Multi-tier Caching and Resource Sharing for Video Streaming in 5G Ultra-dense Networks

Nguyen-Son Vo*, Minh-Phung Bui, Phuc Quang Truong, Cheng Yin, and Antonino Masaracchia

***Abstract*—In this paper, we formulate a multi-tier caching and resource sharing (CRS) optimization problem for a high video streaming performance in 5G ultra-dense networks. Particularly, in user tier, the optimal sets of spectrum owning users (SUs) with available downlink resources, caching users (CUs) with cached video versions, and normal users (NUs) requesting the video versions are tripartite for video streaming over device-to-device (D2D) communications. In femto tier, the optimal caching strategy is found to indicate which video versions and where to cache them in the femtocell base station (FBSs). By cooperating with the video versions cached in macro tier, i.e., macro base station (MBS), the mobile users (MUs) can receive the video versions from D2D communications, FBSs, and MBS flexibly. The objective of the CRS solution is to maximize the video playback quality, while saving the caching storage of FBSs and satisfying a given throughput required by MUs and a given target signal to interference plus noise ratio of SUs.**

***Index Terms*—5G Ultra-dense Networks, Multi-tier Caching, Resource Sharing, Video Streaming.**

I. Introduction

In 5G ultra-dense networks (5G UDNs), various video applications and services (VASs) requested by a proliferation of mobile users (MUs) cause an extreme traffic congestion at the backhaul links of macro base stations (MBSs) and small-cell base stations (SBSs) [1]–[4]. Under this perspective, substantial solutions for the network modifications could be necessary to meet the upcoming surge of VASs' demand. Caching techniques can be labeled as the most efficient solution that does not require any system architecture changes [5]. Amongst caching techniques, multi-tier caching can be deployed in the MBSs, SBSs, and MUs to convey the VASs closer to the MUs for streaming [6]–[9]. In this way, the workload at the backhaul links of MBSs and SBSs is mitigated and the quality of experience (QoE) of MUs is improved thanks to proximity communications.

As regards multi-tier caching techniques, the base stations and MUs can cooperate in caching to reduce the content redundancy and delivery delay [6]. The caching techniques are also deployed at the MBSs and SBSs, e.g., femtocell base stations (FBSs), to gain high hit probability and system capacity by finding optimal caching placements [7], [8]. In addition to the studies in [6]–[8], the available downlink spectrum resources of MUs have been exploited in [9] to offload videos locally over device-to-device (D2D) communications. However, the main problems of the approach proposed in [9] are i) the impossibility to pair the caching users (CUs) that have the videos, with the normal users (NUs) that request the videos, and ii) the inability to select the proper video versions to cache in the FBSs, to achieve higher system performance.

In this paper, we propose a multi-tier caching and resource sharing (CRS) solution that enables the MUs to receive the video versions flexibly from D2D communications (user tier), FBSs (femto tier), and MBS (macro tier) in 5G UDNs. The MUs are categorized into three types including spectrum owning users (SUs) that have downlink spectrum resources, the CUs, and the NUs. The NUs are paired with the CUs for D2D video communications by reusing the downlink resources of the SUs. Both SUs and CUs receive the video versions from the FBSs and MBS, while the NUs are able to receive the video versions further from the CUs. Assuming that the MBS has a large storage to cache all video versions, the problem here is that how to efficiently utilize the system resources, i.e., downlink resources of SUs and caching storage resources of MBS, FBSs, and CUs, to satisfy the MUs.

In particular, we formulate a CRS optimization problem and solve it for 1) the optimal set of SUs, CUs, and NUs that is able to establish D2D video communications by sharing the downlink resources of SUs and cached video versions of CUs and 2) the optimal caching strategy that is able to select the proper video versions and the right FBSs in which to cache them. The objective is to serve the MUs various VASs at maximum playback quality. The CRS problem also considers the constraints on caching storage of the FBSs, throughput required by the MUs, and target signal to interference plus noise ratio (SINR) of the SUs, for saving the caching storage, conserving the system throughput, and limiting the interference effect caused by the transmission of the CUs on the SUs. Simulation results demonstrate that the proposed CRS solution outperforms other benchmarks (e.g., caching, sharing, no caching nor sharing, and no video version selection).

II. System Model and Formulations

A. System Model

We consider the CRS model for VASs in 5G UDNs illustrated in Fig. 1. The model has one MBS in the macro tier, J

Nguyen-Son Vo (*Corresponding Author) is with the Institute of Fundamental and Applied Sciences, Duy Tan University, Ho Chi Minh City, 700000, Vietnam, and is also with the Faculty of Electrical-Electronic Engineering, Duy Tan University, Da Nang, 550000, Vietnam (e-mail: vonguyenson@duytan.edu.vn).

Minh-Phung Bui is with Van Lang University, Ho Chi Minh City, 700000, Vietnam (e-mail: buiminhphung@vanlanguni.edu.vn).

Phuc Quang Truong is with HCMC University of Technology and Education, Ho Chi Minh City, 700000, Vietnam (e-mail: phuctq@hcmute.edu.vn).

Cheng Yin and Antonino Masaracchia are with Queen's University Belfast, Belfast BT7 1NN, UK (e-mail: {cyin01, A.Masaracchia}@qub.ac.uk).

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2018.308.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2020.2983408, IEEE Communications Letters

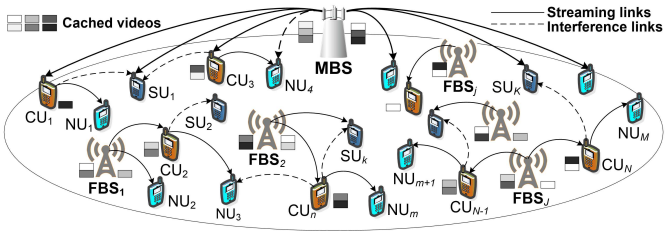

Fig. 1. A CRS model for VASs in 5G UDNs.

FBSs in the femto tier, a number of MUs including K SUs, N CUs, and M NUs in the user tier, and I videos. The video i, $i = 1, 2, ..., I$, has V_i versions encoded at different encoding rates. The video versions are sent from the MBS to the MUs through conventional cellular transmissions, from the FBSs to the MUs by using the channel splitting and F-ALOHA schemes to avoid the interferences [10], and from the CUs to the NUs over D2D communications reusing the downlink resources of SUs. The spatial distribution of CUs sharing the downlink resource from the SU k, $k = 1, 2, ..., K$, is modeled as homogeneous Poisson Point Process with density λ_C^k [11]. We assume that the system parameters remain at least during a streaming session of the longest video version[1]. Whenever the MBS anticipates that there will be an increasing number of video requests, it implements the CRS scheme in three steps: i) updating system parameters, ii) formulating and solving the CRS optimization problem, and iii) delivering the videos.

First, the MBS updates the system parameters such as wireless channel characteristics; caching storage of FBSs and CUs; video information (version, size, and popularity); required throughput of MUs for video playback; and target SINR of SUs. Second, the MBS formulates the constrained CRS optimization problem and solves it for the optimal sharing index $w_{n,m}^k$ and the optimal caching index $u_j^{v_i}$. $w_{n,m}^k$ indicates if the SU k shares its downlink with the CU n to send the video version v_i to the NU m ($w_{n,m}^k = 1$) or not ($w_{n,m}^k = 0$), $n = 1, 2, ..., N$, $m = 1, 2, ..., M$, and $v_i = 1, 2, ..., V_i$. Here the video version v_i is cached in the CU n with probability $p_n^{v_i}$ depending on the remaining storage of the CU and the size and popularity of the video version. $u_j^{v_i}$ indicates if the FBS j caches the video version v_i ($u_j^{v_i} = 1$) or not ($u_j^{v_i} = 0$), $j = 1, 2, ..., J$. Finally, by cooperating with the video versions cached in the MBS, the CRS delivers the video versions flexibly from the MBS, FBSs, and CUs to the MUs.

B. System Formulations

In order to obtain the objective function of the CRS optimization problem, i.e., average quality of received video versions, it is necessary to analyze the capacity success probabilities at the SUs, CUs, and NUs as presented below.

1) Capacity Success Probability at SUs: The SU k receives the video version v_i from the FBS j or the MBS. The capacity success probability at the SU k is derived from its capacities over the channels from the FBS j and the MBS which are given by

$$C_j^{k,v_i} = W \log_2 \Big(1 + u_j^{v_i} P_j G_j^k / N_0\Big), \tag{1}$$

$$C_0^{k,v_i} = W \log_2 \Big(1 + \frac{\prod_{j=1}^{J}(1 - u_j^{v_i}) P_0 G_0^k}{N_0 + I_{C,S}^{k,v_i}}\Big), \tag{2}$$

where W is the system bandwidth; P_j and P_0 are the transmission powers of the FBS j and the MBS; G_j^k and G_0^k are the channel gains from the FBS j and the MBS to the SU k which are Rayleigh fading, independently and identically exponentially distributed with unit mean, and multiplied by a standard power law path loss function with a path loss exponent η; N_0 is the power of additive white Gaussian noise; and $I_{C,S}^{k,v_i} = \sum_{n=1}^{N}\sum_{m=1}^{M} w_{n,m}^k p_n^{v_i} P_C^k G_n^k$, P_C^k is the transmission power of the CUs sharing the downlink from the SU k, G_n^k is the channel gain between the CU n and the SU k, and $p_n^{v_i} = a r_i + b\theta_n^{v_i}$ is the probability that the CU n caches the video version v_i [9], $a, b \in [0, 1]$, $a + b = 1$, $r_i = i^{-\alpha}(\sum_{i=1}^{I} i^{-\alpha})^{-1}$ is Zipf-like distribution [12], and $\theta_n^{v_i} = r_n\big[1 - (L_i^{v_i} - \min\{L_i^{v_i}, v_i = 1, 2, ..., V_i\})/L_i^{v_i})\big]$; here $\alpha \geq 0$ is the skewed popularity among different videos, r_n is the percentage of remaining storage of the CU n, and $L_i^{v_i}$ is the size of video version v_i.

Let $C_{\text{th}}^{v_i}$ be the capacity threshold (measured in Kbps) required to send the video version v_i for playback, the capacity success probabilities are given by [11]

$$p_j^{k,v_i} = \Pr\{C_j^{k,v_i} \geq C_{\text{th}}^{v_i}\} = \exp\Big(-\xi_j^{k,v_i} N_0 / u_j^{v_i} P_j\Big), \tag{3}$$

$$p_0^{k,v_i} = \Pr\{C_0^{k,v_i} \geq C_{\text{th}}^{v_i}\} = \exp\Big\{-\xi_0^{k,v_i}\Big[\lambda_C^{k,v_i}\Big(\frac{P_C^k}{\prod_{j=1}^{J}(1 - u_j^{v_i})P_0}\Big)^{\frac{2}{\eta}}\Big]\Big\}, \tag{4}$$

where $\xi_j^{k,v_i} = (d_j^k)^{\eta}(2^{C_{\text{th}}^{v_i}/W} - 1)$, $\xi_0^{k,v_i} = \pi (d_0^k)^2 \Gamma(1 + \frac{2}{\eta})\Gamma(1 - \frac{2}{\eta})\big(2^{C_{\text{th}}^{v_i}/W} - 1\big)^{2/\eta}$, d_j^k and d_0^k are the distances from the FBS j and the MBS to the SU k, and the density $\lambda_C^{k,v_i} = \sum_{n=1}^{N}\sum_{m=1}^{M} w_{n,m}^k p_n^{v_i}$ within a circular cell area.

The capacity success probability to send the video version v_i from the FBS j and the MBS to the SU k is

$$p_{0,j}^{k,v_i} = 1 - (1 - p_j^{k,v_i})(1 - p_0^{k,v_i}). \tag{5}$$

2) Capacity Success Probability at CUs: Without interference, the capacities at the CU n over the channels from the FBS j and the MBS are expressed as

$$C_{j,n}^{v_i} = W \log_2 \Big(1 + u_j^{v_i} P_j G_j^n / N_0\Big), \tag{6}$$

$$C_{0,n}^{v_i} = W \log_2 \Big(1 + \frac{\prod_{j=1}^{J}(1 - u_j^{v_i}) P_0 G_0^n}{N_0}\Big), \tag{7}$$

where G_j^n and G_0^n are the channel gains from the FBS j and the MBS to the CU n.

We then obtain the corresponding capacity success probabilities at the CU n as follows:

$$p_{j,n}^{v_i} = \Pr\{C_{j,n}^{v_i} \geq C_{\text{th}}^{v_i}\} = \exp\Big(\frac{-\xi_j^{n,v_i} N_0}{u_j^{v_i} P_j}\Big), \tag{8}$$

$$p_{0,n}^{v_i} = \Pr\{C_{0,n}^{v_i} \geq C_{\text{th}}^{v_i}\} = \exp\Big(\frac{-\xi_0^{n,v_i} N_0}{\prod_{j=1}^{J}(1 - u_j^{v_i}) P_0}\Big), \tag{9}$$

[1]Under this assumption, the proposed model efficiently serves the MUs the local VASs in crowded areas such as concert or meeting halls, museums, office buildings, stadiums, hospitals, campuses, etc.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2020.2983408, IEEE Communications Letters

where $\xi_j^{n,v_i} = (d_j^n)^\eta (2^{\frac{C_{\rm th}^{v_i}}{W}} - 1)$ and $\xi_0^{n,v_i} = (d_0^n)^\eta (2^{\frac{C_{\rm th}^{v_i}}{W}} - 1)$, d_j^n and d_0^n are the distances from the FBS j and the MBS to the CU n.

Therefore, the capacity success probability to send the video version v_i from the FBS j and the MBS to the CU n is

$$p_{0,j,n}^{v_i} = 1 - (1 - p_{j,n}^{v_i})(1 - p_{0,n}^{v_i}). \tag{10}$$

3) Capacity Success Probability at NUs: Different from SUs and CUs, the capacities at the NU m, which come from the CUs, FBSs, and MBS, are respectively computed as

$$C_{n,m}^{k,v_i} = W \log_2 \big(1 + \frac{w_{n,m}^k p_n^{v_i} P_C^k G_n^m}{N_0 + P_0 G_0^m + I_{C,C}^{k,v_i}}\big), \tag{11}$$

$$C_{j,m}^{k,v_i} = W \log_2 \big(1 + \frac{u_j^{v_i}(1 - w_{n,m}^k p_n^{v_i}) P_j G_j^m}{N_0}\big), \tag{12}$$

$$C_{0,m}^{k,v_i} = W \log_2 \big(1 + \frac{\prod_{j=1}^{J}(1 - u_j^{v_i})(1 - w_{n,m}^k p_n^{v_i}) P_0 G_0^m}{N_0}\big), \tag{13}$$

where G_n^m, G_j^m, and G_0^m are the channel gains from the CU n, FBS j, and MBS to the NU m, and $I_{C,C}^{k,v_i} = \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} w_{n',m'}^k p_{n'}^{v_i} P_C^k G_{n',m'}$.

So, the corresponding capacity success probabilities are

$$\begin{aligned} p_{n,m}^{k,v_i} &= \Pr\{C_{n,m}^{k,v_i} \geq C_{\rm th}^{v_i}\} \qquad (14)\\ &= \exp\Big\{ - \xi_{n,m}^{v_i} \Big[\lambda_M \Big(\frac{P_0}{w_{n,m}^k p_n^{v_i} P_C^k} \Big)^{\frac{2}{\eta}} + \lambda_C^{'k,v_i} \Big] \Big\}, \end{aligned}$$

$$p_{j,m}^{k,v_i} = \Pr\{C_{j,m}^{k,v_i} \geq C_{\rm th}^{v_i}\} = \exp\Big[\frac{-\xi_{j,m}^{v_i} N_0}{u_j^{v_i}(1 - w_{n,m}^k p_n^{v_i}) P_j} \Big], \tag{15}$$

$$\begin{aligned} p_{0,m}^{k,v_i} &= \Pr\{C_{0,m}^{k,v_i} \geq C_{\rm th}^{v_i}\} \qquad (16)\\ &= \exp\Big[\frac{-\xi_{0,m}^{v_i} N_0}{\prod_{j=1}^{J}(1 - u_j^{v_i})(1 - w_{n,m}^k p_n^{v_i}) P_0} \Big], \end{aligned}$$

where Eq. (14) is similarly computed in [11], $\xi_{n,m}^{v_i} = \pi (d_n^m)^2 \Gamma(1 + \frac{2}{\eta}) \Gamma(1 - \frac{2}{\eta}) \big(2^{\frac{C_{\rm th}^{v_i}}{W}} - 1\big)^{2/\eta}$, $\xi_{j,m}^{v_i} = (d_j^m)^\eta (2^{\frac{C_{\rm th}^{v_i}}{W}} - 1)$, and $\xi_{0,m}^{v_i} = (d_0^m)^\eta (2^{\frac{C_{\rm th}^{v_i}}{W}} - 1)$; d_n^m, d_j^m, and d_0^m are the distances from the CU n, the FBS j, and the MBS to the NU m; and the density $\lambda_M = 1$ (i.e., one MBS) and $\lambda_C^{'k,v_i} = \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} w_{n',m'}^k p_{n'}^{v_i}$ within a circular cell area.

Finally, the capacity success probability at the NU m is

$$p_{0,j,n,m}^{k,v_i} = 1 - (1 - p_{n,m}^{k,v_i})(1 - p_{j,m}^{k,v_i})(1 - p_{0,m}^{k,v_i}). \tag{17}$$

4) Average Quality of Received Videos: To obtain the average quality of received videos, we apply the rate-distortion (RD) model expressed by a decaying exponential function [13] which represents the experimental RD characteristics of the videos encoded by high efficiency video coding standard [14]. In this model, if the video version v_i is played back at rate or capacity $C_{\rm th}^{v_i}$, the corresponding reconstructed distortion is

$$D_i(C_{\rm th}^{v_i}) = \gamma_i (C_{\rm th}^{v_i})^{\beta_i}, \tag{18}$$

where γ_i and β_i are the sequence-dependent parameters selected so that Eq. (18) meets the experimental RD curves.

Based on the aforementioned capacity success probabilities in (5), (10), and (17), we can compute the overall average quality values of received videos at the MUs (i.e., SUs, CUs, and NUs) as below

$$\overline{Q} = \frac{1}{3J} \sum_{j=1}^{J} (Q_S^j + Q_C^j + Q_N^j), \tag{19}$$

where

$$Q_S^j = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} p_{0,j}^{k,v_i} Q\big(D_i(C_{\rm th}^{v_i})\big), \tag{20}$$

$$Q_C^j = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} p_{0,j,n}^{v_i} Q\big(D_i(C_{\rm th}^{v_i})\big), \tag{21}$$

$$Q_N^j = \frac{1}{KMN} \sum_{k=1}^{K} \sum_{N=1}^{N} \sum_{m=1}^{M} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} p_{0,j,n,m}^{k,v_i} Q\big(D_i(C_{\rm th}^{v_i})\big), \tag{22}$$

where $Q\big(D_i(C_{\rm th}^{v_i})\big) = 10 \log_{10} \frac{255^2}{D_i(C_{\rm th}^{v_i})}$ is the peak signal-to-noise ratio (PSNR) measured in dB.

III. CRS Optimization Problem and Solution

So far, we have (19) as the objective function of the CRS optimization problem. We further compute the total storage of FBSs (L_F^i) used to cache all versions of the video i and the total throughput required by MUs (C) which are considered in the constraints of the CRS optimization problem as below

$$L_F^i = \sum_{j=1}^{J} \sum_{v_i=1}^{V_i} u_j^{v_i} L_i^{v_i}, \tag{23}$$

$$C = C_S + C_C + C_N, \tag{24}$$

where

$$C_S = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} (C_j^{k,v_i} + C_0^{k,v_i}), \tag{25}$$

$$C_C = \sum_{j=1}^{J} \sum_{n=1}^{N} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} (C_{j,n}^{v_i} + C_{0,n}^{v_i}), \tag{26}$$

$$C_N = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{N=1}^{N} \sum_{m=1}^{M} \sum_{i=1}^{I} \frac{r_i}{V_i} \sum_{v_i=1}^{V_i} (C_{n,m}^{k,v_i} + C_{j,m}^{k,v_i} + C_{0,m}^{k,v_i}). \tag{27}$$

Finally, let $L_{\max}^i$, C^*, and γ_0 be the constraints on the caching storage of FBSs for saving, the throughput required by MUs for conserving, and the target SINR for guaranteeing the quality of service (QoS) of SUs respectively, the CRS optimization problem is formulated as

$$\max_{u_j^{v_i}, w_{n,m}^k} \ \overline{Q} \tag{28a}$$

$$\text{s.t.} \sum_{v_i=1}^{V_i} u_j^{v_i} \leq 1, i = 1, 2, ..., I, j = 1, 2, ..., J, \quad (28b)$$

$$\sum_{m=1}^{M} w_{n,m}^{k} \leq 1, k = 1, 2, ..., K, n = 1, 2, ..., N, \quad (28c)$$

$$\sum_{n=1}^{N} w_{n,m}^{k} \leq 1, k = 1, 2, ..., K, m = 1, 2, ..., M, \quad (28d)$$

$$L_F^i \leq \mu L_{\max}^i, i = 1, 2, ..., I, \quad (28e)$$

$$C \leq \delta C^*, \quad (28f)$$

$$I_{C,S}^{k,v_i} \leq \frac{P_0 G_0^k}{\gamma_0} - N_0, k = 1, 2, ..., K, i = 1, 2, ..., I, \quad (28g)$$

$$v_i = 1, 2, ..., V_i.$$

where the constraint (28b) is to limit the FBS j to cache up to one proper version of video i. The constraints (28c) and (28d) are to guarantee that the CU n is able to transmit to up to one NU and the NU m is able to receive from up to one CU. $0 < \mu \leq 1$ in (28e) and $0 < \delta \leq 1$ in (28f) are to adjust $L_{\max}^i$ and C^* for the convenience of evaluation, here $L_{\max}^i = r_i \times J \times I \times \max\{L_i^{v_i}, v_i = 1, 2, ..., V_i\}$. The last constraint (28g), which comes from (2), is to limit the interference effect of D2D communications on the QoS of the SU k.

To solve (28), exhaustive binary searching algorithm [9] is applied to finding $u_j^{v_i}$ and $w_{n,m}^k$. Although memory requirement and time complexity of the exhaustive binary searching algorithm (especially in the large scale of 5G UDNs) are higher than other methods, e.g., stochastic global searching methods like genetic algorithms for exact or approximated optimal results, it is acceptable due to the two following reasons. First, it is the simple method to find the global optimal results. Second, the proposed CRS solution is efficiently used to serve the MUs the local VASs in crowded areas, in which the system characteristics change mainly depending on the arrival of MUs in local areas following a daily pattern with the stability on an hourly basis [15]. This pattern gives us the opportunity to relax the strict requirement of time complexity to serve the MUs, while the memory requirement can be powerfully handled at the MBS. In addition, to ensure that the SU k can share its downlink resource with only one pair of CU n and NU m, the searching algorithm follows the rule that if $w_{n,m}^k = 1$, then $w_{n,m}^{k'} = 0, k' = 1, 2, ..., K, k' \neq k$.

IV. Performance Evaluation

The CSR model is deployed with the parameters set in Table I. Assuming that the MBS covers the circular cell area within the radius of 1500m and the relative distances between the MBS and the MUs, the FBSs and the MUs, the CUs and the SUs, and the CUs and the NUs, are randomly distributed in the ranges of [500, 1500]m, [20, 100]m, [50, 150]m, and [1, 20]m, respectively. In addition, we take into account 3 videos, i.e., $V_1^{v_1}$ (Basketballpass), $V_2^{v_2}$ (Racehorses), and $V_3^{v_3}$ (Foreman), to analyze their experimental RD curves by using HM reference software version 12.0 [16] and obtain $L_i^{v_i}$, $C_{th}^{v_i}$, γ_i, and β_i given in Table I. To evaluate the performance of CRS, we compare it to the other three schemes including only caching (OCC), only resource sharing (ORS), and no caching nor resource sharing (NCS).

Fig. 2 shows the performance of CRS, OCC, ORS, and NCS versus the skewed popularity exponent α. The results indicate that the system gains higher performance when α increases. The reason is that increasing α makes the popularity more skewed amongst the videos, and thus less videos are with higher popularity and more videos are with low popularity. In this context, the system focuses on serving the MUs the videos with higher popularity rather than the ones with lower popularity to gain higher performance. The proposed CRS outperforms the other OCC, ORS, and NCS thanks to the joint solution of caching and resource sharing techniques. The OCC is mostly better than the ORS because it is likely to provide more possibilities of caching and transmitting over better channels than the ORS. The NCS is the worst case due to without CRS assisted.

In Fig. 3, we investigate the effects of the constraints on the performance of CRS. To do so, we further consider three cases including 1) CRS_1: decreasing the caching storage of FBSs by changing μ in (28e) from 0.5 to 0.3, 2) CRS_2: decreasing the required throughput of MUs by changing δ in (28f) from 1 to 0.5, and 3) CRS_3: increasing the target SINR of SUs by changing γ_0 in (28g) from 3dB to 5dB. Particularly, in case of CRS_1, the system has less chances to cache and thus provides lower playback quality than that in case of CRS. In CRS_2, the required throughput of MUs decreases due to lower MUs' playback resolutions, the channels that provide higher throughput are not selected for streaming to save the system bandwidth resource. In other words, the system serves the MUs lower quality (compared to the CRS) to meet the

TABLE I
Parameters Setting

Symbols	Specifications
J, K, N, M	3 FBSs, 3 SUs, 4 CUs, 5 NUs
I, V_i	3 videos, each has 3 versions
P_0, P_j, P_C^k, N_0	5W, 1W, 10^{-3}W, 10^{-12}W
W	5MHz
η	4
γ_0	3dB
a, b, μ, δ	0.5, 0.5, 0.5, 1
r_n	Fixed to 1, i.e. all CUs have 100% of storage to cache
$L_i^{v_i}$	[11,867 23,734 35,600; 198,680 264,906 351,000; 33,382 66,763 113,496]Kbit
$C_{\text{th}}^{v_i}$	[1,000 2,000 3,000; 3,000 4,000 5,300; 50,000 100,000 170,000]Kbps
$\{\gamma_i; \beta_i\}$	{9,806 76,520 1,644,000;-0.9972 -1.1530 -1.0920}
C^*	12Gbps, i.e., each MU is served at 1Gbps

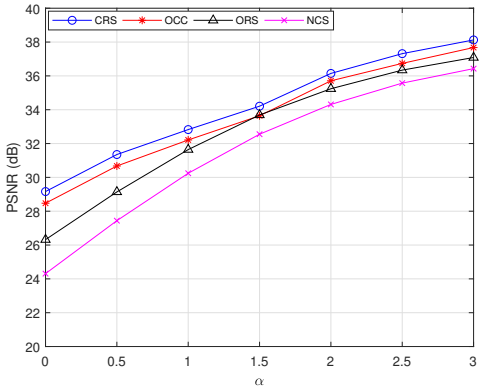

Fig. 2. Performance of CRS, OCC, ORS, and NCS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2020.2983408, IEEE Communications Letters

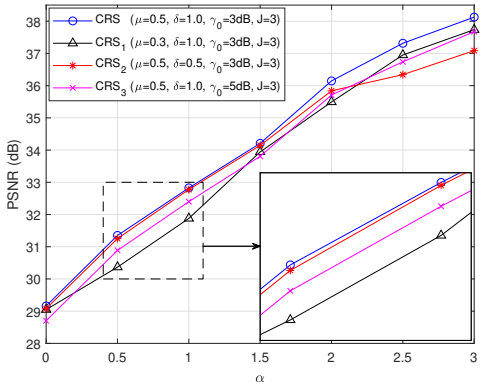

Fig. 3. Performance of CRS versus different system parameters.

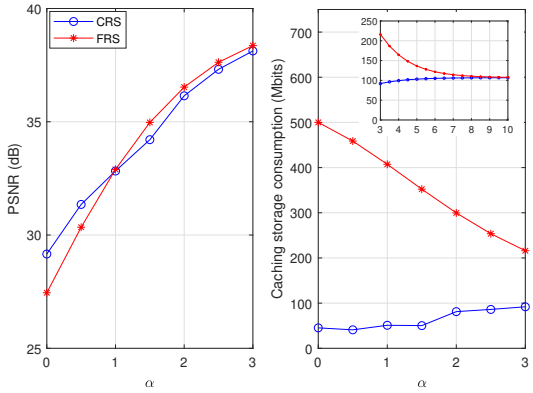

Fig. 4. Performance of CRS and FRS.

playback resolutions of the MUs. In CRS$_3$, when the target SINR γ_0 increases to ensure higher QoS for the SUs, the number of D2D communication sessions is reduced to make less interference impact on the SUs. This in turn reduces the system performance because it cannot exploit the D2D communications for video streaming in close proximity.

To investigate the performance of CRS in terms of playback quality and caching storage consumed in the FBSs, we compare the CRS to the full rate caching and resource sharing scheme (FRS). In FRS, we keep the resource sharing scheme applied while the FBSs always cache the video versions with the highest encoding rates, i.e. no video version selection, instead of selecting proper video versions in CRS. As shown in Fig. 4, the FRS outperforms the CRS when $\alpha > 1$. The reason is that when α gets higher values, the system focuses on serving the video versions with higher popularity. In this context, while the performance of CRS is limited due to the caching storage and required throughput constraints, these constraints are relaxed in the FRS. In consideration of caching storage consumption, the FRS clearly uses higher caching storage than the CRS does. The caching storage consumption of both CRS and FRS converges on a certain value with respect to the increase of α because the proper video versions in CRS and the video versions with the highest encoding rates in FRS are the same. To make clear about the convergence, a further result of caching storage consumption of CRS and FRS versus α extended from 3 to 10 is shown on the top right corner of the Fig. 4.

V. Conclusion

We have proposed the CRS solution that can exploit not only the caching storage of the MBS, FBSs, and MUs but also the downlink resources of the MUs for advanced video streaming applications and services (VASs) in 5G UDNs. Furthermore, the CRS takes into account the process of D2D paring and video version selection as well as the constraints on caching storage of the FBSs, required throughput of the MUs, and target SINR of the SUs, so as to provide the MUs with high performance of advanced VASs. Simulation results are shown to demonstrate the benefits of the CRS compared to other schemes, i.e., caching, resource sharing, no caching nor resource sharing, and no video version selection, in terms of video playback quality and caching storage consumption.

References

[1] T. H. Nguyen, D. Q. Nguyen, and V. D. Nguyen, "Quality of service provisioning for D2D users in heterogeneous networks," *EAI Trans. Industrial Netw. and Intelligent Syst.*, vol. 6, no. 21, pp. 1–7, Oct. 2019.

[2] M. T. Nguyen, "An energy-efficient framework for multimedia data routing in Internet of things (IoTs)," *EAI Trans. Industrial Netw. and Intelligent Syst.*, vol. 6, no. 19, pp. 1–8, Jun. 2019.

[3] H. T. Nguyen *et al.*, "Collaborative multicast beamforming for content delivery by cache-enabled ultra dense networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3396–3406, May 2019.

[4] N.-S. Vo, T. Q. Duong, and M. Guizani, "QoE-oriented resource efficiency for 5G two-tier cellular networks: A femtocaching framework," in *Proc. IEEE Global Commun. Conf.*, Washington, DC, Dec. 2016, pp. 1–6.

[5] H. S. Goian *et al.*, "Popularity-based video caching techniques for cache-enabled networks: a survey," *IEEE Access*, vol. 7, pp. 27 699–27 719, Mar. 2019.

[6] P. Lin, K. S. Khan, Q. Song, and A. Jamalipour, "Caching in heterogeneous ultradense 5G networks: A comprehensive cooperation approach," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 22–32, 2019.

[7] J. Wen, K. Huang, S. Yang, and V. O. K. Li, "Cache-enabled heterogeneous cellular networks: optimal tier-level content placement," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5939–5952, Sep. 2017.

[8] X. Li *et al.*, "Collaborative multi-tier caching in heterogeneous networks: modeling, analysis, and design," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6939, Aug. 2017.

[9] N.-S. Vo, T. Q. Duong, M. Guizani, and A. Kortun, "5G optimized caching and downlink resource sharing for smart cities," *IEEE Access*, vol. 6, pp. 31 457–31 468, May 2018.

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

[11] A. Bhardwaj and S. Agnihotri, "Energy- and spectral-efficiency trade-off for D2D-multicasts in underlay cellular networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 546–549, Aug. 2018.

[12] L. Breslau *et al.*, "Web caching and Zipf-like distributions: evidence and implications," *IEEE INFOCOM*, vol. 1, pp. 126–134, 1999.

[13] W. Xiang *et al.*, "Forward error correction-based 2-D layered multiple description coding for error-resilient H.264 SVC video transmission," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1730–1738, Dec. 2009.

[14] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[15] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proc. of ACM EuroSyst.*, Leuven, Belgium, Apr. 2006, p. 333–344.

[16] *"HM Reference Software Version 12.0."*. [Online]. Available: https://hevc.hhi.fraunhofer.de.