Dynamic Video Streaming in Caching-enabled Wireless Mobile Networks

C. Liang and S. Hu
Depart. of Systems and Computer Eng., Carleton Univ., Ottawa, ON, Canada

Abstract—Recent advances in software-defined mobile networks (SDMNs), in-network caching, and mobile edge computing (MEC) can have great 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. With the objective of maximizing the mean measurement of video quality, 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 dual-decomposition method to decouple those three sets of variables. 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

The video service is replacing voice and other applications to become the fundamental service in mobile networks [1]–[3]. Moreover, high-definition (HD) videos (e.g., 720p, 1080p, and beyond) that request at least 5-20 Mbps user data rate will become ubiquitous [4], 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) [5], in-network caching [6] and mobile edge computing (MEC) [7].

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 [5]. With the programmability, SDN is considered as a promising candidate to enhance traffic engineering [8]. The success of the utilizing SDMNs for traffic engineering depends critically on our ability to jointly provision the backhaul and radio access networks (RANs) for the traffic [9]–[11].

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 [12]. Particularly, 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 [13]–[16].

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) [7]. 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. 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 [17], [18].

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 paper, we jointly consider Mobile Edge Computing and Caching (MECC)-enabled SDMNs 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. In addition, 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 rest of this paper is organized as follows. Section II introduces the system model and formulates the presented problem. Section III describes the proposed algorithms and the corresponding analysis. Simulation results are discussed in Section IV. Finally, we conclude this study in Section V.

II. 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.

A. Network 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 paper 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. For simplicity, vertical handover [19]–[21] is not considered in this paper. To evaluate the gain of the video quality, the measure of each video quality \( q \) is defined as \( s_q \).

In this paper, we consider a 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 BSs, MBS 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) and users formed a set \( \mathcal{J} \) and users formed a set \( \mathcal{L} \). \( \mathcal{N} \) comprised with wired and wireless links denoted by sets \( \mathcal{L}^{\text{wd}} \) and \( \mathcal{L}^{\text{wt}} \), 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 paper, to simply 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 \( \gamma_l = \log (1 + \frac{g_{m_l}m_l}{\sigma_l}) \), where \( g_{m_l}m_l \) is the large-scale channel gain that includes pathloss and shadowing between the transmission node \( n_l \) (the source of link \( l \)) and the receiving node \( m_l \) (the destination of link \( l \)). We deploy the same model used in [22] to calculate the pathloss and applying shadowing. \( p_{m_l} \) (Watt/Hz) is the normalized transmission power on link \( l \). The fixed equal power allocation mechanism is used, which means transmission power \( p_{m_l} \) is the same for all frequencies. \( \sigma_0 \) is the power spectrum density of additive white Gaussian noise. Accordingly, the achievable data rate capacity of link \( l \) is \( R_l = W\gamma_l \).

Each data flow can be split to multiple paths as the user is assumed to be served by multiple BSs through BS cooperation [10] or multistream carrier aggregation. 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, the data flow of user 1 is split to two paths where one is from content source server to MBS 1 then to users 1 and another is directly from SBS 2.

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 \( f_1 \) demanded by user \( i \) is found at node \( j \), namely \( f_1 \in \mathcal{F}_j \), node \( j \) becomes a candidate source. To indicate a hitting event between user \( i \) and node \( j \), we define \( h_{ij} = 1 \) and \( v_{ij} = v_i \) if \( f_i \in \mathcal{F}_j \). \( v_{ij} = v_i \) means the video requested by user \( i \) can be potentially fully provided by node \( j \) and \( h_{ij} = 1 \) means a successful hitting event. If \( f_i \notin \mathcal{F}_j \), \( h_{ij} = +\infty \) that means there is no hitting event (infinite resource is required to process this video) and \( v_{ij} = 0 \) means node \( j \) cannot response the video request \( i \).

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 [17], 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 20 concurrent video processing tasks are allowed.

B. Problem Formulation

The purpose of the 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, \( x_{qi} = 0 \).

To support video services demanded by users, an optimal path set for all data flows should be found by solving the proposed algorithm. Denote \( P_i \) as a path set including all candidate paths for user \( i \) and \( P_j \) as a subset of \( P_i \) including all candidate paths starting from node \( j \). A path \( p_{ij}^k \in P_i \) 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 \( \sum_{p_{ij}^k \in P_i} r_{ij}^k \) that is the aggregated rate of all selected paths.

As we mentioned in above, the computing resource on each node needs to be scheduled to video transcoding tasks. Thus, we define a binary variable \( y_{ij} \in \{0, 1\} \) as the computing task assignment indicator. If \( y_{ij} = 1 \), node \( j \) is able to transcode the video demanded by user \( i \) to a desired quality level; otherwise, the video data cannot be retrieved from node \( j \). To improve the whole network utility by maximizing the overall mean gain of videos, the SDN controller performs traffic engineering to assist users adaptively selecting optimal video quality levels. Thus, the proposed problem can be formed as follows.

\[
\max_{X,Y} \quad U(X) = \frac{1}{|I|} \sum_{i \in I} \sum_{q=1}^{Q} s_q x_{qi} \tag{1a}
\]

subject to

\[
r_{ij}^k \in \mathbb{R}^+, \forall i, j, k, \tag{1b}
\]

\[
x_{qi}, y_{ij} \in \{0, 1\}, \forall q, i, j, k, \tag{1c}
\]

\[
\sum_{q=1}^{Q} x_{qi} = 1, \forall i \in I, \tag{1d}
\]

\[
\sum_{i \in I} y_{ij} h_{ij} e_i \leq C_j, \forall j \in J. \tag{1d}
\]
where \( \{x_{qi}\}, \{y_{ij}\} \) and \( \{r^{k}_{ij}\} \) are elements of \( \mathbf{X}, \mathbf{Y} \) and \( \mathbf{R} \), respectively. Constraint (1c) reflects that only one resolution level can be selected for one user. The computing capacity on each node \( j \) is specified by the constraint (1d) where \( c_i \) (bps) is the computing resource required for transcoding video \( f_i \). The flow conservation law (FCL) of traffic engineering is claimed by constraints (1e) and (1f) where \( \mathcal{P}_l \) is the set of paths that pass link \( l \). (1e) means the allocated data 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, (1f) enforces that the total allocated spectrum cannot exceed the available spectrum bandwidth. The demand constraint for every video flow is given by (1g). The constraint (1h) requires that any candidate path starting from node \( j \) can be selected only when it has the content in the cache (\( h_{ij} = 1 \)) and the computing resource is assigned to transcode this content (\( y_{ij} = 1 \)) or the highest quality is selected (\( x_{Q_i} = 1 \)).

Unfortunately, problem (1) is difficult to solve and implement. Firstly, the mix integer variables result in the problem a mix-integer linear problem (MILP) that generally is NP-complete. Moreover, video resolution, path selection, and resource scheduling are decided by different layers, network nodes, and perform in different time scales. Lastly, the necessary exchange of local information about the network and links affects the performance as overheads are introduced.

### III. Proposed Scheme

In this section, dual-decomposition method is deployed to simplify problem (1).

This paper 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. 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 obtained.

We firstly define independent local feasible sets \( \Pi_r, \Pi_x, \Pi_y \) 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

\[
\Pi_r = \{ \{ r^{k}_{ij} \} \mid \mathbb{R}^+ \}, \quad (1) \\
\Pi_x = \{ \{ x_{qi} \} \mid \{ 0, 1 \} \}, \quad (2) \\
\Pi_y = \{ \{ y_{ij} \} \mid \{ 0, 1 \} \}. \quad (3)
\]

Fortunately, the coupled constraints are (1g) and (1h). Thus, by relaxing constraints (1g) and (1h) with dual variables \( \{ \mu_i \} \) and \( \{ \lambda_{ij} \} \) the Lagrangian can be shown as:

\[
\max_{\mathbf{X},\mathbf{R},\mathbf{Y}} \quad U(\mathbf{X}) + \sum_{i \in \mathcal{I}} \mu_i \left[ \sum_{p^k_{ij} \in \mathcal{P}_i} r^{k}_{ij} - \sum_{q = 1}^Q v_{qi} x_{qi} \right] \\
- \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \lambda_{ij} \left[ \sum_{p^k_{ij} \in \mathcal{P}_i} r^{k}_{ij} - v_{ij} (x_{Q_i} + y_{ij}) \right]
\]

\[
\text{s.t.} \quad X \in \Pi_x, \quad Y \in \Pi_y, \quad R \in \Pi_r.
\]

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 (23). Accordingly, the dual problem (DP) then is:

\[
\text{DP :} \quad \min_{\mu \in \mathbb{R}^+, \lambda \in \mathbb{R}^+} \quad D(\mu, \lambda) = g_x(\mu, \lambda) + g_r(\mu, \lambda) + g_y(\lambda) \quad (6)
\]

where \( g_x(\mu, \lambda), g_r(\mu, \lambda), \) and \( g_y(\lambda) \) are dual functions obtained as the maximum value of the Lagrangians solved in following problems (7), (8) and (9) for given \( \{ \mu_i \} \) and \( \{ \lambda_{ij} \} \).

\[
g_x(\mu, \lambda) = \sup_{x_{qi} \in \Pi_x} \left\{ \frac{U(\mathbf{X}) - \sum_{i \in \mathcal{I}}^I \mu_i \sum_{q} v_{qi} x_{qi}}{\sum_{i \in \mathcal{I}, j \in \mathcal{J}} \lambda_{ij} v_{ij} x_{Q_i}} \right\}, \quad (7)
\]

\[
g_r(\mu, \lambda) = \sup_{r^{k}_{ij} \in \Pi_r} \left\{ \sum_{i \in \mathcal{I}} \sum_{p^k_{ij} \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) r^{k}_{ij} \right\}, \quad (8)
\]

\[
g_y(\lambda, \lambda) = \sup_{y_{ij} \in \Pi_y} \left\{ \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \lambda_{ij} v_{ij} y_{ij} \right\}. \quad (9)
\]

It is observed that \( D(\mu, \lambda) \) 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 (6). Obviously, a sub-gradient of problem (6) for \( \lambda_{ij} \) is:

\[
z_{ij}^{\lambda} = \sum_{p^k_{ij} \in \mathcal{P}_i} r^{k}_{ij} - v_{ij} (x_{Q_i} + y_{ij}), \quad \text{and for} \ \mu_i \quad z_{i}^{\mu} = \sum_{q \in \mathcal{P}_i} r^{k}_{ij} - \sum_{q} v_{qi} x_{qi}. \quad \text{According to dual decomposition, we thus can update} \ \mu_i \text{ and} \ \lambda_{ij} \text{ based on:}
\]

\[
\mu_i^{[t+1]} = \mu_i^{[t]} - \tau_i^{[t]} z_i^{\mu}, \quad (10)
\]

\[
\lambda_{ij}^{[t+1]} = \lambda_{ij}^{[t]} - \tau_{ij}^{[t]} z_{ij}^{\lambda}, \quad (11)
\]

where \( \tau_i^{[t]} \) and \( \tau_{ij}^{[t]} \) are the length of step at iteration step \( [t] \).

Thus, if we are able to solve the inner problems (7), (8) and (9) 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 \( x_{qi}, r^{k}_{ij} \), and \( y_{ij} \). In the remaining of this section, algorithms will be given to solve problems (7), (8) and (9). Observe that problem (7)
can be decoupled to users where the local problem of each user is shown as

$$\max_{x_{qi} \in \{0,1\}} \sum_{q=1}^{Q} s_q x_{qi} - \mu_i \sum_{q} v_q x_{qi} + \sum_{j \in J} \lambda_{ij} v_{ij} x_{Q_i}$$

s.t. $$\sum_{q=1}^{Q} x_{qi} = 1.$$  \hspace{1cm} (12)

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.

Similar to problem (7), problem (9) also can be decoupled to each node $j$ as follows.

$$\max_{y_{ij} \in \{0,1\}} \sum_{i \in I} \lambda_{ij} v_{ij} y_{ij}$$

s.t. $$\sum_{i \in I} y_{ij} h_{ij} v_{ij} \leq C_j,$$  \hspace{1cm} (13)

Obviously, this is a 0-1 knapsack problem that is usually NP-complete. Thus, BnB method or dynamic programming can be used to solve this problem, but they are both computationally intensive and might not be practical for large-scale problem. Therefore, firstly, to ease the size of this problem so that common methods can be used, we form a set $\mathcal{I}_0^j$ including every user who has non-zero gain $\lambda_{ij} v_{ij}$ or finite $h_{ij}$. Formally, $\mathcal{I}_0^j := \{ i | \lambda_{ij} v_{ij} > 0, h_{ij} < \infty \}$. It is easy to see that $y_{ij} = 0$ if $i \notin \mathcal{I}_0^j$. Thus, we only need to consider users that in $\mathcal{I}_0^j$, which leads to a reduction of the problem size.

This problem (3) 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. Denote the data rate of path $p_{ij}^k$ decided by the wired backhaul network is $r_{ij}^k$ and its peer decided by the RAN is $r_{ij}^{k, l}$. Furthermore, the wired backhaul network can be decoupled to links, as capacities of wired links usually are independent with each other. $r_{ij}^k$ denotes that the data rate of path $p_{ij}^k$ allocated by link $l$. It should be noted that $r_{ij}^{k, l}$ does not mean $p_{ij}^k$ passing link $l$, $r_{ij}^{k, l}$ can be considered the opinion for $p_{ij}^k$ given by link $l$. Actually, all local variables of $r_{ij}^{k, l}$ is the opinion or recommendation from one part of the whole network. By defining $r_{ij}^{k, l}$ and $r_{ij}^k$, problem (8) can be revised as

$$\max_{r_{ij}^k, r_{ij}^{k, l} \in \mathbb{R}^+} \frac{1}{2} \sum_{l \in \mathcal{L}^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^k \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) r_{ij}^{k, l}$$

$$+ \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^k \in \mathcal{P}_i} (\mu_i - \lambda_{ij}) r_{ij}^k$$

s.t. $$\sum_{i \in \mathcal{I}} \sum_{p_{ij}^k \in \mathcal{P}_i} \alpha_{ij}^{k, l} r_{ij}^{k, l} \leq B_1, \forall l \in \mathcal{L}^{wd},$$

$$\sum_{l \in \mathcal{L}^{wd}} \sum_{i \in \mathcal{I}} \sum_{p_{ij}^k \in \mathcal{P}_i} \alpha_{ij}^{k, l} r_{ij}^{k, l} \leq W,$$

$$r_{ij}^{k, l} = r_{ij}^k, \forall i, j, k, l.$$

TABLE I: Network parameters settings

| Network parameters | value |
|--------------------|-------|
| 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 |

where $L^{wd} = |\mathcal{L}^{wd}|$ is the number of wired backhaul links. By ussing dual decomposition method, the equality constraints in (14) can be decoupled to each links or RAN. Thus, problem (14) can be solved locally at each wired link or RAN similar to the method we use to decouple video adaptation and computing.

IV. SIMULATION RESULTS AND DISCUSSIONS

In the simulation, we consider a cellular network, consisting of one MBS, 15 SBSs and 15 active users, that covers a 250m-by-250m area. Transmission with single antenna for both transmitter and receiver is considered in our paper. The remaining simulation parameters are summarized in Table I. We assume that the total 1000 videos are in the video library $\mathcal{F}$. In this paper, to adapt ubiquitous HD videos in next generation mobile networks, we refer the measure to U-vMOS proposed in [3], [4]. Each video $f$ can be encoded to 6 levels with constant bit rate (CBR), and has the same length of 600 seconds.

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. The $f$-th most popular video has a request probability of $(f^{-0.56})/(\sum_{j=1}^{|[\mathcal{F}]|} f^{-0.56})$. The default cache capacity $S_n$ of a MBS is 200 video files, and default cache capacity of a 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.

The MEC computing capability of a MBS is set to 150 Mbps that is equivalent to processing 6 videos simultaneously. Due to limitations of SBSs, the computing performance of a MEC server at SBS is only set to 50 Mbps equivalent to 2 videos.

It is also necessary to discuss the distribution of video resolutions of our proposed scheme. As shown in Fig. I overall, 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 much higher ratio, up to 15%, on the 1080p and higher resolutions compared to cache-only case and no MECC case. Some observations can be made from Fig. I. Firstly, the MECC improves the performance of the network-assist video rate adaptation significantly on 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
servers may not always be the bottleneck of the system. In 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.

V. CONCLUSIONS

In this paper, we jointly studied the 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 mean video measurement 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. Simulation results were presented to show that our proposed scheme can significantly improve the mean video measurement.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and Huawei Technologies Canada CO., LTD.

REFERENCES

[1] 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, June 2014.

[2] 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, July 2016.

[3] D. Schoolar. (2015) Whitepaper: Mobile video requires performance and measurement standards. [Online]. Available: http://www-01c.huawei.com/

[4] 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/4.5GWhitpaper.pdf

[5] 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.

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

[7] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, “White paper: Mobile edge computing: A key technology towards 5G,” ETSI, Tech. Rep., Dec. 2015.

[8] A. Mendiola, J. Astorga, E. Jacob, and M. Higuero, “A survey on the contributions of software-defined networking to traffic engineering,” IEEE Commun. Surveys & Tutorials, vol. PP, no. 99, pp. 1–1, 2016.

[9] H. Farmanbar and H. Zhang, “Traffic engineering for software-defined radio access networks,” in Proc. IEEE Netw. Operations & Management Symp. (NOMS), May 2014, pp. 1–7.

[10] 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, June 2014.

[11] N.-D. Dao, 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, June 2015, pp. 191–196.

[12] 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.

[13] G. Paschos, E. Bahtug, I. Land, G. Caire, and M. Debbah, “Wireless caching: Technical misconceptions and business barriers,” IEEE Comm. Mag., vol. 54, no. 8, pp. 16–22, Aug. 2016.

[14] D. Liu, B. Chen, C. Yang, and A. F. Molisch, “Caching at the wireless edge: design aspects, challenges, and future directions,” IEEE Comm. Mag., vol. 54, no. 9, pp. 22–28, Sept. 2016.

[15] 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, 2016.

[16] 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 ICC’17, Paris, France, June 2017.

[17] H. A. Pedersen and S. Dey, “Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing,” ACM/IEEE Trans. Networking, vol. 24, no. 2, pp. 996–1010, May 2016.

[18] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, “Collaborative multi-bitrate video caching and processing in mobile-edge computing networks,” arXiv preprint arXiv:1612.01436, 2016.

[19] 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.

[20] F. Yu and V. Krishnamurthy, “Optimal joint session admission control in integrated WLAN and CDMA cellular networks with vertical handoff,” IEEE Trans. Mobile Computing, vol. 6, no. 1, pp. 126–139, Jan. 2007.

[21] L. Ma, F. R. Yu, and V. C. M. Leung, “Performance improvements of mobile SCTP in integrated heterogeneous wireless networks,” IEEE Trans. Wireless Commun., vol. 6, no. 10, pp. 3567–3577, Oct. 2007.

[22] 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. 2700–2716, June 2013.

[23] 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.