Downlink Power Control for Multi-User VBR Video Streaming in Cellular Networks

Yingsong Huang, *Student Member, IEEE*, and Shiwen Mao, *Senior Member, IEEE*

***Abstract*—We investigate the problem of downlink power control for streaming multiple variable bit rate (VBR) videos in a multicell wireless network, where downlink capacities are limited by inter-cell interference. We adopt a deterministic model for VBR video traffic that considers video frame sizes and playout buffers at the mobile users. The problem is to find the optimal transmit powers for the base stations, such that VBR video data can be delivered to mobile users without causing playout buffer underflow or overflow. We formulate a nonlinear nonconvex optimization problem and prove the condition for the existence of feasible solutions. A centralized branch-and-bound algorithm is then developed, which incorporates the Reformulation-Linearization Technique and can produce $(1-\epsilon)$-optimal solutions. We also propose a low-complexity distributed algorithm with fast convergence as an alternative to the centralized algorithm. Through simulations with VBR video traces under fading channels, we find the distributed algorithm can achieve a performance very close to that of the centralized algorithm.**

***Index Terms*—Cross-layer optimization, downlink power control, variable-bit-rate video, video streaming.**

I. Introduction

A. Motivation

With the dramatic advances in wireless networking technology and wireless communication devices, there is an exponentially increasing demand for wireless video service. This trend is driven by the compelling need for ubiquitous access to video content over wireless access networks, and will significantly stress the capacity of existing wireless networks and strongly influence the design of future wireless networks. Tremendous effort is needed in wireless video research to meet this demand. Researchers have explored new wireless technologies to enable high quality video services. For example, video transmission over cognitive radio networks is an emerging area in video communications, where secondary users sense licensed channels and aim to exploit the transmission opportunities in the spectrum holes [2], [3].

Manuscript received December 01, 2012; revised March 01, 2013; accepted April 09, 2013. Date of publication June 20, 2013; date of current version November 13, 2013. This work was supported in part by the US National Science Foundation (NSF) under Grant CNS-0953513. This work was presented in part at IEEE INFOCOM 2011, Shanghai, China, Apr. 2011 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pal Halvorsen.

The authors are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: yzh0002@tigermail.auburn.edu; smao@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2270457

While it is important to develop new wireless architectures and technologies for higher spectral efficiency, it is equally important to investigate how to support video in existing wireless networks, since the infrastructure will still last for a considerable period of time. In this paper, we consider multi-user video streaming over a multicell wireless network, a wireless network architecture widely deployed all over the world. We consider the typical case of downlink video transmissions. For the multicell system, generally intra-cell interference can be effectively controlled with precise synchronization or the use of guard times. The capacities of the downlinks are mainly limited by the inter-cell interference due to simultaneous base station (BS) transmissions using the same channel. With power control, the Signal to Interference plus Noise Ratio (SINR) of the interfering downlinks can be tuned to combat fading channels and to accommodate video traffic variations. Therefore, effective downlink power control is necessary to support concurrent wireless videos.

We consider the problem of streaming multi-user variable-bit-rate (VBR) videos in the multicell wireless network. This is motivated by the superior perceived quality of VBR videos over constant-bit-rate (CBR) videos. VBR video has stable visual quality for the frames, but at the cost of large variations in the bit rate, while CBR video maintains a stable bit rate, but the frames have large variations in visual quality. We aim to investigate how to provide ubiquitous access to stored VBR videos through existing cellular networks.

It is a challenging problem to support VBR video traffic, which exhibits both strong long-range and short-range dependence. Stochastic models have been developed to capture the burstiness in VBR video traffic. In [4], [5], the authors observed the *long-range-dependence* in VBR video traffic and modeled the autocorrelation with self-similar processes. The stochastic models can be incorporated in QoS mechanisms for VBR videos, and for traffic synthesizing in simulations [6]. Traffic models for MPEG-4 and H.264 are investigated in [7]–[9]. In particular, [7] studied the autocorrelation function (ACF) of frame sizes in MPEG-4 and H.264 VBR video traces. The authors showed that MEPEG-4 and H.264 VBR traces exhibit complex statistic characters, including both long-range dependent and short-range dependent properties. Ref. [8] demonstrated nonstationary statistics of MPEG2 and MPEG4 VBR traffic and developed a bandwidth prediction scheme based on Kalman filter. Ref. [9] provided a survey of VBR traffic models. It was shown that different videos may have quit different characteristics and marginal distributions of frame sizes. Thus it is difficult to find a common statistic model for various VBR videos.

Since it is nontrivial to develop parsimonious traffic models that can accurately capture the auto-correlation structure, and

1520-9210 © 2013 IEEE

the large frame size variations may cause frequent playout buffer underflow or overflow, we address this issue by a deterministic traffic model for stored VBR video, which considers frame size, frame rate, and playout buffers [10]–[13]. Unlike prior work that is focused on a single video session over a given CBR or VBR channel, we exploit power control, a unique capability in wireless networks, to adjust the downlink capacities based on prior knowledge of frame sizes and playout schedules. Usually large frames are rarely transmitted simultaneously. Jointly optimizing the BS transmit powers is, to some extend, analogous to statistically multiplexing VBR videos in the downlink of the cellular network.

B. Related Work

Due to the difficulty in general statistical models for VBR videos, the deterministic model is used to provide a common way to characterize VBR videos, where the varying frame sizes can be represented by the cumulative curves and the frame size variation can be accurately captured. A deterministic model for MPEG4 based neural networks was introduced in [14]. With this approach, the piecewise-constant-rate transmission and transport (PCRTT) method was used, aiming to optimize one or more objectives while preserving continuous video playout. In [10], the authors proposed bandwidth allocation schemes for dynamically sharing a CBR channel among multiple VBR video streams. In [11], Salehi *et al.* considered smoothing VBR video over a CBR link and developed an effective algorithm to achieve the greatest rate smoothness. In [15], McManus and Ross introduced a dynamic programming framework to set PCRTT rates and intervals to optimize different objective functions. These techniques do not directly apply to our problem of VBR over multicell wireless networks, due to the fundamental difference between wireless channels and wired CBR links.

In several recent papers [13], [16]–[18], the authors studied the problem of transmitting one VBR video over wireless networks. In [16], it was shown that the separation between a delay jitter buffer and a decoder buffer is in general suboptimal, and several critical parameters are derived for the system. In [13], the authors studied the frequency of jitters under both network and video system constraint and provided a framework for quantifying the trade-offs among several system parameters. Refs. [17], [18] investigated effective admission control schemes for VBR videos over wireless networks in terms of bandwidth and QoS requirements.

Power control is an important mechanism for interference-limited wireless networks. Most prior work focused on maximizing network utility in the forms of SINR or bit rate [19]–[23]. In [19], the authors presented centralized and distributed power control algorithms for achieving target SINRs. In [20], Chiang studied the problem of joint power control and congestion control, aiming to maximize the throughput of TCP-Vegas over an ad hoc network. Gjendemsj *et al.* [21] presented centralized binary power control algorithms for maximizing the sum rate over multiple interfering links. In [22], the feasibility of distributed target-SIR-tracking power control algorithm was studied and a gradual soft removal method to was proposed to achieve minimum outage. Ref. [23] jointly optimized admission control and power control in cognitive networks. Although laid out the theoretical foundation and developed effective algorithms, these techniques cannot be directly applied for VBR video over multicell wireless networks with buffer and delay constraints.

C. Approach

In this paper, we first present a formulation of multi-user VBR streaming in cellular networks that considers downlink power control, inter-cell interference, VBR video characteristics, and playout buffer requirements. The objective is to achieve high playout buffer utilization, under playout buffer underflow and overflow constraints and peak power constraint. This is a nonlinear nonconvex problem to which traditional convex optimization techniques [20] and low- or high-SINR approximations [20], [21] cannot be directly applied.

We then derive the condition of the existence of feasible power assignments, which can achieve downlink capacities to guarantee no buffer underflow and overflow. We develop a centralized algorithm that can produce solutions with bounded optimality gap. Specifically, we use the Linearization-Reformulation Technique (RLT) to obtain a linear programming (LP) relaxation of the original problem. Solving this LP relaxation yields a lower bound to the original problem. Interestingly, since the constraints are preserved in the relaxation procedure, the lower-bounding solution is also feasible to the original problem; the corresponding objective value with this solution provides a lower bound to the global optimum. The LP relaxation is then incorporated into the branch-and-bound framework to obtain a centralized algorithm, which can produce a solution within the (1-ϵ) range of the global optimal.

To simplify computation and control, we develop a distributed algorithm based on distributed constrained power control (DCPC) [19], where each BS iteratively updates transmit power based on feedback of measured SINR at the target receiver. It is shown that with DCPC, the power vector converges to a unique power vector that can achieve the goal of maximizing playout buffer utilization and avoiding playout buffer underflow and overflow. We evaluate the proposed algorithms with simulations using VBR video traces [24] and fading channels. The distributed algorithm is shown to achieve a performance very close to that of the centralized algorithm. Both algorithms are demonstrated to be highly effective for streaming VBR videos over multicell wireless networks.

In the remainder of this paper, we present the problem formulation in Section II. We describe the centralized algorithm in Section III and the distributed algorithm in Section IV. Simulation results are presented in Section V and related work discussed in Section I-B. Section VI concludes this paper.

II. Problem Statement

A. Network and Video System Model

We consider the downlinks of an M-cell wireless network. In each cell, a BS streams video to mobile users in the cell, each allocated with a downlink channel. A channel is a spectral resource slot, the nature of which depends on the specific multiple access technique adopted for the multicell network. Without loss of generality, we assume that the downlink channels within a cell are orthogonal (e.g., due to perfect synchronization of spreading codes or use of guard times). The main

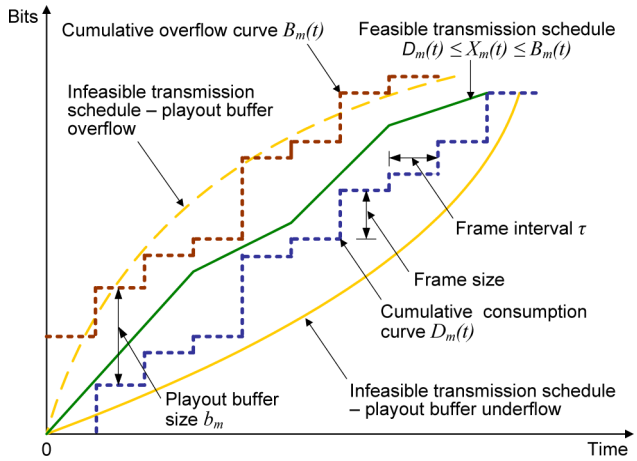


Fig. 1. Feasible and infeasible transmission schedules for video session i.

interference at a user stems from the concurrent downlink transmissions in neighboring cells that use the same channel. There is a need for the BS's to adopt power control to mitigate such inter-cell interference.

We consider the problem of streaming multi-user VBR videos in the multicell network. We assume the wired segment of a video session path is reliable with sufficient bandwidth, while the last-hop wireless link is the bottleneck of the end-to-end path [25]. Thus the corresponding video data is always available at the BS before the scheduled transmission time.

As discussed, it is non-trivial to accurately model VBR video traffic, which exhibits both strong asymptotic self-similarity and short-range correlation [4]. To this end, we adopt a *deterministic model* that considers frame sizes and playout buffers [11]. Let $D_m(t)$ be the *cumulative consumption curve* of user m, representing the cumulative amount of bits consumed by the decoder at time t. The cumulative consumption curve is determined by video characteristics such as frame sizes and rates, and playout schedule. Assume user m's playout buffer is b_m bits and its video has L_m frames. We can derive a *cumulative overflow curve* for user m as

$$B_m(t) = \min\{D_m(t-1)+b_m, D_m(L_m)\},\ 0 \le t \le L_m. \quad (1)$$

$B_m(t)$ is the maximum number of cumulative received bits at time t without overflowing user m's playout buffer. Finally, we define *cumulative transmission curve* $X_m(t)$ as the cumulative amount of bits transmitted to user m at time t. To simplify notation, we assume the video sessions have identical frame rate and the frame intervals are synchronized. Thus a time slot t is equal to the t-th frame interval, denoted as τ, for $0 \le t \le \max_m\{L_m\}$.[1]

Since $D_m(t)$, $B_m(t)$ and $X_m(t)$ are cumulative curves, they are all nondecreasing functions, as illustrated in Fig. 1. A feasible transmission schedule will produce a cumulative transmission curve $X_m(t)$ that lies within $D_m(t)$ and $B_m(t)$, i.e., causing neither underflow nor overflow at the playout buffer. In practice, $D_m(t)$'s are known for stored videos and are delivered to the BS's (or a centralized video scheduler that manages the transmission of multiple BS's) during the session setup phase, and the $B_m(t)$'s are then derived as given in (1).

B. Problem Formation

For the multicell wireless video network, consider a specific channel and let $\mathcal{U} = \{1, 2, \ldots, M\}$ denote the set of users sharing the channel, where user m is located in cell m.[2] Let the BS *transmit power vector* be $\vec{P}(t) = [P_1(t), P_2(t), \ldots, P_M(t)]^T$ in time slot t. The capacity of the downlink from BS m to user m, denoted as $C_m(t)$, depends on the SINR at user m, which can be written as

$$\gamma_m(\vec{P}(t)) = \frac{G_m^m P_m(t)}{\sum_{k\neq m} G_k^m P_k(t) + \eta_m}, \quad (2)$$

where G_k^m is the path gain from BS k to user m and η_m is the noise power at m. We assume slow block fading channels such that the path gains do not change within each time slot [26], but vary over different time slots following a certain distribution. The downlink capacity $C_m(t)$ also depends on the channel bandwidth B_w and the transceiver design, such as modulation and channel coding. Without loss of generality, we use the upper bound as predicted by the Shannon capacity.

$$C_m(\vec{P}(t)) = B_w \log\left(1 + \gamma_m(\vec{P}(t))\right). \quad (3)$$

The impact of fading channels is incorporated in the SINR in (3). For practical systems, the achievable capacity may be a fraction of $C_m(\vec{P}(t))$, but this part is omitted for brevity.

Once the link capacity is determined, $C_m(t)\tau$ video bits will be delivered to user m in that time slot. The cumulative transmission curve $X_m(t)$ can be written as

$$X_m(0) = 0;\ X_m(t) = X_m(t-1) + C_m(t)\tau. \quad (4)$$

Assume peak power constraint $0 \le P_m \le \bar{P}$, for all m. The problem is to determine the transmit power vector $\vec{P}(t)$, for $0 < t \le \max_m\{L_m\}$, such that the resulting cumulative transmission curves satisfy

$$D_m(t) \le X_m(t) \le B_m(t),\ \text{for all } m, t, \quad (5)$$

i.e., without causing playout buffer underflow or overflow. Since the video frames have variable sizes and the video sessions have random phases, large frames from different sessions are less likely to occur in the same time slot. Power control for the downlinks is, in some sense, analogous to exploiting statistical multiplexing gain for VBR video flows.

From (3)–(5), the feasible SINR range at user m is

$$e^{\max\{0, D_m(t)-X_m(t-1)\}/B_w\tau} - 1 \le \gamma_m \le e^{B_m(t)-X_m(t-1)/B_w\tau} - 1. \quad (6)$$

In (6), the lower bound is the SINR that just empties the buffer without causing underflow. The upper bound is the SINR that just fills up the buffer without causing overflow.

Generally, the feasible transmit power vector $\vec{P}(t)$ is not unique for a given set of VBR video sessions. Among the set of feasible solutions, a schedule that transmits more data is more desirable since it provides a larger search space for optimizing

[1]This assumption can be relaxed for more general cases. The time slot duration could be arbitrary as in [26] (i.e., equal to multiple frame intervals).

[2]0–1 index variables can be used to model the case where no user uses the channel in some cells, but are omitted for brevity.

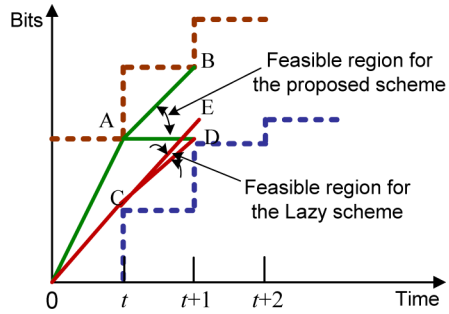


Fig. 2. Illustration the feasible solution spaces with different starting points.

transmit power vectors for future time slots. The main idea is to exploit playout buffers as an effective means to combat fading channels. When the channel is good, the scheme attempts to transmit as much data as possible; when the channel is bad, the video bits stored in the buffer can sustain the playout for some time, thus reducing the buffer underflow rate. The more data transmitted, the bigger the chance to overcome a future underflow event when the channel is bad. This motivation will be verified by the simulation results presented in Section V.

The main idea is illustrated in Fig. 2. The amount of feasible solutions (or, the search space for the optimal solution), depends on the starting point of the transmission scheme in the current time slot (i.e., the buffer occupancy at the beginning of the time slot). Assume that the maximum available power allows a maximum transmission rate as the slope of green line segment A-B in time slot t. If the buffer is full at the beginning of time slot t, the starting point for the transmission schedule is point A. The feasible region for this time slot, under the current power constraint, is in the range between lines A-B and A-D. Alternatively, if the buffer is almost empty at the beginning of time slot t, the starting point of the transmission schedule will be point C. For the given power constraint, the feasible region is in the rage between lines C-D and C-E (note that C-E is parallel to A-B, indicating the same maximum power constraint). Clearly, the feasible region of the latter is much smaller. In the case of starting from point C, if in time slot t there is a deep channel fading, a very large power may be required to achieve the rate indicated by C-D. If this required power is larger than the power constraint, there will be buffer underflow in this time slot. In the case of starting from point A, for the same deep fading channel, a much smaller power can be used to achieve the minimum rate given by A-D, and buffer underflow can be avoided.

Omitting constant B_w, we formulate the optimal power control problem for VBR videos, termed Problem OPT-VBR.

$$\textbf{maximize} \sum_{m\in\mathcal{U}} \log(1+\gamma_m(t)) \tag{7}$$

$$\textbf{subject to}: \gamma_m(t) = \frac{G_m^m P_m(t)}{\sum_{k\neq m} G_k^m P_k(t) + \eta_m}, \ \forall\, m \tag{8}$$

$$\gamma_m^{min}(t) \le \gamma_m(t) \le \gamma_m^{max}(t), \ \forall\, m \tag{9}$$

$$0 \le P_m \le \bar{P}, \ \forall\, m, \tag{10}$$

where $\gamma_m^{max}(t)$ is the upper bound in (6) and $\gamma_m^{min}(t)$ is the larger one between the lower bound in (6) and γ_m^{th}, a minimum SINR requirement imposed by the transceiver design.

In Problem OPT-VBR, the optimization variable are the powers $P_m(t)$. The amount of video data delivered in time slot t is maximized, under playout buffer underflow and overflow constraints and peak transmit power constraints. This is a nonlinear nonconvex problem, to which traditional convex optimization techniques do not directly apply. Furthermore, to achieve the objective of avoiding playout buffer underflow and overflow, the SINRs may assume values ranging from very low to very high. Thus the existing high SINR approximation [20] and low SINR approximation [21] techniques cannot be used. In the following, we first investigate the existence of feasible solutions. We then derive effective centralized and distributed algorithms to solve Problem OPT-VBR in Sections Sections III and IV.

C. Existence of Feasible Solutions

Due to the wide range of VBR video frame sizes, the corresponding SINR requirements also assume a wide range of values. Under conditions where many video sessions coincidently transmit their large frames in the same time slot, Problem OPT-VBR may not have a feasible power assignment to deliver all the frames. In this section, we derive the conditions for the existence of feasible power assignments. We assume a centralized scheduler in the multicell network, which has prior knowledge of all the channel gains and the cumulative consumption and overflow curves.

We define the *minimum required rate* for user m in time slot t, denoted as $C_m^{min}(t)$, as the bit rate such that the playout buffer is just emptied, but without underflow, at the end of time slot t. We have the following result for $C_m^{min}(t)$.

Lemma 1: The largest value for the minimum required rate $C_m^{min}(t)$ is $\bar{C}_m^{min}(t) = [D_m(t) - D_m(t-1)]/\tau$.

Proof: According to the definition of $X_m(t)$ in (4), we have $C_m(t) = [X_m(t) - X_m(t-1)]/\tau$. From the definition of $C_m^{min}(t)$, the playout buffer is emptied at the end of time slot t, i.e., $X_m(t) = D_m(t)$. Therefore, we can derive the minimum required rate as

$$C_m^{min}(t) = \frac{\max\{0, D_m(t) - X_m(t-1)\}}{\tau}. \tag{11}$$

From the feasibility condition (5), we have $X_m(t-1) \ge D_m(t-1)$. Substituting it into (11), we have

$$C_m^{min}(t) \le \frac{[D_m(t) - D_m(t-1)]}{\tau} \equiv \bar{C}_m^{min}(t). \tag{12}$$

Rate $\bar{C}_m^{min}(t)$ occurs when the playout buffer is empty at both the beginning and end of time slot t, but without buffer overflow during the entire time slot. ■

We have the following condition for the existence of a feasible power assignment for Problem OPT-VBR.

Theorem 1: There exits a feasible power assignment for Problem OPT-VBR for time slot t, if there exits a feasible power assignment that can achieve the rate vector $\left[\bar{C}_1^{min}(t), \bar{C}_2^{min}(t), \cdots, \bar{C}_M^{min}(t)\right]$.

Proof: Recall that γ_m^{min} is the SINR corresponding to the minimum required rate $C_m^{min}(t)$. Let $\bar{\gamma}_m^{min}(t)$ be the SINR corresponding to $\bar{C}_m^{min}(t)$. Since (3) is a monotonically increasing function, we have $0 \le \gamma_m^{min}(t) \le \bar{\gamma}_m^{min}(t)$.

We now consider the power assignment that achieves rates $\bar{C}_m^{min}(t)$, or, the corresponding SINRs $\bar{\gamma}_m^{min}(t)$. From (8) and (9), the minimum SINR constraint is

$$\gamma_m(t) = \frac{G_m^m P_m(t)}{\sum_{k\neq m} G_k^m P_k(t) + \eta_m} \geq \bar{\gamma}_m^{min}(t),\ \forall\ m. \quad (13)$$

Equation (13) is a system of linear equations of the power vector $\vec{P}(t)$, which can be written in the matrix form as

$$\left(\mathbf{I} - \bar{\mathbf{\Gamma}}^{min}\mathbf{A}\right)\vec{P}(t) \succeq \bar{\mathbf{\Gamma}}^{min}\vec{\nu}, \quad (14)$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{A}$ is an $M \times M$ matrix with

$$A_{mk} = \begin{cases} 0, & m = k \\ \frac{G_k^m}{G_m^m}, & m \neq k\ , \end{cases} \quad (15)$$

$\bar{\mathbf{\Gamma}}^{min} = \text{diag}\{\bar{\gamma}_1^{min}(t), \bar{\gamma}_2^{min}(t), \ldots, \bar{\gamma}_M^{min}(t)\}$ is a diagonal matrix, and $\vec{\nu} = [\eta_1/G_1^1, \eta_2/G_2^2, \ldots, \eta_M/G_M^M]^T$.

Define $\mathbf{\Gamma}^{min} = \text{diag}\{\gamma_1^{min}(t), \gamma_2^{min}(t), \ldots, \gamma_M^{min}(t)\}$ and $\mathbf{\Delta} = \bar{\mathbf{\Gamma}}^{min} - \mathbf{\Gamma}^{min} \succeq \mathbf{0}$. Assume $\vec{P}$ is a power assignment that achieves $\bar{\gamma}_m^{min}(t)$ for all m, which satisfies (14). Substituting $\bar{\mathbf{\Gamma}}^{min} = \mathbf{\Delta} + \mathbf{\Gamma}^{min}$ into (14), we have $\left(\mathbf{I} - \mathbf{\Gamma}^{min}\mathbf{A}\right)\vec{P} \succeq \mathbf{\Gamma}^{min}\vec{\nu} + \mathbf{\Delta}\left(\vec{\nu} + \mathbf{A}\vec{P}\right)$. Since $\mathbf{\Delta}$, $\vec{\nu}$, $\mathbf{A}$ and $\vec{P}$ all have non-negative elements, we have $\mathbf{\Delta}\left(\vec{\nu} + \mathbf{A}\vec{P}\right) \succeq \mathbf{0}$, and therefore,

$$\left(\mathbf{I} - \mathbf{\Gamma}^{min}\mathbf{A}\right)\vec{P} \succeq \mathbf{\Gamma}^{min}\vec{\nu}. \quad (16)$$

That is, $\vec{P}$ can also achieve $\gamma_m^{min}(t)$ for all m and it satisfies the minimum SINR constraint in (9).

Once the minimum SINR constraint in (9) (i.e., no buffer underflow) is satisfied, the maximum SINR constraint in (9) (i.e., no buffer overflow) can be satisfied since BS m can stop transmission when the playout buffer at user m is full. ■

Theorem 1 allows us to evaluate, for given videos and channel gains, if there is a feasible power assignment for each time slot. There is no need to consider the transmission schedules and playout buffer occupancies in previous time slots. At the beginning of time slot t, we obtain $\bar{\gamma}_m^{min}(t)$ from the cumulative consumption curve $D(t)$ and channel gains. If the linear system (14) is solvable and the resulting $\vec{P}$ satisfies constraint (10), then there is a feasible power assignment for Problem OPT-VBR for this time slot. The following fact from [27] can be used for the feasibility test.

Fact 1: The following statements are equivalent: (i) there exits a feasible power assignment satisfying (14); (ii) the maximum modulus eigenvalue of $\left(\bar{\mathbf{\Gamma}}^{min}\mathbf{A}\right)$ is less than 1; (iii) the reciprocal matrix $(\mathbf{I} - \bar{\mathbf{\Gamma}}^{min}\mathbf{A})^{-1} = \sum_{k=0}^{\infty}\left(\bar{\mathbf{\Gamma}}^{min}\mathbf{A}\right)^k$ exists and is positive component-wise.

D. Comparison With a Lazy Scheme

A "lazy" scheme is proposed in [12] for VBR video transmission over a wired network. This is an ON-OFF scheme and it transmits a video frame as late as possible before its playout deadline at the maximum link speed, which minimizes the required client buffer size. In multicell multi-user wireless VBR video streaming, the maximum link speed varies from time to time due to interference and channel fading. Thus, the original lazy scheme cannot be applied directly.

We enhance the lazy scheme to support multicell multi-user VBR video streaming, termed W-Lazy, where every BS transmits a frame that is needed for playout in the next time slot. Then we can determine the rate vector (and the transmit powers) as given in Theorem 1. We use W-Lazy as a benchmark for comparison and evaluation of the proposed algorithms. We have the following results for W-Lazy.

Corollary 1.1: Problem OPT-VBR has a larger solution space than that of the W-Lazy scheme.

Proof: This result directly follows Theorem 1. ■

Corollary 1.2: If $\vec{C}^*(t) = [C_1^*(t), \ldots, C_n^*(t)]$ is the solution to Problem VBR-OPT, then any other vector $\vec{C}(t)$ that is element-wise smaller than $\vec{C}^*(t)$ has a smaller solution space.

Proof: This result also follows a similar process as in the proof of Theorem 1. ■

III. Centralized Algorithm

In this section, we present a centralized algorithm to provide solutions with bounded optimality gap. We first use RLT to obtain a linear programming (LP) relaxation of Problem OPT-VBR [28]. We then incorporate the linear relaxation into a branch-and-bound framework, which can produce (1-ϵ)-optimal solutions.

A. Reformulation and Linearization

We first apply *polyhedral outer approximation* for the logarithm functions in Problem OPT-VBR to obtain a Polynomial Programming Problem OPT-VBR(p) [28]. We then use *RLT bound-factor product constraints* to relax the quadratic terms to obtain an LP relaxation OPT-VBR(l). The time slot index (t) is dropped in the following to simplify notation.

For the logarithm functions in the objective function, let $u_m = \log(1+\gamma_m)$. We obtain a linear objective function $\sum_{m\in\mathcal{U}} u_m$ and new constraints $u_m = \log(1+\gamma_m)$. We deal with the new constraints using polyhedral outer approximation. Since $\gamma_m^{min} \leq \gamma_m \leq \gamma_m^{max}$, we choose H points, denoted as $\{\gamma_m^h\}$, within this range as

$$\gamma_m^h = (1+\gamma_m^{min})\left(\frac{1+\gamma_m^{max}}{1+\gamma_m^{min}}\right)^{h/H-1} - 1, \quad h = 0, 1, \ldots, H-1, \quad (17)$$

where $\gamma_m^0 = \gamma_m^{min}$ and $\gamma_m^{H-1} = \gamma_m^{max}$. We can obtain a *convex envelop* for the logarithm function in $[\gamma_m^{min}, \gamma_m^{max}]$, which consists of H tangent lines at the H points given in (17) and the line segment connecting the two end points. We relax the logarithm constraint by using its convex envelop, represented by the following new linear constraints:

$$\begin{cases} u_m \geq \frac{\log(1+\gamma_m^{min})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m^{max}-\gamma_m) + \frac{\log(1+\gamma_m^{max})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m-\gamma_m^{min}) \\ u_m \leq \log(1+\gamma_m^h) + \frac{\gamma_m-\gamma_m^h}{1+\gamma_m^h},\ h = 0, 1, \ldots, H-1. \end{cases}$$

The first line is for the segment connecting the two end points, and the second line is for the tangent lines at the H points.

Thus we obtain a polynomial programming problem OPT-VBR(p), as given in (18)∼(25). We can rewrite the last constraint (25) as $\sum_{k\neq m} G_k^m\gamma_m P_k - G_m^m P_m + \eta_m\gamma_m = 0$, which

contains quadratic terms in the form of $\gamma_m P_k$. We next introduce RLT bound-factor product constraints to remove such terms and to obtain an LP relaxation.

$$\textbf{maximize} \sum_{m\in\mathcal{U}} u_m \tag{18}$$

$$\textbf{subject to:}$$

$$G_m^m P_m - \left(\sum_{k\neq m} G_k^m P_k + \eta_m\right)\gamma_m^{min} \geq 0,\ \forall\, m \tag{19}$$

$$G_m^m P_m - \left(\sum_{k\neq m} G_k^m P_k + \eta_m\right)\gamma_m^{max} \leq 0,\ \forall\, m \tag{20}$$

$$0 \leq P_m \leq \bar{P},\ \forall\, m \tag{21}$$

$$u_m \geq \frac{\log(1+\gamma_m^{min})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m^{max}-\gamma_m)+ \frac{\log(1+\gamma_m^{max})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m-\gamma_m^{min}),\ \forall\, m \tag{22}$$

$$u_m \leq \log(1+\gamma_m^h) + \frac{\gamma_m-\gamma_m^h}{1+\gamma_m^h},\ \forall\, m,h \tag{23}$$

$$\gamma_m^h = (1+\gamma_m^{min})\left(\frac{1+\gamma_m^{max}}{1+\gamma_m^{min}}\right)^{h/H-1} - 1, \forall\, m,h \tag{24}$$

$$\gamma_m = \frac{G_m^m P_m}{\sum_{k\neq m} G_k^m P_k + \eta_m},\ \forall\, m. \tag{25}$$

Define substitution variables $v_{mk} = \gamma_m P_k$, for all m, k. Since γ_m and P_k are bounded by their respective lower and upper bounds as $\gamma_m^{min} \leq \gamma_m \leq \gamma_m^{max}$ and $0 \leq P_k \leq \bar{P}$, we obtain the following RLT bound-factor product constraints.

$$\begin{cases} v_{mk} - \gamma_m^{min} P_k \geq 0 \\ \gamma_m^{max} P_k - v_{mk} \geq 0 \\ \gamma_m \bar{P} - v_{mk} - \gamma_m^{min}\bar{P} + \gamma_m^{min} P_k \geq 0 \\ \gamma_m^{max}\bar{P} - \gamma_m^{max} P_k - \gamma_m \bar{P} + v_{mk} \geq 0. \end{cases}$$

The quadratic terms $P_k\gamma_m$ are thus replaced with v_{mk} with the above linear RLT bound-factor constraints, and an LP relaxation OPT-VBR (l) is obtained as given in (26)∼(37).

$$\textbf{maximize} \sum_{m\in\mathcal{U}} u_m \tag{26}$$

$$\textbf{subject to}:$$

$$G_m^m P_m - \left(\sum_{k\neq m} G_k^m P_k + \eta_m\right)\gamma_m^{min} \geq 0,\ \forall\, m \tag{27}$$

$$G_m^m P_m - \left(\sum_{k\neq m} G_k^m P_k + \eta_m\right)\gamma_m^{max} \leq 0,\ \forall\, m \tag{28}$$

$$0 \leq P_m \leq \bar{P},\ \forall\, m \tag{29}$$

$$u_m \geq \frac{\log(1+\gamma_m^{min})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m^{max}-\gamma_m)+ \frac{\log(1+\gamma_m^{max})}{\gamma_m^{max}-\gamma_m^{min}}(\gamma_m-\gamma_m^{min}),\ \forall\, m \tag{30}$$

$$u_m \leq \log(1+\gamma_m^h) + \frac{\gamma_m-\gamma_m^h}{1+\gamma_m^h},\ \forall\, m,h \tag{31}$$

$$\gamma_m^h = (1+\gamma_m^{min})\left(\frac{1+\gamma_m^{max}}{1+\gamma_m^{min}}\right)^{h/H-1} - 1, \forall\, m,h \tag{32}$$

$$v_{mk} - \gamma_m^{min} P_k \geq 0,\ \forall\, m, k\neq m \tag{33}$$

$$(\gamma_m - \gamma_m^{min})\bar{P} - v_{mk} + \gamma_m^{min} P_k \geq 0,\ \forall\, m, k\neq m \tag{34}$$

$$\gamma_m^{max} P_k - v_{mk} \geq 0,\ \forall\, m, k\neq m \tag{35}$$

$$(\gamma_m^{max} - \gamma_m)\bar{P} - \gamma_m^{max} P_k + v_{mk} \geq 0,\ \forall\, m, k\neq m \tag{36}$$

$$\sum_{k\neq m} v_{mk} G_k^m - G_m^m P_m + \eta_m\gamma_m = 0,\ \forall\, m. \tag{37}$$

The LP relaxation OPT-VBR(l) can be effectively solved with an LP solver in polynomial time. The optimal solution to the LP relaxation consists of $\{\vec{P}^*, \vec{u}^*, \vec{\gamma}^*, \vec{v}^*\}$. During the reformulation and linearization procedure, we mainly relax the logarithm function in the objective function of OPT-VBR. The original constraints of OPT-VBR are preserved in OPT-VBR (l). Therefore, we have the following theorem regarding the feasibility of the solution, which greatly simplifies the *local search* procedure of the branch-and-bound algorithm to be presented in Section III-B.

Theorem 2: The optimal transmit power vector $\vec{P}^*$ to the LP relaxation OPT-VBR (l) is a feasible solution to the original problem OPT-VBR.

Proof: The optimal transmit power vector $\vec{P}^*$ of problem OPT-VBR (l) satisfies the power constraint (28), which is identical to the power constraint (10) in problem OPT-VBR. Since the solution $\vec{P}^*$ also satisfies SINR constraints (26) and (27), it can be easily shown to satisfy the SINR constraints (8) and (9) in problem OPT-VBR. Thus the optimal transmit power vector $\vec{P}^*$ of the relaxed problem OPT-VBR (l) is also feasible to the original problem OPT-VBR. ■

B. Branch-and-Bound Algorithm

According to Theorem 2, we can substitute the optimal power assignment $\vec{P}^*$ for the LP relaxation into Problem OPT-VBR to obtain a lower bound, while the LP solution itself provides an upper bound. We next incorporate the LP relaxation into a branch-and-bound framework to obtain an algorithm that can produce (1-ϵ)-optimal solutions.

Branch-and-bound is an iterative method for solving optimization problems, especially for discrete and combinatorial problems. A branch-and-bound procedure has two key components. The first one, called *branching*, is to partition a problem into subproblems. The procedure is repeated recursively to each of the subproblems and all produced subproblems naturally form a tree structure, i.e., the *branch-and-bound tree*. Its nodes are the constructed subproblems. The leaves of the tree is also call the *Problem List*. The other component is *bounding*, which is a fast way of finding upper and lower bounds for the optimal solution for each subproblem. For a maximization problem, an infeasible upper bound (UB) can be found by solving a relaxed problem. A *local search* algorithm is then used to explore the neighborhood, to find a feasible lower-bounding solution (LB). As discussed, we can easily derive upper and lower bounds by solving the LP relaxation (no need for local search). The core of the approach is an observation that, for a maximization task, if the upper bound for a subproblem l_1 is smaller than the lower bound for any other subproblem l_2, then l_1 and the branch rooted at l_1 can be safely discarded from the tree, such that the computational complexity can be reduced. This procedure is called *pruning*.

The algorithm terminates when the upper bound reaches $(1+\epsilon)$ of the lower bound. Let the optimal object value be $O \leq UB$, we have $LB \geq 1/1+\epsilon UB \geq 1/1+\epsilon O = (1-\epsilon+\epsilon^2-\epsilon^3+\cdots)O \approx (1-\epsilon)O$, for $0 \leq \epsilon \ll 1$. The pseudo code for the branch-and-bound algorithm is given in Algorithm 1.

Algorithm 1 Branch-and-Bound Algorithm

1 Initialization

2 Obtain LP relaxation OPT-VBR (l) as Prob 1;

3 Set optimal solution $sol = \phi$, Problem list $\mathcal{S} = \{\text{Prob } 1\}$, $UB = \infty$, and $LB = 0$;

4 Solve Prob 1 for solution $\{\vec{P}', \vec{u}', \vec{\gamma}', \vec{v}'\}$ and upper bound UB_1;

5 Use $\vec{P}'$, **(7)**, and **(8)** to get lower bound LB_1;

6 Set $UB = UB_1$ and $LB = LB_1$;

7 Iteration & pruning

8 Select Prob l with the largest UB_l in $\mathcal{S}$ and set $UB = UB_l$;

9 **If** $LB_l > LB$ **then**

10 Set $sol = \vec{P}'_l$ and $LB = LB_l$;

11 **If** $UB \leq (1+\epsilon)LB$ **then**

12 stop with solution sol;

13 **else**

14 remove all probs k in $\mathcal{S}$ with $UB_k \leq (1+\epsilon)LB$;

15 **end**

16 **end**

17 Partition

18 For Prob l, find the maximum relaxation error among all RLT variables, e.g., $\max_{m,k}\{|\gamma_m P_k - v_{mk}|\}$;

19 **If** $(\gamma_m^{max} - \gamma_m^{min}) \cdot \min\{\gamma'_m - \gamma_m^{min}, \gamma_m^{max} - \gamma'_m\}$ $\geq (P_m^{max} - P_m^{min}) \cdot \min\{P'_m - P_m^{min}, P_m^{max} - P'_m\}$ **then**

20 partition $[\gamma_m^{min}, \gamma_m^{max}]$ into $[\gamma_m^{min}, \gamma'_m]$ and $[\gamma'_m, \gamma_m^{max}]$;

21 **else**

22 partition $[P_m^{min}, P_m^{max}]$ into $[P_m^{min}, P'_m]$ and $[P'_m, P_m^{max}]$;

23 **end**

24 Bounding

25 Solve the partitioned probs l_1 and l_2 to get solutions sol_{l_1}, sol_{l_2} and bounds $UB_{l_1}, UB_{l_2}, LB_{l_1}, LB_{l_2}$;

26 Remove Prob l from $\mathcal{S}$;

27 **If** $(1+\epsilon)LB < UB_{l_1}$ **then**

28 add Prob l_1 into $\mathcal{S}$;

29 **end**

30 **if** $(1+\epsilon)LB < UB_{l_2}$ **then** ;

31 add Prob l_2 into $\mathcal{S}$

32 **end**

33 **if** $\mathcal{S} = \phi$ **then**

34 stop;

35 **else**

36 go to Step 6;

37 **end**

C. Enhancement

We further introduce a heuristic to accelerate the convergence of the branch-and-bound algorithm. At the beginning of time slot t, if the playout buffer level is above a threshold, say, 80%, and $X_m(t-1) \geq D_m(t)$ at user m, we set $P_m(t) = 0$ and remove the link from the optimization process.

Generally the playout buffer size should be at least greater than the largest frame size. Given the large variations in VBR frame sizes, there could be multiple frames stored when the buffer is close to full. When the above conditions are satisfied, there is little chance of buffer underflow at the end of time slot t even if we do not transmit anything to user m. On the other hand, if we schedule a non-zero power $P_m(t)$ for this link, only a small amount of bits can be transmitted due to the buffer overflow constraint, but at the cost of reduced SINRs at all other links. Excluding such links from transmission not only greatly speeds up the convergence of the branch-and-bound algorithm, but also increases the SINR and capacity of other active links.

IV. Distributed Algorithm

A. Develop a Distributed Algorithm

Although the RLT-based branch-and-bound algorithm can provide a $(1-\epsilon)$-optimal solution, it requires a centralized implementation. A centralized controller is needed to collect network, link and video related information, and to update transmit power for each downlink. In this section, we develop a distributed algorithm for Problem OPT-VBR that can be implemented in each BS and operate with local information.

We assume each BS obtains video cumulative consumption curves and playout buffer sizes for its users during the video session initiation phase. At the beginning of time slot t, each BS m computes for user m the minimum rate as $[D_m(t) - X_m(t-1)]/\tau$, i.e., the data rate that empties the playout buffer at the end of time slot t but without underflow, and the maximum rate as $[B_m(t) - X_m(t-1)]/\tau$, i.e., the data rate that makes the playout buffer full at the end of time slot t but without overflow. BS m then translates the minimum and maximum rates to minimum and maximum SINRs, i.e., $\gamma_m^{min}(t)$ and $\gamma_m^{max}(t)$ as given in (6). In the following, we again drop the time slot index (t) to simplify notation.

To maximize objective function (7), BS m sets a target SINR as $\gamma_m^{tar} = \gamma_m^{max}$, and tries to achieve the target SINR by adjusting its transmit power. The problem then becomes a *Distributed Constrained Power Control* (DCPC) problem [19]. BS m first randomly sets its initial transmit power as $0 < P_m^0 \leq \bar{P}$.

Let γ_m^i be the i-th SINR measurement at user m, which is carried in the feedback to BS m. BS m then uses the following DCPC algorithm to update its power after receiving the i-th SINR feedback.

$$P_m^i = \min\left\{\bar{P}, \frac{\gamma_m^{tar}}{\gamma_m^i} P_m^{i-1}\right\},\ i = 1, 2, \ldots. \tag{38}$$

If the γ_m^{tar}'s are feasible (see Section II-C), the power vector series $\{\vec{P}^0, \vec{P}^1, \ldots, \vec{P}^i, \ldots\}$ is proved to converge to a unique positive power vector satisfying the following equation [19].

$$\vec{P} = \min\left\{\vec{\bar{P}}, \mathbf{\Gamma}^{tar}(\mathbf{A}\vec{P} + \vec{\nu})\right\}, \tag{39}$$

where $\mathbf{\Gamma}^{tar} = \mathrm{diag}\{\vec{\gamma}^{tar}\} = \mathrm{diag}\{\gamma_1^{tar}, \gamma_2^{tar}, \ldots, \gamma_M^{tar}\}$. Furthermore, the converged power vector $\vec{P}^*(t)$ also achieves the target SINR $\gamma_m^{tar}(t)$ for each BS m. The convergence result is summarized in the following fact from [19].

Fact 2: With the DCPC algorithm (38), the transmit power vector converges to a unique positive power vector $\vec{P}^*$ satisfying (39). After convergence, either $\vec{P}^*$ achieves $\vec{\gamma}^{tar}$ or at least one of the elements in $\vec{P}^*$ is equal to $\bar{P}$.

The pseudo code for the distributed DCPC algorithm is given in Algorithm 2, where α is a fraction in (0,1) and β is a positive integer.If BS m's transmit power remains at the maximum power $\bar{P}$ for β iterations, while the target SINR γ_m^{tar} is still not achieved, we reset the target SINR as $\gamma_m^{tar} = \gamma_m^{min} + \alpha \cdot (\gamma_m^{tar} - \gamma_m^{min})$ and restart the iterative update process. We choose $\alpha = 0.618$, the reciprocal of the *golden ratio*, and β from 2 to 5 in our simulations.

Algorithm 2: DCPC Algorithm 2

1 BS m obtains b_m, D_m, and B_m for user m;

2 BS m computes SINR bounds γ_m^{max} and γ_m^{min};

3 BS m sets $\gamma_m^{tar} = \gamma_m^{max}$ and $P_m(0) \in (0, \bar{P}]$;

4 **While** *TRUE* **do**

5 BS m receives SINR feedback γ_m^i and updates its power as: $P_m^i = \min\left\{\bar{P}, \left(\gamma_m^{tar}/\gamma_m^i\right) P_m^{i-1}\right\}$;

6 **If** ($P_m^i = \bar{P}$ *for* β *iterations*) & ($\gamma_m^i \neq \gamma_m^{tar}$) **then**

7 reset the target SINR as: $\gamma_m^{tar} = \gamma_m^{min} + \alpha \cdot (\gamma_m^{tar} - \gamma_m^{min}) i = i + 1$;

8 **end**

9 $i = i + 1$;

10 **end**

In practice, the path gains vary over time due to channel fading. It is possible that during some time slot, the transmission is not feasible even for the minimum required rate. It is nontrivial to test the feasibility of the target SINR vector $\vec{\gamma}^{tar}$ in a distributed manner with only local information. In fact, if the target SINR vector is infeasible, the problem of finding the largest set of links that can be supported at the given SINRs is proved to be NP-Complete [29]. Therefore, we adopt the following heuristic strategies to handle the case when the target SINR vector cannot be achieved by a feasible power assignment due to deep fading channels.

i) In the first time slot, if the DCPC algorithm does not converge in a certain number of steps, suspend the transmission of the video with the largest frame size for sometime and retry the algorithm.
ii) Adopt the acceleration enhancement as in the centralized algorithm, which is described in Section III-C .
iii) If the DCPC algorithm does not converge for the reduced γ_m^{tar} (see Lines 6–8 in Algorithm 2), further reduce the target SINR as $\gamma_m^{tar} = \gamma_m^{min} + \alpha \cdot (\gamma_m^{tar} - \gamma_m^{min})$. If still no convergence when $\gamma_m^{tar} = (1+\epsilon) \cdot \gamma_m^{min}$, for $0 < \epsilon \ll 1$, all the links whose buffer will not be empty in the next time slot will pause their transmissions. Since the algorithm always tries to transmit as more data as possible (i.e., by setting a high target SINR whenever possible), it is highly likely that such links won't have buffer underflow in the following time slots.
iv) If all the above steps fail, the BS suspends its transmission and the user freezes the playout precess until the next time slot.

B. Discussions

Channel Estimation: In a real deployment, some effective channel estimation schemes should be adopted to obtain the channel gains. Based on the channel gains, the schedules can be computed and the transmit powers determined.

Scalability Issue: The algorithm is focused on intercell interference in a multi-cell wireless network, while assuming orthogonal channels for users in the same cell. The major interference comes from users in neighboring cells using the same channel. Since in each cell each video session is allocated with one channel, given the hexagon design of cellular networks, there may be at most seven users (six in the six neighboring cells, and one in the center cell) that share the same channel. So, for seven video sessions, the computation complexity should be small. Scalability may become an issue when there are a large number of channels, and the controller needs to solve one problem for each channel. We conjecture that the distributed/heuristic algorithm can handle a considerable large number of concurrent problems due to its low computational complexity. An admission control mechanism may also be necessary to limit the number of concurrent video sessions.

Long Videos: The proposed algorithms do not require a specific video length. If the video length is much longer, we can cut the video to several shorter segments and the proposed scheduling algorithms can be executed for each segment of the video stream.

VCR Control: For VCR control functions, a similar strategy can be adopted that was used in our prior work [30]. For example, if the user would like to fast forward or rewind the video, the commands are then sent to the controller. The controller will locate the rewound or fast forwarded location and shift the underflow and overflow curves to that point. Then, based on the new frame size information and the shifted underflow and overflow curves, a new transmission schedule will be computed. See [30] for details.

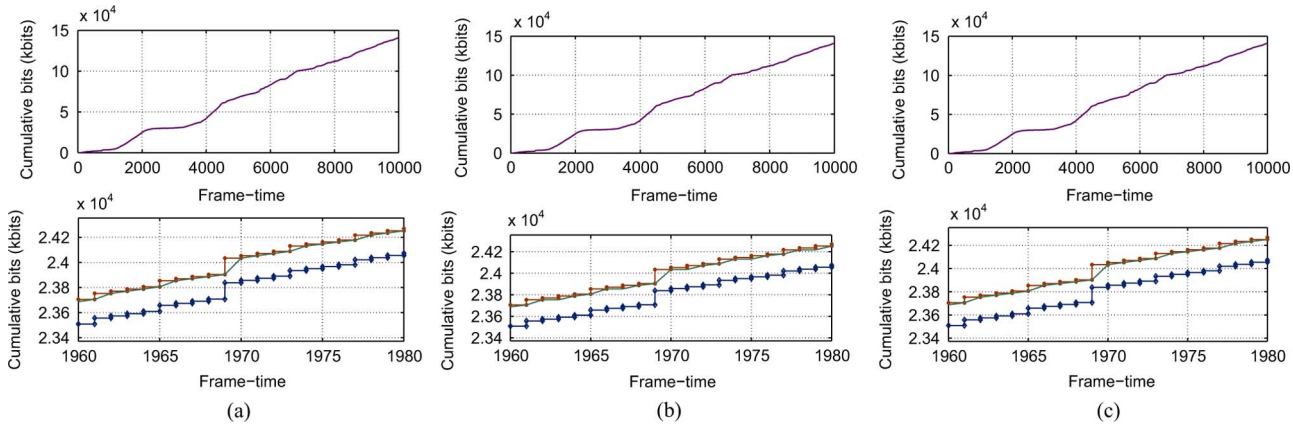


Fig. 3. The cumulative overflow, transmission, and consumption curves when transmitting Star Wars at link 1 in the seven-cell network. The three curves in each lower figure are the cumulative overflow, transmission, and consumption curves from top to bottom. (a) Centralized algorithm. (b) Accelerated centralized algorithm. (c) DCPC.

V. Simulation Results

To evaluate the performance of the proposed algorithms, we simulate VBR video streaming in a 7-cell wireless network. We assume the channels within a cell are orthogonal and inter-cell interference is the major limiting factor. The channel bandwidth is $B_w = 1$ MHz. The path gain averages are set to $\bar{G}_k^m = d_{km}^{-4}$, where d_{km} is the physical distance from BS k to user m. We assume Rayleigh fading channels in all the simulations, where the normalized path gain is exponentially distributed as $f(G_k^m) = \exp\{-G_k^m/\bar{G}_k^m\}$ for $G_k^m \geq 0$. The distance from a user to its corresponding BS is uniformly distributed from 100 m to 1000 m and the inter-cell BS distance is from 1600 m to 2000 m. The temperature is $T_0 = 290$ Kelvin and the equivalent noise bandwidth is also 1 MHz. The peak power constraint is $\bar{P} = 1$ Watt.

In each cell, the channel is dedicated to one mobile user for VBR video streaming. We assume BS's 1, 4 and 7 are streaming movie *Star Wars*, BS's 2 and 5 are streaming *NBC News*, and the remaining links 3 and 6 are transmitting *Tokyo Olympics*. We use the VBR video traces from the Video Trace Library hosted at Arizona State University [24] in all the simulations. The playout buffer size is set to be 1.5 times of the largest frame size in the requested VBR video.

A. Centralized Algorithm

We implement the branch-and-bound centralized algorithm using MATLAB. We choose $\epsilon = 10\%$ for the simulations. From the VBR video traces, we derive the cumulative consumption and overflow curves. The centralized algorithm computes the optimized power assignment for the BS's at beginning of each time slot. In Fig. 3(a), we plot the cumulative consumption, overflow and transmission curves for *Star Wars* transmitted on link 1. The top subfigure is for 10000 frames. We also plot the curves from frame 1960 to frame 1980 in the bottom subfigure, while frame 1969 has the largest size among the 10000 frames. We observe that the cumulative transmission curve $X_1(t)$ is very close to the cumulative overflow curve $B_1(t)$, indicating that the centralized algorithm always aims to maximize the transmission rate as allowed by the buffer and power constraints, and the playout buffer is fully utilized for most of the time. There is no playout buffer overflow or underflow for the entire range of the movies.

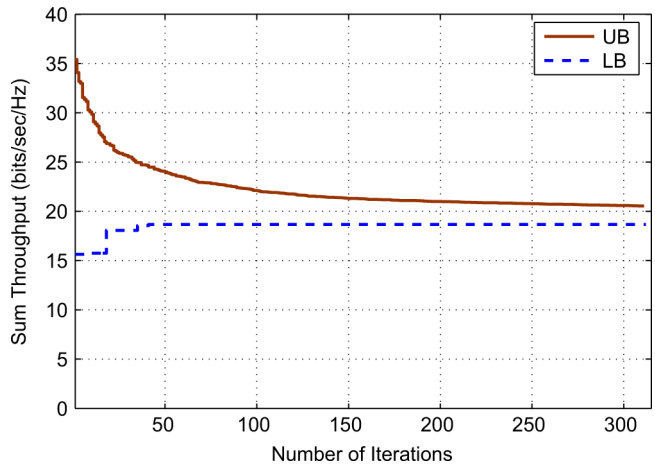


Fig. 4. Convergence of the branch-and-bound algorithm in time slot 1 when all the I-frames are transmitted and all the buffers are empty (i.e., the worst case scenario).

In Fig. 4, we plot the upper and lower bounds for objective function (7) for time slot 1. This is the hardest time slot with respect to power control, since all the sessions are transmitting I-frames and all the playout buffers are empty in this time slot in our simulations. We observe the optimality gap between UB and LB is continuously decreased until the $\epsilon = 0.1$ threshold is reached. In other time slots where the frame sizes are not consistently large and the playout buffers are close to full, it usually takes only a few (e.g., 5 or 6) iterations to reach the optimality gap threshold.

The average computation time are recorded for the default scenario, for a PC with Intel 1.83 GHz CPU and 3.00 G RAM. We find most of the computation time is consumed in the first time slot, because the first frame is the I-frame that usually has the largest size in the sequence. The computation time varied with the channel fading condition and decreases as the optimality tolerance ϵ increases, in the range of 1.5 s to 480 s. The smaller the interference, the smaller the computation time. The performance also significantly depends on the linear solver of

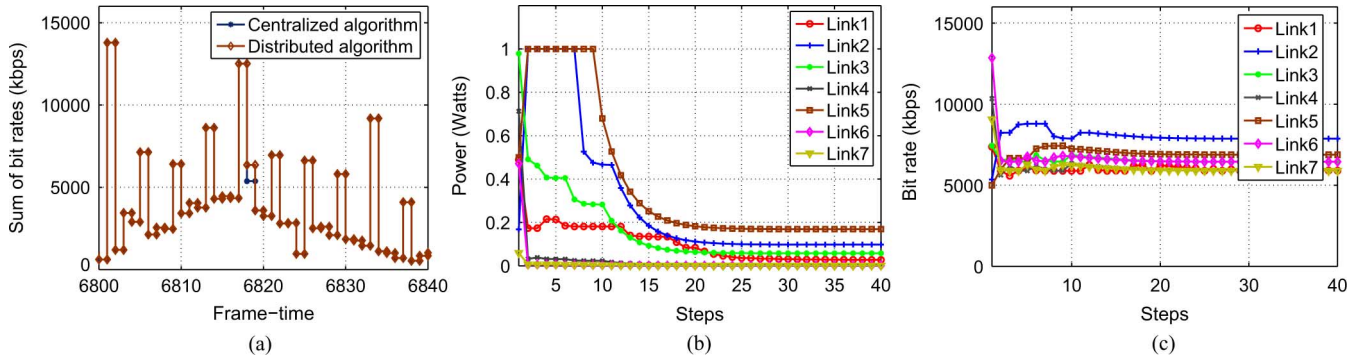


Fig. 5. Simulation results with a seven-link network. (a) Rate sums with the algorithms. (b) Convergence of transmit powers with DCPC. (c) Convergence of bit rates with DCPC.

Matlab. There is a tradeoff between computation time and optimality. The larger the ϵ, the smaller the computation time. Since the branch and bound framework is computation intensive, it is mainly used as a benchmark for performance evaluation.

We also evaluate the accelerated scheme under the same video and network conditions. The curves for link 1 are plotted in Fig. 3(b). It can be seen that during time slots 1963, 1967, and 1971, there is no transmission on link 1 since the playout buffer is over 80% full. Pausing transmission in these time slots makes it easier for other links to transmit large frames and speeds up the convergence of the algorithm, while causing no buffer underflow at link 1. Since large frames rarely occur in the same time slot (except for time slot 1), this is analogous to statistical multiplexing of VBR videos. We find in the simulation, a link can pause in over 60% of the time slots with the acceleration heuristic, resulting in significant reduction in computation time.

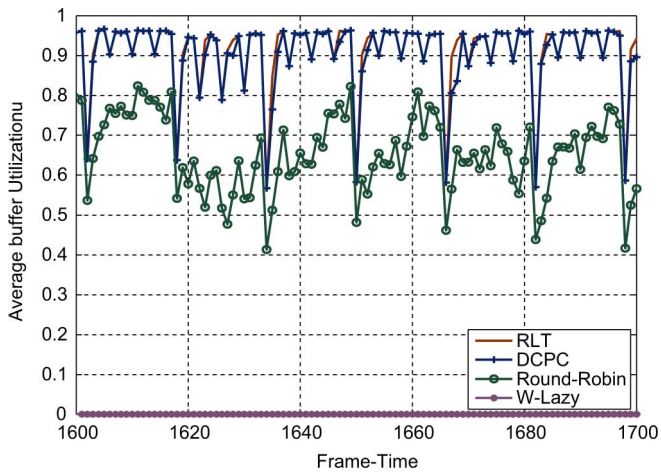


Fig. 6. Average buffer utilization of the four schemes.

B. Distributed Algorithm

We next examine the performance of DCPC. The network and video setups are the same as those in the centralized algorithm simulations. The cumulative overflow, transmission, and consumption curves obtained by DCPC are plotted in Fig. 3(c) for *Star Wars* transmitted on link 1. We observe very similar performance as in the case of the centralized algorithm shown in Fig. 3(a). The cumulative transmission curve is again very close to $B_m(t)$, and there is neither buffer overflow nor underflow during the transmission of 10000 frames.

To compare the distributed and centralized algorithms, we compute the sum of the bit rates of all the links in each time slot. The acceleration scheme is not used for both algorithms in this simulation. The rate sums are plotted in Fig. 5(a) from time slot 6800 to 6840. We observe that the sum rates achieved by the centralized algorithm and that by the distributed algorithm are identical for most part of this interval. Examining the rate sums for the entire 10000 time slots, we find that the rate sum achieved by the DCPC algorithm is within 99% of the corresponding rate sum achieved by the centralized algorithm in over 97% of the time slots.

The convergence of the distributed DCPC algorithm is plotted in Figs. 5(b) and 5(c) for one of the time slots. The accelerated scheme is incorporated with DCPC, such that a link m may pause its transmission if its buffer is over 80% full and $X_m(t-1) > D_m(t)$. The evolution of the BS transmit powers are plotted in Fig. 5, where after 23 steps, all the transmit powers converges to a value between 0 and $\bar{P} = 1$ Watt. The converged power vector is $\vec{P}^* = [0.0023, 0.208, 0.185, 0.0013, 0.163, 7.1 \times 10^{-04}, 0.188]$ Watt. The evolution of the bit rates are plotted in Fig. 5(c). It is interesting to see the data rates converge faster than the transmit powers in this case. All the data rates reach stable values after a few steps.

C. Empirical Performance Evaluation

We evaluate the performance of the proposed schemes by comparing them with the following two schemes.

- A round-robin scheme where the BS allocates power in a QoS based round-robin fashion, which favors the session that would suffer buffer starvation if no transmission is scheduled. When a specific BS is selected for transmission, it transmits the video with maximum power without overflowing the client buffer, and all its neighbors remain silent in the same frame-time slot.
- W-Lazy, as described in Section II-D.

First, we investigate the average buffer utilization at end of each time slot. When underflow happens, the missing frame is discarded, and the next frame will be scheduled for the transmission in the next time slot. We observe that the proposed RLT and DCPC schemes achieve higher average buffer utilization than the other two schemes. Fig. 6 shows the average buffer utilization from frame-time slot 1600 to 1700. We find that the buffer utilization of RLT and DCPC fluctuate around 90% mostly, while

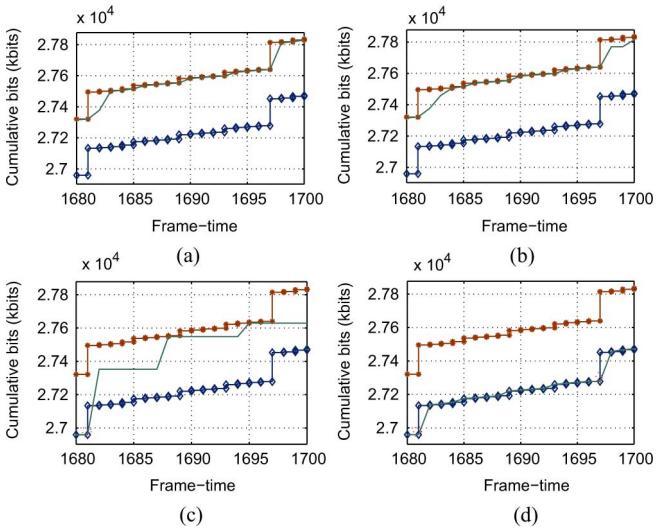


Fig. 7. Illustration of underflow events. As in Fig. 3, the three curves in each figure are the cumulative overflow, transmission, and consumption curves from top to bottom.

TABLE I
NUMBER OF UNDERFLOW EVENTS AND AVERAGE POWER CONSUMPTION

	RLT	DCPC	Round-robin	W-Lazy
Mean	0	0.2	516	1076
Conf. Int.	$[0, 0]$	$[-0.355, 0.755]$	$[442, 591]$	$[514, 1637]$
Avg. Power Cons. (W)	0.012	0.0078	0.0161	0.0039

the utilization of the Round-robin scheme is in the range of 50% to 80%. We also find that the W-Lazy scheme always achieves a zero buffer utilization, since it only transmits each frame as late as possible in each time slot. At the end of a time slot, all the data will be consumed by the user and the buffer is left empty.

We then compare the average number of underflow events in Table I. we find RLT achieves underflow free transmission, while the number of underflow events for DCPC is negligible in the simulations. This is because both schemes aim to transmit as much video data as possible under the feasible condition in each frame-time slot. The extra video data transmitted will be in the playout buffer to provide a cushion to future large frames or network dynamics. On the other hand, both Round-robin and W-Lazy suffers a large number of underflow events. We also illustrate the buffer underflow events in the period from 1680 to 1700 in Fig. 7. The red dot circles indicate the buffer underflow. It can be seen that the cumulative transmission curve lies below the cumulative consumption curve when buffer underflow events occur. This results in an infeasible transmission schedule, which causes frozen playout.

The average power consumption of the schemes are also shown in Table I. W-Lazy consumes the least power. Due to the variation of frame size and network condition, the transmission of W-Lazy are infeasible in many time slots. To prevent the divergence of power allocation, some video sessions should be paused and the power savings of W-Lazy are achieved by pausing video transmissions. However, this is at the cost of significantly more buffer underflow events, which are undesirable for user experience. The Round-robin scheme tries to transmit as much video data as possible. However, it chooses a session greedily and pauses other unselected video sessions. This also causes many underflow events for the unselected sessions. Also, due to the round robin fashion and limited buffer size, when the unselected session become selected, its low buffer utilization will lead to a larger power consumption in order to fill the buffer, especially when it misses the previous good channel condition and the channel condition is worse at the current time slot.

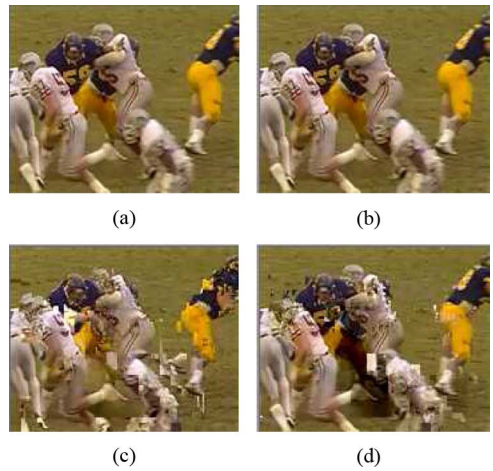


Fig. 8. Perceived quality of decoded videos: *Football* video frame 14. (a) RLT. (b) DCPC. (c) Round-robin. (d) W-Lazy.

TABLE II
STATISTICS OF THE QUALITY OF THE DECODED VIDEOS: *FOOTBALL*

	No. underflow events (No. Lost frames)	Average SNR over all frames (Y dB)
RLT	0	34.39
DCPC	0	34.39
Round-robin	$3(14, 21, 28)$	30.56
W-Lazy	$2(11, 81)$	34.14

The visual quality of the decoded *QCIF* VBR *Football* video under different schemes are presented in Fig. 8. The 128-frame video is encoded in VBR with fixed quantization parameters. We present the perceived quality of the decoded video frames by RLT, DCPC, Round-robin and W-Lazy in the figure. The numbers of underflow events and average SNRs of the decoded videos are presented in Table II.[3] A larger buffer size can achieve a better performance, but it will cause extra cost of the node design in practice. We find that the proposed schemes are free of underflow events and produce the best perceived quality than both the W-Lazy and Round-robin schemes. Lost frames also affect frame decoding with error propagations. For example, frame 11 is lost during W-Lazy transmission, the decoded frame 14 also suffers from the loss.

VI. CONCLUSION

We studied downlink power control for multi-user VBR video streaming in multicell networks. The problem formulation considers downlink power control, inter-cell interference, VBR video characteristics, and playout buffer requirements.

[3]The Round-robin scheme uses a larger buffer than the other schemes in this simulation, otherwise the decode will crash if the same small buffer size as in the other schemes is used for Round-robin.

We developed a centralized algorithm that can provide (1-ϵ)-optimal solutions, and a fast distributed algorithm that only needs local information. The algorithms are evaluated with extensive simulations with VBR video traces and fading channels, and are demonstrated to be effective for streaming VBR videos over multicell wireless networks.

REFERENCES

[1] Y. Huang and S. Mao, "Downlink power control for variable bit rate video over multicell wireless networks," in *Proc. IEEE INFOCOM'11*, Shanghai, China, Apr. 2011, pp. 2561–2569.

[2] D. Hu, S. Mao, Y. T. Hou, and J. H. Reed, "Fine grained scalability video multicast in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, Apr. 2010.

[3] D. Hu and S. Mao, "On medium grain scalable video streaming over cognitive radio femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 641–651, Apr. 2012.

[4] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 24, no. 4, pp. 269–280, 1994.

[5] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, no. 2/3/4, pp. 1566–1579, Feb./Mar./Apr. 1995.

[6] D. P. Heyman and T. V. Lakshmanr, "What are the implications of long-range dependence for VBR-video traffic engineering?," *IEEE/ACM Trans. Netw.*, vol. 4, no. 3, pp. 301–317, Jun. 1996.

[7] M. Dai, Y. Zhang, and D. Loguinov, "A unified traffic model for MPEG-4 and H.264 video traces," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1010–1023, Aug. 2009.

[8] S. Kang and S. L. Y. Won, "On-line prediction of nonstationary variable-bit-rate video traffic," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1219–1237, Mar. 2010.

[9] S. Tanwir and H. Perros, "A survey of VBR video traffic models," *IEEE Commun. Surveys & Tutorials*, pp. 1–25.

[10] S. Liew and H. Chan, "Lossless aggregation: a scheme for transmitting multiple stored VBR video streams over a shared communications channel without loss of image quality," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1181–1189, Aug. 1997.

[11] J. Salehi, Z.-L. Zhang, J. Kurose, and D. Towsley, "Supporting stored video: reducing rate variability and end-to-end resource requirements through optimal smoothing," *IEEE/ACM Trans. Netw.*, vol. 6, no. 4, pp. 397–410, Aug. 1998.

[12] S. Sen, D. Towsley, Z. Zhang, and J. K. Dey, "Optimal multicast smoothing of streaming video over the internet," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1345–1359, Sep. 2002.

[13] G. Liang and B. Liang, "Balancing interruption frequency and buffering penalties in VBR video streaming," in *Proc. IEEE INFOCOM'07*, Anchorage, AK, USA, May 2007, pp. 1406–1414.

[14] E. A. Gharavol, M. Khademi, and M. R. T Akbarzadeh, "A new variable bit rate (VBR) video traffic model based on fuzzy systems implemented using generalized regression neural networks (GRNN)," in *Proc. IEEE Fuzzy Systems'06*, Vancouver, BC, Canada, Jul. 2006, pp. 2142–2148.

[15] J. M. Mcmanus and K. W. Ross, "A dynamic programming methodology for managing prerecorded VBR sources in packet-switched networks," *Proc. SPIE, Performance and Control of Network Syst.*, pp. 140–154, 1997.

[16] T. Stockhammer, H. Jenkac, and G. Kuhn, "Streaming video over variable-bit-rate wireless channels," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 268–277, Apr. 2004.

[17] S. Chatziperis, P. Koutsakis, and M. Paterakis, "A new call admission control mechanism for multimedia traffic over next-generation wireless cellular networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 1, pp. 95–112, Jan. 2008.

[18] F. D. Rango, M. Tropea, P. Fazio, and S. Marano, "Call admission control for aggregate MPEG-2 traffic over multimedia geo-satellite networks," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 612–622, Sep. 2008.

[19] S. Grandhi, J. Zander, and R. Yates, "Constrained power control," *Int. J. Wireless Personal Commun.*, vol. 1, no. 4, pp. 257–270, Apr. 1995.

[20] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 104–116, Jan. 2005.

[21] A. Gjendemsj, D. Gesbert, G. Oien, and S. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3164–3173, Aug. 2008.

[22] M. Rasti and A. R. Sharafat, "Distributed uplink power control with soft removal for wireless networks," *IEEE Trans. Commun.*, vol. 59, no. 3, pp. 833–843, Mar. 2011.

[23] I. Mitliagkas, N. D. Sidiropoulos, and A. Swami, "Joint power and admission control for ad-hoc and cognitive underlay networks: Convex approximation and distributed implementation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4110–4121, Dec. 2011.

[24] M. Reisslein, Video Trace Library, Arizona State Univ. [Online]. Available: http://trace.eas.asu.edu/.

[25] M. Chen and A. Zakhor, "Multiple TFRC connections based rate control for wireless networks," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 1045–1062, Oct. 2006.

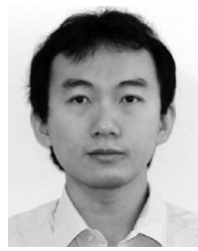
[26] J. Lee, R. Mazumdar, and N. Shroff, "Downlink power allocation for multi-class wireless systems," *IEEE/ACM Trans. Netw.*, vol. 13, no. 4, pp. 854–867, Aug. 2005.

[27] N. Bambos, S. C. Chen, and G. J. Pottie, "Radio link admission algorithm for wireless networks with power control and active link quality protection," in *Proc. IEEE INFOCOM'95*, Boston, MA, USA, Apr. 1995, pp. 97–104.

[28] S. Kompella, S. Mao, Y. Hou, and H. Sherali, "On path selection and rate allocation for video in wireless mesh networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 1, pp. 212–224, Feb. 2009.

[29] M. Andersin, Z. Rosberg, and J. Zander, "Gradual removals in cellular PCS with constrained power control and noise," in *Proc. IEEE PIMRC'95*, Toronto, ON, Canada, Sep. 1995, pp. 56–60.

[30] Y. Huang, S. Mao, and Y. Li, "A majorization approach to downlink multiuser VBR video streaming," *Elsevier Comput. Commun.*, vol. 35, no. 15, pp. 1828–1837, Sep. 2012.



Yingsong Huang (S'12) received the M.S. degree in control theory and control engineering and the B.S. degree in Automation, both from Chongqing University, Chongqing, China. Since 2007, he has been pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL. His research interests include modeling, control and optimization in microgrids, smart grid and computer networks.

Shiwen Mao (S'99–M'04–SM'09) received Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA (now Polytechnic Institute of New York University) in 2004. He was a research staff member with IBM China Research Lab from 1997 to 1998. He was a Postdoctoral Research Fellow/Research Scientist at Virginia Tech, Blacksburg, VA, USA from 2003 to 2006. Currently, he is the McWane Associate Professor in the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA.

His research interests include cross-layer optimization of wireless networks and multimedia communications, with current focus on cognitive radios, femtocells, 60 GHz mmWave networks, free space optical networks, and smart grid. He is on the Editorial Board of IEEE Transactions on Wireless Communications, IEEE Internet of Things Journal, IEEE Communications Surveys and Tutorials, Elsevier Ad Hoc Networks Journal, and Wiley International Journal of Communication Systems. He is the Director of E-Letter of the Multimedia Communications Technical Committee (MMTC), IEEE Communications Society for 2012–2014.

Dr. Mao is a coauthor of *TCP/IP Essentials: A Lab-Based Approach* (Cambridge University Press, 2004). He was awarded McWane Endowed Professorship in the Samuel Ginn College of Engineering for the Department of Electrical and Computer Engineering, Auburn University in August 2012. He received the NSF Faculty Early Career Development Award (CAREER) in 2010. He is a co-recipient of IEEE ICC 2013 Best Paper Award, The 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems and The Best Paper Runner-up Award at ICST QShine 2008. He was named 2012 Exemplary Editor of IEEE Communications Surveys & Tutorials. He also received Auburn Alumni Council Research Awards for Excellence-Junior Award in 2011 and two Auburn Author Awards in 2011. Dr. Mao holds one US patent.