Dynamic Resource Allocation and Layer Selection for Scalable Video Streaming in Femtocell Networks: A Twin-Time-Scale Approach

Jian Yang, *Senior Member, IEEE,* Peng Si, Zilei Wang, *Member, IEEE,*
Xiaofeng Jiang, *Member, IEEE,* Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Scalable video streaming over femtocell networks relying on two-tier spectrum-sharing is designed for coping with time-varying channel conditions, stringent video QoS requirements as well as with strong cross-tier interference between the over-sailing macro- and the femtocells. Dynamic video layer selection and resource allocation are invoked to enable the adaptation of the scalable video streaming service to the dynamics of both channel quality and interference price fluctuations. We formulate the design as a constrained stochastic optimization problem, which strikes a compelling compromise between the perceivable quality of experience and the monetary implications of the interference. Since the time scale of resource allocation is more short-term than that of the video layer selection, we decompose the original long-term utility optimization problem into a pair of readily tractable subproblems with the aid of two different time-scales by invoking the powerful technique of Lyapunov drift and optimization. By exploiting the specific structure of these subproblems, low-complexity algorithms are derived for dynamic video layer selection and resource allocation, which rely on the near-instantaneously available information rather than on any prior statistical knowledge. Finally, we derive the analytical bounds of the theoretically achievable performance. Experimental results are presented for characterizing the performance attained.**

I. Introduction

With the ever-increasing popularity of intelligent wireless devices, accessing flawless video content via wireless network continues to increase the thirst for increased download rates. As reported by Cisco [1], the wireless video tele-traffic has increased at an exponential rate over the recent years, and it is predicted to grow even faster in the near future. Due to the increasingly over-crowded spectrum, this tele-traffic explosion induced by video streaming imposes substantial challenge on the conventional cellular networks [2]. Against this backdrop, the provision of femtocellular coverage constitutes an attractive solution for improving the attainable network capacity and coverage quality. Owing to their compact size, inexpensive construction and low power femtocells have indeed become popular. Customers can utilize their own broadband Internet access as the backhaul link connected to the operator's core network [3]. As a benefit of their relatively compact coverage area, femtocells often provide line-of-sight coverage, hence drastically improving the signal-to-interference-plus noise ratio (SINR) and the area spectral efficiency.

Despite these benefits, femtocells also impose additional technical challenges in terms of inflicting cross-tier interference, which potentially degrades the SINR. These issues are further aggravated by the unplanned random manner of sharing the spectrum [3]. Therefore, efficient cross-tier resource allocation is required in these femtocell networks having a two-tier structure [4]–[6].

In order to address this problem, femtocells have been richly characterized in the literature [7]–[11], but predominantly from the operator's - rather than from the customer's - perspective. However, most femtocells tend to be randomly deployed by individual customers, rather than by the operator [3]. This operating pattern of individual deployment imposes the challenge of acquiring the global network conditions for optimizing the resource allocation from the customer's perspective. Furthermore, the above-mentioned operator-centric approaches do not necessarily represent the interests of the femtocell customers. It should be also noted that the above-mentioned treatises put emphasis on optimizing the transmission of general data, where the end user's satisfaction is characterized by a potentially over-simplified function, for instance by a linear function concerning the achievable bitrate [9] or by the SINR attained [12]. Hence, they do not consider the perceptual quality of the video traffic, and therefore they are inappropriate for supporting the stringent QoS requirements of video streaming services.

Hence, we aim for conceiving video streaming over the above-mentioned two-tier spectrum-sharing femtocell networks. We consider a scenario, where a femto base station (FBS) is installed indoors by a customer and video clips are streamed to the mobile users within the licensed spectrum band of a macrocellular network operator (MNO). Inspired by the spectrum underlay approach of Cognitive Radio Networks (CRN), a femtocell is allowed to transmit its data over the macrocellular band, provided that the total interference is not higher than a tolerable threshold [13]. The price-based interference control mechanism of [9] is a promising technique of controlling the behaviour of the femtocell, which allows the

This work was supported by National Natural Science Foundation of China (No. 61573329), State Key Program of National NSF of China (61233003), and Youth Innovation Promotion Association CAS. L. Hanzo would also like to acknowledge the financial support of the European Research Council's Advanced Fellow Grant Beam-Me-Up.

J. Yang, P. Si, Z. Wang and X. Jiang are with School of Information Science and Technology, University of Science and Technology of China, Jinzhai Road 96, Hefei, Anhui 230027, China. E-mail:jianyang@ustc.edu.cn.

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK. E-mail:lh@ecs.soton.ac.uk.

femtocell to optimize both its transmission power allocation and its channel assignment, subject to an acceptable cost. The basic idea of this interference control is that in order to protect the transmissions of the licensed macrocell user equipment (MUE), the macrocell imposes a high price on the specific femtocell that inflicts co-channel interference upon an MUE, and vice versa. Here, we assume that the two-tier spectrum-sharing femtocell network considered employs this price-based interference control regime.

Due to the high dynamics of the network conditions and channel availabilities, having the ability to promptly adjust the video bitrate is essential for supporting smooth transmissions over wireless/wired networks [14]. Scalable Video Coding (SVC) [15], [16] constitutes a promising technique of supporting smooth video bitrate adjustment. Briefly, SVC encodes the video clips into a base layer providing basic image quality and several enhancement video layers, which may be dropped in case of network-congestion. Receiving more video layers implies having a higher perceivable video quality. The video bitrate adaptation can be achieved by dynamically selecting the number of video layers for transmission. This scalability is particularly beneficial for streaming video over femtocell networks, which enables the system to accommodate both single-link-induced and network-induced channel quality fluctuations [17].

However, the provision of video streaming services over femtocell networks has to address the following key challenges. Firstly, strike a compelling tradeoff between the monetary cost and the perceivable video quality. Higher bitrates offer higher video quality and smooth playback, thus enhancing the quality of experience (QoE) of the users at the cost of a higher payment not only due to its high channel occupancy but also owing to the additional interference. The second challenge is the time-varying dynamics of the femtocell network. The quality of wireless channel has a stochastically fluctuating nature, and the price of interference is time-varying, which is difficult to predict accurately. The third challenge is the time-scale mismatch between the resource allocation and the video layer adjustment. In [18], we have briefly explored the problem of power control and video layer selection in femtocell networks, whilst disregarding both the associated channel allocation and the different time scales at which the wireless resource allocation and the application-level video layer switching operate. Motivated by these challenges, we aim for designing a comprehensive solution for streaming video over the two-tier network considered by relying both on a/ joint dynamic video layer selection, b/ channel allocation and c/ power assignment (JLCP) mechanism. Against the above background, this treatise offers the following three contributions.

- We formulate the problem of scalable video transmission from femtocells as a constrained stochastic optimization problem under a given pricing strategy by the primary network, which maximizes the weighted utility defined as a combination of the perceivable video quality and monetary cost subject to our smooth playback constraint. The monetary cost is commensurate with the co-channel inference imposed on the MUEs.
- By adapting Lyapunov's stochastic optimization technique [19] to the proposed model, the original long-term average utility problem is decomposed into a pair of near-instantaneous optimization subproblems. This assists us in exploiting the specific structure of the subproblems to derive a low-complexity resource allocation and video layer activation strategy that operates at two different time-scales. Furthermore, we derive the analytical bounds of both the time-averaged queue lengths and of the achievable utility of the proposed solution.
- Finally, we conduct experiments based on real video traces to investigate the achievable performance of the proposed method. Our experimental results illustrate that the proposed solution is capable of promptly responding to the variations of the environment. Finally, it is verified that the proposed feasible solution is indeed capable of approaching the optimal solution by controlling a carefully designed tradeoff parameter.

The rest of the paper is structured as follows. We discuss the related work in Section II. Our mathematical model of streaming scalable video over two-tier femtocell networks is formulated by integrating a femtocellular network model, a monetary cost model and a video streaming model in Section III. Section IV is dedicated to deriving the strategy of joint dynamic video layer selection, channel allocation and power assignment, while Section V characterizes the attainable performance. Finally, we offer our conclusions in Section VI.

II. Related Work

The technical challenges of femtocell networks and their potential countermeasures are detailed for example in [3], [20]. Bearing in mind that the additional "tier" of femtocells overlaps with the conventional macrocell network, numerous solutions have been conceived for mitigating the cross-tier interference imposed. Specifically, a spectrum allocation strategy is proposed for maximizing the throughput of femtocell networks having a two-tier structure in [8]. Aiming for mitigating the co-channel interference inflicted upon the macrocells by the femtocells, the authors of [12] propose a utility-driven method for adapting to the SINR fluctuations experienced in the femtocells. In [21], Kim *et al.* investigate the per-tier outage probability in two-tier femtocell networks by introducing an approximate femtocell interference distribution model. In [22], Elsherif *et al.* utilize an adaptive graph-coloring method for developing an innovative resource allocation strategy, in order to efficiently manage the interference among femtocells and to maintain fairness across the different users. Hsieh *et al.* [23] investigate the problem of maximizing the number of customers supported at specific QoS constraints in a dense network associated with an arbitrary topology. However, these centralized algorithms implicitly assume that the macrocells and femtocells are deployed and managed by the same service provider.

Some prior contributions have also considered the problem of two-tier networks, where FBSs are operated by different service providers. In [9], price-based resource allocation has been conceived for femtocell networks to optimally share

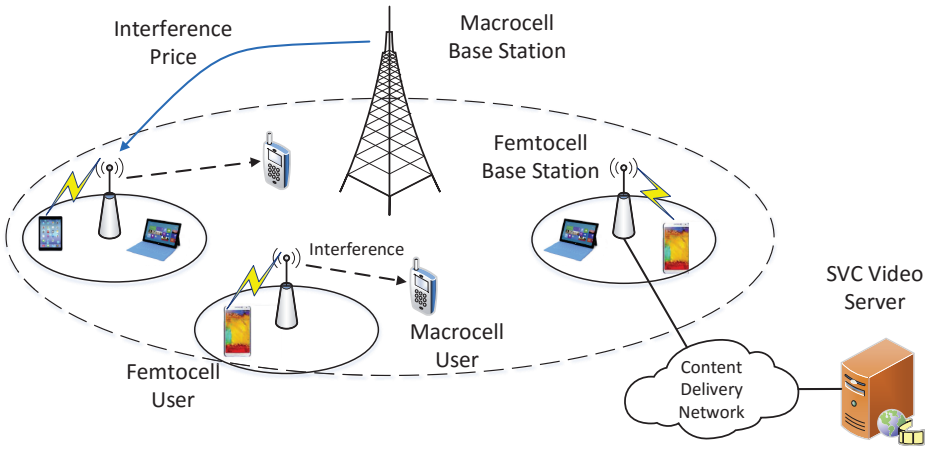


Fig. 1. System architecture for scalable video streaming in a two-tier HetNet.

the spectrum by using game theory. The macrocell protects itself by imposing a price commensurate with the interference arriving from the users attached to femtocells. A hierarchical dynamic game is presented in [24] to model the competition between the macrocell service provider and the small cell providers. In [11], Wang *et al.* consider the problem of maximizing the profit in femtocell networks from the perspective of the network operator. In [10], Jiang *et al.* study the economic issues of cognitive femtocell networks. By relying on a sophisticated game theoretic solution, a twin-tier pricing-based framework is presented for analyzing the Nash-equilibrium price of both gaming parties, *i.e.*, of the network operators and the femtocell users, respectively, corresponding to the macrocells and femtocells. Wang *et al.* [25] consider the downlink distributed power control problem of femtocell networks at a certain QoS provision guaranteed for the macrocell user equipment, where a pair of non-cooperative game formulations are presented to analyze the Nash equilibrium and to develop the corresponding algorithms. ElSawy *et al.* [26] model and analyze the performance of cognitive two-tier networks in a multichannel environment by using stochastic geometry. However, again most of these contributions are investigated from the perspective of the macrocell operators, who do not give much cognizance to the interest of the femtocell.

Several recent treatises have nonetheless investigated two-tier networks from the perspective of femtocells. For example, Ma *et al.* [13] combine the overlay spectrum and underlay spectrum for designing dynamic access to the femtocell networks for the sake of improving the network capacity. Zhang *et al.* [27] propose hybrid access to cognitive femtocells by dynamically allocating the spectral resources. In [28], Wang *et al.* study the average femtocell throughput maximization problem under the constraint of the macrocell's queue stability. However, these contributions mainly focus on general data transmission, which fail to adequately characterize the unique nature of video traffic. Thus, they are not appropriate for supporting the stringent perceptual QoS requirements of the video streaming services. Considering the scalable video streaming problem of femtocell networks, Hu *et al.* [17] formulate it as a stochastic programming problem associated with network dynamics and uncertainties in the cognitive radio networks. However, they mainly focus their attention on scheduling the access to both the Macrocell Base Station(MBS) and to the FBS. In [29] and [30], the authors jointly consider the wireless transmission scheduling and video quality adaptation in dense wireless networks. In [31], time-domain resource partitioning between the macrocell and small cells is invoked for optimizing the delivery of Dynamic Adaptive Streaming over HTTP (DASH). Xu *et al.* [32] present a video quality-aware mobile association, whilst giving cognizance both to the spectral and energy efficiency of two-tier heterogeneous networks. However, although the above-mentioned literature [17], [29], [30], [32] discusses adaptive video transmissions in heterogeneous wireless networks, none of them considers the cost levied due to the inference imposed by the small cells on the macrocell. Furthermore, the time scales of the video adaptation and the wireless resource allocation are very different in practical systems, which should be prudently considered, when streaming video over heterogeneous wireless networks.

III. System Model

We consider an OFDM-based two-tier macro-femto network, where FBSs are installed by residential users within the coverage-range of the MBS for the sake of streaming scalable video to multiple femtocell user equipment (FUE) over orthogonal wireless channels, as shown in Fig. 1. Spectrum-sharing is employed for enhancing the resource-exploitation. Following the framework proposed in [9], we assume that the macrocell operator imposes a certain 'contamination-tax' in form of an interference-price based on the level of interference power imposed on each slot by the FBS in order to motivate the mitigation of interference. The FBS is allowed to transmit video over the channels of the macrocell as long as the FUEs pay a compensation for their interference inflicted upon the macrocell. In this paper, we consider unicast based scalable video streaming for supporting multiple user-specific video streams. The system considered also employs sophisticated error resilient techniques in [33] for the transport of video

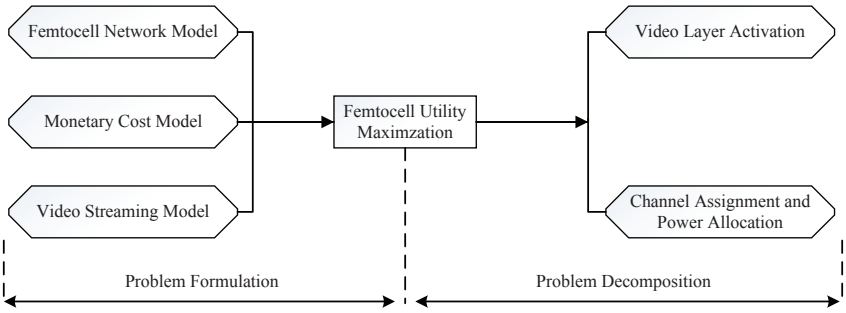


Fig. 2. The road map for conceiving scalable video streaming scheme in a two-tier HetNet.

over wireless networks, so that we can focus our attention on the problem of resource allocation and video layer selection. Specifically, advanced wireless link adaptation based on Adaptive Modulation and Coding (AMC) and Hybrid Automatic Repeat reQuest (HARQ) is applied for keeping the number of transmission errors low enough for video transmission. Furthermore, slice-structured video coding is employed for reducing the visual degradation inflicted by transmission errors. The technique of multiple reference frames is used to limit the error propagation. Due to decoding dependencies among the video layers, FBS schedules the lower video layers for transmission before scheduling any higher layers in order to guarantee their correct decoding order.

Fig. 2 characterizes the road map of our paper for conceiving scalable video streaming scheme in a two-tier HetNet. Its left part is the focus of this section for formulating the problem of femtocell utility maximization which rests on femto network model, monetary cost model and video streaming model, while the right part of Fig. 2 is the topic of Section IV to illustrate the problem solving by invoking classic Lyapunov drift theory [19] to decompose it into two tractable subproblems.

A. Femtocell Network Model

The femtocell operates within the frequency band of macrocell, which consists M orthogonal channels of the set $\mathcal{M} = \{1, \ldots, M\}$. A video server streams the video content through FBS to the FUEs $\mathcal{N} = \{1, \ldots, n\}$. We assume that the channel fading is *i.i.d.* across the time slots. Here, the time slot length is dependent on the specific communication system. For instance, the wireless time slot has a duration of 0.5ms in the context of LTE systems [34]. Different channels have independent and potentially different channel gain distributions for the different users. Here, we consider a block-fading model, where the channel gain remains constant during each time slot, but potentially changes from one slot to another. Let $h_{m,n}$ denote the gain of the channel m between the FBS and the FUE n. By inserting pilot symbols obeying a given time-frequency pattern in the downlink, FUE estimates the channel gain and feeds it back to FBS [35]. Hence, we assume that FBS estimates $h_{m,n}(t)$ at the start of all time slot t. For FBS, transmission over certain channels may cause severe interference to the macrocell, which leads to a high monetary cost. For the sake of reducing the interference contaminating the macrocell as well as achieving satisfactory data rates by exploiting the multiuser diversity and spectral diversity, it is necessary to design an efficient channel allocation and power assignment strategy for FBS.

We use a binary variable $x_{m,n}(t) \in \{0, 1\}$ for representing the channel assignment decision, where $x_{m,n} = 1$ if the channel m is allocated to FUE n, and $x_{m,n} = 0$ otherwise. Here, exclusive channel assignment is enforced in order to avoid interference among FUEs, hence any channel can only be allocated to a single FUE per slot. Hence, the channel assignment decision should satisfy

$$\sum_{n \in \mathcal{N}} x_{m,n}(t) \leq 1, \forall t. \tag{1}$$

Let $p_{m,n} \leq p_{max}$ denote the transmission power of channel m for the FUE n, where p_{max} represents the FBS's maximal transmission power for each channel allowed by the primary network. Then, the total transmission power of FBS satisfies

$$\sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} p_{m,n}(t) \leq P_{max}, p_{m,n}(t) \leq p_{max}, \forall t, \tag{2}$$

where P_{max} is the total transmission power constraint of the FBS. According to Shannon's formula, the maximum achievable rate of the FUE n at the time slot t is

$$r_n(t) = \sum_{m \in \mathcal{M}} x_{m,n}(t) \log \left(1 + \frac{p_{m,n}(t) h_{m,n}(t)}{\sigma^2_{m,n}}\right), \tag{3}$$

where $\sigma^2_{m,n}$ contains the background noise of the FUE n on channel m and the extra interference imposed by the MBS. Without loss of generality, it is assumed for convenience that we have $\sigma^2_{m,n} = \sigma^2, \forall m, n$ in the rest of this paper.

B. Monetary Cost Model

Let us assume that the macrocell announces a set of prices per unit of received interference power imposed by the femtocell at the start of each slot.[1] We use $\mu_m(t)$ to denote the unit interference price of the channel m during the time slot t. Note that the interference prices are determined by the pricing scheme at the macrocell side, but this issue is beyond the scope

[1]For simplicity, here we assume interference prices are announced each time slot. In fact, the derivation in the rest of the paper is still valid for the scenario where the price announcement has a longer and uncertain period of an integral multiple of time slots.

of our paper. A specific dynamic pricing strategy is proposed in [9], which can be straightforwardly employed in this paper. For femtocells, the sequence of consecutive interference prices can be deemed to be a stochastic process fluctuating on the time-slot scale. Let us define g_m as the channel gain between the FBS and the corresponding macrocell user of channel m. Then, for channel $m \in \mathcal{M}$, the interference power imposed by the FBS on the macrocell becomes $\sum_{n\in\mathcal{N}} x_{m,n}p_{m,n}g_m$. Hence, the total monetary cost of FBS during the slot t is

$$C(t) = \sum_{m\in\mathcal{M}} \mu_m(t) \sum_{n\in\mathcal{N}} x_{m,n}(t)p_{m,n}(t)g_m(t). \quad (4)$$

In practice, the total monetary cost of a FBS can be determined with the aid of the MBS. Specifically, the MUE of channel m can estimate the co-channel interference $I_m(t) = \sum_{n\in\mathcal{N}} x_{m,n}(t)p_{m,n}(t)g_m(t)$ imposed on channel m by the FBS, and further report it to the MBS. Having obtained these interference measurements, the MBS calculates the total monetary cost by $C(t) = \sum_{m\in\mathcal{N}} \mu_m(t)I_m(t)$, which is fed back to the FBS for performing its own resource allocation.

C. Video Streaming Model

Both the wireless channel and the interference price fluctuate stochastically over time, inevitably leading to the fluctuation of the maximum achievable rate of the FUEs. Hence, agile bitrate adaptation is crucial for maintaining a smooth video streaming service. In order to enable prompt video bitrate adaptation, we assumed that the quality-scalable layered video sequence has a single base layer plus $(L-1)$ potential enhancement layers. By either adding or dropping the enhancement layers, the video bitrate can be adjusted according to the state of the network.[2]

Generally, the video frames are partitioned into Groups of Pictures (GoPs) and the video bitrate is usually adapted at the boundary of GoPs [36]. In order to achieve high compression efficiency, the duration of a GoP is usually configured to be on the order of hundreds of milliseconds. By contrast, the channel allocation and power assignment are performed at the time scale of the wireless time slot lasting for tens of milliseconds. Hence, the interval of the video bitrate adaptation at the application layer is much longer than that of the resource allocation in the FBS PHY/MAC layer. For simplicity, the duration of each GoP is assumed to be an integer number of time slots. Explicitly, we use T to represent the number of time slots contained within the duration of each GoP. Thus, the video layer selection decision is made every T time slots, as shown in Fig. 3.

Let $d_n^k(l)$ denote the data size of the kth GoP of the FUE n's video, when the first l layers are chosen for streaming. Let $q_n^k(l)$ represent the corresponding video quality, *e.g.*, Peak Signal to Noise Ratio (PSNR) or Mean Opinion Score (MOS), associated with the transmission of the first l layers. FBS takes the action concerning the activation of the video layers $l(t)$ for each FUE only at the time instants of $t = kT$ $(k = 1, 2, \cdots)$.

[2]It should be noted that video bitrate adaptation can also be achieved by other techniques. For instance, DASH supports video bitrate adaptation by providing alternative chunks encoded at different bit rates.

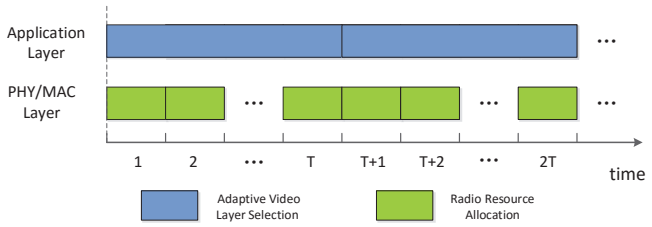


Fig. 3. Illustration of the different time-scales of both video layer activation and radio resource allocation.

Once the number of transmitted video layers has been adjusted, the video layer configuration remains unchanged during the next T time slots. For the FUE n, the video bitrate during the time slot $t \in [(k-1)T+1, kT]$ is kept constant, which is given as

$$\omega_n(t) \triangleq d_n^k(l)/T. \quad (5)$$

Correspondingly, the video quality of the FUE n during the time slot $t \in [(k-1)T+1, kT]$ is defined as

$$q_n(t) \triangleq q_n^k(l). \quad (6)$$

For a specific SVC encoded video, the PSNR/MOS associated with the first l layers is available from the video source [15], [37]. This means that $q_n(t)$ can be straightforwardly obtained according to (6).

If the video bitrate is kept higher than the available transmission rate, the playback buffer in the receiver is likely to become depleted, which results in playback interruption that substantially reduces the QoE of the user. Hence, we impose a constraint both on the channel bitrate and on the video bitrate in order to avoid playback interruptions. Specifically, the time-averaged transmission rate of the FUE n is defined as $\overline{r}_n \triangleq \lim_{T\to\infty} \frac{1}{T}\sum_{\tau=0}^{T-1} \mathbb{E}\big[r_n(\tau)\big]$, while the time-averaged video bitrate of the FUE n is formulated as $\overline{\omega}_n \triangleq \lim_{T\to\infty} \frac{1}{T}\sum_{\tau=0}^{T-1} \mathbb{E}\big[\omega_n(\tau)\big]$. Naturally, the average channel transmission rate should be higher than the video bitrate over a long time scale for the sake of avoiding playback interruptions. Hence, the following constraint should be satisfied,

$$\overline{r}_n \geq \overline{\omega}_n. \quad (7)$$

Although the constraint (7) does not explicitly characterize the average queuing delay constraint of the FBS, it ensures the stability of the queuing system in the FBS. Hence, the average queue length is bounded. This fact assists us in keeping the probability of playback interruption events close to zero by pre-buffering video frames in the queue of the FUE. Hence, applying the constraint (7) implicitly guarantees the delay constraint as a QoE metric.

D. Problem Formulation

The instantaneous system utility of scalable video streaming over femtocells is defined as

$$U(t) = \sum_{n\in\mathcal{N}} q_n(t) - \alpha C(t). \quad (8)$$

The first term in $U(t)$ is the total perceivable video quality of all FUEs, while the second one is the total monetary cost

charged by the macrocell operator in order to curb the co-channel interference. The parameter α in (8) is a weight invoked for striking a compromise between the quality of experience and the monetary cost. Hence, the time-averaged system utility is defined as

$$\overline{U} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{\tau=0}^{T-1} \mathbb{E}\big[U(\tau)\big]. \tag{9}$$

We aim for conceiving a joint video layer selection and resource allocation for maximizing the time-averaged utility, while satisfying the transmission constraint of (7) and the resource allocation constraint of (1) as well as (2). Hence, the femtocell Utility Maximization (FUM) problem is formulated as:

$$\begin{aligned}
\text{Maximize} \quad & \overline{U} && (10)\\
\text{Subject to} \quad & \overline{r_n} \geq \overline{\omega_n}, n \in \mathcal{N} && (11)\\
& x_{m,n}(t) \in \{0,1\}, && (12)\\
& \sum_{n\in\mathcal{N}} x_{m,n}(t) \leq 1, && (13)\\
& \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} p_{m,n}(t) \leq P_{max}, p_{m,n}(t) \leq p_{max}, && (14)\\
& l_n(t) \in \mathbb{L}_n \triangleq \{1,2,...,L_n\}. && (15)
\end{aligned}$$

Frequent adjustment of video enhancement layers incurs a fluctuating video quality, which may degrade the quality experienced by the viewer. This negative impact can be mitigated by incorporating a time-averaged video layer switching rate constraint into the problem (10)-(15), which has been proposed in our previous work [38], [39]. Specifically, we defined the layer difference between two consecutive time slots to quantify the quality variation of FUE n as $F_n(t) = |l_n(t) - l_n(t-1)|$, where $l_n(t)$ is the number of video layers to be transmitted to the FUE n at time slot t. In order to achieve an acceptable video quality variation, the time-averaged video layer variation of the FUE n is expected to remain below a tolerable threshold η_n. Accordingly, we defined the time-averaged video layer variation as $\overline{F}_n \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} F_n(t)$. Hence, the constraint $\overline{F}_n < \eta_n$ on video quality variation can be integrated into the problem (10)-(15) for establishing a problem formulation having quality smoothness.

It should be noted that we apply a weighted objective striking a tradeoff between the video quality maximization and the monetary cost minimization, which has to empirically choose the tradeoff parameter α. An alternative technique is to separate the item of monetary cost from the utility (8) and to formulate the problem in terms of maximizing the perceivable video quality at an acceptable monetary cost. Specifically, let κ represent the acceptable monetary cost, and define the time-averaged monetary cost of FBS as $\overline{C} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} C(t)$. Then, we can employ the cost constraint $\overline{C} \leq \kappa$ to formulate the video quality maximization subject to an acceptable monetary cost as well as to the playback smoothness constraint.

The problem (10)-(15) constitutes a stochastic optimization problem. Generally, solving such a problem requires knowledge of the distributions of the channel dynamics and of the interference price. However, there is no prior knowledge about them. In order to circumvent this difficulty, we opt for online measurements, rather than relying on any prior statistical knowledge, to seek the solution based on Lyapunov's optimization technique [19].

IV. Joint video Layer Selection, Channel allocation and Power assignment (JLCP) strategy

This section elaborates on deriving an online twin-time-scale algorithm for solving the FUM of (10)-(15) by exploiting the classic Lyapunov optimization theory of [19].

A. Problem Reformulation as Lyapunov Optimization

In order to solve the constrained FUM through the Lyapunov method, we construct a virtual queue $H_n(t)$ for transforming the time-averaged rate constraints (11) into the queue stability constraint. The dynamics of the queue $H_n(t)$ are formulated as:

$$H_n(t+1) = \big[H_n(t) + \omega_n(t) - r_n(t)\big]^+, \tag{16}$$

where $(a)^+ \triangleq \max(a, 0)$. From the theorem in [19], the mean rate stability of $H_n(t)$ implies that the time-averaged constraint (11) is satisfied. Then, a Lyapunov function associated with the virtual queues is defined as:

$$L(t) = \frac{1}{2} \sum_{n\in\mathcal{N}} H_n(t)^2.$$

Let $\boldsymbol{H}(t) \triangleq \{H_n(t), n \in \mathcal{N}\}$ represent a vector concatenated by $H_n(t)$. The conditional T-slot Lyapunov drift is defined as the expected variation of the Lyapunov function across T time slots:

$$\Delta_T(t) \triangleq \mathbb{E}\big[L(t+T) - L(t)|\boldsymbol{H}(t)\big]. \tag{17}$$

According to the classic theory of Lyapunov drift [19], minimizing $\Delta_T(t)$ asymptotically satisfies both the time-averaged channel bitrate and the time-averaged video bitrate constraints. Since the objective is to maximize the system utility under the constraints, a T-slot "drift-plus-penalty" is defined in the context of the Lyapunov optimization framework, which combines the Lyapunov drift and the instantaneous T-slot system utility of $\sum_{\tau=t}^{t+T-1} U(\tau)$ as follows:

$$\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U(\tau)|\boldsymbol{H}(t)\right], \tag{18}$$

where $\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U(\tau)|\boldsymbol{H}(t)\right]$ is the expected utility of the femtocell over T slots and V is a tradeoff parameter invoked for striking a compromise between the system's utility and the queue stability in the control policy. Therefore, we can re-interpret the optimization problem (10)-(15) as carrying out a combined decision concerning the video-layer selection, the channel allocation and the power assignment for minimizing the "drift-plus-penalty" of (18).

In order to develop our online JLCP strategy in the next section, we derive the upper bound of (18) as follows. At the start of all time slots, the channel allocation and power

assignment decisions can be made based on the current observations. However, the video layer selection is performed every T slots, which requires the prior knowledge of the future queue length during the interval spanning from the slot t to the slot $(t+T-1)$. Unfortunately, for any $\tau \in [t+1, t+T-1]$, the future queue length $\boldsymbol{H}(\tau)$ depends both on the video layer selection at the slot t and on the channel allocation as well as on the power assignment decisions during $[t, \tau]$, which is difficult to predict. Here, we follow the concept of [40] by approximating the near-future queue length based on the current observation (*i.e.*, $H(\tau) = H(t)$ for all $t < \tau \leq t+T-1$) to derive a relaxed upper bound of the drift-plus-penalty. This upper bound is described by **Lemma 1** as follows.

Lemma 1. *Let $t = kT$, where k is a non-negative integer. For any feasible decision, we have*

$$\begin{aligned}\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U(\tau)|\boldsymbol{H}(t)\right] \leq B \\ + T\sum_{n\in\mathcal{N}} \mathbb{E}\left[H_n(t)\omega_n(t) - Vq_n(t)|\boldsymbol{H}(t)\right] \\ + \sum_{\tau=t}^{t+T-1} \mathbb{E}\left[\alpha V C(\tau) - \sum_{n\in\mathcal{N}} H_n(t) r_n(\tau)|\boldsymbol{H}(t)\right],\end{aligned} \tag{19}$$

where $B \triangleq \frac{1}{2}T^2 N(r_{\max}^2 + \omega_{\max}^2)$ is a constant.

Proof: See the Appendix A. ∎

According to Lyapunov optimization framework [19], the bound (19) can be used as the objective function to develop the joint video layer selection, channel allocation and power assignment strategy.

B. Joint Video Layer Selection, Channel Allocation and Power Assignment Strategy

As shown in (19), the second item of the right part only depends on the video layer activation variables l_n, while the last item only relies on the channel and power decision variables $x_{m,n}$ and $p_{m,n}$. This specific structure provides a basis for us to decompose the optimization problem into two parallel sub-problems, one for the video layer selection and another for resource allocation, as discussed below.

1) Video Layer Activation: Owing to the separable structure of the bound (19), the optimal number of video layers can be obtained by maximizing the second term in the righthand side of (19) at the beginning of each slot $t = kT, k = 0, 1, \cdots$, which is given as:

$$\begin{aligned}\text{Maximize} \quad & \sum_{n\in\mathcal{N}} \left[Vq_n(t) - H_n(t)\omega_n(t)\right] \\ \text{Subject to} \quad & l_n(t) \in \mathbb{L}_n, \forall n.\end{aligned} \tag{20}$$

In (20), the first term is the weighted video quality of FUE n with the aid of $l_n(t)$ video layers, while the second one is interpreted as the corresponding queuing weighted video bitrate, which is determined by the drift quantity of the Lyapunov function for the sake of stabilizing the virtual queues. Increasing V enables the proposed strategy to improve the received video quality of FUEs, while decreasing V is helpful for stabilizing the virtual queues.

It may be observed that the problem (20) can be further separated into N subproblems, each corresponding to a FUE. Hence, the optimal number of video layers for the nth FUE can be obtained by enumerating all possible layers $l_n \in \mathbb{L}_n$. Generally, the number of available enhancement layers is very limited in practical systems, for example to 2-6 enhancement layers. Furthermore, the video layer selection is performed every T slots. Hence, the computational complexity of solving the problem (20) is appealingly low.

2) Channel Allocation and Power Assignment: In this subsection, we derive the optimal channel allocation and power assignment strategy by minimizing the last term at the righthand side of (19) at the start of the slot t. Substituting (3) and (4) into the last term of (19), the channel allocation and power assignment problem is written as:

$$\begin{aligned}\min \quad & \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} \beta_m(t) x_{m,n}(t) p_{m,n}(t) - \\ & \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} H_n(t) x_{m,n}(t) \log\left(1 + \frac{p_{m,n}(t)h_{m,n}(t)}{\sigma^2}\right)\end{aligned} \tag{21}$$

$$\begin{aligned}\text{s.t.} \quad & x_{m,n}(t) \in \{0,1\}, \\ & \sum_{n\in\mathcal{N}} x_{m,n}(t) \leq 1, \forall m, \\ & \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_{m,n}(t) \leq P_{max}, p_{m,n}(t) \leq p_{max},\end{aligned} \tag{22}$$

where the new notation of $\beta_m(t) \triangleq \alpha V \mu_m(t) g_m(t)$ is used to simplify the expression. Similarly, the first term of (21) is the weighted monetary cost charged to the femtocell, while the second term may be viewed as a queuing-weighted transmission rate determined by the Lyapunov drift for stabilizing the virtual queues. For notational simplicity, we omit the time index t in the analysis below.

Due to the binary nature of the channel allocation variables $x_{m,n}$, the problem (21) is a mixed-integer problem, which is challenging. Here, we apply a two-stage approach to solve this problem. The first stage aims for finding the optimal power assignment for any given binary channel allocation, while the second stage is devoted to seeking the optimal channel allocation by utilizing the optimal power solution gleaned from the first stage.

In the first stage, where the channel allocation is determined, the first term in (21) is an affine function while the second term is a concave function. According to convex optimization theory [41], the power assignment problem (21) is a convex one. The optimal solution of this problem is given by **Proposition** 1.

Proposition 1. *For a fixed channel allocation $x_{m,n}$, the optimal power assignment of the problem* (21)-(22) *is expressed as*

$$p_{m,n}^* = \left(\frac{H_n x_{m,n}}{\lambda + \beta_m} - \frac{\sigma^2}{h_{m,n}}\right)^+, \tag{23}$$

where λ represents the Lagrange factor used for weighting the power constraint (22)*.*

Proof: See the Appendix B. ∎

Although the optimal power of (23) in **Proposition** 1 is reminiscent of the classic water-filling algorithm [42], it is much more complex, since the traditional water-filling only has a single water level. Specifically, the water level in the traditional water-filling algorithm relies on a single parameter, λ, while the one in the proposed solution (23) depends both on λ as well as on H_n, α and $\mu_m g_m$. These parameters reflect the mismatch between the video bitrate and transmission rate as well as the mismatch between FUE's willingness-to-pay and the interference price for the unit transmission power, which result in different water levels for different channels. At the same channel quality, an increased value of the water level implies having an increased transmission power, thus enhancing the transmission bitrate [43]. Therefore, the channel's adaptation can be achieved by the optimal power assignment in (23) that is driven by the sensitivity to the specific transmission conditions (e.g., channel quality, rate mismatch) as well as to the monetary costs.

The optimal value of the key parameter λ can be found as follows. Recall the complementary condition (46) from Appendix B, *i.e.*,

$$\lambda \left(\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} p^*_{m,n} - P_{max} \right) = 0, \tag{24}$$

where $\lambda \geq 0$ and $p^*_{m,n}$ is strictly decreasing upon increasing λ, as shown in **Proposition** 1. For $\lambda = 0$, the corresponding optimal power assignment is $p^*_{m,n} = \left(\frac{H_n x_{m,n}}{\beta_m} - \frac{\sigma^2}{h_{m,n}} \right)^+$. If this power assignment satisfies the power constraint, *i.e.*,

$$\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \left(\frac{H_n x_{m,n}}{\beta_m} - \frac{\sigma^2}{h_{m,n}} \right)^+ \leq P_{max}, \tag{25}$$

then this solution is feasible, and the optimal Lagrange multiplier becomes $\lambda^* = 0$. Otherwise, it is an infeasible solution, because the allocated power is higher than the available power, which implies that the optimal value of λ should be larger than 0. In this case, we can find $\lambda > 0$ by satisfying

$$\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} p^*_{m,n} - P_{max} = 0. \tag{26}$$

According to **Proposition** 1, it is true that $\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} p^*_{m,n} = 0 < P_{max}$ when $\lambda \to \infty$, and $\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} p^*_{m,n} > P_{max}$ when $\lambda = 0$. Since $p^*_{m,n}$ is monotonically non-increasing upon increasing λ, there is a unique solution λ^* to (26). Numerically, λ^* can be found by an efficient method, such as the classic bisection search or golden section search [44].

Below, we present the second stage invoked for finding the optimal channel allocation $x_{m,n}$. Substituting the optimal power (23) into (21) as well as exploiting the binary nature of $x_{m,n}$, we formulate the following optimization problem for the channel allocation:

$$\begin{aligned} \text{Minimize} \quad & \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \phi_{m,n} x_{m,n} \\ \text{Subject to} \quad & x_{m,n} \in \{0,1\}, \\ & \sum_{n\in\mathcal{N}} x_{m,n} \leq 1, \forall m, \end{aligned} \tag{27}$$

where

$$\phi_{m,n} \triangleq \beta_m \left(\frac{H_n}{\lambda + \beta_m} - \frac{\sigma^2}{h_{m,n}} \right)^+ - H_n \left(\log \left(\frac{H_n h_{m,n}}{\sigma^2(\lambda + \beta_m)} \right) \right)^+. \tag{28}$$

It should be noted that (28) is derived by exploiting the binary nature of $x_{m,n}$, which assists us in simplifying the objective in (27) as a linear function of $x_{m,n}$. Its minimization can be achieved by independently minimizing $\sum_{n\in\mathcal{N}} \phi_{m,n} x_{m,n}$ of each channel $m \in \mathcal{M}$. This implies that for a particular channel m, we can find its optimal allocation to FUE n^*_m by minimizing $\phi_{m,n}$, *i.e.*,

$$n^*_m \triangleq \arg\min_{n\in\mathcal{N}} \phi_{m,n}. \tag{29}$$

Then, the optimal channel allocation is given by

$$x^*_{m,n} = \begin{cases} 1, & \text{if } n = n^*_m, m \in \mathcal{M} \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Note that λ^* and $x^*_{m,n}$ are coupled with each other, which means that determining λ^* requires the value of $x_{m,n}$, while $x^*_{m,n}$ (or $\phi_{m,n}$) depends on λ. This coupling property imposes additional difficulty on solving the problem. The conceptually simplest way of solving this coupled problem is to consider all the possible channel allocations for finding the optimal solution. However, the computational complexity of this method may become excessive due to the large-scale search-space of the channel allocation combinations (*i.e.*, N^M). Below, we exploit the specific structure of the problem to derive a low-complexity solution.

If $H_n h_{m,n} > \sigma^2(\lambda + \beta_m)$, from (28) we arrive at:

$$\frac{\partial \phi_{m,n}}{\partial H_n} = -\frac{\lambda}{\lambda + \beta_m} - \log \left(\frac{H_n h_{m,n}}{\sigma^2(\lambda + \beta_m)} \right). \tag{31}$$

Otherwise,

$$\frac{\partial \phi_{m,n}}{\partial H_n} = 0. \tag{32}$$

Hence, $\phi_{m,n}$ is non-increasing upon increasing H_n. Similarly, it can be shown that $\phi_{m,n}$ is also non-increasing, when increasing $h_{m,n}$. This implies that assigning the channel to the FUEs having high channel quality or a high queue length is beneficial for improving the system's utility. It provides a basis for designing a simple greedy algorithm for the channel allocation. Intuitively, a particular channel m is assigned to the specific FUE having the longest virtual queue, provided that the channel gain for this FUE satisfies $h_{m,n} > \frac{\sigma^2(\lambda+\beta_m)}{H_n}$. Since the virtual queue length H_n indicates the degree of mismatch between the bitrate of the video stream and the available bitrate of the wireless channel, the above allocation policy can be interpreted by assuming that the FBS prefers to assign channels to specific FUEs, whose transmission rate is much lower than the video bitrate during the past time slots, so that the probability of playback interruptions can be reduced. However, the channel gain for this FUE should be good enough to satisfy $h_{m,n} > \frac{\sigma^2(\lambda+\beta_m)}{H_n}$. Otherwise, the optimal power is zero according to (23), which leads to a waste

of spectral resources. Hence, we define the feasible FUE set for the channel m by

$$\mathcal{N}_m \triangleq \{n \in \mathcal{N} : H_n h_{m,n} > \sigma^2(\lambda + \beta_m)\}. \quad (33)$$

Then, the detailed procedure of the proposed algorithm is presented as follows. Initially, we assign the channels to the FUEs associated with the largest queue length, *i.e.*,

$$\widehat{n}_m = \arg\max_{n\in\mathcal{N}} H_n, m \in \mathcal{M}. \quad (34)$$

Then, calculate λ according to (24). Once λ is obtained, we can determine the feasible FUE set $\mathcal{N}_m, m \in \mathcal{M}$. For each channel m, it is necessary to check whether the selected FUE $\widehat{n}_m$ is in the feasible FUE set $\mathcal{N}_m$. If not, we update the allocation of the channel m by seeking the particular FUE $\widetilde{n}_m$ having the largest queue length in its feasible FUE set, *i.e.*, $\widetilde{n}_m = \arg\max_{n\in\mathcal{N}_m} H_n$. This procedure is repeated by calculating λ and assigning channels iteratively, until (24) and (29) are satisfied. We summarize the above procedure in ***Algorithm 1***.

Algorithm 1 Channel Allocation and Power Assignment

$n_m^* \leftarrow \arg\max_{n\in\mathcal{N}} H_n, \forall m \in \mathcal{M}$

procedure CALCULATING λ
  if inequality (25) holds **then**
    $\lambda \leftarrow 0$
  else
    $\lambda \leftarrow$ the solution of (26)
  end if
  return λ
end procedure

while $H_{n_m^*} h_{m,n_m^*} \leq \sigma^2(\lambda + \beta_m), \forall m \in \mathcal{M}$ **do**
  $n_m^* \leftarrow \arg\max\{H_n | H_n h_{m,n} > \sigma^2(\lambda + \beta_m)\}$
  invoke procedure of calculating λ
end while

Then, achieve the optimal channel allocation by (30), and obtain the optimal power assignment by (23).

C. Performance Analysis

Assumption 1. *If a quantity $\delta > 0$ and an array of actions $l_n(t)$ for $t = kT, k = 0, 1, \cdots$ and $\{x_{m,n}(t), p_{m,n}(t)\}$ for $t = 0, 1, \cdots$ exist, such that the inequalities below are satisfied*

$$\frac{1}{T}\sum_{\tau=KT}^{KT+T-1} \mathbb{E}[\omega_n(\tau)] < \frac{1}{T}\sum_{\tau=KT}^{KT+T-1} \mathbb{E}[r_n(\tau)] - \delta, \forall n \in \mathcal{N}. \quad (35)$$

This assumption implies that there exists at least one control policy that makes the resultant transmission rate strictly higher than the video bitrate for each FUE. Actually, this assumption is not restrictive at all, because setting $l_n = 1$ is a feasible control policy. Hence, the conditions (35) ensure that the channel conditions/interference price is good/low enough for supporting the transmission of the base layer. Otherwise, no scheduling is capable of providing acceptable video quality. This assumption is quite mild and reasonable in practice.

Below, we formulate a theorem for theoretically quantifying the bound on the performance that the proposed method can achieve.

Theorem 1. *Let us assume that the conditions* (35) *are satisfied for $\exists\delta > 0$, and that the virtual queue length is initially zero, i.e., $H(0) = 0$. For $\forall V > 0$, we have two properties concerning the JLCP Policy as follows:*

(a) The average lengths of the virtual queues are upper bounded by

$$\lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\sum_{n\in\mathcal{N}} H_n(kT)] \leq \frac{B + VT(U_{\max} - U_{\min})}{\delta T}, \quad (36)$$

where U_{max} and U_{min} are the maximum and minimum utility, respectively.

(b) The utility achieved by JLCP satisfies

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[U_{JLCP}] \geq \overline{U^*} - \frac{B}{VT}, \quad (37)$$

where $\overline{U^*}$ denotes the theoretically optimal femtocell utility concerning the problem (10-15).

Proof: See the Appendix C. ■

Theorem 1 implies that the average utility of the JLCP strategy asymptotically approximates the optimal system utility of $\overline{U^*}$ upon increasing V, while the average lengths of the virtual queue also increase as V increases. Note that the virtual queue length characterizes the video stream backlog due to the mismatch between the transmission rate and video bitrate. This means that a large value of V leads to a higher queuing delay in the FBS. However, we can reduce the risk of playback interruption by setting an appropriate pre-buffer level in the FUE. Explicitly, initially the received video data will not be played out until the buffer has sufficient video data, since the queue lengths are bounded by $O(V)$, as shown in (36). In general, a long pre-buffer time degrades the users' experience. Therefore, the configuration of V depends on the specific compromise struck between the utility-optimality and the queuing-delay (pre-buffer time). Furthermore, δ characterizes the remaining capability of the system to support the video traffic, and a smaller value of δ results in a higher average queueing delay in FBS.

V. Performance Evaluation

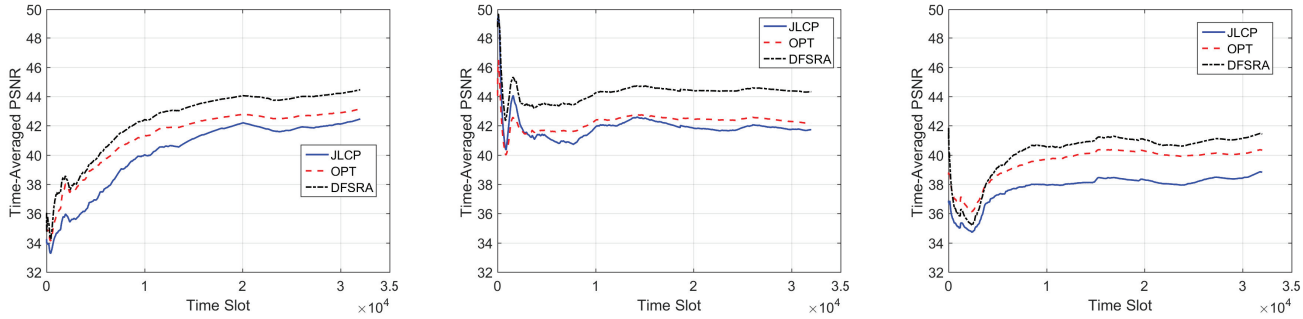
This section is devoted to investigating the attainable performance that the JLCP strategy is able to achieve. The average monetary cost, the PSNR and the playback interruption ratio are adopted as the performance metrics. For comparison, we choose the theoretic solution for achieving optimal utility as a benchmark, which requires the non-causal apriori knowledge of all channel conditions, interference prices and video traffic statistics. Furthermore, the interference-oblivious algorithm, as well as the DASH-Friendly Scheduling and Resource Allocation scheme (DFSRA) proposed in [45], which was specifically designed for improving the video quality and

TABLE I
TWO-TIER FEMTOCELL NETWORK SETUP FOR EXPERIMENTS

Parameters	Values/assumptions
Carrier frequency	2GHz
System bandwidth	2 MHz
Number of sub-channels	10
Number of FUE	3
Power constraint of FBS	23 dbm
Distribution of channel gain between FBS and FUE	Rayleigh with variance 0.376
Distribution of channel gain between FBS and MUE	Rayleigh with variance 0.05
Distribution of interference price	Gaussian (mean of 50 and standard deviation of 10)

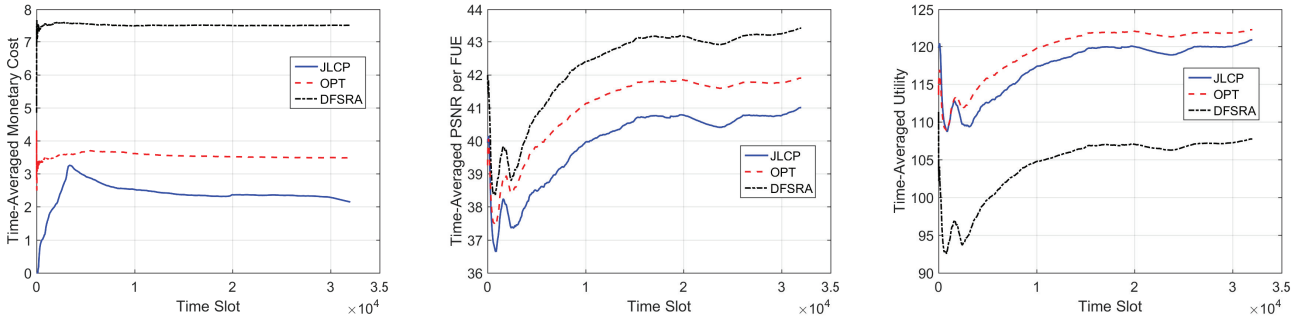
TABLE II
STATISTICS OF THE VIDEO SEQUENCES

Video Sequence \ Layers		Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7
Star War IV	Average PSNR (dB)	37.75	39.42	41.21	42.43	44.01	45.27	45.69
	Average Bitrate (kbps)	61.58	197.24	269.9	318.89	371.08	411.15	434.04
Indiana Jones	Average PSNR (dB)	35.24	36.76	38.61	39.82	41.55	42.93	43.58
	Average Bitrate (kbps)	108.45	323.3	439.77	517.93	605.38	673.65	713.32
Silence of the Lambs	Average PSNR (dB)	38.06	40.39	42.41	43.45	44.65	45.32	45.57
	Average Bitrate (kbps)	58.04	210.15	280.11	324.12	370.99	403.41	424.70



(a) FUE 1 (b) FUE 2 (c) FUE 3

Fig. 4. Average PSNR of the three users for joint video layer selection, channel allocation and power assignment strategy (JLCP), the optimal algorithm (OPT) as well as Dash-friendly scheduling and resource allocation scheme (DFSRA) using the parameters of Table I.



(a) Average monetary cost versus time (b) Average PSNR per FUE versus time (c) Average utility versus time

Fig. 5. Performance comparison of joint video layer selection, channel allocation and power assignment strategy (JLCP), the optimal algorithm (OPT) as well as Dash-friendly scheduling and resource allocation scheme (DFSRA) using the parameters of Table I.

playback smoothness, was also implemented as a benchmark algorithm.

A. Simulation Setup

In the simulations, a single FBS is installed within the range of a single MBS to formulate a twin-tier spectrum-sharing network. There are 3 FUEs in the system, and the system bandwidth is equally divided into 10 channels, each with 200KHz bandwidth. Here, we adopted the classic Rayleigh channel model. Speficially, the downlink spanning from FBS to FUE, namely $h_{m,n}$, has a Rayleigh-distributed received signal associated with a variance of 0.376, while that from FBS

to MUE, namely g_m, has a variance of 0.05. This implies that the received signal power of the FBS - FUE link is higher than that of the FBS - MUE link. The maximum power constraint of the FBS P_{max} is set to 23 dBm. The unit prices of interference μ_m are Gaussian random variables with a mean of 50 and standard deviation of 10. The specific parameters of the two-tier network considered in the simulations were summarized in Table I.

The "Star War IV", "Indiana Jones" and "Silence of the Lambs" [37] clips were adopted to generate the video traffic in the experiment. Each clip contains 1 base layer as well as 6 enhancement, and their video frame rate is 30 fps. The FUEs randomly access one of the above three videos. Each GoP has a pattern of G16B15, *i.e.*, IBBBBBBBBBBBBBBIBB, where hierarchical B frames were used, as recommended in [46]. The statistics concerning the video clips are described in Table II. The number of video layers is only changed GoP-by-GoP. The duration of a time slot equals to one-tenth of that of each GoP.

B. Simulation Results

1) Performance Comparison: For the sake of characterizing the performance improvement of our JLCP strategy, we compared it to the optimal algorithm (OPT) and to the DFSRA. The parameters of JLCP were as follows. The compromise between the system utility and the virtual queue stability V was set to 3×10^6. The reason for assuming such a large value of V is because the queue length has a large value due to using the unit of bits. The tradeoff between the video quality and monetary cost α was set to 1. In order to achieve a smooth playback, the number of pre-buffered frames in the FUE was set to 130.

In Fig. 4 and Fig. 5, we plotted our experimental results concerning the PSNR values of the three users, PSNR per FUE, average monetary cost and average utility. It is seen from Fig. 4 and Fig. 5b that DFSRA achieves the highest PSNR. The basic reason for this is that DFSRA is designed for maximizing the video quality as well as playback smoothness, which does not take cross-tier interference into account. It always tries to increase the transmission power to support a higher video bitrate regardless of the interference price. Consequently, it incurs the highest monetary cost, as shown in Fig. 5a. Fig. 4 and Fig. 5b also illustrates that OPT attains a higher PSNR than that of JLCP, while JLCP achieves a lower monetary cost than that of OPT, as shown in Fig. 5a. The reason is that OPT aims for optimizing the time average utility by using the statistical information of channels, interference price and video sequences. It is insensitive to the short-term fluctuation of these environmental conditions. By contrast, the advantage of JLCP is its capability to adapt to the dynamically fluctuating environmental conditions by leveraging the queueing shift. The fluctuation of these environmental conditions may result in the fluctuation of virtual queues $\boldsymbol{H}(t)$. The virtual queues $\boldsymbol{H}(t)$ will be higher when the channel cannot afford the video traffic, which enables JLCP to choose a reduced-layer configuration. Although the lower layer configuration leads to a lower PSNR, it also incurs a lower monetary cost. From Fig. 5c, we can see that the time-averaged utility achieved by JLCP is much higher than that of DFSRA, although it is a little bit lower than that of OPT. However, due to the adaptability of JLCP, it achieves a much lower playback interruption rate (PIR) per minute, than that of OPT, as shown in Fig. 6. Although the playback interruption rate achieved by DFSRA is a little bit lower than that of JLCP, this is achieved at the cost of an increased average monetary cost. Furthermore, JLCP can achieve a sufficiently low playback interruption rate by setting the appropriate number of pre-buffered frames.

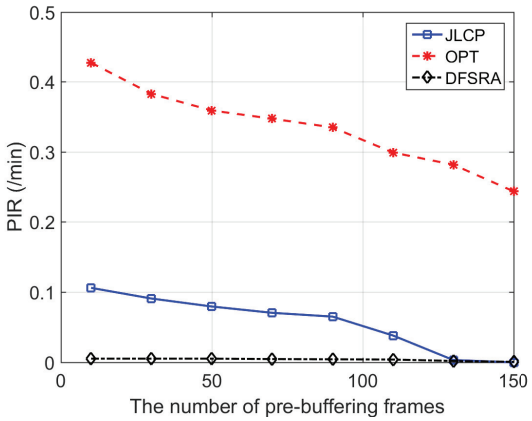


Fig. 6. The sensitivity of average playback interruption rate to different number of pre-buffered video frames in the FUE.

2) Performance Sensitivity to Parameters: In this section we aim for evaluating the effect of the parameters on JLCP. We first conducted experiments for investigating the sensitivity of the playback interruption rate to the initial buffer fullness of the FUE. The number of pre-buffered frames was set to $10, 30, 50, 70, 90, 110, 130, 150$, respectively. The remaining simulation parameters were set to the same values as in Section V-B1. All the simulation results were averaged over 50 independent runs, which can be seen in Fig. 6.

Fig. 6 shows that the average playback interruption rates of JLCP, DFSRA and OPT all decrease upon increasing the initial buffer fullness. This phenomenon coincides with the law of buffering, where more buffered frames have a higher capability to compensate for the fluctuation of the channel quality. It can also be observed that JLCP achieves a sufficiently low playback interruption rate. Although a lower playback interruption rate can be achieved with more pre-buffered frames, there is a trade-off to be considered, because the increasing number of pre-buffered frames results in a longer waiting time, which may severely degrade the viewers' experience. Fig. 6 shows that an extremely low playback interruption rate can be achieved by setting the number of pre-buffered frames to as high as 130.

Next, we performed further experiments for assessing the impact of the tradeoff factor V on JLCP. The values of V were set to $1 \times 10^6, 3 \times 10^6, 5 \times 10^6, 7 \times 10^6, 9 \times 10^6, 11 \times 10^6$, and the rest of the parameters were set to the same values as in Section V-B1. All the simulation results were averaged over 50 independent runs, which are portrayed in Fig. 7 and 8.

It is illustrated in Fig. 7 that increasing V is beneficial for improving the average PSNR at the cost of an increased

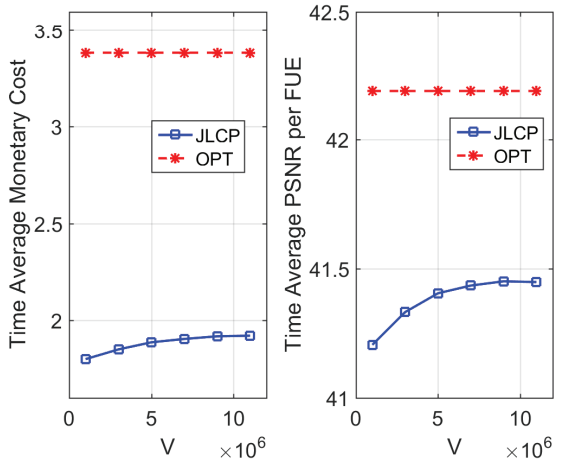


Fig. 7. The sensitivity of average monetary cost and average PSNR to the parameter V of (18).

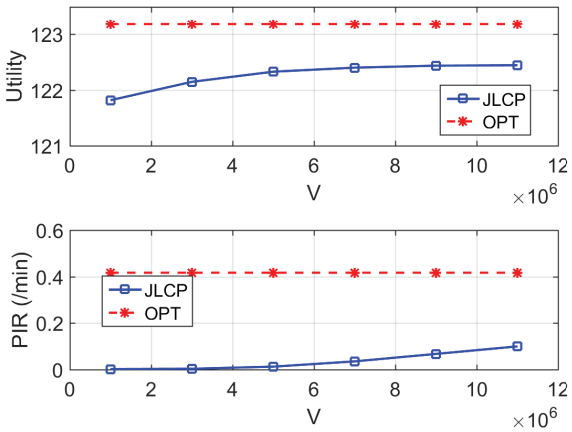


Fig. 8. Sensitivity of average utility and average playback interruption rate to the parameter V of (18).

average monetary cost. This is because a larger V imposes a higher weight on the video quality. Thus, the JLCP strategy is inclined to enhance the video quality by making the layer selection decisions by minimizing (20). However, the video bitrate also has to be increased for increasing the video quality. Hence, the transmit power of the FBS has to be increased for meeting the bitrate requirement of streaming video, which in turn increases the monetary cost. Fig. 8 further illustrates that both the average utility and the average playback interruption rate are increased as the tradeoff factor V increases. This result verifies Theorem 1 that the average length $\boldsymbol{H}(t)$ of the virtual queue and the average utility achieved by JLCP both increase upon increasing V. The higher queue length $\boldsymbol{H}(t)$ implies a higher mismatch between the video bitrate and the available transmission rate, which further increases the playback interruption probability. Hence, there is a compromise between the average utility as well as the playback interruption rate, and the control parameter V allows us to attain different average utility/playback smoothness trade-offs.

Furthermore, we conducted experiments for assessing the impact of the weighting factor α on JLCP. The values of α were set to $0.9, 0.95, 1, 1.05, 1.1$, and the rest of the parameters were set to the same values in Section V-B1. All the simulation results were averaged over 50 independent runs, which were given by Fig. 9.

It is illustrated in Fig. 9a and 9b that as the parameter α increases, both the average PSNR per FUE and the cost decrease. This is because with a larger α, JLCP tends to control the interference imposed on the macrocell by reducing the transmission power, which results in a reduced cost. Nevertheless, the scheme has to select fewer video layers to match the low transmission rate due to a low transmission power, which reduces the video quality.

VI. Conclusions

This paper discussed the problem of streaming scalable videos over two-tier spectrum-sharing femtocell networks. We established a joint resource allocation and video bitrate adaptation model for improving the weighted utility defined as a combination of the video quality and the monetary cost due to the interference imposed by the femtocell on the macrocell. By applying the Lyapunov optimization technique, we decomposed the problem into two subproblems, which can be optimized independently at different time scale. By exploiting the specific structure of subproblems, we derived an online measurement based strategy for resource allocation and layer selection, which operate without any prior statistical knowledge. Analytical bounds were also derived for the proposed algorithm for showing the theoretically achievable performance. We evaluated the efficiency of the proposed online algorithm through simulations with the aid of real video traces. The results illustrated that the proposed algorithm provided a prompt response to the dynamics of the environment. We also demonstrated that the proposed solution was capable of approaching the optimal solution by controlling the tradeoff parameter V.

In this paper, we have discussed the proposed solution in the scenario of sparsely located femtocells, so that the interference imposed by the other femtocells may be deemed negligible. It can also be readily extended to densely deployed scenarios by invoking inter-cell interference mitigation techniques, such as coordinated time-domain muting between small cells relying on enhanced inter-cell interference coordination (eICIC) and coordinated multipoint (CoMP) or enhanced CoMP (eCoMP) techniques. We will investigate these solutions in our future work.

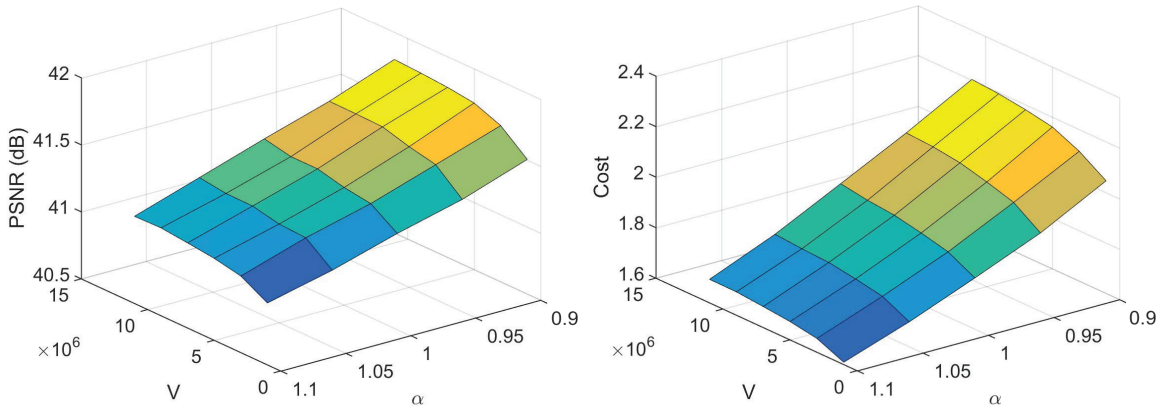
Appendix A
Proof of Lemma 1

Squaring the queue dynamics (16) yields

$$H_n^2(t+1) \leq H_n^2(t) + \omega_n^2(t) + r_n^2(t) + 2H_n(t)[\omega_n(t) - r_n(t)].$$

Then, we have

$$\mathbb{E}\left[\frac{1}{2}H_n^2(t+1) - \frac{1}{2}H_n^2(t)|\boldsymbol{H}(t)\right] \leq \frac{1}{2}N(\omega_{\max}^2 + r_{\max}^2) + \mathbb{E}\left[\sum_{n\in\mathcal{N}} H_n(t)(\omega_n(t) - r_n(t))|\boldsymbol{H}(t)\right].$$



(a) Average PSNR per FUE

(b) Average monetary cost

Fig. 9. Sensitivity of average PSNR and average cost to the parameter α of (8).

By summing this inequality over the span of $[t, t+T-1]$ and subtracting the weighted T-slot expected utility, we attain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}H_n^2(t+T) - \frac{1}{2}H_n^2(t) - V\sum_{\tau=t}^{t+T-1}U(\tau)|\boldsymbol{H}(t)\right] \leq B_1 \\
&+ \mathbb{E}\left[\sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} H_n(\tau)(\omega_n(t) - r_n(\tau))|\boldsymbol{H}(t)\right] \\
&- V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}\left(\sum_{n\in\mathcal{N}} q_n(t) - \alpha C(\tau)\right)|\boldsymbol{H}(t)\right] \\
&= B_1 + \sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} \mathbb{E}\left[H_n(\tau)\omega_n(t) - Vq_n(t)|\boldsymbol{H}(t)\right] \\
&+ \sum_{\tau=t}^{t+T-1}\mathbb{E}\left[\alpha VC(\tau) - \sum_{n\in\mathcal{N}} H_n(\tau)r_n(\tau)|\boldsymbol{H}(t)\right],
\end{aligned} \tag{38}$$

where $B_1 \triangleq \frac{1}{2}TN(r_{\max}^2 + \omega_{\max}^2)$ is a constant.

We approximate $H_n(\tau)$ by $H_n(t)$ in the second term of the right-hand side of (38) to derive a relaxed upper bound. Using the fact that for any $\tau \in [t, t+T-1]$

$$H_n(t) - (\tau - t)r_{\max} \leq H_n(\tau) \leq H_n(t) + (\tau - t)\omega_{\max},$$

we can obtain

$$\begin{aligned}
&\sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} H_n(\tau)\omega_n(t) \leq \sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}}[H_n(t)\omega_n(t) + (\tau-t)\omega_{\max}^2] \\
&= T\sum_{n\in\mathcal{N}} H_n(t)\omega_n(t) + \frac{1}{2}T(T-1)N\omega_{\max}^2
\end{aligned} \tag{39}$$

and

$$\begin{aligned}
&-\sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} H_n(\tau)r_n(\tau) \leq \\
&\sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}}[-H_n(t)r_n(\tau) + (\tau-t)r_{\max}^2] \\
&= -\sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} H_n(t)r_n(\tau) + \frac{1}{2}T(T-1)Nr_{\max}^2.
\end{aligned} \tag{40}$$

Substituting (39) and (40) into the second term of (38), we have

$$\begin{aligned}
&\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}U(\tau)|\boldsymbol{H}(t)\right] \leq B \\
&+ T\sum_{n\in\mathcal{N}}\mathbb{E}\left[H_n(t)\omega_n(t) - Vq_n(t)|\boldsymbol{H}(t)\right] \\
&+ \sum_{\tau=t}^{t+T-1}\mathbb{E}\left[\alpha VC(\tau) - \sum_{n\in\mathcal{N}} H_n(t)r_n(\tau)|\boldsymbol{H}(t)\right],
\end{aligned} \tag{41}$$

where $\triangleq B_1 + \frac{1}{2}T(T-1)N\omega_{\max}^2 + \frac{1}{2}T(T-1)Nr_{\max}^2 = \frac{1}{2}T^2N(r_{\max}^2 + \omega_{\max}^2)$ is a positive constant.

Appendix B
Proof of Proposition 1

As mentioned previously, the original power assignment problem (21) exhibits convexity. Hence, we can solve this problem by using the classic Lagrange duality theory [41]. For simplifying the notation, the index t is omitted in the follow-up derivation below.

We define the Lagrangian L associated with the power assignment problem as

$$\begin{aligned}
&L(\mathbf{p}, \lambda, \boldsymbol{\nu}) = \alpha V\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\mu_m x_{m,n}p_{m,n}g_m - \\
&\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} H_n x_{m,n}\log\left(1 + \frac{p_{m,n}h_{m,n}}{\sigma^2}\right) + \\
&\lambda(\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_{m,n} - P_{max}) - \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\nu_{m,n}p_{m,n},
\end{aligned} \tag{42}$$

where the Lagrange factors λ and $\nu_{m,n}$ are used for weighting the constraints $\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_{m,n} \leq P_{max}$ and $p_{m,n} \geq 0$, respectively.

According to the classic Lagrange dual theory, we have a dual function of $\mathcal{G}(\lambda, \boldsymbol{\nu}) = \inf_{\mathbf{p}} L(\mathbf{p}, \lambda, \boldsymbol{\nu})$. Then, the Lagrange dual problem corresponding to the original power assignment problem is written as:

$$\text{Maximize} \quad \mathcal{G}(\lambda, \boldsymbol{\nu}), \text{s.t.} \lambda \geq 0, \boldsymbol{\nu} \succcurlyeq \mathbf{0}.$$

According to the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_{m,n} - P_{max} \le 0, \tag{43}$$

$$p_{m,n} \ge 0, \quad \forall m, \forall n, \tag{44}$$

$$\lambda \ge 0, \nu_{m,n} \ge 0, \quad \forall m, \forall n, \tag{45}$$

$$\lambda(\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} p_{m,n} - P_{max}) = 0, \tag{46}$$

$$\nu_{m,n} p_{m,n} = 0, \quad \forall m, \forall n, \tag{47}$$

$$\frac{\partial L(\mathbf{p}, \lambda, \boldsymbol{\nu})}{\partial p_{m,n}} = 0, \quad \forall m, \forall n. \tag{48}$$

Equation (48) yields

$$\alpha V \mu_m x_{m,n} g_m - H_n x_{m,n} \frac{h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} + \lambda - \nu_{m,n} = 0. \tag{49}$$

Upon substituting (45) into (49), we can eliminate $\nu_{m,n}$, which yields

$$\alpha V \mu_m x_{m,n} g_m - H_n x_{m,n} \frac{h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} + \lambda \ge 0. \tag{50}$$

From (47) and (49), we have

$$p_{m,n}\left[\alpha V \mu_m x_{m,n} g_m - H_n x_{m,n} \frac{h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} + \lambda\right] = 0. \tag{51}$$

We consider the following specific cases:

- *Case 1:* $\lambda > \frac{H_n x_{m,n} h_{m,n}}{\sigma^2} - \alpha V \mu_m x_{m,n} g_m$. If $p_{m,n} > 0$, then from (51) we have $\lambda = \frac{H_n x_{m,n} h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} - \alpha V \mu_m x_{m,n} g_m < \frac{H_n x_{m,n} h_{m,n}}{\sigma^2} - \alpha V \mu_m x_{m,n} g_m$, which contradicts to the assumption. Thus we have $p_{m,n} = 0$.
- *Case 2:* $\lambda \le \frac{H_n x_{m,n} h_{m,n}}{\sigma^2} - \alpha V \mu_m x_{m,n} g_m$. If $\alpha V \mu_m x_{m,n} g_m - H_n x_{m,n} \frac{h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} + \lambda > 0$, we have $p_{m,n} = 0$ according to (51), and thus $\lambda > \frac{H_n x_{m,n} h_{m,n}}{\sigma^2} - \alpha V \mu_m x_{m,n} g_m$, which again contradicts to the assumption. Thus, it follows $\alpha V \mu_m x_{m,n} g_m - H_n x_{m,n} \frac{h_{m,n}}{\sigma^2 + p_{m,n} h_{m,n}} + \lambda = 0$. Hence, we have $p_{m,n} = \frac{H_n x_{m,n}}{\lambda + \alpha V \mu_m x_{m,n} g_m} - \frac{\sigma^2}{h_{m,n}}$.

Combining these results, we can rewrite the optimal power assignment as

$$p^*_{m,n} = \left(\frac{H_n x_{m,n}}{\lambda + \beta_m} - \frac{\sigma^2}{h_{m,n}}\right)^+, \forall m, \forall n. \tag{52}$$

Appendix C
Proof of Theorem 1

Let $t = kT$, for some $k \in \{0, 1, \cdots\}$. Recall that the proposed JLCP algorithm is developed through minimizing the term on the right side of the following inequality

$$\begin{aligned}
&\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U(\tau)|\boldsymbol{H}(t)\right] \le B \\
&+ \sum_{\tau=t}^{t+T-1}\sum_{n\in\mathcal{N}} \mathbb{E}\left[H_n(t)(\omega_n(t) - r_n(\tau))|\boldsymbol{H}(t)\right] \\
&+ \sum_{\tau=t}^{t+T-1} \mathbb{E}\left[\alpha V C(\tau) - \sum_{n\in\mathcal{N}} V Q_n(\omega_n(t))|\boldsymbol{H}(t)\right].
\end{aligned} \tag{53}$$

Note that the value of the right term of (53) attained through JLCP is not more than that of any other feasible strategy, including the strategy satisfying the conditions specified in Assumption 1. Thus, we have

$$\begin{aligned}
&\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U_{JLCP}(\tau)|\boldsymbol{H}(t)\right] \le B \\
&- \delta T \sum_{n\in\mathcal{N}} H_n(t) - V \sum_{\tau=t}^{t+T-1} U_{SC}(\tau),
\end{aligned} \tag{54}$$

where $U_{JLCP}(\tau)$ and $U_{SC}(\tau)$ denote the utility achieved by the proposed JLCP and the policy satisfying the conditions specified in Assumption 1 , respectively. Rearranging the terms, and exploiting the fact that $U_{JLCP}(\tau) - U_{SC}(\tau) \le U_{\max} - U_{\min}$, we get that

$$\Delta_T(t) \le B + VT(U_{\max} - U_{\min}) - \delta T \sum_{n\in\mathcal{N}} H_n(t), \tag{55}$$

where $U_{\max}$ and $U_{\min}$ are the maximum and minimum utility[3], respectively.

By rearranging the terms and taking the expectation of (55), we obtain:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\sum_{n\in\mathcal{N}} H_n^2(t+T)] - \frac{1}{2}\mathbb{E}[\sum_{n\in\mathcal{N}} H_n^2(t)] + \delta T \mathbb{E}[\sum_{n\in\mathcal{N}} H_n(t)] \\
&\le B + VT(U_{\max} - U_{\min}).
\end{aligned}$$

The above inequality is satisfied for each slot $t = kT$. Summing it over the instants of $t = kT, k = 0, 1, \cdots, K-1$, and dividing both sides by δKT, followed by taking the operator of $\limsup$ as $K \to \infty$ yields:

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\sum_{n\in\mathcal{N}} H_n(kT)] \le \frac{B + VT(U_{\max} - U_{\min})}{\delta T}.$$

This proves part (a) of Theorem 1.

According to Theorem 4.5 in [19], we arrive at the conclusion that there is a stationary optimal ω-only strategy attaining the optimal utility $\overline{U^*}$ while meeting the queue constraint. Hence, for the optimal ω-only policy, the right-side second term of (53) is non-positive due to satisfying the queue constraint, and we have

$$\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} U_{JLCP}(\tau)\right] \le B - V \sum_{\tau=t}^{t+T-1} U^*(\tau),$$

where $U^*(\tau)$ denote the utility achieved by the optimal ω-only policy. Taking expectations of the above inequality and summing it over the instants of $t = kT, k = 0, 1, \cdots, K-1$, yields:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\sum_{n\in\mathcal{N}} H_n^2(KT)] - \frac{1}{2}\mathbb{E}[\sum_{n\in\mathcal{N}} H_n^2(0)] - \\
&V\mathbb{E}\left[\sum_{\tau=0}^{KT-1} U_{JLCP}(\tau)\right] \le KB - V \sum_{\tau=0}^{KT-1} U^*(\tau).
\end{aligned}$$

[3]Note that both $U_{\max}$ and $U_{\min}$ are finite due to boundedness of $q_n(t)$ and $\mu(t)$.

Dividing both sides by VKT, and taking the operator of $\limsup$ as $K \to \infty$, we have:

$$\lim_{K\to\infty} \frac{1}{KT}\mathbb{E}\left[\sum_{\tau=0}^{KT-1} U_{JLCP}(\tau)\right] \geq \lim_{K\to\infty} \frac{1}{KT}\sum_{\tau=0}^{KT-1} U^*(\tau) - \frac{B}{VT}.$$

This proves part (b) of Theorem 1.

REFERENCES

[1] Cisco White Paper, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," 2016.

[2] L. Hanzo, P. J. Cherriman, and J. Streit, *Video Compression and Communications*. IEEE Press-John Wiley & Sons, 2007.

[3] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[4] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA cloud-RAN of small cells overlaying a macrocell," *IEEE Trans. Mob. Comput.*, vol. PP, no. 99, p. 1, 2016.

[5] F. Pantisano, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Interference alignment for cooperative femtocell networks: A game-theoretic approach," *IEEE Trans. Mob. Comput.*, vol. 12, no. 11, pp. 2233–2246, Nov. 2013.

[6] J. H. Yun, "Intra and inter-cell resource management in full-duplex heterogeneous cellular networks," *IEEE Trans. Mob. Comput.*, vol. 15, no. 2, pp. 392–405, Feb. 2016.

[7] P. Semasinghe, E. Hossain, and K. Zhu, "An evolutionary game for distributed resource allocation in self-organizing small cells," *IEEE Trans. Mob. Comput.*, vol. 14, no. 2, pp. 274–287, Feb. 2015.

[8] W. C. Cheung, T. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 561–574, Apr. 2012.

[9] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A Stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538–549, April 2012.

[10] C. Jiang, Y. Chen, K. J. R. Liu, and Y. Ren, "Optimal pricing strategy for operators in cognitive femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5288–5301, Sept. 2014.

[11] C.-Y. Wang and H.-Y. Wei, "Profit maximization in femtocell service with contract design," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1978–1988, May 2013.

[12] V. Chandrasekhar, J. G. Andrews, T. Muharemovic, Z. Shen, and A. Gatherer, "Power control in two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[13] B. Ma, M. H. Cheung, V. W. S. Wong, and J. Huang, "Hybrid overlay/underlay cognitive femtocell networks: A game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3259–3270, Jun. 2015.

[14] Y. J. Yu, P. C. Hsiu, and A. C. Pang, "Energy-efficient video multicast in 4G wireless systems," *IEEE Trans. Mob. Comput.*, vol. 11, no. 10, pp. 1508–1522, Oct. 2012.

[15] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[16] Y. Huo, C. Hellge, T. Wiegand, and L. Hanzo, "A tutorial and review on inter-layer FEC coded video streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1166–1207, Apr. 2015.

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

[18] P. Si, Y. Shi, R. Zhu, J. Yang, and H. Xi, "Dynamic power and layer selection for scalable video streaming in femtocell networks," in *IEEE International Conference on Communication Workshop*, Jun. 2015.

[19] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[20] T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 293–311, Feb. 2013.

[21] Y. Kim, S. Lee, and D. Hong, "Performance analysis of two-tier femtocell networks with outage constraints," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2695–2700, Sept. 2010.

[22] A. R. Elsherif, W. P. Chen, A. Ito, and Z. Ding, "Adaptive resource allocation for interference management in small cell networks," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2107–2125, Jun. 2015.

[23] H. Y. Hsieh, S. E. Wei, and C. P. Chien, "Optimizing small cell deployment in arbitrary wireless networks with minimum service rate constraints," *IEEE Trans. Mob. Comput.*, vol. 13, no. 8, pp. 1801–1815, Aug. 2014.

[24] K. Zhu, E. Hossain, and D. Niyato, "Pricing, spectrum sharing, and service selection in two-tier small cell networks: A hierarchical dynamic game approach," *IEEE Trans. Mob. Comput.*, vol. 13, no. 8, pp. 1843–1856, Aug. 2014.

[25] H. Wang, J. Wang, and Z. Ding, "Distributed power control in a two-tier heterogeneous network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6509–6523, Dec. 2015.

[26] H. ElSawy and E. Hossain, "Two-tier hetnets with cognitive femtocells: Downlink performance modeling and analysis in a multichannel environment," *IEEE Trans. Mob. Comput.*, vol. 13, no. 3, pp. 649–663, Mar. 2014.

[27] L. Zhang, T. Jiang, and K. Luo, "Dynamic spectrum allocation for the downlink of OFDMA-based hybrid-access cognitive femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1772–1781, Mar. 2016.

[28] H. Wang and Z. Ding, "Macrocell-queue-stabilization-based power control of femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5223–5236, Sept. 2014.

[29] K. Miller, D. Bethanabhotla, G. Caire, and A. Wolisz, "A control-theoretic approach to adaptive video streaming in dense wireless networks," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1309–1322, Aug. 2015.

[30] D. Bethanabhotla, G. Caire, and M. J. Neely, "Adaptive video streaming for wireless networks with multiple users and helpers," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 268–285, Jan. 2015.

[31] A. Argyriou, D. Kosmanos, and L. Tassiulas, "Joint time-domain resource partitioning, rate allocation, and video quality adaptation in heterogeneous cellular networks," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 736–745, May 2015.

[32] Y. Xu, R. Q. Hu, Y. Qian, and T. Znati, "Video quality-based spectral and energy efficient mobile association in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 805–816, Feb. 2016.

[33] S. Thomas, H. Miska M., and T. Wiegand, "H. 264/AVC in wireless environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 657–673, Jul. 2003.

[34] "3GPP TS 36.201: Phyical Layer-general description," p. 8. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.201/

[35] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-Part I: theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[36] C. Chen, X. Zhu, G. de Veciana, A. Bovik, and R. Heath, "Rate adaptation and admission control for video transmission with subjective quality constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 22–36, Feb. 2015.

[37] "Video Trace Library." [Online]. Available: http://trace.eas.asu.edu/

[38] P. Si, J. Yang, S. Chen, and H. Xi, "Smoothness contraint based stochastic optimization for wireless scalable video streaming," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 759–762, May 2015.

[39] X. Wu, J. Yang, Y. Ran, and H. Xi, "Adaptive scalable video transmission strategy in energy harvesting communication system," vol. 17, no. 12, pp. 2345–2353, Dec. 2015.

[40] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Power cost reduction in distributed data centers: A two-time-scale approach for delay tolerant workloads," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 200–211, Jan. 2014.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[42] D. Palomar and J. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 686–695, Feb. 2005.

[43] X. Kang, H. Garg, Y.-C. Liang, and R. Zhang, "Optimal power allocation for OFDM-based cognitive radio with new primary transmission protection criteria," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 2066–2075, Jun. 2010.

[44] S. C. Chapra and R. P. Canale, *Numerical Methods for Engineers*. McGraw-Hill, 2012.

[45] M. Zhao, X. Gong, J. Liang, W. Wang, X. Que, and S. Cheng, "QoE-driven cross-layer optimization for wireless dynamic adaptive streaming of scalable videos over HTTP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 451–465, Mar. 2015.

[46] R. Gupta, A. Pulipaka, P. Seeling, L. J. Karam, and M. Reisslein, "H.264 coarse grain scalable (CGS) and medium grain scalable (MGS) encoded video: A trace based traffic and quality evaluation," *IEEE Trans. Broadcast.*, vol. 58, no. 3, pp. 428–439, Sept. 2012.

Jian Yang received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively. From 2006 to 2008, he was a postdoctoral scholar in the Department of Electronic Engineering and Information Science in USTC. Since 2008 he has been employed as an associate professor in the Department of Automation, USTC.

He is currently a professor in the School of Information Science and Technology, USTC. His research interests include future network, distributed system design, modeling & optimization, multimedia over wired & wireless and stochastic optimization. Dr. Yang received Lu Jia-Xi Young Talent Award from Chinese Academy of Sciences in 2009.

Peng Si received the B.S. and Ph.D degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2011 and 2016, respectively. He is currently an engineer in China Power Engineering Consulting Group Co. Ltd (CPECC). His current research interests include M2M communications, stochastic optimization and heterogeneous network.

Xiaofeng Jiang received the B.E. and Ph.D. in information science and technology from University of Science and Technology of China (USTC), Hefei, China, in 2008 and 2013. He is a post doctoral fellow at the department of automation of USTC from 2013. He is an associate research fellow at the department of automation of USTC from 2017. His recent research interests include discrete event dynamic system, future network, cognitive radio and cognitive radar.

Zilei Wang received the B.S. and Ph.D. degrees in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2002 and 2007, respectively.

He is currently an Associate Professor with the Department of Automation, USTC, and the Founding Leader of the Vision and Multimedia Research Group (http://vim.ustc.edu.cn). Before joining USTC as a faculty, he was a postdoc research fellow at National University of Singapore. His current research interests include computer vision, multimedia, and deep learning.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk) FREng, FIEEE, FIET, Fellow of EURASIP, DSc received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 111 PhD students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1700+ research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE VTS. During 2008 - 2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk Lajos has 30 000+ citations and an H-index of 70.