MINISTRY OF EDUCATION AND TRAINING
DUY TAN UNIVERSITY

BUI MINH PHUNG

COOPERATIVE VIDEO CACHING AND DELIVERY OPTIMISATION IN 5G ULTRA-DENSE NETWORKS

MAJOR: Computer Science
CODE: 9.48.01.01

Summary of PhD Dissertation

ACADEMIC SUPERVISOR
Dr. Vo Nguyen Son

DANANG – 2021

INTRODUCTION

1. Motivation

In the new era of the fourth industrial revolution embedded with 5G networks based Internet of things (IoT), about 50 billion devices will be connected to wireless networks for communications and advanced services by 2021 [1]. In this context, a massive number of mobile users (MUs) requesting high data rate services and applications will pose a serious challenge for 5G networks. These certainly make 5G networks deteriorated because of extreme traffic congestion at the backhaul links of macro base stations (MBSs) and small cell base stations (SBSs). While developing more high-speed backhaul links of the MBSs and the SBSs can significantly improve the system delivery capacity, but in a very costly investment with network architecture changes, new solutions and techniques can be alternated more effectively.

Ultra-dense networks (UDNs) have been emerged as a promising architecture to meet the requirements of 5G networks, i.e., increasing the system capacity 1000 times and decreasing the latency to 1ms [2]. However, deploying UDNs requires disruptive technologies, techniques and optimization designs to provide a large number of MUs with high data rate and huge size of services and applications, such as video streaming applications and services (VAS), at high quality of service (QoS) and resource efficiency. Therefore, many technologies, techniques and optimization designs for UDNs have been studied focusing on how to utilize the resources of space, time, code, spectrum, bandwidth, energy and storage, and how to bring advanced services closer to the MUs.

Recently, caching technique has drawn significant attention from researchers in both academic and industrial sectors to benefit the Internet service providers and content providers as well as to satisfy the high demand of MUs for advanced services and applications. In this field, Vietnamese researchers have had the opportunity to cooperate with the world-class experts to develop projects and publish research works in prestigious journals [14, 19, 21-23, 29-31, 35, 38, 44, 50]. However, these research groups have not comprehensively implemented the model, analysis and optimization design of cooperative video caching and delivering in 5G UDNs. Therefore, there have had many challenges to improve the capacity of 5G UDNs to provide the MUs with advanced services at high QoS and high resource efficiency. In developed countries, researches in this field are more in quantity and better in quality, e.g., device-to-device (D2D) caching [15, 20, 33, 42, 45], femto-caching [26, 27, 34], small-cell caching [17, 32, 37, 41, 43, 46], MBS caching [39, 40], and multi-tier caching [16, 18, 24, 25, 28, 36, 47, 48].

Although researches carried out in developed countries are numerous and make more important contributions than Vietnamese researchers do, there are many problems that need more emerging techniques, models, analysis, optimization designs, and breakthrough standards to meet the high and complex requirements of video caching and delivering in 5G UDNs.

2. Objectives, subjects, scope and research methods x

2.1. Objective of research

In this dissertation, the objective is to propose a multi-tier cooperative video caching and delivery optimization in 5G UDNs to provide the MUs with advanced services at high QoS while efficiently exploiting the resources.

2.2. Subjects of research

- 5G UDNs: models, characteristics of multi-tier 5G UDNs including MBS, SBSs (such as microcell, picocell, femtocell) and D2D communications.
- Video: common video standards, rate-distortion (RD) and video encoding rate, caching and delivering video in 5G UDNs.
- Models: social-ware, QoS and resources in 5G UDNs.
- Algorithms: algorithms to solve the optimization problems of caching, resource sharing and video delivering from MBS and SBSs to MUs and among the MUs.

2.3. Scope of research

The dissertation is within the scope of techniques, users, networks and services as well as maths and tools, described as follows:

- Techniques: caching and resource sharing techniques in 5G UDNs.
- Users, networks and services: VAS in 5G UDNs, QoS criteria, resource efficiency, social-aware model, i.e., Indian Buffet Model (IBM).
- Maths and tools: searching algorithms, mathematical analysis tools, video encoding tools and the rate-distortion model of videos.

2.4. Research methods

To obtain the results associated with the research objectives, based on the subjects of research, two research methods are applied including 1) Analytical and synthetic method and 2) Quantitative research method, given as follows:

- Analytical and synthetic methods: Analyzing and evaluating the related works, thereby identifying outstanding issues, questions and ideas, and proposing research hypotheses as well as confirming the requirements for new better emerging models and solutions. And then, the results of analysis and evaluation will be synthesized, linked, combined and reorganized to propose better schemes and solutions based on the research hypothesis.
- Quantitative research methods: In the quantitative method, the proposed models and related factors are quantified by computational expressions. These computational expressions to describe the system model are validated by performing simulations and observing the response of the system. The benefits of the proposed model are verified by comparing with other conventional solutions.

3. Research tasks, achieved results

3.1. Research tasks

- Research task 1: formulate a social-aware caching and resource sharing (SCS) optimization problem for video streaming services in 5G UDNs, compute the system parameters of the proposed model, solve the optimization problem by searching algorithm, and simulate and evaluate the proposed model.
- Research task 2: formulate a user-demand-aware multi-rate cooperative video caching and delivery (CRS) optimization problem for a high video streaming performance in 5G UDNs, compute the system parameters of the proposed model, solve the optimization problem by searching algorithms, and simulate and evaluate the proposed model.

3.2. Achieved results

The two main results of the dissertation are summarized as follows:

- Result 1: Propose the SCS scheme by exploiting the available caching storage and spectrum resources in multi-tier 5G UDNs.
- Result 2: Propose the CRS scheme by efficiently exploiting the available caching storage and spectrum resources in multi-tier 5G UDNs.

4. Dissertation structure

Introduction
Chapter 1: Overview of Video Caching and Delivering in 5G UDNs.
Chapter 2: Social-aware Caching and Resource Sharing Maximized Video Delivery Capacity in Multi-tier 5G UDNs.
Chapter 3: User-demand-aware Multi-rate Cooperative Video Caching and Delivery in Multi-tier 5G UDNs.
Conclusion

CHAPTER 1. OVERVIEW OF VIDEO CACHING AND DELIVERING IN 5G UDNs

1.1. Introduction to 5G UDNs

The model of 5G UDNs is based on a multi-tier architecture to provide the mobile users (MUs) with various services and applications, depicted in Figure 1-1. In this model, the upper tier uses higher spectrum and stronger signal with higher transmission rate, while the service and coverage area will be expanded thanks to the middle and lower tiers [2, 52]. 5G UDNs allow a flexible combination of communication techniques by exploiting the spectrum resources in the network to ensure connectivity and meet the higher user demand. However, 5G UDNs have had many challenges about physical issues as well as using and exploiting the information of MUs, such as:

- Ultra dense connectivity from different access ranges and traffic at different locations. These in turn cause local congestion, unfair performance and access authority of the MUs.
- Private or shared access restrictions at different tiers generate different interference levels. For example, different D2D communications can interfere with others and even disrupt the access.
- The issue of priority over channel access at different frequencies and over resource allocation strategies.

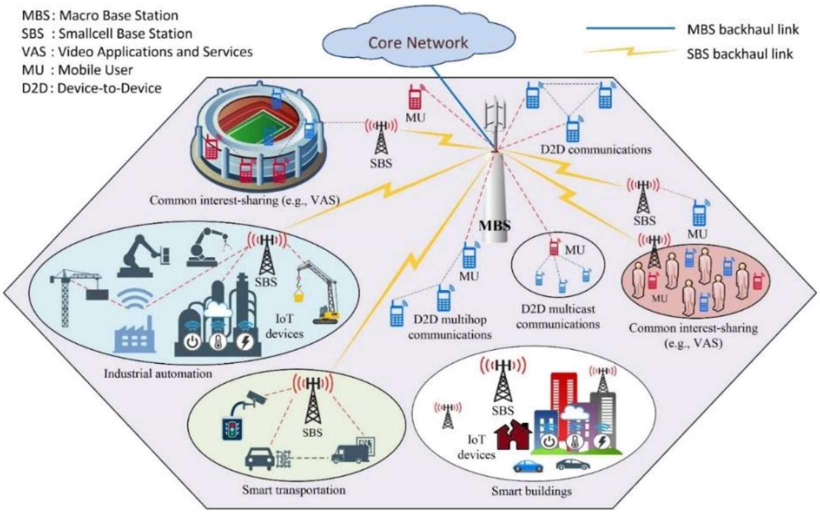

Figure 1-1. Connections among devices, technologies and applications in 5G UDNs [2]

It is obvious that the studies aiming at providing solutions to overcome the aforementioned challenges in 5G UDNs are urgent to optimally using the resources; improve the QoS and even the quality of experience (QoE).

1.2. Caching and delivering video model in 5G UDNs

Caching techniques to bring services/ applications/ contents, especially services and applications with high capacity and data rate such as video streaming applications and services (VAS), closer to the MUs, can be labeled as the most efficient solution to deal with the congestion problem that does not require any system architecture changes.

Caching technique is often joint with downlink spectrum resource sharing to bring more benefits to the Internet service providers and content providers as well as to satisfy the high demand of MUs for advanced services and applications such as video delivering.

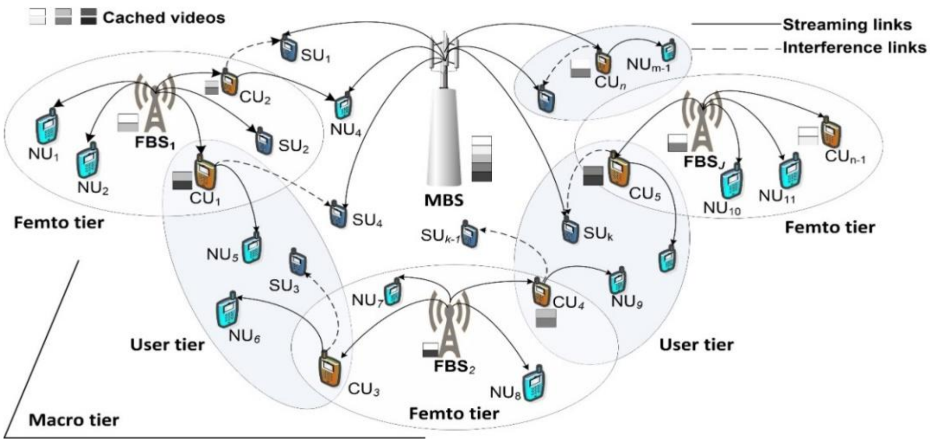

Figure 1-2. Caching and resource sharing model for video services and applications in 5G UDNs

Caching techniques include single-tier and multi-tier caching and delivering. Single-tier video caching and delivering is a technique that allows caching the videos at the MBS, the SBSs, or the MUs with available caching storage. Multi-tier cooperative video caching and delivering is a technique that allows caching the videos at more than one tier in 5G UDNs (Figure 1-2).

Although multi-tier caching is more complex than single-tier caching, the videos can be delivered flexibly by the best tier to the MUs. Studies related to multi-tier caching are mainly deployed in two tiers. In this dissertation, we deployed the caching technique in three tiers jointed with downlink spectrum resources sharing (SUs) for device-to-device (D2D) communications. The tiers cooperate in caching and sharing resources to serve the MUs the highest QoS and resource efficiency.

1.3. Algorithms applied to solve the optimization problems of video caching and delivering

The joint optimization for caching and resource sharing is a technique that has been applied by many researchers to deliver the videos in 5G UDNs with high efficiency and low cost. These caching optimization problems are usually solved by finding the indexes that determine the caching placement and selection/allocation of resources in the system using binary vectors or matrices, i.e., the values of 1 and 0 meaning that the system does (or does not) decide to cache and/or to share. To find the optimal results, optimal search algorithms have been proposed to solve the problems such as exhaustive search [18, 54-57], dynamic programming [50], stochastic learning [40], game learning [43], greedy algorithms [58], and heuristic algorithms [25, 59]. Each algorithm has advantages and disadvantages in terms of accuracy, time, and computational complexity. In this dissertation, a genetic algorithm (GA) is applied to solve a typical multi-tier caching and resource sharing optimization problem for video streaming in 5G UDNs in the following chapters while a bat algorithm (BA) is also implemented to confirm the effectiveness of the GA. In addition, the exhaustive algorithm (EA) is used because of its simplicity, accuracy, and suitability for small search space problems, as well as a benchmark for GA and BA.

1.4. Conclusion of chapter 1

Chapter 1 introduces an overview of video caching and delivering in 5G UDNs with benefits and challenges that need to be researched. Besides, caching and resource sharing in 5G UDNs are also presented in general. The advantages and disadvantages of models, caching and sharing techniques of the related works are also analyzed and evaluated in detail. Chapter 1 also introduces some algorithms such as EA, GA, BA, which will be implemented, compared, and applied to solve optimization problems in Chapters 2 and 3. All is to demonstrate the differences and the contributions of the dissertation.

CHAPTER 2. SOCIAL-AWARE CACHING AND RESOURCE SHARING MAXIMIZED VIDEO DELIVERY CAPACITY IN MULTI-TIER 5G UDNs

2.1. Introduction to SCS

In 5G networks, a massive number of connections of high data rate services, e.g., video streaming services, certainly make the networks deteriorated because of extreme traffic congestion at the backhaul links of macro base stations (MBSs). Although ultra-dense networks (UDNs) have been considered as a promising architecture to stimulate the 5G networks, the congestion problem hampers the UDNs to provide mobile users (MUs). In Chapter 2, a social-aware caching and resource sharing (SCS) strategy for video streaming services is proposed to maximize the video delivery capacity in 5G UDNs. The proposed SCS strategy can relax the workload of the core networks in general and the backhaul link of MBS by allowing an arbitrary MU can retrieve the videos flexibility from the MBSs, FBSs, and TXs at high cache-hit ratio and maximum delivery capacity. In this context, the MUs include Transmitters (TXs) and Receivers (RXs) in D2D communications (TX-RX) and the MUs share their downlink spectrum (SUs) with the D2D pairs for D2D communications. The problem is how the SCS delivers the videos to the MUs with maximum capacity.

To solve the aforementioned problem, the SCS strategy is implemented and solved to indicate which FBSs caches the videos and which SU shares its downlink resource with the D2D pairs, so as to maximize the average system capacity delivered to the MUs. In order to improve the efficiency of the SCS strategy, the SCS problem is formulated by taking into account the social relationship of each D2D pair and the available storage of FBSs. Furthermore, a SU shares its downlink resources for D2D pairs from TXs to RXs, and the TXs reuse this downlink resource will cause interference to the SU. Therefore, the SCS also considers the lower bound constraint of signal to interference plus noise ratio (SINR) at the SU to ensure the quality of the SU. Simulation results are analyzed to show the benefits of the proposed SCS strategy compared to other conventional schemes. The results related Chapter 2 were published in Mobile Networks & Application journal [C1] and presented at the Heterogeneous Networking for Quality, Reliability, Security and Robustness conference (Qshine 2018) [C2] given in the "List of publications related to the dissertation" at the end of the document.

2.2. System model of SCS

In this chapter, the system model of SCS in 5G UDNs consists of a device layer and a social layer as illustrated in Figure 2-1. In the device layer, there are one MBS, *I* videos, *J* FBSs, *K* SUs, and *N* D2D pairs (i.e., *N* TXs and *N* RXs). In the social layer, each D2D pair has its own social relationship based on their history contact duration and encounter. In this model, SCS strategy implemented at the MBS is described in the following three steps:

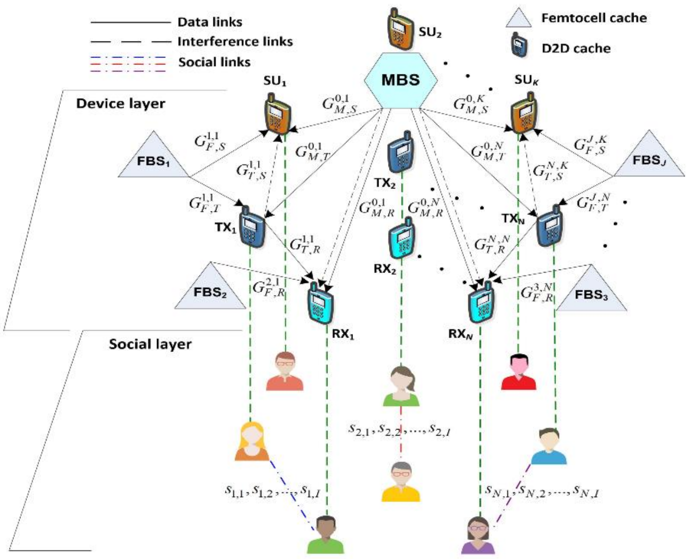

Figure 2-1. System model of SCS

Step 1 – <u>Updating system parameters</u>: In this step, the MBS updates the system parameters of the videos, FBSs, SUs, D2D pairs, characteristics of wireless channels, system bandwidth, etc., as listed in Table 2-1 if there are some significant changes.

Step 2 – <u>Formulating and solving SCS optimization problem</u>: The updated parameters enable the MBS to formulate the SCS optimization problem and then solve it for the optimal results of 1) number of caching copies of each video and caching placements at the FBSs represented by the caching index $u_{j,i}$, $j = 1,2,...,J$ and $i = 1,2,...I$; 2) downlink resource sharing allocation between the SUs and the D2D pairs represented by the caching index $v_{k,n}$, $k = 1,2,...,K$ and $n = 1,2,...N$; to maximize the system delivering capacity, where

$u_{j,i} = 1$ if the FBS j decides to cache video i, otherwise $u_{j,i} = 0$ and
$v_{k,n} = 1$ if the SU k decides to share its downlink resource with D2D pair n, otherwise $v_{k,n} = 0$

Step 3 – <u>Implementing SCS strategy</u>: Finally, based on the optimal results, the MBS assigns the FBSs to cache their corresponding videos and the SUs to share theirs downlink resources with the proper D2D pairs, for delivering the videos to the MUs (SUs, TXs, and RXs).

2.3. SCS formulations

In order to facilitate the formulations, the notations used in this system model are presented in Table 2-1. Following the system model and the objective of the chapter, we formulate all the aspects of the SCS including the social relationship between the TX and the RX of each D2D pair and the wireless channels that allow us to derive the system delivery capacity from the MBS and the FBSs to the MUs and from the TXs to the RXs. Consequently, we derive the overall average system delivery capacity, which is the objective function of the SCS optimization problem.

Table 2-1. Notations of SCS

Symbols	Descriptions
I	Number of videos
J	Number of FBSs
K	Number of SUs

N	Number of D2D pairs, each pair has a D2D transmitter (TX) and a D2D receiver (RX)
T_{min}^{i}	Minimum time required for offloading the video *i*, i.e., depending on the length of the video *i*, *i* = 1, 2,…, *I*
$s_{n,i}$	Social-based probability that the D2D pair *n* is qualified to offload the optimize video *i*, *n* = 1, 2,…, *N*
$u_{j,i}$	Caching index, i.e., $u_{j,i} = 1$ if the FBS *j* decides to cache the video *i*, otherwise $u_{j,i} = 0$
$v_{k,n}$	Downlink resource sharing index, i.e., $v_{k,n} = 1$ if the SU *k* decides to share its downlink resource with the D2D pair *n*, otherwise $v_{k,n} = 0$
P_M^0	Transmission power of the MBS
P_F^j	Transmission power of the FBS *j*, *j* = 1, 2, ..., *J*
P_T^n	Transmission power of the TX *n* of the D2D pair *n*
$G_{X,Y}^{x,y}$	Channel gain between *X* and *Y*, *x* and *y* are the index orders of *X* and *Y*, respectively
N_0	Power of additive white Gaussian noise (AWGN)
W	System bandwidth
r_i	Access rate (popularity) of the video *i*
β_n	Percentage of available storage of the TX *n*
$p_{n,i}$	Probability that the TX *n* decides to cache the video *i*
γ_0	Lower bound constraint of signal to interference plus noise ratio (SINR) at SUs
R_S	Capacity delivered from the MBS and FBSs to the SUs
R_T	Capacity delivered from the MBS and FBSs to the TXs
R_R	Capacity delivered from the MBS, FBSs, and TXs to the RXs

2.3.1. Social relationships of D2D pairs

The social relationship between TX and RX of the D2D pair *n* is used to derive the probability of successfully offloading the video *i* of length T_{min}^{i}.

$$s_{n,i} = 1 - \int_0^{\delta T_{min}^{i}} f(u;\kappa_n,\theta_n)du = 1 - \frac{\gamma\left(\kappa_n, \frac{\delta T_{min}^{i}}{\theta_n}\right)}{\Gamma(\kappa_n)} \tag{2.5}$$

2.3.2. Wireless channels

The wireless channels are characterized by the channel gains given by $G_{X,Y}^{x,y} = h_{X,Y}^{x,y}\, g_{X,Y}^{x,y}$, here:

$X \in$ {M, F, T} stands for the source nodes, i.e., {MBS, FBS, TX}

$Y \in$ {S, T, R} stands for the destination nodes, i.e., {SU, TX, RX}.

where $h_{X,Y}^{x,y}$ is the exponential power fading coefficient (~ exp(1))

And $g_{X,Y}^{x,y} = \|d\|^{-\eta}$ is the standard power law path loss function in which η is the path loss exponent *d* is the distance between *X* and *Y*.

2.3.3. System delivery capacity

2.3.3.1. Capacity delivered to SUs

The SNR at the SU *k* over the channel from the FBS *j* is given by

$$\gamma_{F,S}^{j,k,i} = \frac{u_{j,i} P_F^j G_{F,C}^{j,k}}{N_0} \tag{2.7}$$

The SINR at the SU *k* over the channel from the MBS is given by

$$\gamma_{M,S}^{0,k,i} = \frac{\prod_{j=1}^{J}\left(1-u_{j,i}\right) P_M^0 G_{M,S}^{0,k}}{N_0 + \sum_{n=1}^{N} s_{n,i} v_{k,n} p_{n,i} P_T^n G_{T,S}^{n,k}} \tag{2.8}$$

$p_{n,i}$ the probability of the TX *n* to cache the video *i*, defined as

$$p_{n,i} = ar_i + b\beta_n \tag{2.9}$$

where a, b ∈ [0,1], a + b = 1 and the access rate of the video *i* modelled by Zipf-like distribution [88], is defined as

$$r_i = \frac{i^{-\alpha}}{\sum_{i=1}^{I} i^{-\alpha}} \tag{2.10}$$

here $\alpha \geq 0$ is the skewed access rate among different videos. $\alpha = 0$ means that all videos have the same access rate of $\frac{1}{I}$, while a higher value of α causes higher access rate difference between the first popular videos and the last unpopular ones.

Taking the system bandwidth *W* and the number of videos *I* into account and applying Shannon-like capacity, the capacity delivered to the SUs is obtained as:

$$R_S = W\left\{\sum_{i=1}^{I} r_i \sum_{k=1}^{K}\left[log_2\left(1+\ \gamma_{M,S}^{0,k,i}\right)+\sum_{j=1}^{J} log_2\left(1+\ \gamma_{F,S}^{j,k,i}\right)\right]\right\} \tag{2.11}$$

2.3.3.2. Capacity delivered to TXs

The capacity delivered from the MBS to the TX and that from the FBS *j* to the TX *n* are computed based on the SINRs at the TX *n*, simply given by

$$\gamma_{F,T}^{j,n,i} = \frac{u_{j,i} P_F^j G_{F,T}^{j,n}}{N_0} \tag{2.12}$$

$$\gamma_{M,T}^{0,n,i} = \frac{\prod_{j=1}^{J}(1-u_{j,i}) P_M^0 G_{M,T}^{0,n}}{N_0} \tag{2.13}$$

From Eqs. (2.12) and (2.13), the capacity delivered from the FBSs and the MBS to the TXs is expressed as

$$R_T = W\left\{\sum_{i=1}^{I} r_i \sum_{n=1}^{N}\left[log_2\left(1+\ \gamma_{M,T}^{0,n,i}\right)+\sum_{j=1}^{J} log_2\left(1+\ \gamma_{F,T}^{j,n,i}\right)\right]\right\} \tag{2.14}$$

2.3.3.3. Capacity delivered to RXs

The SINR at the RX *n* over the channel from the TX *n*:

$$\gamma_{T,R}^{n,k,i} = \frac{s_{n,i} v_{k,n} p_{n,i} P_T^n G_{T,R}^{n,n}}{N_0 + P_M^0 G_{M,R}^{0,n} + \sum_{l=1, l\neq n}^{N} s_{l,i} v_{k,l} p_{l,i} P_T^l G_{T,R}^{l,l}} \tag{2.15}$$

The SNR at the RX *n* is computed over the channel from the FBS *j* as

$$\gamma_{F,R}^{j,n,k,i} = \frac{u_{j,i}\left(1 - s_{n,i} v_{k,n} p_{n,i}\right) P_F^j G_{F,R}^{j,n}}{N_0} \tag{2.16}$$

The SNR at the RX *n* is computed over the channel from the MBS as

$$\gamma_{M,R}^{0,n,k,i} = \frac{\prod_{j=1}^{J}\left(1-u_{j,i}\right)\left(1 - s_{n,i} v_{k,n} p_{n,i}\right) P_M^0 G_{M,R}^{0,n}}{N_0} \tag{2.17}$$

So far, the capacity delivered from the MBS, the FBSs, and the TXs to the RXs is expressed as

$$R_R = W\left\{\sum_{i=1}^{I} r_i \sum_{n=1}^{N}\sum_{k=1}^{K}\left[log_2\left(1+\ \gamma_{M,R}^{0,n,k,i}\right)+log_2\left(1+\ \gamma_{T,R}^{n,k,i}\right)+\sum_{j=1}^{J} log_2\left(1+\ \gamma_{F,R}^{j,n,k,i}\right)\right]\right\} \tag{2.18}$$

Finally, from Eqs. (2.11), (2.14) and (2.18), the overall average system delivery capacity per each MU is given by

$$R = \frac{R_s + R_T + R_R}{K + 2N} \tag{2.19}$$

2.4. SCS optimization problem and solution

The SCS optimization problem consists of the objective function (2.19) and the two constraints on 1) the storage capacity of the FBSs and 2) the target SINR of the SUs. This problem can be formulated below

$$\max_{u_{j,i}, v_{k,n}} R \tag{2.20}$$

$$s.t. \begin{cases} \sum_{j=1}^{J} u_{j,i} \le c_i, i = 1,2,\dots,I \\ \sum_{n=1}^{N} s_{n,i} v_{k,n} p_{n,i} P_T^n G_{T,C}^{n,k} \le \dfrac{P_M^0 G_{M,S}^{0,k}}{\gamma_0} - N_0, k = 1,2,\dots,K, i = 1,2,\dots,I \end{cases} \tag{2.21}$$

Where, c_i the optimal number of caching copies of the video *i* are cached how to maximize the average number of copies cached in the FBSs. This problem can be formulated below

$$\max_{c_i} \sum_{i=1}^{I} r_i c_i \tag{2.22}$$

$$s.t. \begin{cases} 0 \le c_i \le J, i = 1,2,\dots,I \\ \sum_{i=1}^{I} c_i \le C^*, I \le C^* \le I \times J \end{cases} \tag{2.23}$$

This problem was solved by an exhaustive matrix search [54] to find U_{JI}^* and V_{KN}^* presented in the Algorithm 2.1.

Algorithm 2.1: Exhaustive matrix search

Input: Initial parameters given in Table 2-2
Generating two feasible matrix search space $\mathcal{U}^* \in \mathcal{U}$ and $\mathcal{V}^* \in \mathcal{V}$ that satisfy (2.23)
Output:
$\mathcal{R}^*$: maximize the average system capacity to MU
$\{U_{JI}^*, V_{KN}^*\}$
For each matrix U_{JI} in $\mathcal{U}^*$ **do**
  For each matrix V_{KN} in $\mathcal{V}^*$ **do**
    $R(U_{JI}, V_{KN}) = R$, computing (2.19)
    $\mathcal{R} \leftarrow \mathcal{R} \cup R(U_{JI}, V_{KN})$
  End for
End for
$R^* = \max \mathcal{R}$
$\{U_{JI}^*, V_{KN}^*\}$ = argmax $\mathcal{R}$

2.5. Performance evaluation

2.5.1. System setting

To perform evaluation, the system parameters are set as shown in Table 2-2. Furthermore, for convenience but without loss of generality, we omit the effect of fading coefficient and only consider the standard power law path loss function in which the path loss exponent $\eta = 4$ and the distances from the MBS to the MUs, the FBSs to the MUs, the SUs to the TXs, and the TXs to the RXs, are randomly distributed from 300m to 1500m, 50m to 250m, 50m to 100m, and 1m to 50m, respectively. To demonstrate the benefits of our proposed SCS strategy (SCS), we compare SCS to the other four schemes including: 1) none social-aware (None-SOA), 2) none downlink resource sharing (None-DRS), 3) average system delivery capacity (AVE), and 4) minimum system delivery capacity (MIN).

Table 2-2. Parameters setting of SCS

Symbols	Specifications
I, J, K, N	5 videos, 3 FBSs, 3 SUs, 5 TX-RX pair
(θ_n)	(5, 10, 20, 15, 25) [46]
(κ_n)	(1, 4, 3, 2, 5) [46]

(T^{i}_{min})	(1, 15, 10, 5, 20) s
(β_n)	(0.1, 0.5, 0.9, 0.3, 0.7)
$\delta, \alpha, \gamma_0, N_0$	10, 2, 5dB, 10^{-13}W
W, P^0_M, P^j_F, P^n_T	5MHz, 5W, 1W, 0.1W
a, b, C^*	0.5, 0.5, 0.7× I×J

2.5.2. System performance of SCS

*2.5.2.1. System performance versus C**

We first evaluate the SCS, None-SOA, None-DRS, AVE, and MIN versus caching storage capacity of the FBSs by setting $C^* = \{0,0.1,0.2,..,1\} \times I \times J$, and compare SCS to None-SOA, None-DRS, AVE and MIN. As shown in Figure 2-2, for proper portion of storage capacity cached in the FBSs, we select $C^* = 0.7 \times I \times J$. When increasing C^*, he proposed SCS gains higher performance and is always the best compared to the others. The None-SOA and None-DRS are similar and better than the AVE and MIN, while the MIN is the worst one. The results of None-SOA and None-DRS mean that social-ware and downlink resource sharing issues play an equivalent important role in improving the system performance.

2.5.2.2. System performance versus δ

Figure 2-3 plots the performance of SCS, None-SOA, None-DRS, AVE, and MIN versus different duration sets of all the considered videos by changing δ in Eq. (2.5) from 1 to 40. In comparison, the proposed SCS is better than the others and its performance is reduced to the performance of None-SOA and None-DRS when δ is too low (or too high). It interestingly means that the duration of videos can be adjusted to meet the social relationships of the D2D pairs, and thus obtaining the highest system performance (i.e., δ = 10).

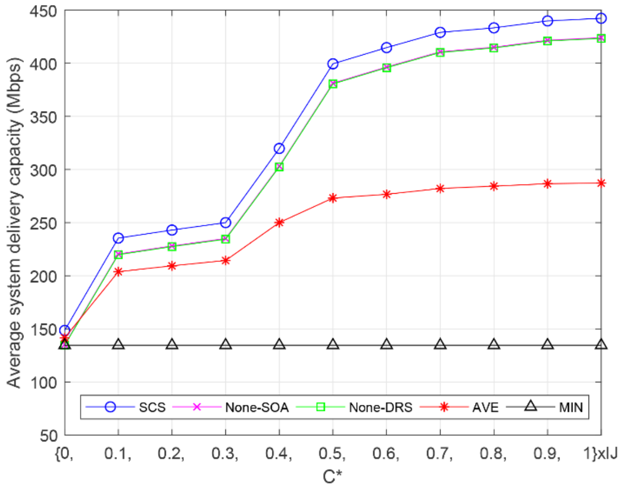

Figure 2-2. Capacity performance versus caching storage capacity of FBSs

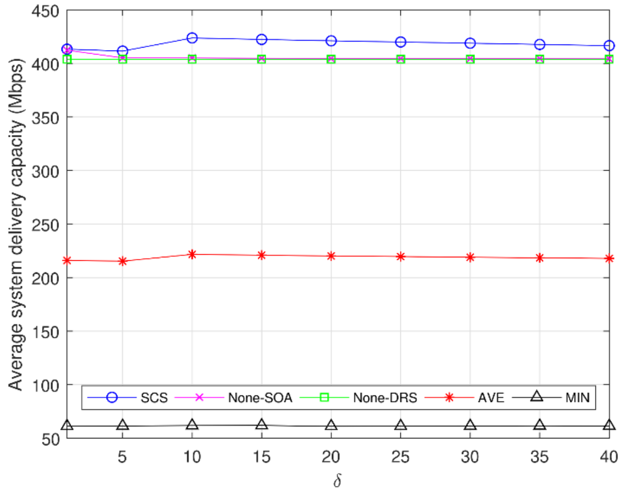

Figure 2-3. Capacity performance versus coefficient of video length adjustment

2.5.2.3. System performance versus α

We evaluate the system performance versus the skewed access rate (α) among different videos (from 0 to 3). The result showed in Figure 2-4. The results demonstrate that exploiting the skewed access rate can improve the system performance. If all the videos have the same access rate, i.e., $\alpha = 0$, the system has a not very high benefit gained from the SCS compared to the None-SOA and None-DRS. However, the SCS gets higher system delivery capacity when α increases. The results imply that focusing on serving less number of higher access rate videos yields higher system performance, and this way, our proposed SCS provides the best system performance compared to the None-SOA, None-DRS, AVE, and MIN.

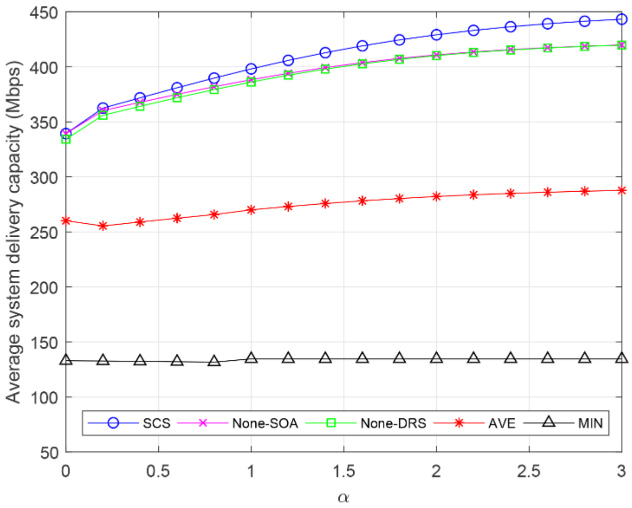

Figure 2-4. Capacity performance versus skewed access rate of videos

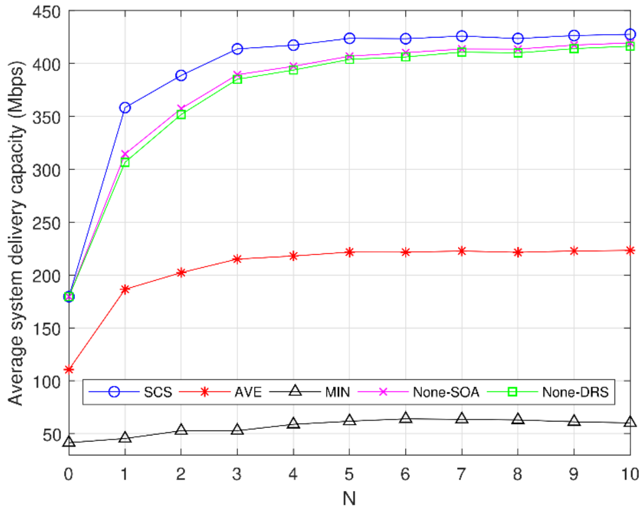

Figure 2-5. Capacity performance versus number of D2D pairs

2.5.2.4. System performance versus N

In Figure 2-5, the system performance is investigated by changing the number of D2D pairs N from 0 to 10. It is easy to see that if $N = 0$, the system does not gain any benefits from downlink resource sharing and social relationship for D2D communications. Therefore, the SCS, None-SOA, and None-DRS have the same result and the system delivery capacity is low. If N increases, the system delivery capacity gets higher and becomes saturated. The saturated situation holds due to the constraint on the target SINR of the SUs that limits the amount of D2D communications. In the context of 5G UDNs with dense D2D pairs, we carefully select an efficient number of D2D pairs such that the target SINR of the SUs is guaranteed and the system delivery capacity is high enough before reaching the saturated situation. The results demonstrate that the SCS gains the best performance, the None-SOA and None-DRS are better than the AVE and MIN.

2.5.2.5. System performance versus J

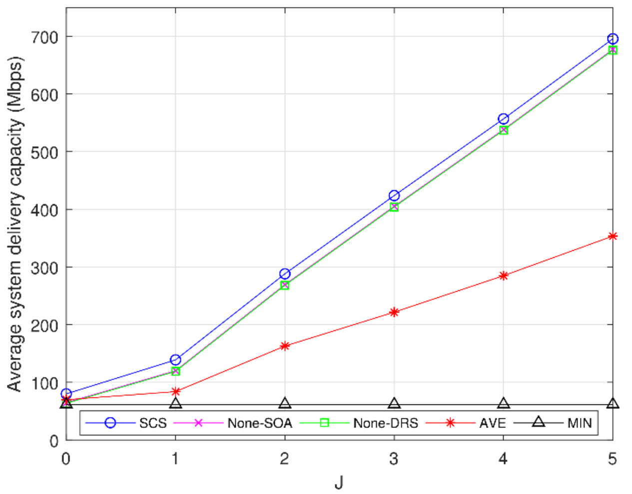

Figure 2-6. Capacity performance versus number of FBSs

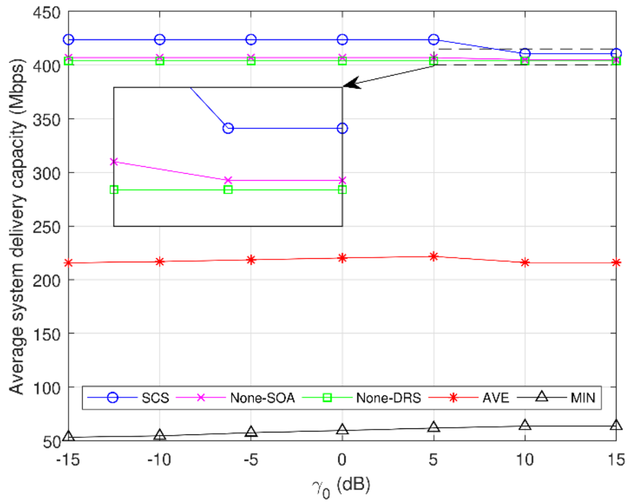

Figure 2-7. Capacity performance versus target SINR of SUs

Figure 2-6 plots the system performance versus the number of FBSs J. The results show that the number of FBSs significantly impacts the system performance. Unlike increasing the number of D2D pairs, increasing the number of FBSs in 5G UDNs makes the system delivery capacity improved rapidly without getting saturated because the constraints on cross-tier and co-tier interferences are not considered thanks to the use of channel splitting and FALOHA schemes [79,80].

2.5.3. System performance versus γ_0

We investigate the system performance under the impact of the target SINR of the SUs γ_0. As shown in Figure 2-7, the system delivery capacity of the SCS and None-SOA decreases and approaches that of the None-

DRS when γ_0 increases. The reason is that if γ_0 is low, more D2D pairs are shared the downlink resource from the SUs to offload the videos for higher system delivery capacity. Otherwise, if γ_0 is high, less D2D pairs offload the videos with lower system delivery capacity. The system delivery capacity of MIN increases because when γ_0 increases, more candidate matrices that cause higher interference impact on the SUs are removed from the search space $\mathcal{V}$. It can be seen obviously that the results of the None-DRS are not affected by γ_0 because there is no interference effect from D2D communications on the SUs. In this scenario, the proposed SCS also surpasses the other None-SOA, None-DRS, AVE, and MIN.

2.5.4. System performance versus distance between MBS and MUs

Finally, we evaluate the system performance versus the distance between the MBS and the MUs d1∊[5, 10]m, d2∊[10, 20]m, d3∊[20, 100]m, d4∊[100, 500]m, d5∊[200, 1000]m, d6∊[300, 1500]m, d7∊[400, 2000]m and d8∊ [500, 2500]m. In this scenario, we only consider the performance of the SCS with different number of FBSs J from 1 to 5. The results in Figure 2-8 show that if the number of FBSs is low (i.e., J = 1 and J = 2), the MUs are more likely to be served by the MBS at low system delivery capacity. It is easy to see that if the MUs are too close to the MBS, all the MUs are served by only the MBS at the same system delivery capacity. However, if the number of FBSs increases (i.e., J = 3, J = 4, and J = 5), the MUs are more likely to be served by the FBSs at higher system capacity delivery. Thanks to the high caching storage capacity of the FBSs (C^* = 0.7 × IJ), the MBS does not involve in serving the MUs event if the MUs are in very close proximity to the MBS. Interestingly, there exists the best region (i.e., d6 ∈ [300, 1500]m) in which the MUs are served the highest system capacity delivery when they are not very close to and not very far from the MBS. In other words, the MUs in this region are optimally served by all the MBS, the FBSs, and the TXs.

For further understanding, we consider the scenarios of lower caching storage capacity of the FBSs C^* = {0.6, 0.5, 0.4, 0.3} × I × J as shown in Figure 2-9.

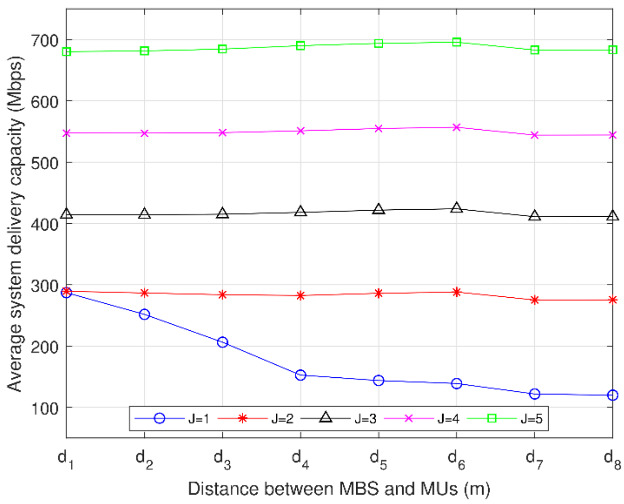

Figure 2-8. Capacity performance versus the distance between MBS and MUs

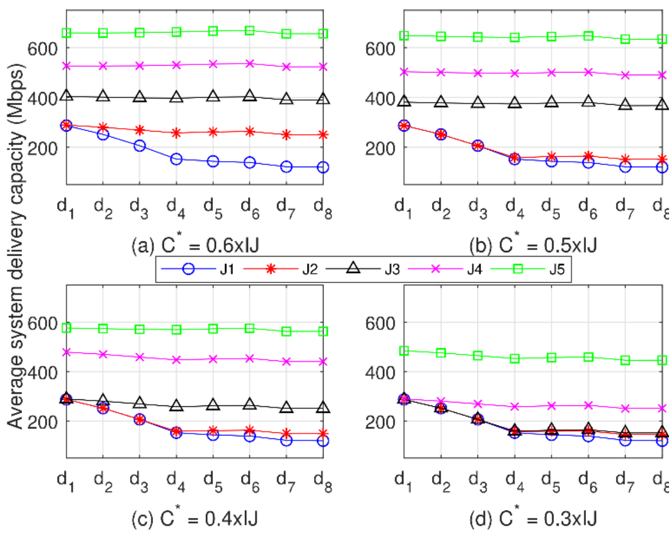

*Figure 2-9. Capacity performance versus the distance between MBS and MUs with lower C**

It can be seen that if the caching storage capacity of the FBSs is too low ($C^* = 0.3 \times I \times J$) (Figure 2-9d), he MUs are mostly served by the MBS when they are closer to the MBS. Especially, in case of d_1 ∊ [5, 10]m, the performance of the SCS is the same for all cases of J = 1, J = 2, J = 3 and J = 4 because all the MUs are served by only the MBS. Obviously, if C^* increases, the probability that the MUs are served by the MBS is reduced. It is important to conclude that the increasing of both C^* and J in 5G UDNs can relax the workload of the MBS.

2.6. Conclusion of chapter 2

We have proposed a social-aware caching and resource sharing (SCS) optimization solution for video delivering at high capacity in 5G UDNs. In particular, the storage of the MBS, the FBSs, and the TXs are utilized to bring the videos closer to the MUs and the downlink resources of the SUs are shared with the D2D pairs for

D2D communications. For more efficiency, we exploit the social relationships of the D2D pairs and the access rate (or the popularity) of the videos in our SCS optimization solution. The SCS optimization solution is carefully analyzed by taking into account the target SINR of the SUs to guarantee their QoS. This way, the SCS strategy can relax the workload at backhaul links of the MBS and the FBSs; provide the MUs with high cache-hit ratio video services by requesting the videos alternately from the MBS, the FBSs, and the TXs; and serve the MUs maximum video delivery capacity. The interesting findings are that 1) a proper duration set of videos selected in accordance with a given set of social relationships of D2D pairs can provide the MUs with the highest system delivery capacity and 2) there exists the best region in which the MUs are served the highest system delivery capacity. These findings can help both the Internet service providers and content providers insightfully understand the 5G UDNs to serve the MUs more efficiently.

However, the SCS scheme still has some limitations such as: 1) the inability to select the proper video versions to cache in the FBSs, to achieve higher system performance, 2) the impossibility to pair the caching users (CUs) that have the videos, with the normal users (NUs) that request the video, and 3) the system evaluation is delivery capacity (expressed in bps) that does not explicitly specify the video quality at the receiver. These existing problems will be solved in Chapter 3.

CHAPTER 3. USER-DEMAND-AWARE MULTI-RATE COOPERATIVE VIDEO CACHING AND DELIVERY IN MULTI-TIER 5G UDNs

3.1. Introduction to CRS

This chapter formulates a user-demand-aware multi-rate cooperative video caching and delivery (CRS) optimization problem for a high video streaming performance in 5G ultra-dense networks. On the one hand, the CRS enables the MUs to receive the video flexibly from MBS (macro tier), FBSs (femto tier), and D2D communications (user tier) in 5G UDNs. On the other hand, the CRS overcomes the issues about caching, sharing and system evaluation parameters of SCS scheme in Chapter 2.

About caching, the CRS considers the playback resolution of MUs, thereby determining the throughput required by MUs to cache different video versions to efficiently exploiting the available caching storage and spectrum resources. About resource sharing, the CRS allows a normal user (NU) to select a caching user (CU) and a user who agrees to share the spectrum (SU) to set up a tripartite (SU, CU, NU) for D2D communications, instead of matching an SU and a pair of (CU, NU) in the SCS scheme. Setting up a tripartite will expand the choices for getting the best D2D communications. About the system evaluation parameters, in this chapter, the system performance is evaluated more explicitly by converting the system delivery capacity to the peak signal-to-noise ratio (PSNR) measured in dB, through the relation between delivery capacity and playback quality using some computations of capacity success probabilities.

Motivated by the aforementioned discussions, the CRS problem is proposed to simultaneously implement 1) from which tier the videos are delivered, 2) which video version will be cached in which FBS, and 3) which tripartite (SU, CU, NU) will be matched? 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 the MUs. The results related to chapter 3 are [C2] and [C3] given in the “List of publication related to the dissertation” at the end of the document.

3.2. System model and Formulations

3.2.1. System Model

In this chapter, the CRS model for VASs in 5G UDNs illustrated in Figure 3-1. The notations used in this system model are presented in Table 3-1.

The model has one MBS in the macro tier, J FBSs in the femto tier, a number of MUs including K SUs, N CUs, and M NUs in the user tier, and I videos. 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 [48] and from the CUs to the NUs over D2D communications reusing the downlink resources of SUs.

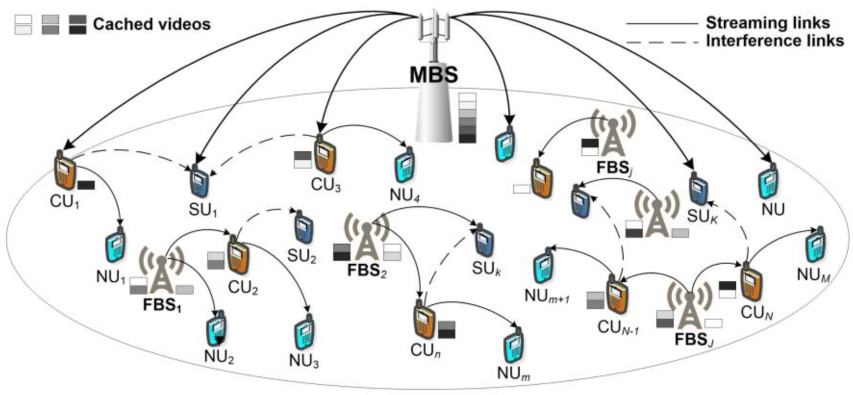

Figure 3-1. System models of CRS

Table 3-1. Notations of CRS

Symbols	Descriptions
I, J, K, N, M	Number of videos, FBSs, SUs, CUs, NUs
V_i	Video version *i, i=1, 2,…, I*
$w_{n,m}^{k}$	Downlink resource sharing index, i.e., $w_{n,m}^{k} = 1$ if SU k decides to share its downlink resource with the CU n and NU m in D2D communication and otherwise $w_{n,m}^{k} = 0$
$u_j^{v_i}$	Caching index, i.e., $u_j^{v_i} = 1$ if FBS j cache video version v_i and otherwise $u_j^{v_i} = 0$
P_M^0	Transmission power of the MBS
P_F^j	Transmission power of the FBS j, j = 1, 2, ..., J
P_T^n	Transmission power of the CU n of the D2D pair n
G_0^k	Channel gain from MBS to SU k
N_0	Power of additive white Gaussian noise (AWGN)
W	System bandwidth
L_F^i	Total storage consumed to cache all versions of the video i in the FBSs
L_{max}^i	Upper limit of total storage consumption of FBSs to cache all the versions of the video i, i.e., $L_{max}^i = r_i \times I \times J \times max\{L_i^{v_i},\ v_i = 1,2,\ \dots, V_i\}$, here $r_i = i^{-\alpha}\left(\sum_{i=1}^{I} i^{-\alpha}\right)^{-1}$ is the popularity of the video i with skewed popularity exponent α modeled as Zipf-like distribution [88]
C, C^*	Total throughput required by MUs and its upper limit
μ and δ	μ and δ where $0 < \mu, \delta \leq 1$ are used to flexibly adjust the upper limits L_{max}^i and C^*

Assume that the system parameters remain at least during a streaming session of the longest video version. Under this assumption, the CRS model efficiently serves the MUs the local VASs in crowded areas such as concert or meeting halls, museums, office buildings, stadiums, hospitals, campuses, etc. 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, as following:

- 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 FBS *j* cache 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.

3.3. System Formulations of CRS

3.3.1. Capacity Success Probability at SUs

The capacity success probability at the SU *k* is derived from its capacities over the channels from the FBS *j*

$$C_j^{k,v_i} = W\, log_2\left(1 + \frac{u_j^{v_i} P_j G_j^k}{N_0}\right) \tag{3.1}$$

And the MBS:

$$C_0^{k,v_i} = W\, log_2\left(1 + \frac{\prod_{j=1}^{J}(1 - u_j^{v_i}) P_0 G_0^k}{N_0 + I_{C,S}^{k,v_i}}\right) \tag{3.2}$$

When $I_{C,S}^{k,v_i} = \sum_{n=1}^{N}\sum_{m=1}^{M} w_{n,m}^{k}\, p_n^{v_i} P_C^k G_n^k$

and $p_n^{v_i}$ is the probability that the CU *n* caches the video version v_i, given by

$$p_n^{v_i} = ar_i + b\theta_n^{v_i} \tag{3.3}$$

when *a*, *b* ∈ [0, 1] and r_i is Zipf-like distribution.

Then, the capacity success probabilities at SU *k* from FBS *j* given by

$$p_j^{k,v_i} = Pr\{C_j^{k,v_i} \geq C_{th^{v_i}}\} = exp\left(\frac{-\xi_j^{k,v_i} N_0}{u_j^{v_i} P_j}\right) \tag{3.6}$$

And the capacity success probabilities at SU *k* from MBS given by

$$p_0^{k,v_i} = Pr\{C_0^{k,v_i} \geq C_{th^{v_i}}\} = exp\left\{-\xi_0^{k,v_i}\left[\lambda_C^{k,v_i}\left(\frac{P_C^k}{\prod_{j=1}^{J}\left(1-u_j^{v_i}\right)P_0}\right)^{\frac{2}{\eta}}\right]\right\} \tag{3.7}$$

when

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

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

$$p_{0,j}^{k,v_i} = 1 - \left(1 - p_j^{k,v_i}\right)\left(1 - p_0^{k,v_i}\right) \tag{3.8}$$

3.3.2. Capacity Success Probability at CUs

The capacities at the CU *n* over the channel from FBS *j* and MBS are expressed as

$$C_{j,n}^{v_i} = W\, log_2\left(1 + \frac{u_j^{v_i} P_j G_j^n}{N_0}\right) \tag{3.9}$$

and

$$C_{0,n}^{v_i} = W\, log_2\left(1 + \frac{\prod_{j=1}^{J}\left(1 - u_j^{v_i}\right) P_0 G_0^n}{N_0}\right) \qquad (3.10)$$

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

- From FBS *j* to CU *n*:

$$p_{j,n}^{v_i} = Pr\{C_{j,n}^{v_i} \geq C_{th}^{v_i}\} = exp\left(\frac{-\xi_j^{n,v_i} N_0}{u_j^{v_i} P_j}\right) \qquad (3.11)$$

- From MBS to CU *n*:

$$p_{0,n}^{v_i} = Pr\{C_{0,n}^{v_i} \geq C_{th}^{v_i}\} = exp\left(\frac{-\xi_0^{n,v_i} N_0}{\prod_{j=1}^{J}\left(1 - u_j^{v_i}\right) P_0}\right) \qquad (3.12)$$

when

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

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

$$p_{0,j,n}^{v_i} = 1 - \left(1 - p_{j,n}^{v_i}\right)\left(1 - p_{0,n}^{v_i}\right) \qquad (3.13)$$

3.3.3. 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

$$C_{n,m}^{k,v_i} = W\, log_2\left(1 + \frac{w_{n,m}^k p_n^{v_i} P_C^k G_n^m}{N_0 + P_0 G_0^m + I_{C,C}^{k,v_i}}\right) \qquad (3.14)$$

$$C_{j,m}^{k,v_i} = W\, log_2\left(1 + \frac{u_j^{v_i}\left(1 - w_{n,m}^k p_n^{v_i}\right) P_j G_j^m}{N_0}\right) \qquad (3.15)$$

and

$$C_{0,m}^{k,v_i} = W\, log_2\left(1 + \frac{\prod_{j=1}^{J}\left(1 - u_j^{v_i}\right)\left(1 - w_{n,m}^k p_n^{v_i}\right) P_0 G_0^m}{N_0}\right) \qquad (3.16)$$

here $I_{C,C}^{k,v_i} = \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} w_{n',m'}^k\, p_{n'}^{v_i} P_C^k G_{n',m'}^k$

So, the corresponding capacity success probabilities are

$$p_{n,m}^{k,v_i} = Pr\{C_{n,m}^{k,v_i} \geq C_{th}^{v_i}\} = exp\left\{-\xi_{n,m}^{v_i}\left[\lambda_M \left(\frac{P_0}{w_{n,m}^k p_n^{v_i} P_C^k}\right)^{\frac{2}{\eta}} + \lambda_C'^{k,v_i}\right]\right\} \qquad (3.17)$$

$$p_{j,m}^{k,v_i} = Pr\{C_{j,m}^{k,v_i} \geq C_{th}^{v_i}\} = exp\left[\frac{-\xi_{j,m}^{v_i} N_0}{u_j^{v_i}\left(1 - w_{n,m}^k p_n^{v_i}\right) P_j}\right] \qquad (3.18)$$

and

$$p_{0,m}^{k,v_i} = Pr\{C_{0,m}^{k,v_i} \geq C_{th}^{v_i}\} = exp\left[\frac{-\xi_{0,m}^{v_i} N_0}{\prod_{j=1}^{J}\left(1-u_j^{v_i}\right)\left(1-w_{n,m}^{k} p_n^{v_i}\right)P_0}\right] \tag{3.19}$$

where

- $\xi_{j,m}^{v_i} = \left(d_j^m\right)^{\eta}\left(2^{\frac{C_{th}^{v_i}}{W}}-1\right)$
- $\xi_{0,m}^{v_i} = \left(d_0^m\right)^{\eta}\left(2^{\frac{C_{th}^{v_i}}{W}}-1\right)$
- $\xi_{n,m}^{v_i} = \pi(d_n^m)^2\, \Gamma\left(1+\frac{2}{\eta}\right)\, \Gamma\left(1-\frac{2}{\eta}\right)\left(2^{\frac{C_{th}^{v_i}}{W}}-1\right)^{2/\eta}$
- d_n^m, d_j^m and d_0^m are the distances from the CU *n*, the FBS *j* and the MBS to NU *m*
- $\lambda'^{k,v_i}_{C} = \sum_{\substack{n'=1\\ n'\neq n}}^{N}\sum_{\substack{m'=1\\ m'\neq m}}^{M} w_{n',m'}^{k}\, p_{n'}^{v_i}$ the density within a circular cell area

Finally, the capacity success probability at the NU *m* is

$$p_{0,j,n,m}^{k,v_i} = 1-\left(1-p_{n,m}^{k,v_i}\right)\left(1-p_{j,m}^{k,v_i}\right)\left(1-p_{0,m}^{k,v_i}\right) \tag{3.20}$$

3.3.4. Average Quality of Received Videos

if the video version v_i is played back at rate or capacity $C_{th}^{v_i}$, the corresponding reconstructed distortion is

$$D_i\left(C_{th}^{v_i}\right) = \gamma_i\left(C_{th}^{v_i}\right)^{\beta_i} \tag{3.21}$$

where γ_i and β_i are the sequence-dependent parameters selected so that Eq. (3.21) meets the experimental RD curves. We can compute the overall average quality values of received videos at the Mus (i.e., SUs, CUs, NUs) as below

$$\overline{Q} = \frac{\sum_{j=1}^{J}\left(Q_S^j + Q_C^j + Q_N^j\right)}{3J} \tag{3.22}$$

where

$$Q_S^j = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i} p_{0,j}^{k,v_i} Q\left(D_i\left(C_{th}^{v_i}\right)\right) \tag{3.23}$$

$$Q_C^j = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i} p_{0,j,n}^{v_i} Q\left(D_i\left(C_{th}^{v_i}\right)\right) \tag{3.24}$$

$$Q_N^j = \frac{1}{KMN}\sum_{k=1}^{K}\sum_{N=1}^{N}\sum_{m=1}^{M}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i} p_{0,j,n,m}^{k,v_i} Q\left(D_i\left(C_{th}^{v_i}\right)\right) \tag{3.25}$$

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

3.4. CRS Optimization and Genetic Algorithms (GA)

3.4.1. CRS Optimization Problem

To implement 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 MU (*C*) which are considered in the constraints of the CRS optimization problem as below

$$L_F^i = \sum_{j=1}^{J}\sum_{v_i=1}^{V_i} u_j^{v_i} L_i^{v_i} \tag{3.26}$$

$$C = C_S + C_C + C_N \tag{3.27}$$

where

$$C_S = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i}\left(C_j^{k,v_i} + C_0^{k,v_i}\right) \tag{3.28}$$

$$C_C = \sum_{j=1}^{J}\sum_{n=1}^{N}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i}\left(C_{j,n}^{v_i} + C_{0,n}^{v_i}\right) \tag{3.29}$$

$$C_N = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{N=1}^{N}\sum_{m=1}^{M}\sum_{i=1}^{I}\frac{r_i}{V_i}\sum_{v_i=1}^{V_i}\left(C_{n,m}^{k,v_i} + C_{j,m}^{k,v_i} + C_{0,m}^{k,v_i}\right) \tag{3.30}$$

Finally, the CRS optimization problem is formulated as

$$\max_{u_j^{v_i}, w_{n,m}^k} \overline{Q} \tag{3.31}$$

$$s.t.\begin{cases} \sum_{v_i=1}^{V_i} u_j^{v_i} \le 1, i = I, \ldots, I, I1,2, \ldots, J, \\ \sum_{m=1}^{M} w_{n,m}^k \le Ik = 1,2, ..I, n = 1,2, \ldots, N, \\ \sum_{n=1}^{N} I \le 1, k = I, \ldots, K, m = 1,2, \ldots, M, \\ \sum_{k=1}^{K} w_{n,m}^k \le 1, n = 1,2, \ldots, N, m = 1,2, \ldots, M, \\ L_F^i \le \mu I, i = 1,2, \ldots, I, \\ C \le \delta C^* \\ I_{C,S}^{k,v_i} \le \frac{P_0 G_0^k}{\gamma_0} - N_0, k = 1,2, \ldots, K, I1,2, \ldots, I, v_i = 1,2, \ldots, V_i \end{cases} \tag{3.32}$$

3.4.2. Genetic Algorithms for CRS problem

We apply the GA tool introduced in [62] to solving the CRS optimization problem. The shortage of this GA tool is that it is able to deal with simple constraints of lower and upper bounds, but not with complicated constraints as aforementioned in. The solution for the complicated constraints is penalty method that converts constrained to an unconstrained optimization problem [99-102].

To do so, we rewrite the constraints (3.32) as below

$$\begin{cases} \Delta u_{i,j} = 1 - \sum_{v_i=1}^{V_i} u_j^{v_i} \ge 0, i = 1,2, \ldots, I, j = 1,2, \ldots, J, \\ \Delta w_{k,n} = 1 - \sum_{m=1}^{M} w_{n,m}^k \ge 0, k = 1,2, \ldots, K, n = 1,2, \ldots, N, \\ \Delta w_{k,m} = 1 - I \ge 0, k = 1,2, \ldots, K, m = 1,2, \ldots, M \\ \Delta w_{n,m} = 1 - \sum_{k=1}^{K} I \ge 0, n = 1,2, \ldots, N, m = 1,2, \ldots, M \\ \Delta L_F^i = \mu L_{max^i} - L_F^i \ge 0, i = 1,2, \ldots, I, \\ \Delta C = \delta C^* - C \ge 0, \\ \Delta I_{C,S}^{k,i,v_i} = \frac{P_0 G_0^k}{\gamma_0} - N_0 - I_{C,S}^{k,v_i} \ge 0, k = 1,2, \ldots, K, i = 1,2, \ldots, I, v_i = 1,2, \ldots, V_i. \end{cases} \tag{3.33}$$

Then, we derive the penalty function consisting of all the above constraints expressed in (3.33)

$$
\begin{aligned}
F = \; & \lambda_1 \sum_{i=1}^{I} \sum_{j=1}^{J} \left(min\{0, \Delta u_{i,j}\}\right)^2 + \lambda_2 \sum_{k=1}^{K} \sum_{n=1}^{N} \left(min\{0, \Delta w_{k,n}\}\right)^2 \\
& + \lambda_3 \sum_{k=1}^{K} \sum_{m=1}^{M} \left(min\{0, \Delta w_{k,m}\}\right)^2 + \lambda_4 \sum_{n=1}^{N} \sum_{m=1}^{M} \left(min\{0, \Delta w_{n,m}\}\right)^2 \\
& + \lambda_5 \sum_{i=1}^{I} \left(min\{0, \Delta L_F^i\}\right)^2 + \lambda_6 (min\{0, \Delta C\})^2 \\
& + \lambda_7 \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{v_i=1}^{V_i} \left(min\{0, \Delta I_{C,S}^{k,i,v_i}\}\right)^2
\end{aligned}
\tag{3.34}
$$

Finally, the GA solution is able to solve the following unconstrained optimization problem

$$
\max_{u_j^{v_i}, w_{n,m}^k} \; \overline{Q}_F = \overline{Q} - F. \tag{3.35}
$$

Algorithm 3.1. Genetic Algorithm

1	**Input**: System parameters as shown in Table 3-2
	N_P = 20,000; Population, i.e. number of individuals
	$PRECI = J*I*V_i + K*N*M$: Number of bits of a chromosome or a binary string
	$P_G = 0.9$: Generation gap
	$P_C = 0.9$: Crossover probability
	$P_m = 10^{-6}$: Mutation probability
	Generating the initial generation including N_P random individuals of solution set $\{X_z\}$, $z = 1, 2, \ldots N_P$
	Gen = 0: Number of Generations
	TC: Termination conditions
2	**Output**: X^* and $\bar{Q}_F^*$
3	**While** *TC* does not hold **do**
4	Gen = Gen + 1
5	Putting $\{X_z\}$ and $\bar{Q}_F(X_z)$ in the mating pool for ranking
6	Select $N_{PG} \times P_G$ better individuals in $\{X_z\}$ with higher $\bar{Q}_F(X_z)$ for breeding the next generation by using stochastic universal sampling operator
7	Selecting a pair of parents to generate a pair of offsprings by using double point crossover operator. The crossover operator is not required to be applied to all the chosen pairs, but done with a crossover probability P_C
8	Mutating the offsprings with a mutation probability P_M to recover good genetic materials
9	Evaluating the fitness values of the offsprings, reinserting them into the present generation
10	**End While**
11	Finding the best fitness value $\bar{Q}_F^*$ with respect to the best individual X^* in the last generation

3.5. Performance Evaluation of CRS and Genetic Algorithm

3.5.1. System parameters and Computer Information

To evaluate the CRS and the GA performance, the CRS model is deployed with the parameters set in Table 3-2. 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, [1, 20]m. In addition, we take into account 3 videos, i.e., $V_1^{v_1}$ (Basketballpass), $V_2^{v_2}$ (Racehourses) and $V_3^{v_3}$ (Foreman) to analyze their experimental RD curves by using HM reference software version 12.0 [52] and obtain $L_i^{v_i}$, $C_{th}^{v_i}$, γ_i and β_i given in Table 3-2.

Table 3-2. System Parameters

Symbols	Specification
J, K, N, M	5 FBSs, 5 SUs, 5 CUs, 10 NUs
I	3 videos
V_i	3 videos, each has 3 versions
W, P_0 , P_j, P_C^k	5MHz, 5W, 1W, 1mW
N_0, η, γ_0	10^{-12}W, 4, 10dB
$\{a, b\}$	{0.5, 0.5}
γ_i	[9806 76520 1644000]
β_i	[-0.9972 -1.1530 -1.0920]
μ, δ	0.5, 1
C^*	10 Gbps, i.e., each MU is served at 0.5 Gbps

Symbols	Specification
r_n	Fix to 1, i.e. all CUs have 100% of storage to cache.
$L_i^{v_i}$	[11867 23734 35600 198680 264906 351000 33382 66763 113496 1172340 1758510 2344680 160410 320820 453960] Kbit
$C_{th}^{v_i}$	[1000 2000 3000 3000 4000 5300 50000 100000 170000 10000 15000 20000 10000 20000 28300]Kbps

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). A common computer with detailed information listed in Table 3-3 was used to simulate and evaluate.

Table 3-3. Computer Information

Computer Information	
Processor	Intel(R) Core (TM) i7-8700 CPU @ 3.20GHz
Processor type	x64 Family 6 Model 14 Stepping 10, Genuine Intel
PHY processor packages	1
Processor cores	6
Logical processors	12
Total PHY memory	16,599,444KB
Operating System	Windows 10

3.5.2. Performance of GA

To evaluate the GA performance, binary bat algorithms (BBA) and exhaustive algorithms (EA) are implemented to compare the accuracy, and time complexity that enable to choose a set of proper parameters for GA to solve the CRS problem. The BBA includes the parameters (L_i^0 = 0.25, E_i^0 = 0.1, f_{min} = 0, f_{max} = 2, ε = 0.9, ς = 0.9) [61]. Similar to the GA, the penalty function method is also applied to the BBA. Due to the difference of input parameters and operators used in GA and BBA, in this dissertation, GA and BBA are evaluated based on 1) equivalence in execution time, eg., N_{PG} =20,000 in GA, the average time complexity is 1,003.68 seconds and N_{PB} =10,000 in BBA, the average time is 957.12 seconds, 2) the accuracy compared to EA and 3) the stability (EA have been implemented in Chapter 2). In this case, the system parameters include (J, N, M, K) set with low values of (3, 3, 5, 3) for implementing EA.

Table 3-4. Comparation between GA, BBA and EA

α	0	0.5	1.0	1.5	2.0	2.5	3.0
EA (dB)	29.0286	31.2176	32.6980	34.1178	36.0832	37.2044	38.0584
BBA (dB)	28.9100	31.2176	32.4708	33.6306	36.0832	36.5090	37.6689
Accuracy	99.59%	100.00%	99.31%	98.57%	100.00%	98.13%	98.98%
GA (dB)	29.0286	31.2176	32.6924	34.1178	36.0716	37.1935	38.0476
Accuracy	100%	100%	99.98%	100%	99.97%	99.97%	99.97%

As we can see in Table 3-4, comparing to the EA, i.e., the exact results, the GA can achieve the accuracy up to 100%. With the same execution time, GA shows more accuracy than BBA. Furthermore, when executing the algorithm 100 times (with α = 1), GA gives results more stable than BBA as shown in Figure 3-2. From these results, GA is selected to apply to solve the CRS problem with larger system parameters.

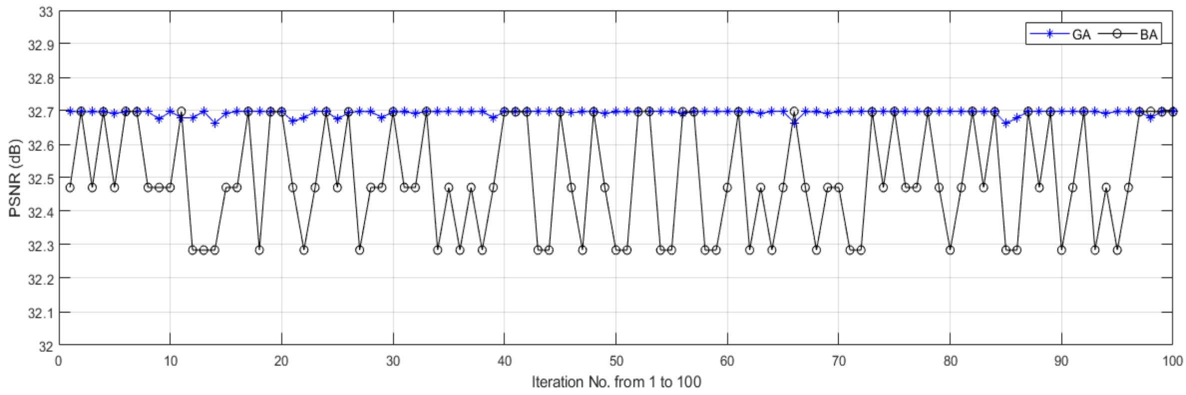

Figure 3-2. Evaluate the stability of GA and BBA

We further evaluate the trade-off between the accuracy and the time complexity of GA versus different populations (N_P) while keeping $\alpha = 1$. As shown in Table 3-5, the GA is done 100 iterations and chooses the worst case with the highest difference in PSNR compared to the EA:

- The results clearly demonstrate that if we accept the low accuracy (i.e., 97.34%), the time complexity is extremely low (19.58s) at $N_P = 1000$.
- The accuracy increases versus the increase of N_P with higher time complexity.

The accuracy cannot be significantly improved when N_P is high enough, i.e., N_P = 10,000. Importantly, at N_P = 10,000, the GA obtains a very high accuracy of 99.23% at very low time complexity of 317.42s that yields a significant time complexity reduction compared to the EA done at 5,664.45s. Based on the results in Table 3-4. Comparation between GA, BBA and EA, we set N_P = 20,000 to implement the GA in the large scale of 5G UDNs to ensure a high accuracy at a reasonable time complexity.

Table 3-5. GA performance in worst cases versus N_P

Metrics	N_P (α = 1)					EA
	1,000	5,000	10,000	15,000	20,000	
Time (s)	19.58	142.50	317.42	556.01	795.63	5,664.45
PSNR (dB)	30.8283	32.2547	32.4474	32.6486	32.6631	32.6980
Accuracy	97.34%	98.64%	99.23%	99.85%	99.89%	100%

To fulfill the GA performance evaluation, the GA with the convergence criteria (or the terminate condition) is presented in the Algorithm 3-1. The GA convergence rate is shown in Figure 3-3, where the "Best" is the highest $\bar{Q}_F(X_z)$ with respect to the best individual of each generation, the "Mean" is average of $\bar{Q}_F(X_z)$ calculated for all individuals of a generation, and the "Error" is the value of penalty function F in (3.34) and (3.35). The results reveal that the GA begins to converge after about 20 generations and actually get converged situation from the 125-th generation when meeting the convergence criteria. Here, "Error" comes to 0 meaning that all constraints are satisfied. The "Best" value does not change and is equal to "Mean", meaning that all individuals are as good as each other in the last generation. As a result, it is certain that the GA is more flexible and feasible to apply to solving the CRS optimization problem in the large scale of 5G UDNs.

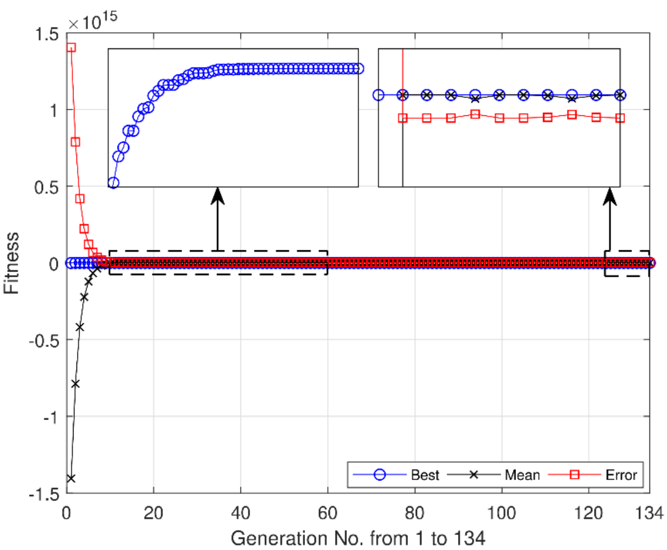

Figure 3-3. GA convergence rate

3.5.3. Performance Evaluation for CRS

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) when changing the number of FBSs in Figure 3-4. If there are not any FBSs caching the video versions (J = 0), the OCC becomes the NCS and both have the same lowest performance due to no caching nor sharing. The CRS without caching becomes the ORS at the same performance but gaining higher performance than the OCC and the NCS since the CRS and the ORS benefit from downlink resource sharing. When increasing J, the performance of ORS and NCS keeps unchanged, while the performance of CRS and OCC significantly increases thanks to caching. The performance of CRS and OCC versus the increase of J gets saturated when J is high enough. It means that it is not necessary to cache many videos in all FBSs. Alternately, the number of FBSs caching the videos must be carefully selected to save the caching storage consumption as well as reduce the complexity of CRS strategy.

Next, the CRS is compared to the OCC, the ORS, the NCS versus the number of SUs agreeing to share the downlink resources in Figure 3-5. If there are not any SUs agreeing to share the downlink resources, the CRS and the ORS become the OCC and the NCS, respectively. The increase of the SUs agreeing to share the downlink resources does not impact the performance of both the OCC and the NCS, but it yields more downlink resource sharing possibilities to increase the performance of CRS and ORS. Importantly, the CRS provides the highest playback quality of VASs compared to the other schemes.

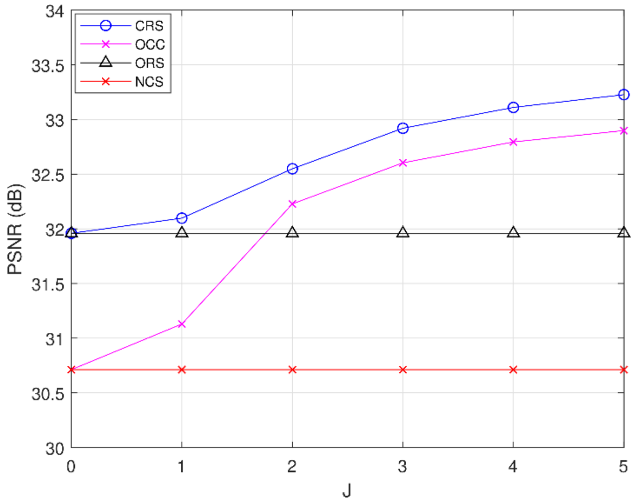

Figure 3-4. Performance of CRS, OCC, ORS, and NCS versus FBSs

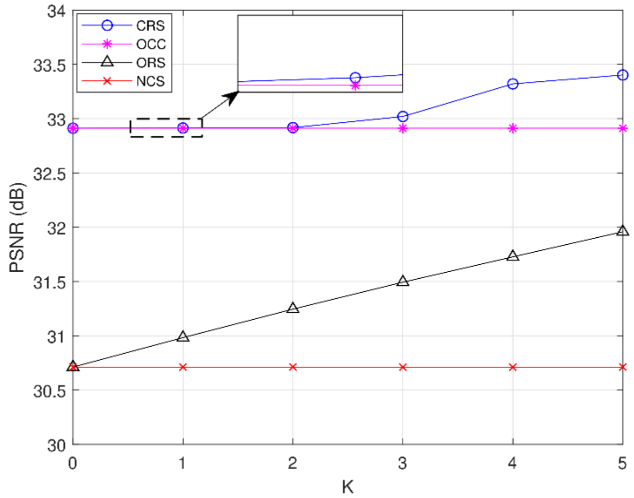

Figure 3-5. Performance of CRS, OCC, ORS, and NCS versus the number of SUs agreeing to share the downlink resources

Figure 3-6 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.

Then, we investigate the effects of the constraints on the performance of CRS in Figure 3-7. To do so, decreasing the required throughput of MUs (δ) from 1 to 0.5, decreasing the caching storage of FBSs (μ) by changing from 0.5 to 0.3 and increasing the target SINR of SUs (γ_0) from 3dB to 5dB.

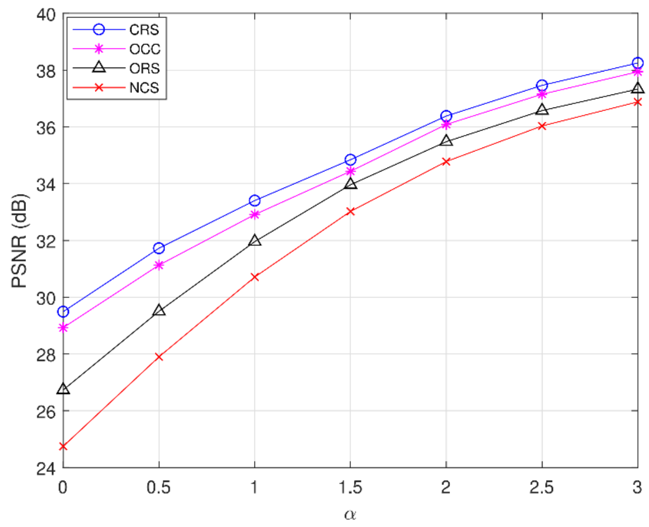

Figure 3-6. Performance of CRS, OCC, ORS, and NCS versus α

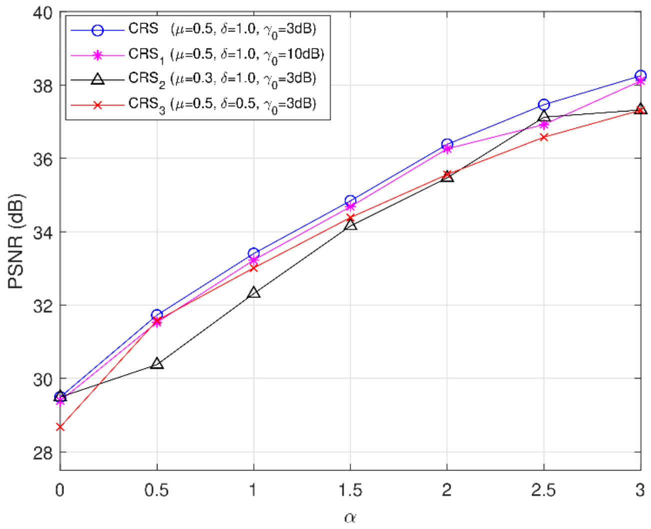

Figure 3-7. Performance of CRS and other system parameters versus α

It is obvious that if 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 playback resolutions of the MUs. In case of decreasing the caching storage of FBSs μ (from 0.5 to 0.3), the system has less chances to cache and thus provides lower playback quality. In addition, when the target SINR γ_0 increases from 3dB to 5dB 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.

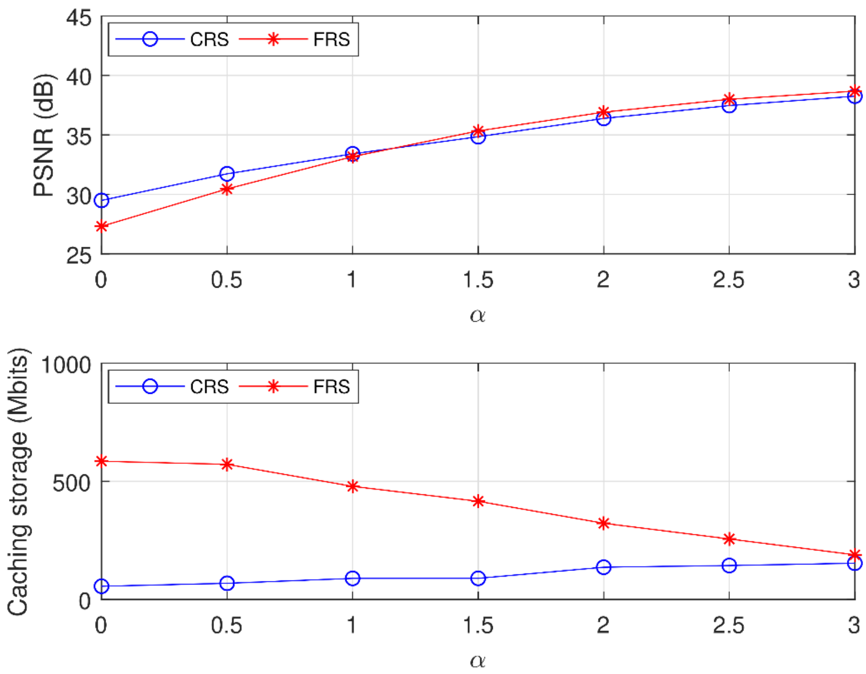

Figure 3-8. Performance of CRS and FRS

Finally, 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 Figure 3-8, the FRS outperforms the CRS when α > 1.

This can be explained that, when α is low, the system will serve MUs with videos version of the same popularity. In this case, selecting the right video version to cache and serve MUs will increase the system performance (satisfying MUs with less storage consumption). On the contrary, 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.

3.6. Conclusion Chapter 3

Chapter 3 have presented a user-demand-aware multi-rate cooperative video caching and delivery (CRS) scheme for high video streaming performance in 5G. In this chapter, the existing problems of Chapter 2 have been improved such as 1) QoS is evaluated more explicitly, 2) efficiently exploiting the available caching storage and spectrum resources of CUs and SUs and 3) the genetic algorithm is applied to solve the CRS problem efficiently.

CONCLUSION

1. Achieved results

The two main contributions of the dissertation are summarized as follows:

1- Propose a social-aware cooperative video caching and video delivering (SCS) scheme by exploiting the available caching storage and spectrum resources in multi-tier 5G ultra-dense network (UDN). The results were published in Mobile Networks & Application journal [C1] and presented at the Heterogeneous Networking for Quality, Reliability, Security and Robustness conference (Qshine 2018) [C2].

2- Propose a user-demand-aware multi-rate cooperative video caching and delivery (CRS) scheme by efficiently exploiting the available caching storage and spectrum resources in multi-tier 5G UDN. In particular, the CRS improves the SCS by 1) considering the video playback quality as the objective function, which is more visual than the video delivery capacity formulated in the SCS, 2) gaining higher opportunistic reuse of storage and spectrum resources, and 3) applying genetic algorithm (GA) to solve the optimal problem in large-scale 5G UDN. The results were published in IEEE Communications Letters [C3] and presented at the Recent Advances on Signal Processing, Telecommunications & Computing conference (SigTelCom2020) [C4].

2. Further research

Some future research directions that can be developed from the dissertation given below:

1- Improving the system model: A more effective system model needs to 1) consider videos with different versions for higher caching performance in multi-tier 5G UDN, 2) utilise spectrum resources for device-to-device (D2D) communications and 3) take into account both social- and user-demand-aware as well as user mobility. In addition, another important caching tier at unmanned aerial vehicles (UAV) must be added.

2- Improving evaluation criteria: A set of parameters to evaluate the quality of user experience (QoE) with strict criteria needs to be proposed. This set is to maximize not only the video delivery capacity or video playback quality, but also the video access rate, continuous playback, and playback stability.

3- Improving algorithms: when the system is expanded and becomes more complicated together with strict QoE evaluated, the GA needs to be further studied in terms of proper accuracy and execution time. In addition, emerging algorithms such as machine learning, deep learning, etc. should be studied and compared to the current GA to select the best algorithm for solving the large-scale optimization problems in 5G UDN.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

[C1] **Minh-Phung Bui**, Nguyen-Son Vo, Sang Quang Nguyen, and Quang-Nhat Tran, "Social-Aware Caching and Resource Sharing Maximized Video Delivery Capacity in 5G Ultra-Dense Networks," *ACM/Springer Mobile Networks & Applications,* pp. 1-13, July 2019;

[C2] **Minh-Phung Bui**, Nguyen-Son Vo, Tien-Thanh Nguyen, Quang-Nhat Tran, and Anh-Tuan Tran, "Social-aware Caching and Resource Sharing Optimization for Video Delivering in 5G Networks," in Proc. *EAI International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (Qshine'18)*, Ho Chi Minh City, Vietnam, Dec. 2018, pp. 73-86;

[C3] Nguyen-Son Vo, **Minh-Phung Bui**, Phuc Quang Truong, Cheng Yin, and Antonino Masaracchia, "Multi-tier Caching and Resource Sharing for Video Streaming in 5G Ultra-dense Networks," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1500-1504 , July 2020;

[C4] **Minh-Phung Bui**, Nguyen-Son Vo, Tien-Vu Truong, Thanh-Hieu Nguyen,Nam Van Nguyen and Cheng Yin, "Genetic Algorithms for Multi-tier Caching and Resource Sharing Optimized Video Streaming in 5G Ultra-dense Networks," in *Proc. of International Conference on Recent Advances on Signal Processing, Telecommunications & Computing* **(SigTelCom'20)**, Ha Noi, Vietnam, Aug. 2020, pp. 66-71;