Enhancing Video Rate Adaptation With Mobile Edge Computing and Caching in Software-Defined Mobile Networks

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

***Abstract*— Recent advances in software-defined mobile networks (SDMNs), in-network caching, and mobile edge computing (MEC) can have significant effects on video services in next generation mobile networks. In this paper, we jointly consider SDMNs, in-network caching, and MEC to enhance the video service in next generation mobile networks. We use a new video experience evaluation standard called U-video mean opinion score (vMOS), which is a more advanced measurement of the video quality based on the well-known vMOS. With the objective of maximizing the mean U-vMOS, an optimization problem is formulated. Due to the coupling of video data rate, computing resource, and traffic engineering (bandwidth provisioning and paths selection), the problem becomes intractable in practice. Thus, we utilize a dual-decomposition method to decouple those three sets of variables. By this decoupling, video rate adaptation is performed at users with network assistants. End nodes can schedule computing resource independently. Traffic engineering is performed by the software-defined networking controller and base stations. Furthermore, to address the challenges of dynamic change of network status and the drawbacks caused by the frequent exchange of information, we design a decentralized algorithm based on alternating direction method of multipliers to solve the traffic engineering problem. Extensive simulations are conducted with different system configurations to show the effectiveness of the proposed scheme.**

***Index Terms*— Video rate adaptation, mobile edge computing, in-network caching, software-defined mobile networks, traffic engineering.**

I. Introduction

GLOBAL mobile data traffic will increase nearly eight-fold from 2015 to 2020, and 75 percent of them will be video [1]. Thus, the video service is replacing voice and other applications to become the major service in mobile networks [1]–[5]. Moreover, high-definition (HD) videos (e.g., 720p, 1080p, and beyond) that request at least 5-20 Mbps user data rate will be ubiquitous [6], which will bring great challenges to the design and operation of next generation mobile networks (e.g., 5G and beyond). To address these challenges, recent advances of information and communications technologies can be explored, such as *software-defined mobile networks* (SDMNs) [7], *in-network caching* [8] and *mobile edge computing* (MEC) [9].

SDMNs have been proposed to fully support SDN design in wireless networks, which enable the programmability in mobile networks so that the complexity and the cost of networks can be reduced [7]. With the programmability, SDN is considered as a promising candidate to enhance traffic engineering, which is a critical component in communication networks [10]. Generally speaking, traffic engineering is used to optimize the network and provision services with requirements by directing traffic in networks [10], [11]. The success of the utilizing SDMNs for traffic engineering depends critically on our ability to jointly provide the backhaul and radio access networks (RANs) for the traffic [11]–[13].

Another promising technology, in-network caching, as one of the key features of *information-centric networking* (ICN), can efficiently reduce the duplicate content transmission in networks [14]. Notably, caching content (e.g., videos) at mobile edge node (e.g., base stations (BSs) and routers) has been proposed as one of the key enablers in next generation mobile networks [8], [15]–[17]. The investigation on exploiting caching in future mobile networks has shown that access delays, traffic loads, and network costs can be potentially reduced by caching contents in mobile networks [8]. In the meantime, MEC has attracted great interest recently as computational resources are moved closer to users, which can efficiently improve the quality of service (QoS) for applications that require intensive computations (e.g., video processing and tracking) [9].

With the widely employed HTTP adaptive streaming, such as Dynamic Adaptive Streaming over HTTP (DASH) (e.g., Google and Netflix), the video client can request the proper quality level adaptively according to the network throughput. However, from the slow reaction of current protocols (e.g., DASH) to the fluctuated mobile network and

Manuscript received December 1, 2017; revised April 28, 2018 and July 27, 2018; accepted July 29, 2018. Date of publication August 21, 2018; date of current version October 9, 2018. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by Huawei Technologies Canada Co., Ltd. The associate editor coordinating the review of this paper and approving it for publication was M. 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.2018.2865354

higher requirements of quality of experience (QoE) challenge the design of next generation mobile networks. To address this problem, MEC that deploys computing servers at BSs of the radio access network (RAN) can proactively optimize the delivery of videos by transcoding videos to desired qualities according to network conditions [18]–[21]. Furthermore, network resources are competed among current DASH selfish clients, which results in instability in QoE, unfairness among clients, and under-utilization of the network resources [22]. To address this problem, network-assisted video streaming schemes are proposed to enable collaboration among DASH video clients and servers [22], [23].

As discussed in [23] and [24], by deploying SDN in wireless network, the management of complex wireless networks can be simplified, which can address the scalability issues. Moreover, since the entire status of the network is available at a SDN controller, the network can have more accurate knowledge on network resources, locations of content (edge) servers and clients, as well as network congestion, which makes E2E QoS provisioning more practical. Thus, to improve the utilization of network resources and E2E service quality, the collaboration among video delivery and network scheduling can be enabled by SDN [22]. Although some works have been done on SDMNs, in-network caching and MEC separately, jointly considering these new technologies to enhance the video service has been largely ignored in the existing research. In this article, we jointly consider SDMNs, in-network caching and MEC to enhance the video service in next generation mobile networks. Specifically, in the proposed framework, a SDN controller is deployed to steer the bandwidth provisioning and traffic paths selection while assisting nodes to perform edge computing and video quality selection. Besides, popular videos can be stored at caches of network nodes (e.g., router and BSs). We design an efficient mechanism that jointly considers network-assisted video rate adaptation, bandwidth provisioning in traffic engineering, and computing resource scheduling in MEC.

The distinctive technical features of this article are listed as follows:

- To specify the network performance for video services, we use a new video experience evaluation standard called U-vMOS proposed by Huawei mLAB [4] to model the network utility. U-vMOS is a more advanced measurement of the video quality based on the well-known video mean opinion score (MOS) introduced in [25] and supports a resolution ranging from 360P to 8K, which can adapt to the evolution of video resolutions.
- With the objective of maximizing the mean U-vMOS of a heterogeneous network (HetNet) comprised of macro BSs (MBSs), small BSs (SBSs) and multiple users, an optimization problem is formulated. The limited backhaul capacity and computing capability of each network node are considered.
- Due to the coupling of video data rate, computing resource, and traffic engineering (bandwidth provisioning and paths selection), the problem becomes intractable in practice. Thus, we utilize dual-decomposition method to decouple those three sets of variables. By this decoupling, video rate adaptation is performed at users with network assistants. End nodes are able to schedule computing resource independently. Traffic engineering is performed by the SDN controller and BSs.
- To address the challenges of dynamic change of network status and the drawbacks caused by the frequent exchange of information, we design a decentralized algorithm based on alternating direction method of multipliers (ADMM) [26] to solve the traffic engineering problem.
- Extensive simulations are conducted with different system configurations to show the effectiveness of the proposed scheme.

The rest of this article is organized as follows. Related works are presented in Section II. Section III introduces the system model and formulates the presented problem. Section IV describes the proposed algorithms and the corresponding analysis. Simulation results are discussed in Section V. Finally, we conclude this study in Section VI.

II. Related Works

TE has been used as a tool to optimize the performance of communication networks and provision service for a long time [10]. By routing data flows and allocating bandwidth of links to each flow, TE guarantees the QoS requirements of services and maximizes the desired network utility. With recent advances of SDN, TE can benefit from the programmability and the flexibility of SDN [27], [28]. An optimization of traffic engineering in the SDN controller is formulated by [27] and it shows that the introduction of SDN gets significant improvements in network utilization. The study in [28] tries to leverage SDN to get the enhancement of broadcast communications. The most related studies of our work are TE in SDMN [11]–[13], [29]. Reference [11] proposes 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 [11], the authors of [13] extend TE of SDMNs 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 SDMNs is investigated in [12] with jointly considering TE and physical layer interference management problem by weighted-minimum mean square error algorithm. Reference [29] proposes a weighted cost-minimization problem of TE with jointly considering the traffic load balancing and control-channel setup cost. Compared to those studies, our work exploits TE of SDMN in a MEC and caching-enabled HetNet where content can be retrieved from multiple places.

Video quality adaptation with radio resource allocation in mobile networks (especially, long-term evolution (LTE)) is studied in a number of studies (e.g., [30]–[34]). Reference [35] provides 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 [36]. This research suggests that SBSs form a wireless distributed caching network that can efficiently transmit video files to users. Meanwhile, Ahlehagh and Dey [37] 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. Reference [38], instead of RANs, moves attention of in-network video caching to core networks of LTE. To utilize new technologies in next generation networks, [39] continues the research in specific to the video. Another research combining the caching and video service is [40] where collaborative caching is studied with jointly considering scalable video coding. However, the RAN and overall video quality requirements are not considered.

In [9], an application called RAN-aware content optimization has been proposed as one of the essential service scenarios of MEC. This application can provide an estimated throughput of the wireless link to the video server so that the video rate selection and congestion control can respond to the network condition fast. Reference [41] extends this application to which realizes context-aware content localization in order to enhance user QoE in video distribution applications. As video analytic and cache are enabled by MEC, to further utilize the powerful computing resource of MEC server at edge nodes, [19], [20], [41] propose to jointly combine the advantages of caching and processing to increase the throughput of mobile networks and QoE of users. In [19], the MEC server transcodes a video with higher rate version in the cache to satisfy a request for a lower rate version according to the optimization of the video rate adaptation and the network condition. Along with this research, [20] designs a scheme where multiple MEC caching and servers are collaborative to provide the caching and processing of videos. Compared to those research, our present study introduces TE and SDMN into this joint framework of video caching and processing to enhance the network performance.

TABLE I

NOTATIONS (ORDERED IN APPEARING SEQUENCE)

Notation	Definition
$\mathcal{F}$	set of video (library)
f	videos
$q \in \{1, ..., Q\}$	video quality level (resolution) indicator
v_q	minimum transmission data rate
s_q	U-vMOS of the resolution
W (Hz)	total spectrum bandwidth
$\mathcal{N}$	set of network nodes
$\mathcal{J}$	set of mobile BS nodes
$\mathcal{I}$	set of UEs
$\mathcal{L}$	set of links
$\mathcal{L}^{wl}$	set of wireless links
$\mathcal{L}^{wd}$	set of wired links
l	directed link
n_l	source node of the link
m_l	destination node of the link
B_l	data rate capacity of the wired link l
γ_l	received spectrum efficiency of the wireless link l
$g_{n_l m_l}$	large scale channel gain of the wireless link l
p_{n_l} (Watt/Hz)	normalized transmission power on link l
$\sigma_0(Watt/Hz)$	noise power spectrum density
$\mathcal{F}_j$	set of video (library) stored at the node
$h_{ij} \in \{0, 1\}$	hitting event indicator
C_j (bps)	computing capacity of MEC server
$x_{qi} \in \{0, 1\}$	resolution indicator of the user
$\mathcal{P}_i$	path set of the user
$\mathcal{P}_{ij}$	path set starting from the node
$p_{ij}^k \in \mathcal{P}_i$	k-th path of the flow i starting from the node j
r_{ij}^k	data rate of path p_{ij}^k
$y_{ij} \in \{0, 1\}$	computing task assignment indicator
$\mathit{\Pi}_r$, $\mathit{\Pi}_x$, and $\mathit{\Pi}_y$	local feasible sets
μ, λ, ν, ϖ	dual variables
z^λ, z^μ	sub-gradients
τ	length of iteration step
$[t]$	iteration step
$\check{r}_{ij}^k$, $\check{r}_{ij}^{k,l}$,$\hat{r}_{ij}^k$	local version of r_{ij}^k
ρ	predefined augmented Lagrangian parameter

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we present the system model of video streaming, mobile network, MEC and caching. The problem is formulated after the related assumptions are given.

In this article, we use a bold capital letter to indicate matrices and vectors. $|\mathcal{S}|$ is used to indicate the number of elements of the set $\mathcal{S}$. $\lfloor a \rfloor$ means taking the maximum integer value that is less than a. $[x]^+ = max\{0, x\}$ denotes the projection function to the nonnegative orthant. Other notations are summarized in Table I.

A. Network Model

1) Video Streaming Model: We assume that users watch streaming videos selected from a source library $\mathcal{F}$ (e.g., Youtube or Netflix) and each video streaming is served by one data flow. For simplicity, one user can only watch one video at the same time, which leads us to use the notation i to index the flow and the corresponding user. Each video file $f \in \mathcal{F}$ is encoded at a finite number of different quality levels (resolutions) $q \in \{1, \ldots, Q\}$, which is similar to DASH service. We assume that each level q requires a minimum data rate v_q (bps) to support smooth playback. Practically, v_q depends on video coding schemes and the video content, which are varying with time. Nevertheless, since the purpose of our article is to maximize the mobile network performance dynamically, we can consider the required data rate is a fixed value v_q when we do the scheduling.

To evaluate the gain of the video quality, the measure of each video quality q is defined as s_q. In this article, to adapt ubiquitous HD videos in next generation mobile networks, we refer the measure to U-vMOS proposed in [4] and [6]. The three key network elements in U-vMOS are video definition, initial buffering delay, and video freeze duration. In this article we only deploy the video definition as the measure due to the following reasons. First, the main focus of this article is cross-scheduling among computing and networking, so we select the definition related to trascoding computing as the performance measurement. Second, sufficiently provisioned network bandwidth ensures smooth streaming. Since sufficient bandwidth is one of the constraints in our proposed problem, we eliminate the consideration of stalling here. Last, considering the portion of users just starting the video is very small in the whole system, we ignore this part in this article. As shown in Table II, q denotes the level of an available resolution shown in the second column. Elements in the third column of Table II are minimum bandwidth (data rate) requirements v_q to guarantee the corresponding resolutions. The fourth column of

TABLE II
U-vMOS OF VIDEO RESOLUTIONS [4]

q	resolution	mean user required data rate v_q	U-vMOS s_q
1	360p	0.5 Mbps	2.8
2	480p	1 Mbps	3.6
3	720p	4 Mbps	4
4	1080p	5 Mbps	4.5
5	2k	10 Mbps	4.8
6	4k	24 Mbps	4.9

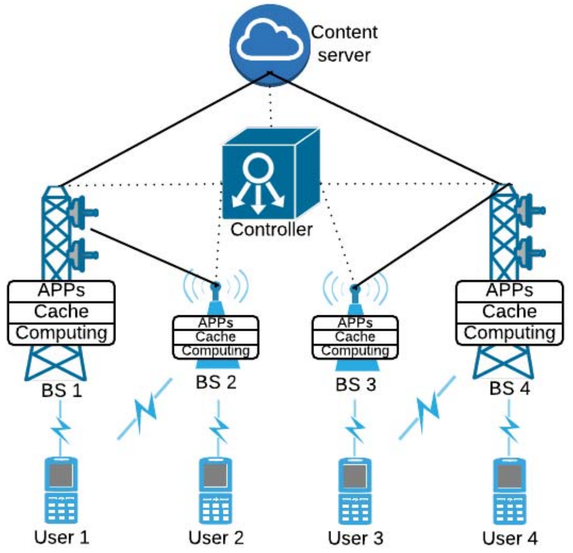


Fig. 1. MEC-enabled software-defined mobile network.

Table II is the U-vMOS of dedicated video resolution, where higher values represent better satisfaction from users.

2) *Network and Communication Model:* In this article, as shown in Fig. 1, we consider a mobile edge computing and caching (MECC)-enabled heterogeneous network (HetNet) with the backhaul network and the radio access network (RAN). The backhaul network is assumed to be a mesh network connecting SBSs, MBSs and Gateways (GWs) by wired links with fixed capacities. Users are connected to BSs with wireless channels sharing total W (Hz) radio spectrum resource. This considered network is modeled by a directed graph $G(\mathcal{N}, \mathcal{L})$. $\mathcal{N}$ includes network nodes (GWs and BSs) formed a set $\mathcal{J}$ and users formed a set $\mathcal{I}$. $\mathcal{L}$ comprised of wired and wireless links denoted by sets $\mathcal{L}^{wd}$ and $\mathcal{L}^{wl}$, respectively. $m_l \in \mathcal{N}$ and $n_l \in \mathcal{N}$ are used to denote the destination node and the source node of link l, respectively.

If link l is a wired link, it is assumed to provide a fixed bandwidth capacity B_l. If link l is a wireless link, the capacity depends on the ratio of the radio resource that the network allocates to this link. In this article, to simplify our analysis, we do not consider any advanced interference management and power allocation schemes. Thus, by using the Shannon bound, the spectrum efficiency of wireless link l is defined as

$$\gamma_l = \log\left(1+\frac{g_{n_l m_l} p_{n_l}}{\sigma_0 + \sum_{\substack{n'_l\in\mathcal{N}\\ n'\neq n}} g_{n'_l m_l} p_{n'_l}}\right) \qquad (1)$$

where $g_{n_l m_l}$ is the channel gain between the transmission node n_l (the source of link l) and the receiving node m_l (the destination of link l) including large-scale pathloss and shadowing, and σ_0 is the power spectrum density of additive white Gaussian noise. $\sum_{\substack{n'_l\in\mathcal{N}\\ n'\neq n}} g_{n'_l m_l} p_{n'_l}$ is the aggregated interference received from any other tranmission node n'_l. This interference model implies that all transmission nodes use the same frequency without any advanced interference mitigation scheme. In this article, p_{n_l} (Watt/Hz) is the normalized transmission power on link l. The fixed equal power allocation mechanism is used, which means transmission power p_{n_l} is the same for all frequencies. Accordingly, the achievable data rate capacity of link l is $R_l = W\gamma_l$.

It should be noted that small-scale fading is not considered when evaluating the SINR in this paper as the small-scale fading varies much faster than caching and bandwidth provisioning. Therefore, the SINR calculated by (1) can be considered as an average SINR over the transmission time. Moreover, in this article, as we mainly consider the benefits from caching and processing scheduling, radio resource allocation is not considered, which leads to the formulation where small-cale fading is ignored. If small-cale fading is considered, the network link capacity (e.g., γ_l) can be modeled as random variables, which leads a stochastic optimization problem where dynamic programming can be one of the methods to solve it.

Each data flow can be split into multiple paths as the user is assumed to be served by multiple BSs through the BS cooperation [12], [42], [43] or the multistream carrier aggregation [44]. Moreover, demanded videos can be potentially retrieved from any nodes (GWs, BSs or the source server) where matched data are found, which means each user can download the data of video from different places. For example, in Fig. 1, the data flow of user 1 is split into two paths where one is from the content source server to MBS 1 then to users 1 and another is directly from SBS 2.

3) *Caching and Computing Model:* The network is equipped with caching and computing functions on network nodes. We assume a subset $\mathcal{F}_j$ of $\mathcal{F}$ is stored at node j. It should be noted that node j always caches the highest quality Q of video file f so that it can be transcoding to a lower quality. As we mentioned above, if video file f_i demanded by user i is found at node j, namely $f_i \in \mathcal{F}_j$, node j becomes a candidate source. To indicate a hitting event between user i and node j, we define $h_{ij} = 1$ if $f_i \in \mathcal{F}_j$; otherwise, if $f_i \notin \mathcal{F}_j$, $h_{ij} = 0$ that means there is no hitting event.

If $h_{ij} = 1$ and node j is selected as one of source nodes of flow i, the video data needs to be transcoded to the required quality level except that the highest quality level is selected. However, unlike the powerful computing resource at the source server (e.g., the data center), due to the computing resource at each node, limited tasks can be activated at the same time. Similar to [19], we define the maximum mobile computing capacity as the number of encoded bits that can be processed per second, denoted by C_j (bps). For example, a 500 Mbps computing capacity means up to 500 Mbits video can be processed in one second. In this article, we do not consider the computing delay because the main benefit of this paper on latency reduction is from backhaul transmission. The same environment is set in [9] and [45]. Moreover, it is straightforward to optimize the extra latency in such MEC-enabled network by applying an existing method [46].

In this study, to simplify our analysis, we assume that, if multiple sources are available for a user, those sources are coordinated perfectly and the transmitted data arrive at the receiving node at the same time. This assumption gives a scenario where the user is not able to receive useless or unordered packets. Practically, the non-synchronous transmission is one of the challenges in multi-path delivery problem. The difficulties come from the content distribution, source coordination, and content request. If the potential sources are not coordinated perfectly, redundant (duplicated) packets may be received by the user, causing inefficient transmission. Another problem is that some packets may not be delivered by the source nodes or delivered in a wrongly scrambled order. Fortunately, network coding [47]–[49] has been proposed as a promising technology to solve this problem. For example, in [47] where Named Data Networking is assumed, after knowing the potential sources, the user will take advantage of multipath communication by sending interests for network coded Data packets over multiple network interfaces and network coded Data packets then follow multiple paths (i.e., the reverse paths followed by the interests) to reach the clients.

B. Problem Formulation

The network-assisted rate adaptation problem in a SDMN with MECC can be stated and formulated as follows.

1) Video Rate Adaptation: The purpose of this considered problem is to find an optimal video quality level for each user with considering network resources and the cached video distribution. We define a binary variable $x_{qi} \in \{0,1\}$ as the resolution indicator of user i. Specifically, if the q-th resolution of the video is selected by user i, $x_{qi} = 1$; otherwise, $x_{qi} = 0$.

2) Path Selection and Resource Allocation: To support video services demanded by users, an optimal path set for all data flows should be found by solving the proposed algorithm. Denote $\mathcal{P}_i$ as a path set including all candidate paths for user i and $\mathcal{P}_{ij}$ as a subset of $\mathcal{P}_i$ including all candidate paths starting from node j. A path $p_{ij}^k \in \mathcal{P}_i$ means the k-th path of flow i starting from node j and the corresponding data rate of this path is denoted by r_{ij}^k if p_{ij}^k is selected. Thus the achievable data rate of flow i is $\sum_{p_{ij}^k \in \mathcal{P}_i} r_{ij}^k$ that is the aggregated rate of all selected paths.

3) Computing Scheduling: As we mentioned in above, the computing resource on each node needs to be scheduled to video trans-coding tasks. In this article, we assume that the computing resource required for trans-coding video from the highest resolution to the requested resolution is c_{qi}. Obviously, $c_{qi} < c_{q'i}$ if $q > q'$. Specially, $c_{Qi} = 0, \forall i$. Thus, we define a variable $y_{ij} \in \Re^+$ (bps) as the computing resource assigned for user i at node j. If $y_{ij} > 0$ and $h_{ij} = 1$, node j is able to trans-code the video demanded by user i to the desired quality level at most to computing speed y_{ij}; otherwise, the video data cannot be retrieved from node j, namely $y_{ij} = 0$. y_{ij} can be zero if the content is cached at node j and the highest resolution is selected $x_{Qi} = 1$ as trans-coding is unnecessary.

4) Proposed SDN-Assisted Rate Adaptation Problem: To improve the whole network utility by maximizing the overall mean U-vMOS, the SDN controller performs traffic engineering to assist users adaptively selecting optimal video quality levels. In this study, the utility function is the average U-vMOS of all video users in the whole considered network instead of the average video quality. One of the features of U-vMOS is the consideration of fairness. The margin increase of scores decreases with the increase of video resolutions, giving logarithm-like presentation. Thus, the proposed problem can be formed as follows.

$$\max_{\mathbf{X},\mathbf{R},\mathbf{Y}} \quad U(\mathbf{X}) = \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \sum_{q=1}^{Q} s_q x_{qi} \tag{2a}$$

subject to

$$r_{ij}^k, y_{ij} \in \Re^+, \quad \forall i, j, k, \tag{2b}$$

$$x_{qi} \in \{0,1\}, \quad \forall q, i \tag{2c}$$

Content constraint:

$$\sum_{q=1}^{Q} x_{qi} = 1, \quad \forall i \in \mathcal{I}, \tag{2d}$$

$$\sum_{p_{ij}^k \in \mathcal{P}_i} r_{ij}^k = \sum_q v_{qi} x_{qi}, \quad \forall i \in \mathcal{I}, \tag{2e}$$

$$\sum_{p_{ij}^k \in \mathcal{P}_{ij}} r_{ij}^k \le h_{ij} y_{ij}, \quad \forall i \in \mathcal{I},\ j \in \mathcal{J}, \tag{2f}$$

$$\sum_q c_{qi} x_{qi} \le y_{ij}, \quad \forall i \in \mathcal{I},\ j \in \mathcal{J}, \tag{2g}$$

physical resource limitations:

$$\sum_{p_{ij}^k \in \mathcal{P}_l} r_{ij}^k \le B_l, \quad \forall l \in \mathcal{L}^{wd} \tag{2h}$$

$$\sum_{l\in\mathcal{L}^{wl}} \frac{\sum_{p_{ij}^k \in \mathcal{P}_l} r_{ij}^k}{\gamma_l} \le W, \tag{2i}$$

$$\sum_{i\in\mathcal{I}} y_{ij} \le C_j, \quad \forall j \in \mathcal{J}, \tag{2j}$$

where $\{x_{qi}\}$, $\{y_{ij}\}$ and $\{r_{ij}^k\}$ are elements of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{R}$, respectively. Constraint (2d) reflects that one and only one resolution level can be selected for one user. The demand constraint for every video flow is given by the (2e), which means the total data rate allocated to the user should be equal to the video data rate requirements. The constraint (2f) requires that the total data rate of any candidate paths starting from node j should be less than the computing speed. Moreover, this constraint also presents if $h_{ij} = 0$ (failed hitting), node j cannot be acted as one of the sources. In addition, the constraint (2g) requires the computing speed should be larger than the requirements of selected transcoding. Meanwhile it also shows that the maximum possible computing speed is the data rate of the highest resolution. Physical resources limitations are claimed by constraints (2h) (2i), and (2j) where $\mathcal{P}_l$ is the set of paths that pass link l. (2h) means the allocated data

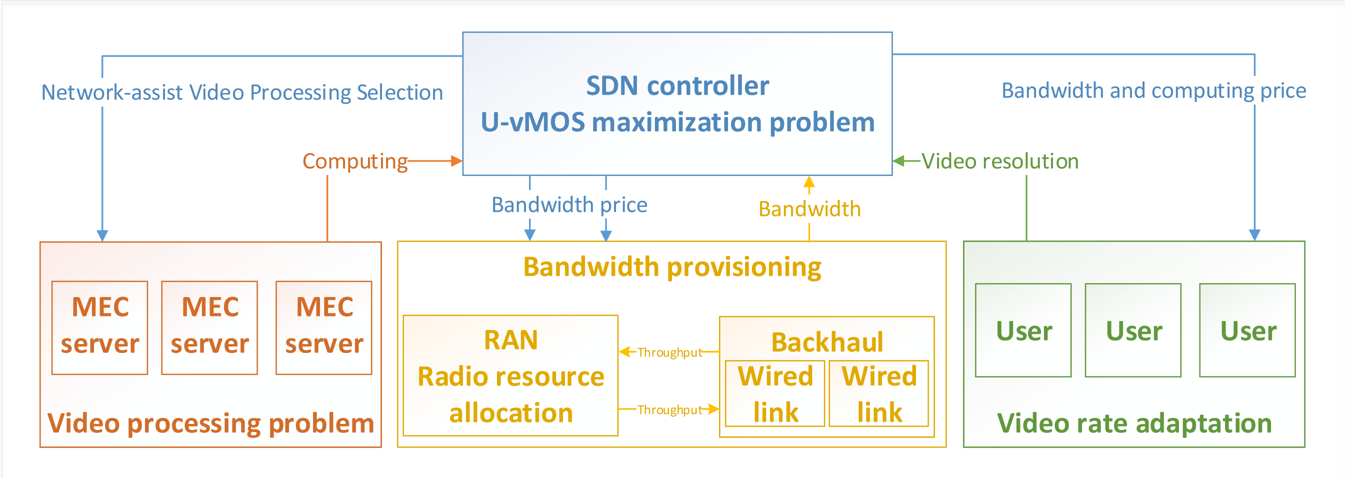


Fig. 2. Structure and information exchange of the proposed scheme.

rate of link l for all passing path should be less than the link capacity. As the radio resource is shared by the whole RAN, (2i) enforces that the total allocated spectrum cannot exceed the available spectrum bandwidth. The computing capacity on each node j is specified by the constraint (2j) where C_j (bps) is the maximum computing resource.

Unfortunately, problem (2) is difficult to solve and implement based on the following observations:

- The mix integer variables result in the problem a mix-integer linear problem (MILP) that generally is NP-complete.
- The complexity of solving (2) by using greedy or genetic methods will increase significantly with the increase of the number of users and (or) BSs.
- Video resolution, path selection, and resource scheduling are decided by different layers, network nodes and perform in different time scales.
- The necessary exchange of local information about the network and links affects the performance as overheads are introduced.

In the next section, we will develop a scheme to solve the problem (2) so that a near-optimal solution can be found.

IV. Proposed SDN-Assisted Video Rate Adaptation With MECC

In this section, dual-decomposition method is deployed to simplify problem (2). An ADMM-based algorithm is designed to perform traffic engineering. An example of the proposed scheme is presented as well.

A. Problem Decomposition

This article aims to give an efficient scheme to help video clients to select appropriate video resolutions while conducting resource scheduling to provision bandwidth and process video data. Observed from problem (2), the network tries to maximize the utility gained by video resolution. However, the video resolution is selected by the video client on each user. Moreover, MEC resource is assigned by each node based on the video distribution and requests, which leads the network hard to steer the traffic. Thus, the network needs to transfer some information to assist users and nodes when they perform video selection and processing so that the optimal network utility can be obtained.

Moreover, in the proposed problem, the channel status of users are changing dynamically and shorter than the change of video resolution selection and video process server scheduling. To decouple the fast changing channel status from other scheduling variables, we should decompose the original problem into a network bandwidth provisioning problem and a resolution selection problem as well as a video processing problem.

Those features of (2) motivate us to adopt dual-decomposition method so that the video quality selection and the MEC scheduling can be separated from the bandwidth provisioning of traffic engineering performed by the SDN controller. The brief idea and overall structure information exchange is given in Fig. 2. The meaning of each part we used in this figure will be introduced and defined in the remaining subsections. The main problem of U-vMOS maximization is solved at the SDN controller. The problem is decoupled into three parts, the computing scheduling problem solved decentralized at MEC servers, the bandwidth provisioning problem solved by the RAN and wired backhaul links, and best video resolutions selected by users. The SDN controller transfers resource prices to network nodes and users to assist them to solve the corresponding problem. After optimizing local variables, network elements feedback them to the SDN controller to help it to adjust the decision.

As we further decouple the RAN and the wired backhaul network, information exchange between them is presented as well. Practically, the bandwidth provisioning problem may be logically solved by the SDN controller. This depends on the level of the deployment of the SDN infrastructure. If all nodes in the network are SDN-enabled, the bandwidth provisioning problem can be solved with the SDN controller. Otherwise, cooperation between SDN-enabled nodes and traditional nodes is necessary.

Before we form the partial Lagrangian function for problem (2), let us define independent local feasible sets Π_r, Π_x, and $\mathit{\Pi}_y$ for variables $\mathbf{R}$, $\mathbf{X}$ and $\mathbf{Y}$, respectively. Those feasible regions only subject to constraints that include one type of variables, which are shown as

$$\mathit{\Pi}_r = \left\{\{r_{ij}^k\} \,\middle|\, \Re^+, (2h), (2i)\ \right\}. \tag{3}$$

$$\mathit{\Pi}_x = \left\{\{x_{qi}\} \,\middle|\, \{0,1\}, (2d)\ \right\}. \tag{4}$$

$$\mathit{\Pi}_y = \left\{\{y_{ij}\} \,\middle|\, \Re^+, (2j)\ \right\}. \tag{5}$$

Fortunately, the coupled constraints are (2e) and (2f). Thus, by relaxing constraints (2e) and (2f) with dual variables $\{\lambda_{ij}\}$, $\{\mu_i\}$, and $\{\nu_{ij}\}$,[1] the Lagrangian can be shown as:

$$\begin{aligned}\max_{\mathbf{X},\mathbf{R},\mathbf{Y}}\ & U(\mathbf{X}) - \sum_{i\in\mathcal{I},j\in\mathcal{J}} \lambda_{ij}\left[\sum_{p_{ij}^k\in\mathcal{P}_{ij}} r_{ij}^k - h_{ij}y_{ij}\right]\\ & + \sum_{i\in\mathcal{I}}\mu_i\left[\sum_{p_{ij}^k\in\mathcal{P}_i} r_{ij}^k - \sum_{q=1}^{Q} v_{qi}x_{qi}\right]\\ & - \sum_{i\in\mathcal{I},j\in\mathcal{J}}\nu_{ij}\left[\sum_{q=1}^{Q} c_{qi}x_{qi} - y_{ij}\right]\\ \text{s.t. }\ & \mathbf{X}\in\mathit{\Pi}_x,\quad \mathbf{Y}\in\mathit{\Pi}_y,\ \mathbf{R}\in\mathit{\Pi}_r.\end{aligned} \tag{6}$$

Thus, the original problem has been separated to two levels of optimization that are higher level for updating dual variables and low level for finding dual functions [50]. Accordingly, the dual problem (DP) then is:

$$\begin{aligned}\textbf{DP}:\quad & \min_{\mu_i,\nu_{ij},\lambda_{ij}\in\mathbb{R}} D(\mu_i,\nu_{ij},\lambda_{ij}) = g_x(\mu_i,\nu_{ij})\\ & \qquad + g_r(\mu_i,\lambda_{ij}) + g_y(\nu_{ij},\lambda_{ij}),\\ & \text{s.t. } \lambda_{ij}\geq 0,\quad \forall i\in\mathcal{I},\ j\in\mathcal{J}\\ & \qquad \nu_{ij}\geq 0,\quad \forall i\in\mathcal{I},\ j\in\mathcal{J}\end{aligned} \tag{7}$$

where $g_x(\mu,\nu)$, $g_r(\mu,\lambda)$, and $g_y(\lambda,\nu)$ are dual functions obtained as the maximum value of the Lagrangians solved in following problems (8), (9) and (10) for given $\{\lambda_{ij}\}$, $\{\mu_i\}$ and $\{\nu_{ij}\}$.

$$g_x(\mu,\nu) = \sup_{x_{qi}\in\mathit{\Pi}_x}\left\{\begin{array}{c} U(\mathbf{X}) - \sum_{i\in\mathcal{I}}^{I}\mu_i\sum_q v_{qi}x_{qi}\\ -\sum_{i\in\mathcal{I},j\in\mathcal{J}}\nu_{ij}\sum_q c_{qi}x_{qi}\end{array}\right\}, \tag{8}$$

$$g_r(\mu,\lambda) = \sup_{r_{ij}^k\in\mathit{\Pi}_r}\left\{\sum_{i\in\mathcal{I}}\sum_{p_{ij}^k\in\mathcal{P}_i}(\mu_i-\lambda_{ij})\,r_{ij}^k\right\} \tag{9}$$

$$g_y(\mu,\nu) = \sup_{y_{ij}\in\mathit{\Pi}_y}\left\{\sum_{i\in\mathcal{I},j\in\mathcal{J}}(\lambda_{ij}h_{ij}+\nu_{ij})\,y_{ij}\right\} \tag{10}$$

It is observed that $D(\mu,\nu,\lambda)$ is not a differentiable function due to the binary variables and candidate path sets. Thus, we can deploy subgradient method to solve the dual problem (7). Obviously, a sub-gradient of problem (7) for λ_{ij} is:

$$z_{ij}^{\lambda} = \sum_{p_{ij}^k\in\mathcal{P}_{ij}} r_{ij}^k - h_{ij}y_{ij}, \tag{11}$$

and for μ_i is

$$z_i^{\mu} = \sum_{p_{ij}^k\in\mathcal{P}_i} r_{ij}^k - \sum_{q=1}^{Q} v_{qi}x_{qi}, \tag{12}$$

and for ν_{ij} is

$$z_{ij}^{\nu} = \sum_{q=1}^{Q} c_{qi}x_{qi} - y_{ij}. \tag{13}$$

According to dual decomposition [51], we thus can update μ_i, ν_{ij} and λ_{ij} based on:

$$\mu_i^{[t+1]} = \mu_i^{[t]} - \tau_\mu^{[t]} z_i^{\mu}, \tag{14}$$

$$\nu_i^{[t+1]} = \nu_i^{[t]} - \tau_\nu^{[t]} z_{ij}^{\nu}, \tag{15}$$

and

$$\lambda_{ij}^{[t+1]} = \left[\lambda_{ij}^{[t]} - \tau_\lambda^{[t]} z_{ij}^{\lambda}\right]^+, \tag{16}$$

where $\tau_\mu^{[t]}$, $\tau_\mu^{[t]}$, and $\tau_\lambda^{[t]}$ are the length of step at iteration step $[t]$.

It can be seen that, after decomposing the original problem into sub-problems, the fast changing channel status has been decoupled from slow changing video resolution. The channel status only appears in the bandwidth provisioning problem (wireless part). The bandwidth provisioning problem aims at find a solution to satisfy the requirement given by the video resolution. If the video resolution of the user does not change, the bandwidth requirements is not changed. Once the bandwidth requirements cannot been satisfied, the network informs the video client to degrade the definition requirement by using the bandwidth price. The same scheme is applied to the computing resource scheduling.

Thus, if we are able to solve the inner problems (8), (9) and (10) in each iteration, the SDN controller can update dual variables and transfer them to nodes and users to assist them to find optimal solutions of their own variables x_{qi}, r_{ij}^k, and y_{ij}. This SDN-assisted optimization scheme is summarized in Alg. 1.

In practice, the convergence speed of subgradient method can be very slow, which gives undesired performance. However, in this study, the subgradient method provides an intuitive and direct method to update the dual prices according the changes of the network and videos. Basically, it can be observed that the steps of solving the dual problem are the steps of adjusting network to provision the service. If the environment (e.g., bandwidth, video resolution) is changed, the controller would not solve the problem from the beginning but only update the dual variables as the reaction to the environment.

In the remaining of this section, algorithms will be given to solve problems (8), (9) and (10).

B. Video Rate Adaptation Based on Network Information

Observe that problem (8) can be decoupled to users where the local problem of each user is shown as

$$\begin{aligned}\max_{x_{qi}\in\{0,1\}}\ & \sum_{q=1}^{Q} s_q x_{qi} - \mu_i\sum_q v_{qi}x_{qi} - \sum_{j\in\mathcal{J}}\nu_{ij}\sum_{q=1}^{Q} c_{qi}x_{qi}\\ \text{s.t. }\ & \sum_{q=1}^{Q} x_{qi} = 1.\end{aligned} \tag{17}$$

[1] Dual variables can be interpreted as costs of bandwidth and computing

Algorithm 1 SDN-Assisted Video Rate Adaptation and Resource Scheduling

```
Initialize: Set stop criteria $\epsilon$ and maximum iterative steps $T$
SDN controller sets initial dual variables $\nu_{ij}^{[0]}, \mu_i^{[0]}, \lambda_{ij}^{[0]}$;
Users (video clients) select initial video quality levels $\mathbf{X}$ and nodes (network function and MEC) set initial resource allocation $\{\mathbf{R}, \mathbf{Y}\}$.
$t = 0$.
for $t \leq T$ or $\epsilon$ is not met do
    SDN controller broadcasts dual variables $\lambda_{ij}^{[t]}, \mu_i^{[t]}, \nu_{ij}^{[0]}$ to users and nodes.
    Each user selects the video quality by solving (8) given $\nu_{ij}^{[t]}$ and $\mu_i^{[t]}$;
    Each node schedules the computing resources by solving (10) given $\lambda_{ij}^{[t]}, \mu_i^{[t]}$;
    The mobile provisions the bandwidth for flows by solving (9) given $\lambda_{ij}^{[t]}$ and $\nu_{ij}^{[t]}$;
    The SDN controller computes the sub-gradient of $\lambda_{ij}$, $\mu_i$ and $\nu_{ij}^{[t]}$ based on (11), (12) and (13).
    The SDN controller updates $\mu_i$, $\nu_{ij}^{[t]}$ and $\lambda_{ij}^{[t+1]}$ according to (14), (15) and (16).
    d) $t = t + 1$.
end for
Update $\{\mathbf{X}, \mathbf{R}, \mathbf{Y}\}$.
```

The above problem can be solved with effortless due to that only one quality level can be selected. Thus, each user only needs to select the level maximizing the utility.

Here we focus on the analysis of the structure of the objective so that the assistant from the SDN controller can be understood. μ_i represents the bandwidth cost (revenue) of user i given by the network. $\mu_i < 0$ means the network can provide more bandwidth for the user by using the revenue to push the user select higher resolution. However, if $\mu_i \geq 0$, the network may be lack of enough resource to support a higher video quality. Similar to μ_i, ν_{ij} represents the computing cost (revenue) at node j of user i given by the network. $\nu_{ij} < 0$ means the node j can provide more computing resource for the user by using the revenue to push the user select higher resolution.

C. Computing Resource Scheduling Based on Network Information

Similar to problem (8), problem (10) also can be decoupled to each node j as follows.

$$\begin{aligned} \max_{y_{ij} \in \{0,1\}} \quad & \sum_{i \in \mathcal{I}} \left(\lambda_{ij} h_{ij} + \nu_{ij} \right) y_{ij} \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}} y_{ij} \leq C_j, \end{aligned} \tag{18}$$

Obviously, this is a linear problem that can be solved easily. However, we can further ease the size of this problem so that common methods can be used, we form a set $\mathcal{I}_j^+$ including every user who has non-zero gain $(\lambda_{ij} h_{ij} + \nu_{ij})$ or non-zero h_{ij}. Formally, $\mathcal{I}_j^+ := \{i | \lambda_{ij} v_{ij} > 0, h_{ij} < \infty\}$. It is easy to see that $y_{ij} = 0$ if $i \notin \mathcal{I}_j^+$. Thus, we only need to consider users that in $\mathcal{I}_j^+$, which leads to a reduction of the problem size.

D. ADMM-Based Algorithms of Bandwidth Provisioning

This problem (9) is easy to solve theoretically as it is a linear problem. However, wired backhaul and radio access links are involved in this problem, which leads the solution hard to achieve in practice. Thus, a decentralized method is appropriate to cope with this problem. In this subsection, we propose to use ADMM as the tool to update r_{ij}^k. ADMM [52] is a simple but powerful algorithm that is well suited to distributed convex optimization. One of the essential properties of ADMM is its quick convergence to a modest accuracy of the optimal solution. It has been successfully used in many cases of mobile networks, such as routing [53], traffic engineering [12], radio resource allocation [54]. A brief overview of ADMM and its application on networks can found in [12] or [54]. We only summarize the main part of ADMM as follows. Generally, ADMM is able to solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ s.t. \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \end{aligned} \tag{19}$$

where $\mathbf{x} \in \mathbb{R}^{q \times 1}$, $\mathbf{z} \in \mathbb{R}^{r \times 1}$, $\mathbf{A} \in \mathbb{R}^{p \times q}$, $\mathbf{B} \in \mathbb{R}^{p \times r}$ and $\mathbf{c} \in \mathbb{R}^{p \times 1}$. There are two basic forms for the ADMM algorithm, namely the unscaled form, and the scaled form. In the unscaled form, the augmented Lagrangian is given as follows.

$$\begin{aligned} L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = & \ f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) \\ & + (\rho/2) \parallel \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c} \parallel_2^2, \end{aligned} \tag{20}$$

where $\mathbf{y} \in \mathbb{R}^{p \times 1}$ is the dual variable vector, $\rho > 0$ is the predefined augmented Lagrangian parameter and $\| \cdot \|_2$ is an Euclidean norm operator. Accordingly, the unscaled ADMM algorithm consists of the following iterations:

$$\mathbf{x}^{t+1} := \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{y}^t, \mathbf{z}^t), \tag{21}$$

$$\mathbf{z}^{t+1} := \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{t+1}, \mathbf{y}^t, \mathbf{z}), \tag{22}$$

$$\mathbf{y}^{t+1} := \mathbf{y}^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{c}), \tag{23}$$

where t is the iteration index.

ADMM can be considered as a combination of the dual decomposition and the augmented Lagrangian methods [52], [55]. To apply ADMM, problem (9) has to be decoupled to subproblems; thus we need first to separate the wired part and the wireless part by introducing some local variables.

We firstly separate the wired backhaul network and the radio access network, which is similar to studies in [12] and [56]. Denote the data rate of path p_{ij}^k decided by the wired backhaul network is $\check{r}_{ij}^k$ and its peer decided by the RAN is $\hat{r}_{ij}^k$. Furthermore, the wired backhaul network can be decoupled to links, as capacities of wired links usually are independent of each other. $\check{r}_{ij}^{k,l}$ denotes that the data rate of path p_{ij}^k allocated by link l. It should be noted that $\check{r}_{ij}^{k,l}$ does not mean p_{ij}^k passing link l. $\check{r}_{ij}^{k,l}$ can be considered the opinion for p_{ij}^k given by link l. All local variables of r_{ij}^k is the opinion or recommendation from one part of the whole network. By defining $\check{r}_{ij}^{k,l}$ and $\hat{r}_{ij}^{k}$, problem (9) can be revised as (24)

$$\begin{aligned}
\max_{\check{r}_{ij}^{k,l}, \hat{r}_{ij}^{k} \in \mathbb{R}^{+}} \quad & \frac{1}{2L^{wd}} \sum_{l \in L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) \check{r}_{ij}^{k,l} \\
& + \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) \hat{r}_{ij}^{k} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \alpha_{ij}^{k,l} \check{r}_{ij}^{k,l} \le B_l, \quad \forall l \in \mathcal{L}^{wd}, \\
& \sum_{l \in \mathcal{L}^{wl}} \frac{\sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \alpha_{ij}^{k,l} \hat{r}_{ij}^{k}}{\gamma_{n_l m_l}} \le W, \\
& \check{r}_{ij}^{l,k} = \hat{r}_{ij}^{k}, \quad \forall i, j, k, l.
\end{aligned} \tag{24}$$

where $L^{wd} = |\mathcal{L}^{wd}|$ is the number of wired backhaul links. Our goal is to develop a decentralized bandwidth algorithm based on ADMM that takes the form of a decomposition-coordination procedure. As the first step of ADMM, we form the augmented Lagrangian function of (24) as (25) as follows,

$$\begin{aligned}
\mathfrak{L}_{\rho}\left(x_l, r_l^f, x_{l,n}\right) = & \sum_{l \in L^{wd}} \frac{1}{2L^{wd}} \sum_{\substack{i \in \mathcal{I}, \\ p_{ij}^{k} \in \mathcal{P}_i}} (\mu_i - \lambda_{ij}) \check{r}_{ij}^{k,l} \\
& + \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) \hat{r}_{ij}^{k} \\
& + \sum_{l \in L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \nu_{ij}^{k,l} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right) \\
& - \frac{\rho}{2} \sum_{l \in L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right)^2,
\end{aligned} \tag{25}$$

where $\rho > 0$ is the predefined augmented Lagrangian parameter. As the augmented Lagrangian function is separable into a wireless part and wired links, ADMM algorithm can be adopted. The algorithm is summarized in Alg. 2. Obviously, problems (26) and (27) are all quadratic programming problems that can be solved completed decentralized across all links and SDN controller at the RAN without effort. Detailed information will be given as follows.

Following the research in [12], $\check{r}_{ij}^{k,l}$ in problem (26) can be updated according to the first-order derivation of the Lagrange shown as follows.

$$\check{r}_{ij}^{k,l} = \frac{1}{2L^{wd}} \left[\frac{\check{A}_{ij}^{k,l}}{\rho}\right]^{+}, \tag{29}$$

where $\check{A}_{ij}^{k,l} = (\mu_i - \lambda_{ij}) + \nu_{ij}^{k,l} + \rho \hat{r}_{ij}^{k} - \varpi^l a_{ij}^{k,l}$ and $\varpi^l \ge 0$ is the Lagrange dual variable of the capacity constraint on link l. Similarly, $\hat{r}_{ij}^{k}$ in problem (27) is calculated as follows.

$$\hat{r}_{ij}^{k} = \frac{1}{2} \left[\frac{\hat{A}_{ij}^{k}}{-\rho}\right]^{+}, \tag{30}$$

where $\hat{A}_{ij}^{k} = (\mu_i - \lambda_{ij}) - \sum_{l \in L^{wd}} \nu_{ij}^{k,l} - \rho \sum_{l \in L^{wd}} \check{r}_{ij}^{k,l} - \varpi \sum_{l \in L^{wl}} \frac{a_{ij}^{k,l}}{\gamma_{n_l m_l}}$ and $\varpi \ge 0$ is the Lagrange dual variable of the spectrum constraint on the RAN.

Algorithm 2 Decentralized Bandwidth Provisioning via ADMM

```
Initialize: Set stop criteria $\epsilon$ and maximum iterative steps $T$
    Set primal and dual variables $\{\check{r}_{ij}^{k,l}, \hat{r}_{ij}^{k}, \nu_{ij}^{k,l}\}$
    $t = 0$
for $t \le T$ or $\epsilon$ is not met do
    The SDN controller of the RAN broadcast $\{\nu_{ij}^{k,l}\}$ and $\{\hat{r}_{ij}^{k}\}$ to each wired link $l$
    Each wired link $l$ solves following problem to update $\{\check{r}_{ij}^{k,l}\}$;
        $$\begin{aligned} \max_{\check{r}_{ij}^{k,l} \in \mathbb{R}^{+}} \quad & \frac{1}{2L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) \check{r}_{ij}^{k,l} \\ & + \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \nu_{ij}^{k,l} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right) \\ & - \frac{\rho}{2} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right)^2 \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \alpha_{ij}^{k,l} \check{r}_{ij}^{k,l} \le B_l, \forall l \in \mathcal{L}^{wd}, \end{aligned} \tag{26}$$
    Each link $l$ send $\{\check{r}_{ij}^{k,l}\}$ to the SDN controller
    The SDN controller of the RAN solves following problem to update $\{\check{r}_{ij}^{k,l}\}$;
        $$\begin{aligned} \max_{\hat{r}_{ij}^{k} \in \mathbb{R}^{+}} \quad & \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) \hat{r}_{ij}^{k} \\ & + \sum_{l \in L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \nu_{ij}^{k,l} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right) \\ & - \frac{\rho}{2} \sum_{l \in L^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right)^2 \\ \text{s.t.} \quad & \sum_{l \in \mathcal{L}^{wl}} \frac{\sum_{i \in \mathcal{I}} \sum_{p_{ij}^{k} \in \mathcal{P}_i} \alpha_{ij}^{k,l} \hat{r}_{ij}^{k}}{\gamma_{n_l m_l}} \le W, \end{aligned} \tag{27}$$
    Update $\nu_{ij}^{k,l}$ according to:
        $$\nu_{ij}^{k,l}(t+1) := \nu_{ij}^{k,l}(t) + \rho\left(\check{r}_{ij}^{k,l} - \hat{r}_{ij}^{k}\right), \forall l \in \mathcal{L} \tag{28}$$
    $t = t + 1$
end for
Output the optimal bandwidth provisioning policy $\{r_{ij}^{k}\}$ .
```

E. Signaling Exchange Example

To illustrate the utilization of Alg. 1 into an SDMN, we give an instance presented by a signaling message exchange example in Fig. 3. It should be noted that video data can be retrieved from the source server and BS. In this example, we assume that the data will be retrieved from a serving BS.

As shown in Fig. 3, a video request (REQ) is sent by the user equipment (UE) and received by the MEC server (*step 1*). The MEC server seeks for matched videos in the local library (*step 2*). As the matched video is found, the MEC server provisions the computing resource to this potential task and

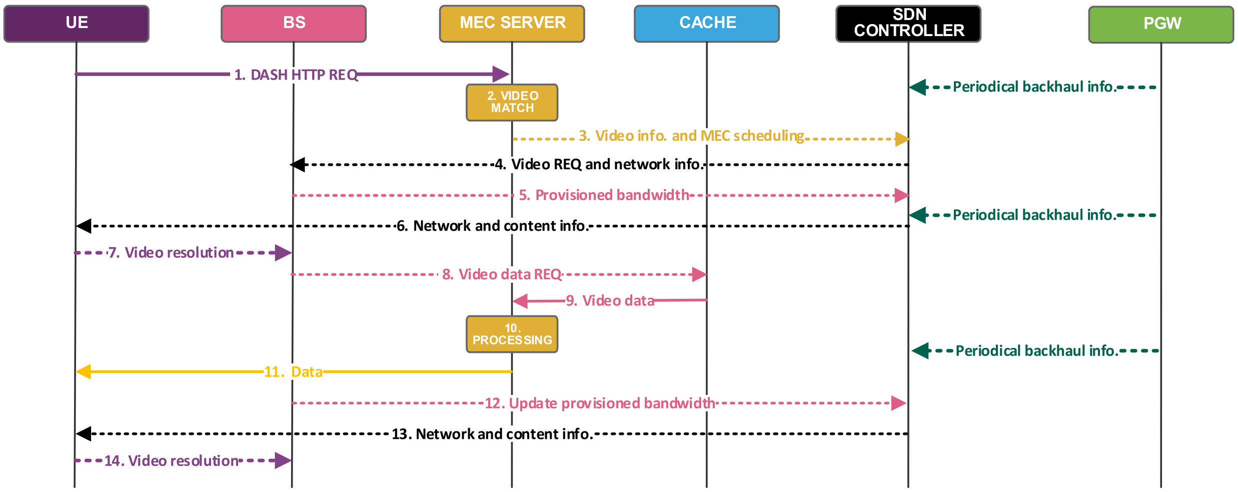

UE
BS
MEC SERVER
CACHE
SDN CONTROLLER
PGW
1. DASH HTTP REQ
Periodical backhaul info.
2. VIDEO MATCH
3. Video info. and MEC scheduling
4. Video REQ and network info.
5. Provisioned bandwidth
Periodical backhaul info.
6. Network and content info.
7. Video resolution
8. Video data REQ
9. Video data
10. PROCESSING
Periodical backhaul info.
11. Data
12. Update provisioned bandwidth
13. Network and content info.
14. Video resolution

Fig. 3. Signaling exchange in cache-enabled SDMNs.

sends info to the SDN controller (*step 3*). The SDN controller updates the network status according to feedbacks from BSs and the PGW. Video REQ and network assist information are delivered to the BS (*step 4*). The MEC server of the BS starts to reserve resource to compute potential task, and corresponding bandwidth is provisioned by related links of the BS (*step 5*). After receiving this information from the BS, the SDN controller sends the network assist information to the UE to help it the select video resolution (*step 6*). The video resolution selection is sent to the BS and cache; then the video is processed to the desired quality level and transmitted to the UE (*steps 7-11*). This procedure continues after the UE ends watching the video ((*steps 12-14*)).

V. Simulation Results and Discussions

In this section, simulation results are presented with different system configurations to demonstrate the performance of the proposed scheme.

A. Parameters Setup

1) Mobile Network Configurations: In the simulation, we consider a cellular network, consisting of one MBS, several SBSs, and multiple active users, that covers a 250m-by-250m area. Transmission with a single antenna for both transmitter and receiver is considered in our article. The remaining simulation parameters are summarized in Table III. Values with $*$ are default values.

2) Video, Cache and MEC Server Configurations: We assume that the total 1000 videos are in the video library $\mathcal{F}$. Each video f can be encoded to 6 levels with constant bit rate (CBR), and has the same length of 600 seconds. Each level maps a resolution in Table II.

Files in $\mathcal{F}$ have been sorted according to the popularity. We assume that the popularity of each video being requested follows a Zipf distribution with exponent 0.56 [36]. The f-th most popular video has a request probability of $(f^{-0.56})/(\sum_{f'=1}^{|\mathcal{F}|} f'^{-0.56})$. The default cache capacity S_n of an MBS is 200 video files, and default cache capacity of an SBS is 100 video files with the highest resolution, which leads the hitting rate of around 50% at MBS and 40% at SBSs. Least Frequently Used (LFU) caching policy is used at the MEC server to place/replace videos in caches, which means each BS stores the most S_n popular video files.

TABLE III
Network Parameters Settings

Network parameters	value
number of SBSs	$5-25$,15*
number of users	$5-25$,15*
frequency bandwidth (MHz)	20
transmission power profile	SISO with maximum power; 49dBm (MBS), 20dBm (SBS)

propagation profile	pathloss:L(distance)=34+40log(distance);
	lognormal shadowing: 8dB;
	no fast fading

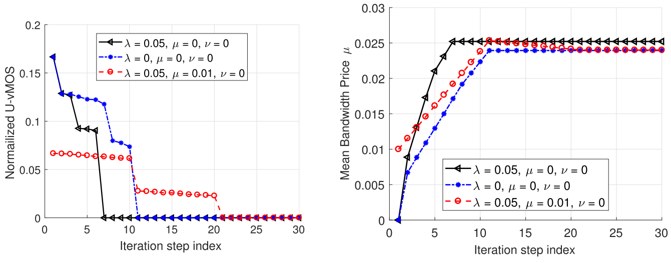
power density of the noise	-174 dBm/Hz
backhaul capacity (Mbps)	MBS to GW:100; SBS to MBS: 50

The MEC computing capability of an MBS is set to 150 Mbps that is equivalent to processing six videos simultaneously. Due to limitations of an SBS, the computing performance of a MEC server at SBS is only set to 50 Mbps equivalent to two videos.

B. Performance Metrics and Schemes

We evaluate the proposed system in three types of performance metrics, each of them focusing on three experiences with different network parameters, such as network load, BS density, computing resource, and cache storage capacity.

- *U-vMOS* is the objective of our proposed scheme. As pointed in [6], U-vMOS of 4.0 can be considered as a minimum requirement of the next generation mobile network.



(a) (b)

Fig. 4. Convergence and optimality of the proposed scheme.

- *Ratio of HD users* refers to the proportion of users that can select HD video (resolution higher than 1080p).
- *Average Load of MEC* refers to the average working load of the MEC servers located at BS nodes.

In our evaluation, we increase the number of users and SBSs to represent the increasing network load and BS density, respectively. Meanwhile, to clearly observe the effects of MEC resource, normalized computing resource and cache storage capacity are used to evaluate our proposed scheme. The term of normalized computing resource (cache storage capacity) means the ratio between the tested computing resource (cache storage capacity) to the default value.

In our simulations, we compare five different schemes shown as follows.

- *MECC* refers to our proposed scheme
- *MECC-MBS* refers to the scheme that has MEC and caches enabled at MBSs, which is proposed in [20].
- *Cache-only* refers to the scheme that has cache enabled but MEC is not available, which has been proposed in [17]. Therefore, videos in caches can be utilized only if the exact quality levels are matched.
- *No MECC* refers to the scheme that all data must be transferred from the source servers, but the source servers are aware of the RAN status. This case refers to a classical MEC use case called RAN-aware video optimization given in [9].

C. Simulation Results

1) Algorithm Performance: Fig. 4 demonstrates the evolution of the proposed dual decomposition algorithm for different initial values of λ and μ. The *normalized U-vMOS* is the ratio of absolute value between the optimal U-vMOS and the current U-vMOS to the optimal U-vMOS, namely $\frac{|U^{[k]}-U^{*}|}{U^{*}}$. The iteration step index refers to the main loop iteration of Alg. 1. At each step we calculate the differences between obtained utility $U^{[k]}$ (U-vMOS) and the optimum utility U^{*} by solving the original problem by exhausted search.[2] In the simulation instances shown in Fig. 4, 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 $\mu = 0$, which means the optimum strategy can be achieved within a few iterations. However, Fig. 4 also suggests that an inappropriate initial value of μ, such as $\mu = 0.01$, may result in a worse convergence speed. In summary, the performance of the proposed algorithm can approach the optimum solution after a few steps and provide a relatively acceptable solution.

2) Video Quality: In Fig. 5, the average U-vMOS of the overall users is presented with different network settings and MECC schemes. We can observe that the U-vMOS of the schemes with MECC is much higher than the schemes without MECC. Specifically, when the number of users increases shown in Fig. 5a, the average U-vMOSs decrease, as the network resources limit the performance. However, by deploying MECC, the average U-vMOS can reach 4 when 25 users are in the network, which means a user is able to have the resolution of 720p on average. To see the effect of the network density, we increase the number of SBSs in the same area as shown in Fig. 5b. Obviously, more SBSs lead to better performance for all cases as they can provide better coverage and channel conditions.

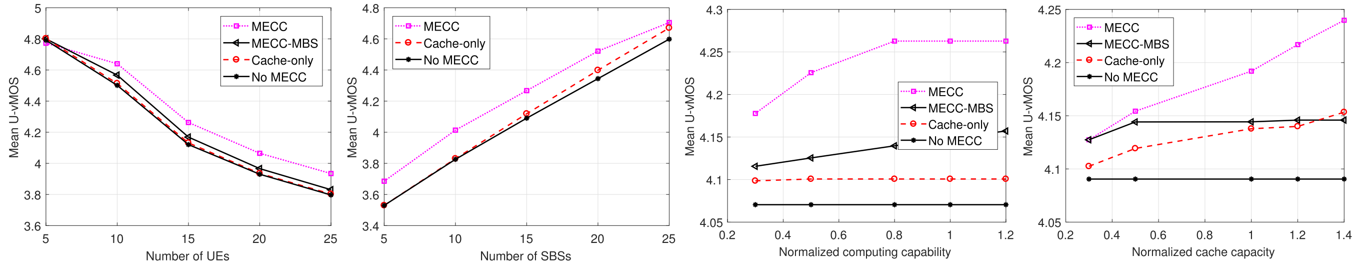
Figs. 5c and 5d show that increasing the capability of MEC servers and the capacity of caches can further improve the performance. It should be noted that Fig. 5c shows that the proposed MECC scheme becomes stable after a certain level of the capability of MEC servers, because other parameters, such as the hitting rate of caches, bandwidth of both wired and wireless links, are dominated the limitation. In Fig. 5d, we can see that more videos stored at caches, which implicitly means higher hitting rate, give better chance to utilize the proposed MECC scheme.

In addition to the mean U-vMOS, it is also necessary to discuss the distribution of video resolutions of our proposed scheme. As shown in Fig. 6, overall, the trend of the distribution of video resolutions is similar to the mean U-vMOS. It can be seen that the ratio of 1080p and higher resolutions decreases with the load of the networks, and increases with the available network resource. The proposed scheme gets a much higher ratio, up to 15%, on the 1080p and higher resolutions compared to the cache-only case and no MECC case. Note that in Figs. 6a and 6b, when the network load is high and network resource is very limited, the proposed MECC scheme still can provide around 20% and 13% of users with resolutions of 1080p and higher.

Some observations can be made from Figs. 5 and 6. Firstly, the MECC improves the performance of the network-assist video rate adaptation significantly on both average U-vMOS and HD videos. Secondly, if only MBSs can provide MEC, the gain of the proposed scheme is not as good as a traditional HetNet. Moreover, compared to traditional networks, in-network caching also can enhance the quality of video services. Furthermore, the capability of MEC servers may not always be the bottleneck of the system. Thus, in the next part, detailed tests are done on the load of MEC servers.

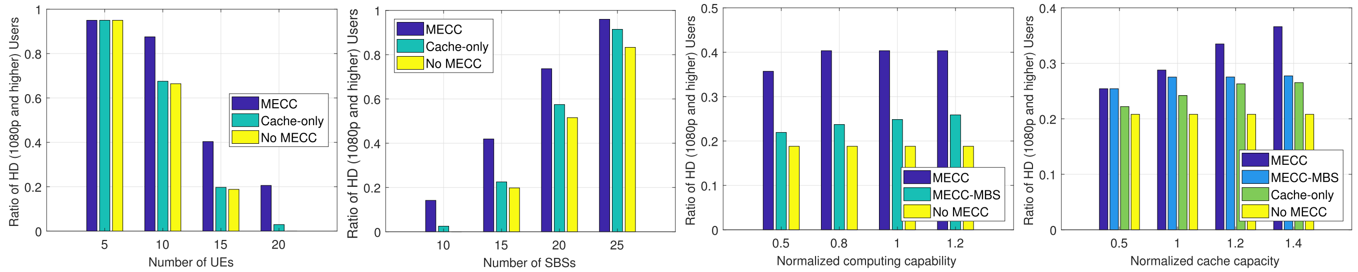
3) The Load of MEC Servers: As mentioned above, results shown in Fig. 7 compare the average load of MEC servers among schemes with different settings. To clearly see the internal relationships, we give two companions, *MECC-MBS-only* and *MECC-SBS-only*, where MEC servers are only deployed at SBSs and the MBS. We can apparently observe that

[2] Due to the time consuming, we reduce the size of the problem to 5 users and 2 BSs



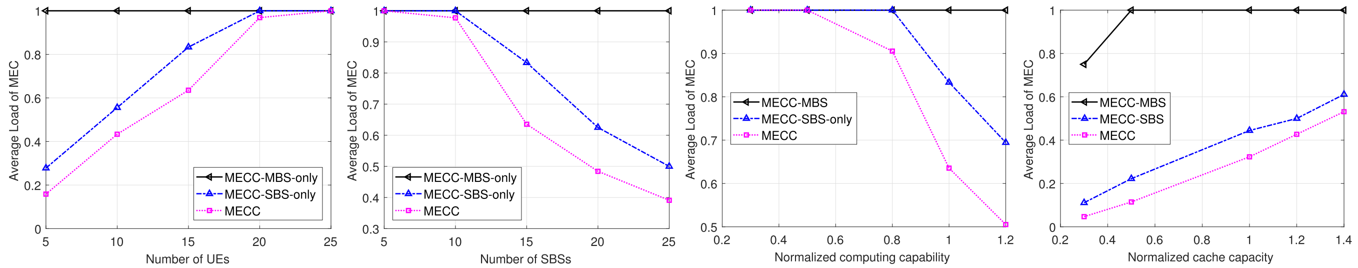
(a) Network load (b) BS density (c) Computing resource (d) Cache storage capacity

Fig. 5. The mean U-vMoS with different network setups.



(a) Network load (b) BS density (c) Computing resource (d) Cache storage capacity

Fig. 6. The distribution of video resolutions with different network setups.



(a) Network load (b) BS density (c) Computing resource (d) Cache storage capacity

Fig. 7. Average load of MEC servers with different network setups.

the load of the MEC server at MBS is always full, which means the MEC at the MBS can be considered as the first choice of the proposed scheme because the MBS has larger coverage and the server has more capability (2 times larger than SBSs). Moreover, as shown in Fig. 7a, the increase of network loads, leads higher computing loads as more tasks are needed to finish. Similarly, higher hitting rate also gives more opportunities to retrieve videos from network nodes, which give more tasks to servers, as shown in Fig. 7b. Both network settings in Figs. 7b and 7c can be treated as the increase of available computing resources. Fig. 7b gives more servers as more SBSs are deployed, while Fig. 7c enhances each server.

VI. Conclusions and Future Work

In this article, we jointly studied the network-assisted video rate adaptation problem in a MEC-enable SDMN where in-network caching was deployed. An optimization problem was formulated with the objective of maximizing the average U-vMOS of a HetNet. Dual-decomposition method has been utilized to decouple video data rate, computing resource, and traffic engineering (bandwidth provisioning and path selection) so that those variables could be obtained independently. To avoid the frequent exchange of network information, a decentralized algorithm based on ADMM was designed to solve the bandwidth provisioning problem. Simulation results were presented to show that our proposed scheme can significantly improve the mean U-vMOS. Future work is in progress to consider big data analytics in the proposed framework.

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2015–2020*, Cisco, San Jose, CA, USA, Feb. 2016.

[2] Y. Wen, X. Zhu, J. J. P. C. Rodrigues, and C. W. Chen, "Cloud mobile media: Reflections and outlook," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 885–902, Jun. 2014.

[3] Z. He, S. Mao, and S. Kompella, "Quality of experience driven multi-user video streaming in cellular cognitive radio networks with single channel access," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1401–1413, Jul. 2016.

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

[5] Z. Guan and T. Melodia, "Cloud-assisted smart camera networks for energy-efficient 3D video streaming," *IEEE Comput.*, vol. 47, no. 5, pp. 60–66, May 2014.

[6] Huawei Technologies Co., Ltd. (2016). *Whitepaper: 4.5G, Opening Giga Mobile World, Empowering Vertical Markets*. [Online]. Available: http://www.huawei.com/minisite/4-5g/img/4.5GWhitepaper.pdf

[7] 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.

[8] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[9] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," ETSI, Sophia Antipolis, France, White Paper 11, Dec. 2015.

[10] A. Mendiola, J. Astorga, E. Jacob, and M. Higuero, "A survey on the contributions of software-defined networking to traffic engineering," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 918–953, 2nd Quart., 2016.

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

[12] 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.

[13] N.-D. Dào, 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.

[14] 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.

[15] G. Paschos, E. Baştuğ, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[16] 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.

[17] C. Liang, Y. He, F. R. Yu, and N. Zhao, "Enhancing QoE-aware wireless edge caching with software-defined wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6912–6925, Oct. 2017.

[18] M. Patel *et al.*, "Mobile-edge computing introductory technical white paper," ETSI, Sophia Antipolis, France, Tech. Rep. 1, Sep. 2014.

[19] 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.

[20] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili. (2016). "Collaborative multi-bitrate video caching and processing in mobile-edge computing networks." [Online]. Available: https://arxiv.org/abs/1612.01436

[21] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint computation offloading and interference management in wireless cellular networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432–7445, Aug. 2017.

[22] K. T. Bagci, K. E. Sahin, and A. M. Tekalp, "Compete or collaborate: Architectures for collaborative DASH video over future networks," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2152–2165, Oct. 2017.

[23] A. Bentaleb, A. C. Begen, R. Zimmermann, and S. Harous, "SDNHAS: An SDN-enabled architecture to optimize QoE in HTTP adaptive streaming," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2136–2151, Oct. 2017.

[24] Z. Guan, L. Bertizzolo, E. Demirors, and T. Melodia. (2017). "WNOS: An optimization-based wireless network operating system." [Online]. Available: https://arxiv.org/abs/1712.08667

[25] *ITU-T: Methods for Subjective Determination of Transmission Quality*, document Rec. 800, Jan. 1996.

[26] J. Eckstein, "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," Rutgers Univ., New Brunswick, NJ, USA, RUTCOR Res. Rep. 32-2012, 2012.

[27] S. Agarwal, M. Kodialam, and T. V. Lakshman, "Traffic engineering in software defined networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2211–2219.

[28] L.-H. Huang, H.-C. Hsu, S.-H. Shen, D.-N. Yang, and W.-T. Chen, "Multicast traffic engineering for software-defined networks," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[29] H. Huang, S. Guo, W. Liang, K. Li, B. Ye, and W. Zhuang, "Near-optimal routing protection for in-band software-defined heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2918–2934, Nov. 2016.

[30] 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.

[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] S. Pudlewski, N. Cen, Z. Guan, and T. Melodia, "Video transmission over lossy wireless networks: A cross-layer perspective," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 6–21, Feb. 2015.

[33] 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.

[34] V. Joseph, S. Borst, and M. I. Reiman, "Optimal rate allocation for video streaming in wireless networks with user dynamics," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 820–835, Apr. 2016.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] C. Ge, N. Wang, S. Skillman, G. Foster, and Y. Cao, "QoE-driven DASH video caching and adaptation at 5G mobile edge," in *Proc. ACM Conf. Inf.-Centric Netw.*, Sep. 2016, pp. 237–242.

[42] 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.

[43] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[44] E Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy-efficient multi-stream carrier aggregation for heterogeneous networks in 5G wireless systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7432–7443, Nov. 2016.

[45] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, "Collaborative multi-bitrate video caching and processing in mobile-edge computing networks," in *Proc. 13th Annu. Conf. Wireless On-Demand Netw. Syst. Services (WONS)*, Feb. 2017, pp. 165–172.

[46] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[47] J. Saltarin, E. Bourtsoulatze, N. Thomos, and T. Braun, "Adaptive video streaming with network coding enabled named data networking," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2182–2196, Oct. 2017.

[48] Q. Wu, Z. Li, and G. Xie, "CodingCache: Multipath-aware CCN cache with network coding," in *Proc. 3rd ACM SIGCOMM Workshop Inf.-Centric Netw.*, 2013, pp. 41–42.

[49] M.-J. Montpetit, C. Westphal, and D. Trossen, "Network coding meets information-centric networking: An architectural case for information dispersion through native network coding," in *Proc. 1st ACM Workshop Emerg. Name-Oriented Mobile Netw. Design-Archit., Algorithms, Appl.*, 2012, pp. 31–36.

[50] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[51] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, "Notes on decomposition methods," Stanford Univ., Stanford, CA, USA, Notes EE364B, 2007, pp. 1–36.

[52] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[53] M. Leinonen, M. Codreanu, and M. Juntti, "Distributed joint resource and routing optimization in wireless sensor networks via alternating direction method of multipliers," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5454–5467, Nov. 2013.

[54] 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.

[55] Q.-D. Vu, L.-N. Tran, M. Juntti, and E.-K. Hong, "Energy-efficient bandwidth and power allocation for multi-homing networks," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1684–1699, Apr. 2015.

[56] H. Farmanbar and H. Zhang, "Cross-layer traffic engineering for software-defined radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3411–3416.

Chengchao Liang (S'15–M'17) received the B.Eng. and M.Eng. degrees 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 received the Senate Medal.

He is currently a Post-Doctoral Fellow with the Department of Systems and Computer Engineering, Carleton University, supported by Mitacs and Huawei Technologies Canada. His research interests include mobile networking, caching and computing with emphasis on cross-layer/system optimization, and convex theory. He is serving and has served as a Reviewer and TPC Member for several journals and conferences, such as the IEEE TRANSACTIONS ON NETWORKING, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE GLOBECOM, and the IEEE VTC.

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. 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–F'18) 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 California, USA. He joined Carleton University in 2007, where he is currently a Professor. His research interests include wireless cyber-physical systems, connected/autonomous vehicles, security, distributed ledger technology, and deep learning. 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 from the IEEE ICNC 2018, VTC 2017 Spring, ICC 2014, Globecom 2012, the IEEE/IFIP TrustCom 2009, and the International Conference on Networking in 2005.

Dr. Yu is a registered Professional Engineer in the province of Ontario, Canada, and a fellow of the Institution of Engineering and Technology. He is a Distinguished Lecturer, the Vice President (Membership), and an elected member of the Board of Governors of the IEEE Vehicular Technology Society. He has served as the technical program committee co-chair for numerous conferences. He serves on the editorial boards of several journals, including the Co-Editor-in-Chief for *Ad Hoc & Sensor Wireless Networks*, and a Lead Series Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and the IEEE COMMUNICATIONS SURVEYS & TUTORIALS.

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 authored over 100 papers in refereed journals and international conferences. His recent research interests include interference alignment, cognitive radio, wireless power transfer, and physical layer security.

Dr. Zhao is a Senior Member of the Chinese Institute of Electronics. He also served as a TPC member for many conferences, such as Globecom, VTC, and WCSP. He received the Top Reviewer Award from the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2016 and was nominated as an Exemplary Reviewer by the IEEE COMMUNICATIONS LETTERS in 2016. He is serving or served on the Editorial Boards of several journals, including the *Journal of Network and Computer Applications*, the IEEE ACCESS, *Wireless Networks*, *Physical Communication*, the *AEU-International Journal of Electronics and Communications*, *Ad Hoc & Sensor Wireless Networks*, and the *KSII Transactions on Internet and Information Systems*.