A Systematic Framework for Dynamically Optimizing Multi-User Wireless Video Transmission

Fangwen Fu, *Student Member, IEEE,* and Mihaela van der Schaar, *Fellow, IEEE,*

***Abstract*—In this paper, we systematically formulate the problem of multi-user wireless video transmission as a multi-user Markov decision process (MUMDP) by explicitly considering the users' heterogeneous video traffic characteristics, time-varying network conditions as well as, importantly, the dynamic coupling among the users' resource allocations across time, which are often ignored in existing multi-user video transmission solutions. To comply with the decentralized wireless networks' architecture, we propose to decompose the MUMDP into multiple local MDPs using Lagrangian relaxation. Unlike in conventional multi-user video transmission solutions stemming from the network utility maximization framework, the proposed decomposition enables each wireless user to individually solve its own local MDP (i.e. dynamic single-user cross-layer optimization) and the network coordinator to update the Lagrangian multipliers (i.e. resource prices) based on not only current, but also the future resource needs of all users, such that the long-term video quality of all users is maximized. This MUMDP solution provides us the necessary foundations and structures for solving multi-user video communication problems. However, to implement this framework in practice requires statistical knowledge of the experienced environment dynamics, which is often unavailable before transmission time. To overcome this obstacle, we propose a novel online learning algorithm, which allows the wireless users to simultaneously update their policies at multiple states during each time slot. This is different from conventional learning solutions, which often update the current visited state per time slot. The proposed learning algorithm can significantly improve the learning performance, thereby dramatically improving the video quality experienced by the wireless users over time. Our simulation results demonstrate the efficiency of the proposed MUMDP framework as compared to conventional multi-user video transmission solutions.**

***Index Terms*—Multi-User Video Transmission, Markov Decision Process, Lagrangian Relaxation, Online Learning.**

I. INTRODUCTION

AS MULTIMEDIA applications continue to proliferate, wireless network infrastructures often need to support multiple simultaneously running applications. Key challenges associated with the robust and efficient multi-user video transmission over wireless networks are the dynamic allocation of the scarce network resources among heterogeneous users experiencing different time-varying network conditions and traffic characteristics and, given the network resource allocation, the dynamic adaptation of the cross-layer transmission strategies of the individual users.

Existing wireless video transmission solutions can be broadly divided into two categories: single-user video transmission solutions, focusing on packet scheduling, error protection or cross-layer adaptation in order to maximize the receiving user's video quality [1][4] and multi-user video transmission, emphasizing multi-user resource allocation among multiple users simultaneously transmitting video and sharing the same wireless resources [5][7]. However, most existing solutions in both categories do not explicitly consider both the heterogeneous characteristics of the video traffic and the time-varying network conditions (e.g. time-varying channel conditions, dynamic multi-user channel access, etc.), thereby often leading to suboptimal performance for wireless media systems. For example, in the single-user video transmission category, most solutions employ Unequal-Error-Protection (UEP) techniques [3] to differentially protect the video packets, or deploy rate-distortion optimization to schedule the video packets [1] based on their distortion impact and delay deadline and the packets' dependencies. However, these solutions assume only simplistic underlying network (channel) models and they do not consider the adaptation of transmission parameters at the other layers of the network stack, besides the application layer. To deal with the wireless network dynamics, cross-layer adaptation methods [2] have been proposed to optimize on-the-fly the transmission parameters at various layers, based on current observations of channel conditions. However, these cross-layer solutions are myopic and result in suboptimal performance because they do not account for the future channel conditions and video traffic.

In the multi-user video transmission category, many current techniques [7][8] are based on the network utility maximization (NUM) framework [6]. In the NUM framework, the basic assumption is that each user has a static utility function of the (average) allocated transmission rate (or QoS). For example, the authors in [7] simply consider the utility to be a function of the average allocated rate. The solutions in [8] defined the utility function (i.e. the average video quality or distortion) as a function of the average rate and packet loss. To deal with the dynamic wireless channel conditions, the resource allocation among the multiple users is repeatedly performed to maximize the current video quality. However, similar to the cross-layer optimization for the single-user video transmission [2][4], these solutions only myopically maximize the video quality for all the users at the current time and do not predict the impact of the current resource allocation on the future video quality of all the users. Therefore, it is crucial to judiciously allocate the limited resources to individual wireless video users such that their long-term utility (i.e. video quality) is maximized.

To address the abovementioned challenges associated with efficient multi-user video transmission over the time-varying

Manuscript received 24 May 2009; revised 20 October 2009.

F. Fu and M. van der Schaar are with the Department of Electrical Engineering, University of California Los Angeles, Los Angeles, CA, 90095 USA (e-mail: {fwfu, mihaela}@ee.ucla.edu).

Digital Object Identifier 10.1109/JSAC.2010.100403.

wireless network, we propose a systematic framework for dynamically and foresightedly optimizing the cross-layer transmission strategies (e.g. packet scheduling and resource acquisition, etc) of multiple users coexisting in the same wireless channel in order to maximize their long-term video quality. In the proposed framework, unlike existing video transmission solutions, we explicitly consider the heterogeneous video traffic characteristics, experienced time-varying network conditions, as well as the dynamic coupling among the users' transmission strategies across time. Our contributions are:

- To characterize the heterogeneous video data, we define a traffic state for each user which considers the number of data units (e.g. video frames or video packets) to be transmitted, their distortion impacts, and the dependencies between them at each transmission time. The traffic state together with the network state (e.g. channel conditions) characterizes the environment dynamics experienced by each user. Using this state definition, we are able to dynamically optimize the resource acquisition and packet scheduling for video transmission over time-varying networks.
- We further formulate the optimization of the packet scheduling and resource allocation for the dynamic multi-user video transmission system as a weakly-coupled MUMDP [10] problem. The MUMDP formulation allows each user to make foresighted transmission decisions by taking into account the future impact of its current decisions on the long-term utilities of all the users. Unlike the conventional centralized solutions [9] to the MDP which have very high computation complexity and unacceptable communication overheads, we propose to decompose this weakly-coupled MUMDP problem using Lagrangian relaxation into multiple local MDPs, each of which can be separately solved by the individual users. This decomposition is different from the conventional dual solutions [7] to the multi-user NUM-based video transmission problem in two ways: (i) instead of maximizing the static utility at each transmission time, our approach allows each wireless user to solve the dynamic cross-layer optimization problem (formulated as the local MDP), which is vital for the delay-sensitive video applications; (ii) instead of updating the Lagrangian multipliers only based on the current resource requirements of all users, our approach updates the multipliers based on not only current, but also future resource needs, such that the long-term video quality of all the users is maximized. To the best of our knowledge, this is the first attempt to formalize the multi-user video communication problem using MUMDP and decompose the MUMDP such that it can be solved in a decentralized manner by autonomous, yet collaborative, users.
- To overcome the obstacle of the unknown environmental dynamics (e.g. state transition probabilities) in real-time video applications operating in dynamic multi-user networks, the MUMDP framework provides the necessary foundations and principles for how the users can autonomously learn on-line to cooperatively optimize the global long-term video quality. Specifically, to deal with the unknown dynamics, each wireless user will deploy online reinforcement learning [16], and the network coordinator will update the resource price dynamically using stochastic subgradient methods [15]. Unlike conventional online learning algorithms [16], which often update the policy for only the visited state during each time slot, our proposed learning algorithm can simultaneously update multiple states during each time slot, which can significantly improve the learning performance. This approach has two advantages: (i) it does not require each user to know the statistical distribution of channel conditions and incoming video traffic beforehand; and, (ii) the wireless user and the network coordinator need to perform only very simple computations in each time slot.

The paper is organized as follows. Section II defines the traffic states, the state transition and the utility function for each wireless user at each time slot. Section III formulates the single-user cross-layer optimization as an MDP and proves that the utility function is concave. Section IV formulates the multi-user video transmission problem as an MUMDP. Section V presents how the MUMDP can be decomposed into multiple local MDPs using the Lagrangian relaxation method and develop the corresponding subgradient method to update the resource price. Subsequently, Section VI describes the proposed distributed online learning algorithm to deal with the unknown video characteristics and channel conditions. Section VII presents numerical results to validate the proposed framework. The conclusions are drawn in Section VIII.

II. Models for Heterogeneous Video Traffic

Unlike traditional traffic models [19], which only characterize the rate changes of video traffic, in this section we aim to develop a general model to represent the encoded video traffic with heterogeneous characteristics (e.g. various delay deadlines, distortion impacts, dependencies, etc.). Using this video traffic model, we will be able to dynamically optimize the resource acquisition and packet scheduling for video transmission over time-varying networks.

A. Attributes of data units

In this section, we discuss how the heterogeneous attributes of the video traffic can be modelled. The video data is often encoded periodically using a Group of Pictures (GOP) structure, which lasts a period of T time slots. The video frames within one GOP are encoded interdependently using motion estimation, while the frames belonging to different GOPs are encoded independently. Note that the prediction-based coding schemes can lead to sophisticated dependencies between the video data.

After being encoded, each GOP contains N data units (DUs), each of which represents one type of DU (e.g. I, P, B frames) and indexed by $j \in \{1, 2, \cdots, N\}$. The types of DUs are determined based on its distortion impact, delay deadline, and dependencies which are illustrated below. We assume that the GOP structure is fixed (i.e. the types of DUs are fixed at each GOP). The set of DUs within GOP $g \in \mathbb{N}$ is denoted by $\{f_1^g, \cdots, f_N^g\}$. The attributes of DU f_j^g are listed below.

Size: The size of DU f_j^g is denoted as l_j^g (measured in packets[1]), where $l_j^g \in [1, l_j^{max}]$, and l_j^{max} is the maximum allowable size for the j-th DU at each GOP. The size of DU f_j^g is determined when DU f_j^g is encoded. To simplify the exposition, l_j^g is generated from an i.i.d. random variable [2] with the probability mass function $PMF_j(l)$. Note that $PMF_j(l)$ is the same for the j-th DU across different GOPs, but is different for different types of DUs.

Distortion impact: Each DU f_j^g has a distortion impact q_j^g per packet, which is assumed to be the same for all the GOPs, i.e. $q_j^g = q_j^{g'}, \forall g, g'$.

Delay deadline: We define the delay deadline of DU f_j^g as the time by which the DU should be decoded in order to be displayed. We denote by d_j^g the delay deadline of DU f_j^g. Since the GOP structure is fixed, the difference between the delay deadlines of the two DUs within one GOP is constant, i.e. $d_j^g - d_{j'}^g = \Delta d_{jj'} > 0$ where $j > j'$, and the delay deadlines of the same DUs from different GOPs satisfy $d_j^g - d_j^{g-1} = T$. In other words, the delay deadline d_j^g is periodic with the period of T, which is the length of one GOP.

Dependency: When one DU f_j^g is encoded based on the prediction from the other DU $f_{j'}^g$, we say that DU f_j^g depends on DU $f_{j'}^g$. Note that the dependencies between DUs only occur within one GOP and DUs from different GOPs can be decoded independently (i.e. no dependency between them.). The dependencies between the DUs within one GOP are expressed as a directed acyclic graph (DAG) [1]. The DAG remains the same for a fixed GOP structure. In this paper, we assume that, if DU f_j^g depends on DU $f_{j'}^g$ (i.e. there exists a path directed from DU f_j^g to DU $f_{j'}^g$ and denoted by $j' \prec j$), then $d_j^g \geq d_{j'}^g$ and $q_j^g \leq q_{j'}^g$. In other words, DU $f_{j'}^g$ should be decoded prior to DU f_j^g and DU $f_{j'}^g$ has higher distortion impact. One illustrative example of DAGs for video data is given in Figure 1.

B. Traffic state representation in each time slot

In this subsection, we discuss how to represent the video traffic which can be potentially transmitted in each time slot. At time slot t, as in [1], we assume that the wireless user will only consider for transmission the DUs with delay deadlines in the range of $[t, t+W]$, where W is referred to as the scheduling time window (STW) and assumed to be given a priori[3]. In this paper, we further assume that STW is chosen to satisfy the following condition: if DU f_j^g directly depends on DU $f_{j'}^g$ (i.e. there is a direct arc from f_j^g to $f_{j'}^g$ in the DAG), then $f_j^g - f_{j'}^g < W$. This assumption ensures that DU f_j^g and $f_{j'}^g$ can be within one STW.

We denote by the set of DUs whose delay deadlines are within the range of $F_t = \{f_j^g | d_j^g \in [t, t+W]\}$. This set F_t is referred to as the DU-type state at time slot t. Since the GOP structure is fixed, it is easy to show that F_t is periodic with the period of T, which means that, for any $f_j^g \in F_t$, there exists $f_j^{g+1} \in F_{t+T}$ and vice versa. Hence, F_t and F_{t+T} have the same types of DUs and the same DAG between these DUs. For example, as shown in Figure 1, $F_t = \{f_1^g, f_2^g, f_3^g\}$ and $F_{t+3} = \{f_1^{g+1}, f_2^{g+1}, f_3^{g+1}\}$ where $T = 3$. It is clear that the DAG between the DUs in F_t is also the same as the one between the DUs in F_{t+3}. Due to the periodicity, there are only T different DU-type states. The transition of F_t is deterministic and denoted by $\delta(F_{t+1} - Next(F_t))$ where $Next(F_t)$ represents the next DU-type state following F_t and $\delta(x) = 1$ if $x = 0$ and otherwise, 0. For example in Figure 1, $Next(F_t) = \{f_2^g, f_3^g, f_4^g, f_5^g\}$.

Furthermore, for each DU $f_j^g \in F_t$, we denote by b_j^g the amount of packets remaining for transmission at time slot t. Note that $b_j^g \leq l_j^g$ which means the remaining packets must be less than the amount of the originally available packets. If DU f_j^g is undecodable, $b_j^g = -1$. We denote the buffer state of the DUs in F_t by $B_t = \{b_j^g | f_j^g \in F_t\}$. The traffic state $\mathcal{T}_t = (F_t, B_t)$ of the video application is then defined as representing the types of DUs, the dependencies between them and the amount of packets remaining for transmission. Hence, the traffic state $\mathcal{T}_t$ is able to capture heterogeneous video traffic and is a super-set of existing well-known single-buffer models (i.e. which ignore packet dependencies and delay deadlines) or multi-buffer models (i.e. which ignore packet dependencies or delay deadlines).

C. Packet scheduling, state transition and immediate reward

Given a transmission rate[4] r_t at time slot t, the wireless user has to determine the amount of packets to be transmitted for each DU in F_t, which we refer to as packet scheduling. The scheduling policy π maps the current traffic state $\mathcal{T}_t$ and transmission rate r_t into the amount of packets transmitted during each time slot, $\mathbf{y}_t = [y_j^g | f_j^g \in F_t]$, i.e. $\pi(\mathcal{T}_t, r_t) = \mathbf{y}_t$. Formally, the scheduling policy π satisfies the following conditions[5]: (i) Underflow constraint: $0 \leq y_j^g \leq b_j^g, \forall f_j^g \in F_t$; (ii) Rate constraint: $\sum_{f_j^g \in F_t} y_j^g \leq r_t$. In other words, we allow partial DUs to be scheduled for transmission. The set of possible policies in each traffic state $\mathcal{T}_t$ given a certain transmission rate r_t is denoted by $\mathcal{P}(\mathcal{T}_t, r_t)$. In this paper, we assume that the packet scheduling policy $\pi(\mathcal{T}_t, r_t)$ is a vector of nonnegative real number, i.e. $\pi(\mathcal{T}_t, r_t) \in \mathbb{R}_+^{|F_t|}$ where $|F_t|$ represents the number of DUs in F_t. This type of packet scheduling policy can be implemented by mixing the packet scheduling policies taking an integer number of packets to transmit. In the rest of paper, we assume that $\pi(\mathcal{T}_t, r_t)$ takes values from the nonnegative real number set. $\mathcal{P}(\mathcal{T}_t, r_t)$ represents the set of feasible mixed packet scheduling policies.

In the following, we discuss the transition of the traffic state $\mathcal{T}_t$, given the transmission rate r_t. First, as discussed in Section II.B, the DU-type state has the deterministic transition $\delta(F_{t+1} - Next(F_t))$, which is independent of the transmission

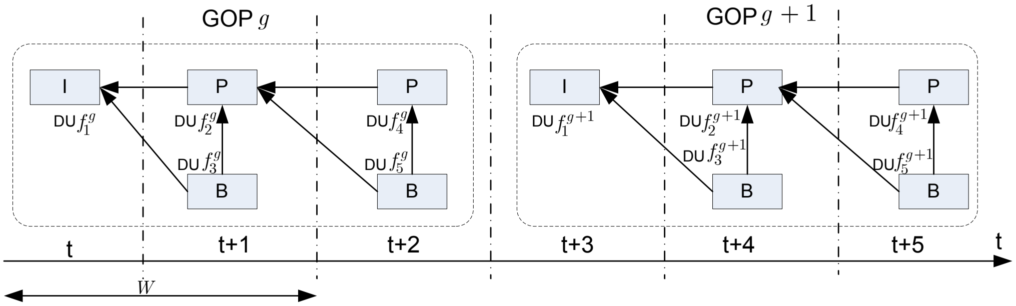
[1] For simplicity, we assume in this paper that each packet has the same length, but this does not affect our proposed solution. It just simplifies our exposition given the space limitations.

[2] The DU size can also be modeled as a random variable depending on the previous DUs [17].

[3] The STW can be determined based on the channel conditions experienced by the user in each time slot. For example, the STW can be set small when the channel conditions are poor (low SNR regime), and to be large whenever the channel condition are good.

[4] The transmission rate can be determined by the allocated network resource and transmission strategies at the layers below application layer.

[5] Similar constraints are also considered in [18]. However, the authors therein did not consider the time-varying transmission rate and foresighted packet scheduling decisions aimed at maximizing the long-term video quality.



$$F_t = \{f_1^g, f_2^g, f_3^g\} \quad F_{t+1} = \{f_2^g, f_3^g, f_4^g, f_5^g\} \, F_{t+2} = \left\{f_4^g, f_5^g, f_1^{g+1}\right\} F_{t+3} = \left\{f_1^{g+1}, f_2^{g+1}, f_3^{g+1}\right\}$$

Fig. 1. DAG-based dependencies and traffic states at each time slot using IBPBP GOP structure

rate r_t. However, the buffer state transition is determined by the scheduling policy, the dependencies between DUs and the PMF of the incoming DUs.

In order to compute the transition from B_t to B_{t+1}, we divide the DUs in F_{t+1} into two disjoint subsets: $F_t \cap F_{t+1}$ and F_{t+1}/F_t[6]. Specifically, $F_t \cap F_{t+1}$ represents the set of DUs that are not expired at time slot $t+1$, and F_{t+1}/F_t represents the newly arriving DUs at time slot $t+1$. We can also divide F_{t+1} into two subsets: the decodable set F_{t+1}^d and undecodable set F_{t+1}^u. F_{t+1}^d represents the set of DUs whose parent DUs in the DAG (if any) have been successfully transmitted, and is computed as $F_{t+1}^d = \{f_j^g | f_j^g \in F_{t+1} \& \forall j' \prec j, b_{j'}^g = 0\}$. F_{t+1}^u represents the set of DUs that cannot be decoded (i.e. $b_{j'}^g = -1$ or $b_{j'}^g > 0$ where $j' \prec j$), and is computed as $F_{t+1}^u = \{f_j^g | f_j^g \in F_{t+1} \& \exists j' \prec j, b_{j'}^g \neq 0\}$. Let $B_{t+1} = \{\hat{b}_j^g | f_j^g \in F_{t+1}\}$ be the buffer state at time slot $t+1$. We are now ready to compute the transition probability of the buffer state from B_t to B_{t+1}, as shown below.

$$\begin{aligned} p(B_{t+1}|B_t, \mathbf{y}_t) = & \prod_{f_j^g \in F_{t+1}^u} \delta(\hat{b}_j^g + 1) \prod_{f_j^g \in F_{t+1}^d \cap (F_{t+1} \cap F_t)} \\ & \delta(\hat{b}_j^g - (b_j^g - y_j^g)) \prod_{f_j^g \in F_{t+1}^d \cap (F_{t+1}/F_t)} PMF_j(\hat{b}_j^g) \end{aligned} \tag{1}$$

where the first product in the right hand side represents the buffer state transition probability of the undecodable DUs, the second one represents the transition probability of the remaining decodable and unexpired DUs, and the third one represents the transition probability of the decodable and newly arriving DUs. The transition of the traffic state $\mathcal{T}_t$ is then computed as

$$p_{\mathcal{T}}(\mathcal{T}_{t+1}|\mathcal{T}_t, \mathbf{y}_t, r_t) = \delta(F_{t+1} - Next(F_t)) p(B_{t+1}|B_t, \mathbf{y}_t). \tag{2}$$

From the computation in Eq. (2), we know that the traffic state transition is Markovian. Given the scheduling policy $\mathbf{y}_t = \pi(\mathcal{T}_t, r_t)$ and transmission rate r_t, the distortion reduction experienced by the wireless user is:

$$u_t(\mathcal{T}_t, \mathbf{y}_t, r_t) = \sum_{f_j^g \in F_t} q_j^g y_j^g. \tag{3}$$

III. Dynamic Optimization for a Single User

In this section, we first consider the optimization of both the packet scheduling and resource acquisition for a single wireless video user experiencing a slow fading wireless channel. In each time slot, the wireless user experiences a channel condition h_t. We assume that the channel condition remains constant within one time slot, but varies across time slots. The changes of can be modelled as a finite state Markov chain (FSMC) [13] with the state transition probability given by $p_h(h_{t+1}|h_t)$, which is independent of the traffic state transition. The transmission rate attained by the wireless user is determined by $r_t(h_t, x_t)$, where $x_t \in X$ represents the amount of network resource (e.g. the transmission time in the TDMA-like network [14] as discussed in Section IV) acquired by the wireless user from the network and X represents the set of possible resource allocations. As we will discuss in Section IV for the multi-user video transmission, the resource acquisition will be affected by other users. The transmission rate function $r_t(h_t, x_t)$ is assumed to be a convex increasing function of x_t, given the channel condition h_t.

We define the state for the wireless user at time slot t as $s_t = (\mathcal{T}_t, h_t) \in S$, which includes the video traffic state and channel state. The state s_t satisfies the Markovian property since both the traffic state and channel state are Markovian. Then, the wireless user state transition is expressed by

$$p(s_{t+1}|s_t, \mathbf{y}_t, x_t) = p_{\mathcal{T}}(\mathcal{T}_{t+1}|\mathcal{T}_t, \mathbf{y}_t, r_t) p_h(h_{t+1}|h_t). \tag{4}$$

At each state s_t, the wireless user takes the actions including the resource acquisition x_t and scheduling $\mathbf{y}_t$, thereby leading to the immediate utility $u_t(s_t, \mathbf{y}_t, x_t) - \lambda_{s_t} x_t$, where λ_{s_t} is interpreted as the resource price as in [1]. Note that we express $u_t(\mathcal{T}_t, \mathbf{y}_t, r_t(h_t, x_t))$ as $u_t(s_t, \mathbf{y}_t, x_t)$ to emphasize that the immediate utility is a function of the state s_t, scheduling action $\mathbf{y}_t$ and allocated time x_t.

In this section, we assume that λ_{s_t} is determined a priori. In Section V, we will discuss how the resource price can be determined in a multi-user scenario. The wireless user's objective is to maximize its expected discounted accumulated utility[7] (we call this "single-user primary problem (SUP)"):

SUP:

$$\max_{\left\{ \substack{\mathbf{y}_t \in \mathcal{P}(s_t, x_t) \\ x_t \geq 0, t \geq 0} \right\}} \sum_{s_0 \in S} v(s_0) E \left\{ \sum_{t=0}^{\infty} \alpha^t (u_t(s_t, \mathbf{y}_t, x_t) - \lambda_{s_t} x_t) | s_0 \right\}$$

where α is the discounted factor in the range[8] of $[0, 1)$, and $v(s_0)$ is the distribution of the initial state. The reasons why

[6] Here $F_{t+1}/F_t = F_{t+1} - F_{t+1} \cap F_t$

[7] In this formulation, we interchangeably express the set of admissible policies as $\mathcal{P}(\mathcal{T}_t, r_t)$ and $\mathcal{P}(s_t, x_t)$.

[8] Our solutions discussed below are also applicable to the problem of maximizing the average accumulated utility by allowing $\alpha \to 1$.

we consider the discounted accumulated utility are: (1) for our considered delay-sensitive applications, the data needs to be sent out as soon as possible to avoid missing delay deadlines; and (2) since a wireless user may encounter unexpected environmental dynamics in the future, it may care more about its immediate reward rather than the future reward. Based on the discussion in Section II, the transition of the state s_t only depends on the current action $a_t = (x_t, \mathbf{y}_t)$. Hence, the problem above can be formulated as an MDP.

Note that unlike the previous video transmission solutions in [2][3][4], here we explicitly take into consideration the heterogeneous characteristics of video traffic (represented by the traffic states) and time-varying channel conditions (represented as channel states). Similar to the work in [1], we optimize the trade-off between the consumed resource and the received reward in terms of distortion reduction, but unlike in [1] we focus on a dynamic setting. The optimization of SUP is called a foresighted optimization for video transmission because it considers the impact of current decisions on the future utility. Based on [9], the optimization of SUP can be solved using the Bellman's equations (5), where $U(s, \boldsymbol{\lambda})$ is the optimal reward-to-go starting at state s, given $\boldsymbol{\lambda} = [\lambda_s]_{s\in S}$. The Bellman's equations can be solved using the value iteration or policy iteration methods [9].

We define (6). Given the resource price $\boldsymbol{\lambda}$, $H(s, \boldsymbol{\lambda}, x)$ represents the utility function the wireless video user obtains in state s under the optimal mixed packet scheduling policy. The optimal mixed packet scheduling policy is characterized in detail in [23] which leads to a concave utility function $H(s, \boldsymbol{\lambda}, x)$.

***Lemma** 1:* $H(s, \boldsymbol{\lambda}, x)$ is a concave function of the allocated resource x.

The proof is given in Appendix A.

The concavity of the utility function $H(s, \boldsymbol{\lambda}, x)$ plays an important role in deriving the optimal resource allocation solution in the multi-user video transmission, which is presented in the subsequent sections.

IV. Multi-User Wireless Video Transmission Formulation

In this section, we aim to formulate the problem of multi-user multimedia transmission over a slowly-fading wireless channel by considering both the heterogeneous traffic and time-varying channel conditions experienced by the users. The users are indexed by $i \in \{1, \cdots, M\}$, where M is the number of users sharing the channel. Similarly, we define the state for the wireless user i at time slot t as $s_t^i = (\mathcal{T}_t^i, h_t^i) \in S^i$ (the superscript i represents user i and the same in the below), which includes the video traffic state $\mathcal{T}_t^i$ and channel state h_t^i. As discussed in Section II, the traffic state $\mathcal{T}_t^i$ models the heterogeneous characteristics of the delay-sensitive video data. The channel state h_t^i represents the channel conditions experienced by the wireless user i. The channel state transition probability $p_h(h_{t+1}^i|h_t^i)$ is independent of the traffic state transition and also other users' state transition.

We assume that the multiple users access the shared channel using the TDMA-like protocol [14]. At each time slot, the portion of time allocated to user i is denoted by $x_t^i \in [0, 1]$. The allocations to all the users satisfy the following inequalities: $\sum_{i=1}^M x_t^i \le 1, \forall t$, which are referred to as the stage resource constraints. Here one stage means one time slot.

We consider a collaborative multi-user video transmission problem aimed at maximizing the expected discounted accumulated video quality of all the users under the stage resource constraints (we call this problem "the multi-user primary problem with stage resource constraints - MUP/SRC"):

MUP/SRC:

$$U^* = \max_{\mathbf{y}_t^i, x_t^i, i=1,\cdots,M} \sum_{s_0^1\in S^1,\cdots,s_0^M\in S^M} \prod_{i=1}^M v(s_0^i) E\left[\sum_{t=0}^{\infty}\sum_{i=1}^M \alpha^t u_t^i(s_t^i, \mathbf{y}_t^i, x_t^i)|\mathbf{s}_0\right]$$

$$s.t. \mathbf{y}_t^i \in \mathcal{P}^i(s_t^i, x_t^i), \sum_{i=1}^M x_t^i \le 1$$

where $v(s_0^i)$ is the distribution of the initial state of user i and is assumed to be independent of other users'.

The multi-user transmission problem in MUP/SRC can be formulated as an MUMDP. Specifically, we define the state of the multi-user system as $mathbfs = (s^1, \cdots, s^M)$. The action performed by each user is $a^i = (\mathbf{y}^i, x^i)$ and the action profile for all the users is $\mathbf{a} = (a^1, \cdots, a^M)$. It is easy to verify that $p(\mathbf{s}'|\mathbf{s}, \mathbf{a}) = \prod_{i=1}^M p^i(s^{i\prime}|s^i, \mathbf{y}^i, x^i)$ since the traffic state and channel state are independent across the users. The reward at each time slot is given by $u_t = \sum_{i=1}^M u_t^i(s_t^i, \mathbf{y}_t^i, x_t^i)$. We note that, when $\alpha = 0$ (i.e. all users make myopic decisions), the MUMDP problem reduces to the traditional multi-user NUM-based resource allocation problems for video transmission [6]-[8], which is performed repeatedly. However, we consider here the dynamic optimization for the multi-user video transmission by taking into account the resource allocation and corresponding scheduling across time (i.e. $\alpha \neq 0$).

From [9], we know that, for this multi-user MDP problem, there is at least one optimal stationary policy that only depends on the current multi-user system state. Hence, in this paper, we restrict our focus to the stationary policies, i.e. the policy only depends on the current state. Then, solving the maximization problem in MUP/SRC is equivalent to solving the following Bellman's equations [9]:

$$U(\mathbf{s}) = \max_{\mathbf{y}^i\in\mathcal{P}^i(s^i,x^i), i=1,\cdots,M, \sum_{i=1}^M x^i \le 1} \left[\sum_{i=1}^M u_i(s^i, \mathbf{y}^i, x^i) + \alpha \sum_{\mathbf{s}'}\prod_{i=1}^M p(s^{i\prime}|s^i, \mathbf{y}^i, x^i) U(\mathbf{s}')\right], \forall \mathbf{s} \quad (7)$$

and $U^* = \sum_{s_0^1\in S^1,\cdots,s_0^M\in S^M} \prod_{i=1}^M v(s_0^i) U(\mathbf{s_0})$.

Based on these Bellman's equations, we can make the following observations: (i) to solve the Bellman's equations, we can use the centralized value iteration or policy iteration [9] to find the optimal state function $U(\mathbf{s})$ for the multi-user MDP problem. However, this solution requires knowing all the users' information (state spaces, action spaces, transition probabilities, and utility functions) and also has a prohibitively high computation complexity. Hence, this centralized solution is not applicable for the multi-user wireless video transmission; (ii) we note that the coupling among the multiple users' video

$$U(s,\boldsymbol{\lambda}) = \max_{\left\{\substack{\mathbf{y}\in\mathcal{P}(s,x)\\ x\geq 0}\right\}} \left[u(s,\mathbf{y},x) - \lambda_s x + \sum_{s'} \alpha p(s'|s,\mathbf{y},x) U(s',\boldsymbol{\lambda}) \right], \tag{5}$$

$$H(s,\boldsymbol{\lambda},x) = \max_{\mathbf{y}\in\mathcal{P}(s,x)} \left[u(s,\mathbf{y},x) - \lambda_s x + \sum_{s'} \alpha p(s'|s,\mathbf{y},x) U(s',\boldsymbol{\lambda}) \right] \tag{6}$$

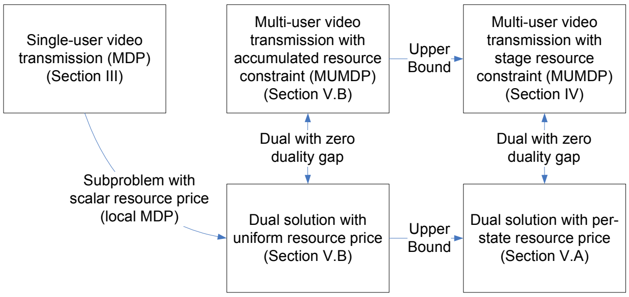

Single-user video transmission (MDP) (Section III)
Multi-user video transmission with accumulated resource constraint (MUMDP) (Section V.B)
Upper Bound
Multi-user video transmission with stage resource constraint (MUMDP) (Section IV)
Dual with zero duality gap
Dual with zero duality gap
Subproblem with scalar resource price (local MDP)
Dual solution with uniform resource price (Section V.B)
Upper Bound
Dual solution with per-state resource price (Section V.A)

Fig. 2. Relationship between the various proposed solutions for the considered multi-user MDP problem

transmission is only through the resource allocation performed at each time slot. The optimal scheduling policy performed by each user i depends on the multi-user system state through the resource allocation x^i. Then, given the resource allocation x^i, the scheduling policy is independent of other users' states. This type of MUMDP is weakly-coupled MDP [10] and thus, the decomposition into multiple local MDPs is possible.

In the next section, we will discuss how the multi-user MDP problem can be decomposed when the resource allocation is dynamic and depends on the multi-user system's state. The relationships among the proposed solutions are illustrated in Figure 2.

V. Dual Decomposition of MUMDP

In this section, we will relax the per-stage resource constraints and show how we can decompose the MUMDP. First, in Subsection A, we introduce a per-state Lagrangian multiplier associated with the resource constraint at each state. This dual solution leads to the zero duality compared to the primary problem MUP/SRC, but requires a centralized solution since the resource price depends on multi-user state which cannot be observed by each individual user. Then, in Subsection B, we impose a uniform resource price, which is independent of the multi-user state. With this uniform resource price, the MUMDP problem can be decomposed into multiple local MDPs, each of which represents a dynamic cross-layer optimization problem that can be separately solved by each individual user. This decomposition is promising since (i) it enables each user to perform the cross-layer optimization independently of other users; and (ii) the network coordinator only needs to simply update the resource price, which involves only very few computations.

A. Dual solution with per-state resource prices

At each state $\mathbf{s}_t$, we introduce a Lagrangian multiplier $\lambda_{\mathbf{s}_t}$ associated with the resource constraint $\left(\sum_{i=1}^M x_t^i - 1\right)$ at each state $\mathbf{s}_t$. Then the dual function is given by

$$U(\boldsymbol{\lambda}) = \max_{\left\{\substack{\mathbf{y}_t^i\in\mathcal{P}^i(s_t^i,x_t^i),x_t^i\geq 0\\ i=1,\cdots,M,t\geq 0}\right\}} \sum_{s_0^i\in S^1,\cdots,s_0^M\in S^M} v(s_0^i)$$
$$E\left[\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{M} \left(u_t^i(s_t^i,\mathbf{y}_t^i,x_t^i) - \lambda_{\mathbf{s}_t} x_t^i + \frac{\lambda_{\mathbf{s}_t}}{M} \right) |\mathbf{s}_0 \right], \tag{8}$$

with $\boldsymbol{\lambda} = [\lambda_{\mathbf{s}}]$. We refer to $\lambda_{\mathbf{s}}$ as ”pre-state resource price”. Then, $\lambda_{\mathbf{s}} x^i$ is the cost user i has to pay in state $\mathbf{s}$ and $\lambda_{\mathbf{s}} \cdot 1$ is the amount of revenue received by the multi-user system by allowing the users to consume the resources (i.e. access the wireless channel). However, we should note that, in the considered collaborative communication scenario, the resource price is used in order to efficiently allocate the limited resource, instead of maximizing the revenue of the multi-user system.

The multi-user dual problem with the per-state resource price (MUD/PSRP) is then given by

MUD/PSRP
$$U^{\boldsymbol{\lambda},*} = \min_{\boldsymbol{\lambda}\geq 0} U(\boldsymbol{\lambda}).$$

The following proposition proves that the dual problem MUD/PSRP has zero duality gap compared to the primary problem in MUP/SRC and thus, the optimal time allocation and scheduling policies corresponding to the optimal resource price $\lambda_{\mathbf{s}}$ at each state are also optimal policies for the primary problem.

***Proposition** 2:* $U^{\boldsymbol{\lambda},*} = U^*$

The sketch of the proof is given in Appendix B.

Similar to the Bellman's equations in Eq. (7) for the primary problem MUP/SRC, we have the following Bellman's equations corresponding to the dual function in MUD/PSRP:

$$U(\mathbf{s},\boldsymbol{\lambda}) = \max_{\mathbf{y}^i\in\mathcal{P}(s^i,x^i),x^i\geq 0,i=1,\cdots,M}$$
$$\left[\left\{ \begin{array}{c} \sum_{i=1}^M (u_i(s^i,\mathbf{y}^i,x^i) - \lambda_{\mathbf{s}} x^i + \frac{1}{M}\lambda_{\mathbf{s}}) + \\ \alpha \sum_{\mathbf{s}'} \prod_{i=1}^M p(s^{i\prime}|s^i,\mathbf{y}^i,x^i) U(\mathbf{s},\boldsymbol{\lambda}) \end{array} \right\} \right], \forall \mathbf{s}. \tag{9}$$

We note that, by setting $\alpha = 0$, the Bellman's equations above degrade to the dual solutions [6][7] to the conventional multi-user video transmission. The degraded Bellman's equations can be decomposed into multiple sub-equations, each corresponding to one user, by letting the user know the resource price. However, in general, this Bellman's equation cannot be decomposed into independent subproblems which can be autonomously solved by each user, since the Bellman's equations are coupled through the resource price $\lambda_{\mathbf{s}}$, which varies with the state of the multi-user system. Hence, a centralized solution has to be deployed by the network coordinator, which requires all users' information, as in the primary solution to MUP/SRC.

B. Dual solution with uniform resource price

In this subsection, we consider a scenario where the same price (referred to as "uniform resource price") is imposed in all the states of the multi-user system, i.e. $\lambda_{\mathbf{s}} = \lambda, \forall \mathrm{s}$. Then, the dual function is given by

$$U(\lambda) = \max_{\left\{\substack{\mathbf{y}_t^i \in \mathcal{P}^i(s_t^i, x_t^i), x_t^i \geq 0 \\ i=1,\cdots,M, t\geq 0}\right\}} \sum_{s_0^1 \in S^1, \cdots, s_0^M \in S^M} \prod_{i=1}^{M} v(s_0^i)$$
$$E\left[\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{M} \left(u_t^i(s_t^i, \mathbf{y}_t^i, x_t^i) - \lambda x_t^i + \frac{\lambda}{M}\right) |\mathbf{s}_0 \right] \quad (10)$$

By minimizing over the uniform resource price λ, we have the multi-user dual problem with uniform resource price (MUD/URP):

MUD/URP

$$U^{\lambda,*} = \min_{\lambda \geq 0} U(\lambda)$$

We would like to note that the Lagrangian relaxation using uniform resource price has been also proposed in [10][11] to decompose a weakly-coupled MUMDP problem. In the below, we mathematically derive the duality and propose a systematic way to compute the subgradient for updating the Lagrangian multiplier which is not explicitly addressed in [10][11]. Interestingly, by setting the uniform resource price, the dual problem MUD/URP is not dual to the primary problem in MUP/SRC. Instead, it is dual to the following problem:

MUP/ARC

$$\hat{U}^* = \max_{\mathbf{y}_t^i, x_t^i, i=1,\cdots,M} \sum_{s_0^1 \in S^1, \cdots, s_0^M \in S^M} \prod_{i=1}^{M} v(s_0^i)$$
$$E\left[\sum_{t=0}^{\infty} \sum_{i=1}^{M} \alpha^t u_t^i(s_t^i, \mathbf{y}_t^i, x_t^i) |\mathbf{s}_0 \right]$$
$$s.t. \mathbf{y}_t^i \in \mathcal{P}^i(s_t^i, x_t^i), \sum_{t=0}^{\infty} \sum_{i=1}^{M} (x_t^i - \frac{1}{M}) \leq 0$$

We call this optimization - "the multi-user primary problem with accumulated resource constraint (MUP/ARC)". The duality between MUD/URP and MUP/ARC can be easily verified. Similar to Proposition 1, we can prove that the duality gap between MUD/URP and MUP/ARC is zero.

We further notice that the resource constraint in the primary problem MUP/SRC satisfies the following condition:

$$\left\{ x_t^i, i = 1, \cdots, M, t \geq 0 | \sum_{i=1}^{M} x_t^i \leq 1 \right\} \subset$$
$$\left\{ x_t^i, i = 1, \cdots, M, t \geq 0 | \sum_{t=0}^{\infty} \sum_{i=1}^{M} x_t^i - \frac{1}{M} \leq 0 \right\}, \quad (11)$$

which means that the feasible resource allocations in the MUP/SRC is a subset of the feasible resource allocations in the MUP/ARC. Then, comparing to the dual solution with state-wise prices, we have the following proposition which shows that $U^{\lambda,*}$ serves as an upper bound of the optimal value for the primary problem.

***Proposition** 3:* $U^{\lambda,*} = \hat{U}^* \geq U^* = U^{\boldsymbol{\lambda},*}$

The proof is given in Appendix C.

The following theorem shows that the Bellman's equations corresponding to the dual function in Eq. (10) can be decomposed into M local Bellman's equations, each corresponding to one user.

***Theorem** 4:* Given $\lambda_{\mathbf{s}} = \lambda, \forall \mathbf{s}$, the optimization in Eq. (10) is given by

$$U(\lambda) = \sum_{i=1}^{M} \sum_{s_0^i} v(s_0^i) U^i(s_0^i, \lambda), \quad (12)$$

with $U^i(s^i, \lambda)$ satisfying the local Bellman's equation:

$$U^i(s^i, \lambda) = \max_{x^i \geq 0, \mathbf{y}^i \in \mathcal{P}(s^i, x^i)}$$
$$\left[\left\{ \begin{array}{c} u^i(s^i, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda + \\ \alpha \sum_{s^{i\prime}} p(s^{i\prime}|s^i, \mathbf{y}^i, x^i) U^i(s^{i\prime}, \lambda) \end{array} \right\} \right] \quad (13)$$

The proof is given in Appendix D.

The key result of Theorem 4 is that $U(\lambda)$ can be decomposed into M local Bellman's equations, which can be solved in a distributed fashion. Each user can solve its own Bellman's equations (and accordingly solve its own cross-layer optimization problem) provided the resource price λ. This local Bellman's equations correspond to the local MDP, which is the single-user cross-layer optimization solved by each individual user (see Figure 2).

Next, we discuss how the resource price can be updated. Given the resource price λ, each user can solve its own Bellman's equations using e.g. value iteration, which results in the optimal resource allocation $x^{i,*}(s^i, \lambda)$ and scheduling policy $\mathbf{y}^{i,*}(s^i, \lambda)$. Note that the resource acquisition is independent of other users' state given the uniform resource price. In the following proposition, we formally compute the subgradient with respect to the resource price λ for the dual problem MUD/URP, which will be used to update the resource price in each iteration.

***Proposition** 5:* The subgradient with respect to λ is given by

$$\sum_{i=1}^{M} Z^i - \frac{1}{1-\alpha}, \quad (14)$$

where $Z^i = \sum_{s_0^i \in S^i} v(s_0^i) \mathbf{e}_{s_0^i}^T (I - \alpha P^i)^{-1} \mathbf{x}^i(\lambda)$ is the expected discounted accumulated resource consumption (note that the expectation is taken over all the possible sample paths), and P^i is the state transition probability matrix, and $\mathbf{e}_{s^i}$ is the vector with the s^i component being 1 and others being zero.

The proof is given in Appendix E.

Using the subgradient method, the resource price is then updated as follows:

$$\lambda^{k+1} = \left[\lambda^k + \beta^k (\sum_{i=1}^{M} Z^i - \frac{1}{1-\alpha}) \right]^+ \quad (15)$$

where β^k is a diminishing step-size which satisfies the following conditions: $\sum_{k=1}^{\infty} \beta^k = \infty, \sum_{k=1}^{\infty} (\beta^k)^2 < \infty$ [15]. One example is $\beta^k = \frac{1}{k}$. We notice that the subgradient computed in Eq. (14) accounts for not only the current resource constraint, but also the future constraints, since MUMDP aims to maximize the long-term utility. The subgradient method

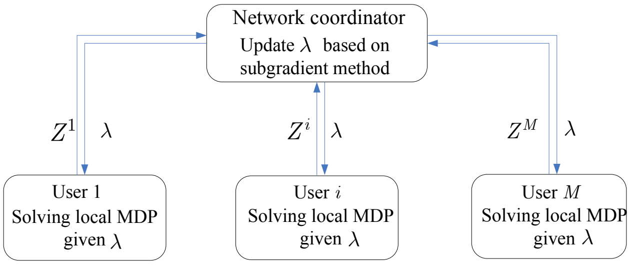

Network coordinator
Update λ based on subgradient method
Z^1
λ
Z^i
λ
Z^M
λ
User 1
Solving local MDP given λ
User i
Solving local MDP given λ
User M
Solving local MDP given λ

Fig. 3. Message exchange in the dual method with uniform price for the multi-user wireless video transmission

shown in Eq. (15) converges to the optimal dual solution due to the concavity of the objective function. The advantages of the proposed decomposition for multi-user video transmission are: (i).given a uniform resource price, each wireless user can solve its own local MDP independently of other users. This enables us to decompose the multi-user video transmission problem by enabling each user to autonomously optimize its packet scheduling and resource acquisition; (ii) this decomposition allows the network coordinator to simply update the scalar resource price. Furthermore, the proposed approach only requires two scalar messages (as shown in Figure 3) to be exchanged between the wireless users and the network coordinator at each iteration. This significantly simplifies the design of the network coordinator (e.g. access points) and reduces the cost of building a wireless network to support video applications.

Previously, we enforced a uniform price for all the states, which enables a decomposition of the dual function computation and provides an upper bound on the utility function U^*. However, the solution to the dual problem may be infeasible (i.e. it may violate the resource constraint in each time slot, $\sum_{i=1}^{M} x^{i,\lambda,*}(s^i) > 1$). Recall that the optimal allocation is given by the solution to Eq. (7). We can approximate the optimal allocation by solving a one-stage multi-user resource allocation problem at each time slot in which we replace the optimal $U(\mathbf{s}')$ in Eq. (7) with the approximated state value function $\sum_{i=1}^{M} U^i(s^{i'}, \lambda)$. This one-stage resource allocation problem becomes:

$$U(\mathbf{s}) = \max_{\sum_{i=1}^{M} x^i \leq 1} \sum_{i=1}^{M} V^i(s^i, x^i), \tag{16}$$

where

$$V^i(s^i, x^i) = \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i)} \left\{ u^i(s^i, \mathbf{y}^i, x^i) + \alpha \sum_{s^{i'}} p(s^{i'} | s^i, \mathbf{y}^i, x^i) U^i(s^{i'}, \lambda) \right\} \tag{17}$$

This one-stage multi-user resource allocation problem is similar to the traditional multi-user NUM-based resource allocation problems. The difference is that, in our formulation, the utility function $V^i(s^i, x^i)$ is the long-term utility instead of the immediate utility as in the conventional NUM formulation. Since the resource allocation is a one-stage optimization and the network is decentralized, it can neither be solved using distributed iteration-based methods as in the conventional NUM, nor directly by the network coordinator, which requires the state transition probability and state value function to compute $V^i(s^i, x^i)$. From Section V.B, we know that each user is able to compute its optimal resource acquisition $x^{i,*}(s^i, \lambda)$ with respect to the current uniform price λ. Comparing Eq. (17) to Eq. (13), we note that

$$x^{i,\lambda,*}(s^i, \lambda) = \arg\max_{x^i} \left\{ V^i(s^i, x^i) - \lambda x^i \right\} \tag{18}$$

Instead of directly solving the one-stage resource allocation in Eq. (16), we propose a heuristic method by scaling down the resource acquisition computed in (18) and weighted by the average gradient of each user, i.e.

$$\hat{x}^{i,\lambda}(\mathbf{s}) = \frac{x^{i,\lambda,*}(s^i, \lambda) \Delta \bar{V}^i}{\sum_{j=1}^{M} x^{j,\lambda,*}(s^j, \lambda) \Delta \bar{V}^j} \tag{19}$$

where $\Delta \bar{V}^i = \frac{1}{2}\left[\Delta V^i(s^i, 0) + \Delta V^i(s^i, x^{i,\lambda,*})\right] = \frac{1}{2}\left[\Delta V^i(s^i, 0) + \lambda\right]$. Here we note that $\Delta V^i(s^i, x^{i,\lambda,*}) = \lambda$ from Eq. (18) and $\Delta V^i(s^i, 0)$ can be computed by user i. We call the resource allocation in Eq. (19) the gradient-based resource allocation scaling where the gradient is the average derivative of the long-term utility of each user.

This scaling can be performed by the network coordinator as follows: at the beginning of each time slot, the users submit the weighted resource acquisition $x^{i,\lambda,*} \Delta \bar{V}^i$ to the network coordinator, and the coordinator performs the resource allocation scaling as in Eq. (19). After the scaling, the network coordinator polls the users according to the scaled resource allocation [14]. Note that, the resource allocations $\hat{x}^{i,\lambda}(\mathbf{s}), \forall i$ satisfy the resource constraints and hence, it provides the lower bound on the optimal utility. In Section VII.B, we qualify the duality gap by showing the upper and lower bounds on the cross-layer optimization performance.

VI. Distributed Online Learning

When implementing the distributed solution proposed in Section V.B in practice, we face the following difficulties: (i) in this distributed solution, each user still has to solve its own local MDP problem for each updated resource price, which still leads to a very high computation complexity for each user; (ii). the multiple iterations introduce the latency which may not be acceptable for real-time video transmission; (iii) the channel state transition probability and incoming DUs' distribution are often difficult to characterize a priori, such that the single-user MDP cannot be explicitly solved online; However, the proposed distributed solution to the MUMDP provides the necessary foundations and principles for how the users can autonomously learn on-line to cooperatively optimize the global long-term video quality. We aim in this section at developing an online learning algorithm. Specifically, we deploy a modified reinforcement learning algorithm [16] to solve the single-user MDP and the stochastic subgradient method [15] to update the uniform resource price. The advantages of the online learning solution are: at each time slot, the wireless user only needs to perform a few simple computations (i.e. incurs a low computation complexity); and, most importantly, there is no a priori knowledge requirement for the channel states and incoming DUs dynamics.

A. Post-decision state-based online learning

Given the uniform resource price λ, user i is able to solve the local MDP problem as in Eq. (13). As discussed in Section II, the incoming data is independent of the buffer size. Furthermore, from Section III, we know that the traffic state transition is also independent of the channel state. In this section, we define a post-decision state $\tilde{s}_t^i = (\tilde{\mathcal{T}}_t^i, \tilde{h}_t^i)$ for user i. The post-decision traffic state $\tilde{\mathcal{T}}_t^i = (\tilde{F}_t^i, \tilde{B}_t^i)$ represents the traffic state after the packet scheduling but before the expired DUs are deleted and the new DUs arrive. The post-decision channel state $\tilde{h}_t^i$ is the same as the current channel state, i.e. $\tilde{h}_t^i = h_t^i$. It is clear that $\tilde{F}_t^i = F_t^i$ because the DU types are not changed after the packet transmission. However, the buffer state $\tilde{B}_t^i$ changes: $\tilde{B}_t^i = B_t^i - \mathbf{y}_t^i$. With the post-decision state in mind, the Bellman equation in Eq. (13) to solve the local MDP can be rewritten as

$$\begin{aligned}U^i(\tilde{s}_t^i, \lambda) = &\sum_{b_j^{i,g}, F_{t+1}^i, h_{t+1}^i} \prod_{f_j^g \in F_{t+1}^{i,d} \cap (F_{t+1}^i / F_t^i)} PMF_j^i(b_j^{i,g}) \\ &\delta(F_{t+1}^i - next(\tilde{F}_t^i)) p(h_{t+1}^i | \tilde{h}_t^i) \max_{x_{t+1}^i \geq 0, \mathbf{y}_{t+1}^i \in \mathcal{P}^i(s_{t+1}^i, x_{t+1}^i)} \\ &\left[\left\{ \begin{array}{l} u^i(s_{t+1}^i, \mathbf{y}_{t+1}^i, x_{t+1}^i) - \lambda x_{t+1}^i + \frac{1}{M}\lambda + \\ \alpha U^i(((F_{t+1}^i, B_{t+1}^i - \mathbf{y}_{t+1}^i), h_{t+1}^i), \lambda) \end{array} \right\} \right],\end{aligned} \tag{20}$$

where the next post-decision state is $\tilde{s}_{t+1}^i = ((F_{t+1}^i, B_{t+1}^i - \mathbf{y}_{t+1}^i), h_{t+1}^i)$ and $U^i(\tilde{s}_t^i, \lambda)$ is the post-decision state value function. Using the post-decision state, we are able to separate the expectation (corresponding to the first line in the right hand side of Eq. (20)) from the optimization. At each time slot, we first perform the cross-layer optimization without worrying about the expectation. After the optimization, we then take the expectation over all the possible state s_t^i. This separation provides us a new way to learn the optimal resource allocation and packet scheduling online, which is discussed below.

B. Batch update of post-decision state value function

From the discussion in Section VI.A, we note that the channel state transition and incoming data dynamics are independent of the buffer state, the resource allocation and the packet scheduling. In the conventional reinforcement learning (e.g. Q-learning, actor-critic learning [16]), the state-value function can be updated only one state per time slot, which leads to slow convergence rate. However, in our context, we are able to update the post-decision state value function at multiple states per time slot.

Specifically, at time slot t, the DU type is F_t^i and channel state is h_t^i, and the set of the feasible post-decision states at time slot t is $\tilde{S}_t^i = \left\{ ((F_t^i, \tilde{B}_t^i), h_t^i), \forall \tilde{B}_t^i \right\}$. Note that the realized post-decision state $\tilde{s}_t^i$ belongs to $\tilde{S}_t^i$. When the expired DUs are deleted, the new DUs arrive and the new channel state h_{t+1}^i is realized, any post-decision state $\tilde{s} \in \tilde{S}_t^i$ transits to the normal state $s^{i'} = ((F_{t+1}^i, B_{t+1}^i), h_{t+1}^i) \in S_{t+1}^i$. In conventional reinforcement learning, the state value function is often updated in the current normal state s_t^i. In contrast, in our cross-layer optimization problem, we can update the post-decision state value function in all the possible post-decision state $\tilde{s}^i \in \tilde{S}^i$ at time slot $t+1$ as follows:

$$\begin{aligned}U^i(\tilde{s}^i, \lambda) = (1 - \gamma_t^i) U^i(\tilde{s}^i, \lambda) + \gamma_t^i &\max_{x^i \geq 0, \mathbf{y}^i \in \mathcal{P}^i(s^{i\prime}, x^i)} \\ &\left[u^i(s^{i\prime}, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda + \alpha U^i(s^{i\prime}, \lambda) \right], \forall \tilde{s}^i \in \tilde{S}^i.\end{aligned} \tag{21}$$

where $\tilde{s}^{i\prime} = ((F^i, B^i - \mathbf{y}^i), h^i)$ and γ_t^i satisfies $\sum_{k=1}^{\infty} \gamma_k^i = \infty, \sum_{k=1}^{\infty} (\gamma_k^i)^2 < \infty$. The maximization in Eq. (21) is performed to obtain the optimal packet scheduling and resource acquisition at the normal state $s^{i\prime}$ which is transited to from the post-decision state $\tilde{s}^i$. Further, to enforce the convergence of the resource price and state value function, γ_k^i should also satisfy $\lim_{k \to \infty} \frac{\beta^k}{\gamma_k^i} = 0$ as shown in [20]. One example is $\gamma_l^i = \frac{1}{k^{0.7}}$. Then, the optimal resource allocation and packet scheduling corresponding to the normal state s_{t+1}^i is implemented at time slot $t+1$, which leads to the actual packet transmission. However, the optimal resource allocation and packet scheduling corresponding to other states is only computed but not implemented. The batch update of the post-decision state-value function in Eq. (21) can significantly improve the learning performance (i.e. it reduces the number of time slots required to achieve a specific distortion reduction) as compared to the conventional on-line learning algorithm. Furthermore, we note that the batch update of the post-decision state value function is not affected by the allocated resource by the network coordinator. Actually, it is only affected by the announced resource price.

C. Stochastic subgradient-based resource price update

From Section V.B, we notice that the subgradient of the dual problem with uniform price is computed as in Eq. (14), which is the expected discounted accumulated resource consumption. Since each wireless user does not know the transition probability, we only use the realized sample path to estimate the subgradient of the dual problem (i.e. using the stochastic subgradient). Specifically, we update the Lagrangian multiplier as follows:

$$\lambda_{k+1} = \left[\lambda_k + \kappa_k \left(\sum_{i=1}^{M} \sum_{t=0}^{\infty} \alpha^t x_t^i - \frac{1}{1-\alpha} \right) \right]^+ \tag{22}$$

where $\sum_{t=0}^{\infty} \alpha^t x_t^i$ is the stochastic subgradient approximating the subgradient Z^i and κ_k is a diminishing step-size satisfying $\sum_{k=1}^{\infty} \kappa_k^i = \infty, \sum_{k=1}^{\infty} (\kappa_k^i)^2 < \infty$. One example is $\kappa_k = \frac{1}{k}$. However, in practice, we cannot wait for an infinite time to update the Lagrangian multiplier. Instead, we update the multiplier every K time slots, i.e. we use $\tilde{Z}^i = \sum_{t=kK}^{(k+1)K-1} \alpha^{t-kK} x_t^i$ instead of $\sum_{t=0}^{\infty} \alpha^t x_t^i$. The proposed online learning algorithm is illustrated in Figure 4. It can be shown that the batch update on the post-decision state value function and subgradient-based resource price update will converge to the optimal solution [21].

VII. Simulation Results

In this section, we present simulation results highlighting the efficiency of the proposed single-user and multi-user video transmission solutions compared to existing solutions. To

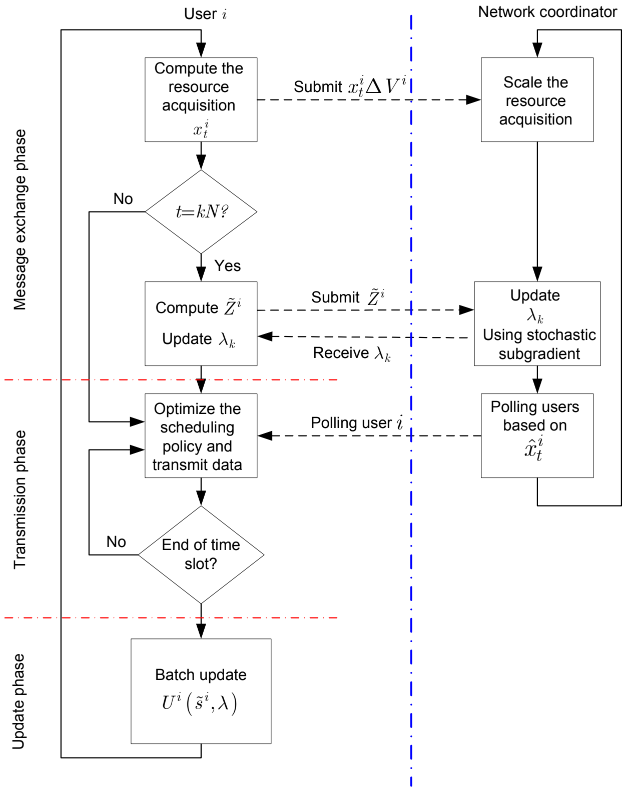

User i
Network coordinator
Message exchange phase
Transmission phase
Update phase
Compute the resource acquisition x_t^i
Submit $x_t^i \Delta V^i$
Scale the resource acquisition
No
$t = kN?$
Yes
Compute $\tilde{Z}^i$
Update λ_k
Submit $\tilde{Z}^i$
Receive λ_k
Update λ_k Using stochastic subgradient
Optimize the scheduling policy and transmit data
Polling user i
Polling users based on $\hat{x}_t^i$
No
End of time slot?
Batch update $U^i(\tilde{s}^i, \lambda)$

Fig. 4. Flowchart of the online learning algorithm

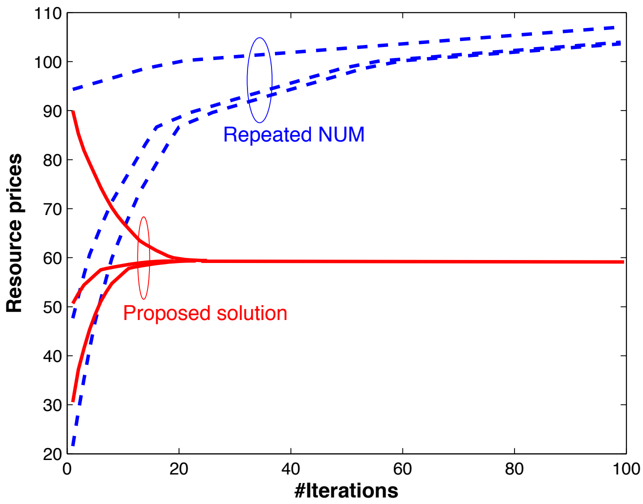


Fig. 5. Convergence of the dual solutions under various initial resource price selection

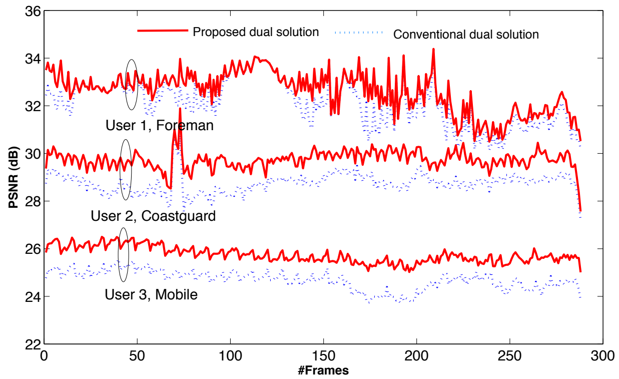


Fig. 6. PSNR of received video sequences

compress the video data, we used a scalable video coding scheme [12], which is attractive for wireless streaming applications because it provides on-the-fly application adaptation to channel conditions, support for a variety of wireless receivers with different resource capabilities and power constraints, and easy prioritization of various coding layers and video packets.

A. Dual solutions with uniform price

In this section, we will verify the convergence of the dual solution with uniform price to the proposed MUMDP. We will further compare the performance of our approach to that of the conventional multi-user dual solution. We first consider three wireless users: User 1 streams the video sequence ”Foreman” (CIF resolution, 30 Hz), User 2 streams the video sequence ”Coastguard” (CIF resolution, 30 Hz) and User 3 streams the video sequence ”Mobile” (CIF resolution, 30 Hz). We compare our proposed dual solution with uniform price to the conventional dual solution [7] based on the NUM framework. Figure 5 shows the convergence of the resource prices with various initial price selections. We notice that, our proposed dual solution with uniform price shows much faster convergence (less than 25 iterations) than the conventional dual solution (having more than 100 iterations). We also note that our solution converges to a lower resource price than the conventional one. This is because that the conventional solution myopically maximizes the video transmission over each time slot. Hence, to achieve a feasible resource allocation, it has to increase its resource price to ensure that the resource allocations over all the states are feasible (corresponding to the worst case scenario.) However, in our solution, we relax the stage resource constraints into the accumulated resource constraint (shown in the problem of MUP/ARC) and the resource price is reduced. We scale down the resource acquisition at each multi-user system state to enforce the feasible allocation. Figure 6 shows the improvements in terms of PSNR when the price converges (i.e. User 1 receives 0.5dB higher PSNR, User 2 receives 1dB higher PSNR and User 3 receives 1.1dB higher PSNR). The improvement is due to the foresighted decisions in our solution, as compared to the myopic decisions in the conventional NUM-based solution.

B. Duality gap of proposed dual decomposition

In this section, we illustrate the duality gap of the proposed the dual decomposition with uniform price. As we know, the resource acquisitions $x^{i,\lambda,*}(s^i), \forall i$ with respect to the optimal uniform price λ^* provide the upper bound on the optimal utility U^*, while the feasible resource allocations $\hat{x}^{i,\lambda}(\mathbf{s}), \forall i$ provide the lower bound on U^*. Thus, the duality gap must be less than the difference between the utilities obtained by $x^{i,\lambda,*}(s^i), \forall i$ and $\hat{x}^{i,\lambda}(\mathbf{s}), \forall i$.

The simulation settings are the same as in Section VII.A. We consider two scenarios in which the users experience average channel conditions of 22dB and 28dB, respectively. The upper and lower bounds of the received video qualities

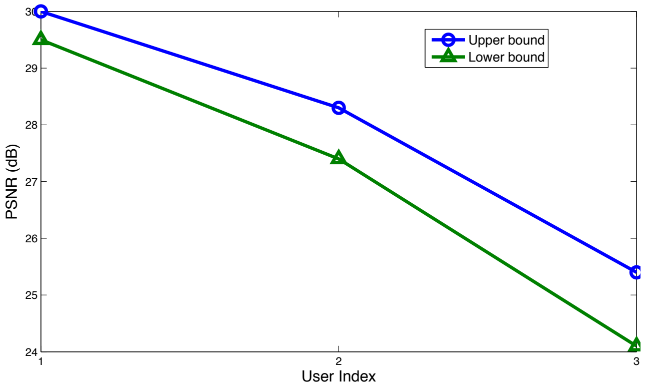


Fig. 7. Upper and lower bounds on the received video qualities under average channel condition of 22dB

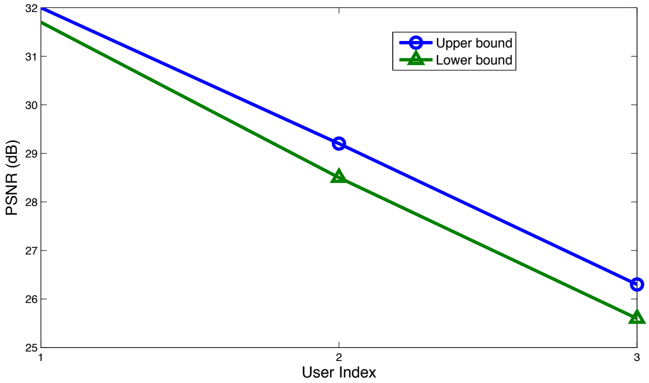


Fig. 8. Upper and lower bounds on the received video qualities under average channel condition of 28dB

for the users are illustrated in Figures 7 and 8. It shows that, when the average channel condition is 22 dB, the difference between the upper bound and the lower bound is 0.5dB, 0.9dB and 1.3 dB for the three users, respectively. While the average channel condition is increased to 28dB, the difference becomes 0.3dB, 0.7dB and 0.7dB, respectively. These simulation results show that, when the channel conditions improve, the difference between the upper and lower bounds decreases. This is because, when the channel conditions improve (or the available resource is more plentiful), the stage resource constraints become loose and the dual decomposition proposed in Section V provides a more accurate approximation on the optimal utility.

C. Online learning

In this section, we will verify the convergence rate of our proposed online learning algorithm and corresponding impact on the video transmission. We also compare our algorithm to the conventional online learning algorithm [16], which is often used to improve the wireless transmission strategies with unknown dynamics [22]. We consider three wireless users streaming video sequences as in Section A. Different from the settings in Subsection A, we assume that all the users initially do not have any statistical information about the channel conditions and incoming data, thereby not knowing the state transitions. Using the proposed online learning, the wireless users keep improving their own resource acquisition policy. The resource price is updated every 2 time slots. Figure 9 shows the average reward received by each user deployed with the proposed online learning and the standard online learning, separately. From this figure, we notice that, compared to the conventional learning algorithm, our proposed method can significantly increase the learning curve (i.e. significantly increasing the average reward). Figure 10 shows the received video quality (in terms of PSNR) of each user over time when using these two learning algorithms. This result further confirms that our proposed learning algorithm can improve the video quality of all the users over time. On average, our proposed algorithm improves the video quality of User 1 by 0.9 dB, User 2 by 1.2 dB and User 3 by 1.4dB in terms of PSNR. This improvement is due to the fact that our proposed approach can update the policy at multiple states during one time slot and hence, exhibits a fast convergence rate.

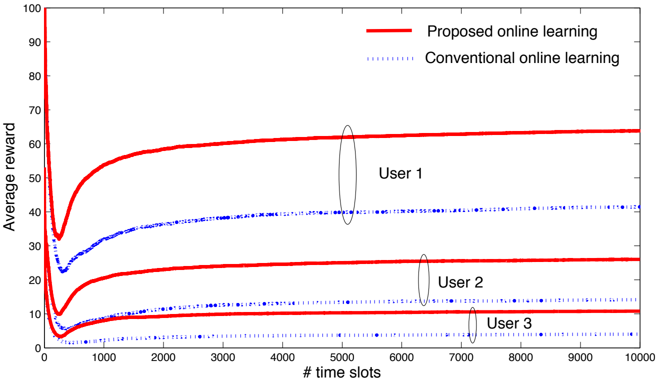


Fig. 9. Learning curves with different online learning algorithms

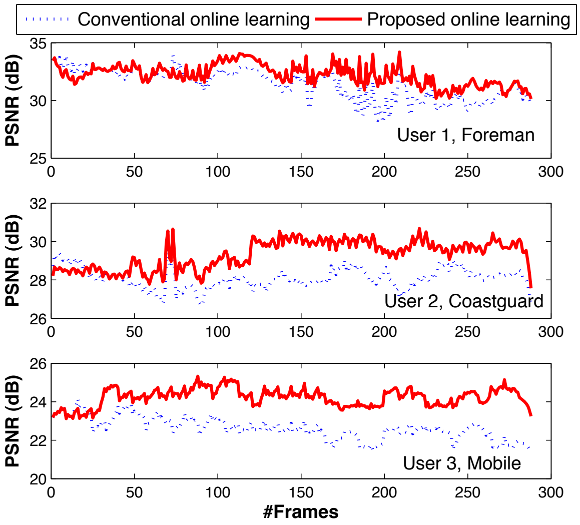


Fig. 10. Received PSNR with different online algorithms

VIII. Conclusion

In this paper, we systematically formulate the dynamic multi-user video transmission as an MUMDP problem in order to account for the heterogeneous video traffic characteristics and dynamic wireless network conditions. This MDP formulation allows the wireless users to make foresighted decisions in order to maximize the long-term video quality instead of the immediate reward, which is essential for video applications. The proposed distributed dynamic optimization approach using Lagrangian relaxation with an uniform resource price allows each wireless user to maximize its own video quality given the resource price. To deal with the unknown video characteristics and channel conditions, and to reduce the computation complexity for each user, a novel online reinforcement learning algorithm has been developed, which allows wireless users to update their transmission policy in multiple states during one time slot, thereby significantly accelerating the learning speed and improving the received video quality.

Appendix A
Proof of Lemma 1

Proof: To prove the concavity of $H(s, \boldsymbol{\lambda}, x)$, we only need to show that the optimal mixed packet scheduling policy always schedules the packet with the highest marginal utility (i.e. the immediate utility minus the future utility if the packet is delayed for transmission, which is computed as in [23]). First, we note that the optimal mixed packet scheduling policy will always schedules the packet which has no parent packets or whose parent packets have been transmitted. This is because scheduling this packet will lead to successful decoding at the receiver side. If $j' \prec j$, then $q_{j'} > q_j$. When both DUs j' and j are available for transmission, the optimal packet scheduling policy transmits DU j' first due to this dependency, which automatically leads to a higher net utility contribution. If DUs j' and j do not depend on each other, the optimal mixed packet scheduling policy will choose the DU with the higher marginal utility for transmission. The utility function is the summation of the marginal utility of the transmitted packets. Hence, the optimal mixed packet scheduling policy gives a concave non-decreasing utility function $H(s, \boldsymbol{\lambda}, x)$ at state s. The properties of optimal packet scheduling policy are described in detail in [23].

Appendix B
Proof of Proposition 2

Sketch of proof: To prove the zero duality gap, we only need to show that the primary MUMDP is a convex optimization with the resource allocation variable $\mathbf{x}$ under the mixed packet scheduling policy. We first notice that the constraint on the variable $\mathbf{x}$ is convex. To prove the concavity of the objective function, we can use the backward induction. We note that the optimal mixed packet scheduling policy always transmits first the packet with higher marginal utility as shown in [23]. Then, at each stage t, given the state value function $U_{t+1}(\mathbf{s})$ computed at stage $t+1$ with initial value being 0, the utility at each state is non-decreasing and concave which is computed as follows.

$$H_t(\mathbf{s}, \mathbf{x}) = \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i), i=1,\cdots,M} \left[\sum_{i=1}^{M} u^i(s^i, \mathbf{y}^i, x^i) + \alpha \sum_{\mathbf{s}'} \prod_{i=1}^{M} p(s^{i'} | s^i, \mathbf{y}^i, x^i) U_{t+1}(\mathbf{s}') \right], \forall \mathbf{s}.$$

This is because that the utility function is the summation of the marginal utility of all the transmitted packets.

Appendix C
Proof of Proposition 3

Proof: From Proposition 2, we know that there is no duality gap between MUP/SRC and MUD/PSRP. We can further show that there is also no duality gap between MUP/ARC and MUD/URP). We also note the fact that the feasible resource allocations in MUP/SRC is a subset of the feasible allocations in MUP/ARC as shown in Eq. (11). Hence, the optimal value given by MUP/SRC is not greater than the one obtained by MUP/ARC, which proves the statement in proposition 3.

Appendix D
Proof of Theorem 4

Proof: We prove this by induction.

We define

$$U_0(\mathbf{s}, \lambda) = \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i), x^i \geq 0} \sum_{i=1}^{M} \left[u^i(s^i, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda \right].$$

It is easy to verify that $U_0(\mathbf{s}, \lambda) = \sum_{i=1}^{M} U_0^i(s^i, \lambda)$ with

$$U_0^i(s^i, \lambda) = \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i), x^i \geq 0} \left[u^i(s^i, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda \right]$$

Similarly we have $U_1(\mathbf{s}, \lambda) = \sum_{i=1}^{M} U_1^i(s^i, \lambda)$ with

$$U_1^i(s^i, \lambda) = \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i), x^i \geq 0} \left[u^i(s^i, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda + \alpha \sum_{s^{i'}} p(s^{i'} | s^i, \mathbf{y}^i, x^i) U_0^i(s^{i'}, \lambda) \right]$$

Recursively, we have

$$U(\mathbf{s}, \lambda) = \lim_{n \to \infty} U_n(\mathbf{s}, \lambda) = \lim_{n \to \infty} \sum_{i=1}^{M} U_n^i(s^i, \lambda) = \sum_{i=1}^{M} U^i(s^i, \lambda).$$

where

$$U^i(s^i, \lambda) = \sum_{\mathbf{y}^i \in \mathcal{P}^i(s^i, x^i), x^i \geq 0} \left[u^i(s^i, \mathbf{y}^i, x^i) - \lambda x^i + \frac{1}{M}\lambda + \alpha \sum_{s^{i'}} p(s^{i'} | s^i, \mathbf{y}^i, x^i) U^i(s^{i'}, \lambda) \right]$$

Appendix E
Proof of Proposition 5

Proof: For each given λ, suppose that $x^{i,*}(s^i, \lambda)$ and $\mathbf{y}^{i,*}(s^i, \lambda), i = 1, cdots, M$ maximize the dual Bellman's equations in Eq. (13) and hence, maximize the objective in Eq. (10). Denote the state transition probability matrix of user i under the actions of $x^{i,*}(s^i, \lambda)$ and $\mathbf{y}^{i,*}(s^i, \lambda)$ by $\mathcal{P}^i$. $P^i_{s^i}$ is the row vector corresponding to the state s^i. Let $U_i^{\lambda',*} = [U_i^{\lambda',*}(s^i)]_{s^i \in S}$ being the column vector. Then, we have

$$U^{\lambda',*}(\mathbf{s}) = \sum_{i=1}^{M} \max_{\mathbf{y}^i \in \mathcal{P}^i(s^i,x^i), x^i \geq 0}$$
$$\left\{ \left[u^i(s^i,\mathbf{y}^i,x^i) - \lambda' x^i + \lambda' \frac{1}{M} \right] \right.$$
$$\left. + \alpha \sum_{s^{i\prime}} p(s^{i\prime}|s^i,\mathbf{y}^i,x^i) U_i^{\lambda',*}(s^{i\prime}) \right\}$$
$$\geq \sum_{i=1}^{M} \left\{ \tilde{U}_0^i(s^i,\lambda) + (\lambda - \lambda')(x^i(s^i,\lambda) - \frac{1}{M}) + \alpha P_{s^i}^i \mathbf{U}^{i,\lambda',*} \right\},$$

Where

$$\tilde{U}_0^i(s^i,\lambda) = \sum_{i=1}^{M} \left[u^i(s^i,\mathbf{y}^i(s^i,\lambda),x^i(s^i,\lambda)) - \lambda x^i(s^i,\lambda) + \lambda \frac{1}{M} \right].$$

Recursively applying this inequality into $U^{i,\lambda',*}(s^{i\prime})$, we further have

$$U^{\lambda',*}(\mathbf{s}) \geq \sum_{i=1}^{M}$$
$$\left\{ \left\{ \begin{array}{c} \tilde{U}_0^i(s^i,\lambda) + \alpha P_{s^i}^i \tilde{\mathbf{U}}_0^i + (\lambda - \lambda') \\ \left[x^i(s^i,\lambda) - \frac{1}{M} + \alpha P_{s^i}^i (\mathbf{x}^i - \frac{1}{M}) \right] + \alpha^2 P_{s^i}^i P^i \mathbf{U}^{i,\lambda',*} \end{array} \right\} \right\}.$$

Finally, we have

$$U^{\lambda',*}(\mathbf{s}) \geq U^{\lambda,*}(\mathbf{s}) +$$
$$(\lambda - \lambda') \left(\sum_{i=1}^{M} \mathbf{e}_{s^i}^T (I - P^i)^{-1} \mathbf{x}^i(\lambda) - \frac{1}{1-\alpha} \right)$$

where $\mathbf{e}_{s^i}$ is a vector with the s^i component being 1 and others being zero. Hence, the subgradient with respect to $\lambda_{\mathbf{s}'}$ is given by

$$\left(\sum_{i=1}^{M} \mathbf{e}_{s^i}^T (I - P^i)^{-1} \mathbf{x}^i(\lambda) - \frac{1}{1-\alpha} \right).$$

REFERENCES

[1] P. Chou, and Z. Miao, "Rate-distortion optimized streaming of packetized media," IEEE Trans. Multimedia, vol. 8, no. 2, pp. 390-404, 2005.

[2] Z. Li, F. Zhai, and A.K. Katsaggelos, "Joint Video Summarization and Transmission Adaptation for Energy-Efficient Wireless Video Streaming," EURASIP J. Advances in Signal Processing, special issue on Wireless Video, vol. 2008.

[3] B. Lamparter, A. Albanese, M. Kalfane, and M. Luby, "PET-priority encoding transmission: a new, robust and efficient video broadcast technology," Proc. ACM Multimedia, 1995.

[4] S. Khan, M. Sgroi, Y. Peng, E. Steinbach, W. Kellerer, "Application-driven Cross Layer Optimization for Video Streaming over Wireless Networks," IEEE Commun. Mag. , pp. 122-130, January 2006.

[5] X. Zhang and Q. Du, "Cross-Layer Modeling for QoS-Driven Multimedia Multicast/Broadcast Over Fading Channels in Mobile Wireless Networks," IEEE Commun. Mag., pp. 62–70, August 2007.

[6] M. Chiang, S. H. Low, A. R. Caldbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," Proc. IEEE, vol. 95, no. 1, 2007.

[7] J. Huang, Z. Li, M. Chiang, and A.K. Katsaggelos, "Joint Source Adaptation and Resource Allocation for Multi-User Wireless Video Streaming," IEEE Trans. Circuits Syst. Video Technol., vol. 18, issue 5, 582-595, May 2008.

[8] G-M. Su, Z. Han, M. Wu, and K.J.R. Liu, "Joint Uplink and Downlink Optimization for Real-Time Multiuser Video Streaming Over WLANs," IEEE J. Sel. Topics Signal Process., vol. 1, no. 2, pp. 280-294, August 2007.

[9] D. P. Bertsekas, "Dynamic programming and optimal control," 3rd, Athena Scientific, Massachusetts, 2005.

[10] J. Hawkins, "A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications," PhD Dissertation, MIT, Cambridge, MA, 2003.

[11] D. Adelman, and A. J. Mersereau, "Relaxation of weakly coupled stochastic dynamic programs," Operations Research, vol. 56, no. 3, pp. 712-727, May-June 2008.

[12] J.R. Ohm, "Three-dimensional subband coding with motion compensation", IEEE Trans. Image Process., vol. 3, no. 5, Sept 1994.

[13] Q. Zhang, S. A. Kassam, "Finite-state Markov Model for Reyleigh fading channels," IEEE Trans. Commun. vol. 47, no. 11, Nov. 1999.

[14] "IEEE 802.11e/D5.0, wireless medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) enhancements for Quality of Service (QoS), draft supplement," June 2003.

[15] D. P. Bertsekas, "Nonlinear programming," Belmont, MA: Athena Scientific, 2nd Edition, 1999.

[16] R. S. Sutton, and A. G. Barto, "Reinforcement learning: an introduction," Cambridge, MA:MIT press, 1998.

[17] D. S. Turaga and T. Chen, "Hierarchical Modeling of Variable Bit Rate Video Sources," Packet Video, 2001.

[18] Q. Li, Y. Andreopoulos, and M. van der Schaar, "Streaming-Viability Analysis and Packet Scheduling for Video over QoS-enabled Networks," IEEE Trans. Veh. Technol., vol. 56, no. 6, pp. 3533-3549, Nov. 2007.

[19] A. Reibman, and A. Berger, "Traffic descriptors for VBR video teleconferencing over ATM networks," IEEE/ACM Trans. Netw., vol. 3, no. 3, pp. 329-339, 1995.

[20] V. Borkar, and V. Konda, "The actor-critic algorithm as multi-time-scale stochastic approximation," Sadhana, vol. 22, part 4, pp. 525-543, Aug. 1997.

[21] V.S. Vorkar, "An actor-critic algorithm for constrained Markov decision processes," System and Control Letters, vol 54, 207-213, 2005.

[22] C. Pandana, and K. J. R. Liu, "Near-optimal reinforcement learning framework for energy-aware sensor communications," IEEE J. Sel. Areas Commun. vol. 23, no. 4, April, 2005.

[23] F. Fu, M. van der Schaar, "Structural solutions for cross-layer optimization of wireless multimedia transmission," Technical Report, June, 2009. Available on line: http://medianetlab.ee.ucla.edu/papers/UCLATechReport-CLO.pdf.

Fangwen Fu (Student Member, IEEE) received the bachelor's and master's degrees from Tsinghua University, Beijing, China, in 2002 and 2005, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, University of California, Los Angeles. During the summer of 2006, he was an Intern with the IBM T. J. Watson Research Center, Yorktown Heights, NY. During the summer of 2009, he was an intern with DOCOMO USA Labs, Palo Alto, CA. He was selected by IBM Research as one of the 12 top Ph.D. students to participate in the 2008 Watson Emerging Leaders in Multimedia Workshop in 2008. He received Dimitris Chorafas Foundation Award in 2009. His research interests include wireless multimedia streaming, resource management for networks and systems, stochastic optimization, applied game theory, video processing, and analysis.

Mihaela van der Schaar (Fellow, IEEE) received the Ph.D. degree from Eindhoven University of Technology, The Netherlands, in 2001. She is currently an Associate Professor with the Department of Electrical Engineering, University of California, Los Angeles. Since 1999, she has been an active participant in the ISO MPEG standard, to which she made more than 50 contributions. She is an Editor (with P. Chou) of Multimedia over IP and Wireless Networks: Compression, Networking, and Systems (New York: Academic, 2007). She has received 30 U.S. patents. Prof. van der Schaar received the National Science Foundation CAREER Award in 2004, the IBM Faculty Award in 2005, 2007, and 2008, the Okawa Foundation Award in 2006, the Best IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Paper Award in 2005, the Most Cited Paper Award from the EURASIP Journal Signal Processing: Image Communications from 2004 to 2006, and three ISO Recognition Awards. She was on the editorial board of several IEEE Journals and Magazines.