Enhancing QoE-Aware Wireless Edge Caching With Software-Defined Wireless Networks

Chengchao Liang, Ying He, F. Richard Yu, *Senior Member, IEEE*, and Nan Zhao, *Senior Member, IEEE*

***Abstract*— Software-defined networking and in-network caching are promising technologies in the next generation wireless networks. In this paper, we propose enhancing the quality of experience (QoE)-aware wireless edge caching with bandwidth provisioning in software-defined wireless networks (SDWNs). Specifically, we design a novel mechanism to jointly provide proactive caching, bandwidth provisioning, and adaptive video streaming. The caches are requested to retrieve data in advance dynamically according to the behaviors of users, the current traffic, and the resource status. Then, we formulate a novel optimization problem regarding the QoE-aware bandwidth provisioning in SDWNs with jointly considering in-network caching strategy. The caching problem is decoupled from the bandwidth provisioning problem by deploying the dual-decomposition method. Additionally, we relax the binary variables to real numbers so that those two problems are formulated as a linear problem and a convex problem, respectively, which can be solved efficiently. Simulation results are presented to show that the latency is decreased and the utilization of caches is improved in the proposed scheme.**

***Index Terms*— Software defined wireless networks, wireless edge caching, bandwidth provisioning, quality of experience, dual-decomposition.**

I. Introduction

ACCORDING to [1], global mobile data traffic will increase nearly sevenfold from 2016 to 2021 and 55 percent of them will be video. Moreover, another prominent feature of next generation wireless networks would be the full support of *software-defined networking* (SDN) [2] design in wireless networks [3]. These trends enforce the optimization of wireless networks to jointly consider the improvement of quality of experience (QoE) for video services and the integration of SDN. New networking technologies in wireless networks, such as *Heterogeneous Networks* (HetNets) [4], [5], *software-defined wireless networks* (SDWNs) [6] and *wireless edge caching* (WEC) [7], [8] that are arising to solve these changes, have been proposed and studied recently.

Generally speaking, SDWNs can enable the reduction of complexity and cost of networks, equip programmability into wireless networks, accelerate evolution of networks, and even further catalyze fundamental changes in the mobile ecosystem [3]. The success of the SDWN will depend critically on our ability to jointly provision the backhaul and radio access networks (RANs) [9].

WEC, as an extension of *in-network caching* that can efficiently reduce the duplicate content transmission [10], has shown that access delays, traffic loads, and network costs can be potentially reduced by caching contents in wireless networks [11]. To successfully combine caching and wireless networks, significant works (e.g., [12], [13]) have been done concerning utilizing and placing contents in the caches of base stations (BSs).

As the QoE of streaming video services mainly includes video resolutions, buffering delays and stalling events [14], [15], the SDWN (e.g., provision QoS [16]) and wireless edge caching (e.g., reduce delay [17]) appear as promising candidates to enhance QoE. However, to the best of our knowledge, QoE-aware joint optimization of the network and cache resources in SDWNs has been largely ignored in the existing research. Unfortunately, the combination of those issues are not straightforward, as several challenges are induced by this joint optimization observed as follows. First, bandwidth provisioning in SDWNs should be content-aware, which means it should assign network resources to users based on the caching status and the improvement of the QoE. Second, to enhance the hitting ratio of caches (utilization), caching strategies should be proactive according to the current traffic and resource status, behaviors of users, as well as the requirements of the QoE. Third, since video SDN flows from service providers usually have minimum requirements, the overall QoE performance of the network needs to be guaranteed.

Thus, to address those issues, in this paper, we propose to jointly optimize QoE of video streaming, bandwidth provisioning and caching strategies in SDWNs with limited network resources and QoE requirements. The distinctive technical features of this paper are listed as follows:

- To decrease the content delivery latency and improve the utilization of the network resources and caches, we design a novel mechanism to jointly provide proactive caching, bandwidth provisioning and adaptive video streaming. BSs are requested to retrieve data in advance dynamically

Manuscript received December 22, 2016; revised May 7, 2017; accepted July 14, 2017. Date of publication August 4, 2017; date of current version October 9, 2017. This work was supported in part by Huawei Technologies Canada Co., Ltd., in part by the Natural Sciences and Engineering Research Council of Canada, and in part by the National Natural Science Foundation of China under Grant 61372089. The associate editor coordinating the review of this paper and approving it for publication was Q. Li. *(Corresponding author: Chengchao Liang.)*

C. Liang, Y. He, and F. R. Yu are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: chengchaoliang@sce.carleton.ca; heying@sce.carleton.ca; richard.yu@carleton.ca).

N. Zhao is with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China (e-mail: zhaonan@dlut.edu.cn).

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/TWC.2017.2734081

according to the behaviors of users, the current traffic and the resource status.
- To cope with the limited resources and the quality of service requirements, we formulate a novel optimization problem regarding the QoE-aware bandwidth provisioning in SDWNs with jointly considering in-network caching strategy.
- The caching problem is decoupled from the bandwidth provisioning problem by deploying the dual-decomposition method. Additionally, we relax the binary variables to real numbers so that those two problems are formulated as a linear problem and a convex problem, respectively, which can be solved efficiently.
- Algorithms are proposed to achieve the sub-optimum solution by solving the relaxed problem and utilizing a rounding up method to recover relaxed variables to binary.

The rest of this paper is organized as follows. Section III introduces the system model and formulate the proposed problem. Section IV presents two proposed algorithms and the corresponding analysis. Simulation results are discussed in Section V. Finally, we conclude this study in Section VI.

II. Related Works

Bandwidth provisioning (flow control) of SDWNs is studied in [16] and [18] through traffic engineering. The authors of [18] propose a multi-path traffic engineering formulation for downlink transmission considering both backhaul and radio access constraints. Moreover, the link buffer status is used as feedback to assist the adjustment of flow allocation. Based on [18], the authors of [16] extend flow control of SDWNs to real-time video traffic specifically. This research proposes an online method to estimate the effective rate of video flows dynamically. Min flow rate maximization of SDWNs is investigated in [9] with jointly considering flow control and physical layer interference management problem using weighted-minimum mean square error algorithm.

In the line of caching at mobile networks, the authors of [19] formulate a delay minimization problem by optimally caching contents at the SBS. This research firstly clusters users with similar content and then deploys a reinforcement learning algorithm to optimize its caching strategy accordingly. Reference [20] proposes a scheme that BSs opportunistically employ cooperative multiple-input multiple-output (Coop-MIMO) transmission by caching a portion of files so that MIMO cooperation gain can be achieved without payload backhaul, while [13] and [21] propose to utilize caching for releasing part of fronthaul in C-RAN. Unfortunately, the research does not take the real-time mobile network and traffic status into account. In [22], the proposed cache allocation policy considers both the backhaul consumption of small BSs (SBSs) and local storage constraints. The authors of [23] introduce in-network caching into SDWNs to reduce the latency of backhaul, but cache strategies are ignored in this study. Reference [24] introduces dynamic caching into BS selection from the aspect of energy saving. In [25], dynamic caching is proposed to consider the mobility of users, but it focuses more on the cache resource and relationship between nodes. Network resources, such as the backhaul capacity and spectrum, are not discussed in this research.

Video quality adaptation with radio resource allocation in mobile networks (especially, long term evolution (LTE)) is studied in a number of studies. The authors of [26] provide a comprehensive survey on quality of experience of HTTP adaptive streaming. Joint optimization of video streaming and in-network caching in HetNets can be traced back to [17]. This research suggests that SBSs form a wireless distributed caching network that can efficiently transmit video files to users. Meanwhile, Ahlehagh and Dey [27] take the backhaul and the radio resource into account to realize video-aware caching strategies in RANs for the assurance of maximizing the number of concurrent video sessions. The authors of [28] move the attention of in-network video caching to the core networks, instead of RANs, of LTE. To utilize new technologies in next generation networks, [29] studies the quality of video in next generation networks. Reference [14] conducts a comprehensive research that investigates opportunities and challenges of combining the advantages of adaptive bit rate and RAN caching to increase the video capacity and QoE of wireless networks. Another research combining the caching and video service is [30], where collaborative caching is studied with jointly considering scalable video coding. However, RAN and overall video quality requirements are not considered.

III. System Model and Problem Formulation

In this section, we present the system model of the HetNet, dynamic caching, service QoE and related assumptions. The notations that will be used in the rest of this paper are summarized in Table I.

A. Network Model

1) Wireless Communication Model: In this paper, we consider the downlink transmission case in a software-defined HetNet comprised of a set $\mathcal{J}$ of cache-enabled BSs, such as macro BSs (MBSs) and small BSs (SBSs). The area covered by SBSs is overlapped with where covered by MBSs. A central SDN controller is deployed to control the network including caching strategies and bandwidth provisioning. BSs connect to the CN through wired backhaul links with fixed capacities (e.g., bps). A content server is physically located at a core router (content delivery networks) in the CN or the source server. Moreover, as shown in Fig. 1, some BSs (e.g., SBSs) connect to the CN through MBSs. Without loss of generality, each BSs may have multiple links to be connected in the network (e.g., SBS 3 in Fig. 1). We use link indicator a_{lj} to denote the link status between BS l and BS j. $a_{lj} = 1$ means BS l is the next hop of BS j with a fixed link capacity R_{lj}^{max} (bps); otherwise $a_{lj} = 0$. Since it is the downlink case, $a_{lj} = 1$ implies $a_{jl} = 0$. Specially, we use $j = 0$ to indicate the CN. Let $r_{lj}^{f} \in \mathbb{R}^{+}$ denote the provisioning bandwidth of BS j for its next hop BS l through wired backhaul links. f in the superscript is used to denote forwarding. Following constraints (backhaul limitation) need to hold

$$r_{lj}^{f} \leq R_{lj}^{max}, \quad \forall j,\ l \in \mathcal{J}. \tag{1}$$

TABLE I
NOTATIONS

Notation	definition
$\mathcal{J}$	the set of mobile BS nodes (e.g., MBSs, SBSs)
$J = \mid \mathcal{J} \mid$	number of BSs
j,l	BS, $j = 0$ means content cloud library
$a_{lj} \in \{0,1\}$	link status between BS j and BS l
R_{lj}^{max} (bps)	backhaul link capacity between BS j and BS l
$\mathcal{I}_j$, $\mathcal{I}$	the set of users
$\mathcal{I}_j^{\pi}$	the predicted (potential) set of users served by BS j
i	wireless user
g_{ij}	large scale channel gain between i and BS j
σ_0	noise power spectrum density
p_j (Watts / Hz)	normalized transmission power of BS j
w_{ij} (Hz)	assigned spectrum bandwidth of user i from j
W_j (Hz)	total spectrum bandwidth of j
γ_{ij}	received SINR of user i served by BS j
Γ_{ij}	spectrum efficiency of user u served by BS j
C	the total number of contents
C_j	the storage capacity of BS j
$h_{ij} \in \{0,1\}$	hitting event indicator between user i and BS j
$c_{ij} \in \{0,1\}$	dynamic caching indicator
$\tilde{c}_{ij} \in [0,1]$	relaxed dynamic caching indicator
δ_0 (seconds)	evaluation time (scheduling cycle time)
π_{ij}	predicted probability of user i served by BS j
Q	number of total video resolution levels
$q \in \{1,...,6\}$	video resolution level indicator
s_q	the MV-MOS of the resolution q
v_q	required transmission data rate of the resolution q
$x_{qi} \in \{0,1\}$	resolution indicator of user i
$\tilde{x}_{qi} \in [0,1]$	relaxed resolution indicator of user i
d_j	the average backhaul transmission delay of BS j
b_0 (seconds)	minimum buffer length to play video
b_{qi} (seconds)	current buffer length of resolution q at user u
S_0 (seconds)	the MV-MOS requirements from service providers
U	*objectives*, utilities
ι	consistence
r_{ij}^c (bps)	required data transmission rate of caching
r_{lj}^f (bps)	provisioned bandwidth for backhaul links
r_{ij}^a (bps)	provisioned bandwidth for air interfaces
λ, μ, ν	dual variables
G, L_j	dual functions
Ω_{ij}	caching margin gain
N_j	number of cachable contents

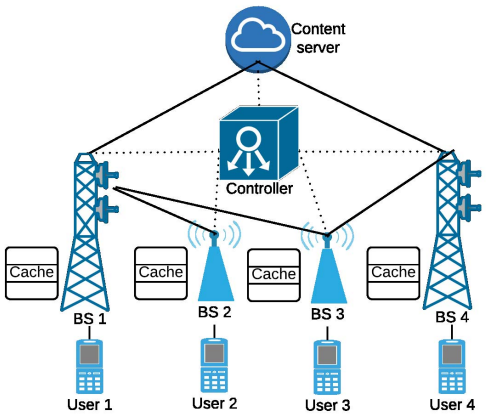


Fig. 1. Network architecture of the cache-enabled SDWN.

Let I_j denote the set of users served by BS j and each user i requests a video flow with certain data rate requirements. The set $I := I_1 \cup \ldots \cup I_j$ of users means the total users set. To enforce single association, we let $I_1 \cap I_j = \varnothing$ and any user is served by the BS that provides the best spectrum efficiency (SE). Association schemes are well-studied in heterogeneous wireless networks [31], [32]. In this paper, to simplify our analysis, we do not consider any advanced interference management and power allocation schemes. We assume that the spectrum used by different users within one BS is orthogonal, which means there is no intra-cell interference between users associated with the same BS. This is similar to a system of frequency-division multiple access, such as LTE. The spectrum reuse factor for all cells is one (overlaid), which means inter-cell downlink interference is considered. A fixed equal power allocation mechanism is used, where the normalized transmit power on BS j is p_j Watts/Hz regardless of positions and allocated spectral of users.

If $i \in I_j$, the signal-to-interference plus noise ratio (SINR) γ_{ij} can be determined by

$$\gamma_{ij} = \frac{g_{ij} p_j}{\sigma_0 + \sum_{l,l \neq j} g_{il} p_l} \tag{2}$$

where g_{ij} is the channel gain between user i and BS j including large-scale pathloss and shadowing, σ_0 is the power spectrum density of additive white Gaussian noise and $\sum_{l,l \neq j} g_{il} p_l$ is the aggregated received interference. As the small-scale fading varies much faster than caching and bandwidth provisioning, small-scale fading is not considered when evaluating the SINR. Therefore, the SINR calculated by (2) can be considered as an average SINR. Moreover, in this paper, as we mainly consider the benefits from dynamic caching and flow-level scheduling, radio resource allocation is not considered, which leads to the formulation where small-cale fading is ignored. However, small-scale fading can have effects on the SINR if the caching process is comparably fast. Thus, if small-cale fading is considered, the network link capacity (e.g., γ_{ij}) can be modeled as random variables, which leads a stochastic optimization problem. The scheme in [33] can provide a suitable solution to this problem.

Accordingly, by using Shannon bound to get SE $\Gamma_{ij} = \log\left(1 + \gamma_{ij}\right)$, the provisioning (achievable) radio access data rate $r_{ij}^a \in \mathbb{R}^+$ for user i with BS j can be calculated as:

$$r_{ij}^a = w_{ij} \log\left(1 + \gamma_{ij}\right) \tag{3}$$

where w_{ij} (Hz) is available spectrum bandwidth of BS j allocated to user i. a in the superscript is used to denote air interface. Since the spectrum bandwidth of each BS j is limited by W_j, the following constraints have to hold

$$\sum_{i \in I_j} \frac{r_{ij}^a}{\Gamma_{ij}} \leq W_j, \quad \forall j \in \mathcal{J}. \tag{4}$$

2) Proactive Wireless Edge Caching Model: To improve the caching performance, BSs can proactive cache contents that may be requested users. The probability that a cached content in BS j will be used by user i is assumed to be π_{ij}. In practice, π_{ij} depends on the user mobility pattern [34] (current location, moving velocity and direction, and unpredictable factors) and the popularity of the content [25]. The research about the behaviors of users has attracted great interests from both academia and industry [35]. Necessary

information for processing prediction can come from real-time data (e.g., UEs and RANs) and historical data (e.g., users database). Fortunately, the improvement of machine learning and big data technologies can help the management and the control of mobile networks [36], [37]. By using those advanced techniques, the network is able to have a deeper view of the traffic and the coverage from the huge volume of historical data, which may help the network to enhance the accuracy of the prediction. It should be noted that the calculation of this probability is beyond the research of this paper and we assume that π_{ij} is known in advanced by the SDWN controller (before each scheduling circle starts). To evaluate the effect of π_{ij}, we test our performance with two different π_{ij} setups in the simulation. Let us use I_j^{π} to denote the potential set of users that may be probably served by BS j, which can be formally defined as $I_j^{\pi} := \{i \mid \pi_{ij} \neq 0, i \in I\}$.

In this proposed scheme, we assume the total number of content is C stored at the content server (cloud centers or the Internet source) and each content has the normalized size of 1. This assumption is reasonable because we can slice the video content into chunks with the same length. Our BSs are assumed to be cache-enabled, so there is a physical cache at each BS j with capacity C_j. As any BS only has limited storage capability, C_j is much smaller than cloud center ($C_j \ll C, \forall j$) and the cache placement decision is made by the SDN controller. We assume that the SDN controller knows all the information about content retrieving requests from users and cache placement status of all BSs. Let us use a binary parameter h_{ij} to denote a hitting event between user i and BS j. $h_{ij} = 1$ means the content requested by user i can be found at BS j and $h_{ij} = 0$ means opposite. According to the control signaling by the SDN controller, contents are pulled actively by BSs through wired backhaul links, an example shown in Section IV-E.

Denote $c_{ij} \in \{0, 1\}$ to be the binary decision variable used to control whether the content i (potentially requested by user i) is going to be placed at BS j or not. The bandwidth required to cache the content requested by i is pre-defined as r_i^c (bps) that is fixed. Hence, the provisioned bandwidth for caching in BS j is $\sum_{i \in I_j^{\pi}} c_{ij} r_i^c$. Note duplicated caching is avoided as the SDN controller knows the cached contents of every BS. In other words, if the content has been cached in BS j, it is unnecessary to cache again, namely $c_{ij} = 0$ if $h_{ij} = 1$. Since each BS j has limited cache space, constraints should hold to ensure caching strategy is limited in the empty space of the cache of each access BS j.

$$\sum_{i \in I_j^{\pi}} c_{ij} \leq C_j, \quad \forall j \in \mathcal{J}. \tag{5}$$

3) Video QoE Model: In this paper, to specify the network performance for video services, we use a novel video experience evaluation method called mobile video mean-opinion-score (MV-MOS) proposed in [38] to model the network utility. MV-MOS is a more advanced measurement of the video quality based on the well-known video mean opinion score (vMOS), and it can adapt to the evolution of video resolutions. We assume the q-th resolution of any video requires a data rate v_q (bps) and gains a MV-MOS s_q (from 1 to 6). For example, [38] points out that 4k video average requires 25Mbps data rate and can be quantified to a MV-MOS 4.9 out of 5. Practically, v_q depends on video coding schemes and the video content, which are varying with time. Nevertheless, since the purpose of our paper is to maximize the wireless network performance dynamically instead of video services, we assume the required data rate is a fixed value v_q over research time for all streams.

We define x_{iq} as the resolution indicator of user i and Q is the highest resolution level. Specifically, if the q-th resolution of the video is selected by user i, $x_{iq} = 1$; otherwise, $x_{iq} = 0$. Thus, for any user i, the experienced MV-MOS is $\sum_q x_{iq} s_q$ and the required data rate is $\sum_q x_{iq} v_q$. Since users only can select one level of resolution at the same time, following constraints should hold:

$$\sum_{q=1}^{Q} x_{iq} = 1, \quad \forall i \in I \tag{6}$$

Moreover, to guarantee the video service QoE of the overall network, we should guarantee the overall average MV-MOS higher an acceptable value S_0, namely,

$$\frac{1}{I} \sum_{i \in I} \sum_{q=1}^{Q} s_q x_{iq} \geq S_0. \tag{7}$$

Usually, the S_0 is provided by service providers or network operators to control the overall performance of a certain area.

In [14] and [15], the authors point out that the stalling (video interruption) degrades users experience more severe even compared to initial latency because it disrupts the smoothness of streaming videos. It happens when the buffer b_{iq} (bits) is exhausted, lower than a threshold b_0. In order to avoid this "annoy" factor, in this paper, we request the wireless network to maintain the buffer higher than a threshold b_0 (bits) at user devices buffer, so that the user can get smooth streaming video experience. This proactive buffer management that sends feedback the buffer status of users to the network has been proposed in the existing studies [16], [39], [40]. To maintain b_0 of video, for any user i, it should follow constraints

$$\delta \sum_{j \in \mathcal{J}} r_{ij}^a + \sum_{q=1}^{Q} x_{iq} b_{iq} - \delta \sum_{q=1}^{Q} x_{iq} v_q \geq b_0, \quad \forall i \in I \tag{8}$$

where δ is a predefined time window unit in second. Every δ seconds, the system adaptively selects the resolution of the demanded video based on buffer status b_{iq} and available data rate $\sum_{j \in \mathcal{J}} r_{ij}^a$. (8) means that the amount of downloaded data and buffered data must support to playback average δ (s) and maintain at least b_0 buffer.

B. Problem Formulation

In this subsection, an optimization problem maximizing the gains from wireless edge caching constrained by physical resource and service QoE is proposed.

The purpose of this study is to improve the overall wireless edge caching performance by considering caching strategies,

network status and QoE provisioning, so it has to choose an appropriate network utility function U. Firstly, the caching utilization should be represented by different metrics (e.g., alleviated backhaul, delay reduction, and reduced network costs). Secondly, as this study is QoE-aware where the delay is one of the most important metrics, we chose delay reduction[1] d_j (seconds) as the gain of caching at the wireless edge [17]. Thus, shown as following, we define a logarithm-based objective function as the utility function, which is equivalent to the proportional fairness scheme in a long term view [41]:

$$U(\mathbf{c}) = \sum_{j \in \mathcal{J}} \log \left(\sum_{i \in I_j^{\pi}} \pi_{ij} \bar{h}_{ij} c_{ij} d_j + \iota \right) \tag{9}$$

where $\bar{h}_{ij} = 1 - h_{ij}$ and $\iota \geq 1$ is consistences to guarantee $U(\mathbf{c})$ does not fall into negative infinity.[2] The product of $\bar{h}_{ij}$ and c_{ij} enforces c_{ij} to be different from h_{ij} if $h_{ij} = 1$, which is equivalent to the case that duplicated caching is avoided. The probability π_{ij} can be interpreted as a weighted factor that represents the success of caching. Therefore, $\pi_{ij}\bar{h}_{ij}d_j$ can be considered as the expected reduced backhaul latency, if $c_{ij} = 1$. To lighten the notations, we let $H_{ij} = \pi_{ij}\bar{h}_{ij}d_j$. We select a logarithm-based objective function due to following features that are perceived [41]:

- *Component-wise monotonic increase* brings that larger caching gain yields larger utility;
- *Convexity* can guarantee the convergence and efficiency of the algorithm;
- *Fairness-awareness* gives a desired balance between the overall network and the single user.

To utilize the physical resource efficiently, the total provisioned bandwidth of flows going out BS j should be less than the total bandwidth of flows coming into the corresponding BS. Otherwise, allocated spectrum or backhaul links will be left unused. The flow conservation constraint for BS j can be written as

$$\underbrace{\overbrace{\sum_{l \in \mathcal{J}} a_{lj} r_{lj}^{f}}^{\text{forwarding}} + \overbrace{\sum_{i \in I_j} \bar{h}_{ij} r_{ij}^{a}}^{\text{radio}} + \overbrace{\sum_{i \in I_j^{\pi}} c_{ij} r_i^{c}}^{\text{caching}}}_{\text{out flows}} \leq \underbrace{\sum_{l \in \mathcal{J} \cup 0} a_{jl} r_{jl}^{f}}_{\text{in flows}}, \quad \forall j \in \mathcal{J}. \tag{10}$$

The first term of left (10) is all reserved bandwidth for flows forwarded to next hop BSs. The second term is the summation of all provisioned bandwidth for bearing users' radio access flows. Obviously, if $\bar{h}_{ij} = 0$ that means we can find existing data (cached in previously caching circles) on BS j for user i so that the backhaul consumption is avoided. The third term is all caching flows. Obviously, the provisioned (reserved) bandwidth of these three kinds of flows cannot be larger than the reserved backhaul bandwidth for this BSs. If $\bar{h}_{ij} = 0$, it is reserved for two kinds of flows (caching and forwarding). The right term of (10) can be as a dynamic backhaul link capacity of BS j.

Thus, given the objective function, and constraints, we can define the joint flow bandwidth provisioning and wireless edge cache placement problem **P0** as follows:

$$\mathbf{P0}: \max_{\substack{\mathbf{r}^a, \mathbf{r}^f \in \mathbb{R}^+ \\ \mathbf{c}, \mathbf{x} \in \{0,1\}}} U \tag{11a}$$

$$s.t.\ (1), (4), (5), (6), (7), (8), (10) \tag{11b}$$

Unfortunately, problem (11), however, is difficult to be solved and implemented based on the following observations:

- The mix integer variables result in the problem a mix-integer non-linear problem (MINLP) that generally is NP-hard and intractable in scheduling [42];
- The complexity of solving (11) by using greedy or genetic methods will increase significantly with the increase of the number of users and (or) BSs. In future cellular networks, the density and number of small cells will rise significantly so that the size of variables will become very large;
- Cache indicators and flows bandwidth are decided by different layers and perform in different time scales.

IV. Joint Bandwidth Provisioning and Wireless Edge Caching With QoE

In this section, an algorithm is proposed to solve the problem **P0** in (11). As those integer variables block us to find a traceable algorithm, we firstly relax binary variables c_{ij} and x_{iq} to real numbers variables $\tilde{c}_{ij}$ and $\tilde{x}_{iq}$ bounded by $[0, 1]$ so that the domain of variables is a convex set. This relaxation is a common way to deal with binary variables in wireless networks [30], [32], [41], [43]. From a long term view, we can interpret this kind of relaxations as a partial allocation or average time sharing portion. For example, a partial caching indicator $\tilde{c}_{ij}$ or resolution $\tilde{x}_{iq}$ means the portion of time when BS j indicate cache to user i or when users i selects the resolution level q, respectively. Moreover, by solving the relaxing binary variables, the upper bound of the proposed can be achieved. We evaluate the gap between this upper bound and the final solution in our simulation.

After relaxing binary variables, it is clear that the feasible set of the revised problem (11) is a convex set because all constraints are linear constraints and variables domain is a convex set. Since the log-based objective function is a strict concave function regarding to $\tilde{c}_{ij}$, the problem (11) is transfered to a convex problem $\tilde{\mathbf{P0}}$ that can be solved effectively without effort by using general solvers. In the remaining subsections, we will present an algorithm based on dual-decomposition to solve the relaxed problem $\tilde{\mathbf{P0}}$ and recover relaxed variables back to binaries.

A. Proposed Caching Decoupling via Dual Decomposition

This study aims to adaptively replace content in caches of BSs while optimize the bandwidth provisioning and the video resolution selection. Observed from the problem **P0**, we can consider that we use the spare backhaul bandwidth to cache

[1] We use the average backhaul downlink latency as the measurement value.

[2] When either $I_j^{\pi} = \varnothing$ or $\pi_{ij}\bar{h}_{ij} = 0, \forall i \in I_j^{\pi}$, the objective function may result in an infeasible problem.

content to BSs before users actually served (or potentially continued being served) with those BSs (or actually request the content) and select a best video resolution for each user based on allocated achievable data rate. Moreover, caches and bandwidth are actually belonging to different layers of the network. Usually, flow bandwidth is restricted by physical resources (provisioned by media access control layer), and x_{iq} depends on the achievable data rate. Joint optimizations of video rate and flow bandwidth are studied in cross-layer design schemes [44]. Differently, in-network cache is placed at the network layer or even higher layer. In addition, scheduling of the caching can be acting in a longer time period compared to flow scheduling. Fortunately, constraints coupling cache and bandwidth are only (10).

Thus, those features of **P0** motivate us to adopt dual-decomposition method so that the caching problem can be separated from the bandwidth provisioning and the resolution selection. We form the partial Lagrangian function for the problem $\tilde{\mathbf{P0}}$ by introducing the dual variables (backhaul prices) $\{\lambda_j\}$ for constraints (10). Then, the partial Lagrangian function can be shown as:

$$G(\lambda)=\sum_{j\in\mathcal{J}}\log\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)-\sum_{j\in\mathcal{J}}\lambda_j\sum_{i\in I_j^{\pi}}\tilde{c}_{ij}r_i^c$$
$$-\sum_{j\in\mathcal{J}}\lambda_j\left(\sum_{l\in\mathcal{J}}a_{lj}r_{lj}^f+\sum_{i\in I_j}\bar{h}_{ij}r_{ij}^a-\sum_{l\in\mathcal{J}\cup 0}a_{jl}r_{jl}^f\right) \quad (12)$$

The dual problem (DP) is thus:

$$\mathbf{DP}:\min_{\lambda\in\mathbb{R}^+}\ G(\lambda)=f_c(\lambda)+f_{r,x}(\lambda) \quad (13)$$

where

$$f_c(\lambda)=\arg\max_{\mathbf{c}\in[0,1]}\left\{\begin{array}{c}\sum_{j\in\mathcal{J}}\log\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)\\ -\sum_{j\in\mathcal{J}}\lambda_j\sum_{i\in I_j^{\pi}}\tilde{c}_{ij}r_i^c,\\ \text{s.t.}\quad (5)\end{array}\right\} \quad (14)$$

and

$$f_{r,x}(\lambda)=\arg\max_{\substack{\mathbf{r}^a,\mathbf{r}^f\in\mathbb{R}^+\\ \mathbf{x}\in[0,1]}}\left\{\begin{array}{c}\sum_{l\in\mathcal{J}\cup 0}\sum_{j\in\mathcal{J}}\lambda_j a_{jl}r_{jl}^f-\\ \sum_{j\in\mathcal{J}}\lambda_j\left(\sum_{l\in\mathcal{J}}a_{lj}r_{lj}^f+\sum_{i\in I_j}\bar{h}_{ij}r_{ij}^a\right),\\ \text{s.t.}\quad (1),(4),(6),(7),(8)\end{array}\right\} \quad (15)$$

Let us introduce a parameter z_n called extra bandwidth for each BS j shown as:

$$z_j=\sum_{l\in\mathcal{J}\cup 0}a_{jl}r_{jl}^f-\sum_{i\in I_j^{\pi}}\tilde{c}_{ij}r_i^c-\sum_{l\in\mathcal{J}}a_{lj}r_{lj}^f-\sum_{i\in I_j}\bar{h}_{ij}r_{ij}^a. \quad (16)$$

According to dual decomposition [45], a subgradient of $G(\lambda)$ is $z=[z_1,\ldots,z_j]^T$, we thus can update λ based on:

$$\lambda^{[k+1]}=\lambda^{[k]}-\alpha_{\lambda}^{[k]}z^{[k]}, \quad (17)$$

where $\alpha_{\lambda}^{[k]}$ is the step length at the iteration step $[k]$.

B. Upper Bound Approach to Solving (14)

In this part, a low complexity algorithm is proposed to solve the problem (14) so that the upper bound can be achieved. Firstly, to make our expression more compact, we define $y_{ij}=\lambda_j r_i^c$ that can be interpreted as the backhaul bandwidth cost of caching. Observe that $f_c(\lambda)$ can be further decoupled to each BS j, thus we focus on solving (14) for one BS. Let $\mu_j\geq 0,\forall j\in\mathcal{J}$ and $\nu_{ij}\geq 0,\forall j\in\mathcal{J},\forall i\in I_j^{\pi}$ be the dual variables associated with constraints (5) and $\tilde{c}_{ij}\leq 1$ of the problem given in (14), respectively. Then, the Lagrangian of (14) can be expressed as

$$\begin{aligned}L_j(\tilde{\mathbf{c}},\mu,\nu)&=\log\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)-\sum_{i\in I_j^{\pi}}\nu_{ij}\left(\tilde{c}_{ij}-1\right)\\&\quad-\sum_{i\in I_j^{\pi}}y_{ij}\tilde{c}_{ij}-\mu_j\left(\sum_{i\in I_j^{\pi}}\tilde{c}_{ij}-S_j\right)\\&=\log\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)\\&\quad-\sum_{i\in I_j^{\pi}}\left(y_{ij}+\mu_j+\nu_{ij}\right)\tilde{c}_{ij}+\mu_j S_j+\sum_{i\in I_j^{\pi}}\nu_{ij}\end{aligned} \quad (18)$$

It is noted that $L_j(\tilde{\mathbf{c}},\mu,\nu)$ is a continuous and differentiable function of $\tilde{c}_{ij}$, μ_j, and ν_{ij}. By differentiating $L_j(\tilde{\mathbf{c}},\mu,\nu)$ with respective to $\tilde{c}_{ij}$, we can have the Karush-Kuhn-Tucker (KKT) conditions as

$$\frac{\partial L_j}{\partial\tilde{c}_{ij}}=\frac{H_{ij}}{\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)}-\left(y_{ij}+\mu_j+\nu_{ij}\right)\leq 0,\quad \forall i\in I_j^{\pi} \quad (19)$$

$$\tilde{c}_{ij}\left[\frac{H_{ij}}{\left(\sum_{i\in I_j^{\pi}}H_{ij}\tilde{c}_{ij}+\iota\right)}-\left(y_{ij}+\mu_j+\nu_{ij}\right)\right]=0,\quad \forall i\in I_j^{\pi} \quad (20)$$

$$\mu_j\left(\sum_{i\in I_j^{\pi}}\tilde{c}_{ij}-S_j\right)=0,\quad \forall i\in I_j^{\pi} \quad (21)$$

$$\nu_{ij}\left(\tilde{c}_{ij}-1\right)=0,\quad \forall i\in I_j^{\pi} \quad (22)$$

From (19) and (20), we obtain an optimal cache allocation for fixed Lagrange multipliers as

$$\tilde{c}_{ij}=\left[\frac{1}{\left(y_{ij}+\mu_j+\nu_{ij}\right)}-\frac{\iota+\sum_{i'\neq i}H_{i'j}\tilde{c}_{i'j}}{H_{ij}}\right]^+, \quad (23)$$

Algorithm 1 Upper Bound Algorithm of the Wireless Edge Caching

Input: Backhaul price $\{\lambda_j\}$ and caching gain $\{H_{ij}\}$
Output: Relaxed cache placement strategy $\{\tilde{c}_{ij}\}$

begin Sovling problem (14)
  Set prescribed accuracy ζ and maximum number of iteration steps T;
  $\mu_j \leftarrow \mu_j^{[0]}, \forall j;\ \ \nu_{ij} \leftarrow \nu_{ij}^{[0]}, \forall i, j;$ // set initial dual variables
  $t \leftarrow 1$;// interation step indicator
  while ζ *is not reached* && $t \leq T$ **do**
    Update $\tilde{c}_{ij}^{[t+1]}$ according to (23);
    Update $\mu_j^{[t+1]}$ according to according to (24);
    Update $\nu_{ij}^{[t+1]}$ according to according to (25);
    $t \leftarrow t + 1$;
  end
end

where $[x]^+ = \max\{x, 0\}$. As $H_{ij} = 0$ leads $\left[\frac{1}{(y_{ij}+\mu_j+\nu_{ij})} - \frac{\iota+\sum_{i'\neq i} H_{i'j}\tilde{c}_{i'j}}{H_{ij}}\right] \rightarrow -\infty$, $\tilde{c}_{ij} = 0$. In other words, since $H_{ij} = 0$, any positive values of $\tilde{c}_{ij}$ do not increase the utility but only waste physical resource instead. To obtain (23), using a gradient-based search, the updated μ_j value is given by

$$\mu_j^{[t+1]} = \left[\mu_j^{[t]} - \alpha_\mu^{[t]}\left(\sum_{i\in I_j^\pi} \tilde{c}_{ij} - S_j\right)\right]^+, \tag{24}$$

where $[t]$ is the iteration index and $\alpha_\mu^{[t]}$ are sufficiently small step sizes. Similar to ν_{ij},

$$\nu_{ij}^{[t+1]} = \left[\nu_{ij}^{[t]} - \alpha_\nu^{[t]}\left(\tilde{c}_{ij} - 1\right)\right]^+, \tag{25}$$

where $\alpha_\nu^{[t]}$ are sufficiently small step sizes. We summarize the procedure used to get the optimal solution of the problem (14) in Alg. 1.

The problem (15) obviously is a linear problem that can be solved effortlessly by general methods (e.g. interior point method) that are polynomial time algorithms in the worst case. Moreover, since the problem (15) is similar to the problem (14), even simpler because of the linear objective function, we can use the same idea to solve the problem (15). Limited by the space, we do not give detailed analysis on it.

C. Rounding Methods Based on Marginal Benefits

Recall that we have relaxed the cache placement indicators c_{ij} and video resolution x_{iq} to real values between zero and one instead of binary variables. Thus, we have to recover them to binary values after we get the relaxed solution. The basic idea of the rounding method is that we select the 'best' users to utilize the caching resource and the 'highest' resolution for users under current resource allocation solution. In this paper, c_{ij} is recovered to binary based on the corresponding marginal benefit [32], [46].

Firstly, we calculate the marginal benefits of each user as

$$\Omega_{ij}^c = \partial U/\partial\tilde{c}_{ij} = \frac{H_{ij}}{\left(\sum_{i\in I_j^\pi} H_{ij}\tilde{c}_{ij} + \iota\right)}. \tag{26}$$

Then, assuming $\tilde{c}_{ij} \in [0, 1]$ is the achieved optimum solution, we can obviously calculate the available number of cache that can be updated at BS j as

$$N_j = \lfloor \sum_{i\in I_j^\pi} \tilde{c}_{ij} \rfloor. \tag{27}$$

where $\lfloor x \rfloor$ means taking the maximum integer value that is less than x. In other words, N_j means the maximum number of content segments that can be cached at e BS j during the scheduling period. N_j users can be selected from I_j^π, whose Ω_{ij}^c are larger than other users. Formally, we can separate all users of BS j to two sets $I_j^{[\pi,+]}$ containing N_j = elements and $I_j^{[\pi,-]}$ containing the remaining users. $I_j^{[\pi,+]}$ and $I_j^{[\pi,-]}$ have following relationship:

$$\max_{i\in I_j^{[\pi,-]}} \Omega_{ij}^c \leq \min_{i\in I_j^{[\pi,+]}} \Omega_{ij}^c \tag{28}$$

Then, the caching indicator c_{ij} can be recovered by

$$c_{ij} = \begin{cases} 1 & \text{if } i \in I_j^{[\pi,+]}, \\ 0 & \text{otherwise,} \end{cases} \tag{29}$$

The recovering of x_{iq} is easier due to constraints (6). By assuming r_{ij}^a is the achieved optimum solution, the method uses following rule:

$$x_{iq} = \begin{cases} 1, & \text{if } q = \arg\max_q\ s_q^v, \text{s.t. } v_q \leq \sum_{i\in\mathcal{I}} r_{ij}^a \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

(30) is used to find the highest resolution whose required data rate is lower than the provided data rate.

As the two relaxed subproblems have been solved and variables are rounded up to binaries, the solution of the original problem **P0** is obtained. A complete description of the overall proposed scheme is stated in Algorithm 2.

D. Computational Complexity, Convergence and Optimality

As we mentioned above, the problems formulated in (11) is nonlinear integer optimization problems and is NP-hard without computational efficient algorithms to obtain the optimal solution [42]. Exhaustive searching method can be used to find a solution. However, for a dynamic caching SDWN system with I users and J BSs, the computational complexity of exhaustive searching method is about $O((J+1)^I)$ even we ignore the calculation of spectrum and backhaul allocation, which is tremendously high and unacceptable for practical implementation. BnB method or dynamic programming can be used to solve this problem, but they are computationally intensive and might not be practical for large-scale problems.

Alg. 1 is a polynomial time algorithm with the computational complexity of $O(IJ)$. The computational complexity

Algorithm 2 The Proposed Joint Allocation Algorithm

Input: Wireless network status ($\{a_{lj}\}$ $\{R_{lj}^{max}\}$ $\{\Gamma_{ij}\}$), caching placement ($\{h_{ij}\}$ $\{\pi_{ij}\}$ $\{C_j\}$), and QoE parameters ($\{v_q\}$ S_0 $\{b_{iq}\}$)

Output: $\{r_{lj}^f\}$ $\{r_{ij}^a\}$ $\{c_{ij}\}$ $\{x_{iq}\}$.

begin Solve the problem **P0**
  Set prescribed accuracy ϵ and maximum number of iteration steps K;
  $\lambda_j \leftarrow \lambda_j^{[0]}, \forall j$; `// set initial bandwidth prices`
  $k \leftarrow 1$; `// interation step indicator`
  while ϵ *is not reached* && $k \leq K$ **do**
    begin bandwidth provisioning and video resolution selection
      Update r_{lj}^f, r_{ij}^a and $\{\tilde{x}_{iq}\}$ by solving the problem (15);
      Rounding $\{\tilde{x}_{iq}\}$ up to $\{x_{iq}\}$ according to (30);
    end
    begin cache placement
      Update $\tilde{c}_{ij}$ by solving the problem (14);
      Calculate the marginal benefits and available cache according to (26) and (27);
      Form $I_j^{[\pi,-]}$ and $I_j^{[\pi,+]}$ based on selecting users according to (28);
      Rounding $\{\tilde{c}_{ij}\}$ up to $\{c_{ij}\}$ according to (29);
    end
    Update $\lambda^{[k+1]}$ according to (17);
    $k \leftarrow k + 1$;
  end
end

of solving the problem (15) is $O(\max\{I, J\}J)$ leading it as a polynomial time algorithm as well. The computational complexity of the proposed dual-decomposition Alg. 2 is $O(\max\{I, J\}J)$. The detailed analysis of computational complexity can be found in Appendix. Obviously, our proposed scheme reduce the computational complexity significantly, compared to exhaustive searching method.

Since either the inner loop in Alg. 1 or the outer loop in the proposed algorithm 2 is used to solve a convex problem which is proved to converge to the exact marginal. However, the precise conditions and the initial bandwidth prices used in dual-decomposition are not able to be decided before a practical implementation. The optimality cannot be guaranteed since we are solving an NP-hard problem. However, as we can get the upper bound of the proposed problem, we can use it as a replacement for the optimal solution to evaluate our proposed algorithm empirically. Thus, simulation results are used to test the convergence and the optimality of the proposed algorithm Fig. 3 in Section V.

E. Implementation Design in SDWNs

To illustrate the utilization of Alg. 2 into SDWNs, we give an instance presented by a diagram in Fig. 2.

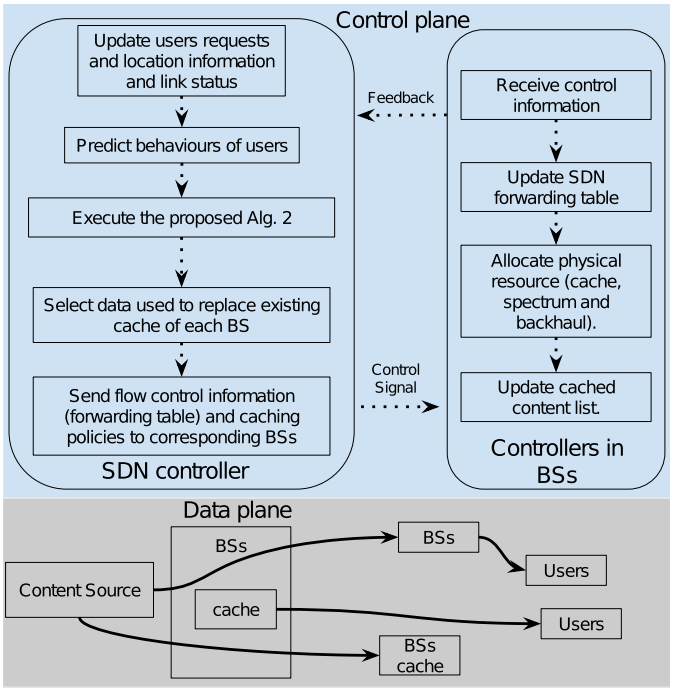


Fig. 2. The flow diagram in the proposed cache-enabled flow control in SDWNs.

Firstly, as shown in Fig. 2, in the control plane, the SDN controller updates the network status according to feedbacks from BSs, then predict the users' behaviors to calculate $\{\pi_{ij}\}$. By using this information, the SDN controller calls our proposed Alg. 2 to calculate the provisioned bandwidth and caching strategy for each flow i and BS j. After this, the SDN controller sends the control information to selected BSs and content source (the cloud center or the public networks). According to the control information, BSs and other network elements (e.g., potential routers) will update their SDN forwarding tables and perform resource allocation (e.g., cell association and scheduling). In the data plane, flows carrying contents requested by users can pass networks elements selected by the control plane and to users.

V. Simulation Results and Discussions

Simulation results are presented in this section to demonstrate the performance of the proposed scheme. The simulator is a Matlab-based system level simulator. Monto-Carlo method is used in the simulations. We run the simulations in an X86 desktop computer with quad core CPU (Intel Q8400), 4GB RAM, and the OS of Microsoft Windows 7 SP1. The positions of MBSs and SBSs are fixed. We randomly deploy users in the covered area in each simulation cycle. Average values are taken to reduce the randomness effects in the simulations.

The number of available videos is 100,000, with popularity following a Zipf distribution with exponent 0.56, following [17]. In the simulation, we consider a cellular network including 57 BSs that cover a 400m-by-400m. 9 BSs are core BSs (macro cells) that connect to the core network and the SDN controller directly. The remaining 48 BSs (small cells)

TABLE II
SIMULATION PARAMETERS

Network parameters	value
geographic area	400m-by-400m
number of BSs	9 macro cells, 48 small cells
number of users	$30-120$,80*
frequency bandwidth (MHz)	20
frequency reuse factor	1
transmission power profile	SISO with maximum power; 49dBm (macro), 20dBm (small)
propagation profile [43]	pathloss:L(distance)=34+40log(distance); lognormal shadowing: 8dB; no fast fading
power density of the noise	-174 dBm/Hz
spectrum efficiency calculation	Shannon bound
backhaul capacity (Mbps)	macro to CN:500; small to macro: $25-160$, 40*
average trasmission latency (ms)	RAN: 50; backhaul: 60
prediction probability range	low:0-0.5; high:0.5-0.99
total content	10^5
content popularity profile	Zipf distribution with exponent 0.56 [17]
cache size at BSs	$300-1200$, 600*

TABLE III
VMOS OF VIDEO RESOLUTIONS [38]

q	resolution	Required data rate v_q	resolution vMOS s_q
1	4k or more	25 Mbps	4.9
2	2k	12 Mbps	4.8
3	1080p	10 Mbps	4.5
4	720p	5 Mbps	4
5	480p	2 Mbps	3.6
6	360p	1.5 Mbps	2.8

connect to the core networks through these 9 BSs. Besides those, perfect synchronization is assumed in this paper. SISO case is considered in our paper, but it is attractive to extend it to MIMO in the future work. The remaining simulation parameters are summarized in Table. II. The values with $*$ mean default values. The information on video streams, such as required data rate and corresponding MV-MOS, are shown in Table III.

In our simulations, we compare five different schemes: the no-caching scheme (baseline (no cache)) that delivers video traffic in wireless networks without caching; the static caching (baseline) scheme (caches is fixed in the research period) with considering resource allocation, similar to what is discussed in [17] and [21], and the proposed dynamic caching scheme that replaces the caches with considering resource allocation and video quality selection. Two scenarios of the content request probability π are tested. One is high probability ranging at $(0.5-0.95]$ and the other is low probability ranging at $(0-0.5]$. The high probability means that the user behavior is easier to predict (e.g., watching popular video and tractable paths) and the low probability means the user may quickly change the watched video and wander in the city.

A. Algorithm Performance

Fig. 3 demonstrates the evolution of the proposed dual decomposition algorithm for different initial values of λ_j. The iteration step k refers to the main loop iteration of Algorithm 2. At each step we calculate the differences between obtained utility $U^{[k]}$ and the optimum utility U^* by solving

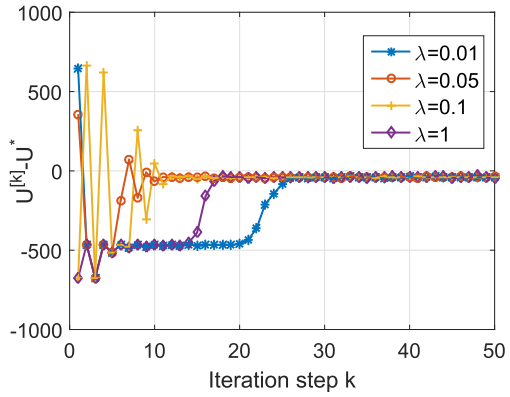


Fig. 3. Convergence and optimality of the proposed scheme.

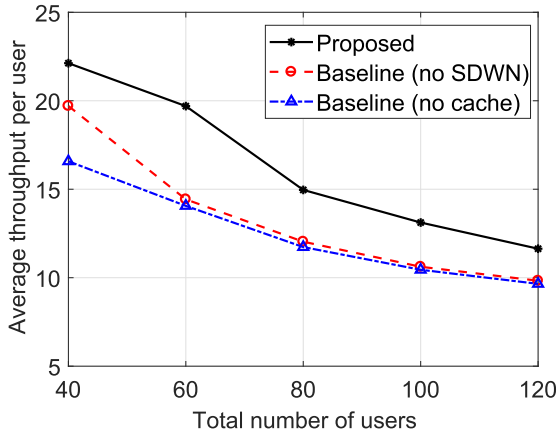


Fig. 4. Average throughput per user with different network loads.

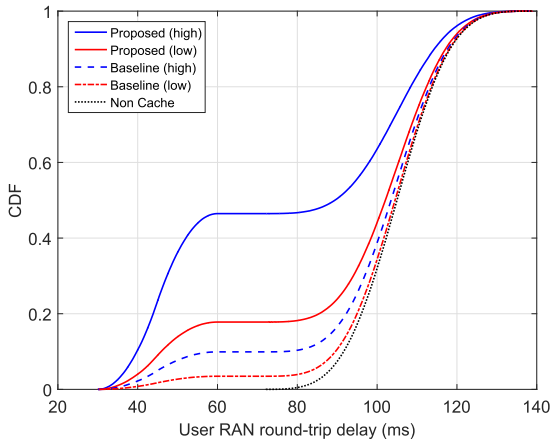
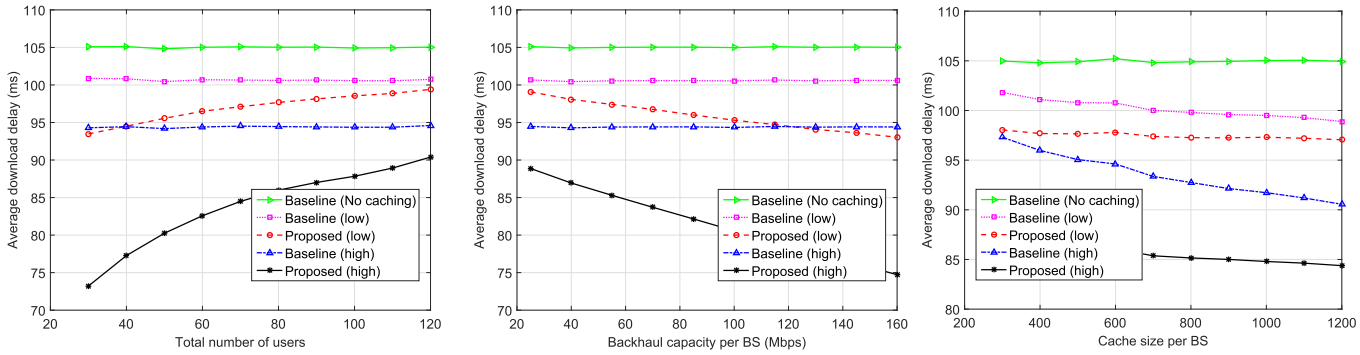


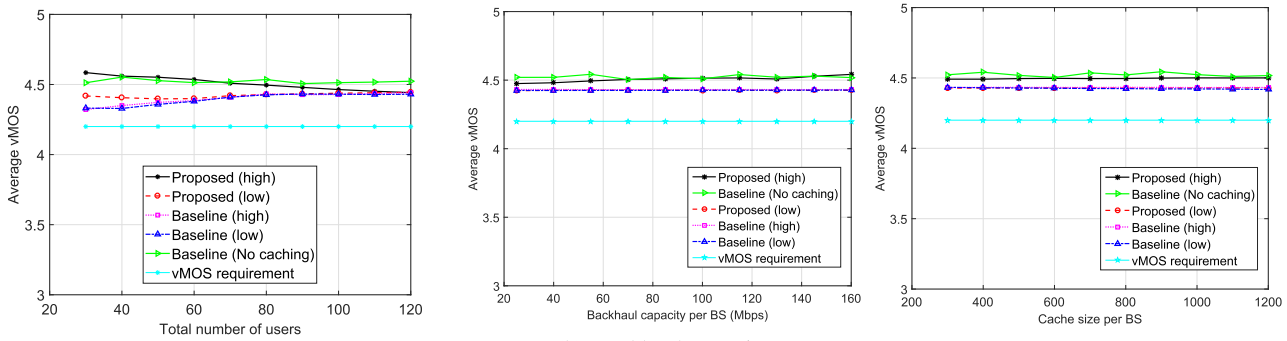
Fig. 5. CDF of round trip delay (backhaul and RAN only) of users.

the relaxed problem (upper bound) $\tilde{\mathbf{P0}}$. In most of simulation instances shown in Fig. 3, we see that the proposed algorithm converges to a fixed point. It can be observed that the iterative algorithm converges to the optimal value within ten steps when $\lambda = 0.05$, which means the optimum dynamic caching strategy and the bandwidth provisioning can be achieved within a few iterations. However, Fig. 3 also suggests that some inappropriate initial values of λ may result in a worse convergence speed. As shown in Fig. 3, its performance can



(a) Network load (b) Backhaul capacity (c) Cache storage size

Fig. 6. Average mobile network delay with different network setups.



(a) Network load (b) Backhaul capacity (c) Cache storage size

Fig. 7. Average vMoS with different network setups.

approach the upper bound with a slight gap, especially after ten steps, which embraces a relatively acceptable solution.

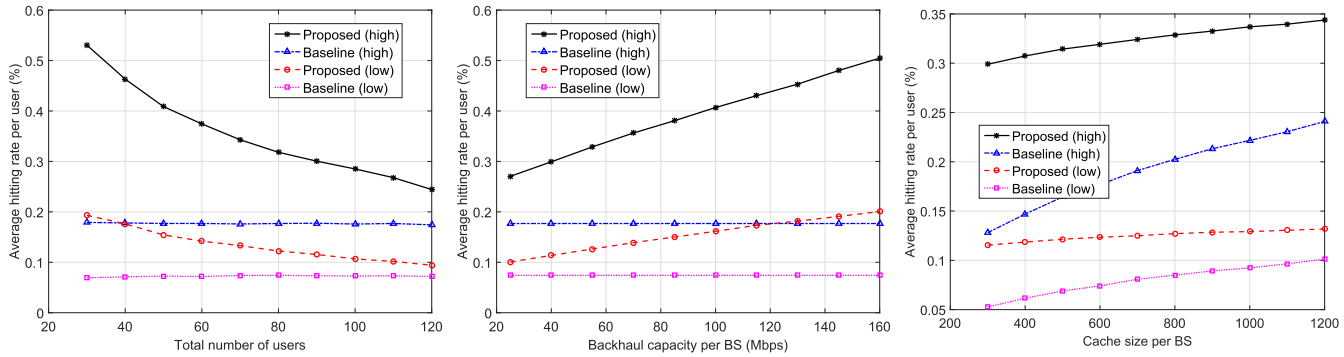
To evaluate the throughput of the proposed scheme, we compare our proposed dynamic caching scheme with two peers. The first one called Baseline (no SDWN) is the proposed caching policy stated in [33] and the second called Baseline (no cache) is the traditional network. As shown in Fig. 4, both two cases with caching show better performance than non-caching case. Moreover, by deploying bandwidth provisioning using SDWN, the proposed scheme shows further improvements on the throughput per user compared to other schemes.

B. Network Performance

1) Delay: Transmission delay is the key performance metric for video streaming in cache-enabled SDWNs. Users may give up watching the video because of the long waiting time for video buffering. In this subsection, we evaluate the average transmission delay performance and the results are shown in Figs. 5 and 6.

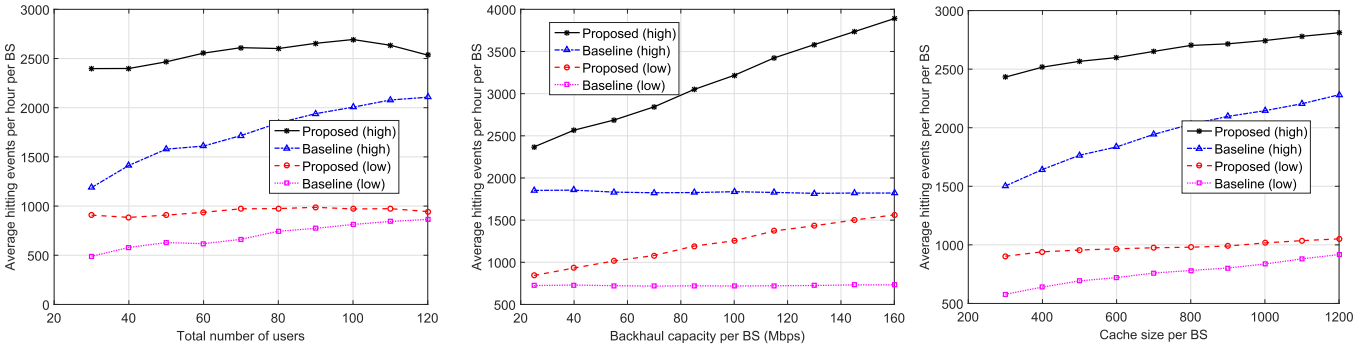
Fig. 5 shows the cumulative distribution function (CDF) of round trip delay of users with different caching schemes and prediction accuracy respectively. The CDFs for dynamic and static caching all improve significantly at delay reduction compared to no in-network caching case, showing a maximum 2.5x gain, in both low and high probability settings. Specifically, the proposed proactive caching boosts the performance of network delay deduction the most when we can keep caching contents with high probability to be requested. Almost 50% users experience delay less than 60ms because their backhaul latency is eliminated by the cache. Moreover, we can see even with low probability setting, the performance surpasses the static caching strategy, which implies that the proposed caching scheme should be deployed no matter the probability. It should be noted that the flat curves appeared in CDFs (between 60 ms to 80 ms) due to the elimination of backhaul latency brought by deploying in-network caching.

Results shown in Fig. 6 compare the average delay performance among different caching schemes and prediction accuracies. In Fig. 6a and Fig. 6b, note that the average delay of users can be further reduced by increasing available physical resource such as available cache space or (and) backhaul bandwidth of each BS because more data are going to be cached when we have more resources. It is observed that the proposed dynamic caching scheme with high probability always achieves the lowest while the proposed dynamic caching scheme with low probability is also better than static caching scheme. However, when the probability is not high, increasing backhaul resource or cache resources has a tiny effect on the delay reduction. Moreover, it can be seen in Fig. 6b that proposed schemes with low probability surpasses the baseline when the backhaul capacity is larger than 90 Mbps. The probable reason is that the bottleneck here is behaviors of



(a) Network load (b) Backhaul capacity (c) Cache storage size

Fig. 8. Average hit ratio with different network setups.



(a) Network load (b) Backhaul capacity (c) Cache storage size

Fig. 9. Average hitting events with different network setups.

users instead of the physical resource. As shown in Fig. 6c, the increase of users degrades the performance of the network, because it implies that more physical resources are put to real-time flows rather than caching data. This results in less hitting events that can reduce the delay. It is interesting to observe in Fig. 6c that the increase of traffic load has very little effect on the cases where no dynamic caching is deployed. The intuitive reason is that since static caching is independent of traffic load.

2) QoE Guarantee: In addition to the mean delay, it is also necessary to analyze the mobile MV-MOS of our proposed scheme. As shown in Fig. 7a, our proposed scheme satisfies the QoE requirements. Specifically, by putting more users into the network, the vMOS requirements are guaranteed even it has a slight decrease. Obviously, as shown in the other two figures, the boost of physical resource provides the better situation where degradation would not happen. Therefore, as mentioned Section IV-A, the proposed proactive caching does not undermine the video quality.

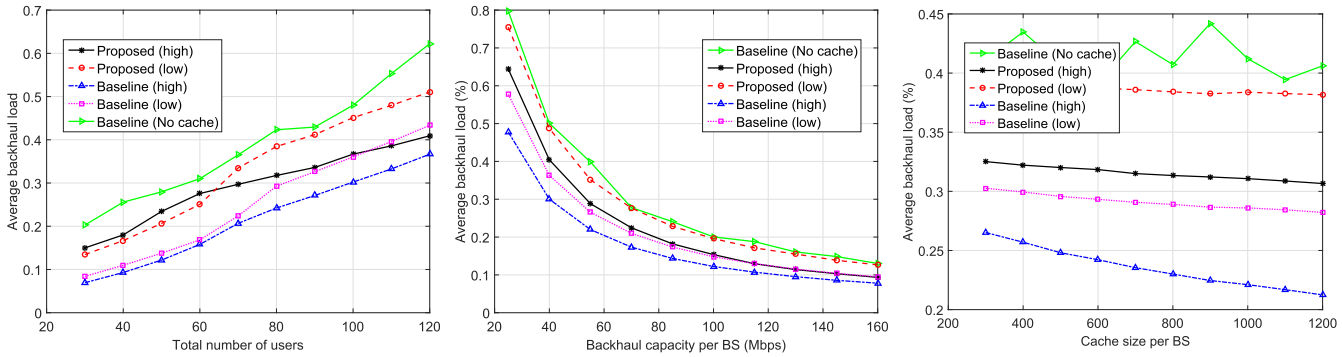
C. Utilization

In this subsection, the utilization of caching resource and backhaul are tested in terms of hitting events, hitting rate and backhaul load.

1) Caching Resource: The hit ratio is widely adopted as the performance metric to evaluate the caching mechanisms [7]. Since our first baseline does not consider in-network caching scheme, consequently, we omit the no-caching scheme in this subsection and illustrate the average hit ratio of the cached contents of the other four schemes in our simulations in Fig. 8. Due to the dynamic caching operations, the proposed scheme improves the average hit ratio performance by around 30% compared to the fixed caching scheme when the probability is high, and by around 10% at most if the probability is lower. In Fig. 8a, it is observed that the hitting ratio of both two proposed cases is decreasing with the total number of users because more users higher the load of backhaul used for streaming videos.

Another performance measurement metric, total hitting events, also can be used to evaluate the utilization of cache resource at each BS. Fig. 9 shows the average total hitting within one hour per BS. Obviously, our proposed schemes show better performance than passive caching schemes. Moreover, with increasing the number of users in the networks, the total cache hitting of fixed caching is increasing due to the larger amount of users boosts the opportunities. Besides network load, with the increase of backhaul capacity of BSs and the cache capacity of each BS, the average hitting events increase significantly by deploying the dynamic caching. The reason is the same as that we mentioned in the analysis of the delay.

2) Backhaul Resource: In this experience, we compare the (normalized) backhaul traffic load of video streaming



(a) Network load (b) Backhaul capacity (c) Cache storage size

Fig. 10. Average backhaul load with different network setups.

with different schemes in Fig. 10 with different parameters. In Fig. 10, we can see no-caching scheme takes the most backhaul load while all other four caching schemes can alleviate the backhaul. Specifically, with varying network status (network load, backhaul capacity and cache size), dynamic caching schemes cost more backhaul as they replace the caches more frequently than fixed caching schemes. Furthermore,the two caching cases (dynamic and fixed) of high probability save more backhaul. Apparently, higher number of users leads higher backhaul load shown in Fig. 10a and more physical resources give more flexible backhaul pressure. shown in Fig. 10b and Fig. 10c.

VI. Conclusions and Future Work

In this paper, we jointly studied radio resource allocation, dynamic caching and adaptive video resolution selection for cache-enabled SDWNs. We proposed a polynomial time algorithm to solve the joint problem. We took the dynamic caching and adaptive video resolution selection into account of the flow control problem in SDWNs and formulated a joint problem. Then, we transferred this problem by relaxation to a convex problem that can be solved efficiently, and an algorithm was designed. Simulation results were presented to show that the performance of our proposed scheme can improve the QoE in terms of delay and video quality as well as efficiency of physical resource utilization. Future work is in progress to consider mobile edge computing in the proposed framework.

Appendix
Computational Complexity Analysis

The complexity of Alg. 1 can be evaluated as follows. Firstly, the number of variables is IJ and the number of dual variables is $(I+1)J$, thus the elementary steps needed for calculating $c_{ij}^{[t+1]}$, $\mu_j^{[t+1]}$ and $v_{ij}^{[t+1]}$ are $(2I+1)J$. Secondly, the maximum loops of iterations for the calculation are T loops. Therefore, as other steps in Alg. 1 is independent from I and J, Alg. 1 runs in polynomial time with the time complexity of $O(T(2I+1)J) = O(IJ)$. Similarly, if Alg. 1 is running in a distributed manner at each BS. The number of variables and dual variables at each BS are at most I and $(1+I)$ respectively. Thus, the elementary steps for calculation are $(2I+1)$, which means the time complexity of each BS is $O(I)$.

The complexity for solving the problem (15) can be evaluated similarly to the above if we use dual methods (e.g., primal-dual interior method). The total number of variables is $QI + IJ + J^2$ and the number of dual variables is $J^2 + J + I + 1 + I + QI$. As a result, the elementary steps needed for calculating those variables at most are $T'(2QI + 2J^2 + IJ + 2I + J + 1)$ where T' is them maximum iterations, which means the time complexity for solving problem (15) is $O(\max\{I, J\}J)$ (a polynomial time algorithm).

In Alg. 2, the number of bandwidth prises is J, which means J steps are needed to update λ_j. As we explained above, the time complexity for updating $r_{lj}^{f,[k+1]}$, $r_{ij}^{a,[k+1]}$, and $x_{iq}^{[k+1]}$ are $O(\max\{I, J\}J)$ and for $c_{ij}^{[t+1]}$ are $O(IJ)$, respectively. Consequently, Alg. 2 obtains the solution with the the time complexity of $O(\max\{I, J\}J)$ as the complexities of rounding $\{c_{ij}\}$ and $\{x_{iq}\}$ up are just $O(IJ)$ and $O(QI)$ (usually $Q \ll J$).

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016–2021 White Paper*. Accessed on Aug. 02, 2017. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[2] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1617–1634, 3rd Quart., 2014.

[3] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, "Software defined mobile networks: Concept, survey, and research directions," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 126–133, Nov. 2015.

[4] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[5] 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. 805–817, Feb. 2016.

[6] Y. Cai, F. R. Yu, C. Liang, B. Sun, and Q. Yan, "Software-defined device-to-device (D2D) communications in virtual wireless networks with imperfect network state information (NSI)," *IEEE Trans. Veh. Tech.*, vol. 65, no. 9, pp. 7349–7360, Sep. 2016.

[7] P. Si, H. Yue, Y. Zhang, and Y. Fang, "Spectrum management for proactive video caching in information-centric cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2247–2259, Aug. 2016.

[8] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[9] W.-C. Liao, M. Hong, H. Farmanbar, X. Li, Z.-Q. Luo, and H. Zhang, "Min flow rate maximization for software defined radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1282–1294, Jun. 2014.

[10] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 26–36, Jul. 2012.

[11] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[12] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[13] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[14] H. A. Pedersen and S. Dey, "Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing," *ACM/IEEE Trans. Netw.*, vol. 24, no. 2, pp. 996–1010, Apr. 2016.

[15] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart., 2015.

[16] N.-D. Đào, H. Zhang, H. Farmanbar, X. Li, and A. Callard, "Handling real-time video traffic in software-defined radio access networks," in *Proc. IEEE ICC Workshops*, Jun. 2015, pp. 191–196.

[17] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[18] H. Farmanbar and H. Zhang, "Traffic engineering for software-defined radio access networks," in *Proc. IEEE Netw. Oper. Manag. Symp. (NOMS)*, May 2014, pp. 1–7.

[19] M. S. El Bamby, M. Bennis, W. Saad, and M. Latva-aho, "Content-aware user clustering and caching in wireless small cell networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 945–949.

[20] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached mimo interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[21] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[22] A. Abboud, E. Baştuğ, K. Hamidouche, and M. Debbah, "Distributed caching in 5G networks: An alternating direction method of multipliers approach," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2015, pp. 171–175.

[23] C. Liang and F. R. Yu, "Bandwidth provisioning in cache-enabled software-defined mobile networks: A robust optimization approach," in *Proc. IEEE Veh. Tech. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–5.

[24] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Joint caching and base station activation for green heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3364–3369.

[25] V. A. Siris, X. Vasilakos, and G. C. Polyzos, "Efficient proactive caching for supporting seamless mobility," in *Proc. IEEE 15th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2014, pp. 1–6.

[26] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2015.

[27] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[28] J. Zhu, J. He, H. Zhou, and B. Zhao, "EPCache: In-network video caching for LTE core networks," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2013, pp. 1–6.

[29] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.

[30] R. Yu *et al.*, "Enhancing software-defined RAN with collaborative caching and scalable video coding," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[31] L. Ma, F. Yu, V. C. M. Leung, and T. Randhawa, "A new method to support UMTS/WLAN vertical handover using SCTP," *IEEE Wireless Commun.*, vol. 11, no. 4, pp. 44–51, Aug. 2004.

[32] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual resource allocation in information-centric wireless networks with virtualization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9902–9914, Dec. 2016.

[33] D. Niyato, D. I. Kim, P. Wang, and M. Bennis, "Joint admission control and content caching policy for energy harvesting access points," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[34] J. Qiao, Y. He, and X. S. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[35] F. Yu and V. C. M. Leung, "Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks," in *Proc. IEEE INFOCOM'*, Anchorage, AK, USA, Apr. 2001, pp. 518–526.

[36] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, "Big data analytics in mobile cellular networks," *IEEE Access*, vol. 4, pp. 1985–1996, Mar. 2016.

[37] Y. He, C. Liang, F. R. Yu, N. Zhao, and H. Yin, "Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, Jun. 2017, pp. 1–6.

[38] D. Schoolar. (2015) *Whitepaper: Mobile Video Requires Performance and Measurement Standards*. [Online]. Available: http://www-file.huawei.com/

[39] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 7071–7076.

[40] A. El Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 988–1001, Jun. 2015.

[41] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, Mar. 2016.

[42] G. Li and H. Liu, "Downlink radio resource allocation for multi-cell OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3451–3459, Dec. 2006.

[43] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[44] F. Yu and V. Krishnamurthy, "Effective bandwidth of multimedia traffic in packet wireless CDMA networks with LMMSE receivers: A cross-layer perspective," *IEEE Trans. Wireless Commun.*, vol. 5, no. 3, pp. 525–530, Mar. 2006.

[45] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, "Notes on decomposition methods," Stanford Univ., Stanford, CA, USA, Tech. Rep. EE364B, 2003. [Online]. Available: https://stanford.edu/class/ee364b/lectures/decomposition_notes.pdf

[46] G. Liu, F. R. Yu, H. Ji, and V. C. M. Leung, "Energy-efficient resource allocation in cellular networks with shared full-duplex relaying," *IEEE Trans. Veh. Tech.*, vol. 64, no. 8, pp. 3711–3724, Aug. 2015.

Chengchao Liang received the B.Eng. degree in communication engineering and the M.Eng. degree in communication and information systems from the Chongqing University of Posts and Telecommunications, China, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from Carleton University, Canada, in 2017. He is currently a Post-Doctoral Fellow with the Department of Systems and Computer Engineering, Carleton University, Canada. His research interests include radio resource management, edge computing and caching, network virtualization, and applications of convex optimization.

Ying He received the B.S. degree in communication and information systems from Dalian Ocean University, Dalian, China, in 2011, and the M.S. degree in communication and information systems from the Dalian University of Technology, Dalian, in 2015, respectively. She is currently pursuing the Ph.D. degree with the Dalian University of Technology and Carleton University. Her current research interests include big data, wireless networks, and machine learning.

F. Richard Yu (S'00–M'04–SM'08) received the Ph.D. degree in electrical engineering from The University of British Columbia in 2003. From 2002 to 2006, he was with Ericsson, Lund, Sweden, and a start-up in CA, USA. He joined Carleton University in 2007, where he is currently a Professor. His research interests include cross-layer/cross-system design, connected vehicles, security, and green ICT. He is a fellow of the Institution of Engineering and Technology. He is also a Distinguished Lecturer and a member of the Board of Governors of the IEEE Vehicular Technology Society. He received the IEEE Outstanding Service Award in 2016, the IEEE Outstanding Leadership Award in 2013, the Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly Premiers Research Excellence Award) in 2011, the Excellent Contribution Award at the IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from the Canada Foundation of Innovation in 2009, and the Best Paper Award at the IEEE ICC 2014, the Globecom 2012, the IEEE/IFIP TrustCom 2009, and the International Conference on Networking 2005. He serves on the editorial boards of several journals, including as Co-Editor-in-Chief of the *Ad Hoc & Sensor Wireless Networks* and a Lead Series Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and the *IEEE Communications Surveys & Tutorials*. He has served as the technical program committee co-chair of numerous conferences. He is a registered Professional Engineer in the province of Ontario, Canada.

Nan Zhao (S'08–M'11–SM'16) received the B.S. degree in electronics and information engineering, the M.E. degree in signal and information processing, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology, Harbin, China, in 2005, 2007, and 2011, respectively. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology, China. He has published more than 100 papers in refereed journals and international conferences. His recent research interests include interference alignment, cognitive radio, wireless power transfer, and physical layer security. He is a Senior Member of the Chinese Institute of Electronics. He received the Top Reviewer Award from the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2016. He was nominated as an Exemplary Reviewer by the IEEE COMMUNICATIONS LETTERS in 2016. He served as a TPC member for many interferences, e.g., Globecom, VTC, and WCSP. He is serving or served on the editorial boards of several journals, including the *Journal of Network and Computer Applications*, IEEE ACCESS, *Wireless Networks*, *Physical Communication*, *AEU-International Journal of Electronics and Communications*, *Ad Hoc & Sensor Wireless Networks*, and the KSII Transactions on Internet and Information Systems.