Multi-rate Selection and Power Allocation Assisted Probabilistic Edge Caching for Cooperative Video Transmission in Dense D2D Networks

Thuong C. Lam, Nguyen-Son Vo*, Viet V. Lam, Trang Hoang, Minh-Phung Bui, and Trung Q. Duong

***Abstract*—In this paper, we consider a cooperative transmission model for video applications and services (VASs) in dense device-to-device (D2D) networks. The model enables the mobile users (MUs) to flexibly receive the videos from macro base station (MBS) and D2D networks with mobile edge caching. Particularly, we formulate a multi-rate selection and power allocation assisted probabilistic edge caching (MPC) optimisation problem under the resource constraints on storage, bandwidth, and power. This problem is solved for the optimal caching probabilities of requested videos corresponding to proper encoding rates selected. The optimal powers of caching MUs and MBS for transmitting the videos are also found to maximise the playback quality, while utilising the system resources. The MPC optimisation problem, which is complicated due to the presence of binary and real variables and various constraints, is feasibly solved by genetic algorithms (GA) with penalty function and truncated chromosome. Simulation results are shown to demonstrate the benefits of both GA and MPC methods compared to other benchmarks. Detailed analyses and interesting findings provide useful insights into the mobile edge caching design of dense D2D networks for VASs.**

***Keywords*—Cooperative video transmission, dense D2D networks, multi-rate encoding, power allocation, probabilistic edge caching.**

I. Introduction

A. General Context

According to the startling statistics reported by Statista Research Department, it is anticipated that in 2033, there will be about 32 billion Internet of Things (IoT) devices making the global data traffic exponentially increased [1]. In the 6G era, wireless networks will be equipped with an ever powerful capability based on communication, sensing, computing, and intelligence [2], [3]. 6G technologies for IoT (6G-enabled IoT) can provide not only new spectrum utilising THz bands; extreme massive connectivity boosting the latency to sub-millisecond level, the density to 10 times higher than 5G, and the peak and experienced rates to Tbit/s and Gbit/s, respectively; but also emerging network architecture integrating with non-terrestrial networks of low-earth orbit (LEO) and very LEO satellites, digital twins, and deep edge intelligence [4]–[7].

In 6G-enabled IoT networks, numerous mobile users (MUs) request the high data rate traffic of video applications and services (VASs), e.g., 360^o videos, video gaming, 4K/8K/12K video streaming, virtual reality, augmented reality, and mixed reality. The VASs, which occupy up to 79% of mobile data traffic [8], are therefore the main focus of the Internet service providers (ISPs) and content providers (CTPs). The proliferation of demand for VASs requires commercial platforms based on emerging 6G technologies to provide the MUs with greater visual and realistic sensations. While the system resources are inherently limited, the numerous MUs requesting high data rate of VASs for greater visual and realistic sensations may cause a serious congestion situation at the backhaul links of 6G networks.

In this context, the most technical challenge to both ISPs and CTPs is how to provide the MUs with high user-perceived quality (UPQ) of VASs at a low cost and in a flexible and quick deployment [9]. It is certain that developing high-speed backhaul links with new network architectures and technologies is costly. Alternately, optimisation designs for VASs can be utilised flexibly and quickly by placing a higher priority on software-defined edge (SDE) techniques [10]. This way, the ISPs and CTPs can further satisfy the high demand of MUs with better UPQ and conserve the system resources (storage, throughput/bandwidth, power, and spectrum), but do not introduce any changes to the network architectures and technologies, thereby remaining low cost.

B. Related Works

One of the most efficient SDE techniques, which has drawn a huge amount of attention from academic and industrial circles, is edge caching. Edge caching technique provides the MUs with local VASs in vicinity for high UPQ. Typical edge

Thuong C. Lam is with the HUTECH Institute of Engineering, HUTECH University, Ho Chi Minh City 70000, Vietnam (e-mail: lc.thuong@hutech.edu.vn).

*Corresponding author: Nguyen-Son Vo is with the Institute of Fundamental and Applied Sciences, Duy Tan University, Ho Chi Minh City, 70000, Vietnam, and also with the Faculty of Electrical-Electronic Engineering, Duy Tan University, Da Nang, 50000, Vietnam (e-mail: vonguyenson@duytan.edu.vn).

Viet V. Lam and Trang Hoang are with Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City 70000, Vietnam (e-mails: {lvviet.sdh212, hoangtrang}@hcmut.edu.vn).

Minh-Phung Bui is with the Faculty of Information Technology, School of Technology, Van Lang University, Ho Chi Minh City 70000, Vietnam (e-mail: phung.bm@vlu.edu.vn).

Trung Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada, and also with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mails: tduong@mun.ca, trung.q.duong@qub.ac.uk).

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2023.25.

Corresponding author is Nguyen-Son Vo.

TABLE I. COMPARISON OF THE MOST IMPORTANT RELATED WORKS.

Ref.	Size	Rate awareness (RA)	Resource allocation	Transmission mode	Objectives
[23]	Different	RA with matching	N/A	D2D with fixed clustering radius	Playback quality (dB)
[24]	Different	RA with matching	N/A	D2D with optimal clustering radius	Playback quality (dB)
[25]	Different	N/A	N/A	Cooperative	Delay
[26]	Different	RA with matching	N/A	D2D	Delay
[27]	Different	RA without matching	N/A	D2D	Playback quality (dB)
[28]	Different	RA without matching	N/A	D2D	System utility
[29]	Different	N/A	PA	D2D	Capacity
[30]	Different	N/A	N/A	Cooperative	Offloading probability
[31]	Different	N/A	PA	Cooperative	Capacity
[32]	Different	RA without matching	N/A	Cooperative	Cache hit
Our work	Different	RA with matching	PA	Cooperative	Playback quality (dB)

caching designs fall into deterministic caching and probabilistic caching. Deterministic caching deploys the contents in stable placements like terrestrial base stations [11]. Meanwhile, probabilistic caching prefers to deploy in dynamic and mobile placements, e.g., device-to-device (D2D) networks, where the MUs can join/leave the caching networks randomly [12]. Thanks to the ease of deployment in dynamic and mobile environments, probabilistic edge caching for content offloading over D2D networks has attracted increasing research interest in the literature [12]–[20].

The probabilistic edge caching in D2D networks is either purely to find the optimal caching policy or caching probability (CP) [12], [13], [19], [21], [22] or to be assisted by other solutions, e.g., spectrum management, channel access control, cooperative transmission with mode selection, clustering, and energy harvesting [14]–[18], [20]. However, these studies mostly consider the quality of service (QoS) based performance metrics, for example, cache hit probability, delay, offloading gain, and cache-added throughput and successful transmission probability for common contents, rather than visual UPQ for VASs [23]. In addition, the works [12]–[20] induce four major concerns about designing an edge caching network including: 1) contents with the same unit size, 2) regardless of matching the capacity of system (represented by the throughput) to the statistical playback resolution of mobile devices, and thus not efficiently applying the rate-distortion model (RDM) to multi-rate encoding adaption/selection for caching, 3) no power allocation (PA), and 4) without the role of macro base station (MBS) for cooperative transmission.

It is certain that the lack of addressing all the aforementioned major concerns makes the conventional methods not take full advantages of system resources, characteristics of videos, diverse capacities of mobile devices, and additional techniques to satisfy the MUs. In fact, there also have a number of studies on probabilistic edge caching in D2D networks addressing some of the concerns such as our previous work in [23], [24] and the other ones in [25]–[32]. The detailed comparison of these most important related works is presented in Table I. We can see that none of them completely address the four major concerns to gain the highest visual UPQ of VASs and efficiently utilise the system resources.

C. Main Contributions and Organisation

In this paper, we fully address all aforementioned major concerns by which the main contributions are summarised as below.

- We propose a multi-rate selection and power allocation assisted probabilistic edge caching (MPC) method for cooperative video transmission in dense D2D networks to satisfy the MUs and benefit the ISPs and CTPs by utilising the storage, throughput, and power resources, without any changes to the network architectures and technologies.
- The MPC method applies not only the multi-rate probabilistic edge caching (MRC) [23] but also the PA to enhance the performance of MRC. Particularly, we deploy a two-mode PA for transmitting the videos by 1) the mobile devices in D2D offloading (DOL) mode of MRC and 2) the MBS in normal cellular transmission (NCT) mode. And then, the MRC and the NCT are synthesised into a cooperative (COP) mode in which an MPC optimisation problem is formulated and solved for the optimal results of each video, i.e., CP, multi-rate encoding selection (RS), and two-mode PA, so as to maximise the playback quality of VASs.
- We modify the genetic algorithms (GA) to feasibly solve the MPC optimisation problem with respect to binary variables for the RS and real variables for the CP and two-mode PA, under complicated constraints on the system resources of storage, throughput, and power.
- Simulation results demonstrate that the GA is feasible and the proposed MPC method outperforms the other conventional ones. In addition, we present the detailed analyses and interesting findings which provide valuable insights into the mobile edge caching design of dense D2D networks for VASs.

The rest of this paper is organised as follows. In Section II, we propose a system model of cooperative video transmission in dense D2D networks with MPC method, introduce the primary notations, definitions, and concepts of the system model, and then describe how it works. Section III is dedicated to deriving the system formulations which enable us to present the MPC optimisation problem and the GA solution in Section IV. The performance evaluations of GA and proposed MPC method in comparison with other conventional ones are shown in Section V. Finally, Section VI concludes the paper.

II. SYSTEM MODEL

The system model of cooperative video transmission in dense D2D networks with MPC method is illustrated in Fig. 1.

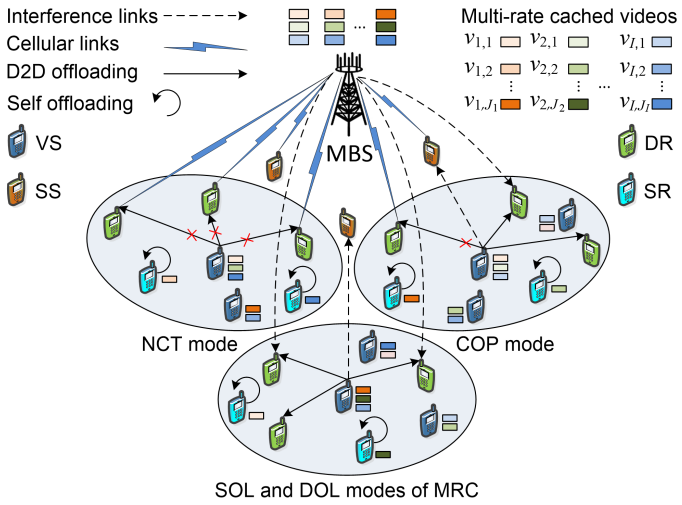

Fig. 1. System model for cooperative video transmission in dense D2D networks with MPC method.

TABLE II. NOTATIONS

Symbols	Descriptions
I	Number of videos
J_i	Number of encoding rates (versions) of video i, $i = 1, 2, ..., I$
$R_{i,j}$	Encoding rate (Kbps) of the version j of video i (namely $v_{i,j}$), $j = 1, 2, ..., J_i$
$S_{i,j}$	Size (Mbits) of $v_{i,j}$
$D_{i,j}$	Reconstructed distortion of $v_{i,j}$ encoded at rate $R_{i,j}$, measured in mean squared error, by following rate-distortion model (RDM)
$w_{i,j}$	Multi-rate encoding selection (RS) index, $w_{i,j} = 1$ means that the $v_{i,j}$ is selected for caching, otherwise $w_{i,j} = 0$
ρ	Probability that an MU serves as a sharing user (SU), otherwise $(1-\rho)$ is the probability that an MU serves as a requesting user (RU)
q_i	Probability that an SU serves as a video sharing user (VS) who caches the video i, otherwise $(1-q_i)$ is the probability that an SU serves as a spectrum sharing user (SS) who shares its downlink spectrum resource for D2D offloading (DOL)
$\lambda_{\rm MU}$	Density of MUs
r_i	Access rate or popularity of video i following Zipf-like distribution
α	Skewed popularity coefficient among different videos
$R_{\rm DOL}$	DOL cluster radius
$R_{\rm NCT}$	Transmission radius covered by MBS
W	System bandwidth
$P_{\rm M}$	Transmission power of MBS to the SSs who share their downlink spectrum resources to establish the DOL mode
$P_i^{\rm NCT}$	Transmission power of MBS to the DOL requesting users (DRs) in NCT mode for video i
$P_i^{\rm DOL}$	Transmission power of mobile devices in DOL mode for video i
N_0	Power of additive white Gaussian noise
η	Path loss exponent

The main notations used for the system are described in Table II. The system includes one MBS, a dense number of MUs, and I videos. We define ρ or $(1-\rho)$ as the probability that an MU is probable to be a sharing user (SU) or a requesting user (RU). An SU can act as a video sharing user (VS) who caches the video i with probability q_i or a spectrum sharing user (SS) who shares its downlink spectrum resource with probability $(1-q_i)$ for DOL mode, $i = 1, 2, ..., I$. There also have two types of RUs that are 1) the RU requesting the video i but it has already cached this requested video, and thus it is served by itself, i.e., self-offloading requesting user (SR), and 2) the RU requesting the video i and it has not cached this one yet, thus it, namely DOL requesting user (DR), is served by either the VSs via DOL mode or the MBS via NCT mode.

Assume that the spatial distribution of MUs is formulated by a homogeneous Poisson point process (PPP) $\Phi_{\rm MU}$ with density $\lambda_{\rm MU}$, the distributions of VSs, SSs, SRs, and DRs also follow the homogeneous PPPs $\Phi_i^{\rm VS}$, $\Phi_i^{\rm SS}$, $\Phi_i^{\rm SR}$, and $\Phi_i^{\rm DR}$ with densities $\lambda_i^{\rm VS}$, $\lambda_i^{\rm SS}$, $\lambda_i^{\rm SR}$, and $\lambda_i^{\rm DR}$, respectively given by

$$\lambda_i^{\rm VS} = \rho q_i \lambda_{\rm MU}, \tag{1}$$
$$\lambda_i^{\rm SS} = \rho(1-q_i)\lambda_{\rm MU}, \tag{2}$$
$$\lambda_i^{\rm SR} = (1-\rho)q_i\lambda_{\rm MU}, \tag{3}$$
$$\lambda_i^{\rm DR} = (1-\rho)(1-q_i)\lambda_{\rm MU}. \tag{4}$$

To meet the diverse playback resolutions of mobile devices, the video i is encoded into J_i rates representing different quality versions, each, namely $v_{i,j}$, $j = 1, 2, ..., J_i$, is of rate $R_{i,j}$ (Kbps) and of size $S_{i,j}$ (Mbits). Furthermore, the video i has its own popularity r_i following Zipf-like distribution [33], [34], given by

$$r_i = \frac{i^{-\alpha}}{\sum_{i=1}^{I} i^{-\alpha}}, \tag{5}$$

where $\alpha \geq 0$, which is the skewed popularity coefficient among different videos, reflects the popularity pattern of videos in accordance with the access behavior of MUs. For example, $\alpha = 0$ represents a special pattern when all the videos are of equal popularity of $1/I$. Meanwhile, increasing α makes the popularity pattern more skewed between the higher popular videos and the lower popular ones.

In this system, given the spatial distributions of all types of MUs, the popularity pattern of videos, the caching storage limit, the system throughput requirement depending on the diverse playback resolutions of mobile devices, and the transmission power limit, the problem is how to serve the RUs the highest playback quality of the received videos. To do so, the MBS is in charge of collecting the system parameters to formulate the MPC optimisation problem. The MPC optimisation problem is then solved for the optimal results of 1) CP (q_i), 2) RS index ($w_{i,j} \in \{0, 1\}$), and 3) two-mode PA, i.e., powers transmitted by the VSs in DOL mode ($P_i^{\rm DOL}$) and by the MBS in NCT mode ($P_i^{\rm NCT}$), for the video i. Here, $w_{i,j} = 1$ means that $v_{i,j}$ is selected for caching with probability q_i. Otherwise, if $w_{i,j} = 0$, it is not cached with probability $(1-q_i)$. Then, given a particular density of MUs, the MBS finds the optimal cluster radius to group the MUs into different clusters. Finally, when the videos are requested by the RUs in any cluster, the system serves the RUs by the COP mode between MRC and NCT as follows:

- MRC: The RUs are served by either themselves, namely self-offloading (SOL) mode, or the VSs in DOL mode. In SOL mode, the RUs are served by themselves because they have cached the requested videos. Meanwhile, in DOL mode, the RUs have not cached the requested videos yet, but at least one of the nearby VSs within a given DOL cluster radius has already cached them. The videos are offloaded from the VSs to the RUs via D2D communications which are established by reusing the same downlink spectrum resources of the SSs.

- NCT: In case the channels for DOL mode cannot be established since the achievable rate probability is lower than a pre-defined threshold, the RUs are served by the MBS via NCT mode.

III. System Formulations

In this section, we derive the formulations of cooperative video transmission system in dense D2D networks with MPC method. These formulations enable us to compute the objective function and the system resource constraints on caching storage, throughput, and power of the MPC optimisation problem.

A. Cache, Spectrum Sharing, and Requesting Hit Probabilities

As aforementioned, there are four types of MUs including VSs, SSs, SRs and DRs, each type has its own hit probability. The VSs provide cache hit probability which consists of self-cache hit and D2D-cache hit. The self-cache hit probability is simply defined by q_i, meanwhile the D2D-cache hit probability is the probability that the video i is cached by at least one VS within a given DOL cluster radius $R_{\rm DOL}$. To compute the D2D-cache hit probability, we follow the homogeneous PPP $\Phi_{\rm MU}$ with density $\lambda_{\rm MU}$ in which the prerequisite probability of having N MUs within $R_{\rm DOL}$ is expressed as [12]

$$\Pr(N, R_{\rm DOL}, \lambda_{\rm MU}) = \frac{(\pi R_{\rm DOL}^2 \lambda_{\rm MU})^N}{N!} e^{-\pi R_{\rm DOL}^2 \lambda_{\rm MU}}. \quad (6)$$

Based on (6), the D2D-cache hit probability is given by

$$p_i^{\rm VS} = 1 - \Pr(N = 0, R_{\rm DOL}, \lambda_i^{\rm VS}) = 1 - e^{-\pi R_{\rm DOL}^2 \lambda_i^{\rm VS}}. \quad (7)$$

Similar to (7), the SUs covered by the MBS within a given radius $R_{\rm NCT}$ provide spectrum sharing hit probability, expressed as

$$p_i^{\rm SS} = 1 - e^{-\pi R_{\rm NCT}^2 \lambda_i^{\rm SS}}, \quad (8)$$

and the SRs and DRs provide SOL and DOL requesting hit probabilities within $R_{\rm DOL}$ which are respectively given by

$$p_i^{\rm SR} = 1 - e^{-\pi R_{\rm DOL}^2 \lambda_i^{\rm SR}}, \quad (9)$$

$$p_i^{\rm DR} = 1 - e^{-\pi R_{\rm DOL}^2 \lambda_i^{\rm DR}}. \quad (10)$$

B. Achievable Rate Probabilities

1) Achievable Rate Probability in MRC: In SOL mode, the SRs request the video i, and then they are served by themselves at rate $R_{i,j}$. In this case, obviously, the achievable rate probability of SOL mode is $p_{i,j}^{\rm SOL} = 1$. Meanwhile, in DOL mode, the video i is transmitted from the VS n ($n \in \Phi_i^{\rm VS}$) to the DR m ($m \in \Phi_i^{\rm DR}$) over D2D communications. The capacity of the wireless channel from the VS n to the DR m used for transmitting the video i is given by

$$C_{n,m}^{(i)} = W \log_2 \big(1 + \frac{P_i^{\rm DOL} g_{n,m}}{N_0 + P_{\rm M} g_{{\rm M},m} + I_i}\big), \quad (11)$$

where W is the system bandwidth; $P_i^{\rm DOL}$, which is optimally found, is the power consumption for transmitting the video i in DOL mode; $P_{\rm M}$ is the power consumption of the MBS transmitted to the SSs who share their downlink spectrum resources to establish the DOL mode between a pair of VS n and DR m; N_0 is the power of additive white Gaussian noise; and by utilising the underlay D2D communications, the DR m suffers from the interference I_i caused by the other pairs who reuse the same downlink spectrum resource to transmit the video i, expressed as

$$I_i = \sum\nolimits_{k \in \Phi_i^{\rm VS} \setminus \{n\}} \sum\nolimits_{l \in \Phi_i^{\rm DR} \setminus \{m\}} P_i^{\rm DOL} g_{k,l}, \quad (12)$$

and the channel gain $g_{{\rm x},m}$, ${\rm x} \in \{\text{VS } n, \text{MBS identified by M}\}$, which is defined as a Rayleigh fading coefficient following an independent and identical exponential distribution with unit mean $h_{{\rm x},m}$ and a standard power law path loss model with exponent η, is given by

$$g_{{\rm x},m} = h_{{\rm x},m} \parallel d_{{\rm x},m} \parallel^{-\eta}, \quad (13)$$

where $\parallel . \parallel$ is the Euclidean norm and $d_{{\rm x},m}$ is the distance from x to the DR m.

In each cluster covered by the circular radius $R_{\rm DOL}$, we assume that the VS nearby (or at) the center is in charge of transmitting the video i. In the worst case, the video i is transmitted over the longest distance of $d_{n,m} = R_{\rm DOL}$. So, the achievable rate probability to transmit the $v_{i,j}$ in DOL mode is the probability that the capacity $C_{n,m}^{(i)}$ is greater than or equal to the rate $R_{i,j}$, given by [11], [35]

$$p_{i,j}^{\rm DOL} = \Pr\{C_{n,m}^{(i)} \geq R_{i,j}\} = {\rm e}^{-\xi_{i,j}^{\rm DOL}\left[\lambda_{\rm NCT}\left(\frac{P_{\rm M}}{P_i^{\rm DOL}}\right)^{\frac{2}{\eta}} + \lambda_i\right]}, \quad (14)$$

where $\lambda_{\rm NCT}$ is the density of the MBS, $\xi_{i,j}^{\rm DOL}$ and λ_i are respectively expressed as

$$\xi_{i,j}^{\rm DOL} = \pi (R_{\rm DOL})^2 \Gamma(1 + \frac{2}{\eta}) \Gamma(1 - \frac{2}{\eta}) (2^{\frac{R_{i,j}}{W}} - 1)^{2/\eta}, \quad (15)$$

and

$$\lambda_i = \max\big\{ \min\{\lambda_i^{\rm VS}, \lambda_i^{\rm DR}\} - \lambda_i^{\rm SS}, 0\big\}. \quad (16)$$

In (16), λ_i is the density of the ones that cause the interference when transmitting the video i.

2) Achievable Rate Probability in NCT: In NCT mode, the DRs are served by the MBS because the achievable rate probability in DOL mode (14) does not hold. To derive the achievable rate probability in NCT mode, it is necessary to compute the wireless channel capacity from the MBS to the DR m for transmitting the video i, given by

$$C_{{\rm M},m}^{(i)} = W \log_2 \big(1 + \frac{P_i^{\rm NCT} g_{{\rm M},m}}{N_0}\big), \quad (17)$$

where $P_i^{\rm NCT}$ is the power of the MBS used to transmit the video i.

Based on (17), we can easily derive the achievable rate probability in the worst case of NCT mode, i.e., $d_{{\rm M},m} = R_{\rm NCT}$, as below

$$p_{i,j}^{\rm NCT} = \Pr\{C_{{\rm M},m}^{(i)} \geq R_{i,j}\} = {\rm e}^{-\xi_{i,j}^{\rm NCT}\left(\frac{N_0}{P_i^{\rm NCT}}\right)}, \quad (18)$$

where $\xi_{i,j}^{\mathrm{NCT}}$ is computed as

$$\xi_{i,j}^{\mathrm{NCT}} = (R_{\mathrm{NCT}})^{\eta}(2^{\frac{R_{i,j}}{W}} - 1). \tag{19}$$

C. Average Playback Quality

To derive the average playback quality of received videos, it is prerequisite to compute the reconstructed distortion $D_{i,j}$ given the rate $R_{i,j}$ of $v_{i,j}$. By following [11], the relationship between $R_{i,j}$ and $D_{i,j}$, namely RDM, is modelled by applying a decaying exponential function, expressed as

$$D_{i,j} = \gamma_i (R_{i,j})^{\beta_i}. \tag{20}$$

To obtain (20), the video i is analysed to generate its practical RDM. Then, we select the values of γ_i and β_i that make (20) agreeable to the practical RDM. For the visual evaluation of UPQ, the quality of $v_{i,j}$ is measured in dB represented by peak signal-to-noise ratio. So, the reconstructed distortion, which is measured in mean squared error is converted into dB by using the expression below

$$\mathcal{Q}_{i,j} = 10 \log_{10} \big(\frac{L^2}{D_{i,j}}\big), \tag{21}$$

where L is the range of the values that an encoded pixel takes.

So far, by utilising the homogeneous PPPs, the RDM and popularity pattern of videos; the cache, spectrum sharing and requesting hit probabilities; and the achievable rate probabilities analysed in the previous sections, we can derive the overall average playback quality of the videos received via the COP mode, i.e., the cooperative transmission between the SOL and DOL modes of MRC and the NCT mode, as below

$$\overline{\mathcal{Q}} = \sum_{i=1}^{I} r_i \Bigg[\underbrace{q_i p_i^{\mathrm{SR}} \sum_{j=1}^{J_i} w_{i,j} \mathcal{Q}_{i,j}}_{\text{SOL mode}} + \underbrace{(1-q_i) p_i^{\mathrm{DR}} \sum_{j=1}^{J_i} p_{i,j}^{\mathrm{CT}} w_{i,j} \mathcal{Q}_{i,j}}_{\text{DOL mode \& NCT mode}} \Bigg], \tag{22}$$

where $w_{i,j}$ is the binary RS index added to make a decision on selecting the rate $R_{i,j}$ associated with the quality $\mathcal{Q}_{i,j}$ of $v_{i,j}$ ($w_{i,j} = 1$) or not ($w_{i,j} = 0$), and $p_{i,j}^{\mathrm{CT}}$ is the successful cooperative transmission (CT) probability computed as

$$p_{i,j}^{\mathrm{CT}} = p_{i,j}^{\mathrm{SUD}} + (1 - p_{i,j}^{\mathrm{SUD}}) p_{i,j}^{\mathrm{NCT}}, \tag{23}$$

and in (23), $p_{i,j}^{\mathrm{SUD}}$ is the successful probability of DOL mode given by

$$p_{i,j}^{\mathrm{SUD}} = p_i^{\mathrm{VS}} p_i^{\mathrm{SS}} p_{i,j}^{\mathrm{DOL}}. \tag{24}$$

It is noted that (22) is the objective function of the MPC optimisation problem which is maximised by finding $w_{i,j}$, q_i, P_i^{DOL}, and P_i^{NCT}, regarding the system resources consumption given in the sequel.

D. System Resources Consumption

In this paper, we utilise three types of system resources consumed including caching storage, throughput, and power. These resources are computed and then limited in the constraints of the MPC optimisation problem. First, the caching storage consumption can be computed on average as below

$$\overline{S} = \sum_{i=1}^{I} \big[q_i p_i^{\mathrm{SR}} + (1-q_i) p_i^{\mathrm{DR}} p_i^{\mathrm{VS}} \big] \sum\nolimits_{j=1}^{J_i} w_{i,j} S_{i,j}. \tag{25}$$

Then, the average throughput consumption, which depends on the demand of SRs and DRs, is computed as

$$\overline{R} = \sum_{i=1}^{I} r_i \Bigg[q_i p_i^{\mathrm{SR}} \sum_{j=1}^{J_i} w_{i,j} R_{i,j} + (1-q_i) p_i^{\mathrm{DR}} \sum_{j=1}^{J_i} p_{i,j}^{\mathrm{CT}} w_{i,j} R_{i,j} \Bigg]. \tag{26}$$

Finally, the average power consumption is separated into DOL power consumption and NCT power consumption, respectively computed as

$$\overline{P}_{\mathrm{DOL}} = \sum_{i=1}^{I} (1-q_i) p_i^{\mathrm{DR}} p_i^{\mathrm{VS}} p_i^{\mathrm{SS}} P_i^{\mathrm{DOL}}, \tag{27}$$

and

$$\overline{P}_{\mathrm{NCT}} = \sum_{i=1}^{I} (1-q_i) p_i^{\mathrm{DR}} (1 - p_i^{\mathrm{VS}} p_i^{\mathrm{SS}}) P_i^{\mathrm{NCT}}. \tag{28}$$

IV. MPC Optimisation Problem with GA Solution

To maximise the overall average playback quality (22) under the constraints on the caching storage, throughput, and power resources consumption given in (25), (26), (27), and (28), the MPC optimisation problem is formulated as below

$$\max_{w_{i,j}, q_i, P_i^{\mathrm{DOL}}, P_i^{\mathrm{NCT}}} \overline{\mathcal{Q}}, \tag{29a}$$

$$\text{s.t. } 0 \le q_i \le 1, \forall i, \tag{29b}$$

$$\textstyle\sum_{j=1}^{J_i} w_{i,j} \le 1, \forall i, \tag{29c}$$

$$\overline{S} \le S, \tag{29d}$$

$$\overline{R} \le R, \tag{29e}$$

$$\overline{P}_{\mathrm{DOL}} \le P_{\mathrm{DOL}}, \tag{29f}$$

$$\overline{P}_{\mathrm{NCT}} \le P_{\mathrm{NCT}}, \tag{29g}$$

where the constraint (29c) is to guarantee that the video i is cached by selecting only one proper version $v_{i,j}$ encoded at rate $R_{i,j}$, S is to limit the storage consumed for caching, R is to limit the system throughput consumption in accordance with the statistics on the diverse playback resolution of mobile devices, and P_{DOL} and P_{NCT} are to limit the transmission power consumption for DOL mode and NCT mode, respectively.

Regarding the GA [36] which is applied to solve the constrained MPC optimisation problem (29), it is able to work

with only simple constraints like (29b), but not with complicated ones such as (29c), (29d), (29e), (29f), and (29g). This shortage of GA can be overcome by using penalty function [34]. To do so, the constrained MPC optimisation problem is first converted into an unconstrained one by rewriting the complicated constraints as below

$$\begin{cases} \Delta w_i = 1 - \sum_{j=1}^{J_i} w_{i,j} \geq 0, \forall i, \\ \Delta S = S - \overline{S} \geq 0, \\ \Delta R = R - \overline{R} \geq 0, \\ \Delta P_{\mathrm{DOL}} = P_{\mathrm{DOL}} - \overline{P}_{\mathrm{DOL}} \geq 0, \\ \Delta P_{\mathrm{NCT}} = P_{\mathrm{NCT}} - \overline{P}_{\mathrm{NCT}} \geq 0. \end{cases} \tag{30}$$

And then, the penalty function is derived as

$$\begin{aligned} F =& \lambda_1 \sum_{i=1}^{I} \big(\min\{0, \Delta w_i\}\big)^2 + \lambda_2 \big(\min\{0, \Delta S\}\big)^2 + \\ & \lambda_3 \big(\min\{0, \Delta R\}\big)^2 + \lambda_4 \big(\min\{0, \Delta P_{\mathrm{DOL}}\}\big)^2 + \\ & \lambda_5 \big(\min\{0, \Delta P_{\mathrm{NCT}}\}\big)^2, \end{aligned} \tag{31}$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the constraint violation degrees used to punish the candidate solutions (namely individuals) that do not satisfy the constraints.

Taking (31) into account, the unconstrained MPC optimisation problem is expressed as

$$\max_{w_{i,j}, q_i, P_i^{\mathrm{DOL}}, P_i^{\mathrm{NCT}}} \overline{\mathcal{Q}}_F = \overline{\mathcal{Q}} - F. \tag{32}$$

We can see that the unconstrained MPC optimisation problem is with respect to both binary variables of RS and real variables of CP and PA. This problem can be solved by GA [34], but the diversity and the complexity of one binary variable matrix ($w_{i,j}$) and three real variable vectors (q_i, P_i^{DOL}, P_i^{NCT}) lead to the result that each individual is represented by a very long chromosome (string). A population of long strings requires an extremely large number of individuals, and thus draining much more time and memory to make GA converged accurately and stably. To overcome this problem, we divide the original string of each individual into four substrings, namely GA with truncated strings (GTS) method. This way, each substring associated with a particular variable is executed by a different set of GA's operators and parameters depending on the characteristic of the variable. As a result, the GTS converges more accurately and stably compared to the implementation of GA with a population of very long strings, namely GA with full strings (GFS). The detailed GTS for solving the unconstrained MPC optimisation problem is presented in Algorithm 1. It is noted in Algorithm 1 that TC is the termination condition of the GTS. The GTS terminates if it satisfies either $Gen = N_G$ or $F \leq 10^{-3}$ together with $\overline{\mathcal{Q}}_F$ kept unchanged in 10 continuous generations.

Algorithm 1 GTS for MPC Optimisation Problem

Input: Parameters of system and GTS in Table III and Table IV
TC: Termination condition
$Gen = 1$: Generation count
Output: $\overline{Q}_F^*\big(w_{i,j}^*, q_i^*, P_i^{\mathrm{DOL},*}, P_i^{\mathrm{NCT},*}\big)$

1: Generate N_P strings randomly, each has (N_{B}+3×N_{R}×B_{P}) bits, namely $\{X_{\mathrm{BR}}^{(z)}\}$, to represent the individual z, $z = 1, 2, ..., N_P$
2: Separate the string $\{X_{\mathrm{BR}}^{(z)}\}$ into 1) the substring $\{X_1^{(z)}\}$ of N_{B} bits and 2) the substrings $\{X_2^{(z)}\}$, $\{X_3^{(z)}\}$, and $\{X_4^{(z)}\}$, each of $N_{\mathrm{R}} \times B_{\mathrm{P}}$ bits
3: Map $\{X_1^{(z)}\}$ into the $I \times J$ binary matrix for $w_{i,j}^{(z)}$ by using b2m operation and map $\{X_2^{(z)}\}$, $\{X_3^{(z)}\}$, and $\{X_4^{(z)}\}$ into the three real vectors for $q_i^{(z)}$, $P_i^{\mathrm{DOL},(z)}$, $P_i^{\mathrm{NCT},(z)}$ respectively by using b2r operation [34], [36], where $J = \max\{J_i\}$ and it is certain that the redundant $(J - J_i)$ bits do not effect the optimal results
4: Compute N_P fitness values based on (32) to have $\overline{Q}_F\big(w_{i,j}^{(z)}, q_i^{(z)}, P_i^{\mathrm{DOL},(z)}, P_i^{\mathrm{NCT},(z)}\big)$
5: **while** TC does not hold **do**
6: Put $\{X_{\mathrm{BR}}^{(z)}\}$ = [$\{X_1^{(z)}\}$, $\{X_2^{(z)}\}$, $\{X_3^{(z)}\}$, $\{X_4^{(z)}\}$] associated with the fitness values $\overline{Q}_F\big(w_{i,j}^{(z)}, q_i^{(z)}, P_i^{\mathrm{DOL},(z)}, P_i^{\mathrm{NCT},(z)}\big)$ into the mating pool for ranking
7: Select $N_{\mathrm{PG}} = N_{\mathrm{P}} \times P_{\mathrm{G}}$ individuals with higher fitness values for breeding by using stochastic universal sampling operator [36], so as to obtain $\{X_{\mathrm{BR}}^{(t)}\}$ = [$\{X_1^{(t)}\}$, $\{X_2^{(t)}\}$, $\{X_3^{(t)}\}$, $\{X_4^{(t)}\}$], $t = 1, 2, ..., N_{\mathrm{PG}}$
8: Choose a pair of parents to create the offsprings by using single point crossover with probability P_{CB} for $\{X_1^{(t)}\}$ and multiple point crossover with probability P_{CR} for $\{X_2^{(t)}\}$, $\{X_3^{(t)}\}$, and $\{X_4^{(t)}\}$
9: Mutate the offsprings $\{X_1^{(t)}\}$ with probability P_{MB} and the offsprings $\{X_2^{(t)}\}$, $\{X_3^{(t)}\}$, and $\{X_4^{(t)}\}$ with probability P_{MR}. This way, the positive genetic features that have been probably lost in the previous steps can be recovered. As a result, we obtain $\{X_{\mathrm{BR}}^{(t),*}\}$ = [$\{X_1^{(t),*}\}$, $\{X_2^{(t),*}\}$, $\{X_3^{(t),*}\}$, $\{X_4^{(t),*}\}$]
10: Repeat the step 3 and step 4 to obtain $\overline{Q}_F\big(w_{i,j}^{(t),*}, q_i^{(t),*}, P_i^{\mathrm{DOL},(t),*}, P_i^{\mathrm{NCT},(t),*}\big)$
11: Reinsert $\{X_{\mathrm{BR}}^{(t),*}\}$ and $\overline{Q}_F\big(w_{i,j}^{(t),*}, q_i^{(t),*}, P_i^{\mathrm{DOL},(t),*}, P_i^{\mathrm{NCT},(t),*}\big)$ into the present generation to have the new sets of $\{X_{\mathrm{BR}}^{(z)}\}$ and $\overline{Q}_F\big(w_{i,j}^{(z)}, q_i^{(z)}, P_i^{\mathrm{DOL},(z)}, P_i^{\mathrm{NCT},(z)}\big)$
12: $Gen = Gen + 1$
13: **end while**
14: Find the highest fitness value $\overline{Q}_F^*\big(w_{i,j}^*, q_i^*, P_i^{\mathrm{DOL},*}, P_i^{\mathrm{NCT},*}\big) \in \overline{Q}_F\big(w_{i,j}^{(z)}, q_i^{(z)}, P_i^{\mathrm{DOL},(z)}, P_i^{\mathrm{NCT},(z)}\big)$ in the last generation

V. Performance Evaluation

A. Parameters Setting

In this paper, the parameters of system and GTS are detailed in Table III and Table IV. We analyse five well-known videos including Basketballdrill, Basketballpass, Foreman, Traffic, and Racehorses to get their practical RDMs by using HM reference software version 12.0 [37]. Then, the five pairs of (γ_i, β_i) that make (20) agreeable to the practical RDMs are computed. Each video is encoded into three rates, i.e., full rate, average rate, and lowest rate corresponding to highest quality, average quality, and lowest quality. The values of (γ_i, β_i), $R_{i,j}$, and $S_{i,j}$ of the five videos are presented in [38]. The GTS is deployed with $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} = \{10^2, 10^{-1}, 10^{-3}, 10^6, 10^4\}$ selected by applying the method given in [39]. It is noted that the mutation probabilities can be ignored, i.e., $P_{\mathrm{MB}} = 0$ and $P_{\mathrm{MR}} = 0$. However, in the MPC optimisation problem, the mutation operator slightly impacts

TABLE III. SYSTEM'S PARAMETERS

Symbols	Specifications
I	5 videos
J_i	3 versions, $\forall i$
R_{NCT}	250 m
R_{DOL}	15 m
λ_{NCT}	$1/(\pi R_{\mathrm{NCT}}^2)$
λ_{MU}	0.01 MUs/m^2
ρ	0.3
η	4
α	1
N_0	10^{-9} W
P_{M}	10 W
W	5 MHz
S	1 Gbits
R	15 Mbps
L	255

TABLE IV. GTS'S PARAMETERS

Symbols	Specifications
N_{P}	15,000 individuals, i.e., population size
N_{B}	$I \times J$ bits in the binary matrix for $w_{i,j}$, $J = \max\{J_i\}$
N_{R}	I variables for each real vector q_i, P_i^{DOL}, or P_i^{NCT}
B_{P}	12 bits representing a real variable
N_{G}	100 generations
P_{G}	0.9, generation gap
P_{CB}	0.6, crossover probability for binary matrices
P_{CR}	0.6, crossover probability for real vectors
P_{MB}	10^{-12}, mutation probability for binary matrices
P_{MR}	10^{-12}, mutation probability for real vectors

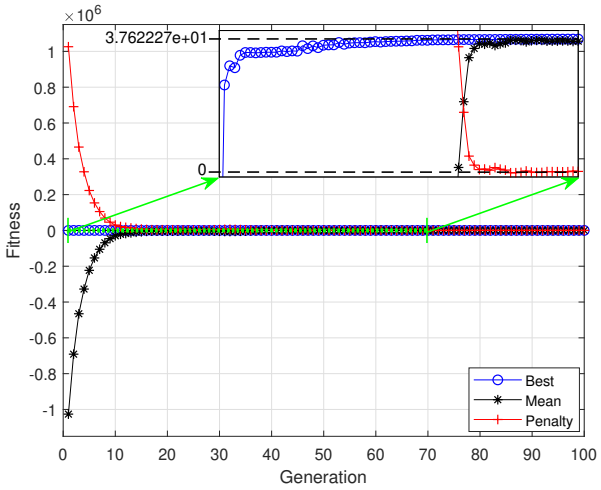

Fig. 2. Convergence rate of GTS.

the convergence of GA, and thus we decrease P_{MB} and P_{MR} started from 10^{-3} until the convergence becomes stable at 10^{-12}.

B. GA Evaluation

The convergence rate of GTS is evaluated by deploying the Algorithm 1 within 100 generations as shown in Fig. 2. In each generation, we find the best fitness value (Best) and the mean fitness value (Mean) with respect to all individuals. Furthermore, to grasp the punishment progress over the generations, the average penalty value (Penalty) for all individuals is computed. The results in Fig. 2 indicate that the GTS starts to converge from the 50th generation. In convergence situation, the Mean gets closer to the Best meaning that all individuals become the best one characterised by the best genetic features. Meanwhile, the Penalty goes to zero to ensure no punishments, i.e., all the constraints are satisfied ($F = 0$).

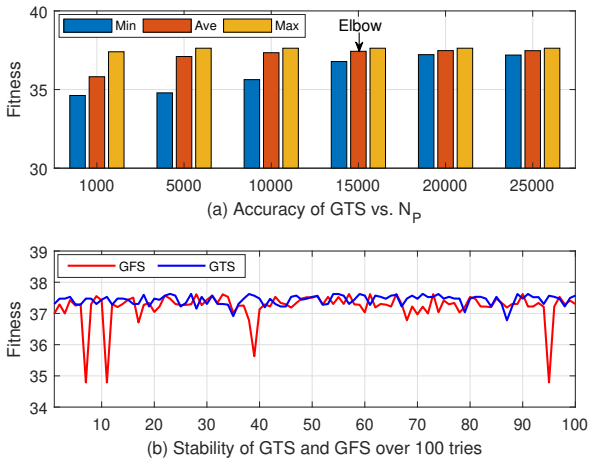

Fig. 3. Accuracy and stability of GTS.

We further evaluate the accuracy and the stability of GTS as illustrated in Fig. 3. In Fig. 3(a), the accuracy of GTS is investigated versus different population sizes by changing N_P from 1,000 to 25,000. For each population size, the Algorithm 1 is repeated 100 tries to yield a set of 100 approximate or exact optimal results, and then, we compute the maximum (Max), average (Ave), and minimum (Min) values of this set. Fig. 3(a) shows that the higher population size we deploy, the higher accuracy of optimal results we obtain but obviously requiring higher time and memory complexity. The proper population size is 15,000 happening at the elbow point of the Ave. From the elbow point, even though we increase the population size, the Ave is not significantly improved but introducing much more time and memory. It is noted that the Max is equivalent to the exact optimal result held if the population size is large enough ($N_P = 5,000$). Regarding the stability, we compare the GTS with the GFS done at $N_P = 15,000$. By repeating 100 tries for both GTS and GFS, we can see in Fig. 3(b) that the GTS is more stable than the GFS is. To make the GFS more stable, it requires a larger population size which must be greater than 15,000 individuals. The benefit of GTS is that it can provide a flexible selection of crossover and mutation features, i.e., operations and probabilities, depending on the different characteristics of optimal variables. In addition, the accuracy and stability analyses confirm that GTS can escape local convergence and reach the global optimum. The GTS is therefore a feasible solution for the complicated MPC optimisation problem.

C. MPC Evaluation

To evaluate the MPC performance, we compare it to the other four methods including 1) multi-rate probabilistic

TABLE V. FEATURES OF DIFFERENT METHODS.

Methods \ Features	CP	RS	CT	PA	S	R	P_{DOL}	P_{NCT}
MPC	✓	✓	✓	✓	1 Gbits	15 Mbps	P^{**}_{DOL}	P^{*}_{NCT}
MRC [23]	✓	✓			1 Gbits	15 Mbps	No	No
MCM [32]	✓	✓	✓		1 Gbits	15 Mbps	P^{*}_{DOL}	No
MPC-FR1 [31]	✓		✓	✓	2 Gbits	20 Mbps	$2\times P^{**}_{\mathrm{DOL}}$	$2\times P^{*}_{\mathrm{NCT}}$
MPC-FR2 [31]	✓		✓	✓	2 Gbits	25 Mbps	$2\times P^{**}_{\mathrm{DOL}}$	$2\times P^{*}_{\mathrm{NCT}}$

Note:
1) P^{*}_{DOL} is the actual DOL power consumption of MRC [23] computed by using (27).
2) P^{**}_{DOL} is the actual DOL power consumption of MCM computed by using (27).
3) P^{*}_{NCT} is the actual NCT power consumption of MCM computed by using (28).
4) CP = Caching probability, RS = Multi-rate encoding selection, CT = Cooperative transmission, PA = Power allocation.

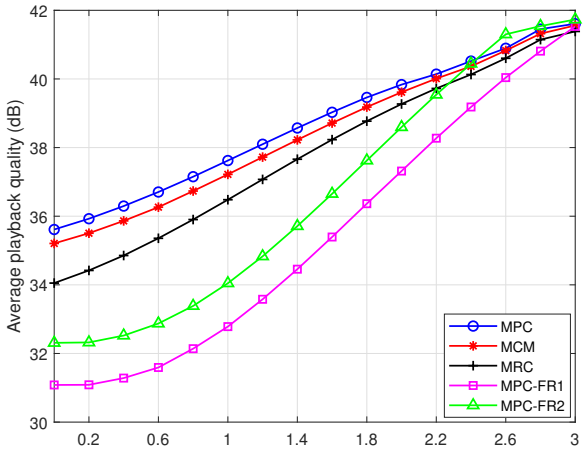

Fig. 4. Playback quality vs. α.

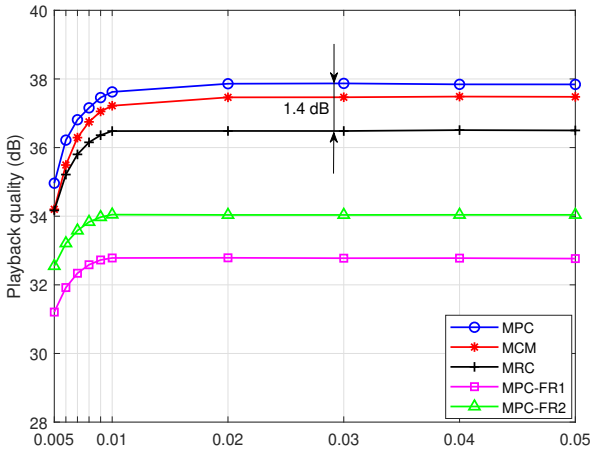

Fig. 5. Playback quality vs. λ_{MU} (MUs/m^2).

caching (MRC) [23], 2) MRC cooperated with MBS (MCM) via NCT mode, 3) MPC with fixed maximum encoding rate and $R = 20$ Mbps (MPC-FR1), and 4) MPC with fixed maximum encoding rate and $R = 25$ Mbps (MPC-FR2). The summary features of MPC, MRC, MCM, MPC-FR1, and MPC-FR2 are listed in Table V. Particularly, in MRC, the multi-rate probabilistic caching for SOL and DOL modes is deployed. In MCM, the MRC cooperates with the MBS via NCT mode to serve the RUs. Meanwhile, the MPC-FR1 and the MPC-FR2 are deployed similarly to the MPC but we force them to cache the fixed and full encoding rate, i.e., always caching the highest quality version. Furthermore, we relax 1) the storage constraint S from 1 Gbits to 2 Gbits and 2) the throughput constraint R from 15 Mbps to 20 Mbps for MPC-FR1 and to 25 Mbps for MPC-FR2. In addition, concerning power consumption constraints, there are no power consumption constraints with MRC, but we can compute the DOL power consumption of MRC by using (27), namely P^{*}_{DOL}. Then, the detailed power consumption constraints (P_{DOL} and P_{NCT}) of MPC, MCM, MPC-FR1, and MPC-FR2 are also listed in Table V. The performance of MPC is evaluated with respect to the playback quality and resource consumption presented below.

1) Playback quality: We first evaluate the playback quality performance of MPC, MRC, MCM, MPC-FR1, and MPC-FR2 versus α as shown in Fig. 4. In comparison, the MPC outperforms all the other methods because it fully takes advantages of CP, RS, CT, and PA to serve the RUs. The MCM is better than the MRC thanks to the cooperative transmission assisted by the MBS via NCT mode. Regarding MPC-FR1 and MPC-FR2, the videos are cached at full rates for the highest resolutions leading to the fact that the wireless channels are not capable of transmitting them well, i.e., low achievable rate probabilities. The playback quality of both MPC-FR1 and MPC-FR2 is therefore lower than that of MPC, MRC, and MCM. However, it is noted that when α is high enough ($\alpha > 2.4$ in Fig. 4), MPC-FR2 becomes better than MPC. The reason is that MPC-FR2 benefits from the caching storage and throughput constraints relaxed to 2 Gbits and 25 Mbps, respectively, while serving the RUs a fewer number of popular videos when α is high. In this case, it is certain that both MPC-R1 and MPC-FR2 consume a huge amount of storage and throughput resources compared to MPC, MRC, and MCM, as shown in Fig. 9.

The playback quality performance is then evaluated versus the density of MUs (λ_{MU}) as shown in Fig. 5. The increase

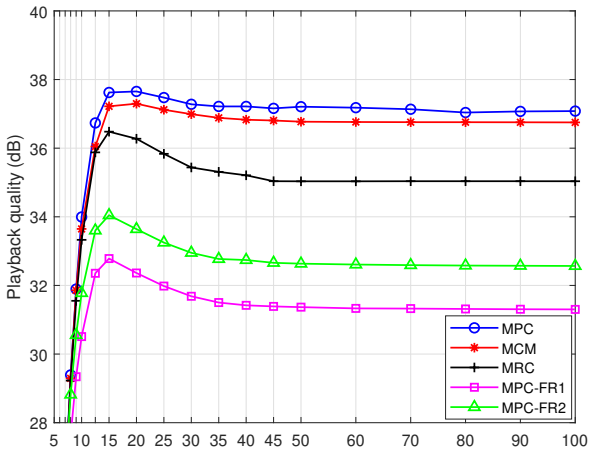

Fig. 6. Playback quality vs. R_{DOL} (m).

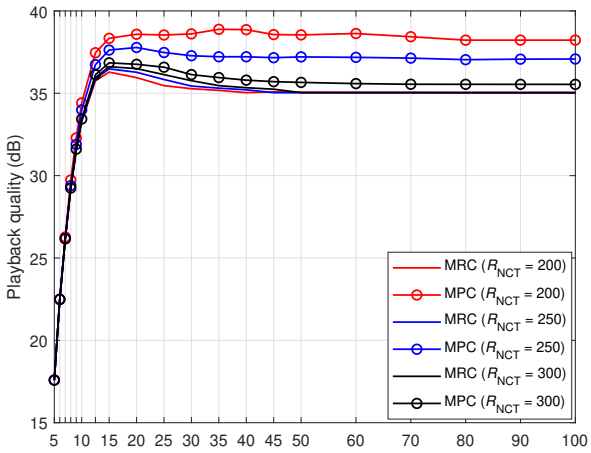

Fig. 7. Playback quality vs. R_{DOL} (m) and various values of R_{NCT} (m).

of λ_{MU} makes the system prefer the SOL and DOL modes to the NCT mode to serve the RUs. All methods increase to the saturated situations in dense D2D networks, i.e., $\lambda_{\mathrm{MU}} > 0.05$ MUs/m^2, due to the physical limits of storage, throughput, and power. Inversely, as λ_{MU} decreases, it is certain that the NCT mode is likely to be used rather than the SOL and DOL modes are. However, due to fewer RUs' demands for the videos when decreasing λ_{MU}, the system is not dedicated to serving them, and thus the performance degrades rapidly. In comparison, the proposed MPC outperforms other methods, especially up to 1.4 dB greater than MRC [23].

Next, Fig. 6 plots the performance of all methods versus different values of DOL radiuses (R_{DOL}). We can observe that the system gains the highest performance at a specific value of R_{DOL}. The reason is that if R_{DOL} decreases, on the one hand, the transmission distance of DOL mode is shorter for higher achievable rate probability. On the other hand, the number of VSs and RUs in both DOL and SOL modes for caching and requesting the videos becomes extremely low leading to the fact that the performance of all methods degrades. If we increase R_{DOL} to the specific value, the DOL mode and the SOL mode are optimally combined in cooperation with the NCT mode to provide the RUs with the highest playback quality. However, if we continue to increase R_{DOL}, the distance transmission of DOL mode is too long making itself useless. In this case, the performance degrades to a saturated situation in which the system serves the RUs by NCT mode and SOL mode (without DOL mode).

Furthermore, in Fig. 7, we plot the performance of MPC and MRC versus R_{DOL} in accordance with different transmission radiuses covered by MBS (R_{NCT}). The results show that the RUs in the clusters located near the MBS are seriously affected by the interference transmission power of the MBS given in (11), making the performance of MRC degraded. Meanwhile, MPC exploits the NCT mode well to significantly enhance the playback quality, and thus increasing the performance gain between MPC and MRC. When the RUs are allocated too far

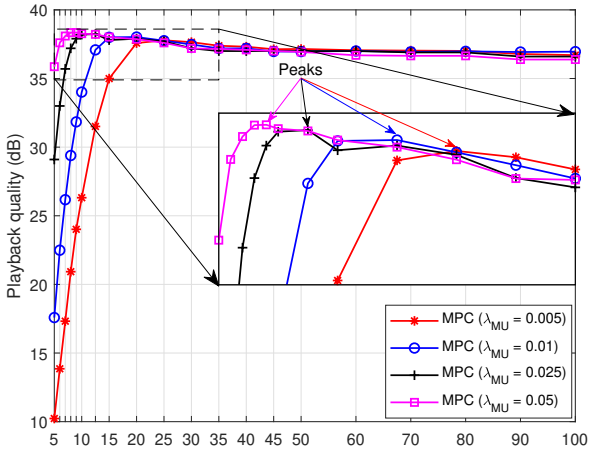

Fig. 8. Playback quality vs. R_{DOL} (m) and various values of λ_{MU} (MUs/m^2).

from the MBS, the NCT mode becomes useless and the gap between MPC and MRC is reduced. In this context, only the PA for DOL mode in each cluster is utilised to improve the performance of MPC. It is clear that the performance of MPC decreases without the NCT mode assisted and gets closer to the performance of MRC when the R_{DOL} is too long.

Finally, the interesting finding is illustrated in Fig. 8 when we evaluate the performance of MPC versus R_{DOL} in accordance with different densities of MUs (λ_{MU}). To do so, we set $P_{\mathrm{DOL}} = 5$ mW and $P_{\mathrm{NCT}} = 7.5$ W. The results indicate that for each density λ_{MU}, we derive a corresponding radius R_{DOL} so that the playback quality can reach its peak. The peak moves from left to right and becomes lower when increasing R_{DOL} and decreasing λ_{MU}. The peak gets lower because the increase of R_{DOL} and the decrease of λ_{MU} both make the successful DOL probability degraded. This finding and the

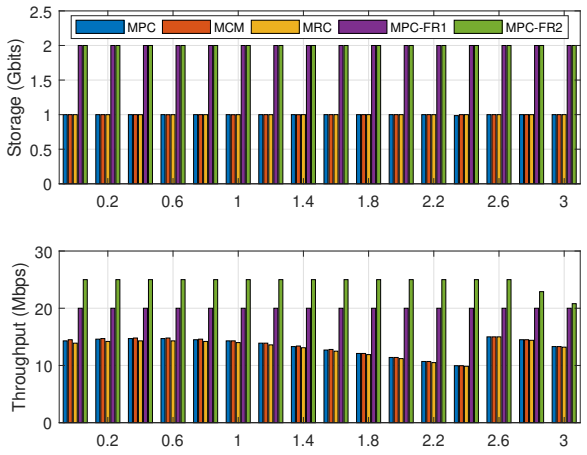

Fig. 9. Storage and throughput consumptions vs. α.

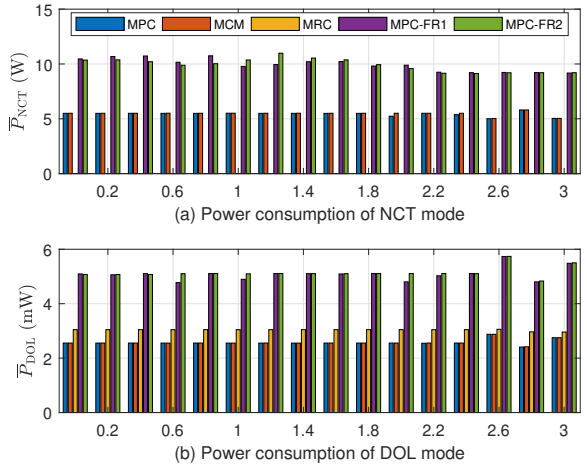

Fig. 10. Power consumption vs. α.

aforementioned analysis enable the system designers to select an optimal DOL radius to group the MUs into different clusters given a specific density of MUs, for the best quality of VASs.

2) Resource consumption: Besides gaining the highest performance of playback quality, the proposed MPC can reasonably utilise the system resources of caching storage, throughput, and power compared to the other methods. The results in Fig. 9 show that the MPC, MCM, and MRC use the same caching storage. Regarding the throughput resource, the MPC consumes a little bit higher than the MRC does, but it is lower than or equal to the MCM is. Meanwhile, the MPC-FR1 and MPC-FR2 consume the highest storage and throughput resources as we relax the constraints $S = 2$ Gbits and $R = 25$ Mbps, instead of 1 Gbits and 15 Mbps in the MPC, MCM, and MRC methods. Obviously, the storage and throughput resource consumption of all methods must satisfy the constraints (29d) and (29e).

With regard to power consumption, there are two types of powers including DOL power (mW) and NCT power (W) as shown in Fig. 10. In Fig. 10(a), it is noted that the MRC method using SOL and DOL modes does not consume any NCT powers. The cooperation between SOL, DOL, and NCT modes further exploits the NCT power to improve the system performance. Therefore, as we can see in Fig. 10(b), it makes the DOL power consumption of MPC and MCM lower than that of MRC. The MPC-FR1 and MPC-FR2 consume the highest power resource because we double their power constraints (P_{DOL} and P_{NCT} in Table V) compared to MPC. Obviously, the power resource consumption of all methods must satisfy the constraints (29f) and (29g).

VI. Conclusion

We have proposed the MPC method for cooperative video transmission in dense D2D networks. The proposed method exploits the benefits of many techniques including probabilistic edge caching, multi-rate encoding, cooperative transmission between SOL, DOL, and NCT modes, and power allocation to provide the RUs with VASs. These techniques together with other aspects of videos, wireless channels, and mobile devices enable us to formulate the MPC optimisation problem. The MPC optimisation problem in the form of binary and real variables under the complicated constraints is efficiently solved by using GA with penalty function and truncated string (GTS) method. The optimal results of caching probability, multi-rate encoding selection, and two-mode power allocation are found to serve the RUs the maximum playback quality of VASs, while utilising the system resources of storage, throughput, and power. Simulation results demonstrate not only the feasibility of GTS in terms of convergence rate, accuracy, and stability, but also the efficiency of the proposed MPC method, compared to other conventional ones. The detailed analyses and interesting findings provide useful insights into the mobile edge caching design of dense D2D networks for VASs.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2023.25.

References

[1] L. S. Vailshery, "Number of IoT connections worldwide 2022-2033," in *Statista*, [Online]. Available: https://www.statista.com, Sep. 2024.

[2] D. V. Huynh, S. R. Khosravirad, S. L. Cotton, T. X. Vu, O. A. Dobre, H. Shin, and T. Q. Duong, "Joint sensing, communications, and computing design for 6G URLLC service-oriented MEC networks," *IEEE Internet of Things J.*, vol. 11, no. 20, pp. 32 429–32 439, Oct. 2024.

[3] X. Lin, L. Kundu, C. Dick, E. Obiodu, T. Mostak, and M. Flaxman, "6G digital twin networks: From theory to practice," *IEEE Commun. Mag.*, vol. 66, no. 11, pp. 72–78, Nov. 2023.

[4] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive IoT toward 6G: A comprehensive survey," *IEEE Internet of Things J.*, vol. 8, no. 15, pp. 11 891–11 915, Aug. 2021.

[5] A. Masaracchia, V. Sharma, B. Canberk, O. A. Dobre, and T. Q. Duong, "Digital twin for 6G: Taxonomy, research challenges, and the road ahead," *IEEE Open J. of the Commun. Society*, vol. 3, pp. 2137–2150, Nov. 2022.

[6] W. Tong and P. Zhu, *"6G: The Next Horizon From Connected People and Things to Connected Intelligence"*, 1st ed. Cambridge University Press, 2021.

[7] T. T. Bui, A. Masaracchia, V. Sharma, O. Dobre, and T. Q. Duong, "Impact of 6G space-air-ground integrated networks on hard-to-reach areas: Tourism, agriculture, education, and indigenous communities," *EAI Endorsed Trans. Tourism, Technol. and Intell.*, vol. 1, no. 1, pp. 1–8, Sep. 2024.

[8] N.-S. Vo, T.-M. Phan, M.-P. Bui, X.-K. Dang, N. T. Viet, and C. Yin, "Social-aware spectrum sharing and caching helper selection strategy optimized multicast video streaming in dense D2D 5G networks," *IEEE Syst. J.*, vol. 15, no. 3, pp. 3480–3491, Sep. 2021.

[9] N. V. Hung, T. D. Chien, P. N. Nam, and T. T. Huong, "Flexible HTTP-based video adaptive streaming for good QoE during sudden bandwidth drops," *EAI Endorsed Trans. Industrial Netw. and Intell. Syst.*, vol. 10, no. 2, pp. 1–15, June 2023.

[10] T. C. Lam, N.-S. Vo, M.-P. Bui, C. D. T. Thai, H. Jung, and V.-C. Phan, "Service time-aware caching, power allocation, and 3D trajectory optimised multimedia content delivery in UAV-assisted IoT networks," *IEEE Trans. Veh. Technol.*, pp. 1–13, Dec. 2024, Early access.

[11] N.-S. Vo *et al.*, "Multi-tier caching and resource sharing for video streaming in 5G ultra-dense networks," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1500–1504, July 2020.

[12] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[13] J. Ma, L. Liu, B. Shang, and P. Fan, "Cache-aided cooperative device-to-device (D2D) networks: A stochastic geometry view," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7444–7455, Nov. 2019.

[14] R. Amer, H. ElSawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Cooperative transmission and probabilistic caching for clustered D2D networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Marrakesh, Morocco, Apr. 2019, pp. 1–6.

[15] R. Amer, H. Elsawy, M. M. Butt, E. A. Jorswieck, M. Bennis, and N. Marchetti, "Optimized caching and spectrum partitioning for D2D enabled cellular systems with clustered devices," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4358–4374, July 2020.

[16] R. Amer *et al.*, "Optimizing joint probabilistic caching and channel access for clustered D2D networks," *Journal of Commun. and Netw.*, vol. 23, no. 6, pp. 433–441, Dec. 2021.

[17] Y. Meng, Z. Zhang, and Y. Huang, "Cache- and energy harvesting-enabled D2D cellular network: Modeling, analysis and optimization," *IEEE Trans. Green Commun. and Netw.*, vol. 5, no. 2, pp. 703–713, June 2021.

[18] M. Naslcheraghi *et al.*, "Probabilistic analysis of operating modes in cache-enabled full-duplex D2D networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6624–6635, June 2022.

[19] X. Song, Y. Geng, X. Meng, J. Liu, W. Lei, and Y. Wen, "Cache-enabled device to device networks with contention-based multimedia delivery," *IEEE Access*, vol. 5, pp. 3228–3239, Feb. 2017.

[20] J. Ma, L. Liu, H. Song, R. Shafin, B. Shang, and P. Fan, "Scalable video transmission in cache-aided device-to-device networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4247–4261, June 2020.

[21] S. Han, F. Xue, C. Yang, J. Liu, and F. Lin, "Data-supported caching policy optimization for wireless D2D caching networks," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7618–7630, Nov. 2021.

[22] S. Batabyal, "On the effect of redundant caching policy on multimedia streaming in D2D underlay network," in *Proc. IEEE Global Commun. Conf.*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 1–6.

[23] Q.-N. Tran, N.-S. Vo, T.-M. Phan, T. C. Lam, and A. Masaracchia, "Multi-rate probabilistic caching optimised video offloading in dense D2D networks," *IEEE Commun. Lett.*, vol. 27, no. 4, pp. 1240–1244, Apr. 2023.

[24] V. V. Lam, T. Hoang, T. Do-Duy, T.-M. Phan, D.-B. Ha, and N.-S. Vo, "Adaptive encoding rate and clustering assisted probabilistic edge caching optimization for video transmission in dense D2D networks," in *Proc. Int. Conf. Advanced Technol. for Commun.*, Ho Chi Minh City, Vietnam, Oct. 2024, pp. 1–5.

[25] J. Wu, T. Xu, Y. Ouyang, J. Tian, and T. Zhou, "Delay-aware probabilistic cache placement for scalable videos in satellite-terrestrial networks," in *Proc. Int. Conf. Wireless Commun. and Signal Processing*, Hefei, China, Oct. 2024, pp. 1–6.

[26] Q. Xia, Z. Jiao, and Z. Xu, "Online learning algorithms for context-aware video caching in D2D edge networks," *IEEE Trans. on Parallel and Distributed Syst.*, vol. 35, no. 1, pp. 1–19, Jan. 2024.

[27] M. Choi, J. Kim, and J. Moon, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245–1257, June 2018.

[28] H. Zhu, Y. Cao, Q. Hu, W. Wang, T. Jiang, and Q. Zhang, "Multi-bitrate video caching for D2D-enabled cellular networks," *IEEE Multimedia*, vol. 26, no. 1, pp. 10–20, Jan.-Mar. 2019.

[29] J. Ma, L. Liu, H. Song, R. Shafin, B. Shang, and P. Fan, "Scalable video transmission in cache-aided device-to-device networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 4247–4261, June 2020.

[30] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. IEEE Inter. Commun. Conf.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[31] X. Zhang and J. Wang, "Statistical QoS-driven power adaptation for distributed caching based mobile offloading over 5G wireless networks," in *Proc. IEEE Int. Conf. Computer Commun. Workshops*, Honolulu, HI, July 2018, pp. 1–6.

[32] Q. Li, Y. Zhang, A. Pandharipande, X. Ge, and J. Zhang, "D2D-assisted caching on truncated Zipf distribution," *IEEE Access*, vol. 7, pp. 13 411–13 421, Jan. 2019.

[33] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *Proc. IEEE Inter. Conf. on Computer Commun.*, New York, NY, Mar. 1999, pp. 126–134.

[34] Q.-N. Tran, N.-S. Vo, M.-P. Bui, T.-M. Phan, Q.-A. Nguyen, and T. Q. Duong, "Spectrum sharing and power allocation optimised multihop multipath D2D video delivery in beyond 5G networks," *IEEE Trans. Cognitive Commun. and Netw.*, vol. 8, no. 2, pp. 919–930, June 2022.

[35] 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.

[36] A. Chipperfield, P. Fleming, H. Pohlheim, and C. Fonseca, *"Genetic Algorithm TOOLBOX For Using with Matlab Ver 1.2 User's Guide"*. University of Sheffield, 1994.

[37] ITU, *HM Reference Software Version 12.0*. http://hevc.hhi.fraunhofer.de, 2016.

[38] N.-S. Vo, T.-H. Nguyen, and H. K. Nguyen, "Joint active duty scheduling and encoding rate allocation optimized performance of wireless multimedia sensor networks in smart cities," *Springer Mobile Netw. Appl.*, vol. 23, no. 6, pp. 1586–1596, Dec. 2018.

[39] N.-S. Vo, T. Q. Duong, H. D. Tuan, and A. Kortun, "Optimal video streaming in dense 5G networks with D2D communications," *IEEE Access*, vol. 6, pp. 209–223, Oct. 2017.