users, and Fig. 2b presents the corresponding cumulative distribution function (CDF). It can be seen that the performance of our proposed QDDRA_DRL exceeds the remaining methods with respect to the overall average QoE. In addition, QDDRA-based methods have better performance than the WMMSE-based methods—showing the effectiveness of our video streaming-tailored resource allocation algorithm. We further analyze the overall average performance on the individual components defined in Eq. (6), as illustrated in Fig. 2c. The QDDRA_DRL performs well in all three items, which is consistent with our intention in designing the cross-layer decision framework. On one hand, with fairness consideration, QDDRA algorithm provides each user with a basic transmission rate to support the effective video delivery. On the other hand, DRL-based method can learn how to maximize the selected video quality, and at the same time, can also learn how much buffer is necessary to alleviate the risk of rebuffering event based on the changing network conditions. Fig. 2d further compares the unfairness in the user-perceived average QoE, where one can see that the QDDRA-based methods exhibits much better fairness. Similar to [5], here we measure the user unfairness using $\sqrt{1 - \text{JainFair}^2}$ where JainFair is the Jain fairness index of obtained QoE for four users.

VI. CONCLUSION

In this work, we proposed a cross-layer framework for multi-user video streaming rate adaption over wireless networks. Our framework employs a tailored physical layer resource allocation algorithm (i.e., QDDRA) to assist optimizing long-term user QoE on the application layer using a deep reinforcement learning approach. The proposed framework offers a simple yet effective way to solve the complex multi-user overall fair long-term QoE maximization task. Extensive simulations show that such cross-layer design is effective and promising.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under No. 61831018, Grant 61831018, Grant 61622112, No. 1931023 and No. 61972256, NSF grants CMMI-1727757 and CCF-1526078, and the China Scholarship Council.

REFERENCES

[1] “Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper,” 2017.
[2] T. Stockhammer, “Dynamic adaptive streaming over HTTP– standards and design principles,” in Proc. ACM MMSys, 2011, pp. 133–144.
[3] C. Liu, I. Bouazizi, and M. Gabbouj, “Rate adaptation for adaptive HTTP streaming,” in Proc. ACM MMSys, 2011. pp. 169–174.
[4] T. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, “A buffer-based approach to rate adaptation: Evidence from a large video streaming service,” ACM SIGCOMM Computer Communication Review, vol. 44, no. 4, pp. 187–198, 2015.
[5] J. Jiang, V. Sokar, and H. Zhang, “Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive,” IEEE/ACM Transactions on Networking, vol. 22, no. 1, pp. 326–340, Feb. 2014.
[6] A. E. Essaili, D. Schroeder, E. Steinbach, D. Staelhe, and M. Shehada, “QoE-based traffic and resource management for adaptive HTTP video delivery in LTE,” IEEE Transaction on Circuits and Systems for Video Technology, vol. 25, no. 6, pp. 988–1001, 2015.
[7] S. Cicalo, N. Changuel, V. Tralli, B. Sayadi, F. Faecheux, and S. Kerbouff, “Improving QoE and fairness in HTTP adaptive streaming over LTE network,” IEEE Transaction on Circuits and Systems for Video Technology, vol. 30, no. 8, 2012.
[8] Y. Lin and W. Yu, “Fair scheduling and resource allocation for wireless cellular network with shared relays,” IEEE Journal on Selected Areas in Communications, vol. 30, no. 8, 2012.
[9] Q. Shi, M. Razaviyayn, Z. Luo, and C. He, “An iteratively weighted mmse approach to distributed sum-utility maximization for a mimo interfering broadcast channel,” IEEE Transactions on Signal Processing, vol. 59, pp. 4331–4340, 2011.
[10] V. Mud, A. P. Badia, M. Mirza, A. Graves, T. Harley, T. P. Lillicrap, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in Proc. ICML, 2016, pp. 1928–1937.