IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2018, accepted May 12, 2018, date of publication May 22, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2839669

5G Optimized Caching and Downlink Resource Sharing for Smart Cities

NGUYEN-SON VO[1], TRUNG Q. DUONG[2], (Senior Member, IEEE), MOHSEN GUIZANI[3], (Fellow, IEEE), AND AYSE KORTUN[2]

[1]Duy Tan University, Da Nang, Vietnam
[2]Queen's University Belfast, Belfast BT7 1NN, U.K.
[3]University of Idaho, Moscow, ID 83844, USA

Corresponding author: Nguyen-Son Vo (vonguyenson@dtu.edu.vn)

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council under Grant EP/P019374/1, in part by a Research Environment Links Grant through the Newton Programme Vietnam Partnership under Grant 339568416, and in part by Newton Prize 2017.

ABSTRACT In smart cities, millions of things, systems, and people are interconnected and communicate with each other over wireless sensor networks, Internet of Things (IoT), and 5G networks. A tremendous amount of data traffic, which is frequently generated by the things in wireless multimedia sensor networks (WMSNs) and/or IoT, is accessed by a massive number of mobile users (MUs). These MUs are all competing to access the 5G network for data as well as urban applications and services. This can in turn cause exhaustion to the 5G network. In such cases, users can experience low data delivery and traffic congestions through backhaul links by macro base stations (MBSs). In this paper, we propose a joint caching and downlink resource sharing optimization framework (CSF) in 5G networks to assist WMSNs to efficiently deliver multimedia contents to the MUs. The CSF enables the MBSs to optimally decide how many replicas of each multimedia content to cache in which fem to base stations for high multimedia content hit rate. It also optimally exploits the MUs that are willing to share their downlink resources and that have retrieved multimedia contents, for offloading with device-to-device communications. The objective is to eventually maximize the system delivery capacity. Simulation results demonstrate that the CSF provides the best performance in terms of hit rate and system delivery capacity.

INDEX TERMS 5G caching, D2D caching, D2D communications, downlink resource sharing, fem to caching, femtocell, macrocell, smart city, wireless multimedia sensor networks.

I. INTRODUCTION

In megacities, the proliferation of urban residents poses various challenges to job pools, economic development, sustainable environment, and welfare and social resilience. These challenges can be overcome in the context of smart cities by adopting disruptive information and communications technologies (ICT). Basically, a multi-layer platform of ICT system is deployed, consisting of the things (sensors, mobile and wearable devices, cameras, actuators, and machines) to capture data from the surroundings, communication systems to deliver the data, and data centers for the purpose of monitoring and analyzing. A smart city can provide many advanced applications and services (AASs). These applications can include surveillance, management, and entertainment, relying on wireless sensor networks (WSNs), Internet of things (IoT), and 5G networks. These can play a prominent role as parts of being "smart" to allow the data to be transmitted efficiently over the system at low cost [2].

By 2020, smart cities will witness the ongoing growth of up to more than 50 billion connected devices together with a huge amount of data conveyed in the era of IoT [3]. The data of structural health of buildings, transportation systems, e-health and e-learning services, smart home and environment, and public safety applications. This has been generated by the things in WSNs and/or IoT and by the AASs in smart cities, is stored in data centers. A large portion of the data traffic is requested by a massive number of mobile users (MUs), which will reach 11.6 billion by 2021 [4], via the macro base stations (MBSs) in 5G networks. Simultaneously, the MBSs superimpose to provide the MUs with conventional cellular traffic. These in turn make 5G networks congested and thus low data delivery capacity performance due to the problem of traffic congestion in the backhaul links of the MBSs.

In fact, it is rigorous to develop such high-speed backhaul links of the MBSs to improve the delivery capacity performance due to their costly production [5]. Meanwhile, several technical solutions have been widely altered by the innovation of network architecture and optimization designs such as ultra-dense small-cell technology [6], device-to-device (D2D) communications [7], massive multiple-input multiple-output (MIMO) [8], and optimization designs [9]. However, these solutions are basically under the constraints on spectrum resources and signaling overhead. Thus, it is not sufficient to tackle the problems of traffic congestion in the backhaul links of the MBSs and high demand of massive number of MUs [10].

Recently, caching and spectrum resource sharing schemes have been applied to all the bodies and tiers of 5G networks, from the user devices and small-cell base stations (SBSs) to the MBSs. Caching, i.e., D2D caching [7], [11]–[14], femtocaching [15]–[18], small-cell caching [5], [19]–[23], MBS caching [24], [25], multi-tier caching [6], [26]–[32], can mitigate the traffic congestion in the backhaul links of the MBSs by placing the strategic data in close proximity to the MUs. And, spectrum resource sharing, i.e., [33]–[38], can improve the performance of spectral use and reuse to serve a massive number of MUs. To the best of our knowledge, in current literature there is a lack of research works in integrating WSNs into 5G networks. We believe this is needed to provide MUs with AASs in smart cities and offer a joint solution of caching and spectrum resource sharing in 5G networks to efficiently improve the data delivery capacity performance and mitigate the backhaul bottlenecks at the MBSs. Even though in many AASs of a smart city, the integration of WSNs and mobile cellular networks (or 5G networks) has been considered as a feasible architecture for collecting sensed data [39]–[42], the benefits of caching and spectrum resource sharing in 5G networks have not been utilized.

In this paper, to provide the MUs with AASs in smart cities, we extend our previous work [1] by taking into account the integration of wireless sensor networks (WSNs) and 5G networks in the context of smart cities with different challenges and more efficient solutions. Particularly, in [1], the scenario was "disaster", and thus the performance metrics focused on ensuring high connectivity and high response due to severe disruptions in the communication networks. However, in smart cities, the connectivity and response metrics are not important. Instead, smart cities face congestion problems because there is a huge amount of rich data generated by WSNs and requested by a massive number of MUs throughout 5G networks. To this end, we focus on how to maximize the average number of replicas of contents to provide the MUs with AASs at high hit rates. Importantly, during providing a high hit rate, we also carefully consider the throughput of each content requested by the MUs from WSNs to limit the number of replicas not exceeding this requested throughput. It makes our proposed caching scheme more efficient corresponding to the demand of MUs.

In addition, in [1], to maximize the system capacity, the decision to cache does not depend on the condition of storage capacity of the D2D transmitters (TXs) and the access rate of multimedia contents. The decision to share does not consider the interference effect of D2D communications on the target SINR of the CUs. In this work extension, all the storage capacity of the TXs and the access rate of contents as well as the interference effect of D2D communications on the target SINR of the CUs are considered to maximize the system delivery capacity more efficiently. The objective is to deliver a huge amount of data traffic to the MUs at a high hit rate and a high system delivery capacity. Meanwhile, we mitigate the tremendous traffic in the backhaul links of the MBSs. We specifically focus on delivering multimedia contents. This accounts for a large portion of data traffic generated by wireless multimedia sensor networks (WMSNs) consisting of WSNs and multimedia devices such as cameras and microphones.

To do so, we design a joint caching and downlink resource sharing optimization framework (CSF) consisting of two optimization solutions. In the first solution, we assume that the hit rate of a multimedia content can be represented by its number of replicas associated with its access rate or popularity and we formulate the number of replicas optimization (NRO) problem. The NRO problem is then solved for the optimal number of replicas to maximize the average number of replicas, and thus providing high hit rate. In the second solution, the system delivery capacity is maximized by solving where to cache and with whom to share optimization (CSO) problem. The objective is to find the optimal set of femto base stations (FBSs) to cache the replicas and the optimal set of MUs. These MUs share their downlink resources and have retrieved multimedia contents for offloading with D2D communications. The contributions of our work are summarized as follows:

- We propose a CSF consisting of NRO and CSO solutions applied to smart cities that can exploit the storage to cache and downlink resource to share in 5G networks. The CSF can convey the WMSNs' multimedia contents to the MUs at a high hit rate and high delivery capacity.
- The NRO problem is formulated and solved for optimal number of replicas of each multimedia content to maximize the average number of replicas for a high hit rate. The optimal results, namely the number of replicas to cache, are found in accordance with the popularity and the demand of multimedia contents to prevent from wasted throughput, e.g., reserved throughput is much larger than requested throughput.
- After knowing how many replicas to cache, we continue to answer the questions of where (which FBSs) to cache and who (which MUs) to share their downlink resources and available retrieved multimedia contents by formulating and solving the CSO problem? The optimal solution to the CSO problem enables the WMSNs'

multimedia contents to be delivered to the MUs from the MBSs, FBSs, and by D2D communications. This way, the system delivery capacity is maximized. In addition, the effect of interference of D2D communications on the MUs that share their downlink resources with D2D communications, is strictly considered as a constraint in the CSO problem to ensure the target signal to interference plus noise ratio (SINR) of the sharing MUs.

- Simulation results are insightfully discussed and comparison is presented between different schemes to demonstrate the benefits of the proposed CSF.

The rest of this paper is organized as follows. In Section II, related works are reviewed. We introduce the system models, i.e., the CSF in 5G networks for WMSNs and the formulations of a hit rate and system delivery capacities, in Section III. Section IV presents the NRO and CSO problems and solutions. Section V is dedicated to demonstrating the benefits of the proposed optimization framework compared to other schemes through simulation results and discussions. Finally, we conclude the paper in Section VI.

II. RELATED WORKS

In this section, we review some key caching [5]–[7], [11]–[32] and spectrum resource sharing [33]–[38] techniques, which are used to mitigate the traffic congestion in the backhaul links of the MBSs and improve the spectrum efficiency, in order to serve a massive number of MUs in 5G networks.

A. CACHING TECHNIQUES

1) D2D CACHING

D2D caching is one of the promising technologies at the edge of 5G networks to tackle the congestion in the backhaul links of the MBSs [7], [11]–[14]. In [7], a multihop network is established relying on D2D communications for fast content distribution, without occupying the backhaul links of the MBSs. A cross-layer optimization is designed by jointly routing the contents at the network layer and allocating the spectrum at the media access control layer to minimize the average delay in multihop D2D networks. Aiming at maximizing the content delivery probability in a stochastic D2D communication network, probabilistic caching placement optimization problems have been studied in [11] and [12]. The results show that a cache-aided throughput-based maximization approach provides higher content delivery probability compared to a cache hit probability-based one [11]. Caching the most popular contents is the simple way to maximize the content delivery probability [12]. For more practical applications, Wang *et al.* [13] further considered the mobility characteristic of the MUs in the caching placement strategy. The proposed mobility-aware caching placement strategy outperforms both random caching and popular caching strategies in terms of data offloading ratio. Park *et al.* [14] have proposed a smart MBS-assisted partial-flow D2D offloading system that can handle the multimedia contents, find the MUs who have cached the contents, and obtain the whole traffic for offloading with D2D communications, so as to support seamless multimedia services. However, the proposed caching placement optimization solutions have not considered the available storage resource in the SBSs and the MBSs, together with the MUs, for cooperatively delivering multimedia contents.

2) FEMTOCACHING

Utilizing considerable storage resource in the FBSs has been studied to alleviate the bottleneck at the MBSs [15]–[18]. In particular, Golrezaei *et al.* [15] have proposed a new multimedia distribution architecture based on the collaboration of femtocaching and D2D communications to increase the system throughput. The problem of which FBSs to cache the multimedia contents associated with their popularities was also solved to minimize the average time for multimedia donwloading [16] and to maximize the number of MUs served by the FBSs [17]. Especially Vo *et al.* [18] have designed a femtocaching framework that not only minimizes the bandwidth consumed at the MBSs and wasted at the FBSs but also provides high playback quality and high hit rate. This can be done by a joint technique of femtocaching and layered multiple description coding with embedded forward error correction for video transmissions. The existing problem in [15]–[18] is a lack of exploiting D2D caching to provide the MUs with high hit rate and high system deliver capacity.

3) SMALL CELL CACHING

Caching in small-cell 5G networks has been drawing much attention as a disruptive solution for offloading the backhaul links of the MBSs [5], [19]–[23]. In [5] and [19], proactively caching in SBSs has been proposed based on the fact that storing popular contents at the SBSs and exploiting the MUs' social relationships can efficiently improve the performance of caching in terms of less backhaul congestion, low delay to the MUs, and high network savings. Stochastic geometry theory has been applied to caching in SBSs to derive theoretical download probability, which enables to analyze the caching performance and optimize the caching probability of each content group [20]. Directly focusing on minimizing the backhaul load, Liao *et al.* [21], [22] have found the optimal content placement in all the SBSs by restructuring the contents with maximum distance separable (MDS) codes. In addition, mobility of the MUs, storage constraint of the SBSs, and delay deadline and popularity of the contents were carefully included in optimal distributed caching policy to minimize the traffic load at the MSBs [23]. However, the aforementioned works have not taken into account the caching collaboration in all the MBSs, SBSs, and MUs, to maximize the system delivery capacity.

4) MBS CACHING

A simple and efficient way to cache in 5G networks can be done at the MSBs [24], [25]. Following the same approach of MDS codes with [21] and [22] for caching, online cache content placement was designed in [24]. The design allows to evaluate the trade-off between caching at the MBSs and caching at the fronthaul of high-capacity core network, which is directly connected to the MBSs. The performance shows that the former, which is not better than the latter in terms of number of served MUs, is still much better to be applied to gain higher spectral efficiency in future 5G networks. It is interesting in [25] that mmWave small cells with directional antennas can be utilized to proactively cache the video contents at the MBSs. Allocating proper storage of each MBS to the MUs can provide the high-mobility MUs with high-quality mobile video streaming at low connection and retrieval delays. It is observed that the joint solution of caching in multiple MBSs, SBSs, and MUs has not been exploited to extend the performance of MBS caching.

5) MULTI-TIER CACHING

A more efficient approach to caching in 5G networks is multi-tier caching [6], [26]–[32]. In [6], caching has been extended from MBSs to MUs using a social group utility maximization game that exploits the social trust and physical reciprocity of the MUs. This way, the social group cost, i.e., the incentive to cache in the MUs for D2D communications, is minimized. By jointly caching in the FBSs (or MBSs) and MUs, Jiang *et al.* [26] (or [27]) have solved an optimal cooperative content caching and delivery problem for the best caching set of FBSs (or MBSs) and MUs to reduce the average downloading latency and enhance the local cache hit rate. In a green approach, maximizing the cache hit rate at high energy efficiency can be achieved by further taking into account the MUs' mobility [28] or mitigating traffic and energy consumption at the backhaul links can be done by a joint optimal technique of transmission and D2D and MBS caching policies [29], [32]. A more complicated solution for optimal caching placement in three-tier 5G networks including MBSs, FBSs, and pico base stations, has been studied to maximize the hit probability [30] and system capacity [31]. Although providing a general multi-tier caching architecture in 5G networks, [6], [26]–[32] have not taken the advantage of both storage and spectrum resource sharing schemes for D2D communications, to gain higher system delivery capacity.

B. SPECTRUM RESOURCE SHARING TECHNIQUES

In 5G networks, solutions for efficient spectrum use, reuse, sharing, and management are essential to meet the proliferation of MUs and of their demand for AASs due to the limitation of spectral resource [33]–[38]. In particular, Akhtar *et al.* [33] have overcome the problem of inaccurate decision of cognitive spectrum sensing, seriously spectrum sharing in dense D2D communications, by proposing a synergistic spectrum sharing mechanism. The results show that the proposed mechanism is reliable for high spectrum efficiency. In [34], storage resource of MUs and spectrum resource in D2D underlaid cellular networks were exploited to cache and share respectively, for high successful transmission probability. Mainly focusing on block-fading environment, an optimal power policy was proposed to minimize the spectrum sharing outage probability of the secondary users [35]. It is interesting that the benefits of spatial gain and spectrum sharing can be utilized to provide a significant improvement of sum rate in full duplex D2D underlaying cellular networks [36] and massive MIMO systems [37]. Based on the fact that mitigating the interference between the SBSs and MBSs can improve the throughput, Mach and Becvar [38] have designed a centralized algorithm to dynamically switch between overlay mode and underlay mode for spectrum sharing. The proposed algorithm can further reduce the energy consumption of the SBSs without degrading the performance of the MBSs. However, spectrum resource sharing has not been integrated with caching to gain higher performance of hit rate and system delivery capacities.

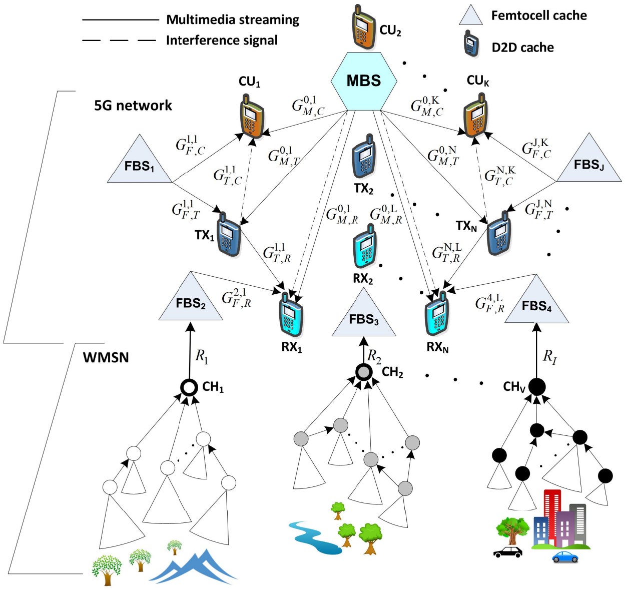


FIGURE 1. 5G caching and downlink resource sharing model for WMSNs in smart cities.

III. SYSTEM MODELS

A. 5G CSF FOR WMSNS

In this paper, we consider an integrated system consisting of a three-tier 5G network and a WMSN as shown in Fig. 1. The three-tier 5G network includes one MBS, J FBSs and $(K + 2N)$ MUs. The MUs are divided into K cellular users (CUs) that share their downlink resources with N D2D pairs, each of a D2D transmitter (TX) and a D2D receiver (RX). A 5G caching model is established by the MBS, FBSs (with femtocaching) and the TXs (with D2D caching) to support the system in delivering multimedia contents captured from the WMSN to the MUs. In this system, the multimedia contents can be streamed 1) from the MBS and FBSs to the CUs and TXs and 2) from the MBS, FBSs, and TXs to the RXs. It is important to note that the femtocaching scheme is deployed to provide high hit rate to access the multimedia contents and high system delivery capacity, while the D2D caching scheme

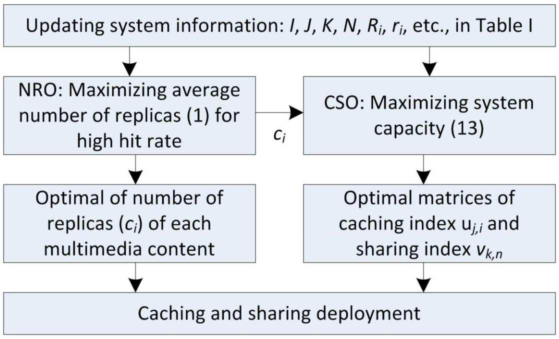

Updating system information: I, J, K, N, R_i, r_i, etc., in Table I
NRO: Maximizing average number of replicas (1) for high hit rate
c_i
CSO: Maximizing system capacity (13)
Optimal of number of replicas (c_i) of each multimedia content
Optimal matrices of caching index $u_{j,i}$ and sharing index $v_{k,n}$
Caching and sharing deployment

FIGURE 2. **Proposed CSF.**

is to further improve the system delivery capacity by utilizing downlink resources shared by the CUs. The WMSN has I sensor clusters, each covers a particular area and sends its captured multimedia content to the 5G network. We assume that the multimedia contents are sent from the WMSN to the 5G network for caching by joint active duty scheduling and encoding rate allocation in [41]. The detailed CSF for high hit rate and high system delivery capacity is shown in Fig. 2 and operates in four steps as follows:

- Step 1: The MBS collects all the system parameters in the whole cell including $I, J, K, N, R_i, r_i, W, G^{\mathsf{x,y}}_{\mathsf{X,Y}}$, etc., shown in Table 1. These parameters are then used (or updated if there are any significant changes) for formulating and solving the NRO and CSO problems.
- Step 2: Formulate and solve the NRO problem for optimal number of replicas of each multimedia content c_i, based on the parameters in step 1. The objective is to maximize the average number of replicas for high hit rate.
- Step 3: Formulate and solve the CSO problem for optimal caching index $u_{j,i}$ and optimal sharing index $v_{k,n}$, based on the parameters set in step 1 and optimal results of c_i found in step 2. The objective is to maximize the overall system delivery capacity.
- Step 4: The MBS deploys the CSF to cache the multimedia contents in the proper FBSs and to share the downlink resources between the right CUs and TX-RX pairs for offloading with D2D communications.

B. HIT RATE

In the hit rate model, let c_i be the number of replicas of the i-th multimedia content cached in all the FBSs, the average number of replicas per each multimedia content, which can be used to represent the hit rate performance, is computed by

$$\bar{I} = \sum_{i=1}^{I} r_i c_i, \tag{1}$$

where $r_i = \frac{i^{-\alpha}}{\sum_{i=1}^{I} i^{-\alpha}}$, namely Zipf-like distribution [43], is the access rate (or popularity) of the i-th multimedia content reflected by the skewed access rate $\alpha \geq 0$ among the contents. $\alpha = 0$ indicates that all multimedia contents have the same access rate of $1/I$, while the higher the value of α increases, the higher the skewed access rate is.

Eq. (1), which is the objective function in NRO problem, is maximized by finding optimal values of c_i for high hit rate. The NRO problem will be discussed in the next section.

TABLE 1. **Notations.**

Symbols	Descriptions
I	Number of multimedia contents or number of sensor clusters in the WMSN
J	Number of FBSs
K	Number of cellular users (CUs) who share their downlink resources with D2D pairs
N	Number of D2D pairs, each has a D2D transmitter (TX) and a D2D receiver (RX)
r_i	Access rate (i.e., the popularity or the importance) of the i-th multimedia content
R_i	Throughput required for the i-th multimedia content streaming, $i = 1, 2, ..., I$
R^i_M, R^i_F, R^i_T	Reserved throughput at the MBS, FBSs, and TXs, to serve the MUs who request for the i-th multimedia content streaming
W	System bandwidth
$G^{\mathsf{x,y}}_{\mathsf{X,Y}}$	Channel gains between X and Y; here $\mathsf{X} \in \{\mathsf{M, F, T}\}$ standing for {MBS, FBS, TX} and $\mathsf{Y} \in \{\mathsf{C, T, R}\}$ standing for {CU, TX, RX}; $\mathsf{x} \in \{j, n\}$, $j = 0, 1, 2, ..., J$, $j = 0$ indicates the MBS, $n = 1, 2, ..., N$ and $\mathsf{y} \in \{k, n\}$, $k = 1, 2, ..., K$
$p_{n,i}$	Probability that the n-th TX decides to cache the i-th multimedia content
β_n	Percentage of available storage of the n-th TX
$v_{k,n}$	Sharing index of the k-th CU, $v_{k,n} = 1$ means that the k-th CU shares its downlink resource with the n-th D2D pair, otherwise $v_{k,n} = 0$
$u_{j,i}$	Caching index of the j-th FBS, $u_{j,i} = 1$ means that the m-th FBS makes a decision on caching the i-th multimedia content, otherwise $u_{j,i} = 0$

C. CHANNEL AND SYSTEM DELIVERY CAPACITY

1) CHANNEL

Because the MBS is overlaid with the FBSs, the channel splitting and F-ALOHA [44], [45] are applied to control the cross-tier and co-tier interference. Each D2D pair (i.e., TX and RX) can share the downlink transmission resource of any CUs. This resource sharing causes the interference effects of MBS on the RXs and of the TXs on the CUs. We denote $G^{\mathsf{x,y}}_{\mathsf{X,Y}}$ as the channel gains between X and Y; here $\mathsf{X} \in \{\mathsf{M, F, T}\}$ standing for {MBS, FBS, TX} and $\mathsf{Y} \in \{\mathsf{C, T, R}\}$ standing for {CU, TX, RX}; $\mathsf{x} \in \{j, n\}, j = 0, 1, 2, \ldots, J, j = 0$ indicates the MBS, $n = 1, 2, \ldots, N$ and $\mathsf{y} \in \{k, n\}, k = 1, 2, \ldots, K$. $G^{\mathsf{x,y}}_{\mathsf{X,Y}}$ is modeled by the exponential power fading coefficient $h^{\mathsf{x,y}}_{\mathsf{X,Y}}$ and the standard power law path loss function $g^{\mathsf{x,y}}_{\mathsf{X,Y}} = ||d||^{-\xi}$, i.e., $G^{\mathsf{x,y}}_{\mathsf{X,Y}} = h^{\mathsf{x,y}}_{\mathsf{X,Y}} g^{\mathsf{x,y}}_{\mathsf{X,Y}}$ [45]. Here, ξ is the path loss exponent, d is the distance between X and Y, and $||.||$ is the Euclidean norm.

2) SYSTEM DELIVERY CAPACITY

To derive the system delivery capacity, we primarily analyze the signal to interference plus noise ratio (SINR) of CUs, TXs, and RXs in the sequel.

For CUs, the k-th CU can simultaneously share its downlink resource with the n-th D2D pair and receive multimedia contents from the MBS and FBSs. The SINRs of the channels from the MBS and FBSs to the k-th CU are given by

$$\gamma_{\mathrm{M,C}}^{0,k,i} = \frac{P_{\mathrm{M}}^{0} G_{\mathrm{M,C}}^{0,k}}{N_0 + \sum_{n=1}^{N} v_{k,n} p_{n,i} P_{\mathrm{T}}^{n} G_{\mathrm{T,C}}^{n,k}}, \tag{2}$$

$$\gamma_{\mathrm{F,C}}^{j,k,i} = \frac{u_{j,i} P_{\mathrm{F}}^{j} G_{\mathrm{F,C}}^{j,k}}{N_0}, \tag{3}$$

where P_{M}^{0} is the transmission power of the MBS, $G_{\mathrm{M,C}}^{0,k}$ is the channel gain between the MBS and the k-th CU, N_0 is the power of additive white Gaussian noise (AWGN), and $v_{k,n}$ is the sharing index used to indicate that the k-th CU agrees to share its downlink resource with the n-th D2D pair ($v_{k,n} = 1$) or not ($v_{k,n} = 0$). If $v_{k,n} = 1$, the k-th CU is affected by the interference from the n-th TX of the n-th D2D pair with transmission power P_{T}^{n} over the channel gain $G_{\mathrm{T,C}}^{n,k}$ between the n-th TX and the k-th CU. In addition, $u_{j,i}$ is the caching index used to indicate that the j-th FBS makes a decision to cache the i-th multimedia content, P_{F}^{j} is the transmission power of the j-th FBS, and $G_{\mathrm{F,C}}^{j,k}$ is the channel gain between the j-th FBS and the k-th CU.

In (2), $p_{n,i}$ is the probability that the n-th TX decides to cache the i-th multimedia content defined by

$$p_{n,i} = a r_i + b \beta_n, \tag{4}$$

here $a, b \in [0, 1]$, $a + b = 1$, and β_n is the percentage of available storage of the n-th TX. It means that the n-the TX decides to cache or not depending on its storage condition and the access rate r_i of the i-th multimedia content.

Similarly, for TXs, the SINRs of the channels from the MBS and FBSs to the n-th TX are given by

$$\gamma_{\mathrm{M,T}}^{0,n} = \frac{P_{\mathrm{M}}^{0} G_{\mathrm{M,T}}^{0,n}}{N_0}, \tag{5}$$

$$\gamma_{\mathrm{F,T}}^{j,n,i} = \frac{u_{j,i} P_{\mathrm{F}}^{j} G_{\mathrm{F,T}}^{j,n}}{N_0}. \tag{6}$$

Considering the RXs of D2D pairs, they are served not only by the MBS, but also by the FBSs and TXs, the SINRs of the channels from the MBS, FBSs, and TXs to the n-th RX are therefore described in sequence as follows:

$$\gamma_{\mathrm{M,R}}^{0,n} = \frac{P_{\mathrm{M}}^{0} G_{\mathrm{M,R}}^{0,n}}{N_0}, \tag{7}$$

$$\gamma_{\mathrm{F,R}}^{j,n,i} = \frac{u_{j,i} P_{\mathrm{F}}^{j} G_{\mathrm{F,R}}^{j,n}}{N_0}, \tag{8}$$

and

$$\gamma_{\mathrm{T,R}}^{n,k,i} = \frac{v_{k,n} p_{n,i} P_{\mathrm{T}}^{n} G_{\mathrm{T,R}}^{n,n}}{N_0 + P_{\mathrm{M}}^{0} G_{\mathrm{M,R}}^{0,n} + \sum_{l=1, l \neq n}^{N} v_{k,l} p_{l,i} P_{\mathrm{T}}^{l} G_{\mathrm{T,R}}^{l,l}}, \tag{9}$$

It is noticed in (9) that the n-th RX is affected by the interference from the MBS with transmission power P_{M}^{0} over the channel gain $G_{\mathrm{M,R}}^{0,n}$ between the MBS and the n-th RX.

From (2)-(9), by using Shannon-like capacity, the capacity of CUs, TXs, and TXs are respectively given by

$$R_{\mathrm{C}} = W \sum_{k=1}^{K} \sum_{i=1}^{I} r_i \left[\log_2 \left(1 + \gamma_{\mathrm{B,C}}^{0,k,i} \right) + \sum_{j=1}^{J} \log_2 (1 + \gamma_{\mathrm{F,C}}^{j,k,i}) \right], \tag{10}$$

$$R_{\mathrm{T}} = W \sum_{n=1}^{N} \left[\log_2 \left(1 + \gamma_{\mathrm{M,T}}^{0,n} \right) + \sum_{j=1}^{J} \sum_{i=1}^{I} r_i \log_2 (1 + \gamma_{\mathrm{F,T}}^{j,n,i}) \right], \tag{11}$$

$$\begin{aligned} R_{\mathrm{R}} = W \sum_{n=1}^{N} \Bigg[& \log_2 (1 + \gamma_{\mathrm{M,R}}^{0,n}) + \sum_{j=1}^{J} \sum_{i=1}^{I} r_i \log_2 (1 + \gamma_{\mathrm{F,R}}^{j,n,i}) \\ & + \sum_{k=1}^{K} \sum_{i=1}^{I} r_i \log_2 (1 + \gamma_{\mathrm{T,R}}^{n,k,i}) \Bigg]. \end{aligned} \tag{12}$$

Finally, the overall average system delivery capacity to each requester (i.e., CU, TX, or RX) is shown as

$$R = \frac{R_{\mathrm{C}} + R_{\mathrm{T}} + R_{\mathrm{R}}}{K + 2N}. \tag{13}$$

We observe that the overall average system capacity (13) can be maximized by finding the optimal caching index $u_{j,i}$ and optimal sharing index $v_{k,n}$. This CSO problem will be formulated in the next section.

IV. NRO AND CSO PROBLEMS AND SOLUTIONS

A. NRO

As we mentioned in (1), a high hit rate can be obtained by finding the optimal number of replicas of each multimedia content c_i to maximize the average number of replicas $\bar{I}$. To do so, we further take into account the constraints of storage of all FBSs; reserved throughput of MBS, FBSs, and TXs; and required throughput of all MUs. The NRO problem is formulated as below.

$$\max_{\mathbf{c_i}} \bar{I}, \tag{14}$$

$$s.t. \begin{cases} 1 \leq c_i \leq J, & i = 1, 2, \ldots, I \\ \sum_{i=1}^{I} c_i \leq \rho I J, & \frac{1}{J} \leq \rho \leq 1 \\ R_{Res}^{i} \leq R_{Req}^{i}, & Ri = 1, 2, \ldots, I \end{cases} \tag{15}$$

where the first and the second constraints are used to ensure that at least one replica of each multimedia content is cached in the FBSs and to limit the number of replicas of each (all) multimedia content(s). The third constraint is to avoid the wasted throughput, in case the reserved throughput (R_{Res}^{i}) is greater than the required throughput (R_{Req}^{i}), i.e., too many FBSs together with the MBS and TXs provide the MUs with replicas of the i-th multimedia content. The R_{Res}^{i} and R_{Req}^{i} for the i-th multimedia content are respectively given by

$$R_{Res}^{i} = R_{M}^{i} + R_{F}^{i} c_i + R_{T}^{i} \sum_{n=1}^{N} p_{n,i}, \tag{16}$$

$$R_{Req}^{i} = (K + 2N) R_i r_i, \tag{17}$$

where R_M^i, R_F^i, and R_T^i are the reserved throughput at the MBS, FBSs, and TXs for the i-th multimedia content to serve the MUs.

Substituting (16) and (17) for the third constraint of (15), we have

$$c_i \le \frac{R_{Req}^i - (R_T^i \sum_{n=1}^{N} p_{n,i} + R_M^i)}{R_F^i}. \tag{18}$$

By combining (18) and the first constraint of (15), the NRO problem can be rewritten as

$$\max_{\mathbf{c_i}} \bar{I} \tag{19}$$

$$s.t. \begin{cases} 1 \le c_i \le C_i, & i = 1, 2, \ldots, I \\ \sum_{i=1}^{I} c_i \le \rho I J, & \frac{1}{J} \le \rho \le 1 \end{cases} \tag{20}$$

where

$$C_i = min\left\{J, max\{1, \frac{R_{Req}^i - (R_T^i \sum_{n=1}^{N} p_{n,i} + R_M^i)}{R_F^i}\}\right\}. \tag{21}$$

The linear programming optimization problem in (19) and (20) can be solved by using primal-dual interior point method (a variant of Mehrotra's predictor-corrector algorithm) [46], [47]. It is noticed that the operator *max* in (21) is to guarantee that the upper bound C_i of the first constraint in (20) cannot be less than 1.

B. CSO

Taking into account the constraints on the optimal number of replicas (i.e., c_i found by solving (14) and (15)) and the target SINR γ_0 of CUs, the CSO problem is formulated and solved to maximize the overall system delivery capacity R in (13) by finding $u_{j,i}$ and $v_{k,n}$. Mathematically, the CSO problem is expressed as follows:

$$\max_{\mathbf{u_{j,i}}, \mathbf{v_{k,n}}} \ R \tag{22}$$

$$s.t. \begin{cases} \sum_{j=1}^{J} u_{j,i} \le c_i, \quad i = 1, 2, \ldots, I \\ \sum_{n=1}^{N} v_{k,n} p_{n,i} P_{\mathrm{T}}^{n} G_{\mathrm{T,C}}^{n,k} \le \frac{P_M^0 G_{M,C}^{0,k}}{\gamma_0} - N_0, \\ \quad k = 1, 2, \ldots, K, i = 1, 2, \ldots, I \end{cases} \tag{23}$$

In (23), the first constraint is to make sure that the number of FBSs decides to cache the i-th multimedia content cannot exceed the optimal number of replicas c_i of the i-th multimedia content found by solving the NRO problem. The second constraint comes from (2) by letting $\gamma_{\mathrm{M,C}}^{0,k,i} \ge \gamma_0$. It is used to limit the effect of interference of D2D pairs on the CUs. In this constraint, to ensure high target SINR of the CUs by increasing γ_0, the number of D2D pairs associated to each CU decreases. Finding the optimal caching index $u_{j,i}$ and optimal sharing index $v_{k,n}$ in (22) and (23) is equivalent to finding two optimal matrices $\mathbf{u}^*_{\mathbf{J\times I}}$ and $\mathbf{v}^*_{\mathbf{K\times N}}$ in two matrix search spaces: $\mathcal{U} = \{u^1_{J\times I}, u^2_{J\times I}, \ldots, u^{2^{J\times I}}_{J\times I}\}$ and $\mathcal{V} = \{v^1_{K\times N}, v^2_{K\times N}, \ldots, v^{2^{K\times N}}_{K\times N}\}$, respectively. Exhaustive binary matrix search, which can be used to solve (22) and (23), is given in **Algorithm** 1.

Algorithm 1 Exhaustive Binary Search for CSO Problem

```
Require: Initial parameters given in Table 2
Ensure: $R^*$, $\mathbf{u}^*_{\mathbf{J\times I}}$, $\mathbf{v}^*_{\mathbf{K\times N}}$
Generating two matrix search spaces
    $\mathcal{U} = \{u^1_{J\times I}, u^2_{J\times I}, \ldots, u^{2^{J\times I}}_{J\times I}\}$ and
    $\mathcal{V} = \{v^1_{K\times N}, v^2_{K\times N}, \ldots, v^{2^{K\times N}}_{K\times N}\}$
$\mathcal{R} \leftarrow \varnothing$
for each matrix $u_{J\times I}$ in $\mathcal{U}$ do
    for each matrix $v_{K\times N}$ in $\mathcal{V}$ do
        if (23) satisfies then
            $R(u_{J\times I}, v_{K\times N}) = R$, computing (13)
            $\mathcal{R} \leftarrow \mathcal{R} \cup R(u_{J\times I}, v_{K\times N})$
        end if
    end for
end for
$R^*$ = max $\mathcal{R}$
$\{\mathbf{u}^*_{\mathbf{J\times I}}, \mathbf{v}^*_{\mathbf{K\times N}}\}$ = argmax $\mathcal{R}$
```

In **Algorithm** 1, the memory and time complexities, which depend on the total search space of $\mathcal{U}$ and $\mathcal{V}$, are equivalent to $\mathcal{O}(2^{J\times I + K\times N})$. In dense 5G networks, the CSO problem is difficult to be solved by centralized search at the MBS. In this scenario, the search space is divided into multiple sub-search spaces, and then distributed exhaustive binary search of each sub-search space can be independently done by each FBS. Afterwards, the sub-optimal results from the FBSs are sent to the MBS for finding global optimal solution.

V. PERFORMANCE EVALUATION

A. SIMULATION SETUP

For simplicity but without loss of generality, we deploy the system with important parameters listed in Table 2. The distances between the MBS and MUs, the FBSs and MUs, the CUs and TXs, and the TXs and RXs, are randomly distributed from 100m to 500m, 50m to 250m, 50m to 100m, and 1m to 50m, respectively.

B. PERFORMANCE METRICS

1) HIT RATE PERFORMANCE

To evaluate the hit rate performance of our NRO, we compare NRO to the other two schemes named equal number of replicas (ENR) and worst number of replicas (WNR). In ENR, the number of replicas of each multimedia content $c^i_{ENR} = \frac{\sum_{i=1}^{I} c_i}{I}$, while in WNR, c^i_{WNR} is inversely proportional to r_i such that $\sum_{i=1}^{I} c^i_{WNR} = \sum_{i=1}^{I} c_i$.

We first evaluate the performance of NRO, ENR, and WNR versus the total caching capacity of the FBSs by changing ρ in the range from $\frac{1}{J} = 0.2$ to 1 (20) and keeping $\alpha = 1$. In Fig. 3, the NRO yields the highest average number of replicas for the highest hit rate compared to both ENR and WNR. The higher caching capacity of FBSs introduces a higher hit rate but getting saturated if each FBS caches all the considered multimedia contents. In comparison versus α

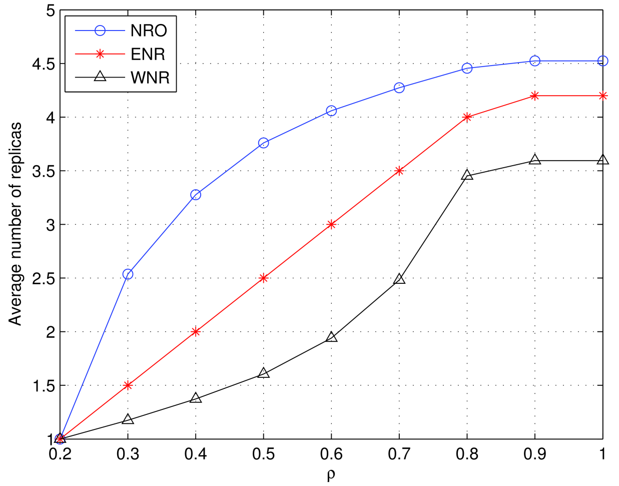


FIGURE 3. Hit rate performance versus FBSs' caching capacity coefficient ρ.

TABLE 2. Parameters setting.

Symbols	Specifications
I	10 Multimedia contents
J	5 FBSs
K	5 CUs
N	4 D2D pairs
$\{R_i\}$	$\{200, 190, 180, 170, 160, 150, 140, 130, 120, 110\}$ Mbps
$\{R_M^i\}$	$\{90, 80, 70, 60, 50, 40, 30, 20, 10, 5\}$ Mbps
$\{R_F^i\}$	$\{50, 45, 40, 35, 30, 25, 20, 15, 10, 5\}$ Mbps
$\{R_T^i\}$	$\{10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$ Mbps
W	5MHz
P_M^0	10W
P_F^j	Fixed to 5W
$\{P_T^n\}$	$\{0.01, 0.05, 0.075, 0.1\}$W
γ_0	10dB
N_0	10^{-13}W
ξ	4 (path loss exponent)
$\{\beta_n\}$	$\{0.2, 0.4, 0.6, 0.8\}$ (path loss exponent)
a, b	0.5

in the range from 0 to 2 and $\rho = 0.5$, the results in Fig. 4 show that while the ENR does not change and the WNR decreases with respect to the increase of α, the average number of replicas of NRO always increases to gain the best hit rate performance. In addition, by keeping $\alpha = 1$ and $\rho = 0.5$, the proposed NRO also outperforms the ENR and WNR versus the number of FBSs (J) and number of multimedia contents (I) as illustrated in Fig. 5 and Fig. 6. The benefit of NRO can be achieved based on the fact that the number of replicas c_i is found directly proportional to the access rate r_i of the i-th multimedia content, meanwhile the ENR and WNR cannot do. Obviously, in Fig. 4 and Fig. 6, the hit rate of WNR degrades versus the increase of r_i because c_{WNR}^i is inversely proportional to r_i.

2) SYSTEM DELIVERY CAPACITY PERFORMANCE

The performance of the system delivery capacity is investigated by comparing our CSO to average delivery capacity (ADC), worst delivery capacity (WDC) schemes, maximum delivery capacity without MUs' caching and downlink resource sharing (None-MUCS) [18], and maximum delivery capacity without femtocaching (None-FBSC) [48]. In ADC and WDC, the system delivery capacity is

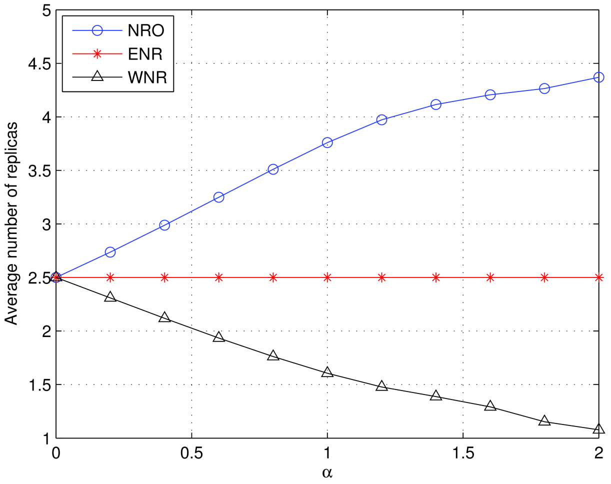


FIGURE 4. Hit rate performance versus skewed access rate coefficient α.

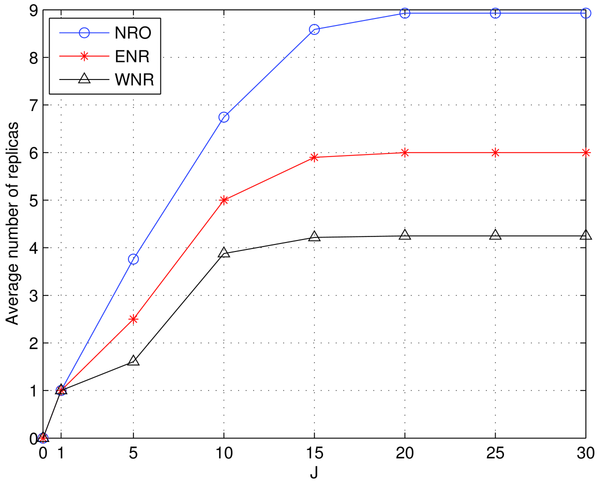


FIGURE 5. Hit rate performance versus number of FBSs J.

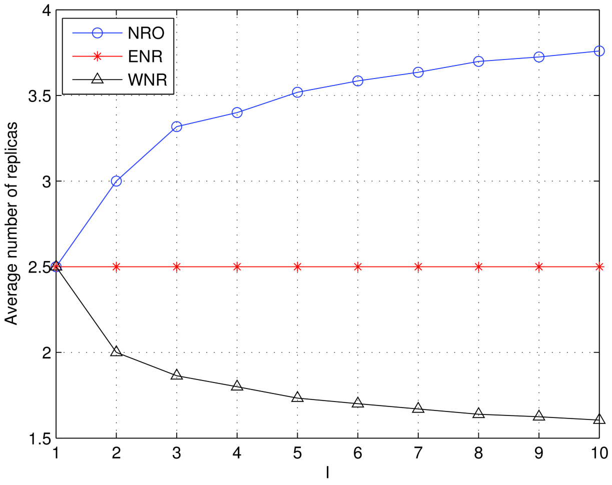


FIGURE 6. Hit rate performance versus number of multimedia contents I.

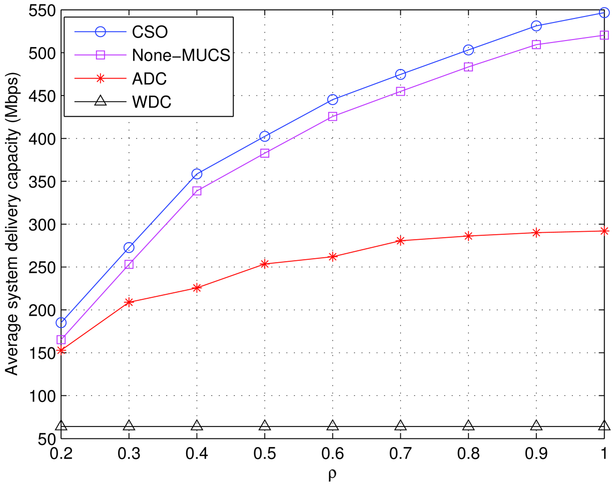


FIGURE 7. Capacity performance versus FBSs' caching capacity coefficient ρ.

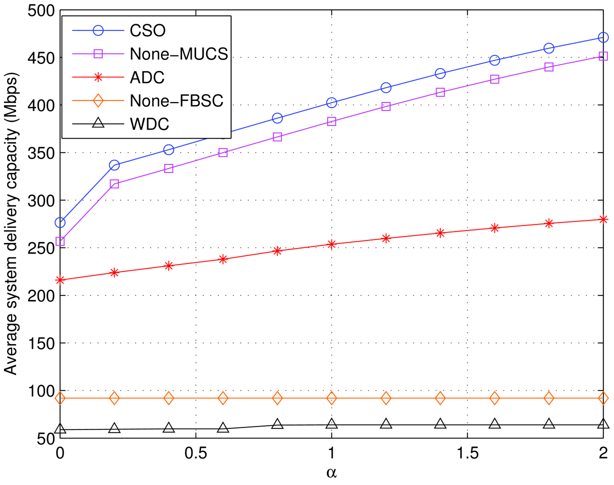


FIGURE 8. Capacity performance versus skewed access rate coefficient α.

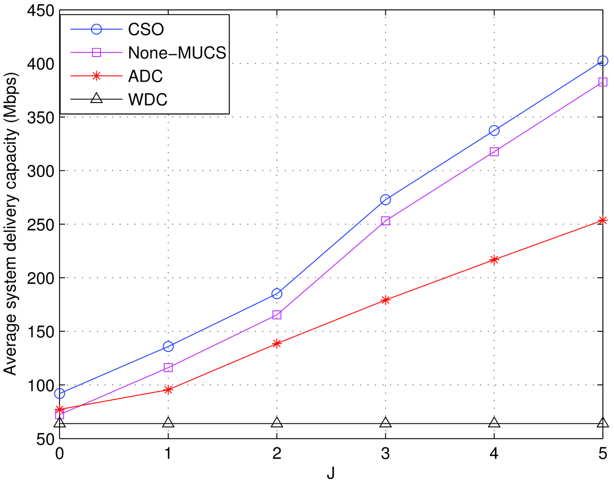


FIGURE 9. Capacity performance versus number of FBSs J.

respectively averaged and minimized over the feasible combinations of the total search space. Meanwhile, in None-MUCS and None-FBSC, the system delivery capacity is maximized over two scenarios: 1) without D2D caching and downlink resource sharing for offloading with D2D communications and 2) without caching multimedia contents in the FBSs. To mitigate the memory and time complexities of searching, we reduce the number of multimedia contents from 10 to 4.

We first evaluate the performance of the system delivery capacity versus the total caching capacity of the FBSs by changing ρ and letting $\alpha = 1$. We do not compare our CSO to None-FBSC because changing ρ is meaningless due to $J = 0$. As shown in Fig. 7, the system delivery capacity of CSO, ADC, WDC, and None-MUCS obviously gets higher and gradually saturated when ρ increases for more number of replicas cached in the FBSs. In addition, we can see that because of without MUs' caching and downlink resource sharing, the None-MUCS results in lower system delivery capacity than the CSO, but it outpaces the ADC and WDC.

Fig. 8 plots the performance of system delivery capacity versus α. We can find that the system delivery capacity significantly decreases in case of None-FBSC, i.e., the None-FBSC is only better than the WDC. It means that femtocaching plays an important role in improving the performance of system delivery capacity compared to D2D caching and downlink resource sharing. It is observed in Fig. 8 that the average system delivery capacity of None-FBSC keeps unchanged because α mostly impacts on femtocaching rather than D2D caching and downlink resource sharing. To insightfully understand the impact of femtocaching on the performance of system delivery capacity, Fig. 9 further depicts the CSO, ADC, WDC, and None-MUCS versus the number of FBSs. Obviously, the results demonstrate that the larger scale of FBSs provides higher system delivery capacity performance, except for the WDC. In all cases, our proposed CSO always gains the highest performance compared to the others.

The impact of the number of CUs, that are willing to share their downlink resources is also investigated as shown in Fig. 10. Fig. 10 reveals that the performance of the system delivery capacity is higher with respect to the increase of the number of CUs K, when more CUs share the downlink resources for offloading with D2D communications. Especially, when $K = 0$, the CSO and None-MUCS have the same result meaning that D2D caching becomes unuseful if there is no CU to share the downlink resource. Similarly, the results of WDC and None-FBSC are the same showing that without both femtocaching and downlink resource sharing make the performance worse.

Finally, we evaluate the performance of the system delivery capacity under the effect of the target SINR of the CUs by changing γ_0 from 0dB to 30dB. It can be seen in Fig. 11 that in case of CSO and None-FBSC, more D2D communications in close proximity are not exploited because they do not satisfy the target SINR of the CUs as formulated in the second constraint of (23). This in turn decreases the system performance if γ_0 increases. The increase of γ also removes

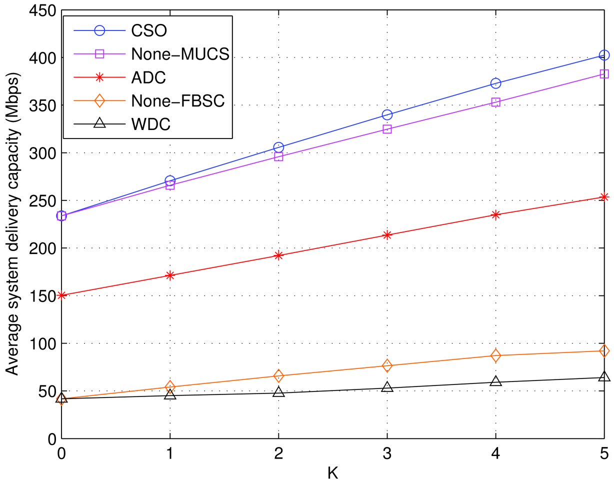


FIGURE 10. Capacity performance versus number of CUs K.

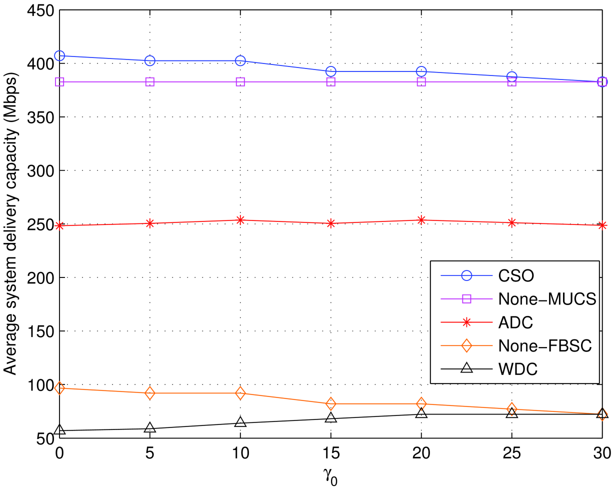


FIGURE 11. Capacity performance versus the target SINR of CUs γ.

more infeasible matrices from the search spaces $\mathcal{U}$ and $\mathcal{V}$ that deteriorate the system performance. For example, when the TX-RX distance for D2D communications is longer than the TX-CU distance, it causes the CUs high interference, but not significantly improving D2D communications capacity. Therefore, the system performance of the WDC is higher when γ increases. The system performance decrease of CSO and increase of WDC keeps the system performance of ADC unchanged versus γ. In addition, the system delivery capacity of None-MUCS does not change because D2D caching and CUs' downlink resource sharing are not considered. It is clear to notice that the system performances of the CSO and the None-FBSC respectively converge on the system performances of None-MUCS and WDC when γ increases.

VI. CONCLUSION

We have designed a joint caching and downlink resource sharing optimization framework (CSF) that is built on exploiting the caching storage of all MBSs, FBSs, and TXs, as well as the downlink resource of the CUs in 5G networks to assist WMSNs in smart cities. The CSF consists of two optimization problems, namely number of replicas optimization problem and where to cache and who to share optimization problem, which are solved to gain a high hit rate and a high system delivery capacity. The simulation results show that our proposed design can serve the MUs the best system performance compared to other schemes.

ACKNOWLEDGMENT

This paper was presented in part at the IEEE Global Communications Conference (GLOBECOM'15), San Diego, CA, USA, December 2015 [1].

REFERENCES

[1] N.-S. Vo, T. Q. Duong, and M. Guizani, "Quality of sustainability optimization design for mobile ad hoc networks in disaster areas," in *Proc. IEEE Global Commun. Conf.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[2] W. Tushar, C. Yuen, K. Li, K. L. Wood, Z. Wei, and L. Xiang, "Design of cloud-connected IoT system for smart buildings on energy management," *EAI Endorsed Trans. Ind. Netw. Intell. Syst.*, vol. 3, no. 6, pp. 1–9, Jan. 2016.

[3] B. Hammi, R. Khatoun, S. Zeadally, A. Fayad, and L. Khoukhi, "IoT technologies for smart cities," *IET Netw.*, vol. 7, no. 1, pp. 1–13, Jan. 2018.

[4] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Cisco, San Jose, CA, USA, White Paper, Mar. 2017. [Online]. Available: http://www.bookref.com

[5] E. Bastug, M. Bennis, and M. Debbah, "Proactive caching in 5G small cell networks," in *Towards 5G: Applications* (Requirements and Candidate Technologies), 1st ed. R. Vannithamby and S. Talwar, Eds. Hoboken, NJ, USA: Wiley, 2016, ch. 6.

[6] K. Zhu, W. Zhi, X. Chen, and L. Zhang, "Socially motivated data caching in ultra-dense small cell networks," *IEEE Netw.*, vol. 31, no. 4, pp. 42–48, Jul./Aug. 2017.

[7] C. Xu, J. Feng, Z. Zhou, J. Wu, and C. Perera, "Cross-layer optimization for cooperative content distribution in multihop device-to-device networks," *IEEE Internet Things J.*, to be published.

[8] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[9] V.-D. Nguyen, T. Q. Duong, H. D. Tuan, O.-S. Shin, and H. V. Poor, "Spectral and energy efficiencies in full-duplex wireless information and power transfer," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2220–2233, May 2017.

[10] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5G cloud radio access networks," *IEEE Netw.*, vol. 31, no. 4, pp. 35–41, Jul./Aug. 2017.

[11] 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.

[12] 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.

[13] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[14] G. S. Park, W. Kim, S. H. Jeong, and H. Song, "Smart base station-assisted partial-flow device-to-device offloading system for video streaming services," *IEEE Trans. Mobile Comput.*, vol. 16, no. 9, pp. 2639–2655, Sep. 2017.

[15] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[16] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[17] Y. N. Shnaiwer, S. Sorour, N. Aboutorab, P. Sadeghi, and T. Y. Al-Naffouri, "Network-coded content delivery in femtocaching-assisted cellular networks," in *Proc. IEEE Global Commun. Conf.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

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

[19] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, ''Pricing and resource allocation via game theory for a small-cell video caching system,'' *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[20] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, ''Probabilistic small-cell caching: Performance analysis and optimization,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, May 2017.

[21] J. Liao, K.-K. Wong, M. R. A. Khandaker, and Z. Zheng, ''Optimizing cache placement for heterogeneous small cell networks,'' *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 120–123, Jan. 2017.

[22] J. Liao, K.-K. Wong, Y. Zhang, Z. Zheng, and K. Yang, ''Coding, multicast, and cooperation for cache-enabled heterogeneous small cell networks,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

[23] E. Ozfatura and D. Gündüz, ''Mobility and popularity-aware coded small-cell caching,'' *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 288–291, Feb. 2018.

[24] W. Han, A. Liu, and V. K. N. Lau, ''PHY-caching in 5G wireless networks: Design and analysis,'' *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[25] J. Qiao, Y. He, and X. S. Shen, ''Proactive caching for mobile video streaming in millimeter wave 5G networks,'' *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[26] G. F. W. Jiang and S. Qin, ''Optimal cooperative content caching and delivery policy for heterogeneous cellular networks,'' *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1382–1393, May 2017.

[27] P. Lin, Q. Song, Y. Yu, and A. Jamalipour, ''Extensive cooperative caching in D2D integrated cellular networks,'' *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2101–2104, Sep. 2017.

[28] M. Chen, Y. Hao, L. Hu, K. Huang, and V. Lau, ''Green and mobility-aware caching in 5G networks,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8347–8361, Dec. 2017.

[29] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, ''Wireless content caching for small cell and D2D networks,'' *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

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

[31] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, ''Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6939, Oct. 2017.

[32] T. X. Vu, S. Chatzinotas, B. Ottersten, and T. Q. Duong, ''Energy minimization for cache-assisted content delivery networks with wireless backhaul,'' *IEEE Wireless Commun. Lett.*, to be published.

[33] A. M. Akhtar, X. Wang, and L. Hanzo, ''Synergistic spectrum sharing in 5G HetNets: A harmonized SDN-enabled approach,'' *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 40–47, Jan. 2016.

[34] Y. Wang, X. Tao, X. Zhang, and Y. Gu, ''Cooperative caching placement in cache-enabled D2D underlaid cellular network,'' *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1151–1154, May 2017.

[35] A. Alabbasi, Z. Rezki, and B. Shihada, ''Outage analysis of spectrum sharing over M-block fading with sensing information,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3071–3087, Apr. 2017.

[36] A. Tang, X. Wang, and C. Zhang, ''Cooperative full duplex device to device communication underlaying cellular networks,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7800–7815, Dec. 2017.

[37] C. Jiang, B. Wang, Y. Han, Z.-H. Wu, and K. J. R. Liu, ''Exploring spatial focusing effect for spectrum sharing and network association,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4216–4231, Jul. 2017.

[38] P. Mach and Z. Becvar, ''Energy-aware dynamic selection of overlay and underlay spectrum sharing for cognitive small cells,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4120–4132, May 2017.

[39] S. Misra, M. Reisslein, and G. Xue, ''A survey of multimedia streaming in wireless sensor networks,'' *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 18–39, 4th Quart., 2008.

[40] N.-S. Vo, D.-B. Ha, B. Canberk, and J. Zhang, ''Green two-tiered wireless multimedia sensor systems: An energy, bandwidth, and quality optimisation framework,'' *IET Commun.*, vol. 10, no. 18, pp. 2543–2550, Dec. 2016.

[41] 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,'' in *Mobile Netw. Appl.*, pp. 1–11, Sep. 2017. [Online]. Available: https://link.springer.com/journal/11036

[42] C. M. Park, R. A. Rehman, and B.-S. Kim, ''Packet flooding mitigation in CCN-based wireless multimedia sensor networks for smart cities,'' *IEEE Access*, vol. 5, pp. 11054–11062, Jun. 2017.

[43] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, ''Web caching and Zipf-like distributions: Evidence and implications,'' in *Proc. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, New York, NY, USA, Mar. 1999, pp. 126–134.

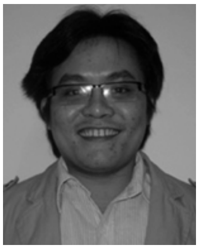
[44] V. Chandrasekhar and J. G. Andrews, ''Spectrum allocation in tiered cellular networks,'' *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[45] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, ''Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks,'' *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 561–574, Apr. 2012.

[46] S. Mehrotra, ''On the implementation of a primal-dual interior point method,'' *SIAM J. Optim.*, vol. 2, no. 4, pp. 575–601, 1992.

[47] Y. Zhang, ''Solving large-scale linear programs by interior-point methods under the MATLAB environment,'' Dept. Math. Statist., Univ. Maryland, College Park, MD, USA, Tech. Rep. TR96-01, Jul. 1995.

[48] X. Cai, J. Zheng, Y. Zhang, and H. Murata, ''A capacity oriented resource allocation algorithm for device-to-device communication in mobile cellular networks,'' in *Proc. IEEE Int. Conf. Commun.*, Sydney, NSW, Australia, Jun. 2014, pp. 2233–2238.



NGUYEN-SON VO received the B.Sc. degree in electrical and electronics engineering and the M.Sc. degree in radio engineering and electronics from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2002 and 2005, respectively, and the Ph.D. degree in communication and information systems from the Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently with Duy Tan University, Da Nang, Vietnam. His current research interests include multimedia streaming over wireless networks, physical layer security, energy harvesting, and optimization design. He was a recipient of the Best Paper Award at the IEEE Global Communications Conference 2016. He has served as a Guest Editor for *ACM/Springer Mobile Networks & Applications* (MONET), Special Issue on Wireless Communications and Networks for Smart Cities, in 2017, and also as a Guest Editor of *IET Communications*, Special Issue on Recent Advances on 5G Communications, in 2018.

TRUNG Q. DUONG (S'05–M'12–SM'13) received the Ph.D. degree in telecommunications systems from the Blekinge Institute of Technology, Sweden, in 2012. He is currently with Queen's University Belfast, U.K., where he was a Lecturer (Assistant Professor) from 2013 to 2017 and has been a Reader (Associate Professor) since 2018. He has authored or co-authored 300 technical papers published in scientific journals (170 articles) and presented at the international conferences (130 papers). His current research interests include Internet of Things, wireless communications, molecular communications, and signal processing.

Dr. Duong received the Best Paper Award at the IEEE Vehicular Technology Conference (Spring) in 2013, the IEEE International Conference on Communications 2014, the IEEE Global Communications Conference 2016, and the IEEE Digital Signal Processing Conference 2017. He was a recipient of the prestigious Royal Academy of Engineering Research Fellowship (2016–2021) and the prestigious Newton Prize 2017. He currently serves as an Editor for the IEEE Transactions on Wireless Communications, the IEEE Transactions on Communications, *IET Communications*, and a Lead Senior Editor for the IEEE Communications Letters.

MOHSEN GUIZANI (S'85–M'89–SM'99–F'09) received the B.S. degree (Hons.) and the M.S. degree in electrical engineering, and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He served as the Associate Vice President of Graduate Studies with Qatar University, the Chair of the Computer Science Department, Western Michigan University, and the Chair of the Computer Science Department, University of West Florida. He also served in academic positions at the University of Missouri–Kansas City, the University of Colorado at Boulder, Syracuse University, and Kuwait University. He is currently a Professor and the ECE Department Chair with the University of Idaho, Moscow, ID, USA. He has authored nine books and over 500 publications in refereed journals and conferences. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He served as a member, the chair, and the general chair of a number of international conferences. He received the teaching award multiple times and the best research award three times. He received the Wireless Technical Committee's Recognition Award in 2017. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker from 2003 to 2005. He was the Founder and the Editor-in-Chief of the *Wireless Communications and Mobile Computing* from 2000 to 2016. He guest edited a number of special issues in IEEE journals and magazines. He is currently the Editor-in-Chief of the IEEE NETWORK. He serves on the editorial boards of several international technical journals.

AYSE KORTUN received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Eastern Mediterranean University, Cyprus, in 2002 and 2004, respectively, and the Ph.D. degree from the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K., in 2012. She was a Lecturer with Cyprus International University, Cyprus. Since 2012, she has been a Research Fellow at the Digital Communication Research Cluster, ECIT, Queen's University Belfast. Her current research interests include spectrum sensing techniques in cognitive radio wireless networks.