Evaluating the Impact of User Behavior on D2D Communications in Millimeter-Wave Small Cells

Chuhan Gao, Yong Li, *Member, IEEE*, Haohao Fu, Yong Niu, Depeng Jin, *Member, IEEE*, Sheng Chen, *Fellow, IEEE*, and Han Zhu, *Fellow, IEEE*

***Abstract*—Small cells in millimeter-wave (mmWave) band are able to provide multigigabit access data rates and have emerged as a cost-efficient solution to offer interference-free device-to-device (D2D) communications. In order to improve system performances and enhance user experiences, direct transmissions between devices need to be scheduled properly. We first propose a transmission scheduling scheme for radio access of small cells in the mmWave band, termed directional D2D medium access control (D3MAC), whereby a path-selection criterion is designed to enable D2D transmissions. Through extensive simulations, we demonstrate that D3MAC achieves near-optimal performances and outperforms other schemes significantly in terms of delay and throughput. Based on this near-optimal scheme, we then evaluate the impact of user behaviors, including the traffic mode and traffic load, as well as user density, denseness, and mobility, on the performance of D2D communications in mmWave small cells. Our study reveals that the performance of D2D communications is improved, as the user density and denseness increase, but this effect is only obvious under heavy traffic loads. Furthermore, user mobility is shown to be another important factor that influences the performance of D2D communications. The system performance is first improved, as the average user speed increases from static, but the performance is degraded significantly when the user speed becomes high.**

***Index Terms*—Device-to-device (D2D) communications, millimeter wave (mmWave), scheduling, small cells, user behavior.**

I. Introduction

MOBILE data traffic is increasing rapidly, and a significant increase in the next few years is predicted [1]. In order to improve mobile network capacity so as to meet the ever-increasing demand, device-to-device (D2D) communication is proposed to enable devices to communicate directly, which offers an underlaying to the cellular network for improving spectral efficiency [2], [3]. Under the control of base stations (BSs), user equipments (UEs) can transmit data to each other through direct links using cellular resources instead of through BSs. Consequently, it enables physical-proximity communication, which saves power while improving spectral efficiency dramatically. D2D communication is expected to be a key feature supported by next-generation cellular networks [3].

On the other hand, there has been an increasing interest in deploying small cells underlying the conventional homogeneous macrocell network in the design of the next-generation mobile communication [4]. This network deployment is usually referred to as heterogeneous cellular networks. Small cell in millimeter wave (mmWave) is a promising technology for future cellular networks to provide high-data-rate communications. Unlike existing communication systems that use lower carrier frequencies (e.g., from 900 MHz to 5 GHz), mmWave-band small cells suffer from high propagation loss. The free-space propagation loss at 60-GHz band is 28 dB more than that occurred at 2.4 GHz [5]. Due to the directionality and high propagation loss, however, the interference between mmWave links is minimal. This is highly advantageous to D2D communications, which involve discovering and communicating with nearby devices. Therefore, the potential of D2D communications in mmWave small cells to enhance the network performance is great. Proper scheduling over radio access for D2D transmissions in mmWave small cells is vital to fully realize this potential.

Moreover, it is necessary to investigate the fundamental factors that influence the achievable performance. Suffering from high propagation loss, mmWave D2D communications can only achieve high transmission rates when two UEs are located near to each other with a line of sight (LOS) between them. The distribution of UEs is, therefore, a key factor, and consequently, the influence of the density and denseness of UEs to the achievable system performance must be carefully investigated. In a cellular network, D2D communications exploit spatial reuse by offloading mobile traffic. However, the offloading capability varies under different traffic loads and modes. These factors also impact the performance of D2D communications. Moreover, current related works mainly considered network scenarios with a static UE distribution. In other words, the issue of mobility has not been fully investigated. Although current research studies are

Manuscript received December 10, 2015; revised May 7, 2016; accepted July 29, 2016. Date of publication December 20, 2016; date of current version July 14, 2017. This work was supported by the National Basic Research Program of China (973 Program) under Grant 2013CB329105, by the National Nature Science Foundation of China under Grant 61301080 and Grant 91338102, and by the Research Fund of Tsinghua University under Grant 20161080099. The review of this paper was coordinated by Prof. S. Tomasin. *(Corresponding author: Yong Li.)*

C. Gao, Y. Li, Y. Niu, and D. Jin are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: gaochthu@126.com; liyong07@tsinghua.edu.cn; yongniu@mails.tsinghua.edu.cn; jindp@tsinghua.edu.cn).

H. Fu is with the Princeton International School of Mathematics and Science, Princeton, NJ 08540 USA (e-mail: 18611608511@163.com).

S. Chen is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K., and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: sqc@ecs.soton.ac.uk).

H. Zhu is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA (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/TVT.2016.2642127

meaningful for studying specific cases or snapshots of D2D communications in real-world cellular networks, they did not unveil the impact of user mobility on the achievable system performance. UE mobility not only causes the dynamic changes of network topologies, but also results in frequent change of D2D pairs. Given that 3GPP has recommended several mobility models for performance evaluation [6], it is meaningful and important to evaluate and assess the impact of mobility on D2D performances.

Aiming to address the above problems, we propose a transmission scheduling scheme, termed directional D2D medium access control (D3MAC), for D2D transmissions in mmWave small cells. In a D3MAC-enabled cellular system, whenever a direct link between the sender and the receiver of a flow has high channel quality, the direct transmission will be adopted instead of transmission through BSs and backhaul networks. Therefore, the proposed D3MAC fully exploits direct transmissions between devices to improve the network performance in terms of throughput and delay. Based on our proposed scheme, we evaluate the D2D communication performance in both static and dynamic networks. Specifically, we assess the system performance under different UE density and denseness, traffic loads, and modes, as well as user mobility, to analyze how these factors affect D2D communications. The contribution of this paper is threefold, as summarized in the following.

1) We formulate the scheduling problem over radio access with direct transmissions between devices into a mixed-integer nonlinear program (MINLP), which minimizes the number of time slots to accommodate the transmission demand. Concurrent transmissions, i.e., spatial reuses, are explicitly considered. To solve this problem, we propose an efficient near-optimal scheduling scheme, referred to as D3MAC, which consists of a path selection criterion and a transmission scheduling algorithm. The priority of D2D transmission is characterized by the path selection parameter of the path selection criterion, while concurrent transmissions are fully utilized in the transmission scheduling to maximize the gain of spatial reuse.
2) We evaluate the impacts of the UE distribution and traffic demand as well as the traffic mode jointly on D2D commutations underlaying mmWave small cells. Specifically, we observe that generally increasing UE density and denseness benefit D2D communications, but, under light traffic loads, this improvement is barely observable. In addition, highly erratically arriving traffic flow degrades the D2D performance, especially under heavy traffic loads and with low UE density and denseness. Explanations are given on how these factors affect D2D communications and network performances.
3) We carry out the study not only in static networks but also in dynamic networks to evaluate the impact of UE mobility on D2D communications. It is observed that the performance is improved under low UE speed, where the mobility enables more D2D pairs to establish, compared with the static case. However, D2D communications perform poorly in high-UE-mobility networks as a result of the frequent changes of D2D pairs.

This paper is organized as follows. After presenting the related work in mmWave cellular networks and D2D communications in Section II, Section III introduces the system model and overview. In Section IV, we formulate the scheduling problem for radio access in mmWave small cells, while Section V is devoted to our proposed D3MAC scheme, which includes a path selection criterion and a transmission scheduling algorithm. In Section VI, we demonstrate that D3MAC is able to achieve a near-optimal performance in terms of network throughput and transmission delay. We also evaluate the impact of UE density and denseness as well as traffic loads and traffic modes on the performance of D2D communications in static networks. We then involve UE mobility in the discussion and evaluate its impact on the D2D communications in Section VII. We conclude the paper in Section VIII.

II. Related Work

Recently, a number of studies have investigated the mmWave technology for cellular networks. Wei *et al.* [7] discussed six key elements to enable mmWave communications in future 5G networks and addressed some possible approaches. Scott-Hayward *et al.* [8] defined and evaluated important metrics to characterize multimedia quality of service (QoS) and designed a QoS-aware scheduling scheme. In terms of small cells in the mmWave bands, most works focused on using bands in 28, 38, and 73 GHz to attain communication ranges in the order of 200 m or even more [9]. Zhu *et al.* [10] proposed a 60-GHz picocell architecture to augment with the existing LTE networks for achieving a significant increase in capacity.

We focus on the performance of D2D communications in mmWave cellular networks. By contrast, the majority of the existing research studies have been conducted on D2D communications at lower frequencies. Lin *et al.* [6] provided an overview of D2D standardization activities in 3GPP and identified several technical challenges. Qiao *et al.* [11] proposed an effective resource-sharing scheme by allowing noninterfering D2D links to operate concurrently. Although D2D communication may bring enhancement for spectral efficiency, it also causes interference as the result of spectrum sharing. For mmWave D2D communications, current research works have mainly studied the problems of power control [12], resource allocation [13], and interference management [14], [15]. Taking advantage of high propagation loss and directional antennas, D2D links can be supported in mmWave 5G networks to enhance network capacity and improve spectrum efficiency. Instead of just focusing on transmission schemes or power control, we further investigate the factors that have important impacts on D2D communication and evaluate how these factors influence the achievable performance.

Some of the existing studies have analyzed the performance of D2D communications underlaying systems. Yu *et al.* [12] evaluated the performance of D2D communication by considering a scenario, where only limited interference coordination between the cellular and D2D communications is possible. Works [16], [17] evaluated the D2D systems under different transmission schemes or mode selection mechanisms. The existing work has

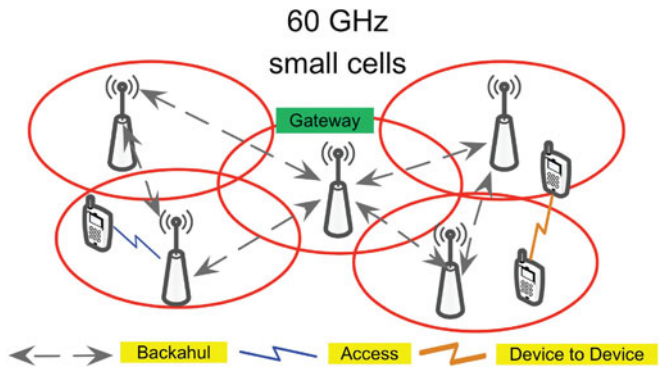

Fig. 1. Illustration of dense deployment of mmWave small cells underlying a macrocell network.

focused on the scenario of microscale cellular networks with lower frequencies. By contrast, we extensively consider the key factors related to the user distribution and traffic demand and investigate their impacts on the performance of D2D communications in mmWave small cells. To the best of our knowledge, there exists no previous study evaluating the influences of user behavior and traffic demands on D2D communications in mmWave small cells.

III. System Model and Problem Overview

A. System Model

Fig. 1 illustrates a typical scenario for dense deployment of mmWave small cells underlying the cellular network. In each small cell, there are several UEs and an access point (AP), which synchronizes the clocks of UEs and provides access services within the small cell. The APs form a mmWave wireless backhaul network, and the backhaul links are optimized in order to achieve high channel quality and reduce interference. Therefore, we assume that the backhaul links are fixed with optimal scheduling, and we focus on the radio access, where D2D communications are enabled and traffic demands can be transmitted through direct links between nearby UEs instead of through the backhaul network. Some APs are connected to the Internet via high-speed wired connections, which are called gateways. The remaining APs must communicate with a gateway in order to send (receive) data to (from) the Internet. To overcome huge path attenuation, both the UEs and APs achieve directional transmissions with electronically steerable directional antennas by beamforming techniques [18].

In the system, there is a centralized controller in the network [19], which usually resides on a gateway. The system resource is partitioned into nonoverlapping time slots of equal length, and the controller synchronizes the clocks of APs. Then, the clocks of UEs are synchronized by their corresponding APs. There is a bootstrapping program in the system, by which the central controller knows the up-to-date network topology and the location information of APs and UEs [20], [21].

In this system, transmissions occur on two types of paths: ordinary and direct paths. A direct path is a direct transmission path from source (a UE) to destination (another UE), which does not pass through the backhaul network. An ordinary path is a transmission path through APs, which may include the access link from the source to its associated AP, the backhaul path from the source's AP to the destination's AP or gateway, and the access link from the destination's associated AP to the destination. The achievable transmission rates for ordinary paths and direct paths can be obtained via a channel transmission-rate measurement procedure [22]. In this procedure, the transmitter of each link transmits measurement packets to the receiver first. Then, with the measured signal-to-noise ratio (SNR) of received packets, the receiver estimates the achievable transmission rate and determines the appropriate modulation and coding scheme by the table related to the SNR value with the appropriate modulation and coding scheme.

With directional transmissions, there exists less interference between links. Under low multiuser interference (MUI), concurrent transmissions can be utilized [23]. The definition of adjacent flows in our system is given as follows: Any two flows are adjacent if the selected links of them share a common node, which can either be an BS or a user device. Two adjacent flows cannot be scheduled concurrently, since all nodes are assumed to be half-duplex, and each of them has at most one connection with one neighbor [24]. We denote link i from sender s_i to destination r_i by (s_i, r_i), and its transmission rate by c_{s_i,r_i}. For two nonadjacent links, we adopt the interference model in [23]. Specifically, for links (s_i, r_i) and (s_j, r_j), the received power from s_i to r_j can be calculated according to

$$P_{r_j,s_i} = f_{s_i,r_j} k_0 P_t l_{s_i,r_j}^{-\gamma} \tag{1}$$

where P_t is the transmission power that is fixed, $k_0 = 10^{\mathrm{PL}(d_0)/10}$ is the constant scaling factor corresponding to the reference path loss $\mathrm{PL}(d_0)$, with d_0 being equal to 1 m, l_{s_i,r_j} is the distance between node s_i and node r_j, and γ is the path loss exponent [23]. The directional indicator f_{s_i,r_j} indicates whether s_i and r_j direct their beams toward each other. If this is the case, $f_{s_i,r_j} = 1$; otherwise, $f_{s_i,r_j} = 0$. Thus, the desired signal-to-interference-plus-noise ratio (SINR) at r_j, denoted by SINR_{s_j,r_j}, can be calculated according to

$$\mathrm{SINR}_{s_j,r_j} = \frac{k_0 P_t l_{s_j,r_j}^{-\gamma}}{WN_0 + \rho \sum_{i \neq j} f_{s_i,r_j} k_0 P_t l_{s_i,r_j}^{-\gamma}} \tag{2}$$

where ρ is the MUI factor related to the cross correlation of the signals from different links, W is the bandwidth, and N_0 is the one-sided power spectra density of white Gaussian noise [23]. For link (s_i, r_i), the minimum SINR to support its transmission rate c_{s_i,r_i} is denoted as $\mathrm{MS}(c_{s_i,r_i})$. Therefore, concurrent transmissions can be supported if the SINR of each link (s_i, r_i) is larger than or equal to $\mathrm{MS}(c_{s_i,r_i})$.

There are two kinds of flows transmitted in the network: the flows between UEs and the flows from or to the Internet (gateway). We assume that there are N_f flows in the network. For flow i, its traffic demand is denoted as d_i. The traffic demand vector for all the flows is denoted by $\mathbf{d}$, a $1 \times N_f$ row vector whose ith element is d_i. For each flow, there are two possible transmission paths in the system: ordinary path and direct path. A flow transmitted via an ordinary path is inherently multihop,

while a flow transmitted via a direct path is single hop. For the lth hop link of the ordinary path for flow i, we denote its sender as $s^o_{l(i)}$ and receiver as $r^o_{l(i)}$ and denote this link as $\left(s^o_{l(i)}, r^o_{l(i)}\right)$. We denote the direct link of flow i as $\left(s^d_i, r^d_i\right)$, with s^d_i as the source and r^d_i as the destination. If no confusion is caused, the superscripts ^{o} and/or ^{d} may be dropped.

Let the maximum number of hops of the ordinary paths be $H_{\max}$. Then, for each flow i, its $1 \times H_{\max}$ transmission-rate vector on the ordinary path is denoted as $\mathbf{c}^o_i$, where each element $c^o_{l(i)}$ represents the transmission rate of the lth hop. We also denote the $N_f \times H_{\max}$ transmission-rate matrix for the ordinary paths of all flows by $\mathbf{C}^o$, whose ith row is simply $\mathbf{c}^o_i$. The transmission rate of the direct path for flow i is denoted as c^d_i, and the $1 \times N_f$ transmission-rate vector for the direct paths of all flows is denoted as $\mathbf{c}^d$, whose ith element is c^d_i.

B. Operation Procedure and Problem Overview

The proposed D3MAC is a frame-based medium access control (MAC) protocol similar to the frame-based scheduling directional MAC (FDMAC) of [24]. Each frame consists of a scheduling phase and a transmission phase, and the scheduling overhead in the scheduling phase can be amortized over multiple concurrent transmissions in the transmission phase as in the FDMAC of [24]. In the scheduling phase, AP polls its associated UEs successively for their traffic demands and reports to the central controller through the backhaul network. Based on the transmission rates of links, the central controller computes a schedule to accommodate the traffic demands of all flows. Then, the central controller pushes the schedule to the APs through the backhaul network, and each AP pushes the schedule to its UEs. In the transmission phase, UEs and APs communicate with each other following the schedule until the traffic demands of all flows are accommodated. The transmission phase consists of multiple stages, and in each stage, multiple flows are activated simultaneously for concurrent transmissions. In a stage, several selected flows transmit the packets through the path selected by the path selection algorithm of D3MAC. Since each flow may consist of different hops, the number of hops of each stage is also unfixed. The number of slots of a stage is not fixed, either due to the fact that the transmissions of each flow require different number of slots. Both the number of hops and the number of slots of a stage depend on which flows are activated. A stage finishes after all the selected flows clear their traffic. In the schedule computation, the transmission path needs be selected optimally between the direct path and ordinary path for each flow, and the schedule should accommodate the traffic demands of flows with a minimum number of time slots to fully exploit spatial reuse. It should be noted that the number of hops of a flow is unknown until D3MAC determines which path to select for transmission.

Let us illustrate the basic idea of D3MAC with Fig. 2, where there are three small cells. In cell 1, UEs A and C are associated with AP1; in cell 2, UE B is associated with AP2; and in cell 3, UE D is associated with AP3. Assume that there are two flows in the network: A→B and C→D. The traffic demands of A→B and C→D are 6 and 8, respectively, and thus, $\mathbf{d} = [6\ 8]$. Numerically, they are equal to the number of packets

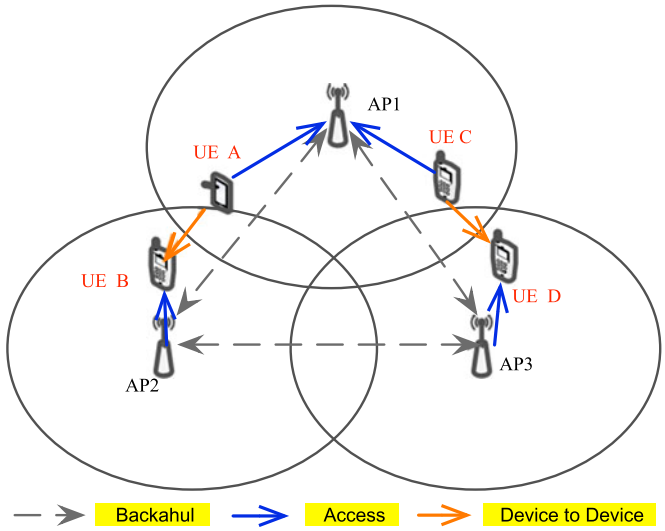

Fig. 2. Example of D3MAC with three small cells.

to be transmitted, assuming that the packet length is fixed. The transmission-rate matrix for the ordinary paths of flows obtained by the measuring procedure is

$$\mathbf{C}^o = \begin{bmatrix} 2 & 3 & 2 \\ 2 & 4 & 2 \end{bmatrix} \tag{3}$$

which indicates that the transmission rates of links A→AP1, AP1→AP2, and AP2→B are 2, 3, and 2, respectively, while the rates of links C→AP1, AP1→AP3, and AP3→D are 2, 4, and 2, respectively. If direct transmission is not enabled, AP1 cannot be scheduled to serve the transmissions of UE A and UE B at the same time due to the half-duplex restriction. In addition, AP 2 should wait to receive all the packets from UE A before starting transmitting them to UE B, since it is not able to transmit to UE B and receive packets from AP 1 at the same time. According to $\mathbf{C}^o$ and $\mathbf{d}$, these two flows need eight and ten time slots to clear all the traffic demands, respectively. Assume that AP 1 serves UE A first and serves UE C after it forwarded the traffic of A to AP2; 15 time slots are required to clear all the traffic in total. On the other hand, the transmission-rate vector for the two direct paths is measured to be $\mathbf{c}^d = [3\ 2]$. This indicates that the direct link of A→B can transmit three packets in one time slot, and the direct link of C→D can transmit two packets in one time slot.

Clearly, for each flow, we need to first select the optimal transmission path between its direct path and ordinary path. The optimal schedule needs to accommodate the traffic demand of flows with a minimum number of time slots. In other words, concurrent transmission should be fully exploited in the schedule. For the example in Fig. 2, the direct links of the both flows should be enabled to enhance performances. According to $\mathbf{c}^d$ and $\mathbf{d}$, these two flows need two and four slots to clear their traffic demands, respectively. Moreover, the two direct paths can be scheduled for concurrent transmission, since they have no common node. Therefore, only four time slots are needed in total. This simple example clearly shows that the selection of transmission paths for flows has a significant impact on the

efficiency of scheduling, and concurrent transmission scheduling needs to be optimized to improve transmission efficiency, which is the basic idea of the proposed D3MAC. It should be noted that the example presented is not a common case in small cells, and our purpose is to illustrate how D2D communications benefit the system performance in our scheme.

IV. Problem Formulation and Analysis

A. Radio Access Without Enabling D2D Communications

We begin by formulating the transmission scheduling problem without introducing D2D communications, where the traffics of all flows are transmitted only through ordinary paths. Since there are only ordinary paths, we drop the superscript ^{o}. Given the traffic demand of flows, to maximize the transmission efficiency, we should accommodate the traffic demand with a minimum number of time slots [24]. Assume that the schedule has K stages, and the number of time slots of the kth stage is δ^k, while the duration of a time slot is denoted as τ. The total number of time slots of a schedule is then $\sum_{k=1}^{K}\delta^k$. For each flow i, we define the number of hops for its ordinary path as its hop number H_i, and we further define a binary variable $b^k_{j(i)}$ to indicate whether the jth hop of the ordinary path for flow i is scheduled to transmit in the kth stage. For any two links (s_i, r_i) and (s_j, r_j), we define a binary variable $I(s_i, r_i; s_j, r_j)$ to indicate whether these two links are adjacent. If they are, $I(s_i, r_i; s_j, r_j) = 1$; otherwise, $I(s_i, r_i; s_j, r_j) = 0$. In a schedule, if a link is scheduled in one stage, it will transmit as many packets as possible until its traffic demand is cleared. Then, the link will not be active in the remaining slots of this stage. Since concurrent transmissions interfere with each other, the SINR of the jth hop of the ordinary path of flow i can be expressed as

$$\mathrm{SINR}_{j(i)} = \frac{P_{r_{j(i)},s_{j(i)}}\cdot b^k_{j(i,)}}{WN_0+\rho\sum_u\sum_{v(u)\neq j(i)}P_{r_{j(i)},s_{v(u)}}\cdot b^k_{v(u)}} \quad (4)$$

where $s_{j(i)}$ and $r_{j(i)}$ denote the transmitter and the receiver of the jth hop of the ordinary path for flow i, respectively. The transmitting rate of link $\left(s_{j(i)}, r_{j(i)}\right)$ is, therefore, expressed as

$$c_{j(i)} = \eta W \log_2\left(1+\mathrm{SINR}_{j(i)}\right) \quad (5)$$

where $\eta \in (0,\ 1)$ is the efficiency of the transceiver design.

Regarding the system constraints, first, all traffic demands should be scheduled, which can be expressed as

$$\sum_{k=1}^{K}\delta^k\tau b^k_{j(i)}c_{j(i)} \geq d_i,\ \forall i \text{ and } j(i)=1,2,\ldots,H_i. \quad (6)$$

To avoid frequent beamforming or steering, each link can be activated at most once in a schedule, which means that

$$\sum_{k=1}^{K} b^k_{j(i)} = \begin{cases} 1, & \text{if } d_i > 0 \\ 0, & \text{otherwise,} \end{cases} \quad \forall i \text{ and } j(i)=1,2,\ldots,H_i. \quad (7)$$

Adjacent links cannot be scheduled concurrently in the same stage due to half-duplexing, which can be expressed as

$$\forall i, j(i), u, v(u), k \text{ and } v(u)\neq j(i):$$
$$b^k_{j(i)} + b^k_{v(u)} \leq 1 \text{ if } I\left(s_{j(i)}, r_{j(i)}; s_{v(u)}, r_{v(u)}\right) = 1. \quad (8)$$

The jth hop of the ordinary path for flow i should be scheduled ahead of the $(j+1)$th hop, which means that

$$\forall i,\ j(i)=1,2,\ldots,H_i-1 \text{ and } \tilde{K}=1,2,\ldots,K:$$
$$\sum_{k=1}^{\tilde{K}} b^k_{j(i)} \geq \sum_{k=1}^{\tilde{K}} b^k_{(j+1)(i)} \text{ if } H_i > 1. \quad (9)$$

Therefore, the optimal scheduling problem without enabling D2D communications (P1) is formulated as follows:

$$\min \sum_{k=1}^{K}\delta^k,$$
$$\text{s.t. constraints (6)–(9) hold.} \quad (10)$$

B. Radio Access With Enabling D2D Communications

We now include D2D communications into the scheduling, where each flow can choose using either the original path or the direct path. Define a binary variable a^k_i to indicate whether the direct link of flow i is scheduled to transmit in the kth stage, i.e., if this is the case, $a^k_i = 1$; otherwise, $a^k_i = 0$. The SINRs of the jth hop of the ordinary path for flow i and the direct path for flow i in the kth stage, denoted by $\mathrm{SINR}^o_{j(i)}$ and SINR^d_i, respectively, can be expressed as

$$\mathrm{SINR}^o_{j(i)} = \frac{P_{r^o_{j(i)},s^o_{j(i)}}\cdot b^k_{j(i)}}{WN_0+\rho\sum_u\sum_{v(u)\neq j(i)}P_{r^o_{j(i)},s^o_{v(u)}}\cdot b^k_{v(u)}+\rho\sum_p P_{r^o_{j(i)},s^d_p}\cdot a^k_p} \quad (11)$$

$$\mathrm{SINR}^d_i = \frac{P_{r^d_i,s^d_i}\cdot a^k_i}{WN_0+\rho\sum_u\sum_{v(u)}P_{r^d_i,s^o_{v(u)}}\cdot b^k_{v(u)}+\rho\sum_{p\neq i}P_{r^d_i,s^d_p}\cdot a^k_p}. \quad (12)$$

The transmitting rate of link $(s^o_{j(i)}, r^o_{j(i)})$ is, therefore, given by $c^o_{j(i)} = \eta W\log_2(1+\mathrm{SINR}^o_{j(i)})$, while the transmitting rate of link (s^d_i, r^d_i) is $c^d_i = \eta W\log_2(1+\mathrm{SINR}^d_i)$.

Next, let us analyze the system constrains. First, each flow can choose either an ordinary path or a direct path to transmit data, which indicates that

$$\forall i,k:\ \sum_{j(i)=1}^{H_i} b^k_{j(i)} = \begin{cases} H_i, & \text{if } a^k_i = 0 \\ 0, & \text{if } a^k_i = 1. \end{cases} \quad (13)$$

Second, all traffic demands should be scheduled; hence

$$\sum_{k=1}^{K}\delta^k\tau\left(c^o_{j(i)}b^k_{j(i)}+c^d_i a^k_i\right)\geq d_i,\ \forall i \text{ and } j(i)=1,2,\ldots,H_i. \quad (14)$$

Each link can be activated at most once in a schedule, which can be expressed as

$$\sum_{k=1}^{K} b_{j(i)}^{k} + a_i^k = \begin{cases} 1, & \text{if } d_i > 0 \\ 0, & \text{otherwise,} \end{cases} \quad \forall i \text{ and } j(i) = 1, 2, \ldots, H_i. \tag{15}$$

Adjacent links cannot be scheduled concurrently in the same stage, which requires the following:

$$\forall i, j(i), u, v(u), k \text{ and } v(u) \neq j(i):$$

$$b_{j(i)}^{k} + b_{v(u)}^{k} \leq 1 \text{ if } I\left(s_{j(i)}^{o}, r_{j(i)}^{o}; s_{v(u)}^{o}, r_{v(u)}^{o}\right) = 1; \tag{16}$$

$$\forall i, u, k \text{ and } u \neq i: a_i^k + a_u^k \leq \text{ if } I\left(s_i^d, r_i^d; s_u^d, r_u^d\right) = 1; \tag{17}$$

$$\forall i, u, v(u), k \text{ and } u \neq i:$$

$$a_i^k + b_{v(u)}^{k} \leq 1 \text{ if } I\left(s_i^d, r_i^d; s_{v(u)}^{o}, r_{v(u)}^{o}\right) = 1. \tag{18}$$

To enable concurrent transmissions, the SINR of each link in the same stage should be able to support its transmission rate, which means that for both direct path and ordinary path, the following must hold:

$$\text{SINR}_{j(i)}^{o} \geq \text{MS}\left(c_{j(i)}^{o}\right) \cdot b_{j(i)}^{k} \quad \forall i, j(i), k \tag{19}$$

$$\text{SINR}_{i}^{d} \geq \text{MS}\left(c_{i}^{d}\right) \cdot a_{i}^{k} \quad \forall i, k \tag{20}$$

where $\text{MS}(c_{j(i)}^{o})$ denotes the minimum SINR required for the jth ordinary-path link to support flow i at the rate $c_{j(i)}^{o}$, while $\text{MS}(c_i^d)$ is the minimum SINR required for the direct-path link to support flow i at the rate c_i^d. Finally, constraint (9) is still required.

Therefore, the problem of optimal scheduling (P2), where D2D communications are enabled, is formulated as follows:

$$\min \sum_{k=1}^{K} \delta^k,$$

$$\text{s.t. constraints } (9) \text{ and } (13)\text{–}(20) \text{ hold.} \tag{21}$$

By solving Problem P2, we can obtain the optimal scheduling solution for the network, which minimizes the slots needed for transmissions. However, Problem P2 is an MINLP, as some of the constraints in P2 are nonlinear constraints.

V. D3MAC Scheme

It is computationally unacceptable to use an exhaustive search to solve P2 for practical networks with mmWave small cells, where the duration of a time slot is only a few microseconds. Therefore, we construct a heuristic algorithm with low complexity to obtain a near-optimal solution so that the scheduling scheme can be implemented easily in practice. We solve this MINLP in two steps. At the first step, we select an appropriate transmission path, either a direct path or an ordinary path, for each flow. If the direct path has high channel quality and can achieve higher transmitting capability than the ordinary path, we choose the direct transmission. At the second step, we accommodate the traffic demand of flows with as few time slots as possible by making full use of every slot and enabling concurrent D2D links as many as possible.

Algorithm 1: Path Selection.

Input: Sets P^b and P^d;
Output: The set of selected paths for all the flows P;
Initialization: $P = \emptyset$;
for *each flow* i **do**
  Obtain $A(p_i^o)$ and $A(p_i^d)$;
  if $\frac{A(p_i^d)}{A(p_i^o)} \geq \beta$ **then**
    $P = P \cup p_i^d$;
  else
    $P = P \cup p_i^b$;
Return P.

A. Path Selection Criterion

For flow i, let p_i^d denote its direct path and p_i^o denote its ordinary path. For the direct transmission path p_i^d with the transmission rate $c_{p_i^d}^d$, we define its transmission capability as

$$A\left(p_i^d\right) = c_{p_i^d}^d. \tag{22}$$

Furthermore, we assume that the ordinary path p_i^o has $H_{\max_i}$ hops, and the transmission rate of its jth hop is $c_{j(p_i^o)}^o$. We can define the transmission capability of the ordinary path p_i^o as

$$A\left(p_i^o\right) = \frac{1}{\sum_{j=1}^{H_{\max_i}} \frac{1}{c_{j(p_i^o)}^o}}. \tag{23}$$

For each flow i, we choose the path with higher transmission capability between its direct path and ordinary path. Therefore, the path selection criterion can be expressed as

$$\forall i: \begin{cases} \text{if } \frac{A\left(p_i^d\right)}{A\left(p_i^o\right)} \geq \beta, & \text{choose } p_i^d \\ \text{otherwise,} & \text{choose } p_i^o \end{cases} \tag{24}$$

where $\beta \geq 1$ is the path selection parameter. The smaller the β, the higher the priority of direct transmissions between devices.

The pseudocode of the path selection process is presented in Algorithm 1, where P^o and P^d denote the sets of all potential paths p_i^o and p_i^d, respectively, while the set P contains all the selected paths for all the flows.

B. Heuristic Transmission Scheduling Algorithm

We propose a heuristic transmission scheduling algorithm to accommodate the traffic demand of flows with as few time slots as possible by fully exploiting spatial reuse. In order to manage the interference effectively by choosing proper hops for concurrent transmissions, we introduce a contention graph to depict the contention relationship between hops. In the contention graph, each vertex represents a hop in the network, and there is an edge between two vertices if there exists severe interference between these two hops. For hop l and hop j, we define the maximum interference between them as

$$\omega_{l,j} = \max\left\{P_{r_j,s_l}, P_{r_l,s_j}\right\}. \tag{25}$$

To control the interference, we set a threshold $\sigma_{l,j}$, and the contention graph is constructed in the way that if the maximum interference between two vertices is less than the threshold, i.e.,

Algorithm 2: Heuristic Transmission Scheduling.

1 **Input**: The set of selected paths for all the flows in stage k, denoted by P^k;
2 **Output**: $\{\mathcal{H}_u^k\}_{u=1}^{U^k}$, δ^k;
3 **Initialization**: Obtain the set of all the hops in P^k, denoted by $\mathcal{H}^k$; $u=0$; $\delta^k=0$;
4 **while** $|\mathcal{H}^k|>0$ **do**
5 $u=u+1$; $\mathcal{H}_u^k=\emptyset$; $\zeta_u^k=0$;
6 Obtain all the hops that can be scheduled currently into the set $\widetilde{\mathcal{H}}$;
7 Obtain $G_u^k(V_u^k,E_u^k)$ based on $\widetilde{\mathcal{H}}$;
8 **while** $|\widetilde{\mathcal{H}}|>0$ **do**
9 Obtain $v\in V_u^k$ with the largest weight W_v;
10 $\mathcal{H}_u^k=\mathcal{H}_u^k\cup v$;
11 **for** *each hop* $(s_{i(j)},r_{i(j)})$ *in* $\mathcal{H}_u^k$ **do**
12 Obtain $\text{SINR}_{s_{i(j)},r_{i(j)}}$;
13 **if** $\text{SINR}_{s_{i(j)},r_{i(j)}}<MS(c_{i(j)})$ **then**
14 $\mathcal{H}_u^k=\mathcal{H}_u^k\setminus v$;
15 Go to line 18;
16 $\zeta_u^k=\max\left\{\zeta_u^k,\lfloor\frac{d_v}{c_{i(j)}\tau}\rfloor\right\}$;
17 $\widetilde{\mathcal{H}}=\widetilde{\mathcal{H}}\setminus N(v)$;
18 $\widetilde{\mathcal{H}}=\widetilde{\mathcal{H}}\setminus v$;
19 Obtain $G_u^k(V_u^k,E_u^k)$ based on $\widetilde{\mathcal{H}}$;
20 $\delta^k=\delta^k+\zeta_u^k$;
21 $\mathcal{H}^k=\mathcal{H}^k\setminus\mathcal{H}_u^k$;
22 $U^k=u$;
23 Return $\{\mathcal{H}_u^k\}_{u=1}^{U^k}$, δ^k.

if $\omega_{l,j}<\sigma_{l,j}$, there will be no edge between these two vertices. Otherwise, there will be an edge between them. There is always an edge between any two adjacent hops, since they cannot be scheduled for concurrent transmissions due to the half-duplex assumption. We denote the contention graph by $G(V,E)$, where V denotes the set of vertices in the contention graph, and E denotes the set of edges in the contention graph. We refer to two vertices as neighbors if there is one edge between them in the contention graph. For any vertex $v\in V$, we denote the set of its neighboring vertices by $N(v)$. We further define the weight of vertex v as the number of time slots that hop v needs for transmission, denoted by W_v.

The pseudocode of our transmission scheduling algorithm is presented in Algorithm 2. To manage the interference between concurrent transmitting hops, the hops with an edge between them in the contention graph should not be scheduled in the same time. Hence, we first obtain all the hops that can be scheduled currently at stage k and build the contention graph (lines 6 and 7), based on the which we schedule the unscheduled hops of flows iteratively in a nonincreasing order of weight with the conditions for concurrent transmissions satisfied (lines 8–19). In line 11, a hop in $\mathcal{H}_u^k$ can either be a hop of an ordinary path or a link of a direct path, and we have draped the corresponding superscript ^{o} or ^{d}. Also, if it is a direct link, we have $i(j)=i$. In line 16, $\lfloor\ \rfloor$ is the integer floor operator. In this inner loop, scheduling stops when no possible hop can be scheduled concurrently any more. The algorithm carries out this process iteratively until all the hops of all the flows considered in stage k are properly scheduled (lines 4–21). In the output of the algorithm, U^k is the maximum number of hops for the longest multihop flow scheduled at stage k, and for $1\le u\le U^k$, $\mathcal{H}_u^k$ contains the hops or links that are scheduled for concurrent transmissions, while δ^k is the number of time slots required for $\left\{\mathcal{H}_u^k\right\}_{u=1}^{U^k}$. Note that if n denotes the number of UEs in the network, the number of concurrent transmission links should be no more than $\lfloor n\rfloor$ [24], due to the nonadjacent constraint. The computational complexities of the path selection algorithm and the transmission scheduling are $\mathcal{O}(N_f)$ and $\mathcal{O}({N_f}^3)$, respectively. Therefore, D3MAC has a complexity of $\mathcal{O}({N_f}^3)$, which is a pseudopolynomial time solution and can be implemented in practice.

VI. Static Network Evaluation and Analysis

A. Comparison With Optimal Solution and Other Protocols

Under a static network environment, we first give the extensive performance evaluation for our proposed D3MAC scheme, given various traffic patterns. Specifically, we compare its performance with those of the optimal solution, obtained by solving the problem P2 with YALMIP [25], and some of the existing protocols. In the simulation, the transmission rate R between UEs as well as between UEs and associated APs is set to 2, 4, and 6 Gbps, respectively, according to the distances between devices, path loss as well as the antenna gain. Due to better channel qualities, the transmission rate of backhaul links is set to 6 Gbps. With $R=2$ Gbps, a packet can be transmitted in one time slot. The packets with transmission delay larger than the delay threshold ϱ are declared as unsuccessful transmission and discarded. Generally, the central controller is able to complete traffic polling, schedule computation, and schedule pushing in a few time slots. Two traffic modes, the Poisson process (PP) and interrupted PP (IPP), are used in the performance evaluation.

1) PP packets arrive at each flow following the PP with arrival rate λ. The traffic load, denoted by T_{load}, in a PP traffic is defined as
$$T_{\text{load}}=\frac{\lambda\cdot L\cdot N_f}{R} \tag{26}$$
where L is the size of data packets.
2) IPP packets arrive at each flow following the IPP with parameters λ_1, p_1, λ_2, and p_2. The arrival intervals of the IPP obey the second-order hyperexponential distribution with a mean of
$$E(X)=\frac{p_1}{\lambda_1}+\frac{p_2}{\lambda_2}. \tag{27}$$
The traffic load T_{load} in this case is defined as
$$T_{\text{load}}=\frac{L\cdot N_f}{E(X)\cdot R}. \tag{28}$$

We do not consider fading in our simulations due to the fact that 60-GHz channels are extremely sparse. The spatial channel response is dominated by a few paths from a few angular directions [29], [30]. Such sparsity is because mmWave signal energy tends to concentrate around the LOS path and a few non-LOS paths from strong reflectors due to beamforming achieved by phased array antennas, which causes channel hardening to weaken fading significantly.

In present LTE systems, the amount of D2D traffic is generally smaller than in our traffic model. We adopted such a traffic model in order to prove that D3MAC is able to exploit the potential and benefit of D2D communications to a great extent.

TABLE I
PARAMETERS OF A SIMULATED NETWORK.

Parameter	Symbol	Value
Duration of one time slot	τ	5 μs
Data packet size	L	1000 bytes
Delay threshold	ϱ	1000 time slots
Contention graph threshold	σ	0.1 mW
PHY data rate	R	2 Gbps, 4 Gbps, 6 Gbps
Propagation delay	dl_p	50 ns
PHY overhead	T_{PHY}	250 ns
Short MAC frame Tx time	T_{ShFr}	$T_{\mathrm{PHY}} + 14 * 8/R + dl_p$
Packet transmission time	T_{packet}	$1000 * 8/R$
SIFS interval	T_{SIFS}	100 ns
ACK Tx time	T_{ACK}	T_{ShFr}
Path selection parameter	β	2

SIFS: Short InterFrame Space

Meanwhile, we expect D2D traffic will experience a significant increase in future communication systems with the potential development of various new applications, which will benefit from D2D communications. The achievable system performance is assessed by the following two metrics.

1) *Average transmission delay*: This is the average transmission delay of the received packets from all the flows, and we evaluate it in units of time slots.
2) *Network throughput*: This is the total number of the successful transmissions of all the flows over the duration of the simulation. For each received packet, if its delay is less than or equal to the threshold ϱ, it is counted as a successful transmission.

1) Comparing With Optimal Solution: We first compare the D3MAC with the optimal solution. Since obtaining the optimal solution is NP hard, we only simulate a scenario of three cells with four users. We consider a network with two APs and seven users. The distance between APs is 20 m, and each of them has a coverage radius of 10 m. There are $N_f = 4$ flows in the network, one of which is the flow between a user and the gateway and the other are the flows among users. The simulation length is set to 0.025 s. The relevant parameters of the simulated network are listed in Table I. Under heavy loads, the execution time of obtaining the optimal solution becomes prohibitively long. Consequently, we can only obtain and present the results under light loads.

Fig. 3 compares the achieved throughput and delay performance by the proposed D3MAC with those of the optimal solution under Poisson traffics, where it can be observed that the performance gap between the D3MAC and the optimal solution is negligible. Even under the traffic load of 2.8, the D3MAC only increases the average transmission delay by less than 10% and reduces the network throughput by less than 3%, compared with the optimal solution. We point out that by optimizing the path selection parameter β, the performance gap between the D3MAC and the optimal solution can be further reduced. The results of Fig. 3, therefore, demonstrate that the D3MAC achieves a near-optimal performance.

2) Comparison With Other Protocols: Next, we compare the D3MAC with the following three benchmark schemes.

1) *Ordinary directional MAC (ODMAC):* In the ODMAC [24], [26], [27], D2D transmissions are not enabled, and all the flows are transmitted through their ordinary paths. The scheduling algorithm of the ODMAC is the same as that used in the proposed D3MAC. This benchmark scheme represents the current state of the art in terms of scheduling the access or backhaul without considering D2D transmissions.
2) *Random path directional MAC (RPDMAC):* The RPDMAC selects the transmission path for each flow randomly from its direct path and ordinary path. Although random selection is rarely adopted in practical schemes, RPDMAC's scheduling algorithm is the same as that of the D3MAC. Thus, it is a good benchmark scheme to show the advantages of the path selection criteria in the D3MAC.
3) *FDMAC-E:* This is an extension of the FDMAC [24], and to the best of our knowledge, the FDMAC achieves the highest efficiency in terms of spatial reuse. In the FDMAC-E, the transmission path is selected in the same way as the D3MAC with the path selection parameter $\beta = 2$. However, in order to show the role of backhaul optimization, the access links and backhaul links are separately scheduled in the FDMAC-E. The access links from UEs to APs are scheduled by the greedy coloring (GC) algorithm of the FDMAC [24]. The backhaul links on the transmission path are scheduled by the time-division multiple access. The access links from APs to UEs are also scheduled by the GC algorithm.

A typical dense deployment of mmWave small cells is simulated, where nine APs, i.e., nine small cells, are uniformly distributed in a square area of 50 m $\times$ 50 m, and the gateway is located at the center of the area. Forty users are uniformly randomly distributed in the simulated area. We believe that this is actually a potential realistic scenario for the next-generation communication system (5G). For example, multiple users within close proximity request to download a same popular content, such as a video, in which case D2D communications have the potential to benefit the system performance. The simulation duration is set to 0.5 s, and the delay threshold ϱ is set to 10^4 time slots, while the rest of the simulation parameters are listed in Table I. Fig. 4 compares the network throughputs as functions of traffic load for the four protocols under the PP traffic. It can be seen from Fig. 4 that under the light load from 0.5 to 1.5, all the four schemes achieve similar performance. The performance of the ODMAC protocol degrades considerably when the traffic load increases beyond 1.5, and it attains the worst performance. The RPDMAC protocol only begins degrading when the network load increases beyond 2, and it outperforms the ODMAC scheme which confirms that enabling D2D transmissions improves the network throughput. For the FDMAC-E protocol, the rate of increase in the throughput begins reducing as the load becomes larger than 2, and its throughput becomes saturated around 400 000 for the high traffic load, which is significantly larger than those of the RPDMAC and ODMAC schemes. The proposed D3MAC protocol attains the best performance. Specifically, the throughput of the D3MAC increases

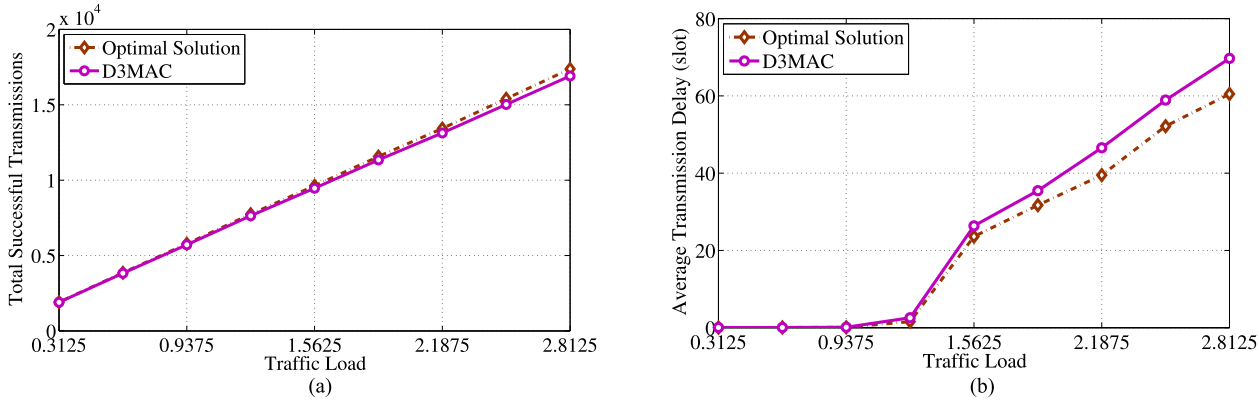

Fig. 3. Comparison with the optimal solution under the Poisson traffic. (a) Network throughput. (b) Transmission delay.

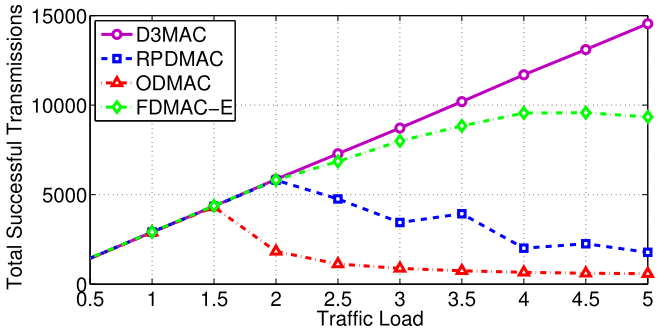

Fig. 4. Network throughputs as functions of traffic load achieved by the four schemes under the Poisson traffic.

linearly with the traffic load. At the high load of 5, the achievable network throughput of the D2MAC is 56% higher than that of the FDMAC-E.

B. Performance Evaluation in Static Networks

We now evaluate the impact of UE behaviors. The network topology is the same as used in obtaining the results of Fig. 4. We concentrate on investigating the influences of the UE density and the UE denseness, under different traffic loads and different traffic modes, on the achievable system performance.

1) Influence of UE Density: To evaluate how the density of UEs influences D2D communications, we deploy 20, 30, 40, 50, and 60 UEs uniformly in the area, which makes the average density 0.008, 0.012, 0.016, 0.020, and 0.024 UE/m^2, respectively. Each UE is associated with the nearest AP. Thirty flows are set in the simulation, 20 of which are between UEs, while 10 of which are between UEs and the gateway, including uploading and downloading. Other simulation parameters are identical to those used in obtaining the results of Fig. 4.

Fig. 5 depicts the network throughputs as the function of the UE density obtained by the D3MAC with three different traffic loads, while Fig. 6 shows the network throughputs as the function of the traffic load with five different UE densities, under both the PP and IPP traffics. We observe that, basically, the network throughput increases with the density of UEs as well as with the traffic load, but the impact of the UE density is heavily influenced by the traffic loads and vice versa.

More specifically, it can be seen from Fig. 5(a) that, when the UE density is very low, the achievable throughputs are similar under all the three PP traffic loads, and additionally increasing the UE density has the same positive impact for all the three cases of traffic loads. However, for the relatively light traffic load of 3, when the UE density exceeds 0.012 UE/m^2, the network throughput becomes saturated. Also, the throughput increase in the case of traffic load 3.5 becomes slower when the UE density exceeds 0.016 UE/m^2. By contrast, the rate of the throughput increase with the traffic load of 5 actually increases when the UE density exceeds 0.016 UE/m^2. Similarly, as can be seen from Fig. 6(a), under light traffic loads, the network throughput increases with the traffic load, and this trend is independent of the UE density. When the traffic load exceeds a certain critical value, which is different for different UE densities, the throughput starts growing slowly and eventually becomes saturated under heavy loads. The exception is the very high UE density case of 0.024 UE/m^2, where the network throughput keeps increasing with the traffic load. The throughputs of the IPP-traffic network given in Figs. 5(b) and 6(b) show the same trends of the PP-traffic network, but the system under IPP traffic attains slightly lower throughput than the PP-traffic network, which is caused by less stable arrival of IPP packets.

It can be seen that the influence of the UE density and traffic load on the achievable network throughput is highly complicated. The UE density has significant impacts on D2D communication in heavily loaded systems, and high UE density increases the system throughput greatly. Increasing the UE density may also increase the throughput performance of D2D communication under light loads, but the improvement is less obvious and smaller. High UE density increasing the system throughput can be explained as follows. The average distance between UEs decreases, as the number of UEs increases in the network, which improves the channel qualities between UEs and increases the transmission rates of D2D links. This allows more flows to transmit packets through D2D links instead of ordinary paths, leading to a higher throughput. However, this impact on throughput is heavily depended on the traffic load. The increase of throughput brought by high UE density is more profound under heavy traffic

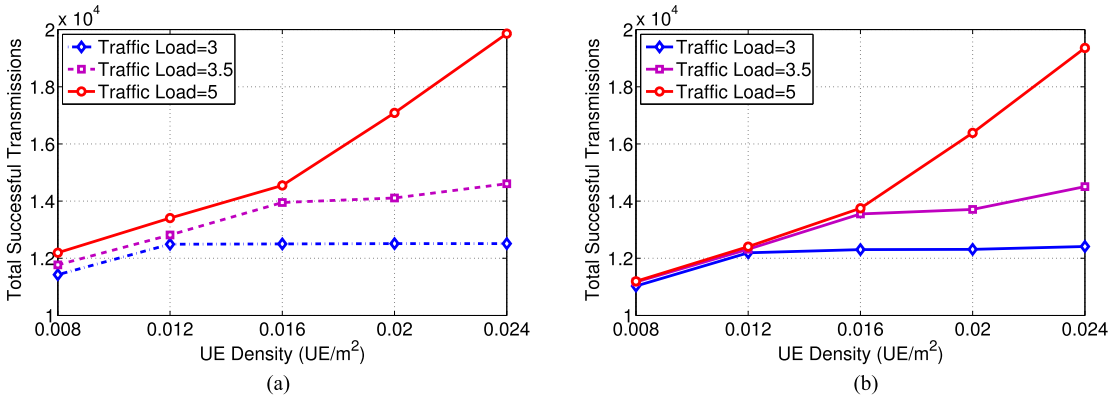

Fig. 5. Network throughput as the function of UE density attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

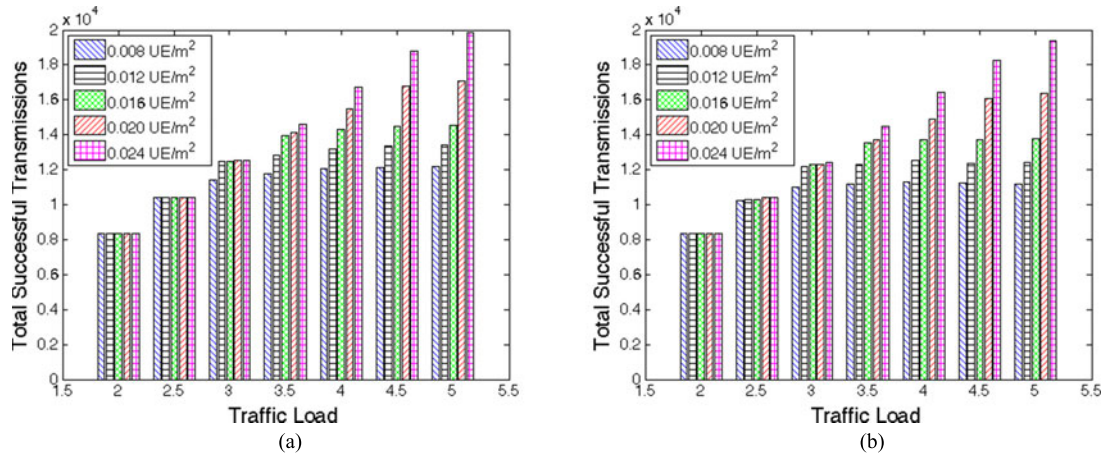

Fig. 6. Network throughput as the function of traffic load attained by the D3MAC, given different UE densities. (a) PP traffic. (b) IPP traffic.

loads. This is because in a heavily loaded system, the network may originally be unable to transmit all packets within the delay threshold. By increasing the UE density, these originally unsuccessful transmissions can be transmitted successfully through more and better D2D links. Under light loads, on the other hand, the network is originally able to transmit almost all packets within the delay threshold, and there is less need to relying on the improvement brought by high UE density for increasing successful transmissions. Thus, the UE density has less impact on light-load networks.

Fig. 7 shows the average transmission delays as the function of the UE density obtained by the D3MA with three different traffic loads, while Fig. 8 depicts the average transmission delays as the function of the traffic load with five different UE densities. We can clearly see that the transmission delay is reduced as the UE density increases, while the transmission delay increases with the network traffic load. We also observe that the average transmission delay performance of the IPP-traffic network is worse than that of the PP-traffic network, since the IPP traffic has lower arriving stability than the PP traffic, and this erratically arriving traffic demand causes longer waiting time at the transmitting devices' queues and hence longer transmission delay.

The reason why the transmission delay decreases as the UE density increases is simple. With the increase of the UE density, the transmission opportunities increases, and this reduces the waiting time of the packets to be transmitted. In addition, a higher UE density also improves the channel quality, which improves the successful transmission rate and reduces retransmissions. This also helps improve the transmission delay performance. The reason why increasing the traffic load worsens the transmission delay performance is also obvious. Increasing the network traffic load simply means more packets to be transmitted, which increases the packets' waiting time in the transmission queues of the transmitting devices, leading to a higher transmission delay.

In the above analysis, 20 out of the 30 flows are between UEs, and the rest are flows between UEs and the gateway. In order to investigate the impact of portions of traffic between UEs, we plot the throughput and delay of D3MAC under different portions of flows between UEs in Fig. 9. The results are obtained under PP traffic. The traffic load is set to 5, and the user densities are set to 0.008, 0.016, and 0.024 UE/m^2. We can observe that increasing the portions of traffic between UEs actually benefits D3MACs throughput and delay performance under all three user densities we set. However, the impact of traffic portion is not

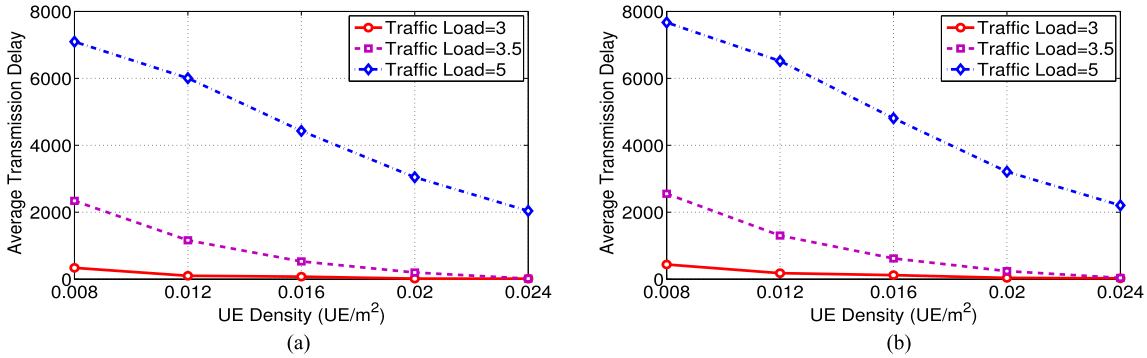

Fig. 7. Average transmission delay as the function of UE density attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

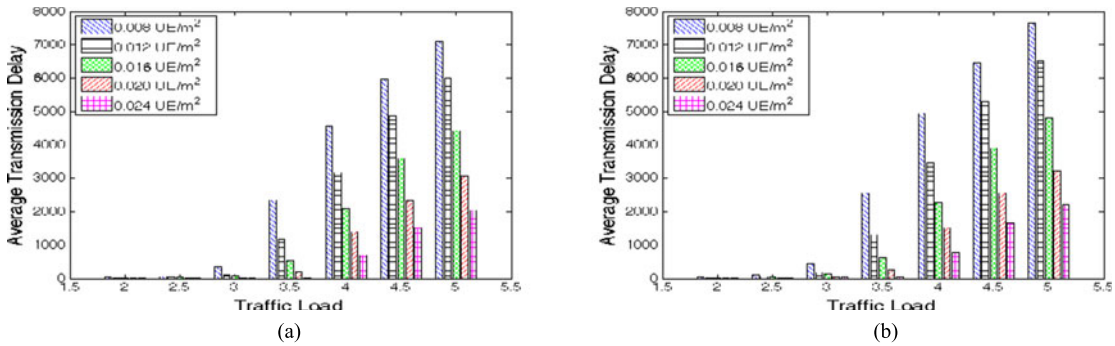

Fig. 8. Average transmission delay as the function of traffic load attained by the D3MAC, given different UE densities. (a) Poisson traffic. (b) IPP traffic.

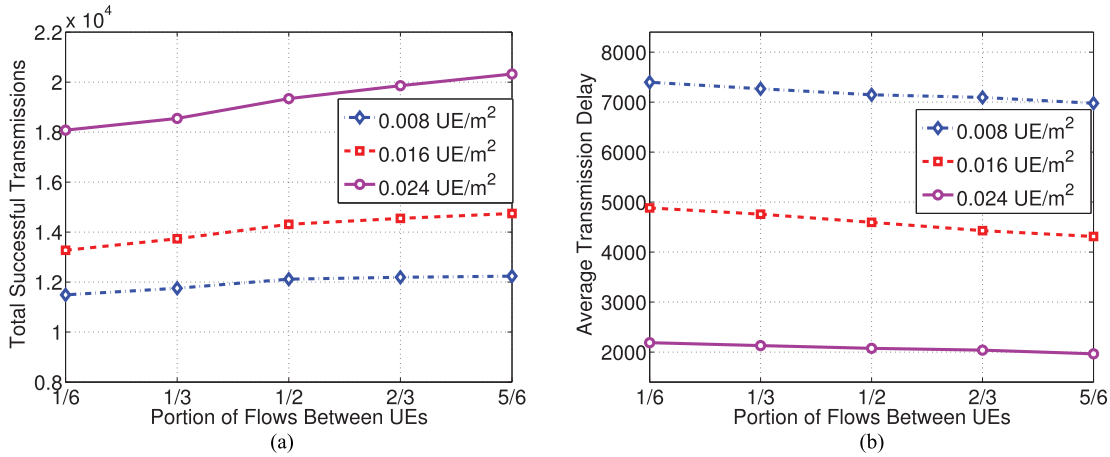

Fig. 9. Throughput and delay performance of D3MAC under different portions of traffic between UEs. (a) Throughput performance. (b) Delay performance.

significant. For example, under the user density of 0.016/m^2, the throughput under of flows between UEs is increased by 11.0% compared with that under of flows between UEs, while the delay is decreased by 13.4%.

2) Influence of UE Denseness: The denseness of UEs is another important factor that affects the performance of D2D communication. With a fixed UE density, high UE denseness indicates that UEs are distributed unevenly, and low denseness means that UEs are near-uniformly distributed. In order to find out how the UE denseness impacts on D2D communication, we deploy the same simulated network with the UE density fixed to 0.016 UE/m^2. However, in each small cell, the locations of UEs follow the 2-D normal distribution with the mean at the cell center and the standard deviation ς per dimension. The denseness of UEs can be adjusted by changing the value of ς, where a small ς indicates a large UE denseness, and vice versa. We set the levels of UE denseness from 1 to 5, which correspond to the values of ς equal to 20, 18, 16, 14, 12, and 10.

Figs. 10 and 11 show that the network throughput and average transmission delay achieved by the D3MAC as the functions of

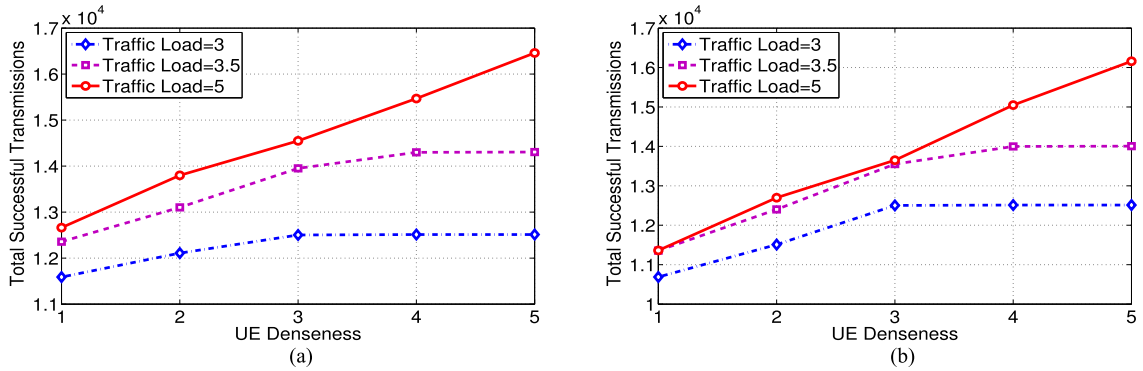

Fig. 10. Network throughput as the function of UE denseness attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

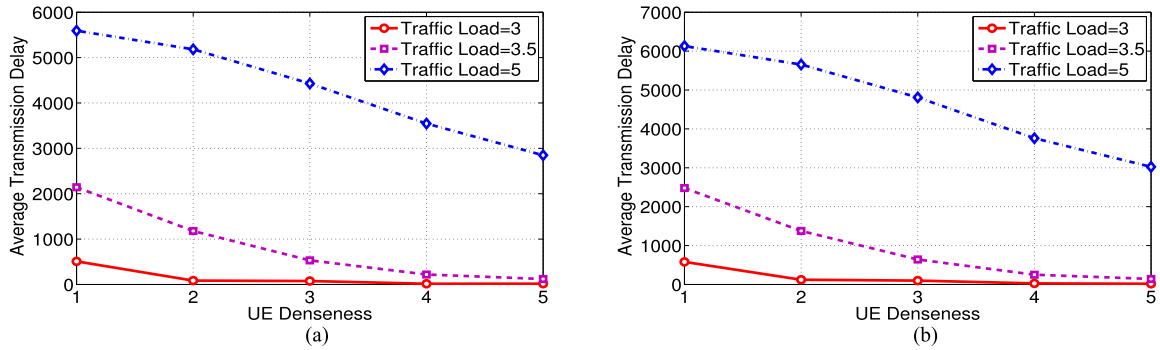

Fig. 11. Average transmission delay as the function of UE denseness attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

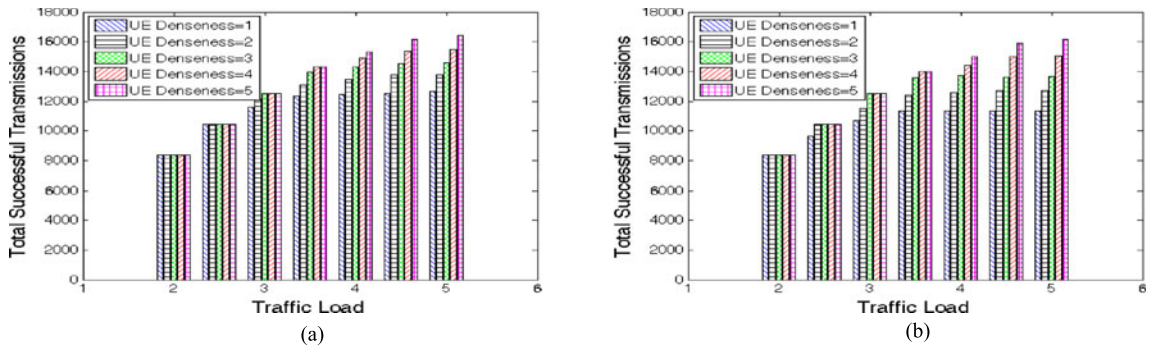

Fig. 12. Network throughput as the function of traffic load attained by the D3MAC, given different levels of UE denseness. (a) PP traffic. (b) IPP traffic.

UE denseness, respectively, given three different traffic loads. The results of Fig. 10 indicate that the throughput increases with UE denseness. Given the UE density, the distances between UEs in high denseness systems become shorter, and therefore, the channel qualities are improved. However, for light-load systems, the throughput stops increasing after the UE denseness reaches a certain level. By contrast, for heavy-load networks, the throughput keeps increasing with a similar rate. Under heavy loads, the traffic demand is beyond the transmitting capability of the original system. Thus, higher UE denseness improves the performance by increasing transmission rates. On the other hand, in a lightly loaded system, the traffic demand may almost be met by the transmitting capability of the original system, and there is less need to rely on the increase of UE denseness for improving transmission rates. Similarly, the average transmission delay is reduced as the UE denseness increases, due to the increase in the transmission rates. Furthermore, the transmission delay decreases more rapidly, as the UE denseness increases, under heavy loads. This rapid reduction in transmission delay as the UE denseness increases under heavy loads is mainly due to the large reduction of the packets' waiting time. By contrast, such an effect is small and not so obvious under light loads. The results of Figs. 10 and 11 also show that the performance of the PP-traffic network is better than that of the IPP-traffic network.

The complicated coupling influences of the traffic load and UE denseness on the performance of D2D communication are

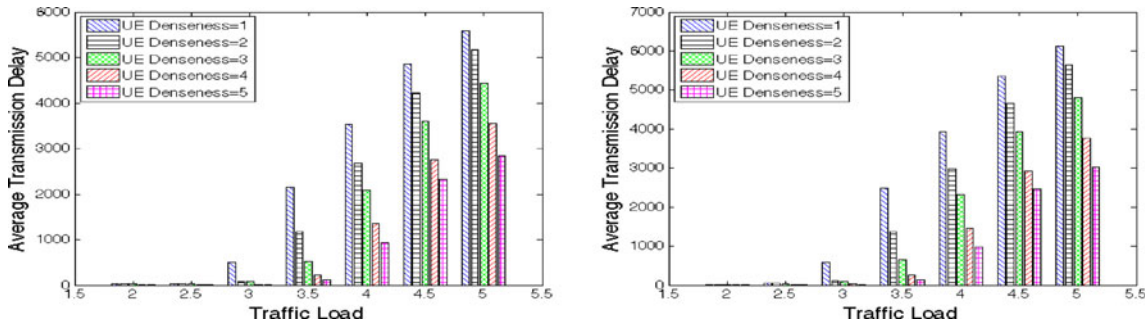

Fig. 13. Average Transmission delay as the function of traffic load attained by the D3MAC, given different levels of UE denseness. (a) PP traffic. (b) IPP traffic.

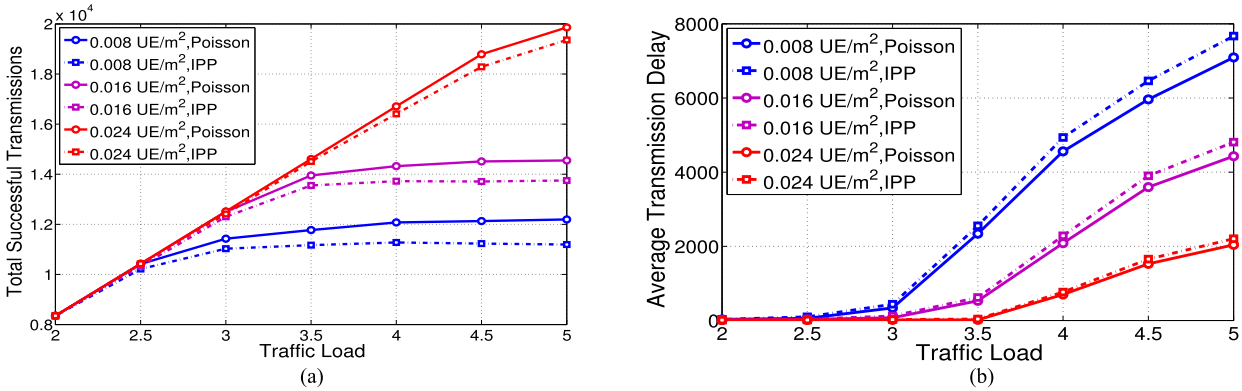

Fig. 14. Network performance as the function of traffic load under PP and IPP traffics and with different UE densities. (a) Network throughput. (b) Average transmission delay.

further illustrated in Figs. 12 and 13, which depict the network throughput and average transmission delay as the functions of the traffic load, respectively, with different levels of UE denseness. Again, we observe that under light loads, the average transmission delay is very small. When the traffic load exceeds a certain value, which is different for different levels of UE denseness, the transmission delay begins increasing rapidly. Also, under light loads, the network throughput increases with the traffic load. When the traffic load exceeds a certain value, which again is different for different levels of UE denseness, the throughput grows slower and eventually becomes saturated.

3) Influence of Traffic Mode: As have shown about, the performance of the PP-traffic network is better than that of the IPP-traffic network, because the IPP packets have lower arriving stability than the PP packets. We now have a closer look into the impact of the traffic mode as well as its coupling influence with other network key parameters. Specifically, Fig. 14 shows the network performance as the function of the traffic loads under PP and IPP traffics and with different UE densities, while Fig. 15 illustrates the network performance as the function of traffic load under PP and IPP traffics and with different levels of UE denseness. We can clearly observe that the PP-traffic system outperforms the IPP-traffic system in terms of achieving higher throughput and lower transmission delay. Moreover, the impact caused by traffic mode varies with traffic load, UE density, and UE denseness. The performance gap between the two systems is larger under heavy traffic load, low UE density, and low UE denseness.

VII. Dynamic Network Evaluation and Analysis

The distribution of UEs is usually dynamic in real-world networks, and the dynamic changes of the network topology have profound impacts on the achievable performance. In order to evaluate the impact of user mobility on the performance of D2D communication, we adopt the realistic human mobility model SLAW proposed by Brockmann *et al.* [28] in our simulation, where the distribution of user traveling distance decays at a power law with the parameter λ_{decay}, and the probability of user remaining in a small spatially confined region for a time period T_{stay} is dominated by algebraically long tails that attenuate with the super diffusive spread. UEs are initially distributed uniformly with the UE density equal to 0.016 UE/m^2, and their mobility traces are generated by the random walk model. Other simulated network parameters remain the same as before. Based on the scale of the simulated network, we set $\lambda_{\mathrm{decay}} = 0.2$, which makes the average moving distance of UE 5 m, and the mean value of T_{stay} is set to 1 s. In order to schedule the system with the D3MAC, we consider the network topology and UE locations static within a rather short duration of time, which is 0.1 s in the simulation, and the transmission rates of all links are updated every 0.1 s. Since the speed of UE is one of the key

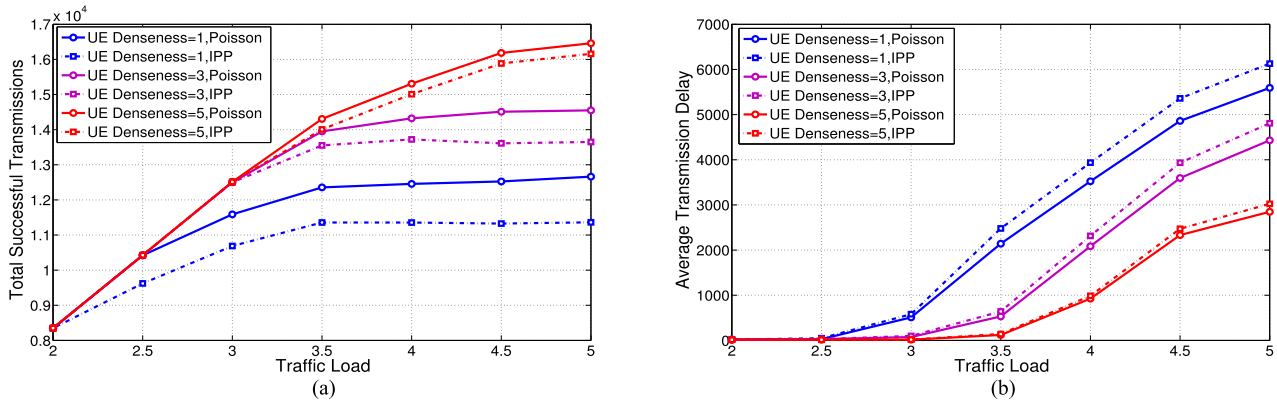

Fig. 15. Network performance as the function of traffic load under PP and IPP traffics and with different levels of UE denseness. (a) Network throughtput. (b) Average transmission delay.

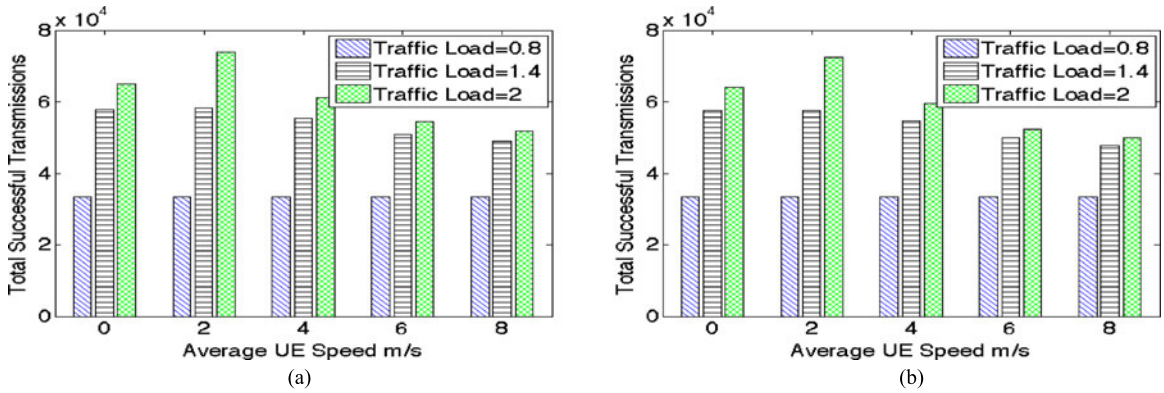

Fig. 16. Network throughput as the function of average UE velocity attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

factors that determine the network topology, we evaluate the throughput and delay performance of the system under different UE velocities to unveil the impact of UE mobility.

Fig. 16 depicts the network throughput as the function of average UE speed, given three different traffic loads. Observe that the impact of UE mobility on the achievable network throughput is heavily influenced by the network traffic load. Specifically, with a very light traffic load of 0.8, the traffic demand can easily be met, and the UE mobility appears to have no impact on the achievable network throughput. With the traffic load of 1.4, increasing the average UE speed from 0 to 2 m/s has a little impact on the network throughput, but the network throughput begins to drop when the average UE speed further increases. By contrast, for the case of the traffic load equal to 2, the achievable system throughput actually increases considerably as the average UE speed increases from 0 to 2 m/s. The system with low average UE speed outperforms the network with a static UE distribution, because the mobility of UEs allows UEs to move around and, thus, to increase the probability of encountering other UEs. This enables more new D2D links and, therefore, improves the system performance. However, the network throughput begins to drop rapidly as the UE speed further increases. The frequent changes of the network topology brought by high UE speed apparently changes D2D pairs too frequently. As a consequence, many D2D links are frequently disabled before the transmission on them is completed, which has detrimental effects on the D2D communication performance.

Fig. 17 plots the average transmission delay as the function of UE mobility, given three different traffic loads. We observe that the transmission delay first decreases when the average UE speed increases from 0 to 2 m/s. The improvement in delay performance at low UE mobility over the static network is caused by newly established D2D links brought by UE mobility. However, as the average UE speed increases further, the transmission delay begins to grow rapidly, especially under heavy traffic loads. This is because high UE mobility causes frequent changes of UE locations, which frequently disables D2D links before the transmissions on them are completed. These unsuccessful transmissions then need to be rescheduled, leading to the increase of transmission delay. This detrimental effect is further amplified by heavy traffic loads.

From the above results and analysis, we can draw the general conclusion that low UE speed improves the performance of D2D communication, in terms of throughput and transmission delay, due to the establishment of new D2D links. When the average UE speed exceeds a certain value, the system performance suffers considerably owing to the frequent change of network topology, which results in highly unstable D2D pairs.

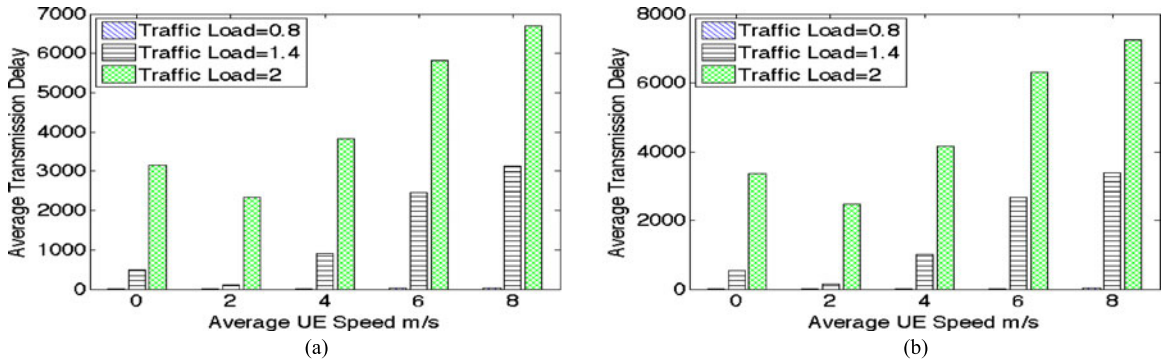

Fig. 17. Average transmission delay as the function of average UE velocity attained by the D3MAC, given different traffic loads. (a) PP traffic. (b) IPP traffic.

This detrimental impact of high UE speed is further amplified under heavy traffic loads.

VIII. Conclusion

In this paper, we have investigated the optimal scheduling problem of radio access of small cells in the mmWave band that enables D2D communications and fully utilizes the spatial reuse. Because the optimal solution for this scheduling problem is computationally prohibitive for large-scale networks, we have proposed a centralized MAC scheduling scheme, referred to as the D3MAC, to solve this scheduling problem efficiently. Through extensive simulations, we have demonstrated that the D3MAC achieves a near-optimal performance, in terms of network throughput and transmission delay, and it outperforms other existing protocols. Our other main contribution has included the comprehensive evaluation of how the user behavior impacts the achievable D2D communication performance in the network consisting of mmWave small cells, based on our proposed D3MAC protocol. Specifically, we have investigated the five key factors related to user behaviors, namely, UE density, UE denseness, UE mobility, traffic load, and traffic mode. Our study has unveiled intriguing patterns and complex interactions of these factors in influencing the achievable performance of D2D communications in mmWave small cells. Our results and observations thus offer valuable lessons and useful guidelines in designing future networks of mmWave-based small cells.

References

[1] J. Andrews, "How can cellular networks handle 1000x the data?" *Technical Talk*, Univ. Notre Dame, Notre Dame, IN, USA, 2011.

[2] G. Fodor *et al.*, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[3] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.

[4] M. N. Islam, A. Sampath, A. Maharshi, O. Koymen, and N. B. Mandayam, "Wireless backhaul node placement for small cell networks," in *Proc. 48th Annu. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 19–21, 2014, pp. 1–6.

[5] 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.

[6] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[7] L. Wei, R. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[8] S. Scott-Hayward and E. Garcia-Palacios, "Multimedia resource allocation in mmwave 5G networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 240–247, Jan. 2015.

[9] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[10] Y. Zhu *et al.*, "Demystifying 60 GHz outdoor picocells," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Network.*, Maui, HI, USA, Sep. 7–11, 2014, pp. 5–16.

[11] 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.

[12] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "On the performance of device-to-device underlay communication with simple power control," in *Proc. IEEE 69th Veh. Technol. Conf.*, Barcelona, Spain, Apr. 26–29, 2009, pp. 1–5.

[13] C. Xu *et al.*, "Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 348–358, Sep. 2013.

[14] P. Janis, V. Koivunen, C. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proc. IEEE 69th Veh. Technol. Conf.*, Barcelona, Spain, Apr. 26–29, 2009, pp. 1–5.

[15] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *Proc. IEEE 72nd Veh. Technol. Conf.*, Ottawa, ON, Canada, Sep. 6–9, 2010, pp. 1–5.

[16] S. Hakola, T. Chen, J. Lehtomaki, and T. Koskela, "Device-to-device (D2D) communication in cellular network—Performance analysis of optimum and practical communication mode selection," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Sydney, Australia, Apr. 18–21, 2010, pp. 1–6.

[17] M. Belleschi, G. Fodor, and A. Abrardo, "Performance analysis of a distributed resource allocation scheme for D2D communications," in *Proc. GLOBECOM Workshops*, Houston, TX, USA, Dec. 5–9, 2011, pp. 358–362.

[18] 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 multi hop MAC design," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1400–1413, Oct. 2009.

[19] Y. Niu, Y. Li, and D. Jin, "Poster: Promoting the spatial reuse of millimeter wave networks via software-defined cross-layer design," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, Maui, HI, USA, Sep. 7–11, 2014, pp. 395–396.

[20] J. Ning, T.-S. Kim, S. V. Krishnamurthy, and C. Cordeiro, "Directional neighbor discovery in 60 GHz indoor wireless networks," in *Proc. 12th ACM Int. Conf. Model., Anal. Simul. Wireless Mobile Syst.*, Tenerife, Spain, Oct. 26–30, 2009, pp. 365–373.

[21] 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.
[22] 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.
[23] 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. IEEE Int. Conf. Commun.*, Ottawa, ON, Canada, Jun. 10–15, 2012, pp. 5221–5225.
[24] 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. 25–30, 2012, pp. 2149–2157.
[25] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. IEEE Int. Conf. Robot. Autom.*, Taipei, China, Sep. 4, 2004, pp. 284–289.
[26] A. Lebedev *et al.*, "Feasibility study and experimental verification of simplified fiber-supported 60-GHz picocell mobile backhaul links," *IEEE Photon. J.*, vol. 5, no. 4, pp. 1–14, Aug. 2013.
[27] D. Bojic *et al.*, "Advanced wireless and optical technologies for small-cell mobile backhaul with dynamic software-defined management," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 86–93, Sep. 2013.
[28] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.
[29] S. Sur, X. Zhang, P. Ramanathan, and R. Chandra, "BeamSpy: Enabling robust 60 GHz links under blockage," in *Proc. 13th Usenix Conf. Netw. Syst. Des. Implementation*, Santa Clara, CA, USA, Mar. 16–18, 2016, pp. 193–206.
[30] S. Sur, V. Venkateswaran, X. Zhang, and P. Ramanathan, "60 GHz indoor networking through flexible beams: A link-level profiling," in *Proc. ACM SIGMETRICS Int. Conf. Meas, Model. Comput. Syst.*, Portland, OR, USA, Jun. 15–19, 2015, pp. 71–84.

Authors' photographs and biographies not available at the time of publication.