# D2D Assisted Beamforming for Coded Caching

Hamidreza Bakhshzad Mahmoodi⋆, Jarkko Kaleva⋆, Seyed Pooya Shariatpanahi∗ and Antti Tölli⋆
⋆ Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014, Finland
∗ School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran,
firstname.lastname@oulu.fi, p.shariatpanahi@ut.ac.ir

***Abstract*—Device-to-device (D2D) aided beamforming for coded caching is considered in a finite signal-to-noise ratio regime. A novel beamforming and resource allocation scheme is proposed where the local cache content exchange among nearby users is exploited. The transmission is split into two phases: local D2D content exchange and downlink transmission. In the D2D phase, users can autonomously share content with the adjacent users. The downlink phase utilizes multicast beamforming to simultaneously serve all users to fulfill the remaining content requests. A low complexity D2D-multicast mode selection algorithm is proposed with comparable performance to the optimal exhaustive search. We first explain the main procedure via one simple example and then present the general formulation. Furthermore, D2D transmission scenarios and conditions useful for minimizing the overall delivery time are identified. By exploiting the direct D2D exchange of file fragments, the common multicasting rate for delivering the remaining file fragments in the downlink phase is increased, providing greatly enhanced overall content delivery performance.**

## I. Introduction

Caching popular content near or at the end-users is a widely accepted solution for supporting high-quality content delivery in next generation networks. This solution benefits from off-peak hours of the network to move some contents closer to the end-users, which later on will be used to mitigate the content delivery burden in network peak hours. Many recent papers have investigated the potentials of this paradigm to improve wireless network performance, such as [1]. A new promising scheme in this context is proposed in [2], which is known as the so-called *coded caching (CC)* approach. In this scheme, instead of locally caching some entire files at the end-user [3], fragments of all the files in the library are stored in the users' cache memories. In the delivery phase, carefully formed coded messages are multicasted to groups of users, which results in *global caching gain* [2].

CC has been shown to be greatly beneficial for both wired and wireless content delivery, under various assumptions [2], [4]–[6]. The original coded caching setup is extended in [4] to a multiple server scenario under different network topologies, aiming to further minimize the required delivery time of requested content. For high signal-to-noise ratio (SNR) regime, [5], [6] show that coded caching can boost the performance of the wireless network in terms of Degrees-of-Freedom (DoF). Specifically, in wireless broadcast channels with a multiple-antenna base station, the global coded caching gain and the spatial multiplexing gain are shown to be additive which will further increase the network data rate [4], [5].

In order to bridge the gap between high-SNR analysis of CC and the practical finite-SNR scenarios, recent works on finite SNR regime have also shown CC to be greatly beneficial when the interference is properly accounted for [7]–[11]. While, the works [7] and [8] use a rate-splitting approach to benefit from the global caching gain and the spatial multiplexing gain at finite SNR, the work [9] follows a zero-forcing (ZF) based approach (extending the ideas in [4] to the finite-SNR setup), which is also order-optimal in terms of DoF. Moreover, the work [10], [11] extends [9] to a general beamformer solution which manages the interaction between interference and noise in a more optimal manner. The general interference management framework proposed in [10], [11], improves the finite-SNR performance of the coded caching in wireless networks significantly. However, optimal beamforming solution can be very complex in some scenarios depending on the network size. The complexity issues associated with the corresponding optimization problem are addressed in [11].

In order to improve the per-user rate and mitigate the complexity of the beamformers in the finite SNR region, this paper extends the work [11] to device-to-device (D2D) assisted scenarios. In this manner, the multicast beamforming of the file fragments in [11] is complemented by allowing a direct exchange of local cache contents. Though D2D delivery has been considered before in caching network (e.g., [3], [12]), in this work (unlike [12] and [3]) we consider CC approach [2] to fill up users' cache (thus every two users have some contents that can be transmitted to the other users) similar to [13]. To this end, authors in [14] have extended the approach in [13] to more realistic scenarios where they use placement delivery array (PDA) for cache placement in order to decrease the file division requirement. However, unlike [13], [14], we proposed to use a combination of D2D and DL delivery methods. The main idea is to allow closely-located users to carry out a portion of the delivery task locally, and hence, significantly improving the content delivery performance. In this paper, we assume that the D2D and downlink (DL) phases are orthogonal in the time domain.

Finding the optimal D2D opportunities in finite SNR is particularly challenging due to the high computational complexity for the DL multicast beamformer design. The optimal D2D/DL mode selection requires an exhaustive search for

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

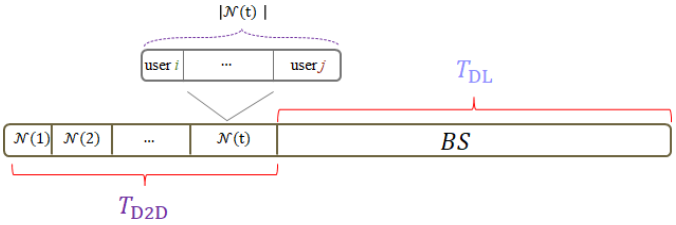


Fig. 1. Time division in D2D assisted transmission. Total time needed to transmit all fragments of files to the users is $T_{\mathrm{D2D}} + T_{\mathrm{DL}}$.

D2D opportunities over a group of users, which quickly becomes computationally intractable. To overcome these practical limitations, we provide a low complexity mode selection algorithm, which allows efficient determination of D2D opportunities even for a large number of users. The computational complexity of the proposed algorithm is significantly reduced with respect to the exhaustive search baseline while retaining comparable performance. Allowing direct D2D exchange of file fragments, the interference management between different downlink multicast streams becomes more relaxed and more efficient as compared to the multicast only case [11]. At the same time, the complexity of the delivery scheme is reduced both at the base station and at the end-users.

The results show that introducing the new D2D phase to the scheme [10] will enhance the performance of the network significantly. Moreover, we have shown the benefits of using D2D transmissions in the complexity reduction of the beamformers [15]. Therein, we provide upper/lower bounds on the number of conditions that should be considered in the design of the beamformers. It should be noted that, to the best of our knowledge, this paper is the first work proposing and analyzing D2D transmissions in a MISO-BC content delivery setup with coded caching techniques.

## II. System Model

We consider a system consisting of a single $L$ antenna base station (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 function of the files, denoted by $Z_k(W_1, \ldots, W_N)$, which is stored in the *cache content placement* phase during off peak hours (cache content placement is identical to [2]). 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 $\mathcal{N}(t)$, are 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 $X_i^{\mathrm{D2D}}$ to an intended set of receivers $\mathcal{R}^{\mathcal{N}}(i) \subseteq \mathcal{N}(t)$, which are interested in decoding $X_i^{\mathrm{D2D}}$. Thus, the message $X_i^{\mathrm{D2D}}$ 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 $h_{ik} \sim \mathcal{CN}(0,1)$ 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 the rest of the group members. Thus, the rate is limited by the weakest receiver

After all the D2D transmissions are done, 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 terminal $k = 1, \ldots, K$ is given by

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

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 $\mathbf{h}_k \in \mathcal{C}^L$, and the receiver noise is given by $z_k \sim \mathcal{CN}(0, N_0)$. 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 $T_{\mathrm{D2D}}$ and $T_{\mathrm{DL}}$ denote the time used for the D2D and downlink (DL) transmission sub-phases, respectively.

## III. D2D Aided Beamforming Explained: Example

Consider the scenario where $K = 4$ users and a library $\mathcal{W} = \{A, B, C, D\}$ of $N = 4$ files, where each user has a cache for storing $M = 2$ files. Also, the base station is equipped with $L = 2$ transmit antennas. Following the same placement as in [2], each file is split into $\binom{K}{\tau} = \binom{4}{2} = 6$ subfiles, as follows

$$\begin{aligned} A &= \{A_{1,2}, A_{1,3}, A_{1,4}, A_{2,3}, A_{2,4}, A_{3,4}\}, \\ B &= \{B_{1,2}, B_{1,3}, B_{1,4}, B_{2,3}, B_{2,4}, B_{3,4}\}, \\ C &= \{C_{1,2}, C_{1,3}, C_{1,4}, C_{2,3}, C_{2,4}, C_{3,4}\}, \\ D &= \{D_{1,2}, D_{1,3}, D_{1,4}, D_{2,3}, D_{2,4}, D_{3,4}\}. \end{aligned}$$

Each file $W_{\mathcal{T}}$ is cached at user $k$ if $k \in \mathcal{T}$. Let us assume that users $1-4$ request files $A-D$, respectively.

In this example, we suppose that users 1, 2, and 3 are close to each other, while user 4 is far from them. (see Fig. 2). Then, the D2D sub-phase consists of exchanging information between the first three users locally (collected in $\mathcal{N} = \{1, 2, 3\}$) [2] in three orthogonal D2D transmissions. More

[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.

[2] Since we are following cache placement in [2], in order to eliminate a term (as a whole) from (2), we set $|\mathcal{N}(t)| = \tau + 1$, $\tau = \frac{KM}{N}$.

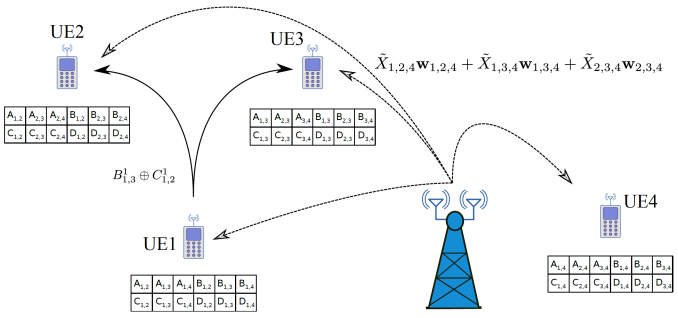


Fig. 2. Example: D2D enabled downlink beamforming system model.

specifically, each subfile is further divided into $\tau = 2$ fragments which are discriminated by their superscript indices[3]. Then, in the first D2D transmission of length $T\left(1 \to \mathcal{R}^{\mathcal{N}}(1)\right)$ seconds, user 1 multicasts $X_1^{D2D} = B^1_{1,3} \oplus C^1_{1,2}$ to $\mathcal{R}^{\mathcal{N}}(1) = \{2,3\}$. In the second D2D transmission, user 2 transmits $X_2^{D2D} = A^1_{2,3} \oplus C^2_{1,2}$ to $\mathcal{R}^{\mathcal{N}}(2) = \{1,3\}$, which will take $T\left(2 \to \mathcal{R}^{\mathcal{N}}(2)\right)$ seconds. Finally, in the third D2D transmission of length $T\left(3 \to \mathcal{R}^{\mathcal{N}}(3)\right)$ seconds, user 3 transmits $X_3^{D2D} = A^2_{2,3} \oplus B^2_{1,3}$ to $\mathcal{R}^{\mathcal{N}}(3) = \{1,2\}$. These transmissions require the total time of

$$\begin{aligned} T_{D2D} = \; & T\left(1 \to \mathcal{R}^{\mathcal{N}}(1)\right) + T\left(2 \to \mathcal{R}^{\mathcal{N}}(2)\right) + \\ & T\left(3 \to \mathcal{R}^{\mathcal{N}}(3)\right) \end{aligned} \tag{4}$$

in which $T\left(i \to \mathcal{R}^{\mathcal{N}}(i)\right) = \frac{F/12}{R_i^{\mathcal{N}}}, \quad i = 1,2,3$ and $R_i^{\mathcal{N}}, i = 1,2,3$ are determined by (1). Then, in the DL sub-phase, the BS transmits a message comprised of the remaining subfiles

$$\mathbf{x}_{DL} = \tilde{X}_{1,2,4}\mathbf{w}_{1,2,4} + \tilde{X}_{1,3,4}\mathbf{w}_{1,3,4} + \tilde{X}_{2,3,4}\mathbf{w}_{2,3,4}, \tag{5}$$

where $\tilde{X}_{1,2,4} = A_{2,4} \oplus B_{1,4} \oplus D_{1,2}$, $\tilde{X}_{1,3,4} = A_{3,4} \oplus C_{1,4} \oplus D_{1,3}$, and $\tilde{X}_{2,3,4} = B_{3,4} \oplus C_{2,4} \oplus D_{2,3}$[4]. At the end of this sub-phase, user 1 is interested in decoding $\{X_{1,2,4}, X_{1,3,4}\}$, user 2 is interested in decoding $\{X_{1,2,4}, X_{2,3,4}\}$, user 3 is interested in decoding $\{X_{1,3,4}, X_{2,3,4}\}$, and finally, user 4 is interested in decoding all the three terms $\{X_{1,2,4}, X_{1,3,4}, X_{2,3,4}\}$. Thus, from the perspective of users 1, 2, and 3, we have a MAC channel with two useful terms and one interference term. However, from the perspective of the user 4, we have a MAC channel with three useful terms. Thus, for users 1, 2, and 3 we have MAC rate region

$$R^k_{\text{MAC}} = \min(\frac{1}{2}R^k_{\text{sum}}, R^k_1, R^k_2), \qquad k = 1,2,3. \tag{6}$$

For example, for $k = 1$, we have $R^1_1 = \log\left(1 + \frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,2,4}|^2}{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{2,3,4}|^2 + N_0}\right)$, $R^1_2 = \log\left(1 + \frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,3,4}|^2}{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{2,3,4}|^2 + N_0}\right)$ and $R^1_{\text{sum}} = \log\left(1 + \frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,2,4}|^2 + |\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,3,4}|^2}{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{2,3,4}|^2 + N_0}\right)$.

[3]In our approach, each subfile is transmitted $\tau$ times in a D2D time slot. Thus, we further split each subfile into $\tau$ file fragments. Then, the user $i$ transmits new file fragments in $X_i^{D2D}$, which has not been transmitted before.

[4]For convenience superscript $\mathcal{S} = \{1,2,3,4\}$ in $\mathbf{w}^{\mathcal{S}}_{\mathcal{T}}\tilde{X}^{\mathcal{S}}_{\mathcal{T}}$ has been omitted in this example.

In order to derive the fourth user's 3-stream rate region, we face a MAC with three messages. Thus, we have 7 MAC region inequalities, which will result in $R^4_{\text{MAC}}$ (the details are omitted here due to lack of space. For details refer to [10], [11]). When all the MAC inequalities for all the users are considered together, we can derive the common multicast rate, which is shown in the corresponding downlink beamformer design problem as follows

$$\begin{aligned} & \max_{\mathbf{w}_{i,j,l}, \gamma^k_m, r} \quad r \\ & \text{subject to} \\ & r \le \tfrac{1}{2}\log(1+\gamma^k_1+\gamma^k_2), \;\; k = 1,2,3 \\ & r \le \log(1+\gamma^k_m), \;\; k = 1,2,3, m = 1,2 \\ & r \le \tfrac{1}{3}\log(1+\gamma^4_1+\gamma^4_2+\gamma^4_3), \;\; r \le \tfrac{1}{2}\log(1+\gamma^4_1+\gamma^4_2) \\ & r \le \tfrac{1}{2}\log(1+\gamma^4_1+\gamma^4_3), \;\; r \le \tfrac{1}{2}\log(1+\gamma^4_2+\gamma^4_3) \\ & r \le \log(1+\gamma^4_m), \;\; m = 1,2,3 \\ & \gamma^1_1 \le \frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,2,4}|^2}{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{2,3,4}|^2 + N_0}, \gamma^1_2 \le \frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,3,4}|^2}{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{2,3,4}|^2 + N_0} \\ & \gamma^2_1 \le \frac{|\mathbf{h}_2^{\text{H}}\mathbf{w}_{1,2,4}|^2}{|\mathbf{h}_2^{\text{H}}\mathbf{w}_{1,3,4}|^2 + N_0}, \gamma^2_2 \le \frac{|\mathbf{h}_2^{\text{H}}\mathbf{w}_{2,3,4}|^2}{|\mathbf{h}_2^{\text{H}}\mathbf{w}_{1,3,4}|^2 + N_0} \\ & \gamma^3_1 \le \frac{|\mathbf{h}_3^{\text{H}}\mathbf{w}_{1,3,4}|^2}{|\mathbf{h}_3^{\text{H}}\mathbf{w}_{1,2,4}|^2 + N_0}, \gamma^3_2 \le \frac{|\mathbf{h}_3^{\text{H}}\mathbf{w}_{2,3,4}|^2}{|\mathbf{h}_3^{\text{H}}\mathbf{w}_{1,2,4}|^2 + N_0} \\ & \gamma^4_1 \le |\mathbf{h}_4^{\text{H}}\mathbf{w}_{1,2,4}|^2/N_0, \gamma^4_2 \le |\mathbf{h}_4^{\text{H}}\mathbf{w}_{1,3,4}|^2/N_0 \\ & \gamma^4_3 \le |\mathbf{h}_4^{\text{H}}\mathbf{w}_{2,3,4}|^2/N_0 \\ & \|\mathbf{w}_{1,2,4}\|^2 + \|\mathbf{w}_{1,3,4}\|^2 + \|\mathbf{w}_{2,3,4}\|^2 \le P. \end{aligned}$$

Where $P$ is the total available power at the BS. Finally, the delivery time of the DL sub-phase is $T_{\text{DL}} = \frac{F/6}{r}$. It should be noted that, compared to the solution proposed in [10], one term is removed from the downlink transmission, i.e., $\tilde{X}_{1,2,3}\mathbf{w}_{1,2,3}$. We have taken care of this term in the D2D phase, which in turn enhances the performance of the downlink phase. First, since we have removed one term from DL transmission (if D2D transmission was not available, the BS had to transmit four terms [10]), the remaining beamformers will be allocated more power, so the DL rate is enhanced. Second, since one term is removed, the number of conditions is less compared to [10] (refer to [15] for more details on complexity analyzes).

## IV. D2D Aided Beamforming: The General Case

In this section, we formulate and analyze the proposed scheme in the general setting. The cache content placement phase is identical to the one proposed in [2]. In general, in each data transmission, $\min(\tau + L, K)$ users can be served simultaneously [11]. Thus, when $\tau + L < K$, $\binom{K}{\tau+L}$ transmission phases are required in total. Unlike in [11], here, the data delivery is split into D2D and DL sub-phases.

To examine the optimal user allocation for the D2D phase, we need to perform an exhaustive search among the D2D subsets. In total, there are $\binom{\tau+L}{\tau+1}$ different user subsets (of size $\tau+1$) among $\tau+L$ number of users in each transmission phase. Thus, the exhaustive search would require $2^{\binom{\tau+L}{\tau+1}}$ evaluations of (3). In each of these evaluations, all the beamformers must be solved, and the total rate computed. Then, the highest

one should be chosen. To simplify the notation, we consider an indication function $I_{D2D}(\mathcal{T})$, which specifies whether the corresponding subset has been allocated for D2D transmission. We define $C(K,\tau,L) = \frac{F}{\binom{K}{\tau}\binom{K-(\tau+1)}{L-1}}$ as the size of the transmitted subfile [11].

### A. Total delivery time $T_{\mathrm{D2D}} + T_{\mathrm{DL}}$

Now, for a given D2D mode allocation, the D2D delivery time is given as

$$T_{\mathrm{D2D}} = \sum_{\mathcal{T}\subseteq\overline{\Omega^{\mathcal{S}}}} \sum_{k\in\mathcal{T}} \frac{C(K,\tau,L)/\tau}{\mathcal{R}_k^{\mathcal{N}}}, \tag{7}$$

where $\overline{\Omega^{\mathcal{S}}} := \{\mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}| = \tau+1, I_{\mathrm{D2D}}(\mathcal{T}) = 1\}$ and $\mathcal{R}_k^{\mathcal{N}}$ is from (1). Since in each D2D subset, each subfile is transmitted by $\tau$ users, we further divide each subfile into $\tau$ file fragments so that we can transmit a distinct file fragment by each user (see the example in Section III).

The beamformers for the DL phase are solved using the SCA approach from [11]. The main difference, in contrast to [11], is that we should not consider all the $\tau+1$ subsets. Here, only those subsets $\mathcal{T}$ for which $I_{\mathrm{D2D}}(\mathcal{T}) = 0$ should be involved in the DL phase. This will reduce the interference between parallel streams significantly. The DL sub-phase throughput is given by

$$\begin{aligned} &R_C\left(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}| = \tau+1, I_{D2D}(\mathcal{T}) = 0\}\right) = \\ &\min_{k\in\mathcal{S}} R_{MAC}^k\left(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T}\subseteq\mathcal{S}, I_{\mathrm{D2D}}(\mathcal{T}) = 0\}\right) \end{aligned} \tag{8}$$

where

$$\begin{aligned} &R_{MAC}^k\left(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T}\subseteq\mathcal{S}, I_{\mathrm{D2D}}(\mathcal{T}) = 0\}\right) \\ &= \min_{\mathcal{B}\subseteq\Omega_k^{\mathcal{S}}} \left[\frac{1}{|\mathcal{B}|}\log\left(1 + \frac{\sum_{\mathcal{T}\in\mathcal{B}} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}|^2}{N_0 + \sum_{\mathcal{T}\in\Omega_{\mathcal{S}}\setminus\Omega_k^{\mathcal{S}}} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}|^2}\right)\right] \end{aligned} \tag{9}$$

where $\Omega^{\mathcal{S}} := \{\mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}| = \tau+1, I_{\mathrm{D2D}}(\mathcal{T}) = 0\}$ is the set of all the user subsets with size $\tau+1$ that will be served in DL phase[5]. The cardinality $|\Omega^{\mathcal{S}}|$ indicates the total number of messages delivered by the BS. Finally $\Omega_k^{\mathcal{S}} := \{\mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}| = \tau+1, I_{\mathrm{D2D}}(\mathcal{T}) = 0 \mid k\in\mathcal{T}\}$ is the set of all the messages required by user $k$.

After computing the rate for DL sub-phase the $T_{\mathrm{DL}}$ is computed as $T_{\mathrm{DL}} = \frac{C(K,\tau,L)}{R_C}$, then the achievable symmetric rate per user is computed using (3). For a large number of users and transmit antennas, solving (8) requires a considerable amount of computation, due to the iterative convex approximation for each subset evaluation [11]. In the following, we provide a low complexity heuristic solution for the proposed mode assessment problem.

### B. Heuristic D2D mode selection with low complexity

In order to decrease the computational load of evaluating $T_{\mathrm{D2D}}$ and $T_{\mathrm{DL}}$ for different D2D mode allocations, we provide a throughput approximation for the D2D mode allocations without having to rely on the general SCA solution for the DL beamformer design. The D2D transmissions occur in orthogonal time slots. The accumulated D2D phase duration is denoted by $T_{\mathrm{D2D}}$. Each successful D2D exchange reduces the remaining number of file fragments to be transmitted by the BS. Thus, there are fewer multicast messages and corresponding beamforming vectors $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ in the DL optimization problem. This allows a more efficient (less restricted) multicast beamformer design, which results in reduced DL phase duration $T_{\mathrm{DL}}$. The D2D mode selection is iteratively carried out as long as the following condition holds:

$$\frac{\hat{T}_{\mathrm{DL}}^i}{N_{\mathrm{F}} - (\tau+1)(i-1)} \geq \hat{T}_{\mathrm{D2D}}^i, i\in\left[1, \binom{\tau+L}{\tau+1}\right], \tag{10}$$

where $N_{\mathrm{F}} = (\tau+1)\binom{\tau+L}{\tau+1}$ is the total number of subfiles that should be delivered to all the users so that they can decode their intended files. Moreover, $\hat{T}_{\mathrm{DL}}^i$ and $\hat{T}_{\mathrm{D2D}}^i$ are the coarse approximated delivery times in the $i^{\text{th}}$ iteration. In (10), we check if any D2D user subset will reduce the DL duration $T_{\mathrm{DL}}$ more than the duration of the corresponding D2D transmission. If a specific subset $\mathcal{T}$ in iteration $i$ satisfies (10), then the D2D transmission for this subset is done following the approach proposed in [13] and $I_{\mathrm{D2D}}(\mathcal{T})$ is set to one for this subset.

In each D2D time slot, $\tau+1$ subfiles are delivered by $\tau+1$ orthogonal D2D transmissions. On the other hand, in the DL sub-phase, all the remaining subfiles $(N_{\mathrm{F}} - (\tau+1)(i-1))$ are delivered simultaneously. Thus, in (10), the average delivery time for a single subfile in the D2D and DL phases are compared. In each iteration, we choose a subset for D2D candidate, i.e., the subset, which provides the highest rate. If at any specific iteration, (10) does not hold, using more D2D transmissions will not improve the overall rate, and the iterative process is terminated. Therefore, at most $\binom{\tau+L}{\tau+1}$ iterations are required compared to $2^{\binom{\tau+L}{\tau+1}}$ needed for the exhaustive search.

The D2D delivery time is coarsely approximated as

$$\hat{T}_{\mathrm{D2D}}^i = \min_{\mathcal{T}\subseteq\Omega^{\mathcal{S}}} \hat{T}_{\mathrm{D2D}}^{\mathcal{T}}, \tag{11}$$

$$\hat{T}_{\mathrm{D2D}}^{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{k\in\mathcal{T}} \frac{C(K,\tau,L)/\tau}{\mathcal{R}_k^{\mathcal{N}}}. \tag{12}$$

Since, in each D2D transmission (e.g., user $i$'s transmission in Fig. 1), $1/\tau$ fraction of each subfile is delivered, $\hat{T}_{\mathrm{D2D}}^{\mathcal{T}}$ is considered as $\frac{C(K,\tau,L)/\tau}{\mathcal{R}_k^{\mathcal{N}}}$ to scale the delivery time. Here, the approximated D2D time for each subset is the average time which is needed to deliver a single subfile. In each D2D subset, there are $|\mathcal{T}| = \tau+1$ number of subfiles that are delivered by D2D transmissions, thus in (12) we have divided the total required time by $|\mathcal{T}|$ to compute the average time for single subfile. Note that, for each iteration $i$, we only consider those subsets that have not already been allocated for D2D.

The DL delivery time is coarsely approximated as

$$\begin{aligned} &\hat{T}_{\mathrm{DL}}^i = \frac{C(K,\tau,L)}{\hat{R}_{\mathrm{DL}}^i}, \quad \hat{R}_{\mathrm{DL}}^i = \min_{k\in[\mathcal{S}]} \hat{R}_k^i, \\ &\hat{R}_k^i = \frac{1}{|\Omega_k^{\mathcal{S}}|}\log\left(1 + \frac{|\Omega_k^{\mathcal{S}}|\|\mathbf{h}_k\|^2\mathrm{SNR}}{|\Omega^{\mathcal{S}}|}\right), \end{aligned} \tag{13}$$

[5]In Section III, $\Omega^{\mathcal{S}} = \{\{1,2,4\},\{1,3,4\},\{2,3,4\}\}$ and $|\Omega^{\mathcal{S}}| = 3$.

where $\hat{R}_k^i$ is the approximated rate of user $k$ considering that $(i-1)$ subsets had been chosen for D2D transmission in the previous iterations. Here, we have assumed that the beamformer $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ is able to completely remove the interference and is matched to the channel of the users in subset $\mathcal{T}$user $k \in \mathcal{T}$ receives the message $\hat{X}_{\mathcal{T}}^{\mathcal{S}}$ by SNR proportional to $\frac{\mathrm{P}_{\mathcal{T}}}{N_0}$. Where, $\mathrm{P}_{\mathcal{T}}$ is the dedicated power to the message $\hat{X}_{\mathcal{T}}^{\mathcal{S}}$, for simplicity we have assumed that the allocated powers to all the messages are equal (which is almost true for most of the cases), thus the received $\mathrm{SINR}_k \propto \frac{\mathrm{SNR}}{|\Omega^{\mathcal{S}}|} \propto \frac{P}{N_0|\Omega^{\mathcal{S}}|}$. In general, beamformer $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ should be designed in such a way that all the users in subset $\mathcal{T}$ can decode the message $\hat{X}_{\mathcal{T}}^{\mathcal{S}}$. Therefore, for the heuristic mode selection process we use the user's channel gain, assuming maximum ratio transmitter (MRT) beamforming ($\mathrm{SINR}_k = \frac{\|\mathbf{h}_k\|^2 \mathrm{SNR}}{|\Omega^{\mathcal{S}}|}$), and limit the rate to the weakest user [6] (to coarsely indicate the multicast beamforming potential for a given subset).

Once the users for D2D mode transmission are found based on (10), the final delivery time and the rate are computed as described in Section IV-A. In comparison to [11], for any specific subset $\mathcal{T}$, such that, $I_{D2D}(\mathcal{T}) = 1$, the coded messages corresponding to this subset are already delivered in the D2D phase. Thus, we can ignore such subsets, which results in less inter-message interference, and thus a lower delivery time at the DL sub-phase than in [11]. Finally, the complete algorithm is given in Algorithm 1.

## V. Numerical Examples

In this section, we provide numerical examples for two scenarios with $K = 3$ and $K = 4$ users. In these scenarios, we consider a circular cell with a radius of $R = 100$ meters, while the BS is in the center of the cell. In order to see the effect of D2D transmission in different situations, we introduce a smaller circle with radius $r$ within the cell area, wherein the users are randomly scattered. Thus, the maximum distance between any pair of two users is $2r$, while the users' distance to BS varies between 0 and $R$. By changing $r$, the maximum users' separation in D2D mode can be controlled. This helps us determine the beneficial users' distance in the D2D phase. In the simulations, we set the pathloss exponent 2 for D2D channels $h_{ik}$ and 3 for DL channels $\mathbf{h}_k$.

Transmit powers for D2D transmission at user side are adjusted in a way that the received SNR at the receiver is 0 dB at 10 meter distance. Also, the transmit power at the BS is adjusted such that the received SNR is 0 dB at 100 meter distance. For comparison, we have also considered the state-of-the-art (SoA) [10], where no D2D is allowed (Multicasting only). We also consider another SoA scenario [13] ( *D2D only*) for comparison.

Fig. 3 shows the per user rate for $K = 3$, $L = 2$ and $\tau = 1$ case as a function of inner circle radius (the example in [15]). The figure demonstrates that, when users are close to each other, there is a significant gain from using a combination of multicasting and D2D transmissions. When the maximum distance among the users starts to increase, the *D2D only* rate decreases drastically. However, the proposed approach in this paper shows more robust behavior as the users start to increase their distances. The beneficial range for D2D transmission in this particular scenario appears to be between $r = 0$ and 5m (10m maximum distance). Obviously, the range can change significantly if pathloss exponent, D2D and DL available power, $\tau$, etc. are varied. As the simulation results show, sending all the data through D2D transmissions or sending them only through multicasting results in a lower rate compared to our approach, which is the optimum combination of these two.

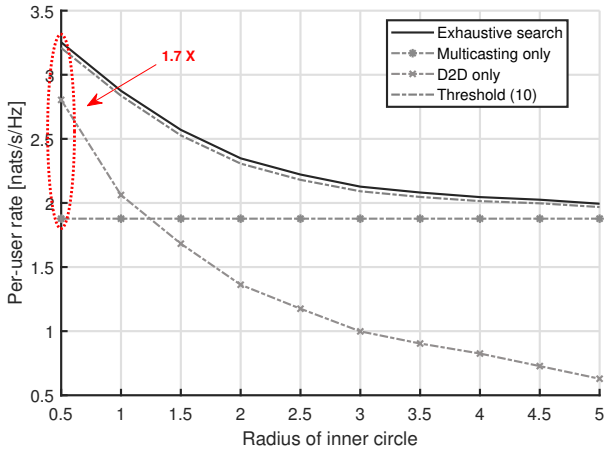


Fig. 3. Per user rate vs. small circle radius $r$ for $K = 3$ and $t = 1$.

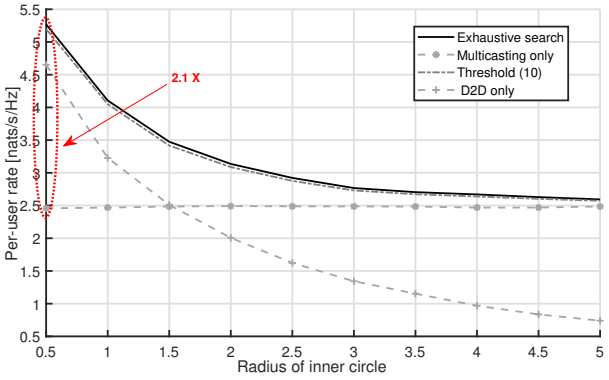


Fig. 4. Per user rate vs. small circle radius $r$ for $K = 4$ and $t = 2$.

Fig. 4 shows the per-user rate versus inner circle radius for $K = 4$, $\tau = 2$, and $L = 2$ (section III). For a higher number of users, the gain from using D2D transmission among nearby users is larger than the case $K = 3$ due to more D2D transmission opportunities. However, the gain of D2D transmission decreases more rapidly compared to the case $K = 3$; the reason is that, since $\tau = 2$, we need more users to be closer to each other in order to be able to perform the D2D transmission efficiently. In general, increasing the number of

[6] Another interpretation for (13) is that the beamformer $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ is assumed to be matched to the weakest user in subset $\mathcal{T}$ without rate loss for other users with better channel condition in the subset.

**Algorithm 1** D2D Assisted Multi-Antenna Coded Caching

**procedure** DELIVERY($W_1, \ldots, W_N$, $d_1, \ldots, d_K$, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$)
  $\tau \leftarrow MK/N$
  **for** $i \in \left[1, \binom{\tau+L}{\tau+1}\right]$ **do** ▷ This is the beginning of the D2D Phase.
    **if** $\frac{\hat{T}^i_{\mathrm{DL}}}{N_{\mathrm{F}}-(\tau+1)(i-1)} \geq \hat{T}^i_{\mathrm{D2D}}$ **then**

      **for all** $k \in \mathcal{T}$ **do** ▷ Each loop pass is one D2D transmission. In each loop a subset $\mathcal{T}$ is selected based on (11).

        Each sub-file is divided into $\tau$ mini-file fragments.
        $X_k^{\mathcal{T}} \leftarrow \oplus_{i \in \mathcal{T}\backslash\{k\}} NEW(W_{d_j, \mathcal{T}\backslash\{i\}})$
        User $k$ multicasts $X_k^{\mathcal{T}}$ to $\mathcal{R}^{\mathcal{T}}(k) = \mathcal{T}\backslash\{k\}$ with the rate $R_k^{\mathcal{T}}$ stated in (1)
        $I_{D2D}(\mathcal{T}) = 1$
      **end for**
    **end if**

  **end for** ▷ This end of the D2D Phase, which was based on the approach used in [13].
  **for all** $\mathcal{S} \subseteq [K], |\mathcal{S}| = \min(\tau + L, K)$ **do** ▷ This is the beginning of the DL phase.
    **for all** $\mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = \tau + 1, I_{D2D}(\mathcal{T}) = 0$ **do**
      $X_{\mathcal{T}}^{\mathcal{S}} \leftarrow \oplus_{k \in \mathcal{T}} NEW(W_{d_k, \mathcal{T}\backslash\{k\}})$
    **end for**
    $\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}\} = \underset{\{\underline{\mathbf{w}}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}|=\tau+1, I_{D2D}(\mathcal{T})=0\}}{\arg\max} R_C\left(\mathcal{S}, \{\underline{\mathbf{w}}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = \tau + 1, I_{D2D}(\mathcal{T}) = 0\}\right)$ ▷ $R_C$ is defined in (8).
    $\underline{\mathbf{X}}(\mathcal{S}) \leftarrow \sum_{\mathcal{T}\subseteq\mathcal{S}, |\mathcal{T}|=\tau+1, I_{D2D}(\mathcal{T})=0} \mathbf{w}_{\mathcal{T}}^{\mathcal{S}} \tilde{X}_{\mathcal{T}}^{\mathcal{S}}$
    **transmit** $\underline{\mathbf{X}}(\mathcal{S})$ with the rate $R_C(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = \tau + 1, I_{D2D}(\mathcal{T}) = 0\})$.
  **end for** ▷ This is the end of the DL phase, which was based on the approach used in [10].
**end procedure**

users in the network will result in a higher gain. However, increasing $\tau$ will result in decreasing the opportunities for D2D transmission, therefore, increasing the slope of the rate curve.

It is worth to mention that, using the heuristic D2D mode selection criteria (defined in Section IV) results in a minimal loss in per-user rate, with a greatly reduced complexity, as compared to the exhaustive search. Simulation result show that the approximated rate (13) is very close the actual rate (9) for different $\Omega^{\mathcal{S}}$ and different network parameters (i.e., $\tau$, $L$, $K$, etc), so the proposed approach follows exhaustive search closely.

## VI. Conclusions

A novel delivery scheme optimized for finite SNR region was proposed, where the multicast beamforming of file fragments is complemented by allowing direct D2D exchange of local cache content. The benefits of partial D2D offloading of multicast delivery of coded caching content were investigated. A straightforward example scenario was assessed in detail, and a generalized formulation was also provided. Moreover, we proposed a heuristic low complexity mode selection scheme with comparable performance to the optimal exhaustive search. We showed that using D2D greatly enhances the per-user rate and results in less restricted (less complex) beamforming problem. In future work, we are going to analyze the effects of D2D communication in the energy efficiency aspect of the network.

## References


[1] 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.

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

[3] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Inter-cluster cooperation for wireless d2d caching networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6108–6121, Sep. 2018.

[4] S. P. Shariatpanahi *et al.*, "Multi-server coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec 2016.

[5] 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.

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

[7] 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.

[8] 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.

[9] S. P. Shariatpanahi *et al.*, "Multi-antenna coded caching," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun 2017, pp. 2113–2117.

[10] 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)*, Vail, USA, Jun. 2018.

[11] A. Tölli *et al.*, "Multi-antenna interference management for coded caching," *CoRR*, vol. abs/1711.03364, 2018. [Online]. Available: http://arxiv.org/abs/1711.03364

[12] X. Li, X. Wang, P. Wan, Z. Han, and V. C. M. Leung, "Hierarchical edge caching in device-to-device aided mobile networks: Modeling, optimization, and design," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1768–1785, Aug 2018.

[13] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inform. Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[14] J. Wang, M. Cheng, Q. Yan, and X. Tang, "Placement delivery array design for coded caching scheme in D2D networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3388–3395, May 2019.

[15] H. B. Mahmoodi, J. Kaleva, and A. Tölli, "Complexity reduction in multicast beamforming for D2D assisted coded caching," in *Proc. IEEE Int. Symp. on Wireless Commun. Systems*, Aug. 2019, pp. 239–243.