

Device-to-Device Communications Enabled Energy Efficient Multicast Scheduling in mmWave Small Cells

Yong Niu, *Member, IEEE*, Yu Liu, Yong Li^{ID}, *Senior Member, IEEE*, Xinlei Chen, Zhangdui Zhong, and Zhu Han, *Fellow, IEEE*

Abstract—To keep pace with the rapid growth of mobile traffic demands, dense deployment of small cells in millimeter wave (mmWave) bands has become a promising candidate for next-generation wireless communication systems. With a greatly increased data rate from huge bandwidth of mmWave communications, energy consumption should be mitigated for higher energy efficiency. Due to content popularity, many content-based mobile applications can be supported by the multicast service. mmWave communications exploit directional antennas to overcome high path loss, and concurrent transmissions can be enabled for better multicast service. On the other hand, device-to-device (D2D) communications in physical proximity should be exploited to improve multicast performance. In this paper, we propose an energy-efficient multicast scheduling scheme, referred to as EMS, which utilizes both D2D communications and concurrent transmissions to achieve high energy efficiency. In EMS, a D2D path planning algorithm establishes multi-hop D2D transmission paths, and a concurrent scheduling algorithm allocates the links on the D2D paths into different pairings. Then, the transmission power of links is adjusted by the power control algorithm. Furthermore, we theoretically analyze the roles of D2D communications and concurrent transmissions in reducing energy consumption. Extensive simulations under various system parameters demonstrate the superior performance of EMS in terms of energy consumption compared with the state-of-the-art schemes. Furthermore, we also investigate the choice of the interference threshold to optimize network performance.

Index Terms—Millimeter wave communications, multicast, D2D, concurrent transmissions, energy efficiency.

I. INTRODUCTION

ONE of the critical goals for future wireless communication systems is to mitigate the energy consumption in light of a greatly increased data rate. With abundant spectrum in the millimeter wave (mmWave) band, mmWave communications are able to provide multi-gigabit communication services, and thus become a hot topic [1]. Moreover, research progress on integrated circuits for mmWave communications, including on-chip and in-package antennas, radiofrequency (RF) power amplifiers (PAs), low-noise amplifiers (LNAs), voltage-controlled oscillators (VCOs), mixers, and analog-to-digital converters (ADCs), has paved the way for electronic products in the mmWave band [2]. There are also several standards defined for indoor wireless personal area networks (WPAN) or wireless local area networks (WLAN), such as ECMA 387 [3], IEEE 802.15.3c [4], IEEE 802.11ad [5], and IEEE802.11ay [6]. Densely deploying small cells in the mmWave band underlying the conventional macrocell network has been proposed to improve network capacity, and this deployment has become a promising candidate for future wireless communication systems. With millions more base stations and billions of connected devices in the 5G era, energy efficiency or energy consumption optimization for mmWave communication systems becomes a critical problem to be investigated.

mmWave communications with higher carrier frequencies experience higher path loss than low carrier frequency communications. For example, the free space path loss at 60 GHz band is 28 decibels (dB) more than that at 2.4 GHz [7]. To combat high channel attenuation, the analog beamforming technique is exploited to synthesize directional antennas in a small platform at both the transmitter and receiver for high antenna gain [8]–[10]. Consequently, the omnidirectional carrier sensing is disabled, and there is the deafness problem. However, there is less interference between directional links, and concurrent transmissions (spatial reuse) can be exploited to improve network performance in terms of throughput or energy efficiency.

On the other hand, content popularity is found in mobile networks, which follows the classic Zipf's law [11]. In other words, most of the requests are for a small amount of content.

Manuscript received May 9, 2017; revised September 11, 2017; accepted November 2, 2017. Date of publication November 14, 2017; date of current version March 15, 2018. This study was supported by the Fundamental Research Funds for the Central Universities Grant 2016RC056; and by the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2017ZT009), Beijing Jiaotong University; and by the China Postdoctoral Science Foundation under Grant 2017M610040; and by the National Natural Science Foundation of China Grants 61725101. The research is partially supported by US NSF CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607, ECCS-1547201, CMMI- 1434789, CNS-1443917, and ECCS-1405121. The associate editor coordinating the review of this paper and approving it for publication was L. Chen. (*Corresponding author: Yong Li.*)

Y. Niu and Z. Zhong are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications, the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: niuy11@163.com).

Y. Liu and Y. Li are with the State Key Laboratory on Microwave and Digital Communications, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn).

X. Chen is with the Department of Electronic and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Z. Han is with the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 02447, South Korea (e-mail: zhan2@uh.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCOMM.2017.2773529

Therefore, many content-based applications like TV content streaming, advertising messages broadcasting, and broadcast communication services can be supported by the multicast service, where the base station (BS) provides multiple users (a multicast group) with the same data [12], [13]. At the same time, there will be many user devices located near to each other in the user-intensive region. In this case, device-to-device (D2D) communications in physical proximity can be exploited to save power and improve energy efficiency [14], [15]. Compared with D2D communications sharing the cellular frequency band, there is less interference between D2D communications in the mmWave band and the cellular systems. In multicast services, the BS needs to serve each user one by one serially in the traditional way, and since the transmission links are adjacent, concurrent transmissions cannot be exploited to achieve better network performance. With D2D communications enabled, users with the multicast data are able to forward the multicast data to other users using better D2D channels, and there will be more nonadjacent links, and concurrent transmissions can be enabled to achieve better performance. As in the standard of IEEE 802.11ad [5], the PBSS (Personal Basic Service Set) is a type of IEEE 802.11 LAN ad hoc network in which stations are able to communicate directly with each other. In the PBSS, one STA is required to assume the role of the PBSS central point (PCP). Thus, we assume the same capability for the BS and user stations in this paper.

In this paper, we study the problem of energy efficient multicast scheduling for mmWave small cells, which exploits concurrent transmissions and D2D communications to achieve high energy efficiency via power control. Our contribution is four-fold, and is summarized as follows.

- The optimal multicast scheduling problem with D2D communications and concurrent transmissions considered is formulated into a mixed integer nonlinear program (MINLP), which minimizes the total energy consumption of multicast transmissions by power control with the throughput larger than or equal to that of the serial unicast scheme.
- We propose an energy efficient and practical multicast scheduling scheme, called EMS, to solve the formulated problem. EMS consists of D2D path planning algorithm, concurrent scheduling algorithm, and power control algorithm. The D2D path planning algorithm establishes the multi-hop D2D transmission paths. The concurrent scheduling algorithm concurrently schedules the links on the D2D paths into different pairings, while the power control algorithm adjusts the transmission power of links for lower energy consumption with the achieved network throughput ensured.
- We demonstrate the roles of D2D communications and concurrent transmissions in reducing energy consumption via theoretical analysis.
- Extensive evaluations under various system parameters demonstrate EMS achieves the lowest energy consumption with the throughput ensured compared with other schemes. Based on the results, we also find that D2D communications play a big role in reducing energy

consumption in EMS. Moreover, we study the impact of the interference threshold on network performance.

We organize the rest of this paper as follows. Section II presents the related work on directional MAC protocols for WPANs or WLANs in the mmWave band. Section III introduces the system model, and analyzes the energy efficient multicast scheduling problem by an example. In Section IV, the optimal energy efficient multicast scheduling problem is formulated into a MINLP. Our proposed EMS scheme is presented in Section V. We theoretically analyze the roles of D2D communications and concurrent transmissions in reducing energy consumption in Section VI. Performance evaluation of EMS is conducted in Section VII. Section VIII concludes this paper.

II. RELATED WORK

There has been some related works on directional MAC protocols for WPANs or WLANs in the mmWave band [16]–[20]. Since the standards of ECMA 387 and IEEE 802.15.3c adopt TDMA, some works are based on TDMA. Cai *et al.* [17] derived the ER conditions that concurrent transmissions always outperform TDMA for both omni-antenna and directional-antenna models, and proposed the REX scheduling scheme (REX) to achieve significant spatial reuse gain. There are also two protocols based on IEEE 802.15.3c, which exploit concurrent transmissions to improve performance when the multi-user interference is below a specific threshold. In the scenario of an indoor WPAN, Qiao *et al.* [16] proposed a concurrent transmission scheduling algorithm to maximize the number of flows with the quality of service requirement of each flow satisfied. Furthermore, a multi-hop concurrent transmission scheme is proposed to address the link outage problem and combat huge path loss. For bursty data traffic, TDMA based protocols may allocate not enough medium time for some flows, while overmuch medium time for others [21].

Some centralized scheduling protocols are also proposed for WPANs or WLANs in the mmWave band [21]–[26]. Gong *et al.* [22] proposed a directional CSMA/CA protocol, which exploits virtual carrier sensing to address the deafness problem. Singh *et al.* [23] proposed a multihop relay directional MAC protocol (MRDMAC), which exploits relaying to overcome blockage. The frame based directional MAC protocol (FDMAC) is proposed in [21], where the greedy coloring algorithm exploits concurrent transmissions for high efficiency. In the scenario of an IEEE 802.11ad WLAN, Chen *et al.* [24] proposed a directional cooperative MAC protocol, D-CoopMAC, to coordinate the uplink channel access. Niu *et al.* [25] proposed a blockage robust and efficient directional MAC protocol (BRDMAC) to overcome the blockage problem by two-hop relaying. In the scenario of heterogeneous cellular networks, Niu *et al.* [26] proposed a joint transmission scheduling scheme for the radio access and backhaul of small cells in 60 GHz band (D2DMAC), where a path selection criterion is designed to enable device-to-device transmissions for performance improvement.

In terms of multicast communication, Naribole and Knightly [13] design, implement, and experimentally evaluate scalable directional multicast

(SDM) to train the access point with per-beam per-client RSSI measurements via partially traversing a codebook tree. Based on the training information, a scalable beam grouping algorithm is designed to achieve the minimum multicast group data transmission time. Park *et al.* [27] proposed an incremental multicast grouping scheme, which generates adaptive beamwidths depending on the locations of multicast devices to maximize the sum rate of devices. However, D2D communications are not considered in this scheme. An efficient scheduling scheme for popular content downloading (PCDS) is developed in [12], where users far from the AP obtain the popular content from nearby users via D2D communications. At the same time, concurrent transmissions are also enabled to improve performance.

In terms of energy efficient MAC protocols for wireless networks in the mmWave band, Niu *et al.* [28] proposed an energy efficient scheduling scheme for the mmWave backhauling network, which exploits concurrent transmissions to achieve higher energy efficiency. However, D2D communications are not considered in that scheme.

III. SYSTEM OVERVIEW

A. System Model

We consider an mmWave small cell with one BS and multiple users (UEs). System time is partitioned into time slots of equal length. We assume the clocks of UEs are synchronized by the BS, and the BS schedules the medium access of all UEs to accommodate their traffic demands. In the small cell, directional transmissions are supported via electronically steerable directional antennas at the BS and UEs. Referring to the personal basic service set (PBSS) in the standard of IEEE 802.11ad, we assume uniform configuration for the BS and UEs, and one UE is required to assume the role of the BS [5]. In addition, there is a bootstrapping program run in the system, from which the BS knows the up-to-date network topology and the locations of UEs [29], [30]. The network topology can be obtained by the neighbor discovery schemes in [29] and [31]–[34]. Location information of nodes can be obtained based on wireless channel signatures, such as angle of arrival (AoA), time difference of arrival (TDoA), or the received signal strength (RSS) [30], [35]–[38]. In our system, the bootstrapping program adopts the direct discovery scheme to discover the network topology [29]. In the direct discovery scheme, a node is in the transmitting or receiving state at the beginning of each time slot. In the transmitting state, a node transmits a broadcast packet with its identity in a randomly chosen direction. In the receiving state, a node listens for broadcast packets from a randomly chosen direction. If a collision happens, the node fails to discover any neighbor; otherwise, if the transmitter is unknown, the receiver discovers a new neighbor by recording the angle of arrival (AOA) and the transmitter's identity. After the direct discovery, the nodes report discovered neighbors to the BSs. At the same time, the BSs obtain the location information of UEs by a maximum-likelihood (ML) classifier based on changes in the second-order statistics and sparsity patterns of the beamspace multiple input multiple output (MIMO) channel matrix [30].

On the other hand, for mmWave communication systems, it has been proposed to separate the C-plane (control plane) and the U-plane (user plane), where mmWave communications are used for data transmissions of users, and control signalling information is transmitted using high-quality lower frequency bands to handle mobility [39], [40]. In this case, advanced localization techniques in lower frequency bands can be utilized to obtain the location of nodes [41], [42]. As shown in [41], through spatio-temporal cooperation, high-accuracy location information can be obtained. With the location information of nodes, the mmWave beam alignment overhead can be significantly reduced [43], and we assume after the bootstrapping program, the BS and UEs are able to point their beams towards each other.

In our system, the multicast service is completed in a frame [21], and each frame consists of the scheduling part and the transmission part. Considering the relative low mobility of UEs, the bootstrapping program will be executed periodically, and the network status will be updated periodically. If the network status changes during a frame, there will be failed transmissions. In the next frame, the receivers of failed transmissions will report the failed transmissions to the BS, and the network status will be updated before next frame. Considering the receivers may be lost from the BS due to blockage by obstacles, the receivers may use the low frequency band to report the network status changes to the BS. With a transmission rate at the order of gigabit per second, the information exchange and training between the BS and UEs or two UEs can be completed in a short time, and the transmission part occupies the most of a frame. Thus, in this paper, we focus on the energy consumption in the transmission part. We assume half-duplex nodes for the BS and UEs, and at most one connection can be supported simultaneously for each node.

As shown in [44]–[46], non-line-of-sight (NLOS) transmissions suffer higher propagation loss than line-of-sight (LOS) transmissions for mmWave small cells. Due to the LOS path is strongest, LOS transmissions can also improve the power efficiency. Suffering from a shortage of multipath for NLOS transmissions, we only consider the LOS transmissions in this paper, and the directional LOS link budget is calculated according to the additive white Gaussian noise channel model [23].

The directional link from nodes i to j is denoted by (i, j) . After the beamforming process, nodes i and j point their antennas towards each other. Then received power P_{ij}^r (mW) at node j from node i can be estimated according to the path loss model, which can be expressed as

$$P_{ij}^r = k_0 G_t(i, j) G_r(i, j) l_{ij}^{-\tau} P_t, \quad (1)$$

where we denote the transmission power by P_t (mW), the distance between nodes i and j by l_{ij} (m), the path loss exponent by τ , the transmit antenna gain of node i in the direction of $i \rightarrow j$ by $G_t(i, j)$, the receive antenna gain of node j in the direction of $i \rightarrow j$ by $G_r(i, j)$, and k_0 is a constant coefficient and $k_0 \propto (\frac{\lambda_c}{4\pi})^2$ (λ_c is the wavelength) [16].

On the other hand, there is less interference between directional links. In this case, we can exploit concurrent

transmissions to improve network performance. Adopting the interference model in [16], for links (u, v) and (i, j) , we can obtain the received interference power at node j from node u as

$$I_{uvij} = \rho k_0 G_t(u, j) G_r(u, j) l_{uj}^{-\tau} P_t, \quad (2)$$

where ρ is related to the cross correlation of signals from different links [16]. Then we can obtain the interference power I_{ij} as

$$I_{ij} = \sum_{(u,v) \in \mathbb{C}_{ij}} I_{uvij}, \quad (3)$$

where we denote the set of links that transmit concurrently with link (i, j) by \mathbb{C}_{ij} .

At the same time, adjacent links cannot be scheduled for concurrent transmissions due to the half-duplex assumption. Therefore, links that are adjacent to link (i, j) are not included in \mathbb{C}_{ij} . Then the received signal to interference plus noise ratio (SINR) at receiver j can be expressed as

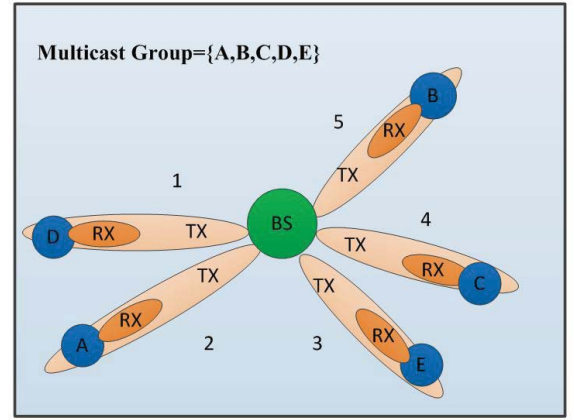
$$\Gamma_{ij} = \frac{k_0 G_t(i, j) G_r(i, j) l_{ij}^{-\tau} P_t}{N_0 W + \rho \sum_{(u,v) \in \mathbb{C}_{ij}} k_0 G_t(u, j) G_r(u, j) l_{uj}^{-\tau} P_t}, \quad (4)$$

where we denote the bandwidth by W (Hz), and the one-sided power spectra density of white Gaussian noise by N_0 (mW/Hz) [16]. Considering the reduction of multipath effect for directional mmWave links [23], link (i, j) is able to achieve a data rate of

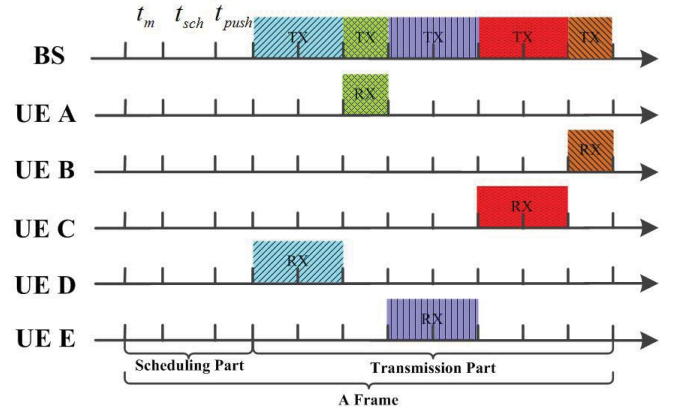
$$R_{ij} = \eta W \log_2 \left(1 + \frac{k_0 G_t(i, j) G_r(i, j) l_{ij}^{-\tau} P_t}{N_0 W + \rho \sum_{(u,v) \in \mathbb{C}_{ij}} k_0 G_t(u, j) G_r(u, j) l_{uj}^{-\tau} P_t} \right), \quad (5)$$

where $\eta \in (0, 1)$ denotes the transceiver design efficiency [16].

In the system, we consider the BS transmits the multicast traffic to a multicast group. EMS exploits both concurrent transmissions and D2D transmissions in close proximity to reduce energy consumption while ensuring the network throughput. We illustrate the operation of EMS in a small cell of five users in Fig. 2 (b). EMS is frame based, and each frame has two parts, scheduling part and transmission part [21]. In the first stage of scheduling part, BS obtains the multicast traffic and corresponding multicast group from upper layers, which takes time t_m ; Then in the second stage, the BS computes a schedule to accommodate the multicast traffic demands, which takes time t_{sch} ; in the third stage, the BS pushes the schedule to the users in the multicast group, which takes time t_{push} . In the second stage of the scheduling part, after establishing the D2D paths, beam training between D2D pairs on the paths is executed to select beams for the transmitters and receivers. In the transmission part, all nodes transmit according to the schedule until the multicast traffic demands are accommodated. We define the period during which a group of concurrent links are activated as one pairing, and multiple pairings may exist in the transmission part. In the



(a) Network Topology and the serial unicast scheme operation



(b) Timeline operation of the serial unicast scheme

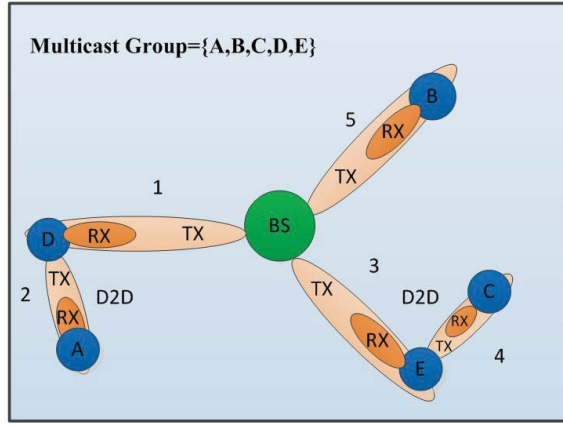
Fig. 1. An example of the serial unicast scheme operation in a small cell of five users.

following section, we discuss the key mechanisms in the problem by this example.

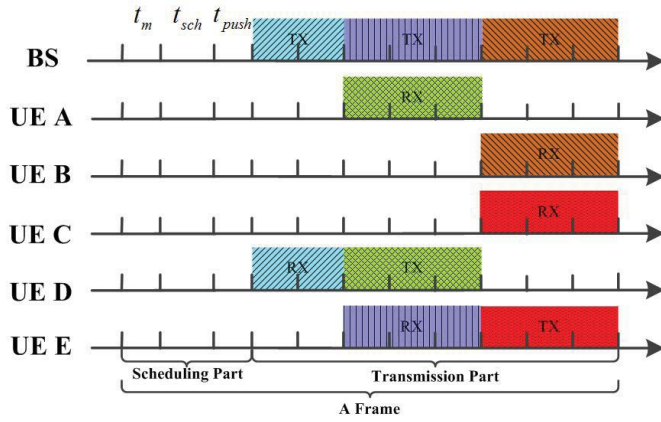
B. Problem Overview

In Fig. 1, we show the multicast service scenario by the serial unicast scheme in a small cell of five users. The multicast group is UEs A, B, C, D, and E. In the serial unicast scheme, the BS serially directs its directional beam towards each user, and the multicast data is transmitted to each user, as shown in Fig. 1 (a). In Fig. 1 (b), we illustrate the timeline operation of the serial unicast scheme.

As comparison, we plot the multicast service by EMS in Fig. 2. In the EMS scheme, we select three transmission paths, $BS \rightarrow B$, $BS \rightarrow D \rightarrow A$, and $BS \rightarrow E \rightarrow C$ as in Fig. 2 (a). Fig. 2 (b) gives the timeline operation of the schedule, which has three pairings in the transmission part. In the first pairing, the BS transmits the multicast data to D. In the second pairing, the BS transmits the multicast data to E, and D transmits to A. In the third pairing, link $BS \rightarrow B$ and link $E \rightarrow C$ are activated to distribute the multicast data to B and C. We can observe that in the second and third pairing, two links transmit concurrently. At the same time, the two-hop



(a) Network Topology and EMS operation



(b) Timeline operation of EMS

Fig. 2. An example of EMS operation in a small cell of five users.

D2D transmission paths, $BS \rightarrow D \rightarrow A$ and $BS \rightarrow E \rightarrow C$, are established.

As we can see, the schedule completes the transmission in the same total time as the serial unicast scheme, which indicates that the achieved throughput of EMS is not less than that by the serial unicast scheme. However, each link in the schedule gets more or equal time slots for transmission than those in the serial unicast scheme. With more time slots for each transmission, lower transmission power can be achieved while completing the transmission of multicast data, and the energy consumption can be reduced accordingly. With the same number of time slots allocated to the multicast service in EMS, each transmission can obtain the most time slots, and the transmission power can be reduced as much as possible. For example, if the transmission time doubles compared with the serial unicast scheme, the transmission power usually can be reduced by more than half. Consequently, the energy consumption can be reduced. From the example, we can observe that there are two key mechanisms to be exploited to reduce energy consumption. The first one is D2D communications, and multi-hop D2D transmission paths should be established. The second one is concurrent transmissions, and interference between concurrent links should be managed appropriately to fully reap the benefits of concurrent transmissions.

IV. PROBLEM FORMULATION AND ANALYSIS

To achieve as high energy efficiency as possible, we minimize the energy consumption of multicast transmission with the throughput ensured. With the advantages of D2D communications and concurrent transmissions fully exploited, the transmit power of each transmission can be adjusted to achieve lower energy consumption. Now, we formulate the problem of optimal multicast scheduling in terms of energy consumption with D2D communications and concurrent transmissions enabled.

A. Problem Formulation

We consider the multicast traffic in a mmWave small cell. There is one multicast group in the network, and we denote the traffic demand for the multicast group by D . \mathbb{U} denotes the set of users in the multicast group. For each user $u \in \mathbb{U}$, we denote the transmit node that serves u by s_u . Since D2D communications are enabled, s_u may be the BS or other users. We assume the schedule for the multicast transmission period has K pairings, and the number of time slots for the k th pairing is denoted by δ^k [21]. The duration of one time slot is denoted by Δ . The binary variable a_u^k is defined to indicate whether the multicast transmission for user u is scheduled in the k th pairing. If it is, a_u^k is equal to 1; otherwise, a_u^k is equal to 0. We denote the transmission power of s_u to u by P_t^u . Then from (5), the achievable transmission rate for user u in the k th pairing can be calculated as

$$R_{s_u u}^k = \eta W \log_2 \left(1 + \frac{a_u^k k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau} P_t^u}{N_0 W + \rho \sum_{(s_v, v) \in \mathbb{C}_{s_u u}} a_v^k k_0 G_t(s_v, u) G_r(s_v, u) l_{s_v u}^{-\tau} P_t^v} \right), \quad (6)$$

where we use (s_v, v) to denote another link scheduled concurrently with link (s_u, u) in the k th pairing.

To achieve high energy efficiency, we try to minimize the energy consumption with the achieved throughput not less than that of the serial unicast scheme. In the serial unicast scheme, the BS transmits the multicast data to each user in the multicast group serially. Thus, the objective function is expressed as

$$\sum_{u=1}^{|\mathbb{U}|} \sum_{k=1}^K P_t^u a_u^k \delta^k \Delta. \quad (7)$$

We can observe that the objective function is defined as the total energy consumed to accommodate all users in the multicast group. In our scheme, the transmitter may be the BS or other users via D2D communications.

The constraints of this energy consumption minimization problem is analyzed as follows. First, to reduce complexity and beamforming overhead, we consider the case where the transmission for each user is scheduled only once in the transmission period, which is expressed as

$$\sum_{k=1}^K a_u^k = 1, \quad \forall u. \quad (8)$$

Second, due to the half-duplex assumption, adjacent links cannot be scheduled concurrently [21], which can be expressed as

$$a_u^k + a_v^k \leq 1, \text{ if links } (s_u, u) \text{ and } (s_v, v) \text{ are adjacent; } (9)$$

Third, the multicast demand of each user should be accommodated by the schedule, which can be expressed as

$$\sum_{k=1}^K R_{s_u u}^k \delta^k \Delta \geq D, \quad \forall u. \quad (10)$$

Fourth, to exploit D2D communications to improve energy efficiency, only the user with the multicast data is able to serve other users. Thus, the multicast transmission for s_u should be scheduled prior to the multicast transmission for user u , i.e.,

$$\sum_{k=1}^{K^*} a_{s_u}^k \geq \sum_{k=1}^{K^*} a_u^k, \quad \forall u, \quad K^* = 1 \sim K. \quad (11)$$

Fifth, the transmission power of s_u for user u should not exceed the maximum allowed transmission power, denoted by P_{max} , which is expressed as follows

$$P_t^u \leq P_{max}, \quad \forall u. \quad (12)$$

Finally, to achieve high energy efficiency with the throughput ensured, we require the achieved throughput should be greater than or equal to that achieved by the serial unicast scheme. The number of time slots needed for the serial unicast scheme to complete the multicast service is denoted by T_s . With the multicast demand fixed as D , we can infer that this throughput requirement is equivalent to the constraint in terms of the occupied number of time slots below. It can be expressed as

$$\sum_{k=1}^K \delta^k \leq T_s. \quad (13)$$

Therefore, the problem of optimal multicast scheduling (P1) can be formulated as

$$\begin{aligned} \text{(P1)} \quad & \min \sum_{u=1}^{|\mathbb{U}|} \sum_{k=1}^K P_t^u a_u^k \delta^k \Delta, \\ & \text{s. t. Constraints (8)–(13).} \end{aligned} \quad (14)$$

In problem P1, we can observe that the objective function (7) and constraint (10) have nonlinear terms, especially $R_{s_u u}^k$ in constraint (10) has a complex form. Considering the binary variable a_u^k , the integer variable δ^k , and the real variable P_t^u , problem (P1) is a mixed integer nonlinear program (MINLP) and is more complex than the NP-complete 0–1 Knapsack problem [47]. Considering the complex form in the objective function (7) and constraint (10), relaxation techniques like the Reformulation-Linearization Technique (RLT) cannot be applied to the nonlinear terms, and it is even difficult to obtain approximate solutions of the original problem [21]. To achieve a practical solution, we propose the energy efficient multicast scheduling scheme for problem P1 in the following.

V. ENERGY EFFICIENT MULTICAST SCHEDULING SCHEME

In this section, we propose the energy efficient multicast scheduling scheme, EMS, for the formulated problem. Both D2D communications and concurrent transmissions are enabled in EMS to improve the energy efficiency. First, we propose a D2D path planning algorithm to establish the multi-hop D2D transmission paths. Then a concurrent scheduling algorithm is proposed to schedule the links on the D2D paths concurrently into each pairing with the interference controlled. Finally, a power control algorithm adjusts the transmission power to realize energy consumption reduction.

A. D2D Path Planning Algorithm

The advantage of D2D communications relies on the better channel conditions between devices in physical proximity. D2D communications between users nearby are preferred due to less propagation loss for saving energy. The D2D path planning algorithm establishes multiple D2D transmission paths from the BS, and by finding the nearest user to the last user on one of the allocated D2D transmission paths, this path is extended by including this new user. If one unallocated user is nearest to the BS, one new path from the BS to this user will be established. The number of hops on each path cannot exceed a predetermined value.

We denote the BS by A , and the set of selected D2D paths by \mathbb{P} . The maximum hop number for each path $p \in \mathbb{P}$ is denoted by H_m . For each path $p \in \mathbb{P}$, the last node on p is denoted by L_p . The set of last nodes of paths in \mathbb{P} is denoted by \mathbb{P}_L . In our algorithm, \mathbb{P}_L represents the set of nodes with the multicast traffic and the ability to serve other users by D2D communications. Since the AP A is the source with the multicast traffic, A is included in \mathbb{P}_L .

The pseudo-code of the D2D path planning algorithm is presented in Algorithm 1. The algorithm iteratively allocates each UE on the D2D transmission paths until all UEs are allocated, as in line 3. The transmission paths in \mathbb{P} extend to unallocated UEs by searching the nearest neighbors of their last nodes, as in lines 4–7. Line 8 obtains node s and its UE with the shortest distance. When node s is the BS, a new path from the BS to the selected UE is generated in \mathbb{P} , as in lines 9–10. When node s is not the BS, the algorithm extends the transmission paths in \mathbb{P} to the nearest unallocated UE from the last node, and removes node s from \mathbb{P}_L , as in lines 11–14. In line 15, the selected UE is removed from the \mathbb{U} . Thus, the selected UE will not be on another D2D path, and will be only scheduled once, which is required by constraint (8). In lines 16–17, the selected UE is added to \mathbb{P}_L if the hop number of the path is less than H_m .

For the example in Fig. 2, the D2D path planning algorithm establishes three transmission paths, $BS \rightarrow D \rightarrow A$, $BS \rightarrow B$, and $BS \rightarrow E \rightarrow C$, for the multicast group of $\{A, B, C, D, E\}$. From Algorithm 1, we can observe that the outer while loop has $|\mathbb{U}|$ iterations, and the inner for loop also has at most $|\mathbb{U}|$ iterations. Thus, the worst case computational complexity of the D2D path planning algorithm is $O(|\mathbb{U}|^2)$, which can be implemented in practice.

Algorithm 1: D2D Path Planning Algorithm

```

1 Input: the multicast group  $\mathbb{U}$ ;
2 Initialization:  $\mathbb{P}_L = \{A\}$ ;  $\mathbb{P} = \emptyset$ ;
3 while  $|\mathbb{U}| > 0$  do
4   for each node  $i \in \mathbb{P}_L$  do
5     Find user  $u$  with the shortest distance to node  $i$ .
6      $r_i = l_{ui}$ .
7      $c_i = u$ .
8   Find node  $s \in \mathbb{P}_L$  with the minimum  $r_s$ ;
9   if  $s == A$  then
10     $\mathbb{P} = \mathbb{P} \cup \{A \rightarrow c_s\}$ ;
11   else
12    Find the path  $p \in \mathbb{P}$  with  $L_p = s$ ;
13    Update  $p$  by extending  $p$  to  $c_s$ ;
14     $\mathbb{P}_L = \mathbb{P}_L - s$ ;
15     $\mathbb{U} = \mathbb{U} - c_s$ ;
16    if  $H(p) < H_m$  then
17       $\mathbb{P}_L = \mathbb{P}_L \cup c_s$ ;
18 Return  $\mathbb{P}$ .
    
```

B. Concurrent Multi-Hop Scheduling

After obtaining the D2D transmission paths by Algorithm 1, the advantage of concurrent transmissions should be further exploited to improve energy efficiency. Thus the concurrent multi-hop scheduling algorithm is proposed to schedule the links on the D2D transmission paths into the transmission period. The algorithm controls the interference via the contention graph, and the maximum independent set (MIS) is utilized to achieve high efficiency.

1) *Contention Graph:* As in [28] and [48], we adopt the contention graph to model the contention relationship among links. Each vertex in the contention graph represents one link in the network. There will be contention (one edge) between two vertices if severe interference between these two links exists. In other words, if the interference between two links is less than a predetermined threshold, we assume no contention between these two links, and concurrent transmissions of these two links are allowed. Instead, if there is severe interference between links, their concurrent transmissions are disabled.

Concretely, we construct the contention graph as follows. To simplify denotation, we denote link (s_u, u) as vertex u . For links (s_u, u) and (s_v, v) , we define the maximum of the interference between them as the weight of the edge between them, i.e.,

$$W_{uv} = \max\{P_{s_u v}^r, P_{s_v u}^r\}. \quad (15)$$

The transmission power here is the maximum transmission power P_{max} . Then the interference threshold σ is defined to control the interference. If $W_{uv}/P_{max} < \sigma$ for links

(s_u, u) and (s_v, v) , no edge exists between these two vertices. Otherwise, there will be contention between these two links. Since concurrent transmissions for adjacent links are disabled, there will always be contention between adjacent links, which is required by constraint (9).

2) *MIS Based Multihop Scheduling Algorithm:* With the contention graph constructed, a maximum independent set (MIS) based multihop scheduling algorithm is proposed to allocate the links on the D2D paths to different pairings. To fully reap the benefits of concurrent transmissions, the MIS based multihop scheduling algorithm schedules as many links into each pairing as possible by obtaining the maximum independent set of the contention graph. With more concurrent transmissions enabled, more time slots can be allocated to each pairing, and lower energy consumption can be achieved by power control. The MIS of the contention graph is a set of isolated vertexes (links) with the maximum cardinality [48]. Since it is NP-complete to obtain the MIS of a general graph, the minimum-degree greedy algorithm is exploited to approximate the maximum independent set. In [49], a performance ratio of $(\Omega + 2)/3$ can be achieved by the minimum-degree greedy algorithm for the graphs of degree bounded by Ω . On the other hand, as required by constraint (11), only the first unscheduled links on the D2D paths can be scheduled in the current pairing.

The contention graph constructed by these links (vertices) in the t th pairing is denoted by $G_t(V_t, E_t)$, where V_t denotes the set of vertices, and E_t denotes the set of edges. Two vertices are referred to as neighbors if one edge exists between them. We denote the set of neighboring vertices for any vertex $v \in V_t$ by $N(v)$. The degree for any vertex $v \in V_t$ is denoted by $d(v)$. We denote the set of links on paths in \mathbb{P} by V . The set of links scheduled in the t th pairing is denoted by V^t , and the set of unvisited and candidate links for the t th pairing is denoted by V_u^t .

As the scheduling process goes on, the algorithm iteratively finds the maximum independent set from $G_t(V_t, E_t)$ for each pairing, and $G_t(V_t, E_t)$ is constructed from the first unscheduled links on the D2D paths from \mathbb{P} in the t th pairing. Then the links in the maximum independent set are scheduled in the same pairing for concurrent transmissions. After the scheduling of each pairing, the contention graph is updated for next pairing after removing the scheduled links from V and \mathbb{P} .

Algorithm 2 presents the pseudo-code of the MIS based multihop scheduling algorithm. The algorithm iteratively schedules the links in V into each pairing, as shown in line 4. In the scheduling for each pairing, the algorithm obtains the first unscheduled links on paths of \mathbb{P} , and construct the contention graph for this pairing, as indicated by line 7. In this way, preceding links on each path will be scheduled prior to links behind, which is required by constraint (11). Then from line 8 to line 12, the minimum-degree greedy algorithm obtains the MIS for each pairing. In lines 13–14, the set of scheduled links V^t is subtracted from V and \mathbb{P} . Since V and \mathbb{P} are updated in lines 13 and 14, the set of first-hop links on the D2D paths is the set of first unscheduled links on the D2D paths, as indicated in line 7.

Algorithm 2: MIS Based Multihop Scheduling

```

1 Input: The set of selected D2D paths,  $\mathbb{P}$ ;
2 The set of links on paths in  $\mathbb{P}$ ,  $V$ ;
3 Initialization:  $t = 0$ ;
4 while  $|V| > 0$  do
5    $t = t + 1$ ;
6   Set  $V^t = \emptyset$ ;
7   Obtain the set of first-hop links on the D2D paths from
    $\mathbb{P}$  and  $V$ ,  $V_t$ ;
8   Set  $V_u^t$  with  $V_u^t = V_t$ ;
9   while  $|V_u^t| > 0$  do
10    Obtain  $v \in V_u^t$  such that  $d(v) = \min_{w \in V_u^t} d(w)$ ;
11     $V^t = V^t \cup v$ ;
12     $V_u^t = V_u^t - \{v \cup N(v)\}$ ;
13   $V = V - V^t$ ;
14  Remove the links in  $V^t$  from  $\mathbb{P}$ ;
15 Return  $V^t$  of each pairing.

```

For the example in Fig. 2, the MIS based multihop scheduling algorithm obtains three pairings. In the first pairing, there is only one link, $BS \rightarrow D$. In the second pairing, there are two links, $BS \rightarrow E$ and $D \rightarrow A$. In the third pairing, there are two links, $BS \rightarrow B$ and $E \rightarrow C$. From Algorithm 2, we can observe that the outer while loop has $|V|$ iterations, and the inner while loop also has at most $|V|$ iterations. Thus, the computational complexity of the MIS based multihop scheduling algorithm is $O(|V|^2)$, which can also be implemented in practice.

C. Power Control Algorithm

With links on the paths scheduled into each pairing by Algorithm 2, a power control algorithm is proposed to adjust the transmission power of links for lower energy consumption, which is also used in [28]. To ensure the throughput achieved by our scheme not less than the serial unicast scheme, we require the number of time slots occupied by our scheme not larger than that by the serial unicast scheme, T_s . With D2D communications and concurrent transmissions enabled by our scheme, the number of time slots allocated to each link by our scheme is larger than that by the serial unicast scheme with the same number of total occupied time slots. With more time slots occupied by each link, lower transmission power can be achieved, and thus the energy consumption can be reduced. For example, if we have twice as much time for transmission, half transmission rate is needed to ensure the throughput. Then from the Shannon's channel capacity, the transmission power P_t is proportional to $(2^{R/W} - 1)$. Under relatively high SINR and low interference, P_t can be reduced by more than half. On the other hand, the better channel conditions provided by D2D links can also help to achieve lower energy consumption. Consequently, the energy efficiency can be improved due to reduced energy consumption and ensured throughput.

In Algorithm 3, we present the pseudo-code of the power control algorithm. For simplicity, we use $u \in V^k$ to denote the link (s_u, u) in the k th pairing. When links' transmission

power is equal to P_{max} , the transmission rate of link (s_u, u) can be obtained as

$$R'_{s_u u} = \eta W \log_2 \left(1 + \frac{k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau} P_{max}}{N_0 W + \rho \sum_{v \in V^k \setminus \{u\}} k_0 G_t(s_v, u) G_r(s_v, u) l_{s_v u}^{-\tau} P_{max}} \right). \quad (16)$$

The number of time slots needed for it to complete the multicast transmission as required by constraint (10), T_u^k , can be calculated as

$$T_u^k = \frac{D}{R'_{s_u u} \Delta}. \quad (17)$$

Thus, we can obtain the maximum needed number of time slots for links in the k th pairing as $T^k = \max_{u \in V^k} \{T_u^k\}$. To reduce the energy consumption as much as possible, we should exploit the time slots available fully while ensuring the throughput as indicated by constraint (13). Thus the T_s time slots is distributed proportionally to each pairing according to T^k . Thus the number of time slots for the k th pairing δ^k can be expressed as

$$\delta^k = \left\lfloor \frac{T^k}{\sum_k T^k} \cdot T_s \right\rfloor, \quad (18)$$

where the floor operation is on pairings before the final pairing, and the remaining time slots are allocated to the final pairing.

Algorithm 3: Power Control Algorithm

```

1 Input: The set of links scheduled in each pairing,  $V^k$ ;
2 The number of pairings,  $K$ ;
3 Initialization:  $k = 0$ ;
4 while  $k < K$  do
5    $k = k + 1$ ;
6   for each link  $u \in V^k$  do
7     Calculate its transmission rate under  $P_{max}$ ,  $R'_{s_u u}$ ;
8     Obtain the number of time slots to complete multi-
     cast transmission by  $R'_{s_u u}$ ,  $T_u^k$ ;
9   Obtain  $T^k = \max_{u \in V^k} \{T_u^k\}$ ;
10   $k = 0$ ;
11 while  $k < K$  do
12   $k = k + 1$ ;
13  Calculate the number of time slots for the  $k$ th pairing,
   $\delta^k$ ;
14  for each link  $u \in V^k$  do
15    Calculate the transmission rate to complete multicast
    transmission,  $R''_{s_u u}$ ;
16    Calculate the transmission power,  $P_t^u$ ;
17 Return  $\delta^k$  of each pairing and  $P_t^u$  of each link.

```

After allocating time slots for each pairing, we can activate the transmission of each link during the whole period of its corresponding pairing to reduce transmission power. The

number of time slots allocated to link (s_u, u) in the serial unicast scheme is denoted by θ_u . Since the threshold σ keeps the interference between concurrent links low, $\delta^k \geq T^k \geq T_u^k \geq \theta_u$ holds generally, and more time slots can be allocated to each link compared with the serial unicast scheme. For link (s_u, u) scheduled in the k th pairing, its needed transmission rate to complete multicast transmission can be expressed as

$$R''_{suu} = \frac{D}{\delta^k \Delta}. \quad (19)$$

With other concurrent links' transmission power equal to P_{max} , the transmission power needed for link (s_u, u) to achieve R''_{suu} can be obtained (20), as shown at the bottom of the next page.

For the example in Fig. 2, the power control algorithm adjusts the transmission power of each link, and allocates time slots for each pairing. For the first pairing, two time slots are allocated. Three time slots are allocated to the second pairing, and three time slots are allocated to the third pairing. In each pairing, each transmission occupies the time slots, and the transmission power is adjusted to reduce energy consumption. Since the algorithm performs power control on each link, it achieves a computational complexity of $O(|V|)$.

VI. PERFORMANCE ANALYSIS

In this section, we demonstrate the roles of D2D communications and concurrent transmissions in reducing energy consumption via theoretical analysis.

For each user u , in the serial unicast scheme, we can obtain the transmission rate from the BS to u as

$$R_{au} = \eta W \log_2 \left(1 + \frac{k_0 G_t(\alpha, u) G_r(\alpha, u) l_{au}^{-\tau} P_{max}}{N_0 W} \right), \quad (21)$$

where we use α to denote the base station. Then we can obtain the energy consumption to provide multicast service for user u as

$$E_u^s = P_{max} \cdot \frac{D}{R_{au}}. \quad (22)$$

In EMS, in contrast, the transmission power to serve user u is (23), as shown at the bottom of the next page.

With (19) incorporated, we can obtain (24), as shown at the bottom of the next page.

Then we can obtain the energy consumption to serve u in EMS as $E_u^d = P_t^u \cdot \delta^k \cdot \Delta$. With (24) incorporated, we can obtain

$$E_u^d = \left(2^{\frac{D}{\delta^k \Delta \eta W}} - 1 \right) \cdot \frac{\left(N_0 W + \rho \sum_{v \in V^k \setminus \{u\}} k_0 G_t(s_v, u) G_r(s_v, u) l_{s_v u}^{-\tau} P_{max} \right)}{k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau}} \cdot \delta^k \cdot \Delta. \quad (25)$$

Since the interference between links is controlled by the interference threshold σ , we can obtain

$$E_u^d < \frac{\left(2^{\frac{D}{\delta^k \Delta \eta W}} - 1 \right) (N_0 W + (|V^k| - 1) \sigma P_{max})}{k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau}} \cdot \delta^k \cdot \Delta. \quad (26)$$

The right side of (26) can be regarded as an upper bound of E_u^d . To minimize the energy consumption, we can try to minimize the upper bound. First, we investigate the role of D2D communications in our scheme. With the antennas between the transmitter and receiver towards each other, the benefits of D2D communications rely on the term $l_{s_u u}^{-\tau}$. Thus, where D2D communications are enabled, we usually have $l_{s_u u} < l_{au}$, and the shorter $l_{s_u u}$ is, the more benefits we can obtain from D2D communications. Therefore, in the D2D path planning algorithm, we establish the D2D paths by searching the nearest neighbors of last nodes on paths.

Then we analyze the role of concurrent transmissions on our scheme. To minimize the energy consumption, we can try to minimize the right side of (26). In the serial unicast scheme, there is only one link in each pairing, and $|V^k|$ is equal to 1. Through concurrent transmissions, the interference increases as shown in (25), but the interference is controlled by the threshold σ . After concurrent transmission scheduling, more time slots can be distributed to each link, and δ^k increases. Thus, we can observe concurrent transmissions increase the number of time slots scheduled for each user with the cost of increased interference between links. With the interference between links controlled by the interference threshold, the energy consumption can be reduced by increasing the number of time slots significantly. To simplify the notation, we define γ as

$$\gamma = \frac{(N_0 W + (|V^k| - 1) \sigma P_{max})}{k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau}} \cdot \Delta. \quad (27)$$

Then we can denote the right side of (26) as E_u , which can be obtained as

$$E_u = \gamma \cdot \left(2^{\frac{D}{\delta^k \Delta \eta W}} - 1 \right) \cdot \delta^k. \quad (28)$$

Taking the derivative of E_u respect to δ^k , we have

$$\frac{dE_u}{d\delta^k} = \gamma \left(2^{\frac{D}{\delta^k \Delta \eta W}} \left(1 - \ln 2 \frac{D}{\delta^k \Delta \eta W} \right) - 1 \right). \quad (29)$$

Denoting $\frac{D}{\delta^k \Delta \eta W}$ by x , we know $2^x (1 - \ln 2 \cdot x) - 1$ is a strictly monotone decreasing function, and since $2^x (1 - \ln 2 \cdot x) - 1$ is equal to 0 when x is equal to 0, we know $\frac{dE_u}{d\delta^k} < 0$ when $\frac{D}{\delta^k \Delta \eta W} > 0$, which always holds in our case. Therefore, we can always reduce the energy consumption by allocating more time slots to each user. After concurrent transmission scheduling, links are grouped into a few number of pairings, and more time slots are allocated to each transmission, which will reduce the energy consumption. To ensure the multicast throughput not less than that achieved in the serial unicast scheme, we require the total number of time slots cannot exceed that in the serial unicast scheme, and thus we distribute the T_s time slots to all pairings. To ensure the energy consumption for each user is reduced, we distribute the time slots to pairings proportionally to their needed number of time slots as indicated by (18).

For each user u , to ensure the energy consumption in EMS is less than that in the serial unicast scheme, we require $E_u^d < E_u^s$. Denoting the $\rho \sum_{v \in V^k \setminus \{u\}} k_0 G_t(s_v, u) G_r(s_v, u) l_{s_v u}^{-\tau} P_{max}$

by I_u to simplify notation, we can obtain the condition for I_u to satisfy $E_u^d < E_u^s$ as follows.

$$I_u < \frac{DP_{max}k_0G_t(s_u, u)G_r(s_u, u)l_{s_uu}^{-\tau}}{R_{au}\left(2^{\frac{D}{\delta^k\Delta\eta W}} - 1\right)\delta^k\Delta}. \quad (30)$$

From the equation, we can observe that for shorter D2D links, the tolerant interference for this link is larger. Besides, since $\left(2^{\frac{D}{\delta^k\Delta\eta W}} - 1\right)\delta^k$ decreases with δ^k , and when allocating more time slots to this link, the tolerant interference is larger.

After power control, more time slots are allocated to each user, and the transmission power is reduced, and the achieved SINR is low. If we approximate $\left(2^{\frac{D}{\delta^k\Delta\eta W}} - 1\right)$ by $\frac{D}{\delta^k\Delta\eta W}\ln 2$, then (25) can be expressed as

$$E_u^d \approx \frac{D\ln 2(N_0W + I_u)}{\eta Wk_0G_t(s_u, u)G_r(s_u, u)l_{s_uu}^{-\tau}}. \quad (31)$$

We can observe that the energy consumption increases linearly with the multicast traffic demand D , Which is consistent with our performance evaluation results in Fig. 5.

VII. PERFORMANCE EVALUATION

In this section, we evaluate the performance of EMS and compare it against other schemes under various system parameters. Besides, the impact of the threshold on the throughput of EMS is also investigated.

A. Simulation Setup

In a typical mmWave small cell, we assume the BS is located in the center of a square area of $20m \times 20m$, where several users are uniformly distributed. In the simulation, we adopt the reference antenna model with side lobe in IEEE 802.15.3c, which consists of a main lobe of the Gaussian form and constant level of side lobes [50]. The antenna gain in decibels (dBS), $G(\theta)$, can be expressed as

$$G(\theta) = \begin{cases} G_0 - 3.01 \cdot \left(\frac{2\theta}{\theta_{-3dB}}\right)^2, & 0^\circ \leq \theta \leq \theta_{ml}/2; \\ G_{sl}, & \theta_{ml}/2 \leq \theta \leq 180^\circ, \end{cases} \quad (32)$$

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Value
Maximum transmission power	P_{max}	30 dBm
Bandwidth	W	2160 MHz
Noise power spectra density	N_0	-134dBm/MHz
Path loss exponent	τ	2
Time slot duration	Δ	18 μ s
MUI factor	ρ	1
Half-power beamwidth	θ_{-3dB}	15°
Efficiency of the transceiver design	η	0.5
Interference Threshold	σ	10^{-12}
Maximum number of hops	H_m	6
Multicast data size	D	10^9 bit
Multicast group size	$ U $	15

where θ is an arbitrary angle within the range $[0^\circ, 180^\circ]$. θ_{-3dB} is the angle of the half-power beamwidth, and θ_{ml} is the main lobe width in units of degrees. θ_{ml} is related to θ_{-3dB} via $\theta_{ml} = 2.6 \cdot \theta_{-3dB}$. The maximum antenna gain G_0 can be obtained from θ_{-3dB} as $G_0 = 10 \log \left(\frac{1.6162}{\sin(\theta_{-3dB}/2)}\right)^2$, while the side lobe gain G_{sl} can be obtained by $G_{sl} = -0.4111 \cdot \ln(\theta_{-3dB}) - 10.579$. Of course, there are differences between theoretical directional antenna model and practical phased-array beam patterns, and strong side lobes of consumer-grade antennas may weaken the advantages of concurrent transmissions in our scheme [51].

We summarize the simulation parameters in Table I. For every result, we perform 50 independent experiments, and the mean of the results are plotted in the figures.

In the evaluation, our scheme EMS is compared with the following two multicast schemes:

- 1) *FDMAC*: the frame-based scheduling directional MAC protocol [21]. In FDMAC, since D2D communications are not considered, and links from the BS to users are adjacent, FDMAC will be reduced to the serial unicast scheme. From the comparison with FDMAC, we can observe the advantages of concurrent transmissions and D2D communications in our proposed scheme.
- 2) *D2D*: the D2D multicast scheme, where D2D communications are exploited as EMS to improve system

$$P_t^u = \frac{\left(2^{\frac{R_{s_uu}''}{\eta W}} - 1\right) \left(N_0W + \rho \sum_{v \in V^k \setminus \{u\}} k_0G_t(s_v, u)G_r(s_v, u)l_{s_vu}^{-\tau} P_{max}\right)}{k_0G_t(s_u, u)G_r(s_u, u)l_{s_uu}^{-\tau}}. \quad (20)$$

$$P_t^u = \frac{\left(2^{\frac{R_{s_uu}''}{\eta W}} - 1\right) \left(N_0W + \rho \sum_{v \in V^k \setminus \{u\}} k_0G_t(s_v, u)G_r(s_v, u)l_{s_vu}^{-\tau} P_{max}\right)}{k_0G_t(s_u, u)G_r(s_u, u)l_{s_uu}^{-\tau}}. \quad (23)$$

$$P_t^u = \frac{\left(2^{\frac{D}{\delta^k\Delta\eta W}} - 1\right) \left(N_0W + \rho \sum_{v \in V^k \setminus \{u\}} k_0G_t(s_v, u)G_r(s_v, u)l_{s_vu}^{-\tau} P_{max}\right)}{k_0G_t(s_u, u)G_r(s_u, u)l_{s_uu}^{-\tau}}. \quad (24)$$

performance. In the D2D scheme, the D2D paths are selected the same as EMS, but the concurrent transmissions are not enabled in the D2D scheme. After selecting the D2D paths, the links on the D2D paths are scheduled into each pairing in sequence (the inherent transmission order of links on each D2D path should be ensured), and there is one link in each pairing. Then the transmission power of each link is adjusted in the same way as EMS with the only difference of only one link in each pairing. From the comparison with the D2D multicast scheme, we can observe the role concurrent transmissions play in our scheme.

In the evaluation, we consider three performance metrics as follows.

- 1) *Energy Consumption*: Total energy consumption of multicast transmissions in the network, which is denoted by EC , and can be expressed as

$$EC = \sum_{k=1}^K \sum_{u \in V^k} P_t^u \delta^k \Delta. \quad (33)$$

- 2) *Energy Ratio*: Energy consumption of EMS divided by that of the D2D Scheme. We denote energy ratio by ER , which can be expressed as

$$ER = \frac{EC_{EMS}}{EC_{D2D}}, \quad (34)$$

where we denote the energy consumption of EMS and D2D scheme by EC_{EMS} and EC_{D2D} , respectively.

- 3) *D2D Ratio*: Energy consumption using D2D communications divided by the total energy consumption for multicast service.

In this paper, we define the energy efficiency as the achieved throughput divided by the consumed energy. With the throughput ensured, the energy efficiency is mainly determined by the energy consumption since in our scheme the time slots consumed by the serial unicast scheme is distributed to the multicast service for users as indicated by (18).

B. Time and Energy Overhead in Beam Training

For the beam training in the scheduling part of EMS, we assume the beam training is done with the help of location information from location techniques [42], [43]. Compared with the serial unicast scheme, EMS and the D2D scheme need to perform beam training between D2D pairs after the D2D paths are established. When the beam training for D2D pairs on the path is completed, the BS computes a schedule by the MIS based multihop scheduling algorithm and the power control algorithm, and then pushes the schedule to users in the multicast group.

With the location information of nodes, the average number of training beam pairs can be significantly reduced, and we adopt 10 candidate beam pairs in the simulation [43]. For each D2D pair, the BS first transmits a small control packet to inform the transmitter of the candidate beam pairs and the corresponding receiver. Then the BS also transmits a small control packet to inform the receiver of the candidate beam pairs and the corresponding transmitter. Afterwards, the transmitter and

TABLE II
BEAM TRAINING PARAMETERS

Parameter	Symbol	Value
Transmission rate in beam training	R	2 Gbps
Propagation delay	δ_p	50ns
PHY overhead	T_{PHY}	250ns
Short MAC frame Tx time	T_{ShFr}	$T_{PHY} + 14 * 8 / R + \delta_p$
SIFS interval	T_{SIFS}	100ns
ACK Tx time	T_{ACK}	T_{ShFr}

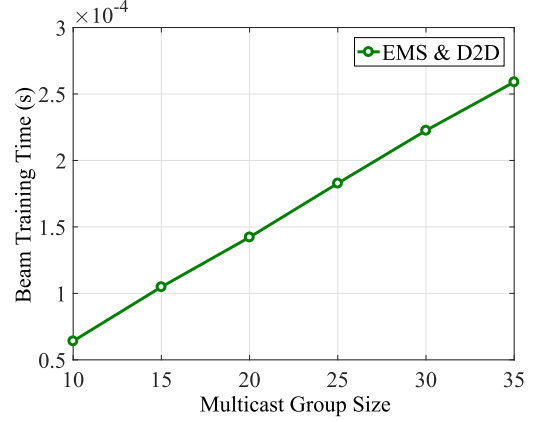


Fig. 3. The time overhead for beam training between D2D pairs in EMS and the D2D scheme under different multicast group sizes.

receiver transmit one small control packet using each training beam pair, and the receiver records the received SNR and transmits an acknowledgement packet with the recorded SNR to the transmitter. After all the candidate training beam pairs are used to perform beam training, the beam pair with the highest received SNR is adopted for the transmission between the transmitter and the receiver. We adopt the simulation parameters in [23], and beam training using one beam pair can be completed within $T_{ShFr} + T_{SIFS} + T_{ACK}$. The transmission power in beam training is 30 dBm. Detailed parameters for beam training is listed in Table II. For EMS and the D2D scheme, additional beam training between D2D pairs is needed to execute D2D communications in the transmission part.

In Fig. 3, we plot the additional time overhead for beam training between D2D pairs in EMS and the D2D scheme under different multicast group sizes. With the increase of multicast group size, the time overhead increases due to more D2D pairs on the D2D paths. When the multicast group size is 35, the time overhead for beam training of D2D pairs is about 2.59×10^{-4} s, while the duration of the transmission part is about 7.94 s. We can also observe that the time overhead is significantly smaller than the duration of the transmission part. Therefore, the additional time overhead for beam training of D2D pairs has a marginal impact on the overall throughput.

We also plot the additional energy consumption for beam training between D2D pairs in EMS and the D2D scheme under different multicast group sizes in Fig. 4. The energy consumption also increases with the multicast group size due to more D2D pairs on D2D paths. When the multicast group size is 35, the energy consumption is about 2×10^{-4} J, and the energy consumption in the transmission part is about 1.2243 J,

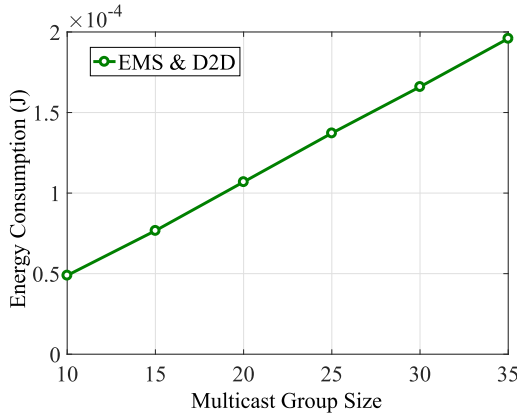


Fig. 4. The energy consumption for beam training between D2D pairs in EMS and the D2D scheme under different multicast group sizes.

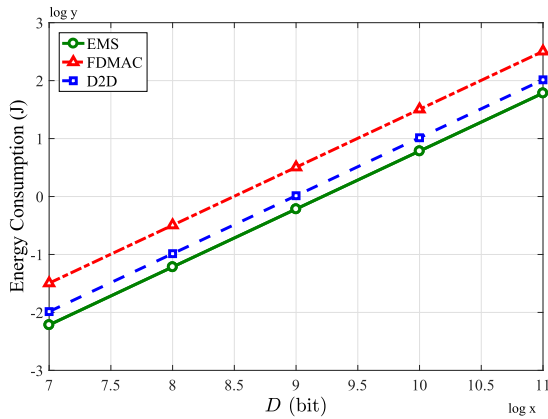


Fig. 5. The energy consumption comparison of three schemes under different multicast data sizes.

which is shown in Fig. 7. Thus, there is a minor increase in the energy consumption due to beam training of D2D pairs. Therefore, combining the results in Fig. 3, beam training of D2D pairs in EMS and the D2D scheme has a marginal impact on the overall throughput, energy consumption, and the energy efficiency. In the following, we focus on the energy consumption, throughput, and energy efficiency in the transmission part.

C. Energy Consumption Comparison in the Transmission Part

The energy consumption of three schemes under different multicast data sizes is plotted in Fig. 5. To show the gap between different schemes more clearly, we show the results with logarithmic coordinates. Other parameters are the same as Table I except the multicast data size. We can observe that EMS achieves the lowest energy consumption, and the gap between the D2D scheme and FDMAc shows the role of D2D communications in reducing the energy consumption. Considering the three schemes achieve the same throughput in the transmission part since the duration of the transmission part is the same, the energy efficiency of EMS is the highest due to the lowest energy consumption. At the same time, the gap between the D2D scheme and EMS shows the role of concurrent transmissions in reducing the energy consumption.

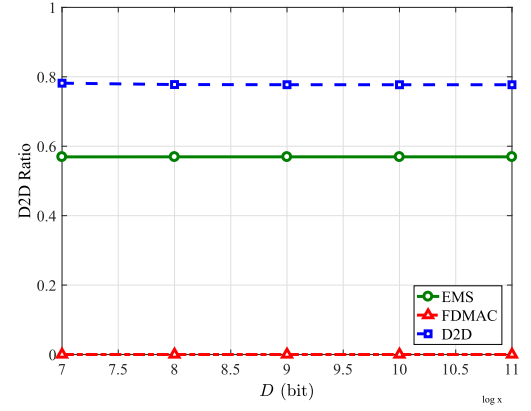


Fig. 6. The D2D ratio comparison of three schemes under different multicast data sizes.

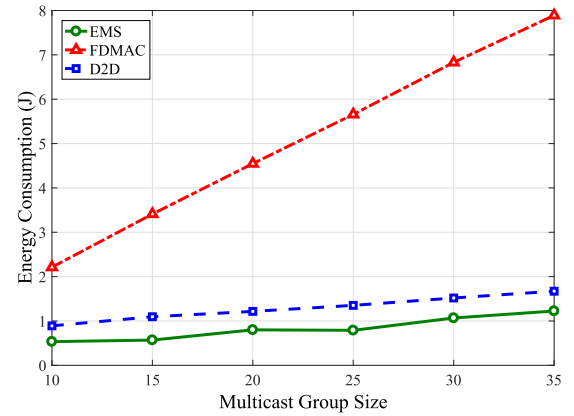


Fig. 7. The energy consumption comparison under different multicast group sizes.

We can also observe that the energy consumption increases with the increase of multicast data size. When the multicast data size is larger, more transmission time is needed to complete the multicast task, and thus the energy consumption increases. Compared with the D2D scheme, EMS reduces the energy consumption by about 41.1% when the multicast data size is 10^{11} bit. When the multicast data size is 10^{11} bit, EMS reduces the energy consumption by about 81% compared with FDMAc.

The D2D ratio comparison of three schemes under different multicast data sizes is plotted in Fig. 6. From the results, we can observe that the D2D scheme achieves the highest D2D ratio among the schemes. Since D2D communications are not enabled in FDMAc, its D2D ratio is 0. In contrast, our proposed EMS achieves a relatively low D2D ratio. In EMS, most multicast transmissions are via D2D communications, and due to the concurrent transmissions, energy consumption using D2D communications is reduced much lower than the D2D scheme without concurrent transmissions. Thus, EMS achieves lower D2D ratio than the D2D scheme. For each transmission, the energy consumption is proportional to the multicast data size, and thus the D2D ratio remains constant with the multicast data size.

The energy consumption comparison of three schemes under different multicast group sizes is plotted in Fig. 7. With the increase of the multicast group size, the energy consumption

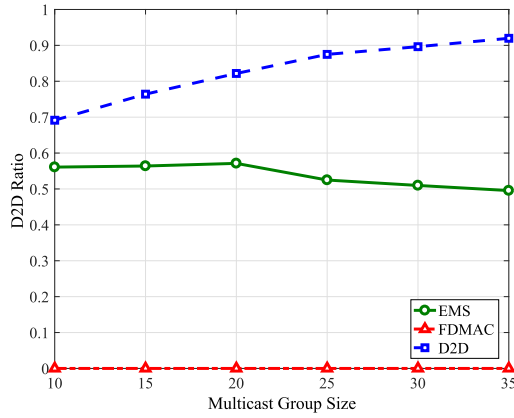


Fig. 8. The D2D ratio comparison under different multicast group sizes.

of three schemes increases due to more users should be served. However, the energy consumption of EMS and the D2D scheme increases with the multicast group size slowly, which demonstrates the advantages of D2D communications and concurrent transmissions in our scheme. As the multicast group size increases, the density of devices increases, and more interference can be observed. Since the contention graph is constructed based on the interference between links, fewer concurrent transmissions are allowed. On the other hand, with the increase of density of devices, there are better channels between devices, and the advantages of D2D communications become larger. Therefore, concurrent transmissions and D2D communications affect the energy consumption in an opposite way, and the energy consumption has a slow rising tendency with the increase of the multicast group size, which indicates a bigger role played by fewer concurrent transmissions and more users. As we can observe, the D2D scheme already improves the performance to a large extent, and when the multicast group size is 35, the D2D scheme reduces the energy consumption by about 78.8% compared with FDMAC, which demonstrates obvious advantage of exploiting D2D communication in improving energy efficiency. Compared with the D2D scheme, EMS further reduces the energy consumption by about 27% due to the concurrent transmission mechanism in EMS. From the results, we can observe that D2D communications play a bigger role in reducing energy consumption than concurrent transmissions for EMS.

The D2D ratio comparison under different multicast group sizes is plotted in Fig. 8. Consistent with the results in Fig. 6, EMS achieves a relatively low value due to the enabled concurrent transmissions. The tendency for EMS and the D2D scheme is different. With more users to serve, there are more multicast services are via D2D communications, and thus the D2D ratio for the D2D scheme increases. For EMS, more users lead to more concurrent transmissions, and the energy consumption using D2D communications can be reduced to a larger extent. Thus, the achieved D2D ratio decreases slowly with the multicast group size.

In Fig. 9, we plot the energy consumption comparison under different maximum transmission power. The results are shown with Y-axis using the logarithmic coordinates. With

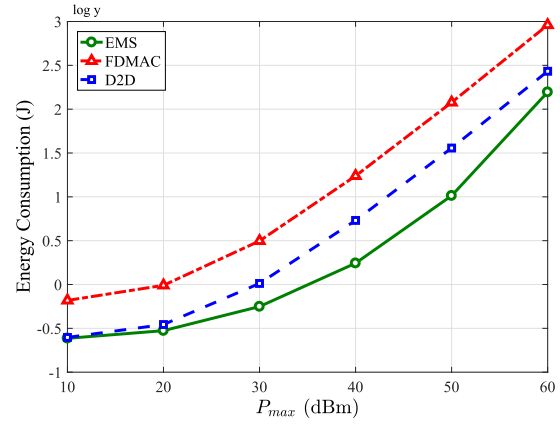


Fig. 9. The energy consumption comparison under different maximum transmission power.

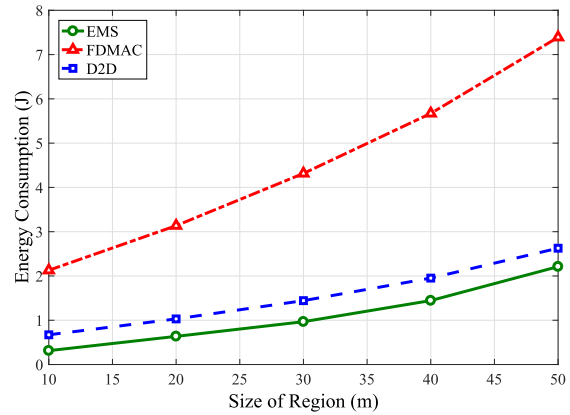


Fig. 10. The energy consumption comparison under different region sizes.

the increase of the maximum transmission power, the energy consumption of three schemes increases since the number of time slots occupied by FDMAC, T_s , decreases due to higher transmission rates. With lower T_s , fewer time slots can be allocated to each transmission in EMS and the D2D scheme. Thus, much reduction in the transmission power cannot be achieved, and the power control mechanism in EMS and the D2D scheme cannot play a big role in reducing the energy consumption. Therefore, the energy consumption increases with the maximum transmission power. Although lower maximum transmission power reduces the energy consumption, it also leads to lower network throughput. Thus, the maximum transmission power should be selected according to practical throughput and energy consumption requirements. As we can observe, EMS still has the lowest energy consumption, and the highest energy efficiency.

In Fig. 10, we plot the energy consumption comparison under different region sizes, where X-axis represents the side length of the square region in the unit of meter (m). EMS achieves the lowest energy consumption among the three schemes under different region sizes. With larger region size, the energy consumption increases since users are distributed more dispersedly, and the link length also increases, which increases the propagation loss. With higher propagation loss,

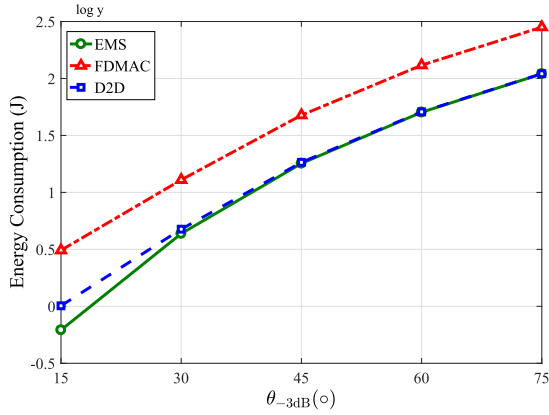


Fig. 11. The energy consumption comparison under different θ_{-3dB} .

more energy is needed to complete the multicast service. When users are distributed more dispersedly, less interference exists between links, and more concurrent transmissions can be enabled. However, more energy consumption from larger link length plays a dominated role considering the high propagation loss at mmWave bands, and the energy consumption increases with the size of region. When the size of region is 50m, EMS reduces the energy consumption by about 70.1% compared with FDMAC, and by about 16% compared with the D2D scheme. From the results, we also can observe that the advantage of EMS compared with FDMAC is mainly because of the D2D communication mechanism since the big performance improvement achieved by the D2D scheme.

The energy consumption comparison under different θ_{-3dB} is plotted in Fig. 11. We examine five cases, with θ_{-3dB} equal to 15°, 30°, 45°, 60°, and 75°. The results are shown with Y-axis using the logarithmic coordinates. We can observe that the energy consumption increases with the θ_{-3dB} . With larger θ_{-3dB} , lower antenna gain can be achieved to compensate the propagation loss as indicated by the antenna model, and more energy is needed to complete the multicast service. As before, EMS achieves the lowest energy consumption. The gap between EMS and the D2D scheme is larger when θ_{-3dB} is smaller. Narrow antenna beams have higher directivity and lead to less interference between links, which is beneficial for concurrent transmissions in EMS. We also observe the big role of D2D communications in reducing energy consumption.

Fig. 12 presents the energy consumption comparison under different maximum numbers of hops, H_m . With the increase of H_m , the energy consumption of EMS and D2D decreases. With a larger maximum number of hops, more users can be served via D2D communications, and more energy can be saved from better channel conditions provided by D2D communications. We also observe that when H_m is 1, three schemes achieve the same performance since both EMS and D2D reduce to the unicast scheme. When H_m is 1, the concurrent transmission mechanism in EMS is also disabled due to the half-duplex constraint. When H_m increases from 1 to 2, we observe a big decrease of energy consumption for EMS and D2D. When H_m increases to 6, the energy consumption decreases slowly. Since more hops may lead to higher overhead in establishing D2D paths, we select H_m to

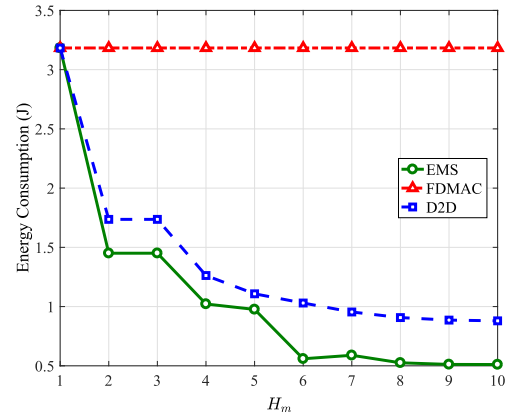


Fig. 12. The energy consumption comparison under different maximum number of hops.

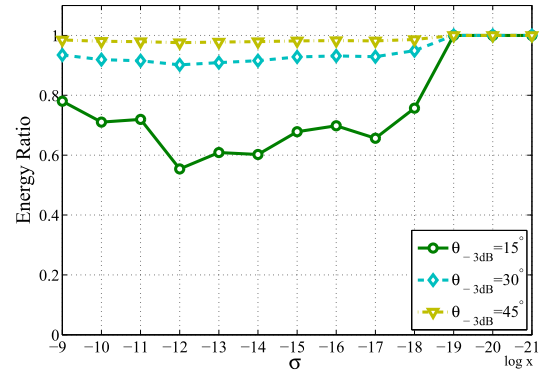


Fig. 13. The energy ratio performance of EMS with different θ_{-3dB} under different interference thresholds.

be 6 to obtain most benefits from D2D communications. Since D2D communications are not enabled in FDMAC, its energy consumption does not change with H_m .

Summarizing the results above, EMS reduces energy consumption, and thus improves energy efficiency via D2D communications and concurrent transmissions, and D2D communications play a significant role in reducing energy consumption.

D. Choice of the Interference Threshold

Since the choice of threshold affects the concurrent transmission mechanism in EMS, which is the difference between EMS and the D2D scheme, the energy ratio under different system parameters should be investigated. Fig. 13 presents the energy ratio of EMS with different θ_{-3dB} under different interference thresholds. The results are shown with X-axis using the logarithmic coordinates. Other parameters are given in Table I except θ_{-3dB} . We can observe that EMS achieves different lowest energy ratio under different θ_{-3dB} . When θ_{-3dB} is smaller, the achieved lowest energy ratio can be lower. If the interference threshold is very small as 10^{-19} , the concurrent transmission mechanism is disabled, and EMS reduces to the D2D scheme. Thus, the energy ratio becomes 1. When the threshold is 10^{-12} , the energy ratio achieves the almost lowest value in three cases, and therefore we select σ to be 10^{-12} in the comparison above.

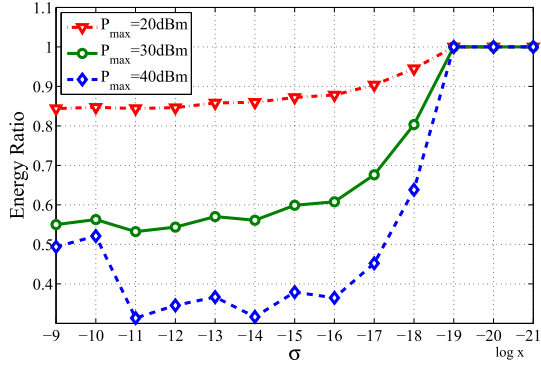


Fig. 14. The energy ratio performance of EMS with different P_{max} under different interference thresholds.

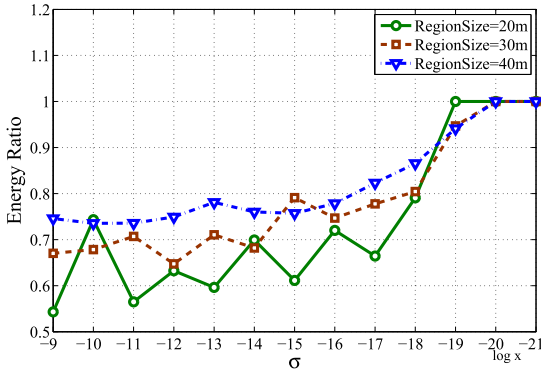


Fig. 15. The energy ratio performance of EMS with different region sizes under different interference thresholds.

Fig. 14 gives the energy ratio of EMS with different P_{max} under different interference thresholds. Other parameters are given in Table I except P_{max} . With the increase of P_{max} , the achieved lowest energy ratio can be lower. The optimal threshold selection is different for different P_{max} . Thus, the threshold should be optimized for different network settings.

The energy ratio of EMS with different region sizes under different interference thresholds is plotted in Fig. 15. With the decrease of the region size, the achieved lowest energy ratio is lower. At the same time, the optimized thresholds for different region sizes are also different. Generally speaking, with the decrease of the region size, the interference power relative to P_{max} increases due to less propagation loss, and the optimal threshold also increases.

VIII. CONCLUSIONS

In this paper, we proposed EMS for energy efficient multicast scheduling in mmWave small cells, which exploits both D2D communications and concurrent transmissions to reduce energy consumption. EMS establishes multi-hop D2D transmission paths by the D2D path planning algorithm. The MIS based concurrent scheduling algorithm schedules the links on the D2D paths into different pairings. Due to more time slots are allocated to each link, lower transmission power can be achieved, and total energy consumption is reduced accordingly. Performance evaluation demonstrates EMS achieves highest energy efficiency compared with other schemes.

TABLE III
NOTATION IN PROBLEM FORMULATION

Symbol	Description
D	The traffic demand for the multicast group
\mathcal{U}	The set of users in the multicast group
u	One user in the multicast group
s_u	The transmit node that serves u
δ^k	The number of time slots for the k th pairing
Δ	The duration of one time slot
a_u^k	A binary variable to indicate whether the multicast transmission for user u is scheduled in the k th pairing
P_t^u	The transmission power of s_u to u
$R_{s_u u}^k$	The transmission rate for user u in the k th pairing
T_s	The number of time slots needed for the serial unicast scheme to complete the multicast service

TABLE IV
NOTATION IN EMS

Symbol	Description
A	The base station
\mathbb{P}	The set of selected D2D paths
p	A path in \mathbb{P}
H_m	The maximum hop number for each path $p \in \mathbb{P}$
L_p	The last node on p
\mathbb{P}_L	The set of last nodes of paths in \mathbb{P}
σ	The interference threshold
$G_t(V_t, E_t)$	The contention graph in the t th pairing
V_t	The set of vertices
E_t	The set of edges
$d(v)$	The degree for any vertex $v \in V_t$
$N(v)$	The set of neighboring vertices for any vertex $v \in V_t$
V_u^t	The set of unvisited and candidate links for the t th pairing
T_u^k	The number of time slots needed for link (s_u, u) to complete the multicast transmission
T^k	The maximum needed number of time slots for links in the k th pairing
$R''_{s_u u}$	The needed transmission rate for link (s_u, u) to complete multicast transmission

Considering the differences between theoretical directional antenna model and practical phased-array beam patterns, we will evaluate the performance of our scheme on a test bed using practical phased-array antennas in the future. Since beamforming in our system relies on the locations of users, we will also investigate the energy consumption involved in finding out pairwise device-device locations. Furthermore, we will also investigate using the multi-level codebook to improve network performance.

APPENDIX A NOTATION IN PROBLEM FORMULATION

In order to facilitate the reader to understand the notations in Problem Formulation, we list the notations in Table III.

APPENDIX B NOTATION IN EMS

To facilitate the understanding of EMS, we also list the notations in Table IV.

REFERENCES

- [1] M. ElKashlan, T. Q. Duong, and H.-H. Chen, "Millimeter-wave communications for 5G—Part 2: Applications [Guest editorial]," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 166–167, Jan. 2015.
- [2] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, Jr., "State of the art in 60-GHz integrated circuits and systems for wireless communications," *Proc. IEEE*, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.

- [3] *High Rate 60 GHz PHY, MAC and HDMIPAL*, Standard ECMA-387, 2010.
- [4] *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*, IEEE Standard 802.15.3c, 2009.
- [5] *Wireless LAN Medium Access Control (MAC) Physical Layer (PHY) Specifications—Amendment 3: Enhancements for Very High Throughput 60 GHz Band*, IEEE Standard 802.11ad, 2012.
- [6] *IEEE P802.11-Task Group ay*. Accessed: Sep. 19, 2017. [Online]. Available: http://www.ieee802.org/11/Reports/tgay_update.htm
- [7] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE/ACM Trans. Netw.*, vol. 19, no. 5, pp. 1513–1527, Oct. 2011.
- [8] J. Wang *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.
- [9] Y. Tsang, A. Poon, and S. Addepalli, "Coding the beams: Improving beamforming training in mmWave communication system," in *Proc. IEEE Global Telecommun. Conf.*, Houston, TX, USA, Dec. 2011, pp. 1–6.
- [10] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.
- [11] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, and Y. Gruenberger, "Is there a case for mobile phone content pre-staging?" in *Proc. CoNEXT*, Santa Barbara, CA, USA, Dec. 2013, pp. 321–326.
- [12] Y. Niu, L. Su, C. Gao, Y. Li, D. Jin, and Z. Han, "Exploiting device-to-device communications to enhance spatial reuse for popular content downloading in directional mmWave small cells," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5538–5550, Jul. 2016.
- [13] S. Naribole and E. W. Knightly, "Scalable multicast in highly-directional 60 GHz WLANs," in *Proc. SECON*, London, U.K., Jun. 2016, pp. 1–9.
- [14] Y. Li, Z. Wang, D. Jin, and S. Chen, "Optimal mobile content downloading in device-to-device communication underlying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3596–3608, Jul. 2014.
- [15] L. Song, D. Niyato, Z. Han, and E. Hossain, *Wireless Device-to-Device Communications and Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
- [16] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks," in *Proc. ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5221–5225.
- [17] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "REX: A randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 113–121, Jan. 2010.
- [18] C.-S. Sum *et al.*, "Virtual time-slot allocation scheme for throughput enhancement in a millimeter-wave multi-Gbps WPAN system," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1379–1389, Oct. 2009.
- [19] C.-S. Sum *et al.*, "A multi-Gbps millimeter-wave WPAN system based on STDMA with heuristic scheduling," in *Proc. GLOBECOM*, Honolulu, HI, USA, Nov./Dec. 2009, pp. 1–6.
- [20] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.
- [21] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2149–2157.
- [22] M. X. Gong, R. Stacey, D. Akhmetov, and S. Mao, "A directional CSMA/CA protocol for mmWave wireless PANs," in *Proc. WCNC*, Sydney, NSW, Australia, Apr. 2010, pp. 1–6.
- [23] S. Singh, F. Ziliotto, U. Madhow, E. M. Belding, and M. Rodwell, "Blockage and directivity in 60 GHz wireless personal area networks: From cross-layer model to multihop MAC design," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1400–1413, Oct. 2009.
- [24] Q. Chen, J. Tang, D. T. C. Wong, X. Peng, and Y. Zhang, "Directional cooperative MAC protocol design and performance analysis for IEEE 802.11ad WLANs," *IEEE Trans. Veh. Technol.*, vol. 62, no. 6, pp. 2667–2677, Jul. 2013.
- [25] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "Blockage robust and efficient scheduling for directional mmWave WPANs," *IEEE Trans. Veh. Technol.*, vol. 64, no. 2, pp. 728–742, Feb. 2015.
- [26] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2052–2069, Oct. 2015.
- [27] H. Park, S. Park, T. Song, and S. Pack, "An incremental multicast grouping scheme for mmWave networks with directional antennas," *IEEE Commun. Lett.*, vol. 17, no. 3, pp. 616–619, Mar. 2013.
- [28] Y. Niu *et al.*, "Energy-efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2674–2687, Mar. 2017.
- [29] J. Ning, T.-S. Kim, S. V. Krishnamurthy, and C. Cordeiro, "Directional neighbor discovery in 60 GHz indoor wireless networks," *Perform. Eval.*, vol. 68, no. 9, pp. 897–915, Sep. 2011.
- [30] H. Deng and A. Sayeed, "mm-Wave MIMO channel modeling and user localization using sparse beamspace signatures," in *Proc. SPAWC*, Toronto, ON, Canada, Jun. 2014, pp. 130–134.
- [31] F. Yildirim and H. Liu, "A cross-layer neighbor-discovery algorithm for directional 60-GHz networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4598–4604, Oct. 2009.
- [32] S. Vasudevan, J. Kurose, and D. Towsley, "On neighbor discovery in wireless networks with directional antennas," in *Proc. INFOCOM*, Mar. 2005, pp. 2502–2512.
- [33] G. Jakllari, W. Luo, and S. Krishnamurthy, "An integrated neighbor discovery and MAC protocol for ad hoc networks using directional antennas," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 1024–1114, Mar. 2007.
- [34] G. Pei, M. Albuquerque, J. H. Kim, D. P. Nast, and P. R. Norris, "A neighbor discovery protocol for directional antenna networks," in *Proc. MILCOM*, Oct. 2005, pp. 487–492.
- [35] T. S. Rappaport, J. H. Reed, and B. D. Woerner, "Position location using wireless communications on highways of the future," *IEEE Commun. Mag.*, vol. 34, no. 10, pp. 33–41, Oct. 1996.
- [36] C. K. Seow and S. Y. Tan, "Non-line-of-sight localization in multipath environments," *IEEE Trans. Mobile Comput.*, vol. 7, no. 5, pp. 647–660, May 2008.
- [37] H. Miao, K. Yu, and M. J. Juntti, "Positioning for NLOS propagation: Algorithm derivations and Cramer–Rao bounds," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 2568–2580, Sep. 2007.
- [38] T. Roos, P. Myllymaki, and H. Tirri, "A statistical modeling approach to location estimation," *IEEE Trans. Mobile Comput.*, vol. 99, no. 1, pp. 59–69, Jan./Mar. 2002.
- [39] L. Yan, X. Fang, and Y. Fang, "Control and data signaling decoupled architecture for railway wireless networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 103–111, Feb. 2015.
- [40] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.
- [41] M. Z. Win *et al.*, "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.
- [42] Z. Liu, W. Dai, and M. Z. Win, "Mercury: An infrastructure-free system for network localization and navigation," *IEEE Trans. Mobile Comput.*, to be published.
- [43] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, Jr., "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.
- [44] S. Geng, J. Kivinen, X. Zhao, and P. Vainikainen, "Millimeter-wave propagation channel characterization for short-range wireless communications," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 3–13, Jan. 2009.
- [45] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, Apr. 2002.
- [46] R. C. Daniels and R. W. Heath, "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, Sep. 2007.
- [47] D. Pisinger, "Where are the hard knapsack problems?" *Comput. Oper. Res.*, vol. 32, no. 9, pp. 2271–2284, Sep. 2005.
- [48] H. Luo, S. Lu, and V. Bhargavan, "A new model for packet scheduling in multihop wireless networks," in *Proc. 6th Annu. Int. Conf. Mobile Comput. Netw.*, Boston, MA, USA, Aug. 2000, pp. 76–86.
- [49] M. M. Halldórsson and J. Radhakrishnan, "Greed is good: Approximating independent sets in sparse and bounded-degree graphs," in *Proc. ACM STOC*, 1994, pp. 439–448.

- [50] Q. Chen, X. Peng, J. Yang, and F. Chin, "Spatial reuse strategy in mmWave WPANs with directional antennas," in *Proc. GLOBECOM*, Anaheim, CA, USA, Dec. 2012, pp. 5392–5397.
- [51] T. Nitsche, G. Bielsa, I. Tejado, A. Loch, and J. Widmer, "Boon and bane of 60 GHz networks: Practical insights into beamforming, interference, and frame level operation," in *Proc. Conf. Emerg. Netw. Experim. Technol.*, 2015, Art. no. 17.



Yong Niu (M'17) received the B.E. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 2011, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, in 2016.

From 2014 to 2015, was a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently an Associate Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His research interests

are in the areas of networking and communications, including millimeter wave communications, device-to-device communication, medium access control, and software-defined networks. He received the Ph.D. National Scholarship of China in 2015, the Outstanding Ph.D. Graduates and the Outstanding Doctoral Thesis of Tsinghua University in 2016, and the Outstanding Ph.D. Graduates of Beijing in 2016. He has served as a Technical Program Committee Member for the IWCMC 2017, VTC2018-Spring, and ICC 2018, and also the Session Chair for the IWCMC 2017.



Yu Liu is currently pursuing the B.E. degree in electronic engineering from Tsinghua University, Beijing, China. His research interests include wireless communications, edge computing, and network optimization.



Yong Li (M'09–SM'16) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. In 2012 and 2013, he was a Visiting Research Associate with Telekom Innovation Laboratories and the HK University of Science and Technology, respectively. From 2013 to 2014, was a Visiting Scientist the University of Miami, Coral Gables, FL, USA. He is currently a

Faculty Member of the electronic engineering with Tsinghua University.

He has authored over 100 research papers and has 10 granted and pending Chinese and international patents. His Google Scholar Citation is about 440 with H-index of 11, also with over 120 total citations without self-citations in Web of Science. His research interests are in the areas of networking and communications, including mobile opportunistic networks, device-to-device communication, software-defined networks, network virtualization, and future Internet. He received the Outstanding Postdoctoral Researcher, the Outstanding Ph.D. Graduates and the Outstanding Doctoral Thesis of Tsinghua University, and his research was granted by the Young Scientist Fund of Natural Science Foundation of China, the Post-Doctoral Special Fund of China, and industry companies of Hitachi, ZET. He has served as the Technical Program Committee (TPC) Chair for the WWW workshop of Simplex 2013, and the TPC of several international workshops and conferences. He is also a Guest-Editor for the *ACM/Springer Mobile Networks and Applications*, Special Issue on Software-Defined and Virtualized Future Wireless Networks. He is currently an Associate Editor of the *EURASIP Journal on Wireless Communications and Networking*.



Xinlei Chen received the B.E. and M.S. degrees in electrical engineering from Tsinghua University, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests are in the areas of networking and communications, including millimeter wave communications, device-to-device communication, medium access control, mobile embedded system, including collaborative drone swarms, and big data.



Zhangdui Zhong is currently a Professor and an Advisor of Ph.D. candidates with Beijing Jiaotong University. He is a Chief Scientist with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He is also the Director of the Innovative Research Team, Ministry of Education, and a Chief Scientist of the Ministry of Railways, China. He has authored or co-authored 7 books, 5 invention patents, and over 200 scientific research papers in his research area. His interests

are wireless communications for railways, control theory and techniques for railways, and GSM-R system. His research has been widely used in the railway engineering, such as Qinghai-Xizang Railway, Datong-Qinhuangdao Heavy Haul Railway, and many high-speed railway lines of China. He is an Executive Council Member of the Radio Association of China, and the Deputy Director of the Radio Association of Beijing. He received the MaoYiSheng Scientific Award of China, the Zhan-TianYou Railway Honorary Award of China, and the Top ten Science/Technology Achievements Award of Chinese Universities.



Zhu Han (S'01–M'04–SM'09–F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland at College Park, MD, USA, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer with JDS Uniphase Corporation, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland at College Park. From 2006 to 2008,

he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a Professor with the Electrical and Computer Engineering Department and the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (the Best Paper Award in the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in the IEEE conferences. He is currently the IEEE Communications Society Distinguished Lecturer.