Complexity Reduction in Multicast Beamforming for D2D Assisted Coded Caching

Hamidreza Bakhshzad Mahmoodi⋆, Jarkko Kaleva⋆, and Antti Tölli⋆
⋆ Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014, Finland firstname.lastname@oulu.fi

***Abstract*—A novel D2D aided beamforming scheme is proposed, where the local cache content exchange among nearby users is exploited. The transmission phase is split into two sub-phases: local D2D content exchange and downlink transmission. In the D2D sub-phase, users can autonomously share content with the adjacent users. In the downlink sub-phase, the Base station (BS) simultaneously serves some number of users to fulfill their remaining content requests by utilizing multicast beamforming. We first explain the main procedure via two simple examples and show the complexity involved in beamformer design. Then, we present the general formulation and detailed complexity analysis. We show analytically that, by exploiting the D2D exchange, the complexity of downlink multicast beamformer design can be greatly decreased. This provides significantly enhanced overall content delivery performance in terms of computational complexity and total delivery time.**

I. Introduction

High quality content delivery is one of the important requirements for the next generation networks. Caching popular content near end-users is a widely accepted technique to fulfill this requirement. This technique uses the off-peak hours of the network to move the content closer to the end-users, which mitigates the content delivery load in network peak hours. Many recent papers, such as, [1]–[3] have explored the potentials of this paradigm for improving the wireless networks performance. A promising scheme in this context is proposed in [4], which is also known as the *Coded caching (CC)* approach. In this scheme, instead of locally caching complete files at the end-users, file fragments throughout the whole library are carefully placed in the users' caches. Thus, in the delivery phase, instead of serving individual users, coded messages can be simultaneously multicast to groups of users. This results in significant performance increases by providing *global caching gain* [4].

CC has been shown to be greatly beneficial for both wired and wireless content delivery under various assumptions [4]–[9]. For example, [5] consider a two-level hierarchical setup in order to increase the multicasting opportunities for coded messages. The original CC setup is extended in [7] to investigate the effects of different multiserver network models on the content delivery time. Furthermore, [8]–[10] show that CC can boost the performance of the wireless network in terms of Degrees-of-Freedom (DoF) in the high signal-to-noise ratio (SNR) regime. Specifically, in wireless broadcast channels with a multiple-antenna broadcast channel (BC), the global coded caching and the spatial multiplexing gains are shown to be additive [7], [8], [10].

In order to bridge the gap between the high SNR analysis of CC and the practical finite SNR scenarios, recent works on finite SNR regime have also shown that, when the interference is properly accounted for, CC can be greatly beneficial [11]–[15]. The works [11] and [12] use a rate-splitting approach to benefit from the global caching gain and the spatial multiplexing gain at finite SNR. On the other hand, [13] follows a zero-forcing (ZF) based approach by extending the ideas in [7] to the finite SNR setup. This approach is also order-optimal in terms of DoF. Moreover, in [14], [15], the authors extend [13] to a general beamforming solution, which manages the interaction between interference and noise in more efficient manner. The general interference management framework proposed in [14], [15], improves the finite SNR performance of the CC in wireless networks significantly. However, the beamformer design complexity in [14] increases by the number of users and messages. This makes the approach impractical for more complex scenarios. The complexity issues, associated with the corresponding optimization problem, are somewhat addressed in [15]. However, the overall complexity with downlink CC beamformer design is still one of the limiting factors hindering the practical deployment.

This paper further investigates the beamformer design complexity proposed in [14], [15] by considering a D2D assisted delivery scheme. In this manner, the multicast beamforming [15] of file fragments is complemented by allowing direct device-to-device (D2D) exchange of local cache contents. By allowing direct D2D exchange of file fragments, the interference management between different downlink multicast streams becomes easier and more efficient as compared to the multicast only case [15], which results in a higher per user rate as demonstrated in [16]. At the same time, the complexity of the delivery scheme can be reduced both at the BS and at the end users. The results in this paper show that introducing the D2D phase to [14] significantly reduces the optimization complexity in addition to improving the delivery performance [16]. Moreover, we provide upper/lower bounds on the number of conditions which bound the computation complexity of the problem. These should be considered in the design of the beamformers when D2D transmission is available.

This work was supported by the Academy of Finland under grants no. 319059 (Coded Collaborative Caching for Wireless Energy Efficiency) and 318927 (6Genesis Flagship).

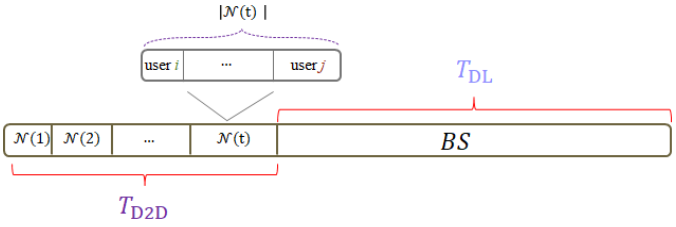

Fig. 1. Time division in D2D assisted transmission. The total time that is needed to transmit all fragments of files to the users is $T_{\mathrm{D2D}} + T_{\mathrm{DL}}$.

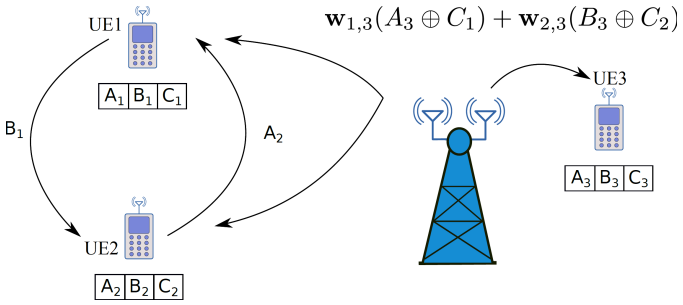

Fig. 2. Example 1: D2D enabled downlink beamforming system model.

II. System Model

We consider a system consisting of a single L antennas BS and K single antenna users. The BS has a library of N files, namely $\mathcal{W} = \{W_1, \ldots, W_N\}$, where each file has the size of F bits. The normalized cache size (memory) at each user is M files. Each user k caches a portion of the files, denoted by $Z_k(W_1, \ldots, W_N)$, which are stored in the *cache content placement* phase during off peak hours. At the *content delivery phase*, user $k \in \{1, \ldots K\}$ makes a request for the file $W_{d_k}, d_k \in [1 : N]$.

Upon the requests arrival, first, we have a D2D sub-phase, which is divided into a number of D2D time slots. In each time slot t, a group of nearby users, denoted by set $\mathcal{N}(t)$, is instructed by the BS to locally exchange data (see Fig. 1). Furthermore, each D2D time slot is divided into $|\mathcal{N}(t)|$ individual D2D transmissions. In each D2D transmission, a user $i \in \mathcal{N}(t)$ transmits a coded message denoted by X_i^{D2D} to an intended set of receivers $\mathcal{R}^{\mathcal{N}}(i) \subseteq \mathcal{N}(t)$, which are interested in decoding X_i^{D2D}. Thus, message X_i^{D2D} can be transmitted at rate[1]

$$R_i^{\mathcal{N}} = \min_{k \in \mathcal{R}^{\mathcal{N}}(i)} \log\left(1 + \frac{P_d \|h_{ik}\|^2}{N_0}\right), \tag{1}$$

where P_d is the device's transmit power constraint, and h_{ik} is the channel response from user i to user k. It should be noted that, in each D2D transmission, we assume that each user in $\mathcal{N}$ multicasts a message to a group of user. Thus, the rate is limited by the weakest receiver.

In the downlink phase, the BS multicasts coded messages containing all the remaining file fragments. Such that, all of the users will be able to decode their requested content. The received downlink signal at user $k = 1, \ldots, K$ is given by

$$y_k = \mathbf{h}_k^{\mathrm{H}} \sum_{\mathcal{T} \subseteq \mathcal{S}} \mathbf{w}_{\mathcal{T}}^{\mathcal{S}} \tilde{X}_{\mathcal{T}}^{\mathcal{S}} + z_k, \tag{2}$$

where $\tilde{X}_{\mathcal{T}}^{\mathcal{S}}$ is the modulated version of the intended message $X_{\mathcal{T}}^{\mathcal{S}}$ to be decoded by all the users in subset $\mathcal{T}$ of set $\mathcal{S} \subseteq [1 : K]$, and $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ is the corresponding beamforming vector. The channel vector between the BS and user k is $\mathbf{h}_k \in \mathcal{C}^L$. The receiver noise is given by $z_k \sim \mathcal{N}(0, N_0)$. The channel state information at the transmitter (CSIT) of all K users is assumed to be perfectly known. The final achievable rate (per user) over the above-described two phases is given by

$$R_U = \frac{F}{T_{\mathrm{D2D}} + T_{\mathrm{DL}}}, \tag{3}$$

where T_{D2D} and T_{DL} denote the time used for the D2D and downlink (DL) transmission sub-phases, respectively.

[1]In this paper, for simplicity, we assume that all D2D user groups $\mathcal{N}(t)$ are served in a TDMA fashion. Further improvement can be achieved by allowing parallel transmissions within multiple groups.

III. D2D Aided Beamforming Explained: Examples

In this section, we discuss the main concepts and principal trade-offs of the considered D2D transmission scheme via two examples. We provide examples for D2D enabled transmission and DL only transmission. For the sake of simplicity, we consider a network of $K = 3$ users.

A. D2D enabled transmission

In this example, illustrated in Fig. 2, we have $K = 3$ users and a library $\mathcal{W} = \{A, B, C\}$ of $N = 3$ files, where each user has the cache size of $M = 1$. The BS is equipped with $L = 2$ transmit antennas. To begin with, the cache content Z_k at each user $k = 1, \ldots, K$ is

$$Z_1 = \{A_1, B_1, C_1\}, Z_2 = \{A_2, B_2, C_2\}, Z_3 = \{A_3, B_3, C_3\}.$$

Here, we assume that each file is divided into three equal-sized sub-files. This follows the same cache placement as in [4]. We further assume that users 1 and 2 are in close proximity, while user 3 is far from them (see Fig. 2). To describe the idea, let us assume that users 1, 2, and 3 request files A, B, and C, respectively. Now, the actual transmission strategy is split into two phases. In the first phase, which is called as the D2D sub-phase, users 1 and 2 are assumed to be using D2D transmission to share their local cache content. Thus, the D2D sub-phase consists of a single D2D time slot with $\mathcal{N} = \{1, 2\}$. It is evident that user 2 would request B_1 from user 1 and user 1 would request A_2 from user 2. Since the D2D transmission is assumed to be half duplex and requires TDMA, this single time slot constitutes of two D2D transmissions. The time required for the D2D sub-phase is then given by

$$\begin{aligned} T_{\mathrm{D2D}} &= T\left(1 \to \mathcal{R}^{\mathcal{N}}(1)\right) + T\left(2 \to \mathcal{R}^{\mathcal{N}}(2)\right) \\ &= \frac{F/3}{R_1^{\mathcal{N}}} + \frac{F/3}{R_2^{\mathcal{N}}}, \end{aligned} \tag{4}$$

where $\mathcal{R}^{\mathcal{N}}(1) = \{2\}$, $\mathcal{R}^{\mathcal{N}}(2) = \{1\}, R_1^{\mathcal{N}} = \log\left(1+\frac{P_d\|h_{12}\|^2}{N_0}\right)$ and $R_2^{\mathcal{N}} = \log\left(1+\frac{P_d\|h_{21}\|^2}{N_0}\right)$. Note that, in each transmission, $\frac{F}{3}$ fraction of the corresponding file is transmitted.

In the following DL sub-phase, the BS simultaneously multicasts the remaining requested content to the users via coded messages. User 3 was not active in the D2D phase and still requires contents C_1 and C_2. However, users 1 and 2 only require A_3 and B_3, respectively. This content is XOR coded over two messages for user pairs $(1,3)$ and $(2,3)$. Namely, the messages are $X_{1,3} = A_3 \oplus C_1$ and $X_{2,3} = B_3 \oplus C_2$. Here, $X_{1,3}$ is a coded message, which would benefit users 1 and 3. Similarly, $X_{2,3}$ is a coded message intended for users 2 and 3. Thus, in order to deliver the correct coded message to each user, multicast beamformer vectors $\mathbf{w}_{1,3}$ and $\mathbf{w}_{2,3}$ are associated with messages $X_{1,3}$ and $X_{2,3}$, respectively. The downlink signal follows as $\mathbf{x}_{DL} = \tilde{X}_{1,3}\mathbf{w}_{1,3} + \tilde{X}_{2,3}\mathbf{w}_{2,3}$, where $\tilde{X}_{1,3}$ and $\tilde{X}_{2,3}$ are the modulated messages (for more details see [15]). Note that, here, user 3 is assumed to use SIC receiver to decode both intended messages (interpreted as a multiple access channel (MAC)), while, users 1 and 2 only get served with a single message with the other seen as interference.

Suppose now user 3 can decode *both* of its required messages $X_{1,3}$ and $X_{2,3}$ with the equal rate[2]

$$R_{MAC}^3 = \min\left(\frac{1}{2}R_{Sum}^3, R_1^3, R_2^3\right), \tag{5}$$

where the rate region corresponding to $\tilde{X}_{1,3}$, and $\tilde{X}_{2,3}$, is limited by $R_1^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{1,3}|^2}{N_0}\right)$, $R_2^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{2,3}|^2}{N_0}\right)$ and $R_{Sum}^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{1,3}|^2+|\mathbf{h}_3^H\mathbf{w}_{2,3}|^2}{N_0}\right)$.

Accordingly, the corresponding downlink beamformer design problem can be expressed as

$$\max_{\mathbf{w}_{2,3},\mathbf{w}_{1,3}} \min(R_{\text{MAC}}^3, R_1^1, R_1^2), \tag{6}$$

where the rates of users 1 and 2 are given as

$$R_1^1 = \log\left(1+\frac{|\mathbf{h}_1^H\mathbf{w}_{1,3}|^2}{|\mathbf{h}_1^H\mathbf{w}_{2,3}|^2+N_0}\right) \tag{7}$$

$$R_1^2 = \log\left(1+\frac{|\mathbf{h}_2^H\mathbf{w}_{2,3}|^2}{|\mathbf{h}_2^H\mathbf{w}_{1,3}|^2+N_0}\right). \tag{8}$$

Due to D2D transmissions, the beamformer design problem is different as compared to [15]. The partial file exchange in the D2D phase alleviates the interference conditions of the DL phase, thus, making the DL multicasting more efficient and less complex. On the other hand, the D2D transmission requires an orthogonal allocation in time domain. This introduces an inherent trade-off between the amount of resources allocated to the D2D and DL phases.

[2]Symmetric rate is imposed to minimize the time needed to receive both messages $\tilde{X}_{1,3}$, and $\tilde{X}_{2,3}$.

Finally, the corresponding symmetric DL rate maximization is given as

$$\begin{aligned}
\max_{\mathbf{w}_{i,j},\gamma_l^k,r} \quad & r \\
\text{s. t.} \quad & r \le \frac{1}{2}\log(1+\gamma_1^3+\gamma_2^3) \\
& r \le \log(1+\gamma_1^3),\ r \le \log(1+\gamma_2^3) \\
& r \le \log(1+\gamma_1^1),\ r \le \log(1+\gamma_1^2) \\
& \gamma_1^1 \le \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2+N_0}, \gamma_1^2 \le \frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3}|^2+N_0} \\
& \gamma_1^3 \le \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{N_0}, \gamma_2^3 \le \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{N_0} \\
& \|\mathbf{w}_{1,3}\|^2 + \|\mathbf{w}_{2,3}\|^2 \le \text{SNR}.
\end{aligned} \tag{9}$$

The rate constraints can be written as convex second-order cone constraints as shown in [15]. However, the signal-to-interference-plus-noise ratio (SINR) constraints are non-convex and require an iterative solution. A successive convex approximation (SCA) solution for the SINR constraints can be found, e.g., in [15]. Please notice that, due to D2D transmission in the the first phase, we have only two beamformers ($\mathbf{w}_{1,3}$ and $\mathbf{w}_{2,3}$), which means that we can dedicate more power to our intended signals ($X_{1,3}$ and $X_{2,3}$) when compared to [15]. The time required for the DL phase is

$$T_{\text{DL}} = \frac{F/3}{r} = \frac{F/3}{\max_{\mathbf{w}_{2,3},\mathbf{w}_{1,3}} \min(R_{\text{MAC}}^3, R_1^1, R_1^2)}. \tag{10}$$

Note that, also in this phase, all users are served with coded messages of size $\frac{F}{3}$ bits, which are multiplexed with the help of the beamforming vectors. Finally, the achievable rate over the D2D and DL phases is given in (3).

B. No D2D transmission available

Here, we make the same assumptions on the network parameters as in Section III-A. Further, we assume that the D2D transmission is not available. Thus, we have only the DL phase and the corresponding DL transmission is $\mathbf{x}_{DL} = \tilde{X}_{1,2}\mathbf{w}_{1,2} + \tilde{X}_{1,3}\mathbf{w}_{1,3} + \tilde{X}_{2,3}\mathbf{w}_{2,3}$. Therefore, all the users are served by 2 intended messages and see one message as interference. Now, suppose that all the users are able to decode *both* of their required messages with equal rate

$$R_{MAC}^i = \min\left(\frac{1}{2}R_{Sum}^i, R_1^i, R_2^i\right), \quad \text{for } i = 1,2,3, \tag{11}$$

where, for example, the rate region corresponding to $\tilde{X}_{1,3}$ and $\tilde{X}_{2,3}$, for user 3, is limited by $R_1^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{1,3}|^2}{N_0+|\mathbf{h}_3^H\mathbf{w}_{1,2}|^2}\right)$, $R_2^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{2,3}|^2}{N_0+|\mathbf{h}_3^H\mathbf{w}_{1,2}|^2}\right)$ and $R_{Sum}^3 = \log\left(1+\frac{|\mathbf{h}_3^H\mathbf{w}_{1,3}|^2+|\mathbf{h}_3^H\mathbf{w}_{2,3}|^2}{N_0+|\mathbf{h}_3^H\mathbf{w}_{1,2}|^2}\right)$. The rate region

for users 1 and 2 are similar to user 3. As in (6), the symmetric DL rate maximization, for this scenario, is given as

$$\begin{aligned}
\max_{\mathbf{w}_{i,j},\gamma_l^k,r} \quad & r \\
\text{s. t.} \quad & r \leq \frac{1}{2}\log(1+\gamma_1^i+\gamma_2^i) \text{ for } i=1,2,3 \\
& r \leq \log(1+\gamma_1^i),\ r \leq \log(1+\gamma_2^i) \text{ for } i=1,2,3 \\
& \gamma_1^1 \leq \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{N_0 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2}, \gamma_2^1 \leq \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{N_0 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2} \\
& \gamma_1^2 \leq \frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{N_0 + |\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3}|^2}, \gamma_2^2 \leq \frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{N_0 + |\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3}|^2} \\
& \gamma_1^3 \leq \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{N_0 + |\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2}, \gamma_2^3 \leq \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{N_0 + |\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2} \\
& \|\mathbf{w}_{1,2}\|^2 + \|\mathbf{w}_{1,3}\|^2 + \|\mathbf{w}_{2,3}\|^2 \leq \text{SNR}.
\end{aligned} \tag{12}$$

The time required for the DL phase is given by

$$T_{\text{DL}} = \frac{F/3}{r} = \frac{F/3}{\max_{\mathbf{w}_{1,2},\mathbf{w}_{1,3},\mathbf{w}_{2,3}} \min_{i\in\{1,2,3\}} R_{\text{MAC}}^i}. \tag{13}$$

Now, we can compare the computational complexity involved in both of these examples by comparing (12) and (9). It can be seen that, by only one D2D transmission, the total number of MAC conditions (rate inequalities) has reduced from 9 in (12) to 5 in (9), which is by almost half. On the other hand, The total number of quadratic terms ($\mathbf{w}_{i,j}$ in $|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{i,j}|^2$), has reduced from 12 (when D2D is not available) to 6 when user 1 and 2 are close to each other (Fig. 2). It should be noted that the number of MAC and quadratic terms greatly dominates the computational complexity when solving these problems via successive convex approximation [15]. As such, the D2D transmission can be considered as a means to reduce the beamforming complexity very efficiently in addition to the improved total delivery time.

IV. General analysis for complexity reduction in beam design

So far, we have shown the effects of D2D transmission in beamforming complexity for a simple example. In this section, we investigate the effects of D2D transmission in computational complexity for the general case. A more thorough investigation from the performance (delivery time, symmetric rate) improvement perspective is provided in [16]. It was shown in [15], that the number of MAC conditions and quadratic terms in the SINR constraints dominate the complexity of the DL beamformer design in finite SNR CC. To this end, we first introduce two boundaries for the number of MAC conditions, then discuss the effects of D2D on the beamformer design complexity.

Theorem 1. *Maximum and minimum number of MAC conditions for the DL Phase when i subsets of users have been chosen for D2D transmission are*

$$MAC_{min}^i = (t+L-b)(2^a-1) + b(2^{a+1}-1), \tag{14}$$

$$a = \left\lfloor \frac{(t+1)(\binom{t+L}{t+1} - i)}{t+L} \right\rfloor, \quad \textit{for} \quad i \leqslant \binom{t+L}{t+1}, \tag{15}$$

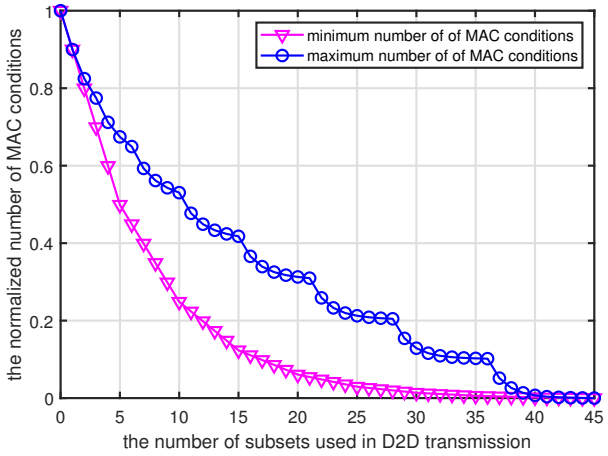

Fig. 3. The normalized number of MAC conditions Vs the number of subsets chosen for D2D transmission for $k=10,\ L=9$ and $t=1$.

$$b = (t+1)(\binom{t+L}{t+1} - i) - a(t+L), \tag{16}$$

$$\begin{aligned}
MAC_{max}^i = {} & (t+L-U)(2^{\binom{t+L-1}{t}} - 1) + \\
& (U-(U_1+1))(2^{(\binom{t+L-1}{t} - \binom{U-2}{t})} - 1) + \\
& U_1(2^{(\binom{t+L-1}{t} - (\binom{U-2}{t} + \binom{U_1-1}{t-1} - Y))} - 1) + \\
& (2^{\binom{t+L-1}{t} - X} - 1),
\end{aligned} \tag{17}$$

$$X = i - \binom{U-1}{t+1}, \textit{ for } \binom{U-1}{t+1} < i \leqslant \binom{U}{t+1},$$

$$U \leqslant (t+L), \tag{18}$$

$$\binom{U_1-1}{t} < X \leqslant \binom{U_1}{t}, \tag{19}$$

$$Y = \binom{U_1}{t} - X, \tag{20}$$

where i is the number of subsets (time slots) that have been chosen for D2D transmission and $t = \frac{KM}{N}$ (M is the normalized user cache size). Please note that, in this paper we assume $\binom{a}{b} = 0$, for $b > a$.

Proof of the theorem is omitted due to the lack of space, proofs are available in extended version of this paper [16]. The number of MAC conditions vary between these two boundaries based on which subsets have been chosen for D2D subphase. Fig. 3 shows the normalized maximum and minimum number of MAC conditions ($K=10,\ L=9,\ t=1$) for different number of D2D transmissions (the number of time slots). It is evident, that the number of MAC conditions decreases drastically by using just few D2D transmissions, which greatly reduced the complexity of the DL beamformer design. For the case depicted in Fig. 3, by choosing only 5 different subsets of users among 45 available subsets, the number of MAC conditions can be reduced to half. Therefore, D2D transmission have a significant gain from the aspect of complexity reduction in beamformer design.

Another important factor in beamforming complexity is the number of quadratic terms in SINR constraints. Next,

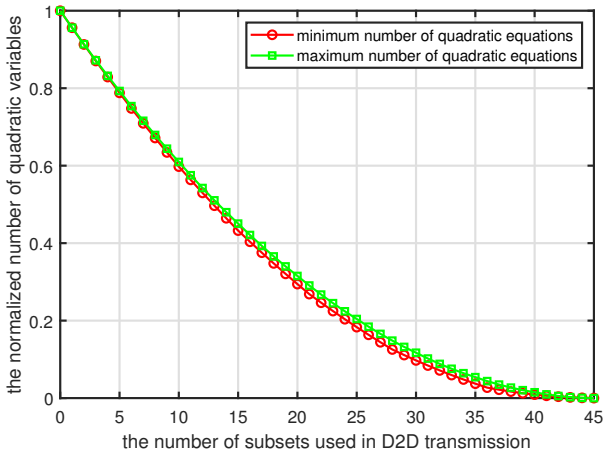

Fig. 4. Normalized number of quadratic variables vs the number of subsets used in D2D transmission.

we provide the boundaries for quadratic number of quadratic terms in the SINR constraints.

Theorem 2. *Maximum and minimum number of quadratic terms, when i subsets of users have been chosen for D2D sub-phase, is*

$$Q^i_{max} = bA_2B_2 + (t+L-b)A_1B_1 \tag{21}$$
$$A_1 = a, \quad A_2 = a+1,$$
$$B_1 = \mathbf{M}^i_T - A_1 + 1, \quad B_2 = \mathbf{M}^i_T - A_2 + 1$$
$$\mathbf{M}^i_T = \binom{t+L}{t+1} - i,$$
$$Q^i_{min} = (t+L-U)A^{'}_1B^{'}_1 + (U-(U_1+1))A^{'}_2B^{'}_2 + \tag{22}$$
$$U_1A^{'}_3B^{'}_3 + A^{'}_4B^{'}_4,$$
$$A^{'}_1 = \binom{t+L-1}{t}, \quad A^{'}_2 = A^{'}_1 - \binom{U-2}{t},$$
$$A^{'}_3 = A^{'}_2 - \binom{U_1-1}{t-1} + Y, \; A^{'}_4 = A^{'}_1 - X,$$
$$B^{'}_1 = \mathbf{M}^i_T - A^{'}_1 + 1, \quad B^{'}_2 = \mathbf{M}^i_T - A^{'}_2 + 1,$$
$$B^{'}_3 = \mathbf{M}^i_T - A^{'}_3 + 1, \quad B^{'}_4 = \mathbf{M}^i_T - A^{'}_4 + 1,$$

where Q^i_{max} is the maximum number of quadratic terms, Q^i_{min} is the minimum number of terms, $\mathbf{M}^i_T$ is the total number of messages that are sent by BS. Moreover, a is defined in (15)*, b is defined in* (16)*, X and U are defined in* (18)*, U_1 is defined in* (19) *and Y is defined in* (20)*.*

See [16] for proof. Fig. 4 depicts the upper and lower boundaries for the same scenario as Fig. 3. The bounds for the quadratic terms are fairly tight and the gap between these two bounds is not as large as for the MAC conditions. Thus, the number of MAC conditions is more affected by the way we choose different subsets for D2D transmission (compared to the quadratic terms). Moreover, the role of D2D transmission in the reduction of total number of quadratic terms is major. By choosing 13 different subsets, we can reduce the total number of quadratic terms (for this case) to half.

It is worth to mention that, in general, we assume that $t + L \leqslant K$, when $t + L < K$. Then, we have $\binom{K}{t+L}$ different transmission phases. Thus, all the equations in this paper are valid for each of these transmission phases.

V. Conclusions

Complexity analysis for CC beamforming assisted with D2D has been provided. In this manner, complexity bounds have been provided for computationally dominant constraints in DL multicast CC beamformer design. These bounds provide trade-off between computational complexity and benefits of DL multicast CC. Further, they provide important insight on the benefits of D2D enabled CC beyond the potential throughput improvement. The provided results assist in mode selection design for practical deployments, where the computational complexity has significant implications regarding transmission scheme feasibility.

References

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

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.

[3] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16–22, August 2016.

[4] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[5] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3212–3229, Jun 2016.

[6] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, Apr 2016.

[7] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec 2016.

[8] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inform. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[9] ——, "Cache-aided interference management in wireless cellular networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–7.

[10] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1176–1188, June 2018.

[11] K. H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[12] E. Piovano, H. Joudeh, and B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun 2017, pp. 2795–2799.

[13] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna coded caching," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun 2017, pp. 2113–2117.

[14] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multicast beamformer design for coded caching," in *2018 IEEE International Symposium on Information Theory (ISIT) (ISIT'2018)*, Vail, USA, Jun. 2018.

[15] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *CoRR*, vol. abs/1711.03364, 2018. [Online]. Available: http://arxiv.org/abs/1711.03364

[16] H. B. Mahmood, J. Kalev, S. P. Shariatpanahi, B. Khalaj, and A. Tölli, "D2D assisted beamforming for coded caching," *arXiv preprint arXiv:1905.05446*, 2019. [Online]. Available: http://arxiv.org/abs/1905.05446