Exploiting Device-to-Device Communications to Enhance Spatial Reuse for Popular Content Downloading in Directional mmWave Small Cells

Yong Niu, Li Su, Chuhan Gao, Yong Li, *Member, IEEE,* Depeng Jin, and Zhu Han, *Fellow, IEEE*

Abstract—With the explosive growth of mobile demand, small cells in millimeter wave (mmWave) bands underlying the macro-cell networks have attracted intense interest from both academia and industry. MmWave communications in the 60 GHz band are able to utilize the huge unlicensed bandwidth to provide multiple Gbps transmission rates. In this case, device-to-device (D2D) communications in mmWave bands should be fully exploited due to no interference with the macrocell networks and higher achievable transmission rates. In addition, due to less interference by directional transmission, multiple links including D2D links can be scheduled for concurrent transmissions (spatial reuse). With the popularity of content-based mobile applications, popular content downloading in the small cells needs to be optimized to improve network performance and enhance user experience. In this paper, we develop an efficient scheduling scheme for popular content downloading in mmWave small cells, termed PCDS (popular content downloading scheduling), where both D2D communications in close proximity and concurrent transmissions are exploited to improve transmission efficiency. In PCDS, a transmission path selection algorithm is designed to establish multi-hop transmission paths for users, aiming at better utilization of D2D communications and concurrent transmissions. After transmission path selection, a concurrent transmission scheduling algorithm is designed to maximize the spatial reuse gain. Through extensive simulations under various traffic patterns, we demonstrate PCDS achieves near-optimal performance in terms of delay and throughput, and also superior performance compared with other existing protocols, especially under heavy load. The impact of the maximum number of hops of transmission paths on its performance is also analyzed for a better understanding of the role of D2D communications.

I. INTRODUCTION

Mobile data is growing explosively. Some industry and academic experts predict a 1000-fold demand increase by 2020 [1]. In order to meet such sharp growth, there is increasing interest in deploying small cells in higher frequency bands, such as the millimeter wave (mmWave) bands between 30 and 300 GHz, underlying the conventional macrocell networks to significantly boost the network capacity. This deployment is usually referred to as heterogeneous cellular networks (HCNs). With huge unlicensed bandwidth (e.g., the 7 GHz spectrum between 57 GHz and 64 GHz approved by the Federal Communications Commission), small cells in the 60 GHz band have gained considerable attention from academia, industry, and standardization bodies. 60 GHz communications enable multi-gigabit data rates, and broadband applications like high-speed data transfer between devices (e.g., cameras, ipads, tablets, and notebooks), real-time streaming of both compressed and uncompressed high definition television (HDTV), wireless gigabit ethernet, and wireless gaming can be supported. Recently, rapid progress in 60 GHz mmWave circuits, including on-chip and in-package antennas, radio frequency power amplifiers, low-noise amplifiers, voltage-controlled oscillators, mixers, and analog-to-digital converters has paved the way for more cost-effective devices in the 60 GHz band [2], [3], [4]. Several standards have been defined for indoor wireless personal area networks (WPAN) or wireless local area networks (WLAN), for example, ECMA-387 [5] , IEEE 802.15.3c [6], and IEEE 802.11ad [7].

However, to popularize and standardize products in the 60 GHz band worldwide, several technical challenges need to be addressed. Due to high carrier frequency, 60 GHz communications suffer from high propagation loss. The free space loss in the 60 GHz band is 21.6 dB worse than 5 GHz for omnidirectional communications. Thus, directional antennas should be synthesized at both transmitter and receiver to form directional high gain beams to combat the significant propagation loss [8], [9], [10]. However, with directional transmissions, the third party nodes cannot perform carrier sensing to avoid contention with the current transmissions, which is referred to as "deafness" [11]. In addition, mmWave links are also vulnerable to blockage by obstacles such as humans and furniture due to weak diffraction ability [12]. Blockage by a human penalizes the link budget by 20-30 dB.

On the other hand, content popularity in mobile networks has been found to follow the classic Zipf's law [13]. It is demonstrated that a small amount of content accounts for the majority of requests, which is very popular among the majority of users. In the 60 GHz small cells, popular content downloading is widely used in many cases, such as the Group/Broadcast communication services like police, fire, and ambulance in public safety networks, device discovery, and advertising messages broadcasting [14]. With directional transmissions, the wireless broadcast channel in 3G/4G networks is not feasible in the 60 GHz band [13], [15]. At the same

Y. Niu, L. Su, Chuhan Gao, Y. Li, D. Jin are with State Key Laboratory on Microwave and Digital Communications, Tsinghua National Laboratory for Information Science and Technology (TNLIST), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mails: liyong07@tsinghua.edu.cn).

Z. Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA (e-mail: zhan2@uh.edu).

time, in the user-intensive region, there is a high probability that two user devices are located near to each other. In this case, device-to-device (D2D) communications in physical proximity can be exploited for content downloading as well as saving power and improving the spectral efficiency [16]. D2D communications in the same carrier frequencies as today's cellular systems have significant interference to the cellular users, which limits the benefits of D2D communications. However, for D2D communications in the mmWave bands, there will be no interference with the cellular systems, and the higher achievable transmission rates in mmWave bands also further increase the benefits of D2D communications. Multi-hop D2D communications in mmWave bands can also be utilized to overcome blockage by obstacles [12], [17]. In addition, in the directional communication scenario, there is less interference between links, and concurrent transmissions are enabled to unleash the potential of spatial reuse. With these fundamental differences between mmWave communications and existing systems using lower carrier frequencies (e.g., from 900 MHz to 5 GHz), new scheduling schemes or MAC protocols for popular content downloading are needed to fully reap the benefits of D2D communications in mmWave bands.

In this paper, we propose an efficient scheduling scheme for popular content downloading in mmWave small cells, termed PCDS, where D2D communications and concurrent transmissions are fully exploited to improve transmission efficiency. In PCDS, users far from the access point (AP) receive the popular content from neighboring users in close proximity that have received the content. Meanwhile, concurrent link transmissions are exploited to significantly improve network capacity. The contributions of this paper are four-fold, which are summarized as follows.

- We design a transmission path selection algorithm to establish multi-hop transmission paths for users, aiming at exploiting better channel conditions (higher transmission rate) between nodes in close proximity to improve transmission efficiency. In addition, the better use of concurrent transmissions (spatial reuse) is also considered in the algorithm to enhance scheduling efficiency.
- After transmission path selection, we formulate the optimal multi-hop transmission scheduling problem into a mixed integer linear programming (MILP), i.e., to minimize the number of time slots to send the popular content from the AP to all users. Concurrent transmissions are explicitly considered under the signal to interference plus noise ratio (SINR) interference model in this formulated problem.
- We propose an efficient and practical concurrent transmission scheduling algorithm to solve the formulated NP-hard problem with low complexity.
- Extensive simulations under various traffic patterns are carried out to demonstrate near-optimal performance of PCDS in terms of delay and throughput, and the superior network performance of PCDS compared with other schemes. In addition, we also analyze the impact of the maximum number of hops in the transmission path selection algorithm on the performance of PCDS, which

provides references for the choice of this parameter in practice.

The rest of this paper is organized as follows. The related work on directional MAC protocols for small cells in the 60 GHz band is introduced and discussed in Section II. Section III introduces the system model, and illustrates the procedure and problems of PCDS. Section IV presents the proposed transmission path selection algorithm to fully exploit D2D transmission and spatial reuse. After transmission path selection, we formulate the problem, and propose a concurrent transmission scheduling algorithm in Section V. Section VI presents the performance evaluation of PCDS under various traffic patterns. Finally, we conclude this paper in Section VII.

II. RELATED WORK

To keep up with surging demand, small cells densely deployed underlying the conventional homogeneous macrocell network have been proposed to create the dual benefits of higher quality links and more spatial reuse [18], [19]. However, reducing the radii of small cells in the same carrier frequencies as today's cellular systems to reap the spatial reuse benefits is fundamentally limited by interference constraints [20]. By using higher frequency bands, such as the millimeter wave (mmWave) bands between 30 and 300 GHz, small cells can significantly boost the overall network capacity due to less interference with macrocells and higher achievable data rates [18], [20].

There has been some related work on directional MAC protocols for small cells in the 60 GHz band [21], [22], [23], [24], [25]. Some work is based on TDMA [5], [6]. Cai *et al.* [22] derived exclusive region (ER) conditions that concurrent transmissions always outperform TDMA. In two protocols [23], [24] based on IEEE 802.15.3c, concurrent transmissions are enabled if the multi-user interference (MUI) is below a specific threshold. Qiao *et al.* [21] proposed a concurrent transmission scheduling algorithm for an indoor IEEE 802.15.3c WPAN, where concurrent transmissions are optimized to maximize the number of flows with the quality of service requirement of each flow satisfied. Qiao *et al.* [25] also proposed a multi-hop concurrent transmission scheme to address the link outage problem and combat huge path loss to improve flow throughput. For the TDMA based protocols, unfair medium time allocation problem exists for individual users under bursty data traffic [15].

There are also other centralized work on MAC protocols for small cells in the 60 GHz band. Gong *et al.* [26] proposed a directive CSMA/CA protocol, which solves the deafness problem by the virtual carrier sensing. However, the spatial reuse is not considered. In the multihop relay directional MAC (MRDMAC), relay paths are established to steer around obstacles [12]. However, since most transmissions are through the piconet coordinator (PNC), concurrent transmissions are also not considered in MRDMAC. Chen *et al.* [27] proposed a spatial reuse strategy for an IEEE 802.11 ad WPAN, where two different service periods (SPs) are scheduled to overlap with each other. Since only two links are considered for concurrent transmissions, the spatial reuse is not fully

exploited. Son *et al.* [15] proposed a frame based directional MAC protocol (FDMAC), which amortizes the scheduling overhead over multiple concurrent transmissions in a row to achieve high efficiency. The core of FDMAC is the Greedy Coloring (GC) algorithm, which fully exploits spatial reuse and greatly improves the network throughput. FDMAC also has a good fairness performance and low complexity. Chen *et al.* [28] proposed a directional cooperative MAC protocol (D-CoopMAC) to coordinate the uplink channel access in an IEEE 802.11ad WLAN. In D-CoopMAC, a two-hop relay path of high channel quality from the source to the destination is established to replace the direct path of poor channel quality for higher transmission efficiency. Niu *et al.* [17] proposed a blockage robust and efficient directional MAC protocol (BRDMAC), which overcomes the blockage problem by two-hop relaying. In BRDMAC, relay selection and spatial reuse are jointly optimized to achieve robust network connectivity and also improve network performance. Niu *et al.* [29] also proposed a channel transmission rate aware directional MAC protocol, RDMAC, where both the multirate capability of links and spatial reuse are exploited to improve network performance. There are two stages in RDMAC. The first stage measures the channel transmission rates of links by a heuristic algorithm, which can compute near-optimal measurement schedules with respect to the total number of measurements. The second stage accommodates the traffic demand of links by a heuristic transmission scheduling algorithm, which can compute near-optimal transmission schedules with respect to the total transmission time. Recently, Niu *et al.* [30] proposed a joint transmission scheduling scheme for the radio access and backhaul of small cells in 60 GHz band, termed D2DMAC, where a path selection criterion is designed to enable D2D transmissions for performance improvement.

There are also distributed MAC protocols for small cells in the 60 GHz band [31], [32]. The memory-guided directional MAC (MDMAC) alleviates the deafness problem by incorporating a Markov state transition diagram, and employs memory to achieve approximate time division multiplexed (TDM) schedules [31]. The directional-to-directional MAC (DtDMAC) solves the asymmetry-in-gain problem with both senders and receivers operate in a directional-only mode [32]. DtDMAC adopts an exponential backoff procedure for asynchronous operations, and the deafness problem is also alleviated by a Markov state transition diagram.

To the best of our knowledge, none of the work above considers the content popularity in the problem, and we try to exploit the D2D transmission in close proximity to enhance spatial reuse for popular content downloading in the 60 GHz small cells.

III. SYSTEM OVERVIEW

A. System Model

We consider a mmWave small cell of n nodes, one of which is the AP, and the rest are users (UEs). The system time is partitioned into non-overlapping time slots of equal length, and the AP synchronizes the clocks of UEs and schedules the medium access of all nodes to accommodate their traffic

demands. With electronically steerable directional antennas equipped at the AP and UEs, directional transmissions are supported between any pair of nodes. In addition, a bootstrapping program is run in the system such that the AP knows the up-to-date network topology and the location information of other nodes [33], [34]. The network topology can be obtained by the neighbor discovery schemes in [33]. Location information can be obtained based on wireless channel signatures, such as angle of arrival, time difference of arrival, or the received signal strength [34]. We also assume all nodes are half-duplex, and each node has at most one connection with one neighbor simultaneously.

For small cells in the 60 GHz band, non-line-of-sight (NLOS) transmissions suffer from higher attenuation than line-of-sight (LOS) transmissions [35], [36], [37]. In Ref. [35], the path loss exponent in the LOS hall is 2.17, while the path loss exponent in the NLOS hall is 3.01. If the distance between the transmitter and the receiver is 10 m, the gap in the path loss is about 10 dB for the LOS hall and the NLOS hall. When operating in a power-limited regime, a 10 dB power loss requires a 10-fold reduction in transmission rate to maintain the same reliability. On the other hand, restricting to the LOS path can maximize the power efficiency since the LOS path is strongest. In addition, NLOS transmissions in the 60 GHz band also suffer from a shortage of multipath, and it is reasonably accurate to calculate link budget for a directional LOS link based on the simple additive white Gaussian noise channel model [12]. Thus, we only consider the Line-of-Sight (LOS) transmissions to achieve high transmission rates and maximize power efficiency [12]. In [12], it has been demonstrated that relaying using LOS links is able to obtain robust network connectivity and high throughput, and multiple schemes have been proposed to overcome blockage by multi-hop relaying [17], [38].

We denote the directional link from node i to j by (i, j), and assume the beamforming process between node i and j has been completed. According to the path loss model, the received power at node j for (i, j), P_{ij}^r (mW) can be estimated by

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

where P_t (mW) denotes the transmission power, k_0 is a constant coefficient and proportional to $(\frac{\lambda}{4\pi})^2$ (λ denotes the wavelength), $G_t(i, j)$ denotes the transmit antenna gain of node i in the direction from node i to j, $G_r(i, j)$ denotes the receive antenna gain of node j in the direction from node i to j, l_{ij} (m) denotes the distance between transmitter i and receiver j, and τ denotes the path loss exponent [21].

Due to directional transmissions, there is less interference between links, and concurrent transmissions (spatial reuse) are enabled to improve network capacity. However, due to the limited communication range, the interference between links cannot be neglected. In this paper, we adopt the interference model in [21]. Then for link (u, v) and (i, j), the received interference power at node j from node u can be calculated as

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

where ρ denotes the MUI factor related to the cross correlation

of signals from different links [21]. If we denote the set of links that transmit concurrently with link (i, j) by $\mathbb{C}_{ij}$, then the interference power I_{ij} can be calculated as

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

Since each node has at most one connection with one neighbor, adjacent links cannot be scheduled concurrently [15]. Thus, $\mathbb{C}_{ij}$ should not include links that are adjacent to link (i, j). The received SINR at receiver j can be calculated as

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

where W (Hz) is the bandwidth, and N_0 (mW/Hz) is the one-sided power spectra density of white Gaussian noise [21].

For each link (i, j), we denote the minimum SINR to support its transmission rate c_{ij} by $\gamma(c_{ij})$. Therefore, link (i, j)'s SINR Γ_{ij} should be greater than or equal to $\gamma(c_{ij})$ to support its concurrent transmissions with other links.

Due to the difference in distance between the transmitter and receiver, the channel transmission rates of different links vary significantly. We denote the $n \times n$ channel transmission rate matrix by $\mathbf{C}$, whose (i, j) element is denoted by c_{ij}. c_{ij} indicates the channel transmission rate of link (i, j), and numerically it is equal to the number of packets that link (i, j) can transmit in one time slot. We assume that a channel transmission rate measurement procedure executes in the system to update the channel transmission rate matrix [29]. This process is summarized as follows: the transmitter of each link firstly transmits measurement packets to the receiver. Then with the measured signal to noise ratio (SNR), the receiver obtains the achievable transmission rate and appropriate modulation and coding scheme (MCS) according to the SNR and MCS correspondence table. Lastly, the receiver will transmit an acknowledgement packet to inform the sender of the transmission rate and MCS. Under low user mobility, the procedure is executed periodically, and the measurement results will be reported to the AP.

In this paper, we consider the popular content downloading traffic from the AP to all users. Traditionally, the traffic is distributed to each UE from the AP one by one in sequence. In PCDS, both D2D transmissions in close proximity and concurrent transmissions of links are exploited to improve transmission efficiency significantly. In Fig. 1 (a), we give a time-line illustration of PCDS, where there are six UEs and an AP. In PCDS, time is divided into a sequence of non-overlapping frames [15], and there are two phases in each frame, the scheduling phase and transmission phase. In the scheduling phase, the AP obtains the packets to be downloaded by users from the network layer, which takes time t_d; then the AP computes the content downloading source for each UE, and calculates a schedule to distribute these packets to all UEs, which takes time t_{sch}; lastly the AP pushes the schedule and selected downloading sources to the UEs in sequence, which takes time t_{push}. In the transmission phase, all nodes start transmissions following the schedule until the

packets are distributed to all UEs. There are multiple pairings in the transmission phase, and in each pairing, multiple links are activated simultaneously for concurrent transmissions. We show an example of PCDS operation in a small cell of six UEs in Fig. 1, and the more detailed explanation of this example can be found in Section III-B.

B. Problem Overview

To improve transmission efficiency, appropriate transmission paths from the AP to UEs need to be selected to fully exploit the potential of D2D communications, and also the benefits of concurrent transmissions. After transmission paths are selected, efficient concurrent transmission scheduling algorithm should be designed to unleash the potential of spatial reuse.
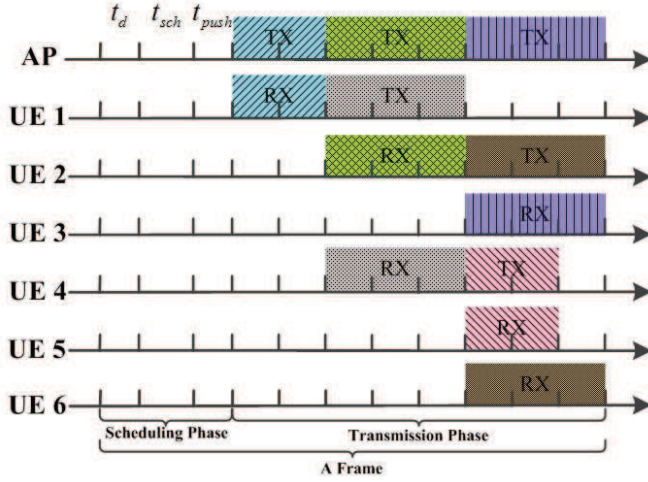
Now, we consider a small cell of an AP and six UEs, whose topology is plotted in Fig. 1 (b). We denote the content downloading traffic by d, and there are $d = 6$ packets to be distributed to all UEs. The channel transmission rate matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 1 & 2 & 2 & 1 & 3 \\ 1 & 0 & 1 & 1 & 1 & 2 & 3 \\ 1 & 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 1 & 1 \\ 2 & 1 & 1 & 3 & 0 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 0 & 1 \\ 3 & 3 & 2 & 1 & 1 & 1 & 0 \end{pmatrix}, \tag{5}$$

where the first six rows/columns are UEs, and the AP is the last row/column. We can observe that the $(7,1)$ element of $\mathbf{C}$ is 3, which means that the link from AP to UE 1 is able to transmit 3 packets in one time slot. If we select three transmission paths, AP $\rightarrow$ UE1 $\rightarrow$ UE4 $\rightarrow$ UE5, AP $\rightarrow$ UE2 $\rightarrow$ UE6, and AP $\rightarrow$ UE3 as in Fig. 1 (b), then we can obtain a schedule to complete traffic downloading of all UEs as shown in Fig. 1 (a). This schedule has three pairings, and in the first pairing, the AP transmits the packets to UE 1 for two time slots. In the second pairing, the AP transmits to UE 2, and UE1 transmits the packets to UE 4 for three time slots. In the third pairing, link AP $\rightarrow$ UE3, UE2 $\rightarrow$ UE6, and UE4 $\rightarrow$ UE5 are activated to distribute the packets to UE 3, UE 5, and UE 6. As we can see, the schedule completes traffic downloading in 8 time slots. However, if these packets are distributed from the AP to UEs one by one without exploiting the D2D communications and spatial reuse, we will need at least 25 time slots to distribute these packets to all UEs. Since selection of transmission paths has a big impact on the efficiency of spatial reuse, we should optimize the transmission paths for higher transmission efficiency. Then efficient transmission scheduling should also be investigated to fully explore the potential of spatial reuse.

IV. TRANSMISSION PATH SELECTION

In this section, we propose a heuristic transmission path selection algorithm to establish multi-hop transmission paths for better use of D2D communications and spatial reuse to improve transmission efficiency.
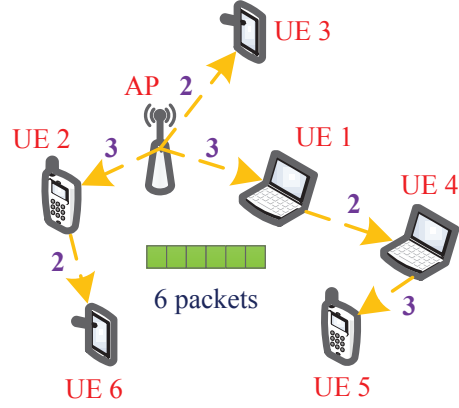
(a) Time-line illustration of PCDS operation

(b) The network topology

Fig. 1. An example of PCDS operation in a small cell of six UEs.

In the algorithm, the AP and UEs that have received the content downloading packets are scheduled to be the downloading sources of the nearest UEs to fully exploit the advantages of D2D communications. To reduce the number of adjacent links for better usage of spatial reuse in transmission scheduling, each UE is allowed to be the downloading source of one nearby UE once. The AP is allowed to be the downloading sources of multiple UEs. We denote the set of UEs in the small cell by $\mathbb{U}$. We also denote the set of UEs whose downloading sources have been selected by $\mathbb{U}_b$, and the set of UEs whose downloading sources are not selected is denoted by $\mathbb{U}_c$. When $|\mathbb{U}_b| < |\mathbb{U}_c|$, i.e., there are more UEs that are not selected than those UEs that selected, the algorithm reaches out from the nodes in $\mathbb{U}_b$ to find the UEs in $\mathbb{U}_c$ with the largest transmission rates to establish downloading links. When $|\mathbb{U}_b| \geq |\mathbb{U}_c|$, the algorithm reached out from the nodes in $\mathbb{U}_c$ to find the UEs in $\mathbb{U}_b$ with the largest transmission rates to establish downloading links.

For each UE $u \in \mathbb{U}$, we define a binary variable b_u to indicate whether the downloading source of UE u has been selected. If so, b_u is equal to 1; otherwise, b_u is equal to 0. The set of UEs whose downloading sources are selected currently is denoted by $\mathbb{U}_t$. We denote the set of the selected transmission paths by $\mathbb{P}_b$. We also denote the AP in the small cell by α. For each UE u, we define a binary variable r_u to indicate whether UE u has been the downloading source for other UEs. If UE u has been the downloading source for other UEs, $r_u = 1$; otherwise, $r_u = 0$. For the AP, r_α is set to 0. For each path $p \in \mathbb{P}_b$, its number of hops is denoted by H_p, and its last node is denoted by l_p. We also define a parameter, H_{max}, to denote the maximum possible number of hops for each path in $\mathbb{P}_b$.

The pseudo-code of the transmission path selection algorithm is presented in Algorithm 1. The algorithm iteratively schedules UEs into the transmission paths until all UEs are scheduled, as in line 3. When $|\mathbb{U}_b| < |\mathbb{U}_c|$, the algorithm extends the transmission paths in $\mathbb{P}_b$ by searching the neighbors of nodes in $\mathbb{U}_b$ with the largest transmission rates, as in lines 5–12. Lines 6–7 obtain the UE with the largest transmission rate from the AP to the UE, and generate a new path from the AP to the selected UE in $\mathbb{P}_b$. Lines 8–12 extend the paths in $\mathbb{P}_b$ to schedule the UEs in $\mathbb{U}_c$ into $\mathbb{P}_b$. All paths in $\mathbb{P}_b$ are restricted to have at most H_{max} hops, as in line 10 and 16. In line 10, the condition of $r_u = 0$ is required since each UE is allowed to be the downloading source of one neighboring UE once. In line 12, the transmission paths in $\mathbb{P}_b$ are extended to the UE with the largest transmission rate from the last node to the node in $\mathbb{U}_c$, and r_u is set to 1 since UE u is the downloading source of the selected UE. When $|\mathbb{U}_b| \geq |\mathbb{U}_c|$, the algorithm extends the transmission paths in $\mathbb{P}_b$ by searching the neighbors of nodes in $\mathbb{U}_c$ with the largest transmission rates, as in lines 14–25. In lines 14–18, the set of possible downloading sources for UEs in $\mathbb{U}_c$ is denoted by $\mathbb{U}_r$, which is obtained by checking the number of hops of paths in $\mathbb{P}_b$ and whether the UE has been the downloading source for another UE, as in line 16. In lines 19–25, the downloading source for each UE in $\mathbb{U}_c$ is selected by obtaining the node in $\mathbb{U}_r$ with the largest transmission rate to the UE in $\mathbb{U}_c$. In line 26, $\mathbb{U}_t$ (the set of UEs that have been scheduled into $\mathbb{P}_b$) is added to $\mathbb{U}_b$ (the set of UEs whose downloading sources have been selected), and removed from $\mathbb{U}_c$.

Applying the transmission path selection algorithm with $H_{max} = 3$ to the example in Section III-B, we obtain the transmission paths as path AP $\rightarrow$ UE1 $\rightarrow$ UE4 $\rightarrow$ UE5, path AP $\rightarrow$ UE2 $\rightarrow$ UE6, and path AP $\rightarrow$ UE3, which have been illustrated in Fig. 1 (b). The computational complexity of Algorithm 1 is $\mathcal{O}(|\mathbb{U}|^2)$.

V. CONCURRENT TRANSMISSION SCHEDULING

After transmission path selection, we schedule the downloading links for each UE in the transmission phase. The content downloading for all UEs needs to be completed with the minimum number of time slots to maximize transmission efficiency. In this section, we first formulate the optimal multi-hop transmission scheduling problem into a mixed integer program based on the problem formulation in FDMAC (the frame-based scheduling directional MAC protocol) [15], and then propose a practical concurrent transmission scheduling

algorithm to fully exploit spatial reuse for maximizing transmission efficiency.

A. Problem Formulation

We denote the transmission schedule in the transmission phase by $\mathbf{S}$, which has K pairings. For each UE u, we denote its downloading source by s_u, and s_u may be the AP or other UEs. The traffic demand to be distributed to all UEs is denoted by d. For each pairing, we define an $n \times (n-1)$ matrix $\mathbf{A}^k$ to indicate the links scheduled to communicate in the kth pairing. Since we focus on popular content downloading in this paper, where the traffic is from the AP to UEs, the links from users to the AP are not considered in matrix $\mathbf{A}^k$. The rows of $\mathbf{A}^k$ indicate the downloading sources of UEs, including all UEs and the AP, while the columns of $\mathbf{A}^k$ indicate all UEs. The (i, j) element of $\mathbf{A}^k$, a_{ij}^k indicates whether link (i, j) is scheduled for transmission in the kth

---

Algorithm 1: Transmission Path Selection.

---

1 **Input:** The set of UEs in the small cell $\mathbb{U}$;
2   Channel transmission rate matrix $\mathbf{C}$;
3 **Initialization:** $\mathbb{U}_b = \emptyset$; $\mathbb{U}_c = \mathbb{U}$; $\mathbb{U}_t = \emptyset$; $\mathbb{P}_b = \emptyset$; $r_\alpha = 0$;
4          $b_u = 0$ and $r_u = 0$ for each $u \in \mathbb{U}$;
5 **while** $|\mathbb{U}_c| > 0$ **do**
6    $\mathbb{U}_t = \emptyset$;
7   **if** $|\mathbb{U}_b| < |\mathbb{U}_c|$ **then**
8   Obtain UE $u \in \mathbb{U}_c$ with the largest $c_{\alpha u}$;
9      $\mathbb{P}_b = \mathbb{P}_b \cup \{\alpha \to u\}$; $b_u = 1$; $\mathbb{U}_t = \mathbb{U}_t \cup \{u\}$;
10   **for** each $u \in \mathbb{U}_b$ **do**
11   Obtain the path $p \in \mathbb{P}_b$ with $l_p = u$;
12   **if** $r_u = 0$ and $H_p < H_{max}$ **then**
13   Obtain UE $v \in \mathbb{U}_c$ with the largest c_{uv} and $b_v = 0$;
14   Extending $p \in \mathbb{P}_b$ to v; $b_v = 1$; $r_u = 1$; $\mathbb{U}_t = \mathbb{U}_t \cup \{v\}$;
15   **else**
16      $\mathbb{U}_r = \emptyset$;
17   **for** each $p \in \mathbb{P}_b$ **do**
18   **if** ($H_p < H_{\max}$ and $r_{l_p} = 0$) **then**
19          $\mathbb{U}_r = \mathbb{U}_r \cup \{l_p\}$;
20      $\mathbb{U}_r = \mathbb{U}_r \cup \{\alpha\}$;
21   **for** each $u \in \mathbb{U}_c$ **do**
22   Obtain UE $v \in \mathbb{U}_r$ with the largest c_{vu} and $r_v = 0$;
23   **if** v is α **then**
24          $\mathbb{P}_b = \mathbb{P}_b \cup \{\alpha \to u\}$; $b_u = 1$; $\mathbb{U}_t = \mathbb{U}_t \cup \{u\}$;
25   **else**
26   Obtain the path $p \in \mathbb{P}_b$ with $l_p = v$;
27   Extending $p \in \mathbb{P}_b$ to u; $b_u = 1$; $r_v = 1$; $\mathbb{U}_t = \mathbb{U}_t \cup \{u\}$;
28    $\mathbb{U}_b = \mathbb{U}_b \cup \mathbb{U}_t$; $\mathbb{U}_c = \mathbb{U}_c - \mathbb{U}_t$;
29 **Return** $\mathbb{P}_b$.

---

pairing. If link (i, j) is scheduled for transmission in the kth pairing, $a_{ij}^k = 1$; otherwise, $a_{ij}^k = 0$. We denote the number of time slots for the kth pairing by δ^k [15]. To maximize transmission efficiency, the transmission schedule should complete the traffic downloading for all UEs with a minimum number of time slots. Therefore, the objective function to be minimized is $\sum_{k=1}^{K} \delta^k$ [15]. Now, we analyze the system constraints of this problem.

First, all UEs can download the traffic once in the schedule, which can be expressed as follows.

$$\sum_{k=1}^{K} a_{s_u u}^k = 1, \ \forall \, u. \tag{6}$$

Second, the schedule should complete the traffic downloading for all UEs, which can be expressed as follows.

$$\sum_{k=1}^{K} (\delta^k \cdot a_{s_u u}^k \cdot c_{s_u u}) \geq d, \ \forall \, u. \tag{7}$$

Third, since UE u obtains the common packets after UE s_u received the common packets, the downloading of UE s_u should be scheduled ahead of the downloading for UE u, which can be formulated as follows. This constraint represents a group of constraints since $\widetilde{K}$ varies from 1 to K,

$$\sum_{k=1}^{\widetilde{K}} a_{s_{s_u} s_u}^k \geq \sum_{k=1}^{\widetilde{K}} a_{s_u u}^k, \ \forall \, u, \ \widetilde{K} = 1 \sim K. \tag{8}$$

Fourth, due to the half-duplex assumption, adjacent links cannot be scheduled for concurrent transmissions. Thus, the links that share common nodes cannot be scheduled in the same pairing, which can be expressed as follows.

$$a_{s_u u}^k + a_{s_v v}^k \leq 1, \ \text{if } (s_u, u) \text{ and } (s_v, v) \text{ are adjacent}. \tag{9}$$

Finally, to enable concurrent transmissions, the SINR of each link in the same pairing should be able to support its transmission rate, which can be formulated as follows.

$$\frac{k_0 G_t(s_u, u) G_r(s_u, u) l_{s_u u}^{-\tau} P_t a_{s_u u}^k}{W N_0 + \rho \sum_{s_v} \sum_{v} k_0 G_t(s_v, u) G_r(s_v, u) l_{s_v v}^{-\tau} P_t a_{s_v v}^k} \geq \gamma(c_{s_u u})$$
$$\times a_{s_u u}^k, \ \forall \, u, k. \tag{10}$$

If link (s_u, u) is not scheduled in the kth pairing, $a_{s_u u}^k$ is 0, and this constraint does not apply. Otherwise, the SINR of link (s_u, u) should be greater than or equal to $\gamma(c_{s_u u})$.

Therefore, the problem of optimal transmission scheduling (P1) can be formulated as follows.

$$(\text{P1}) \quad \min \sum_{k=1}^{K} \delta^k, \tag{11}$$

s. t.   Constraints (6)–(10).

B. Problem Reformulation

Since constraints (7) and (10) are nonlinear, problem P1 is a mixed integer nonlinear programming (MINLP), which is generally NP-hard. By the Reformulation-Linearization Technique (RLT) [15], [39], we linearize constraints (7) and (10). The RLT procedure is used to produce tight linear programming relaxations for an underlying nonlinear and non-convex polynomial programming problem. In the RLT procedure, for each nonlinear term, a variable substitution is applied to linearize the objective function and the constraints. In addition, nonlinear implied constraints for each substitution variable are generated by taking the products of bounding terms of the decision variables, up to a suitable order [15].

For the second order term in constraint (7), we define a substitution variable $\xi^k_{s_u u} = \delta^k \cdot a^k_{s_u u}$. δ^k is bounded as $0 \le \delta^k \le \overline{T}$, where $\overline{T} = \max\{\lceil \frac{d}{c_{s_u u}} \rceil, \forall\ u\}$ [15]. With $0 \le a^k_{s_u u} \le 1$, we can obtain the *RLT bound-factor product constraints* for $\xi^k_{s_u u}$ as

$$\begin{cases} \xi^k_{s_u u} \ge 0 \\ \delta^k - \xi^k_{s_u u} \ge 0 \\ \overline{T} \cdot a^k_{s_u u} - \xi^k_{s_u u} \ge 0 \\ \overline{T} - \delta^k - \overline{T} \cdot a^k_{s_u u} + \xi^k_{s_u u} \ge 0 \end{cases} \quad \forall\ u, k. \quad (12)$$

For constraint (10), we first convert it to

$$(k_0 G_t(s_u, u)G_r(s_u, u)l_{s_u u}{}^{-\tau}P_t - \gamma(c_{s_u u})WN_0) \times a^k_{s_u u}$$
$$\ge \gamma(c_{s_u u})\rho \sum_{s_v} \sum_v k_0 G_t(s_v, u)G_r(s_v, u)l_{s_v u}{}^{-\tau}P_t a^k_{s_u u}a^k_{s_v v},$$
$$\forall\ u, k. \quad (13)$$

For the second order term $a^k_{s_u u}a^k_{s_v v}$, we define $\omega^k_{s_u u s_v v} = a^k_{s_u u}a^k_{s_v v}$ as the substitution variable. Since $0 \le a^k_{s_u u} \le 1$ and $0 \le a^k_{s_v v} \le 1$, the *RLT bound-factor product constraints* for $\omega^k_{s_u u s_v v}$ are

$$\begin{cases} \omega^k_{s_u u s_v v} \ge 0 \\ a^k_{s_u u} - \omega^k_{s_u u s_v v} \ge 0 \\ a^k_{s_v v} - \omega^k_{s_u u s_v v} \ge 0 \\ 1 - a^k_{s_u u} - a^k_{s_v v} + \omega^k_{s_u u s_v v} \ge 0 \end{cases} \quad \forall\ u, v, k. \quad (14)$$

By substituting $\xi^k_{s_u u}$ and $\omega^k_{s_u u s_v v}$ into constraints (7) and (10), problem P1 can be reformulated into a mixed integer linear programming (MILP) as

$$\min \sum_{k=1}^{K} \delta^k \quad (15)$$

s. t.

$$\sum_{k=1}^{K} (\xi^k_{s_u u} \cdot c_{s_u u}) \ge d, \ \forall\ u; \quad (16)$$

$$(k_0 G_t(s_u, u)G_r(s_u, u)l_{s_u u}{}^{-\tau}P_t - \gamma(c_{s_u u})WN_0) \times a^k_{s_u u}$$
$$\ge \gamma(c_{s_u u})\rho \sum_{s_v} \sum_v k_0 G_t(s_v, u)G_r(s_v, u)l_{s_v u}{}^{-\tau}P_t \omega^k_{s_u u s_v v},$$
$$\forall\ u, k; \quad (17)$$

Constraints (6), (8), (9), (12), (14).

Considering the example in Section III-B, with the selected transmission paths by Algorithm 1, we solve the MILP using an open-source MILP solver, YALMIP [40]. The optimal schedule consists of three pairings, and completes traffic downloading with eight time slots, which has been illustrated in Fig. 1 (a). However, problem P1 has $\mathcal{O}((|\mathbb{U}|)^2 K)$ decision variables, and $\mathcal{O}((|\mathbb{U}|)^2 K)$ constraints, and using the optimization software to solve the problem takes significant computation time, which is unsuitable for mmWave systems [15]. Therefore, to implement efficient concurrent transmission scheduling in practical mmWave small cells, heuristic concurrent transmission scheduling algorithms with low computational complexity are needed, which will be constructed in the following subsection.

C. Concurrent Transmission Scheduling Algorithm

After the transmission path selection by Algorithm 1, we propose a heuristic concurrent transmission scheduling algorithm to compute the near-optimal transmission schedules with much lower complexity than that of optimization software, borrowing the design ideas of the Greedy Coloring (GC) algorithm. Since adjacent links cannot be scheduled concurrently in the same pairing, the set of links in each pairing can be represented by a matching, and thus the maximum number of links in the same pairing is $\lfloor n/2 \rfloor$ [15]. We denote the set of links scheduled in the tth pairing by $\mathbb{E}^t$, and the set of vertices of the links in $\mathbb{E}^t$ is denoted by $\mathbb{V}^t$. Thus, the problem of optimal transmission scheduling is to obtain the matching in each pairing to complete traffic downloading with a minimum of time slots. In each pairing, our algorithm first obtains transmission paths with the largest number of hops, and then the hop with the largest weight among the first unscheduled hops of these paths will be visited first [15]. To maximize spatial reuse, the algorithm iteratively allocates as many links as possible into each pairing with the concurrent transmission conditions satisfied. For the hops on the same path, the preceding hops should be scheduled first since each UE can be the downloading source for other UEs only after it has received the common packets.

We denote the set of transmission paths selected by Algorithm 1 by $\mathbb{P}_b$. For each path $p \in \mathbb{P}_b$, we denote its number of hops by H_p. The set of hops in $\mathbb{P}_b$ is denoted by $\mathbb{E}_b$. We denote the hth hop of path p by (p, h), and define its weight w_{ph} as the number of time slots to complete traffic downloading. We denote the first unscheduled hop on path p by (p, F_p), where F_p indicates its hop number. We also denote the transmitter of (p, F_p) by s_{pF_p}, and the receiver by r_{pF_p}. In the tth pairing, we denote the set of paths that are not visited yet by $\mathbb{P}^t_u$.

The pseudo-code of the concurrent transmission scheduling algorithm is presented in Algorithm 2. After obtaining the set of the selected transmission paths, $\mathbb{P}_b$, we obtain the set of hops in $\mathbb{P}_b$ as $\mathbb{E}_b$. The algorithm iteratively schedules the hops in $\mathbb{E}_b$ into each pairing until all hops in $\mathbb{E}_b$ are scheduled, as in line 6. In each pairing, we first visit the paths with the largest number of unscheduled hops, as in line 11. Then the first unscheduled hop with the largest weight is visited and selected as the candidate hop of this pairing, as in line 12. In line 13, the algorithm examines whether the candidate hop is adjacent to the hops already in this pairing. If the candidate hop is not adjacent to the hops already in this pairing, this candidate hop will be added to this pairing to check whether the concurrent transmission conditions of this pairing are satisfied, as in line 14 and lines 15–18. If the SINR of one link in this pairing cannot support its transmission rate, the candidate hop will be removed from this pairing, as in line 17 and 23. Otherwise, the number of time slots of this pairing is updated to accommodate the traffic demand of this candidate hop, and this hop is removed from $\mathbb{E}_b$ in line 19. The visited path is removed from $\mathbb{P}_u^t$ in line 24. If the number of links in each pairing reaches $\lfloor n/2 \rfloor$ or there is no path unvisited, the algorithm will start scheduling for the next pairing as in line 10, and the scheduling results for this pairing will be outputted in line 25.

For the example in Section III-B, with the transmission paths selected by Algorithm 1, Algorithm 2 gives the same schedule as YALMIP [40]. However, since the outer while loop has $|\mathbb{E}_b|$ iterations, which at most is $|\mathbb{U}|$, and the inner while loop and for loop have $|\mathbb{E}^t|$ iterations, which is at most $\lfloor n/2 \rfloor$, our algorithm has the computational complexity of $\mathcal{O}(|\mathbb{U}|^3)$, which is much lower than YALMIP.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the delay and throughput of our proposed popular content downloading scheme under various traffic patterns. We compare it with two existing schemes, and also investigate the impact of the maximum number of hops H_{max} on the performance of our scheme.

A. Simulation Setup

In the simulation, we consider a typical mmWave small cell of an AP and ten UEs. We assume the AP is located in the center of a square area with $10m \times 10m$, and the ten UEs are uniformly distributed in the area. Adopting the simulation parameters in Table II of [12], the duration of a time slot is set to 5 μs. The packet size is set to 1000 bytes. According to the distances between nodes, we set three transmission rates, 2 Gbps, 4 Gbps, and 6 Gbps. With a transmission rate of 2 Gbps, a packet can be transmitted in a time slot [15]. The AP obtains the common packets from the upper layer or pushes the schedule to UEs in one time slot [15]. For the simulated small cell, it takes a few time slots for the AP to compute the transmission paths and transmission schedule. In the simulation, we assume nonadjacent links are able to be scheduled for concurrent transmissions.

For the downloading traffic, we adopt two modes:

1) Poisson Process: The packets arrive following a Poisson process with arrival rate λ. The traffic load, denoted by T_p, is defined as

$$T_p = \frac{\lambda \times L \times |\mathbb{U}|}{R}, \tag{18}$$

where L denotes the size of data packets, $|\mathbb{U}|$ denotes the number of UEs, and R is set to 2 Gbps.

2) Interrupted Poisson Process (IPP): The packets arrive following an interrupted Poisson process (IPP). The parameters of the interrupted Poisson process are λ_1, λ_2, p_1 and p_2, and the arrival intervals of an IPP obey the second-order hyper-exponential distribution with a mean of

$$E(X) = \frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2}. \tag{19}$$

Since the IPP can also be represented by an ON-OFF process, IPP traffic is typical bursty traffic. The traffic load T_i in this mode is defined as

$$T_i = \frac{L \times |\mathbb{U}|}{E(X) \times R}. \tag{20}$$

The simulation length is 10^5 time slots, and the delay threshold is set to 2.5×10^4 time slots. The packets with delay larger than the threshold are discarded by UEs. We evaluate the system by the following three performance metrics:

Algorithm 2: Concurrent Transmission Scheduling.

1 Input: The set of selected transmission paths, $\mathbb{P}_b$;
2   The set of hops in $\mathbb{P}_b$, $\mathbb{E}_b$;
3   The number of hops for each path $p \in \mathbb{P}_b$, H_p;
4   The weight of each hop $(p, h) \in \mathbb{E}_b$, w_{ph};
5 Initialization: Set $F_p = 1$ for each $p \in \mathbb{P}_b$; $t=0$;
6 while $|\mathbb{E}_b| > 0$ **do**
7 $t=t+1$;
8   Set $\mathbb{V}^t = \emptyset$, $\mathbb{E}^t = \emptyset$, and $\delta^t = 0$;
9   Set $\mathbb{P}_u^t$ with $\mathbb{P}_u^t = \mathbb{P}_b$;
10   **while** $|\mathbb{P}_u^t| > 0$ and $|\mathbb{E}^t| < \lfloor n/2 \rfloor$ **do**
11   Obtain the set of unvisited paths with the largest number of unscheduled hops, $\mathbb{P}_{mh}$;
12   Obtain the hop (p, F_p) of path $p \in \mathbb{P}_{mh}$ with the largest weight, w_{pF_p};
13   **if** $s_{pF_p} \notin \mathbb{V}^t$ and $r_{pF_p} \notin \mathbb{V}^t$ **then**
14 $\mathbb{E}^t = \mathbb{E}^t \cup \{(p, F_p)\}$; $\mathbb{V}^t = \mathbb{V}^t \cup \{s_{pF_p}, r_{pF_p}\}$;
15   **for** each link (p, h) in $\mathbb{E}^t$ **do**
16   Calculate the SINR of link (p, h), Γ_{ph};
17   **if** $\Gamma_{ph} < \gamma(c_{ph})$ **then**
18   Go to line 23
19 $\delta^t = \max(\delta^t, w_{pF_p})$, $\mathbb{E}_b = \mathbb{E}_b - (p, F_p)$;
20   **if** $F_p == H_p$ **then**
21 $\mathbb{P}_b = \mathbb{P}_b - p$;
22 $F_p = F_p + 1$; Go to line 24
23 $\mathbb{E}^t = \mathbb{E}^t - \{(p, F_p)\}$; $\mathbb{V}^t = \mathbb{V}^t - \{s_{pF_p}, r_{pF_p}\}$;
24 $\mathbb{P}_u^t = \mathbb{P}_u^t - p$;
25   Output $\mathbb{E}^t$ and δ^t;
26 Return $\mathbb{E}^t$ and δ^t for each pairing.

---

1) **Average Transmission Delay:** The average traffic downloading delay from the AP to UEs, which is in units of time slots.

2) **Network Throughput:** The number of successfully transmitted packets to all UEs until end of the simulation. With the constant simulation length and fixed packet size, the total number of successfully transmitted packets is a good indication to show the throughput performance.

3) **D2D Ratio:** The fraction of packets transmitted by device-to-device links over the total number of successfully transmitted packets. This metric is used to evaluate the role of D2D communications in the traffic downloading.

In the simulation, we compare our scheme with the following two transmission schemes:

1) *SBTS*: The serial broadcasting transmission scheme, where the packets are transmitted to UEs from the AP serially without exploiting the D2D communications.

2) *FDMAC-H*: In FDMAC-H, the concurrent transmission scheduling algorithm borrows the design ideas of the greedy coloring (GC) algorithm in FDMAC [15], and performs edge coloring iteratively on the first unscheduled hops of transmission paths in a non-increasing order of weight with the conditions for concurrent transmissions satisfied. After each coloring, the set of first unscheduled hops is updated since the hops after these scheduled links become the first unscheduled hops. In FDMAC-H, the transmission path selection algorithm is the same as PCDS, and the detailed pseudo-code illustration can be found in [30].

B. Comparison with the Optimal Solution

To show the gap between the concurrent transmission scheduling algorithm and the optimal solution of the MILP, we first compare PCDS (Section V.C.) with the optimal solution of the MILP (Section V.B.). Since obtaining the optimal solutions is time-consuming, we simulate a smaller scenario of six UEs. All other simulation parameters are as defined in Section VI-A.

We plot the delay and throughput comparison of PCDS and the optimal solution under Poisson traffic in Fig. 2. From the results, we can observe that the gap between the delay of PCDS and the optimal solution is small under light load, and the gap increases slowly with the traffic load. The gap is only about 17.2% at the traffic load of 3.33. In terms of network throughput, the gap is also very small. The gap is only about 2.8% at the traffic load of 3.33. Therefore, we have demonstrated that PCDS achieves near-optimal performance in terms of concurrent transmission scheduling.

(a) Average transmission delay (b) Network throughput
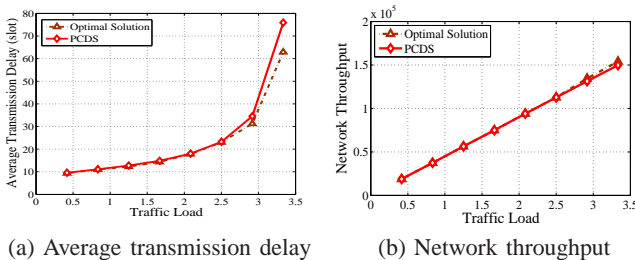
Fig. 2. Delay and throughput comparison of Optimal Solution and PCDS.

We also plot the average execution time of PCDS and the optimal solution under Poisson traffic in Fig. 3. We can

observe that the optimal solution takes much longer to execute than PCDS, and the gap increases with increasing traffic load, which indicates that PCDS has much lower computational complexity.
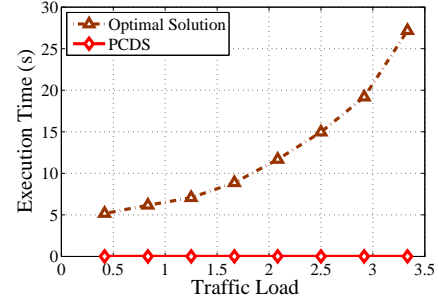

Fig. 3. Execution time comparison of Optimal Solution and PCDS.

C. Comparison with Existing Schemes

We compare the delay and throughput of SBTS, FDMAC-H, and PCDS under different traffic patterns. The maximum number of hops of transmission paths in Algorithm 1, H_{max} is set to 4.

1) Delay: We plot the average transmission delay of three schemes under different traffic loads in Fig. 4. We can observe that the delay curves of the three protocols rise slowly under light traffic loads. The average delay of SBTS and FDMAC-H begin to increase rapidly when the traffic load exceeds 1.5 and 2, while the delay of PCDS doesn't have obvious growth until the traffic load is greater than 3. Under heavy loads, PCDS outperforms SBTS and FDMAC-H significantly. With the traffic loads varying from 3 to 5, PCDS reduces the average delay by about 69.2% and 68.6% under Poisson and IPP traffic compared with FDMAC-H, respectively. Compared with SBTS, PCDS reduces the average delay by about 75.5% and 75.5% under Poisson and IPP traffic, respectively. In SBTS, since the links from the AP to UEs are adjacent, spatial reuse cannot be exploited. Furthermore, for UEs that have low transmission rates from the AP, their downloading suffers from higher delay. However, by D2D transmissions, PCDS allows UEs to receive packets from other UEs that have received the packets as well as the AP, and the low rate links from the AP to UEs are broken up into multiple hops of high rate, where more efficient spatial reuse can be exploited to improve transmission efficiency and reduce delay. The reason that PCDS outperforms FDMAC-H is that our concurrent transmission scheduling algorithm computes schedules with higher efficiency for the transmission paths selected by Algorithm 1.

2) Throughput: The network throughput achieved by the three schemes are plotted in Fig. 5. All three protocols have similar performance under light loads, and the throughput increases almost linearly with the traffic load. This is due to the fact that almost all arriving packets are distributed to all UEs successfully within the simulation duration. When the traffic load exceeds 1.5 and 2, the increases of throughput under SBTS and FDMAC-H fail to keep up with the traffic load, and the throughput curves begin to drop when the traffic
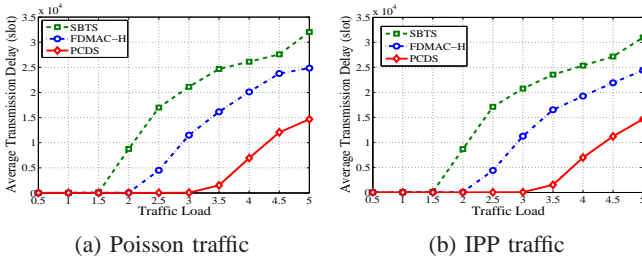
Fig. 4. Average transmission delay of the three transmission schemes under different traffic loads.

load is over 2 and 2.5, respectively. However, the curves of PCDS only begin to drop at the traffic load of 4. Compared with SBTS, PCDS increases the network throughput by about 282.5% on average with the traffic load varying from 3 to 5 under Poisson traffic, and about 275.1% under IPP traffic. PCDS also increases the network throughput by about 107.2% on average with the traffic load varying from 3 to 5 under Poisson traffic compared with FDMAC-H, and about 98.5% under IPP traffic. With the increase of the traffic load, the delay increases, and a considerable number of packets cannot be transmitted within the simulation duration. In addition, there are also more packets that are not counted as successful transmissions due to their delays exceeding the delay threshold. D2D transmissions enable the AP to offload a part of traffic distribution to UEs, and thus the potential of concurrent transmissions is fully unleashed.

3) Complexity: In terms of complexity, the SBTS scheme has much lower complexity compared with PCDS, but its performance is much poorer. FDMAC-H has the same transmission path selection algorithm as PCDS, which is $\mathcal{O}(|\mathbb{U}|^2)$. For the transmission scheduling algorithm, FDMAC-H has the complexity of $\mathcal{O}(|\mathbb{U}|^3)$, which is the same as that of PCDS. Therefore, FDMAC-H has the same complexity as PCDS, but its performance is poorer than PCDS.
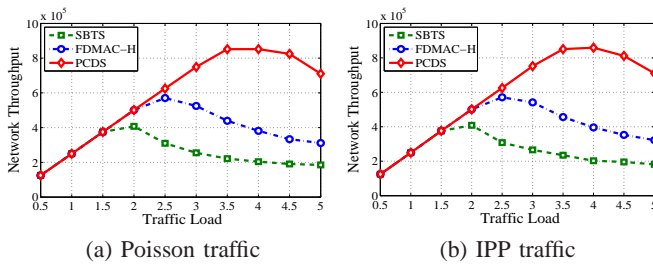

Fig. 5. Network throughput of the three transmission schemes under different traffic loads.

D. Impact of the Maximum Number of Hops

In this Section, we investigate the performance of PCDS under different maximum number of hops in the algorithm of transmission path selection, H_{max}, and denote PCDS with H_{max} equal to 2, 3, and 4 by PCDS-2, PCDS-3, and PCDS-4, respectively.

In Fig. 6, we plot the average transmission delay of PCDS with different H_{max}. Under light traffic loads, PCDS-4 and PCDS-3 have similar performance, but outperform PCDS-2 significantly. Under heavy loads between 3 and 5, PCDS-4 has more significant advantages in terms of delay. PCDS-4 reduces

the delay by about 62.1% on average with the traffic load between 3 and 5 under Poisson traffic compared with PCDS-3, and by about 71% compared with PCDS-2. With larger H_{max}, the links of low channel quality from the AP to UEs are broken up into more hops, and concurrent transmissions can be exploited more fully to improve performance.
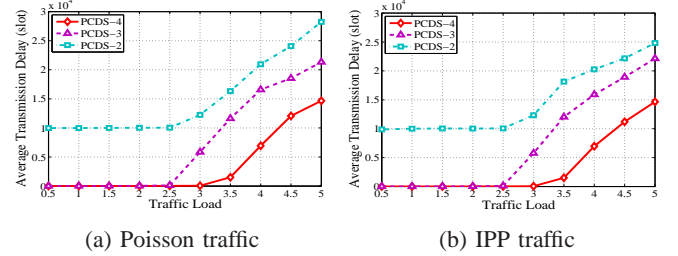

Fig. 6. Average transmission delay of PCDS with different H_{max}.

We plot the network throughput of PCDS with different H_{max} in Fig. 7. We can observe that the throughput is consistent with the delay in Fig. 6. With the traffic load varying from 3 to 5, PCDS-4 increases the network throughput by about 55.5% on average compared with PCDS-3, and about 57.9% compared with PCDS-2, respectively, under IPP traffic. With larger H_{max}, more traffic downloading is offloaded from the AP to UEs, which exploits D2D transmissions with higher transmission rate to improve network throughput.
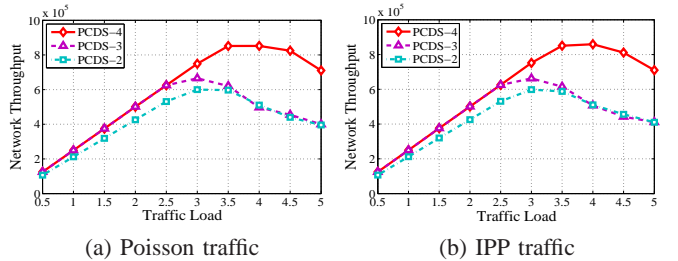

Fig. 7. Network throughput of PCDS with different H_{max}.

In Fig. 8, we plot the D2D ratios of PCDS with different H_{max}. We can observe PCDS-4 and PCDS-3 have higher D2D ratios than PCDS-2, especially under heavy loads. Thus, D2D transmissions are exploited more fully and efficiently in PCDS-4 than PCDS-3 and PCDS-2, which is also consistent with the results in Fig. 6 and Fig. 7.

Since each UE is only allowed as the downloading source of one neighboring UE once, the number of hops of the selected paths increases as a tolerance of 1 arithmetic progression starting from 1. If the number of UEs can be expanded to the sum of an arithmetic progression, the maximum number of hops of paths can be inferred as $\frac{\sqrt{1+8|\mathbb{U}|}-1}{2}$. Otherwise, the maximum possible number of hops of the selected paths is $\left\lceil \frac{\sqrt{1+8|\mathbb{U}|}-1}{2} \right\rceil$. Thus, although larger H_{max} usually indicates better performance, the maximum number of hops of transmission paths in Algorithm 1 is limited by $\left\lceil \frac{\sqrt{1+8|\mathbb{U}|}-1}{2} \right\rceil$. Therefore, H_{max} should be selected according to actual network conditions in practice.
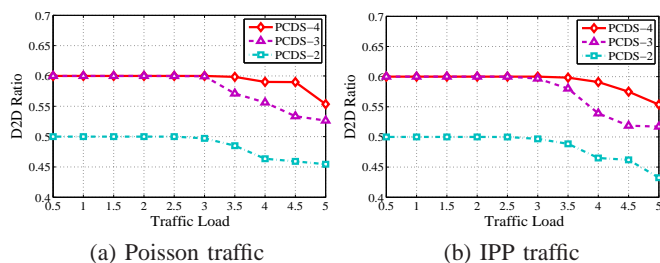
Fig. 8. D2D Ratios of PCDS with different H_{max}.

VII. Conclusion

In this paper, we proposed PCDS for popular content downloading in mmWave small cells in the 60 GHz band, which exploits both D2D transmission in close proximity and concurrent transmissions to improve transmission efficiency. Transmission path selection is optimized for better use of D2D communications and concurrent transmissions in content downloading. Then a concurrent transmission scheduling algorithm is designed to exploit spatial reuse to improve transmission efficiency. Finally, extensive simulations under various traffic patterns demonstrate PCDS reduces transmission delay and improves network throughput significantly compared with other existing schemes, especially under heavy load. Compared with FDMAC-H, PCDS improves the network throughput by about 102.9% on average with the traffic load between 3 and 5.

References

[1] "Qualcomm data challenge," accessed February 2014, http://www.qualcomm.com/media/documents/wireless-networks-risingmeet-1000x-mobile-data-challenge.

[2] C. H. Doan, S. Emami, D. A. Sobel, A. M. Niknejad, and R. W. Brodersen, "Design considerations for 60 GHz CMOS radios," *IEEE Communications Magazine*, vol. 42, no. 12, pp. 132–140, Dec. 2004.

[3] F. Gutierrez, S. Agarwal, K. Parrish, and T. S. Rappaport, "On-chip integrated antenna structures in CMOS for 60 GHz WPAN systems," *IEEE J. Selected Areas in Communications*, vol. 27, no. 8, pp. 1367–1378, Oct. 2009.

[4] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, "State of the art in 60-GHz integrated circuits and systems for wireless communications," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.

[5] ECMC TC48, ECMA standard 387, "High rate 60 GHz PHY, MAC and HDMI PAL," Dec. 2008.

[6] IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group 3c (TG3c). Available: http://www.ieee802. org/15/pub/TG3c.html.

[7] *Draft Standard for Information Technology–Telecommunications and Information Exchange Between Systems–Local and Metropolitan Area Networks–Specific Requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 4: Enhancements for Very High Throughput in the 60 Ghz Band*, IEEE P802.11ad/D9.0, Oct. 2012.

[8] J. Wang, Z. Lan, C. Pyo, T. Baykas, C. Sum, M. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Selected Areas in Communications*, 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. of IEEE Global Telecommunications Conference*, Houston, TX, Dec. 5-9, 2011.

[10] Z. Xiao, L. Bai, and J. Choi, "Iterative Joint Beamforming Training with Constant-Amplitude Phased Arrays in Millimeter-Wave Communications," *IEEE Communications Letters*, vol. 18, no. 5, pp. 829–832, May 2014.

[11] Y.-B. Ko, V. Shankarkumar, and N. F. Vaidya, "Medium access control protocols using directional antennas in ad hoc networks," in *Proc. 19th IEEE INFOCOM*, Tel Aviv, Israel, Mar. 26-30, 2000.

[12] 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.

[13] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, Y. Grunenberger, V. Erramilli, "Is there a case for mobile phone content pre-staging?," in *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, Santa Barbara, CA, Dec. 9–12, 2013.

[14] X. Lin, R. Ratasuk, A. Ghosh, et. al., "Modeling, analysis, and optimization of multicast Device-to-Device transmissions," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4346–4359, Aug. 2014.

[15] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. IEEE INFOCOM*, Orlando, FL, Mar. 25-30, 2012.

[16] Y. Li, Z. Wang, D. Jin, and S. Chen, "Optimal mobile content downloading in Device-to-Device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3596–3608, Jul. 2014.

[17] 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. PP, no. 99, Jun. 2014.

[18] M. N. Islam, A. Sampath, A. Maharshi, O. Koymen, N. B. Mandayam, "Wireless backhaul node placement for small cell networks," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, Mar. 19–21, 2014.

[19] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[20] Y. Zhu, Z. Zhang, et. al., "Demystifying 60GHz outdoor picocells," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, Hawaii, HI, Sep. 7–11, 2014.

[21] J. Qiao, L. X. Cai, X. Shen, and J. Mark, "STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks," in *Proc. IEEE ICC*, Ottawa, Canada, Jun. 10–15, 2012.

[22] 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.

[23] C. Sum, Z. Lan, R. Funada, J. Wang, T. Baykas, M. A. Rahman, and H. Harada, "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.

[24] C.-S. Sum, Z. Lan, M. A. Rahman, et. al., "A multi-Gbps millimeter-wave WPAN system based on STDMA with heuristic scheduling," in *Proc. IEEE GLOBECOM*, Honolulu, HI, Nov. 30-Dec. 4, 2009.

[25] J. Qiao, L. X. Cai, and X. Shen, et. al., "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.

[26] M. X. Gong, R. J. Stacey, D. Akhmetov, and S. Mao, "A directional CSMA/CA protocol for mmWave wireless PANs," in *Proc. IEEE WCNC'10*, Sydney, Australia, Apr. 18–21, 2010.

[27] Q. Chen, X. Peng, J. Yang, and F. Chin, "Spatial reuse strategy in mmWave WPANs with directional antennas," in *Proc. 2012 IEEE GLOBECOM*, Anaheim, CA, Dec. 3-7, 2012.

[28] Q. Chen, J. Tang, D. Wong, X. Peng, and Y. Zhang, "Directional cooperative MAC protocol design and performance analysis for IEEE 802.11 ad WLANs," *IEEE Trans. Veh. Tech.*, vol. 62, no. 6, pp. 2667–2677, Jul. 2013.

[29] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "A Two Stage Approach for Channel Transmission Rate Aware Scheduling in Directional mmWave WPANs," *Wireless Communications and Mobile Computing*, DOI: 10.1002/wcm.2521.

[30] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device transmissions in joint scheduling of access and backhaul for small cells in 60 GHz band," to appear in *IEEE Journal on Selected Areas in Communications*.

[31] S. Singh, R. Mudumbai, and U. Madhow, "Distributed coordination with deaf neighbors: efficient medium access for 60 GHz mesh networks," in *Proc. IEEE INFOCOM* (San Diego, CA), Mar. 14-19, 2010, pp. 1–9.

[32] E. Shihab, L. Cai, and J. Pan, "A distributed asynchronous directional-to-directional MAC protocol for wireless ad hoc networks," *IEEE Trans. Veh. Tech.*, vol. 58, no. 9, pp. 5124–5134, Nov. 2009.

[33] J. Ning, T.-S. Kim, S. V. Krishnamurthy, and C. Cordeiro, "Directional neighbor discovery in 60 GHz indoor wireless networks," in *Proc. ACM MSWiM '09*, Tenerife, Spain, Oct. 26-30, 2009.

[34] H. Deng and A. Sayeed, "Mm-wave MIMO channel modeling and user localization using sparse beamspace signatures," in *2014 IEEE 15th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Toronto, Canada, Jun., 2014, pp. 130–134.

[35] S. Y. Geng, J. Kivinen, X. W. 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.

[36] 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.

[37] R. C. Daniels and R. W. Heath, "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Veh. Tech. Mag.*, vol. 2, no. 3, pp. 41–50, Sep. 2007.

[38] Y. Niu, C. Gao, Y. Li, L. Su, and D. Jin, "Exploiting multi-hop relaying to overcome blockage in directional mmWave small cells," to appear in *Journal of communications and networks*.

[39] H. D. Sherali and W. P. Adams, *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Boston, MA: Kluwer Academic, 1999.

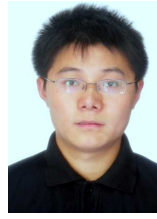[40] YALMIP YALMIP Wiki [Online]. Available: http://users.isy.liu.se/johanl/yalmip/.

Yong Niu received B.E. degree in Beijing Jiaotong University, China in 2011. He is currently working toward his PhD degree at the Department of Electronic Engineering, Tsinghua University, China. His research interests include millimeter wave communications, medium access control, and software-defined networks.

Li Su received the B.S. degree from Nankai University, Tianjin, China, in 1999 and Ph.D. degree from Tsinghua University, Beijing, China in 2007 respectively both in electronics engineering. Now he is a research associate with Department of Electronic Engineering, Tsinghua University. His research interests include telecommunications, future internet architecture and on-chip network.

Chuhan Gao is an undergraduate student from Tsinghua University, Beijing, China. He is currently working toward the B.E. degree. His research interests include millimeter-wave wireless communications, wireless networks, and software-defined networks.

Yong Li (M'2009) received his B.S. degree in Electronics and Information Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007, and his Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. During July to August in 2012 and 2013, he worked as a Visiting Research Associate in Telekom Innovation Laboratories (T-labs) and HK University of Science and Technology, respectively. During December 2013 to March 2014, he visited University of Miami, FL, USA as a Visiting Scientist. He is currently a faculty member of the Electronic Engineering at the Tsinghua University.

His research interests are in the areas of networking and communications, including mobile opportunistic networks, device-to-device communication, software-defined networks, network virtualization, future Internet, etc. He received Outstanding Postdoctoral Researcher, Outstanding Ph. D Graduates and Outstanding Doctoral thesis of Tsinghua University, and his research is granted by Young Scientist Fund of Natural Science Foundation of China, Postdoctoral Special Find of China, and industry companies of Hitachi, ZET, etc. He has published more than 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 more than 120 total citations without self-citations in Web-of Science. He has served as Technical Program Committee (TPC) Chair for WWW workshop of Simplex 2013, served as the TPC of several international workshops and conferences. He is also a guest-editor for ACM/Springer Mobile Networks & Applications, Special Issue on Software-Defined and Virtualized Future Wireless Networks. Now, he is the Associate Editor of EURASIP journal on wireless communications and networking.

Depeng Jin received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering.

He is an associate professor at Tsinghua University and vice chair of Department of Electronic Engineering. Dr. Jin was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future Internet architecture.

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

From 2000 to 2002, he was an R&D Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, wireless multimedia, security, and smart grid communication. He serves as an Associate Editor for IEEE Transactions on Wireless Communications since 2010. He received IEEE Fred W. Ellersick Prize in 2011 and the NSF CAREER award in 2010.