Mobile Networks and Applications (2019) 24:1603–1617
https://doi.org/10.1007/s11036-018-1181-1

Device-to-Device Communications Enabled Multicast Scheduling with the Multi-level Codebook in mmWave Small Cells

Yong Niu[1] · Liren Yu[2] · Yong Li[2] · Zhangdui Zhong[1] · Bo Ai[1] · Sheng Chen[3,4]

Published online: 20 November 2018

Abstract

To keep up with the rapid growth of mobile data, there are increasing interests to deploy small cells in millimeter wave (mmWave) bands to underlay the conventional homogeneous macrocell network as well as in exploiting device-to-device (D2D) communications to improve the efficiency of the multicast service that supports content-based mobile applications. To compensate for high propagation loss in the mmWave band, high-gain directional antennas have to be employed, while it is critical to optimize multicast service in order to improve the network performance. In this paper, we develop an efficient multicast scheduling scheme for small cells in the mmWave band, called MD2D, where both D2D communications in close proximity and multi-level antenna codebook are utilized. Specifically, a user partition and multicast path planning algorithm is proposed to partition the users in the multicast group into subsets and to determine the transmission node for each subset, so as to achieve optimal utilization of D2D communications and multi-level antenna codebook. Then a multicast scheduling algorithm schedules the transmission for each subset. Furthermore, in order to optimize the network performance, the optimal choice of user partition thresholds is analyzed. Extensive simulations demonstrate that the MD2D achieves the best performance, in terms of network throughput and energy efficiency, compared with other existing state-of-the-art schemes. MD2D improves the network throughput compared with the second-best scheme by about 27%.

Keywords Millimeter wave communication · Device-to-device communication · Small cells · Multicast service · Directional antenna · Multi-level codebook

1 Introduction

Mobile traffic demands are explosively increasing over the past decade [1]. In order to improve mobile network capacity accordingly, small cells underlying the macrocell network have been proposed and received much attention. This new network deployment is usually referred to as heterogeneous cellular networks (HCNs). However, reducing the radii of small cells in the carrier frequencies employed in today's cellular systems to reap the spatial reuse benefits is limited by interference constraints [2]. By utilizing higher frequency bands, such as the millimeter wave (mmWave) bands, and bringing the network closer to users by a dense deployment of small cells [2, 3], HCN can significantly boost the overall network capacity while imposing less interference, compared to a conventional HCN deployment. With abundant amount of spectrum in the mmWave band, such as the 28 GHz band, the 38 GHz band, the 60 GHz band, and the E-band (71–76 and 81–86 GHz), multi-Gbps communication services, e.g., uncompressed and high-definition video transmission and wireless gigabit Ethernet for laptops and desktops, can easily be supported [4]. Moreover, rapid progress in complementary metal-oxide-semiconductor technology for radio frequency integrated circuits promotes the popularity of electronic products in the mmWave band [5]. Several standards have been defined for indoor wireless personal area networks (WPAN) and wireless local area networks (WLAN), such as IEEE 802.15.3c [6] and IEEE 802.11ad [7]. The standard for the next generation ultra high data rate communications in the 60 GHz band, IEEE802.11ay, is targeting at least one mode of operation capable of supporting a maximum throughput of no less than 20 Gbps [8].

Different from conventional communication systems in lower frequency bands, mmWave communications suffer from higher propagation loss. Consequently, directional antennas are synthesized at both transmitter and receiver to

✉ Yong Niu
niuy11@163.com

Extended author information available on the last page of the article.

form directional high-gain beams in order to combat the high propagation loss [9–12]. It is well-known that wide beams have low antenna gains and can only support low transmission rates, while narrow beams have high antenna gains and can support high transmission rates. With the aid of the multi-level antenna codebook, therefore, transmitters can use wide beams to communicate with multiple low-rate receivers, and use narrow beams to communicate at a high transmission rate. Hence, by optimizing the beam selection in a multi-level antenna codebook, the flexibility of directional beams can be exploited to improve the achievable network performance.

In content downloading, a small amount of popular content typically account for the majority of requests, and the content popularity in mobile networks is found to obey the Zipf's law [13]. Multicast service, which provides multiple users of a multicast group with the same data from the access point (AP), can support applications like TV content streaming, advertising messages broadcasting, and broadcast communication services for police, fire and ambulance [14, 15]. To serve more users simultaneously, wider beams are preferred. But wider beams with lower antenna gain can only support lower transmission rate, which degrades multicast efficiency. Thus, it is necessary to partition the multicast group into multiple subsets, and select an appropriate beam to serve each subset. On the other hand, in the user-intensive region, where the multicast service is usually applied, the probability that two user devices are located near to each other will be high. In this case, device-to-device (D2D) communications in physical proximity can be exploited for improving multicast efficiency, owing to better channel conditions as well as power saving and enhanced spectral efficiency [16]. For the multicast traffic, users already with the multicast content can use the D2D communications to forward the content to users nearby. Thus, D2D communications usually have shorter distance, and thus the propagation loss is less. Consequently, higher transmission rate can be obtained, and thus less transmission time is needed. Thus, network throughput can be increased. With less time needed to accommodate the multicast demands, the energy consumption can be reduced accordingly if the transmission power is fixed. Therefore, the energy efficiency can be increased.

Against the above background, in this paper, we investigate the problem of optimal multicast scheduling in mmWave small cells underlaid by D2D communications. This problem is challenging because to serve users efficiently, users in the multicast group must be partitioned optimally into subsets and beams must be selected optimally to serve users in each subset. Moreover, user device with the multicast traffic must be capable of serving other subsets efficiently by exploiting better D2D channels. The D2D communication here is for multicast transmission, and "D2D" here means "one device to multiple devices using directional beams", which is very different from previous works. To obtain a practical solution, we propose an efficient multicast scheduling scheme, referred to as MD2D, where appropriate beams are selected to serve users efficiently in each multicast transmission, while D2D communications are utilized to improve multicast efficiency. Our contribution is three-fold as summarized below.

- We formulate the problem of the optimal multicast scheduling with D2D communications and beam selection in a multi-level codebook considered into a mixed integer nonlinear program that minimizes the total multicast transmission time, by efficient multicast group partition, beam selection, and D2D communication utilization.
- To obtain a practical solution to this challenging problem, an efficient multicast scheduling scheme is proposed, called MD2D, where two algorithms are proposed, user partition and multicast path planning, and multicast scheduling. The first algorithm appropriately partitions the users in the multicast group into subsets and determines the transmission node for each subset, while the second one schedules the transmission for each subset efficiently.
- We further investigate the optimal selection of user partition thresholds to optimize the achievable network performance. Extensive evaluations under various system settings demonstrate that our proposed MD2D achieves the best performance, in terms of network throughput and energy efficiency, compared with other existing state-of-the-art schemes.

The rest of this paper is organized as follows. Section 2 reviews the related work on the media access control (MAC) protocols for mmWave small cells. Section 3 introduces the system model and illustrates the basic idea of our MD2D. Section 4 formulates the optimal multicast scheduling problem with D2D communications, multicast group partition and beam selection in a multi-level codebook. Section 5 is entirely devoted to our proposed multicast scheduling scheme, namely, MD2D. Section 6 evaluates the performance of our MD2D scheme, in terms of network throughput and energy consumption, using there existing schemes as the benchmarks. Section 7 gives the conclusion.

2 Related work

There exist some related works on directional MAC protocols for WPANs and WLANs in the mmWave band [17–21]. In WPANs and WLANs, time division multiple access (TDMA) protocol is traditionally adopted [6, 22]. Cai et al. [18] derived the conditions of exclusive region that concurrent transmission always outperforms TDMA. Based on IEEE 802.15.3c, two protocols utilize concurrent transmissions are enabled when the multi-user interference is below a specific threshold [19, 20]. In an indoor IEEE 802.15.3c WPAN, the work of [17] proposed a concurrent transmission scheduling algorithm, where concurrent transmissions are optimized to maximize the number of flows with the quality of service requirement for each flow satisfied. Qiao et al. [21] also proposed a multi-hop concurrent transmission scheme to address the link outage problem and to improve flow throughput. For TDMA based protocols, there is an unfair medium time allocation problem for individual users under bursty data traffic [23].

There also exist some centralized protocols proposed for WPANs or WLANs in the mmWave band [23–26]. Specifically, Gong et al. [24] proposed a scheduling scheme based on the traditional carrier sense multiple access with collision avoidance (CSMA/CA) protocol. In the protocol, the piconet coordinator (PNC) distributes the network allocation vector information and utilizes the virtual carrier sensing to overcome the deafness problem. In [25], a multihop-relay based directional MAC (MRDMAC) protocol is proposed, where the PNC applies a weighted round robin scheduling to overcome the deafness problem. If a wireless terminal (WT) is lost due to blockage, the AP will choose a live WT to act as a relay to the lost WT. Son et al. [23] proposed a frame based directional MAC protocol (FDMAC), which amortizes the scheduling overhead over multiple concurrent transmissions in a row to achieve high efficiency. The core of this FDMAC is the greedy coloring algorithm, which utilizes concurrent transmissions to improve the network throughput. Chen et al. [26] proposed a directional cooperative MAC protocol, called D-CoopMAC, to manage the uplink channel access in an IEEE 802.11ad WLAN. Niu et al. [27] proposed a blockage-robust and efficient directional MAC (BRDMAC) protocol to overcome the blockage problem by two-hop relaying. In BRDMAC, relay selection and spatial reuse are jointly optimized to achieve a near-optimal network performance, in terms of delay and throughput. Recently, Niu et al. [28] proposed a joint transmission scheduling protocol for the radio access and backhaul of small cells in mmWave band, called D2DMAC. In D2DMAC, a path selection criterion is proposed to exploit D2D transmissions when performance improvement is available. Zhang et al. [29] investigate user association and power allocation in mmWave-based ultra dense networks with attention to load balance constraints, energy harvesting by base stations, user quality of service requirements, energy efficiency, and cross-tier interference limits. To solve the joint user association and power optimization problem, they proposed an iterative gradient user association and power allocation algorithm to achieve an optimal point.

In terms of multicast communication, there also exist a few works on MAC protocols for WPANs and WLANs in the mmWave band. Naribole et al. [15] implemented a technique called scalable directional multicast (SDM) to train the AP with per-beam per-client received-signal-strength-indicator measurements via partially traversing a codebook tree. Based on the training information, they proposed a scalable beam grouping algorithm to obtain the minimum multicast group data transmission time. Evaluation results presented in [15] show that the gain provided by the SDM increases with the group size, and it provides a near-optimal group selection with significantly reduced training time. Park et al. [30] proposed an incremental multicast grouping (IMG) scheme where the beamwidths are adaptively assigned via the locations of multicast devices to maximize the sum rate. However, D2D communications were not enabled in this IMG.

It is clear that jointly utilizing multi-level codebook and D2D communications to maximize multicast efficiency for small cells in the mmWave band is challenging and has not been exploited in the open literature. To the best of our knowledge, we are the first to address this problem.

3 System overview

3.1 System model

We consider a mmWave small cell consisting of n nodes, one of which is the AP and the rest are user equipments (UEs). The system time is partitioned into non-overlapping time slots of equal length, and the AP synchronizes the clocks of UEs as well as schedules the medium access of all the nodes to accommodate the multicast demand of users. Equipped with steerable directional antennas, the AP and users generate the directional beams of different beamwidths via a multi-level codebook. We assume that a bootstrapping program is run in the system so that the AP knows the up-to-date network topology and the location information of UEs [31, 32]. For example, the network topology can be obtained by the neighbor discovery schemes of [31], while the location information can be obtained via wireless channel signatures, such as angle of arrival, time difference of arrival, or the received signal strength [32].

With the location information of nodes, the mmWave beam alignment overhead can be significantly reduced [33]. We assume the mmWave small cells are deployed underlying the macrocell to form the heterogeneous cellular network (HCN), and the small-cell APs and UEs are also equipped with omnidirectional antennas for 4G communications. Thus, the location information can also be obtained by the localization techniques in the cellular bands. Meanwhile, the transmission requests and some signaling information for mmWave small cells can be collected by the reliable 4G communication. Under relatively low mobility, the network topology and location information will be updated periodically by about 10 ms.

Because non-line-of-sight transmissions suffer from very high attenuation, mmWave communications in small cell mainly rely on line-of-sight (LOS) transmissions. Therefore, we assume that a LOS path is available for each transmission [34].

Denote the directional link from node i to j by (i, j). In directional beamforming, both nodes i and j point toward each other via a beam from an L-level codebook. Assume that transmit node i adopts the tth beam in the lth level of the codebook, which is denoted as $\varphi(t, l)$, and the antenna gain of $\varphi(t, l)$ in the direction of $i \rightarrow j$ is $G_{ij}^{(\mathrm{T})}(\varphi(t, l))$, while receive node j adopts the sth beam in the hth level of the codebook, which is denoted by $\varphi(s, h)$, and the antenna gain of $\varphi(s, h)$ in the direction of $i \rightarrow j$ is $G_{ij}^{(\mathrm{R})}(\varphi(s, h))$. Then based on the path loss model [34], the received power at node j for link (i, j) is given by

$$P_{ij}^{(\mathrm{R})} = k_0 G_{ij}^{(\mathrm{T})}(\varphi(t, l)) G_{ij}^{(\mathrm{R})}(\varphi(s, h)) d_{ij}^{-\tau} P_t, \qquad (1)$$

where P_t is the transmission power and k_0 is a constant that is proportional to $\left(\frac{\lambda}{4\pi}\right)^2$ with λ being the carrier wavelength, while d_{ij} is the distance between transmitter i and receiver j and τ is the path loss exponent [17]. Hence the received signal to noise ratio (SNR) of link (i, j) is calculated according to

$$\mathrm{SNR}_{ij} = \frac{P_{ij}^{(\mathrm{R})}}{N_0 W} = \frac{k_0 G_{ij}^{(\mathrm{T})}(\varphi(t, l)) G_{ij}^{(\mathrm{R})}(\varphi(s, h)) d_{ij}^{-\tau} P_t}{N_0 W}, \qquad (2)$$

where W [Hz] is the bandwidth and N_0 [mW/Hz] is the one-sided power spectra density of the link's white Gaussian noise [17]. Considering the reduction of multipath effect for directional mmWave links [25], the achievable data rate of link (i, j) can be estimated based on Shannon's channel capacity as

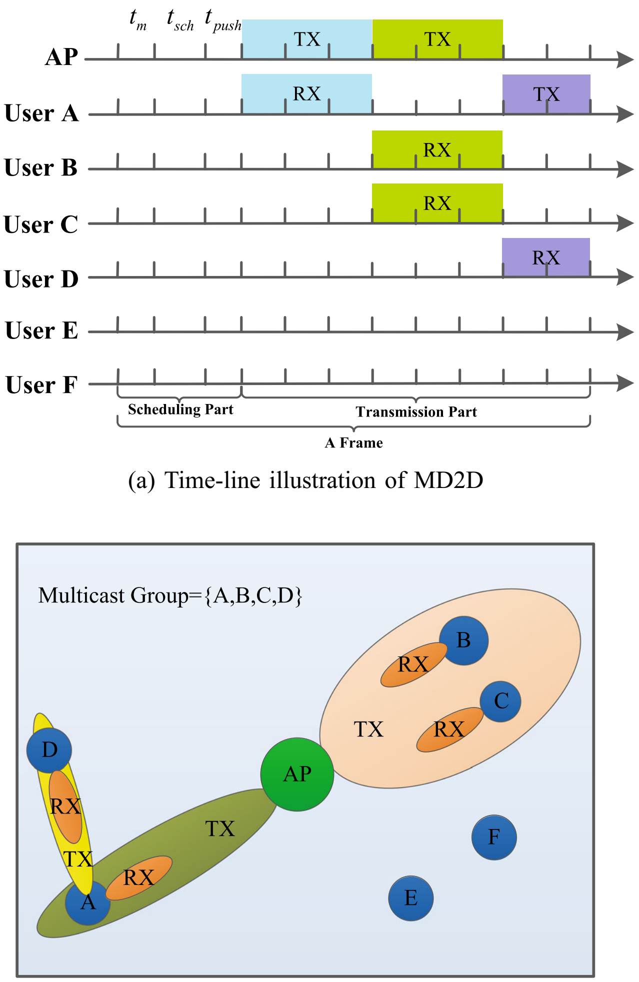
$$R_{ij} = \eta W \log_2\left(1 + \mathrm{SNR}_{ij}\right), \qquad (3)$$

where $\eta \in (0, 1)$ denotes the efficiency of the transceiver [17].

3.2 Problem overview

We consider the multicast service transmitted from the AP to a multicast group. To improve multicast efficiency, appropriate beams should be selected from the multi-level codebook for each multicast transmission. How to select appropriate beams is illustrated in Section 5.1. Besides, users should be able to receive the multicast service from nearby users through D2D communications. To illustrate our MD2D scheme, which exploits both D2D transmissions and multi-level codebook, we use the simple example depicted in Fig. 1, which is a small cell of six users, with the multicast group consisting of users A to D.

Figure 1a depicts the time-line illustration of this example. In the system, time is divided into a sequence of non-overlapping frames [23], and each frame consists of two periods, multicast scheduling period and multicast transmission period. In the scheduling period, the AP first obtains the multicast traffic and the information of the associated multicast group from the network layer, which needs time t_m; next the AP computes a schedule required to complete the multicast service, which needs time t_{sch}; then AP pushes the schedule to the users in the multicast group in sequence, which needs time t_{push}.



(a) Time-line illustration of MD2D

(b) Network topology and MD2D operation

Fig. 1 An example of MD2D operation in a small cell of six users, with the multicast group consisting of users A to D

In the transmission period, all the nodes in the multicast group begin transmissions following the schedule until the multicast traffic is distributed to all the users in the group. As illustrated in Fig. 1a, a multicast transmission period is naturally divided into multiple phases according to the schedule, and in each phase, the multicast transmission occupies several time slots.

Figure 1b illustrates the network topology of this simple example. To serve this multicast group efficiently, the multicast group can be partitioned into three subsets, namely, user A, users B and C, and user D. Since users B and C are located sufficiently close with a very small angle difference, the AP can serve them with a wide beam simultaneously. By contrast, since user D and user A span a large angle, serving them simultaneously with a wide beam will lead to low transmission rate and, consequently, requires many more time slots. Instead, the AP may serve user A with a narrow beam, and then by exploiting the close proximity of users A and D, the AP can let user A to serve user D via the D2D communication.

Thus, the schedule as shown in Fig. 1a is obtained, which consists of three multicast transmission phases in a frame. In the first phase, the AP directs its narrowest beam towards user A to achieve the highest transmission rate, as illustrated in Fig. 1b, which occupies three time slots. In the second phase, AP serves users B and C simultaneously with a wide beam, which occupies three time slots. Finally, user A serves user D with the narrowest beam to achieve highest transmission rate in the third phase. Because users A and D are very close, this D2D based third phase only occupies two time slots. Thus, a total of eight time slots are needed to serve the multicast group. By contrast, if the AP serves the four users sequentially with the narrowest beams, a total of twelve time slots are needed.

From this simple example, we can clearly observe that there are two key problems to solve in order to maximize multicast efficiency. The first one is how to partition the multicast group into subsets and serve each subset via an appropriate beam to achieve high multicast efficiency. The second one is how to effectively utilize D2D communications by exploiting the physical proximity of devices to further improve multicast efficiency as much as possible.

4 Problem formulation

The set of users in the multicast group is denoted by $\mathbb{U}$ and partition $\mathbb{U}$ into S subsets, i.e., $\mathbb{U} = \mathbb{U}_1 \cup \mathbb{U}_2 \cdots \mathbb{U}_S$. Note that the users in each subset receive multicast service simultaneously. We denote the jth user in the ith subset $\mathbb{U}_i$ by u_{ij}. Since $|\mathbb{U}_i|$ is at least 1, $1 \leq S \leq |\mathbb{U}|$. The traffic demand for the multicast group is denoted by D. The schedule for the multicast transmission period of a frame contains K phases, and each phase lasts several consecutive time slots.

For each phase, we define the size $1 \times S$ vector $\mathbf{a}^k$ to indicate whether the subsets of the multicast group are scheduled in the kth phase, where $1 \leq k \leq K$. Specifically, let the ith element of $\mathbf{a}^k$ be denoted by a_i^k, where $1 \leq i \leq S$. If $\mathbb{U}_i$ is scheduled in the kth phase, $a_i^k = 1$; otherwise, $a_i^k = 0$. We denote the transmit node for the subset $\mathbb{U}_i$ by s_i. Since D2D communication is enabled, s_i may be the AP or a user with the multicast data. When s_i is a user with the multicast data, we denote the multicast subset that s_i belongs to by $\mathbb{U}_{f(s_i)}$, i.e., $f(s_i)$ denotes the subset number of s_i. For the user $u_{ij} \in \mathbb{U}_i$, the achievable transmission rate provided by s_i is R_{ij} as given in (3), which is rewritten here

$$R_{ij} = \eta W \log_2\left(1 + \frac{k_0 G^{(\mathrm{T})}_{s_i u_{ij}}(\varphi(t,l)) G^{(\mathrm{R})}_{s_i u_{ij}}(\varphi(s,h)) d^{-\tau}_{s_i u_{ij}} P_t}{N_0 W}\right), \quad (4)$$

where again $\varphi(t,l)$ denotes the transmit beam of s_i and $\varphi(s,h)$ is the receive beam of u_{ij}. We denote the set of beams, i.e., the codebook with L levels, by $\mathbb{C}_L$. Therefore, $\varphi(t,l), \varphi(s,h) \in \mathbb{C}_L$. We further denote the required transmission rate to serve the users in $\mathbb{U}_i$ simultaneously by R_i. We denote the number of time slots scheduled for the kth phase by δ^k, and the time slot duration is Δ.

To maximize the multicast efficiency or throughput, the transmission schedule should accommodate the multicast demand of all the users with a minimum number of time slots. Therefore, the objective function to be minimized is simply

$$\mathcal{J} = \sum_{k=1}^{K} \delta^k. \quad (5)$$

Next we analyze the system constraints of this multicast transmission optimization problem.

First, each multicast group is partitioned into several subsets, which is expressed as the following two constraints

$$\mathbb{U} = \mathbb{U}_1 \cup \mathbb{U}_2 \cup \cdots \cup \mathbb{U}_S, \quad (6)$$

$$1 \leq S \leq |\mathbb{U}|. \quad (7)$$

Second, since a_i^k is a binary variable, we have the constraint

$$a_i^k = \{0, 1\}, \ \forall i \in \{1, 2, \cdots, S\}, \forall k \in \{1, 2, \cdots, K\}. \quad (8)$$

Third, to reduce beamforming overhead and system complexity, we restrict to the case that the multicast transmission for each subset is only scheduled once in one frame of the schedule. Thus, we have

$$\sum_{k=1}^{K} a_i^k = 1, \ \ \forall i \in \{1, 2, \cdots, S\}. \quad (9)$$

Fourth, to serve the users in each subset simultaneously, the required transmission rate must meet the condition

$$R_i = \min_{u_{ij} \in \mathbb{U}_i} R_{ij}, \ \ \forall i \in \{1, 2, \cdots, S\}, \quad (10)$$

where R_{ij} is given by (4) with the constraint

$$\forall \varphi(t,l) \in \mathbb{C}_L,\ \forall \varphi(s,h) \in \mathbb{C}_L. \tag{11}$$

Fifth, the schedule must meet the multicast demand, and therefore we must have

$$\sum_{k=1}^{K} \left(a_i^k \cdot R_i \cdot \delta^k \cdot \Delta \right) \geq D,\ \ \forall i \in \{1, 2, \cdots, S\}. \tag{12}$$

Lastly, to be able to exploit D2D communications, the transmit node of each subset should obtain the multicast data first. Thus, if a subset has a user that transmits multicast data to another subset, then the multicast transmission to this subset should be scheduled prior to the D2D based multicast transmission to the other subset. This constraint can be expressed as

$$\sum_{k=1}^{K^*} a_{f(s_i)}^k \geq \sum_{k=1}^{K^*} a_i^k,\ \ \forall i,\ \ K^* \in \{1, \cdots, K\}. \tag{13}$$

Thus, the optimal multicast scheduling problem (P1) can be expressed as follows

$$\text{(P1)}\ \ \min \mathcal{J} = \sum_{k=1}^{K} \delta^k,$$
$$\text{s.t. Constraints (6) to (13) are met.} \tag{14}$$

Note that constraints (10) and (12) are nonlinear, and there exists set partitioning operation in constraint (6). Thus the problem (P1) is a mixed integer nonlinear programming, where a_i^k are binary variables, δ^k are integer variables and S is an integer variable, while $\varphi(t,l)$ and $\varphi(s,h)$ are discrete variables. This problem is more complex than the NP-complete 0-1 Knapsack problem [17, 35]. The optimal solution can be obtained via the exhaustive search, which has high computational complexity, and cannot be applied in practice.

5 Proposed multicast scheduling scheme

As stated previously, there are two key mechanisms to improve multicast efficiency. First, the potential of multi-level codebook should be unleashed. Wide beams are able to cover larger angle range and may serve more users simultaneously. However, wide beams have low antenna gains, and achievable transmission rates may be low. By contrast, narrow beams have high antenna gains and are able to support high transmission rates. But narrow beams have limited coverage in angle range and may not be able to serve many users simultaneously. Second, the advantages of D2D communications in physical proximity should be reaped. If the users are located sufficiently close, multicast transmission using a wide beam may be more efficient. Clearly, optimizing the network performance based on these two mechanisms is a complex problem, which requires elaborate design for user partition, multicast path planning and beam selection for multicast transmission. To reduce complexity and achieve practical solutions, we propose the heuristic multicast scheduling scheme, MD2D, for the optimization problem (P1). Specifically, we first propose a user-clustering and multicast-path-planning algorithm to partition the multicast group into appropriate subsets and to decide the multicast transmission paths for the multicast group, which is required by constraints (6) and (7). Then, we propose a multicast scheduling algorithm to schedule the multicast transmissions into the transmission period.

5.1 User partition and multicast path planning

We start from the AP to find the nearest user subset. Users that are located very close to each other and span a limited angle range will be put into a subset to be served simultaneously. In this way, we can realize the potential of multi-level codebook, and use wide beams to serve more users. We continue to expand the set of the already allocated subsets including the AP by finding the possible user subset nearby. If such a newly selected subset is very close to one of the allocated subsets, then we enable the D2D communications between these two subsets to serve this new subset more efficiently. Of course, the multicast transmission to this new subset should be scheduled behind the transmission to the allocated subset. In this way, the advantages of D2D communications are exploited to enhance multicast efficiency. In other words, the objective function in problem (P1) can be optimized via D2D communications.

Let us denote the AP in the small cell as $\mathbb{U}_0$ and the subset of users that the algorithm allocates in the tth iteration by $\mathbb{U}_t$. We also denote the set of the subsets that have been allocated and thus are able to serve other unallocated users by $\mathbb{U}_\mathcal{M}$. Since the AP has the multicast data to serve users, it can be regarded as a subset that has been allocated, and thus $\mathbb{U}_\mathcal{M}$ is initialized to $\mathbb{U}_\mathcal{M} = \{\mathbb{U}_0\}$. The allocated multicast transmission path from allocated subset $\mathbb{U}_s$ to the new subset $\mathbb{U}_t$ in the tth iteration is denoted by $\mathbb{P}^t$. For each subset $\mathbb{U}_t \in \mathbb{U}_\mathcal{M}$ with $t > 0$, we define the polar coordinates of user i relative to the center of $\mathbb{U}_t$ by $\left(r_i^t, \theta_i^t\right)$. For each subset $\mathbb{U}_t$, we denote its transmit node by s_t, and the beam selected for s_t to serve $\mathbb{U}_t$ is denoted by $\varphi(b_t, l_t)$. We assume that all the users in $\mathbb{U}_t$ point to s_t with the narrowest or finest receive beams. The radius threshold and angle threshold for user partition are denoted by r_{th} and θ_{th}, respectively.

Algorithm 1 User partition and multicast path planning

```
1  Input: the multicast group $\mathbb{U}$;
2  Initialization: $t$=0; $\mathbb{U}_{\mathcal{M}} = \{\mathbb{U}_0\}$;
3  while $|\mathbb{U}| > 0$ do
4      $t = t + 1$; $\mathbb{U}_t = \emptyset$; $\mathbb{P}^t = \emptyset$;
5      for each $\mathbb{U}_s \in \mathbb{U}_{\mathcal{M}}$ do
6          Find the user $i$ in $\mathbb{U}$ with the minimum $r_i^s$;
7          $r^s = r_i^s$;
8      Find the subset $\mathbb{U}_s \in \mathbb{U}_{\mathcal{M}}$ with the minimum $r^s$;
9      for each user $j \in \mathbb{U}$ do
10         Find the maximum angle difference
           $\theta_{\max} = \max_{i \in \mathbb{U}_t} |\theta_j^s - \theta_i^s|$;
11         if $\mathbb{U}_t = \emptyset$ then
12             $\theta_{\max} = 0$;
13         if $|r_j^s - r^s| \leq r_{th}$ and $\theta_{\max} \leq \theta_{th}$ then
14             $\mathbb{U}_t = \mathbb{U}_t \cup j$;
15             $\mathbb{U} = \mathbb{U} - j$;
16     $\mathbb{U}_{\mathcal{M}} = \mathbb{U}_{\mathcal{M}} \cup \{\mathbb{U}_t\}$; $\mathbb{P}^t = \{\mathbb{U}_s \rightarrow \mathbb{U}_t\}$;
17     for each node $i \in \mathbb{U}_s$ do
18         Obtain $R_t^i$, its maximum achievable rate to
           serve $\mathbb{U}_t$;
19         Obtain $\varphi(b_t^i, l_t^i)$, its corresponding selected
           beam to serve $\mathbb{U}_t$;
20     $s_t = \arg\max_{i \in \mathbb{U}_s} R_t^i$;
21     $\varphi(b_t, l_t) = \varphi(b_t^{s_t}, l_t^{s_t})$;
22 Return $\mathbb{U}_{\mathcal{M}}$, $\mathbb{P}^t$, $s_t$ and $\varphi(b_t, l_t)$ for each $\mathbb{U}_t$.
```

The pseudo-code of this user partition and multicast path planning is listed in Algorithm 1. Starting from line 3, it iteratively partitions the users in $\mathbb{U}$ into subsets and schedules the multicast transmission for each subset until all the users are scheduled. Specifically, in the tth iteration, we first find a user with the shortest distance from a subset $\mathbb{U}_s \in \mathbb{U}_{\mathcal{M}}$ in lines 5–8. Then we allocate the users that are close to this user into the subset $\mathbb{U}_t$ in lines 9–15. We measure the closeness in terms of distance and angle with respect to the reference subset $\mathbb{U}_s$ identified in lines 5–8. The angle that the current users span is denoted by $\theta_{\max}$, as indicated in lines 10–12. As shown in line 13, the users selected should be located not far from the reference radius r^s by a threshold r_{th}, and the angle that the current users in the subset span after the candidate user j is added should be no more than a threshold θ_{th}. If the candidate user j meets these two conditions, it is added into $\mathbb{U}_t$ and also removed from $\mathbb{U}$, as shown in lines 14–15. In line 16, the newly allocated subset $\mathbb{U}_t$ is added to $\mathbb{U}_{\mathcal{M}}$ and the multicast transmission from $\mathbb{U}_s$ to $\mathbb{U}_t$ is recorded by $\mathbb{P}^t$. Lines 17–21 determine the transmit node and select beam for $\mathbb{U}_t$. We first obtain the maximum achievable rate and corresponding beam for each user in $\mathbb{U}_s$ to serve $\mathbb{U}_t$, denoted by R_t^i and $\varphi(b_t^i, l_t^i)$. Then, we select the user in $\mathbb{U}_s$ with the highest maximum achievable rate as the transmit node for $\mathbb{U}_t$, and the corresponding beam is recorded, which is the appropriate beam we referred to before. The algorithm is completed in line 22 by returning $\mathbb{U}_{\mathcal{M}}$, $\mathbb{P}^t$, and the selected transmit node and beam for each subset.

The outer loop of lines 3–21 has $|\mathbb{U}|$ iterations. Each of the three inner loops, lines 5–7, lines 9–15 and lines 17–19, has at most $|\mathbb{U}|$ iterations. Moreover, the operations inside each inner loop impose at most the complexity on the order of $|\mathbb{U}|$. Therefore, the worst-case computational complexity of Algorithm 1 is on the order of $\mathsf{O}(|\mathbb{U}|^3)$. This pseudo-polynomial time solution can be implemented in practice.

5.2 Multicast scheduling algorithm

The proposed multicast scheduling algorithm iteratively allocates the multicast transmission for each subset into each phase until all the subsets are scheduled. We will denote the scheduled multicast transmission in the kth phase by $\mathbb{E}^k$. The pseudo-code of this multicast scheduling is given in Algorithm 2. Note that to meet the requirement of constraint (13), only the user with the multicast data is able to serve other users. Thus, if user i in subset $\mathbb{U}_s$ is the transmit node for subset $\mathbb{U}_t$, the multicast transmission to $\mathbb{U}_s$ must be scheduled before the transmission to $\mathbb{U}_t$. In Algorithm 1, we obtain $\mathbb{U}_t$ after $\mathbb{U}_s$, and this order naturally meets constraint (13). Therefore, we can simply schedule the transmissions to the subsets one by one by following the same order as recorded by Algorithm 1, as indicated in lines 4–6 of Algorithm 2, which select the subset for the kth phase. The transmit node and beam determined by Algorithm 1 for this subset are then used for the multicast transmission to this subset in the kth phase, as indicated in lines 7–8. As shown in line 7, there is only one multicast transmission for each phase, which is required by constraint (9). The scheduling results for all the phases are outputted in line 9.

Algorithm 2 Multicast scheduling

```
1  Input: $\mathbb{U}_{\mathcal{M}}$; $\mathbb{P}^t$, $s_t$ and $\varphi(b_t, l_t)$ for each $\mathbb{U}_t \in \mathbb{U}_{\mathcal{M}}$;
2  Initialization: $k$=0;
3  while $k < |\mathbb{U}_{\mathcal{M}}| - 1$ do
4      $k$=$k$+1;
5      Set $\mathbb{E}^k = \emptyset$;
6      Find the $k$th transmission, $\mathbb{P}^k$;
7      $\mathbb{E}^k = \{s_k \rightarrow \mathbb{U}_k\}$;
8      Set the transmit beam for $s_k$ to $\varphi(b_k, l_k)$;
9  Return $\mathbb{E}^k$ for each phase.
```

The computational complexity of Algorithm 2 is obviously on the order of $\mathrm{O}(|\mathbb{U}_{\mathcal{M}}|)$, where $|\mathbb{U}_{\mathcal{M}}| \le |\mathbb{U}|$, which is negligible compared with the computational complexity of Algorithm 1. Therefore, the computational complexity of our proposed multicast scheduling scheme MD2D, which consists of Algorithm 1 and Algorithm 2, is on the order of $\mathrm{O}(|\mathbb{U}|^3)$.

6 Performance evaluation

This section evaluates the performance of our MD2D scheme. We also investigate the impact of the two thresholds in Algorithm 1, r_{th} and θ_{th}, on the achievable system throughput, energy consumption, and energy efficiency.

6.1 Simulation setup

In an mmWave small cell with $|\mathbb{U}|$ users, the AP is located in the center of a 20 m × 20 m square area and the users are uniformly and randomly distributed in the area. After the bootstrapping program, we assume the network topology and location information of nodes have been collected by the AP, and the information will be updated periodically. During each frame, due to Gbps transmission rate in the mmWave band, the signalling overhead involving D2D path planning and transmission scheduling is small, and does not have a significant impact on system performance [23]. Besides, the difference in overhead for different schemes is small, and thus we mainly consider the transmission part for performance evaluation.

We adopt the directional antenna model from IEEE 802.15.3c with a main lobe of Gaussian form in linear scale and constant level of side lobes [36]. The gain of the directional antenna in dB, denoted by $G(\theta)$, can be expressed as

$$G(\theta) = \begin{cases} G_0 - 3.01 \cdot \left(\frac{2\theta}{\theta_{-3\text{dB}}}\right)^2, & 0° \le \theta \le \theta_{ml}/2, \\ G_{sl}, & \theta_{ml}/2 \le \theta \le 180°, \end{cases} \tag{15}$$

where the angle θ takes the value in [0°, 180°], $\theta_{-3\text{dB}}$ is the angle of the half-power beamwidth, and the main lobe width θ_{ml} is related to $\theta_{-3\text{dB}}$ by $\theta_{ml} = 2.6 \cdot \theta_{-3\text{dB}}$, while G_0 is the maximum antenna gain given by $G_0 = 10\log_{10}\left(\frac{1.6162}{\sin(\theta_{-3\text{dB}}/2)}\right)^2$, and the side lobe gain G_{sl} can be expressed as $G_{sl} = -0.4111 \cdot \ln(\theta_{-3\text{dB}}) - 10.579$. In the simulation, we adopt the four-level codebook, where the half-power beamwidth $\theta_{-3\text{dB}}$ is equal to 15°, 30°, 45° and 60°, respectively. The parameters of the simulated mmWave small cell are summarized in Table 1 [27, 28]. For each experiment, we perform one hundred independent simulations and take the average of the results.

Table 1 Simulation Parameters of the simulated mmWave small cell

Parameter	Symbol	Value
System bandwidth	W	2160 MHz
Background noise	N_0	−134 dBm/MHz
Path loss exponent	τ	2
Number of users	$\lvert\mathbb{U}\rvert$	$5 \sim 30$
Maximum Transmission power	P_t	$30 \sim 40$ dBm
Time slot duration	Δ	18 μs
Efficiency of the transceiver design	η	0.5
Multicast data size	D	$1 \sim 10$ Gb

To evaluate the our MD2D scheme in terms of energy consumption and throughput, we adopt the following three performance metrics.

1) **Network Throughput**: The achieved multicast throughput of all the users in the network, expressed as

$$\mathrm{NT} = \frac{|\mathbb{U}| \cdot D}{\sum_{k=1}^{K} \delta^k \cdot \Delta} \ \text{[b/s]}, \tag{16}$$

2) **Energy Consumption**: The total energy consumption of all the multicast transmissions, given by

$$\mathrm{EC} = \sum_{k=1}^{K} \frac{D}{R_k} \cdot P_t \ \text{[J]}, \tag{17}$$

where R_k is the transmission rate in the kth phase.

3) **Energy Efficiency**: The ratio of the achieved network throughput over the consumed energy, which is expressed as

$$\mathrm{EE} = \frac{\mathrm{NT}}{\mathrm{EC}} = \frac{|\mathbb{U}| \cdot D}{\sum_{k=1}^{K} \delta^k \cdot \Delta} \cdot \frac{1}{\sum_{k=1}^{K} \frac{D}{R_k} \cdot P_t} \ \text{[b/s/J]}. \tag{18}$$

As reviewed in Section 2, there exists no previous work in the existing literature that joint exploits D2D communications and multi-level codebook for improving multicast efficiency in mmWave based networks. In order to demonstrate the advantages of utilizing both D2D communications and multi-level codebook in our MD2D scheme, we compare our scheme with the following three multicast schemes.

1) **FDMAC**: In the FDMAC scheme, the AP sequentially transmits the multicast data to the users one by one using the finest-level beam. This is the baseline scheme which exploits neither multi-level codebook nor D2D communications [23].
2) **MC**: In the multi-level codebook scheme, the multicast group is divided into different subsets, and the AP selects an optimal beam from the multi-level codebook to serve each subset. In this scheme, the multi-level

codebook is exploited but D2D communications are not enabled. Through comparison with the MC scheme, the advantages of using D2D communications in our scheme can be observed.

3) **D2D**: In the D2D multicast scheme, D2D communications are utilized to improve the system performance, similar to the MD2D. However, for this D2D-only scheme, the finest-level beam is always used for each transmission, where only one user is served. Therefore, unlike our MD2D, this scheme does not utilize the multi-level codebook. Through the comparison with the D2D scheme, the advantages of our scheme due to the multi-level codebook are demonstrated.

To evaluate the performance of our scheme under NLOS transmissions, we adopt the NLOS parameters in [34], where the path loss exponent is equal to 3.01, and the shadowing effect is also considered. In the following performance evaluation, the results under NLOS transmissions are also presented.

6.2 Impact of the user partition thresholds

Intuitively, the choice of the radius threshold r_{th} and angle threshold θ_{th} in Algorithm 1 of user partition and multicast path planning will seriously impact the performance of our MD2D scheme. For the system specified in Section 6.1, Fig. 2 plots the network throughput performance achieved by our MD2D with various radius and angle threshold values.

In terms of the impact of angle threshold, too small or too large θ_{th} will degrade the achieved NT performance metric. More specifically, with a small angle threshold of $\theta_{th} = 1°$, only a small number of users located nearby can be allocated into a same subset, and MD2D will choose narrow beams to serve each subset. Consequently, the potential of the multi-level codebook is not fully exploited, and the number of transmissions increases, which degrades the network throughput. On the other hand, with a large angle threshold of $\theta_{th} = 15°$, many more users will be allocated into a same subset. MD2D will serve such a subset via a wide beam, and the transmission rate will decrease. Besides, some users served in this way would be much better served via more efficient D2D communications. For this system, it can be observed that with $\theta_{th} = 10°$, MD2D achieves the best performance.

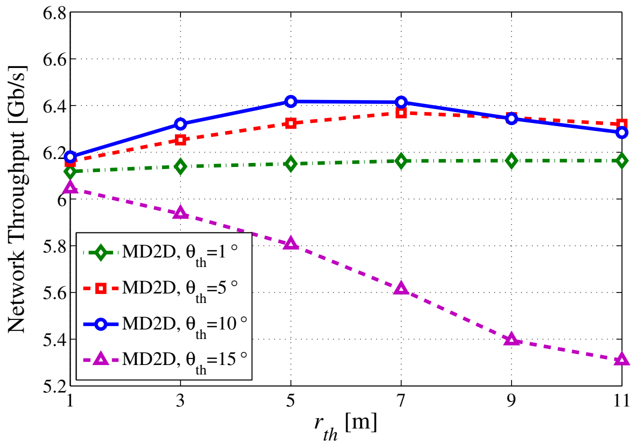


Fig. 2 The network throughput achieved by our proposed MD2D given different radius and angle thresholds, given $\mathbb{U} = 9$, $P_t = 30$ dBm and $D = 1$ Gb

In terms of the impact of radius threshold, given a too small angle threshold of $\theta_{th} = 1°$, the influence of r_{th} to the achieved NT performance metric is very small. By contrast, given a too large angle threshold of $\theta_{th} = 15°$, increasing r_{th} degrades the achieved NT performance metric seriously. With the 'optimal' angle threshold $\theta_{th} = 10°$, MD2D with $r_{th} = 5$ m achieves the best performance and, moreover, for r_{th} between 5 m and 7 m, the network throughputs are all very good. Observe that in the case of $\theta_{th} = 10°$, when r_{th} is small, the network throughput increases with the radius threshold. This is because more users are allocated into a same subset, and wide beams are able to serve these subsets simultaneously, which unleashes the potential of the multi-level codebook. However, when r_{th} is large, increasing r_{th} degrades the performance. This is because similar to the case of too large angle threshold, too many users will be allocated into a same subset which reduces the transmission rate and, moreover, D2D communications could not be exploited fully to improve the network performance. With the angle threshold $\theta_{th} = 5°$, the best choice of radius threshold is $r_{th} = 7$ m.

Besides, in Fig. 3, we plot the energy consumptions achieved by our MD2D with different radius thresholds and angle thresholds. We can obverse that MD2D with angle threshold $\theta_{th} = 10°$ and radius threshold r_{th} between

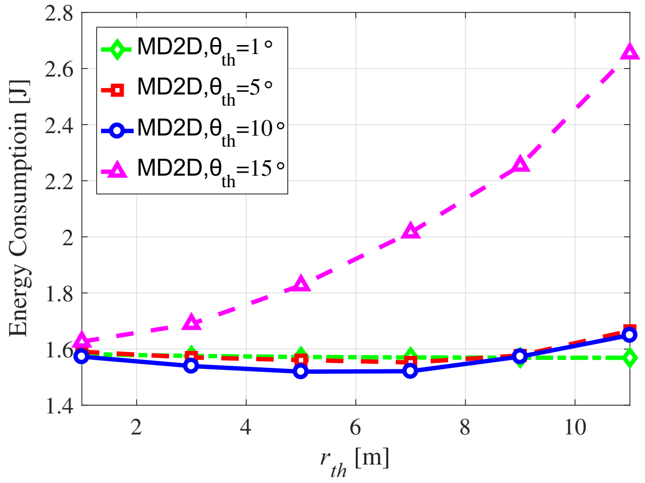


Fig. 3 The energy consumptions achieved by our proposed MD2D given different radius and angle thresholds, given $\mathbb{U} = 9$, $P_t = 30$ dBm and $D = 1$ Gb

5 m and 7 m consumes least energy. For this system, higher throughput means that it can finish task scheduling faster, which causes that the system consumes less energy. According to the above analysis and simulation results, with the 'optimal' angle threshold $\theta_{th} = 10°$, MD2D with radius threshold r_{th} between 5 m and 7 m, achieves the best network throughput performance. Hence, MD2D with the same threshold ought to consume least energy, which agrees with the simulation results.

For this system, it can be seen that the optimal choice of angle threshold and radius threshold is θ_{th} around 10° and r_{th} in the range of 5 m to 7 m. From this experiment, we may conclude that the angle threshold and radius threshold should be optimized according to the network environment, in order to maximize the achieved network throughput performance. Although the optimal threshold may be different under each case, the threshold can be selected to be optimized for the most of the cases. On the other hand, when the performance degrades too much in some cases, the system is assumed to adapt to the changes, and adjust the thresholds in our scheme.

6.3 Comparison with other schemes

We now compare the NT, EC and EE performance of our MD2D scheme with those of the other three schemes, i.e., **FDMAC**, **MC** and **D2D**. The user partition thresholds of our MD2D, r_{th} and θ_{th}, are set to 6 m and 10°, respectively, according to the investigation of the previous section.

6.3.1 Network throughput

Figure 4 compares the NT performance metrics achieved by the four schemes for different numbers of users.

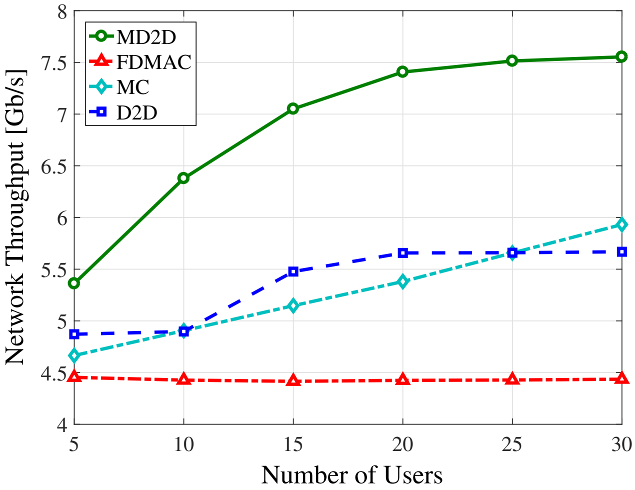


Fig. 4 Comparison of the network throughputs as the functions of user number for four schemes under LOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

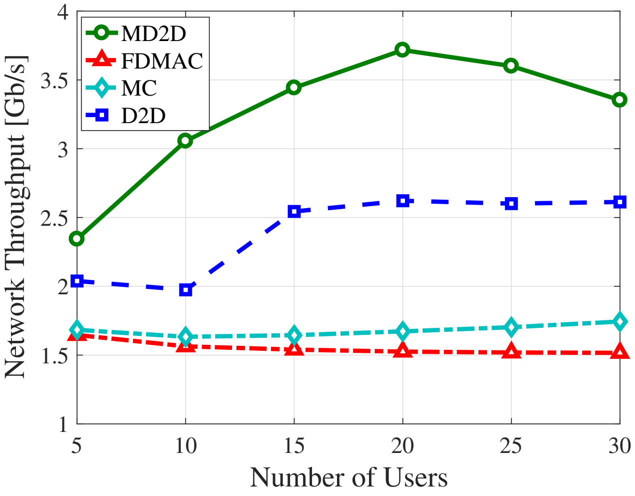


Fig. 5 Comparison of the network throughputs as the functions of user number for four schemes under NLOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

As expected, the FDMAC scheme attains the worst performance and its NT metric remains constant with the increase of $|\mathbb{U}|$. By contrast, the NT metric of the MC scheme increases linearly with $|\mathbb{U}|$. This is because as the number of users increases, the MC scheme can exploit the multi-level codebook more effectively. As for the D2D scheme, its NT metric is relatively low when $|\mathbb{U}|$ is small. But when there are sufficient users in the network, the benefit of D2D communications becomes significant, leading to the considerable increase in the achieved NT performance. However, as $|\mathbb{U}|$ increases further, its NT performance becomes saturated. The results of Fig. 4 also confirm that our MD2D scheme achieves the best performance among the four multicast schemes. For example, given $|\mathbb{U}| = 5$, the MD2D scheme improves the network throughput by about 10% over the second-best D2D scheme, while with $|\mathbb{U}| = 30$, our MD2D scheme outperforms the second-best MC scheme by 27%.

In Fig. 5, the network throughput comparison of our scheme and three other schemes with different number of users under NLOS transmission assumption is presented. The results show that our scheme performs best under NLOS transmission assumption. Compared with results under LOS assumption, MD2D scheme achieves lower throughput since links suffer higher propagation loss under NLOS transmission assumption.

Figure 6 depicts the NT metrics as the functions of P_t achieved by the four schemes. Since the P_t is in dBm in the figure, the relationship between NT and transmission power is still consistent with Shannon law. As expected, the FDMAC scheme attains the worst performance, while our MD2D achieves the best performance. Specifically, the performance gap between our MD2D scheme and the

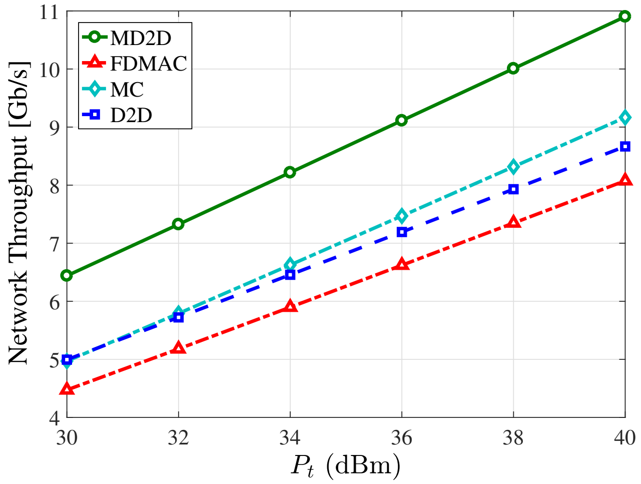


Fig. 6 Comparison of the network throughputs as the functions of transmission power for four schemes under LOS transmission assumption, given $\mathbb{U} = 9$ and $D = 1$ Gb

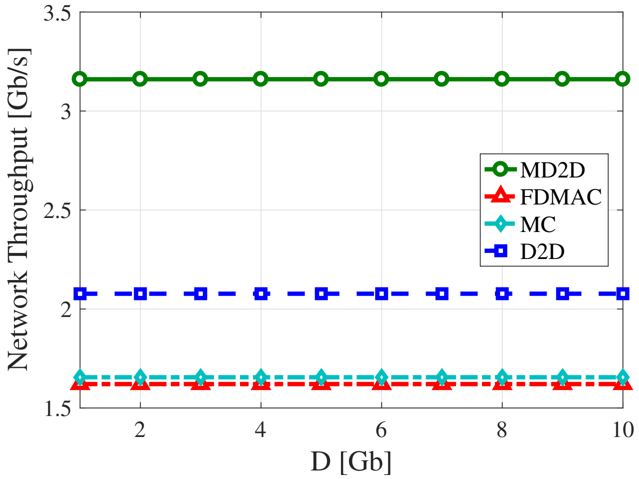


Fig. 8 Comparison of the network throughputs as the functions of multicast data size for four schemes under NLOS transmission assumption, given $\mathbb{U} = 9$ and $P_t = 30$ dBm

second-best MC scheme increases from 1.5 Gb/s to 1.9 Gb/s as P_t increases from 30 dBm to 40 dBm.

Figure 7 compares the NT performance metrics achieved by the four schemes, given different multicast data sizes. Since the number of time slots required scales with the multicast data size, the network throughput remains constant when the multicast data size changes. Again the FDMAC scheme attains the worst performance and our MD2D achieves the best performance. The performance gap between our MD2D scheme and the second-best D2D scheme is 1.4 Gb/s.

In Fig. 8, we plot the network throughput comparison with different multicast data sizes under NLOS transmission assumption. We can observe that our scheme achieves best, and still lower throughput is achieved due to higher propagation loss.

6.3.2 Energy consumption

Figure 9 compares the energy consumptions of the four schemes under different numbers of users. Clearly, the energy consumption increases linearly with the number of users. Observe from both Figs. 9 and 4 that our MD2D scheme consumes the lowest energy consumption and achieves the highest network throughput. This is because higher throughput means that the system can finish the task scheduling faster, which causes that the system consumes less energy. For example, when the number of users is $|\mathbb{U}| = 30$, our MD2D consumes 1 J less energy, while

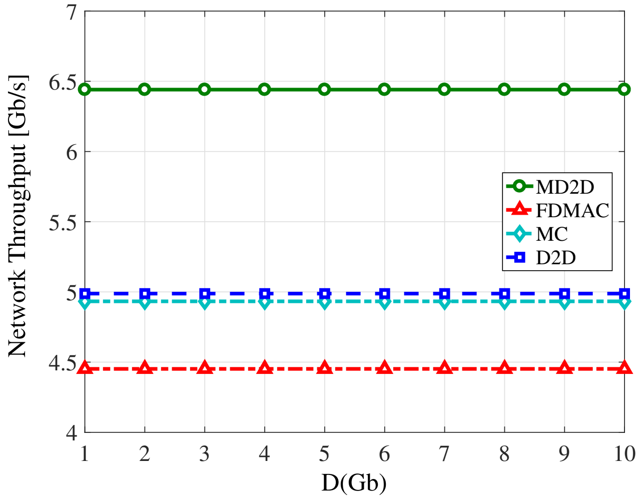


Fig. 7 Comparison of the network throughputs as the functions of multicast data size for four schemes under LOS transmission assumption, given $\mathbb{U} = 9$ and $P_t = 30$ dBm

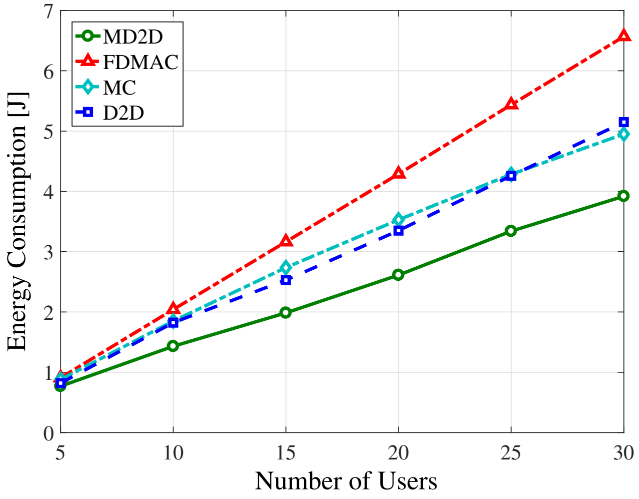


Fig. 9 Comparison of the energy consumptions as the functions of user number for four schemes under LOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

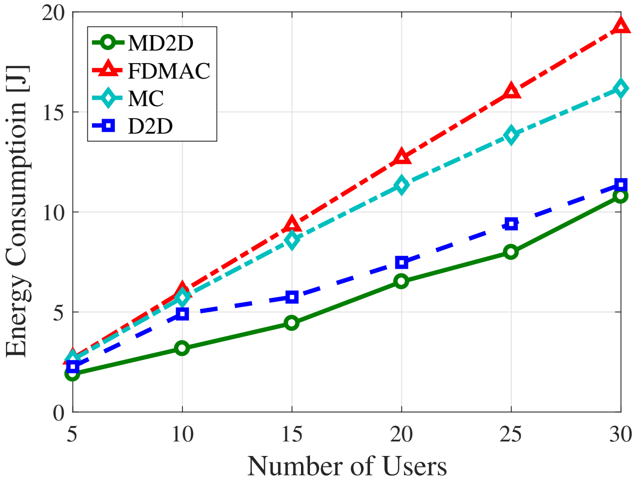


Fig. 10 Comparison of the energy consumptions as the functions of user number for four schemes under NLOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

increasing the network throughput by 1.6 Gb/s, compared with the second-best MC scheme. This clearly demonstrates the significant benefits of jointly exploiting multi-level code book and D2D communications.

Figure 10 plots the energy consumption comparison with different number of users under NLOS transmission assumption. From the results, we can observe that our scheme achieves lower energy consumption compared with other schemes. Compared with the LOS case, the energy consumption is much higher for all schemes since worse channel conditions under NLOS transmission assumption to complete the same multicast task.

Figure 11 plots the energy consumptions of the four schemes given different multicast data sizes. The energy consumption increases with D since the system needs more time to transmit data. Not surprisingly, our MD2D scheme consumes the least energy, while attaining the highest network throughput. Also observe that the gaps of energy consumption between our MD2D and other schemes increase with the multicast data size. In particular, given the multicast data size $D = 10$ Gb, our MD2D consumes 22% less energy than the second-best D2D scheme.

6.3.3 Energy efficiency

The EE metric combines both the network throughput and energy consumption performance, which evaluates how efficient the energy is consumed to achieve the network throughput. In Fig. 12, we plot the EE metrics achieved by the four schemes given different numbers of users. With the increase of users, the traffic load of the network increases and the energy efficiency generally decreases, as can be seen from Fig. 12. Since network throughput affects energy consumption, and network throughput and energy consumption affect energy efficiency, as expected, MD2D achieves the highest energy efficiency performance because MD2D achieves the best NT performance and consumes the least energy among the four multicast schemes. In particular, given the number of users $|\mathbb{U}| = 30$, our MD2D improves the energy efficiency by about 72% compared with the second-best MC scheme.

Figure 13 plots the energy efficiency comparison with different number of users under NLOS transmission assumption. We can observe that our scheme has the highest energy efficiency among all the schemes. Compared with the results under the LOS assumption, the achieved energy

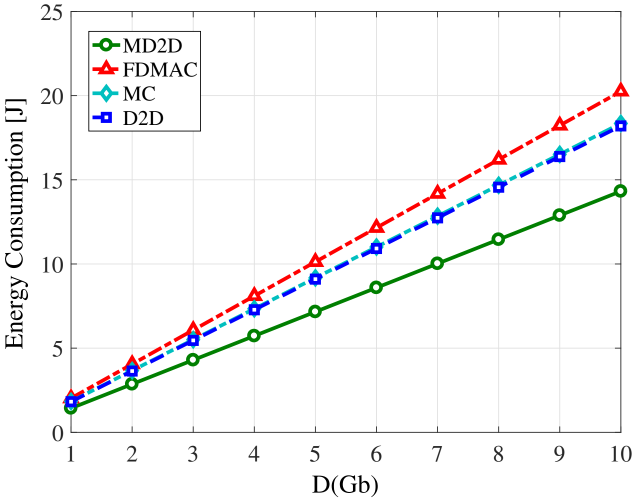


Fig. 11 Comparison of the energy consumptions as the functions of multicast data size for four schemes, given $\mathbb{U} = 9$ and $P_t = 30$ dBm

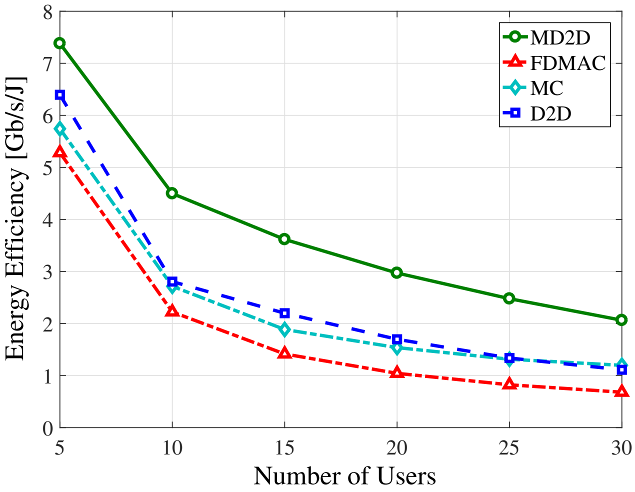


Fig. 12 Comparison of the energy efficiencies as the functions of user number for four schemes under LOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

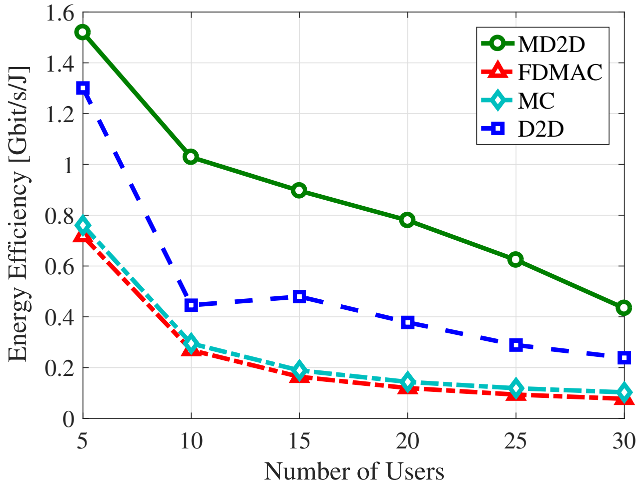


Fig. 13 Comparison of the energy efficiencies as the functions of user number for four schemes under NLOS transmission assumption, given $P_t = 30$ dBm and $D = 1$ Gb

efficiency is lower since lower throughput and higher energy consumption under NLOS transmissions.

Figure 14 compares the EE metrics of the four schemes under different multicast data sizes. Since higher energy consumption is necessary for larger multicast data size, the energy efficiency generally decreases with the multicast data size. Not surprisingly, our MD2D achieves the highest energy efficiency among the four schemes. Compared to the second-best D2D scheme, our MD2D improves the energy efficiency by about 64% and 66%, respectively, given the multicast data sizes $D = 1$ Gb and $D = 10$ Gb.

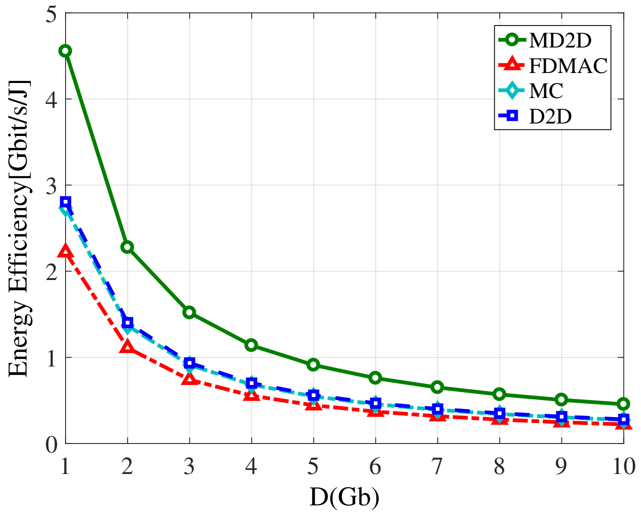


Fig. 14 Comparison of the energy efficiencies as the functions of multicast data size for four schemes, given $\mathbb{U} = 9$ and $P_t = 30$ dBm

7 Conclusions

In this paper, we have proposed an efficient multicast scheduling scheme for mmWave small cells, which jointly exploits both D2D transmissions and multi-level antenna codebook to improve multicast efficiency. Our novel contribution has been twofold. Firstly, we have shown that the optimal multicast scheduling problem by jointly optimizing the utilizations of D2D transmissions and multi-level antenna codebook is NP-hard. Secondly, in order to obtain practical and efficient solution, we have developed a novel multicast scheduling scheme, called MD2D. More specifically, in our MD2D solution, an efficient user-partition and multicast-path-planning algorithm partitions users in the multicast group into subsets and selects the transmit node for each subset. Then an effective multicast scheduling algorithm schedules the transmission for each subset into each transmission phase. Extensive simulation results have verified that our MD2D multicast scheduling scheme significantly outperforms the other two multicast scheduling schemes relying on D2D communications and multi-level antenna codebook alone, respectively, in terms of network throughput and energy efficiency. In the future work, we will analyze the relationship between the objective function and the parameters such as angle and radius thresholds in a theoretically way.

Acknowledgments This study was supported by the National Natural Science Foundation of China Grants 61725101 and 61801016; and by the China Postdoctoral Science Foundation under Grant 2017M610040 and 2018T110041; and by National key research and development program under Grant 2016YFE0200900; and by the Beijing Natural Fund under Grant L172020; and by Major projects of Beijing Municipal Science and Technology Commission under Grant No. Z181100003218010.

References

1. Andrews JG (2011) Can cellular networks handle 1000× the data? Seminar given at University of Texas at Austin
2. Zhu Y, Zhang Z, Marzi Z, Nelson C, Madhow U, Zhao BY, Zheng H (2014) Demystifying 60GHz outdoor picocells. In: Proceedings of MobiCom 2014 (Maui, Hawaii), Sep 7–11, pp 5–16
3. Chandrasekhar V, Andrews JG, Gatherer A (2008) Femtocell networks: a survey. IEEE Commun Mag 46(9):59–67
4. Elkashlan M, Duong TQ, Chen H-H (2015) Millimeter-wave communications for 5G – part 2: applications [guest editorial]. IEEE Commun Mag 53(1):166–167
5. Rappaport TS, Murdock JN, Gutierrez F (2011) State of the art in 60-GHz integrated circuits and systems for wireless communications. Proc IEEE 99(8):1390–1436
6. IEEE 802.15.3c Standard (2009) Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs)—amendment 2: millimeter-wave based alternative physical layer
7. IEEE 802.11ad Standard (2012) Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications –

amendment 3: enhancements for very high throughput in the 60 GHz Band

8. IEEE P802.11-Task Group ay http://www.ieee802.org/11/Reports/tgay_update.htm
9. Wang J, Lan Z, Pyo C-W, Baykas T, Sum C-S, Rahman MA, Gao J, Funada R, Kojima F, Harada H, Kato S (2009) Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems. IEEE J Sel Areas Commun 27(8):1390–1399
10. Xiao Z, He T, Xia P, Xia X-G (May 2016) Hierarchical codebook design for beamforming training in millimeter-wave communication. IEEE Trans Wireless Commun 15(5):3380–3392
11. Xiao Z, Zhu L, Choi J, Chao X, Xia X (2018) Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter-wave communications. IEEE Trans Wirel Commun 17(5):2961–2974
12. Xiao Z, Xia P, Xia X (2016) Enabling UAV cellular with millimeter-wave communication: potentials and approaches. IEEE Commun Mag 54(5):66–73
13. Finamore A, Mellia M, Gilani Z, Papagiannaki K, Erramilli V, Grunenberger Y (2013) Is there a case for mobile phone content pre-staging? In: Proceedings of CoNEXT 2013 (Santa Barbara, CA), Dec 9–12, pp 321–326
14. Niu Y, Su L, Gao C, Li Y, Jin D, Han Z (2016) Exploiting device-to-device communications to enhance spatial reuse for popular content downloading in directional mmwave small cells. IEEE Trans Veh Technol 65(7):5538–5550
15. Naribole S, Knightly E (2016) Scalable multicast in highly-directional 60 GHz WLANs. In: Proceedings of SECON 2016 (London, UK), Jun 27–30, pp 1–9
16. Li Y, Wang Z, Jin D, Chen S (2014) Optimal mobile content downloading in device-to-device communication underlaying cellular networks. IEEE Trans Wireless Commun 13(7):3596–3608
17. Qiao J, Cai LX, Shen X, Mark JW (2012) STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks. In: Proceedings of ICC 2012 (Ottawa, Canada), Jun 10–15, pp 5221–5225
18. Cai LX, Cai L, Shen X, Mark JW (2010) Rex: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna. IEEE Trans Wireless Commun 9(1):113–121
19. Sum C-S, Lan Z, Funada R, Wang J, Baykas T, Rahman M, Harada H (2009) Virtual time-slot allocation scheme for throughput enhancement in a millimeter-wave multi-Gbps WPAN system. IEEE J Sel Areas Commun 27(8):1379–1389
20. Sum C-S, Lan Z, Rahman MA, Wang J, Baykas T, Funada R, Harada H, Kato S (2009) A multi-Gbps millimeter-wave WPAN system based on STDMA with heuristic scheduling. In: Proceedings of GLOBECOM 2009 (Honolulu, Hawaii), Nov. 30–Dec. 4, pp 1–6
21. Qiao J, Cai LX, Shen XS, Mark JW (2011) Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks. IEEE Trans Wireless Commun 10(11):3824–3833
22. Standard ECMA-387 (2010) High Rate 60 GHz PHY, MAC and HDMIPAL
23. Son IK, Mao S, Gong MX, Li Y (2012) On frame-based scheduling for directional mmWave WPANs. In: Proceedings of INFOCOM, 2012 (Orlando, FL), Mar. 25–30, pp 2149–2157
24. Gong MX, Stacey R, Akhmetov D, Mao S (2010) A directional CSMA/CA protocol for mmWave wireless PANs. In: Proceedings of WCNC 2010 (Sydney Australia), Apr. 18–21, pp 1–6
25. Singh S, Ziliotto F, Madhow U, Belding EM, Rodwell M (2009) Blockage and directivity in 60 GHz wireless personal area networks: from cross-layer model to multihop MAC design. IEEE J Sel Areas Commun 27(8):1400–1413
26. Chen Q, Tang J, Wong DTC, Peng X, Zhang Y (2013) Directional cooperative MAC protocol design and performance analysis for IEEE 802.11 ad WLANs. IEEE Trans Veh Technol 62(6):2667–2677
27. Niu Y, Li Y, Jin D, Su L, Wu D (2015) Blockage robust and efficient scheduling for directional mmWave WPANs. IEEE Trans Veh Technol 64(2):728–742
28. Niu Y, Gao C, Li Y, Su L, Jin D, Vasilakos AV (2015) Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells. IEEE J Sel Areas Commun 33(10):2052–2069
29. Zhang H, Huang S, Jiang C, Long K, Leung VCM, Poor HV (2017) Energy efficient user association and power allocation in millimeter wave based ultra dense networks with energy harvesting base stations. IEEE J Sel Areas Commun 35(9):1936–1947
30. Park H, Park S, Song T, Pack S (2013) An incremental multicast grouping scheme for mmWave networks with directional antennas. IEEE Commun Let 17(3):616–619
31. Ning J, Kim T-S, Krishnamurthy SV, Cordeiro C (2011) Directional neighbor discovery in 60 GHz indoor wireless networks. Perform Eval 68(9):897–915
32. Deng H, Sayeed A (2014) Mm-Wave MIMO channel modeling and user localization using sparse beamspace signatures. In: Proceedings SPAWC 2014 (Toronto, Canada), Jun. 22–25, pp 130–134
33. Choi J, Va V, Gonzalez-Prelcic N, Daniels R, Bhat CR, Heath RW (2016) Millimeter-wave vehicular communication to support massive automotive sensing. IEEE Commun Mag 54(12):160–167
34. Geng S, Kivinen J, Zhao X, Vainikainen P (2009) Millimeter-wave propagation channel characterization for short-range wireless communications. IEEE Trans Veh Technol 58(1):3–13
35. Pisinger D (2005) Where are the hard knapsack problems? Comput Oper Res 32(9):2271–2284
36. Chen Q, Peng X, Yang J, Chin F (2012) Spatial reuse strategy in mmWave WPANs with directional antennas. In: Proceedings of GLOBECOM 2012 (Anaheim, CA), Dec. 3–7, pp 5392–5397

Affiliations

Yong Niu[1] · Liren Yu[2] · Yong Li[2] · Zhangdui Zhong[1] · Bo Ai[1] · Sheng Chen[3,4]

Liren Yu
ylr14@mails.tsinghua.edu.cn

Yong Li
liyong07@tsinghua.edu.cn

Zhangdui Zhong
zhdzhong@bjtu.edu.cn

Sheng Chen
sqc@ecs.soton.ac.uk

[1] State Key Laboratory of Rail Traffic Control and Safety, Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

[2] State Key Laboratory on Microwave and Digital Communications, Tsinghua National Laboratory for Information Science and Technology (TNLIST), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

[3] School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

[4] King Abdulaziz University, Jeddah 21589, Saudi Arabia