Mobility-Aware Transmission Scheduling Scheme for mmWave Cells

Article *in* IEEE Transactions on Wireless Communications · July 2018

DOI: 10.1109/TWC.2018.2853748

CITATIONS
25

READS
199

6 authors, including:

Yong Niu
Beijing Jiaotong University
179 PUBLICATIONS **4,576** CITATIONS
SEE PROFILE

bo ai
Beijing Jiaotong University
1,027 PUBLICATIONS **24,778** CITATIONS
SEE PROFILE

Depeng Jin
Tsinghua University
514 PUBLICATIONS **16,208** CITATIONS
SEE PROFILE

Mobility-Aware Transmission Scheduling Scheme for Millimeter-Wave Cells

Yu Liu, Xinlei Chen, Yong Niu, *Member, IEEE*, Bo Ai, *Senior Member, IEEE*, Yong Li, *Senior Member, IEEE*, and Depeng Jin, *Member, IEEE*

***Abstract*— With the explosive growth of mobile traffic, diverse mobile applications with high throughput demands have gained considerable attention from both academia and industry. However, dynamics due to human mobility impose serious issues on high throughput communications. Although millimeter-wave (mm-wave), relays, and concurrent transmission have been applied for throughput improvement, how to achieve high throughput transmission in mobility-aware scenarios is still challenging. In this paper, we propose a throughput-efficient service scheduling (TESS) scheme, which exploits multi-hop relay and concurrent transmissions with the consideration of human mobility. In TESS, we first propose a mobility-aware transmission scheduling scheme in single mm-wave cell. The proposed scheduling scheme consists of a relay path planning algorithm and a global time scheduling algorithm. In the relay path planning algorithm, we establish multi-hop relay transmission paths from base station to service points. In the global time scheduling algorithm, we schedule concurrent transmission in relay paths. Furthermore, the proposed TESS scheme is extended for multi-cell scenarios. Through extensive performance evaluations under realistic human mobility trajectories, we demonstrate the superior performance of TESS in terms of system throughput compared with the state-of-the-art schemes.**

***Index Terms*— Millimeter wave communications, multi-hop relay transmissions, concurrent transmissions, mobility-aware transmission scheduling.**

I. Introduction

MOBILE traffic is growing explosively and diversely. According to Cisco's forecast, global mobile data traffic will increase sevenfold between 2016 and 2021, reaching 49% exabytes per month by 2021 [1]. Numerous mobile applications with the demand for a great amount of data spring up, such as videos, high definition maps, image recognition etc. Moreover, more than three-fourths of the world's mobile data traffic will be video by 2021, which indicates the importance of content downloading services [1]. In particular, people use these applications on buses, trains, running, as well as walking on sidewalks. This causes the service continuity problem, which becomes a key issue in mobility scenarios. Accordingly, nowadays mobile communications are badly in need of high throughput and mobility-aware transmission.

Based on the high throughput demand, mm-wave communications have been widely researched [2], [3]. With huge bandwidth from 30 GHz to 300 GHz, the mm-wave communication is commited as a promising 5G technology to provide multi-gigabyte broadband communication services. Especially, high gain directional antennas with analog beamforming are utilized to overcome high channel attenuation [4], [5]. Besides, the topology and architecture of cellular networks are undergoing a major paradigm shift in 5G era, from traditional grid networks to 5G ultra-dense cellular networks [6], [7]. More specifically, densely deployed small cells will play a major role thanks to their ability to increase the spatial reuse of the available resources.

Furthermore, multi-hop relay transmissions have proved their advantages in throughput improvement [8]. Since relays allow for reducing the end-to-end path loss between an information source and its destination, they are able to extend cell range, improve system throughput as well as provide a better QoS [9]. For instance, the multi-hop relay communication model was proposed and evaluated in [10] for higher data rate and better coverage. In addition, several multi-hop relay protocols have been proposed to accommodate the multi-hop relay architecture [8], [9]. Meanwhile, concurrent transmissions are also exploited for capactiy improvement and spatial reuse [11]–[13]. The system throughput is significantly improved with several transmission links working simultaneously.

Nevertheless, human mobility poses several challenges to the high throughput transmission. To be specific, human mobility is associated with the frequent position change, which makes it more difficult to accurately estimate, track and predict the channel state. Moreover, frequent handovers in multiple cells have a serious impact on the quality of service (QoS) and the transmission efficiency. As a result, both the channel quality and the mobile network performance are highly sensitive to the movements of users. On the other hand, since we integrate the multi-hop relay and the concurrent transmission

Manuscript received April 12, 2018; revised June 15, 2018; accepted June 25, 2018. Date of publication July 16, 2018; date of current version September 10, 2018. This work was supported in part by the National Nature Science Foundation of China under Grant 61861136003, Grant 61621091, Grant 61673237, and Grant 61725101, in part by the National Key Research and Development Program of China under Grant 2017YFE0112300 and Grant 20031887521, in part by the Research Fund of Tsinghua University, Tencent Joint Laboratory for Internet Innovation Technology, in part by Beijing Natural Fund under Grant L172020, and in part by the China Postdoctoral Science Foundation under Grant 2018T110041. The associate editor coordinating the review of this paper and approving it for publication was G. Mao. *(Yu Liu and Xinlei Chen contributed equally to this work.) (Corresponding author: Yong Li.)*

Y. Liu, X. Chen, Y. Li, and D. Jin are with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liuyu2419@126.com; xinlei.chen@sv.cmu.edu; liyong07@tsinghua.edu.cn).

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

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/TWC.2018.2853748

to achieve high throughput, the difficulty of the mobility-aware transmission is multiplied. As for the multi-hop relay transmission, mobility has to be considered in routing schemes with the relay selection and the path construction. Besides, media access control (MAC) is necessary in concurrent transmission for collision avoidance. Therefore, the problem of when and through which paths the data will be transmitted, needs to be solved in the mobility-aware transmission. To achieve the high throughput transmission, existing challenges are summarized as follows,

- Dynamics in human mobility lead to uncertainties in communications. The frequent position change affects the channel state impairing the QoS and the throughput. Besides, frequent handovers complicate the problem in multi-cell scenarios.
- Path construction and concurrent scheduling are tightly coupled in the mobility-aware transmission. The complex interaction in routing and MAC problems increases difficulties in the modelling design, as well as the solution.

In this paper, we study the problem of throughput efficient scheduling for mm-wave cells, where multi-hop relays and concurrent transmission are exploited to achieve the mobility-aware transmission and enhance the content downloading service. Our novel contributions are listed as follows.

- We formulate the problem of optimal mobility-aware transmission into a mixed integer nonlinear program (MINLP), which maximizes the system throughput based on users' mobility trajectories with multi-hop relays and concurrent transmissions exploited.
- We propose a throughput-efficient service scheduling scheme, termed TESS, to solve the formulated problem. TESS takes advantage of prior-known mobility trajectories and selects the relay near the service point for content caching to overcome mobility dynamics.
- The proposed TESS consists of relay path planning algorithm, global time scheduling algorithm, and multi-cell scheduling algorithm. In order to solve the tight coupling in the mobility-aware transmission, TESS decouples the scheduling in space and time domains sequentially. In space domain, the relay path planning algorithm establishes multi-hop relay transmission paths for service points. Meanwhile, in time domain, the global time scheduling algorithm schedules concurrent transmission in relay paths. Furthermore, the multi-cell scheduling algorithm solves the handover problems in multiple mm-wave cells.
- Extensive evaluations under various system parameters demonstrate TESS achieves higher system throughput compared with state-of-the-art schemes. Typically, it improves the system throughput by over 100% in single mm-wave cell. For the multi-cell scenario, TESS provides a more than 3x increase in system throughput.

The rest of the paper is organized as follows. Section II discusses the related work on directional MAC protocols in the mm-wave band, 5G cellular networks and mobility-aware communications. Section III introduces the system model, and formulates the problem of multi-hop transmission scheduling with human mobility considered into a MINLP problem. Section IV presents our proposed TESS scheme, and explains the algorithms in detail. Section V evaluates the performance under various system parameters. In the light of our results, we conclude the paper in Section VI.

II. Related Work

Directional protocols and scheduling schemes for WPANs or WLANs have been extensively researched in the mm-wave band [14]–[16]. For example, based on IEEE 802.15.3c, Qiao *et al.* [14] introduced a concurrent transmission algorithm to maximize the number of flows with the throughput requirement of each flow satisfied. Besides, Singh *et al.* [15] exploited multi-hop relay to overcome blockage in the multi-hop relay directional MAC protocol (MRDMAC). Niu *et al.* [16] proposed an energy-efficient multicast scheduling scheme for the radio access in the mm-wave band, which utilize both D2D communications and concurrent transmissions for high energy efficiency. These protocols and standards make the concurrent relay transmissions possible.

Recently, 5G cellular networks have attracted much interest [17]–[20]. With the concept of network densification and small cells, 5G ultra-dense cellular networks are proposed for raising throughput and saving energy consumption in cellular scenarios. To answer the important question of how much densification can be deployed for ultra-dense cellular networks, Ge *et al.* [17] investigated the backhaul network capacity as well as energy efficiency, and densification limits are obtained. To exploit the benefits of small cell deployments, Calabuig *et al.* [18] proposed a mobility management scheme, which utilized context information to anticipate handovers and future demands, radio fingerprints to discover nearby small cells. To integrate mm-wave with 5G ultra-dense cellular networks, Giordani *et al.* [19] analysed the initial access techniques based on the coverage and delay, where a trade-off between initial access delays and misdetection probability is observed. Moreover, Matsumura *et al.* [20] proposed the concept of the shared base station in 5G cellular networks, and prototyped a shared base station supporting mm-wave communications. These works provide the fundamental design of mm-wave cells for 5G ultra-dense cellular networks.

Also, there are a few works on mobility-aware communications. In particular, the concept of proactive caching is widely discussed for mobility-aware communications [21]. Considering the human mobility, the core concept is caching at the wireless edge in order to reduce the transmission latency. For instance, Dai and Liu [22] investigated the mobility property, and proposed a mobility-aware proactive caching strategy, in which a novel controller was designed for making cache decisions. Liu *et al.* [23] utilized the coded caching, and proposed a mobility-aware caching scheme in small cells. Besides, Wang *et al.* [24] combined proactive caching with D2D communications, and proposed the incentive mechanism to motivate users to cache. However, these existing works in proactive caching are focus on the cache strategy without considering the transmission scheduling before caching.

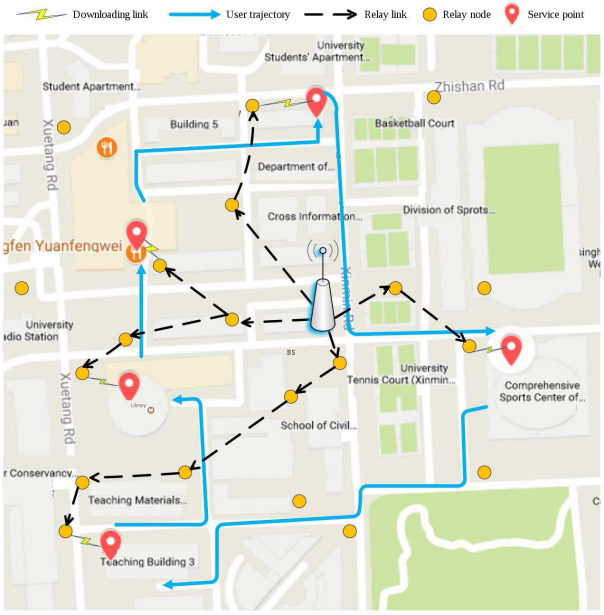


Fig. 1. The typical scenario of the system model: Jack's weekday life and the illustration of TESS scheme.

Actually, mobility-aware transmission plays an important role in addressing the human mobility.

Nevertheless, most existing works consider mm-wave transmissions in a static scenario with fixed nodes, which do not exploit human mobility to provide high throughput service. For example, Niu *et al.* [12] developed multi-hop relaying to achieve throughput-efficient transmissions, assuming all users are static. First interest in the mobility-aware transmissions dates back to the work of [25]. The authors proposed a trade-off between data rate and mobility by means of a set of transmission classes that use smart antennas in different configurations, including smart antennas at both BS and the mobile user. These considerations provide a typical case for nowadays directional antennas in mobility-aware transmissions. In the scenario of LTE-A networks, Yildiz and Sokullu [26] proposed a mobility-aware downlink scheduling algorithm to enhance the system performance. Khadka *et al.* [27] proposed an autocorrelation based feedback reduction technique in order to reduce the channel mismatch due to human mobility. However, relay and concurrent transmissions are not considered in these works for performance improvement.

To the best of our knowledge, we are the first to integrate concurrent relay communications to achieve mobility-aware transmission scheduling in mm-wave cells.

III. System Overview and Problem Formulation

A. System Model

Fig. 1 shows a typical scenario for our considered system model. The story's major role, Jack, is a college student, whose weekday scenario is almost similar. In the early morning, while working out in the gym, Jack usually listens to some online music. Then he walks into the teaching building for lessons, where Internet searching is always necessary. After school, Jack often learns online courses in the library, and next enjoys Internet surfing in the canteen and apartment. For simplicity, places in mobility trajectories where users demand for content services, are referred to as service points. Based on the prior-known trajectory and deployed relays (orange nodes in the figure), the BS selects relays near service points for content caching. Multi-hop relay paths from the BS to selected relays are established, and scheduled for concurrent transmission, shown as black relay links in Fig. 1. When Jack reaches service points, the downloading link is built from the corresponding relay to the user. Therefore, the user would enjoy high quality of services cached at relays.

Based on the above scenario, we consider a mm-wave cell with one BS, static relay nodes (REs), and one mobile user (UE), where the BS is located at the cell center and REs are deployed around the cell. The system time is partitioned into non-overlapping time slots of equal length, and clocks of REs and UE are synchronized by the BS. Equipped with electronically steerable directional antennas, directional transmissions are supported between any pair of BS, REs and UE. Besides, all nodes are assumed to be half-duplex, i.e., each node has at most one connection with another node simultaneously.

The directional link from node i to j is denoted by (i, j). After the beamforming procedure, node i and j point their antennas towards each other. Then based on the path loss model, the received power at node j for (i, j), denoted by P_{ij}^r (mW), can be obtained as

$$P_{ij}^r = k_0 G_t(i,j) G_r(i,j) d_{ij}^{-\gamma} P_t, \tag{1}$$

where we denote the transmission power by P_t (mW), the transmit antenna gain of node i by $G_t(i, j)$, the receive antenna gain of node j by $G_r(i, j)$, the distance from transmitter i to receiver j by d_{ij} (m), the path loss exponent by γ, and k_0 is a constant coefficient and proportional to $(\frac{\lambda}{4\pi})^2$ (λ stands for the wavelength) [14].

Considering the directional transmission in a large range, concurrent transmissions are adopted for spatial reuse as well as capacity improvement. Due to the long distance between mm-wave links in the large region as well as the directional antenna technology, the interference among mm-wave links is little enough to be omitted [28]. For link (i, j), we can obtain the SNR at receiver j as

$$SNR_{ij} = \frac{k_0 G_t(i,j) G_r(i,j) d_{ij}^{-\gamma} P_t}{N_0 W}, \tag{2}$$

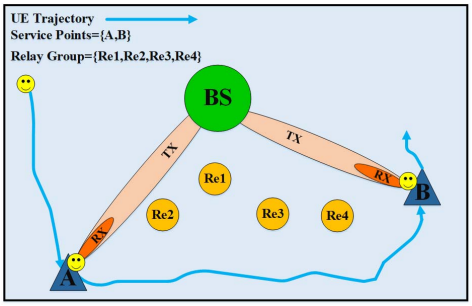
where we denote the bandwidth by W (Hz), and the one-sided power spectra density of white Gaussian noise by N_0 (mW/Hz) [14].

Based on the multipath effect of directional mm-wave links [15], as well as Shannon's channel formula, link (i, j) is able to achieve a data rate of

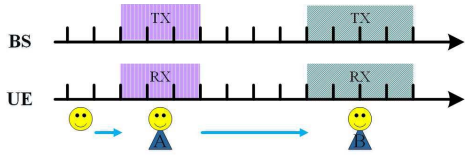
$$R_{ij} = \eta W \log_2 \left(1 + \frac{k_0 G_t(i,j) G_r(i,j) d_{ij}^{-\gamma} P_t}{N_0 W}\right), \tag{3}$$

where $\eta \in (0, 1)$ depicts the transceiver efficiency [14].

In the system, we consider the BS transmits service content to the mobile UE directly. As mentioned before, UE's

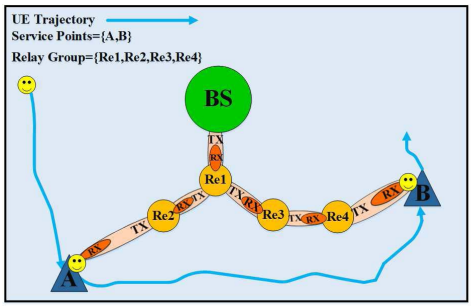


(a) Network topology and the unicast scheme operation

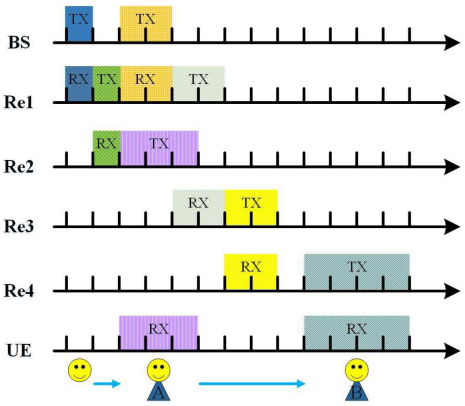


(b) Timeline operation of the unicast scheme

Fig. 2. An example of the unicast scheme operation in a small cell with two service points.



(a) Network topolgy and TESS operation



(b) Timeline operation of TESS

Fig. 3. An example of TESS operation in a small cell with two service points.

mobility trajectory can be derived by trajectory data mining in GPS location and other information. In our solution, taking advantage of prior-known UE's mobility trajectory with BS's and REs' position information, we utilize both relaying and concurrent transmissions of links to achieve high system throughput. To illustrate the key mechanisms of the problem, we present an example in Section III-B.

B. Problem Overview

We consider a small mm-wave cell with one BS and four relays deployed, whose topology is shown in Fig. 2(a) and Fig. 3(a). A UE's trajectory and two service points are also plotted in figures. The relay group is Re1, Re2, Re3, Re4, while the service points are A and B, where UE passes by and requires for communication and network service. For the numerical example, we denote a channel transmission rate matrix $\mathbf{C}$ in unit of packet, which is

$$\mathbf{C} = \begin{pmatrix} 0 & \mathbf{6} & \mathbf{6} & 3 & 1 & 1 & 6 \\ 6 & 0 & 3 & 1 & \mathbf{2} & 0 & 4 \\ 6 & 3 & 0 & \mathbf{6} & 1 & 2 & 3 \\ 3 & 1 & 6 & 0 & 0 & \mathbf{3} & 1 \\ 1 & 2 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 2 & 3 & 0 & 0 & 1 \\ \mathbf{6} & 4 & 3 & 1 & \mathbf{1} & \mathbf{1} & 0 \end{pmatrix}, \tag{4}$$

where the first four rows/columns are relays, the fifth and sixth rows/columns are UE at service points A and B, and the BS is the last row/column. For instance, (1,2) element of $\mathbf{C}$ is equal to 6, indicating that relay Re1 can transmit 6 packets to relay Re2 in one time slot. Additionally, (7,5) element of $\mathbf{C}$ is 1, which means that BS can transmit 1 packet per time slot to UE if UE is at service point A.

Based on the channel transmission rate matrix $\mathbf{C}$, we show the execution of unicast scheme in Fig. 2. In particular, only when UE is at service points can the BS transmit data to UE. Thus, when UE arrives at the service point, the BS directs its directional beam towards UE, and executes the data transmission until UE leaves the service point, as shown in Fig. 2(a). Fig. 2(b) exhibits the timeline operation of the unicast scheme, in which UE receives 3 packets at A and 4 packets at B, respectively.

As comparison, the scheduling and execution of TESS are plotted in Fig. 3. According to the network topology and TESS operation in Fig. 3(a), TESS extends two multi-hop relay path from BS to the service points A and B, BS $\rightarrow$ Re1 $\rightarrow$ Re2 $\rightarrow$ A and BS $\rightarrow$ Re1 $\rightarrow$ Re3 $\rightarrow$ Re4 $\rightarrow$ B. Combining the timeline operation of TESS in Fig. 3(b), we can observe that before UE arrives at A, BS $\rightarrow$ Re1, and Re1 $\rightarrow$ Re2 are correspondingly activated for data transmission. When UE arrives at A, Re2 transmits data to UE. Meanwhile, BS $\rightarrow$ Re1 transmits data concurrently. Re1 $\rightarrow$ Re3, and Re3 $\rightarrow$ Re4 are also activated subsequently. Due to the enough data cached at Re4 and UE's not arriving, there is a blank slot without any link transmissions before UE arrives at B. Finally, UE receives 6 packets at A, and 12 packets at B.

As can be observed, both the unicast scheme and TESS complete the data transmission at the same time. However, TESS transmits much more data than the unicast scheme, leading to high system throughput. The enhancement in system performance is mainly supported by two key mechanisms, relay communications and concurrent transmissions. Furthermore, it can be obtained that the multi-hop relay path selection plays an important role in spatial reuse, which should be optimized in the space scheduling. Also, appropriate management

in concurrent transmissions is quite necessary for efficient data forwarding in the time scheduling.

C. Problem Formulation

We divide the session of UE mobility trajectory into K time slots and denote the duration of each time slot by Δ. In the trajectory, there are several service points, as mentioned in Section I. During the time that UE stays at each service point, it downloads content from REs. We denote the set of UE's service points by $\mathbb{U}$, and each point is also regarded as one node in the system. We denote the set of REs by $\mathbb{V}$, and the BS in the network by $\mathbb{B}_s = \{B_s\}$. For each node w in the system ($\mathbb{U} \cup \mathbb{V}$), we denote $s_w \in \mathbb{V} \cup \mathbb{B}_s$ as the transmitter node to w. Besides, for each service point $u \in \mathbb{U}$, we denote t_u as UE's arrival time at u, and τ_u as staying time at u (time here means time slots). Moreover, a bool variable b_{ij}^k is defined to indicate whether the transmission for link (i, j) is scheduled in the kth time slot. If so, b_{ij}^k equals to 1; otherwise, b_{ij}^k equals to 0. With (3), the achievable transmission rate at node w in the kth time slot can be obtained as,

$$R_{s_w w}^k = \eta W \log_2 \left(1 + \frac{b_{s_w w}^k k_0 G_t(s_w, w) G_r(s_w, w) d_{s_w w}^{-\gamma} P_t}{N_0 W}\right). \tag{5}$$

Since the data are transmitted from the BS to service point u through a specific multi-hop path, we denote H_u as the number of hops for this path, and denote $w_j(u)$ as the receiver node in the jth hop. For each $w_j(u)$, we define integer variables $t_{w_j(u)}^s$ and $t_{w_j(u)}^e$. Then $w_j(u)$ starts receiving at the $t_{w_j(u)}^s$th time slot and stops receiving at the $t_{w_j(u)}^e$th time slot. The amount of received data for UE at service point u is denoted by D_u.

To achieve high performance for content downloading for UE, we should maximize the system throughput, defined as the total amount of the received data divided by the total time of its mobility trajectory. Thus, the objective function can be obtained as $\sum_{u=1}^{|\mathbb{U}|} \frac{1}{K\Delta} \cdot D_u$. Moreover, the amount of received data D_u can be calculated according to the achievable transmission rate $R_{s_u u}^k$ in time slots, which means,

$$\sum_{u=1}^{|\mathbb{U}|} \frac{1}{K\Delta} \cdot D_u = \sum_{u=1}^{|\mathbb{U}|} \sum_{k=1}^{K} \frac{1}{K} \cdot R_{s_u u}^k. \tag{6}$$

We now consider the system constraints. First, the transmission at each service point can be scheduled only when UE stays there, which can be expressed as,

$$b_{s_u u}^k = 0, \quad \forall\ u, \text{ if } k \le t_u \text{ or } k \ge t_u + \tau_u. \tag{7}$$

Second, according to the half-duplex assumption, we cannot schedule adjacent links concurrently, which can be obtained as,

$$b_{s_u u}^k + b_{s_v v}^k \le 1, \quad \text{if link}(s_u, u) \text{ and } (s_v, v) \text{ are adjacent}, \quad \forall u, v \in \mathbb{U} \cup \mathbb{V}. \tag{8}$$

Third, the jth hop in the path for service point u should be scheduled before the $(j+1)$th hop, and only when UE stays at u, can the last hop transmit cache data, which can be acquired as,

$$t_{w_j(u)}^s \le t_{w_j(u)}^e \le t_{w_{(j+1)}(u)}^s \le t_{w_{(j+1)}(u)}^e, \quad \forall u,\ j = 1, 2, \cdots, H_u - 1. \tag{9}$$

$$t_u \le t_{w_{H_u}(u)}^s \le t_{w_{H_u}(u)}^e \le t_u + \tau_u, \quad \forall u. \tag{10}$$

Fourth, UE's received data at u cannot exceed the data forwarded by each hop of the path, which means that,

$$\sum_{k=t_{w_j(u)}^s}^{t_{w_j(u)}^e} R_{s_{w_j(u)} w_j(u)}^k \Delta \ge \sum_{k=t_{w_{H_u}(u)}^s}^{t_{w_{H_u}(u)}^e} R_{s_u u}^k \Delta, \quad \forall u,\ j = 1, 2, \cdots, H_u. \tag{11}$$

Therefore, the mobility-aware optimal scheduling problem (P1) can be formulated as follows,

$$\text{(P1)} \ \max \sum_{u=1}^{|\mathbb{U}|} \sum_{k=1}^{K} \frac{1}{K} \cdot R_{s_u u}^k, \quad \text{s. t. Constraints } (7)\text{–}(11). \tag{12}$$

From problem P1, we can observe that the constraint (11) and the objective function (12) have nonlinear terms. Accordingly, it is a MINLP, where $b_{s_u u}^k$ is the binary variable, $t_{w_j(u)}^s$ and $t_{w_j(u)}^e$ are integer variables, and $R_{s_u u}^k$ is the real variable. The transmission data for each service point cannot be determined until transmission path and scheduling time are determined. This problem is more complex compared with the 0-1 Knapsack problem, which is NP-complete [14], [29]. Thus, we propose a high throughput service scheduling scheme to solve it with good performance as well as low complexity in next section.

IV. The TESS Scheme

In this section, we propose a throughput efficient scheduling scheme for the formulated problem above. As mentioned before, to overcome dynamics due to human mobility, we achieve the transmission scheduling of each service point in space and time. For the space scheduling, we develop a relay path planning algorithm. For the time scheduling, we propose a global time scheduling algorithm. Finally, we extend and strengthen the scheme to the scenario of multiple cells with multiple BSs.

To illustrate this further, we introduce the decoupling of space and time first. For each relay node, we need to determine its activity in the whole scheduling, including when to transmit data and its corresponding receiver node. In fact, determining the receiver node is to construct multi-hop relay paths to maximize the system throughput with objective function (12), while determining the transmission time is to satisfy the time constraints (7)-(11) in Section III-C. Thus, the whole scheduling can be decoupled. Without considering the time constraints, we first utilize the relay path planning algorithm to establish the multi-hop relay path from the BS to each service point. Then we apply the global time scheduling algorithm for the established path to accommodate the time constraints. The details of algorithms are as follows.

A. Relay Path Planning

For the best channel condition, the nearest RE to u is selected as s_u, responsible for caching and serving UE. Based on this, we establish a multi-hop relay path from the BS to s_u with pre-determined hops, borrowing the ideas of Dijkstra Algorithm [30]. If the found path's hop number does not equal to the pre-determined one, we remove the path, and re-find a new path with pre-determined hops.

We denote the destination node s_u by A. The set of selected paths is denoted by $\mathbb{P}$, and the hop number for each path $p \in \mathbb{P}$ is denoted by H_p. For each path $p \in \mathbb{P}$, the last node on p is denoted by L_p. Especially, the selected path from B_s to A is denoted by p_*, and the pre-determined hop number for p_* is denoted by H_*. Besides, we denote $\mathbb{N}_V$ as the set of visited nodes in planning. We also denote d_{AV} as the minimum distance from A to other nodes in $\mathbb{V}$, to indicate whether there exists the path to A. If d_{AV} equals to infinity, no path exists from B_s to A with H_* hops.

The pseudo-code of the relay path planning algorithm is shown in Algorithm 1. We name the algorithm as a function **PathPlanning**, as in line 3. The algorithm iteratively searches for the relay path with H_* hops, until the path is found or there is no path can be found, as illustrated in line 5. To be specific, lines 6-9 extend the relay path by finding the next node c_s with the shortest distance r_s to visited nodes. When the visited node is B_s, the algorithm generates a new path from B_s to c_s, and adds to $\mathbb{P}$, as indicated lines 10-11. Otherwise, the algorithm finds the corresponding path that the visited node is on, and extends the path to c_s, as in lines 13-15. If the newly added path reaches the determined node A, i.e., c_s is A, lines 16-20 examine whether the hop number equals to H_*. If so, the relay path p_* is obtained and returned, as in lines 16-18; otherwise, the algorithm sets the distance of the last hop on p, d_{As}, to infinity, so that the failed path can be excluded from the next iteration. Then the algorithm sets $\mathbb{P}$, $\mathbb{N}_V$, and $\mathbb{V}$ as the initial value for searching the possible path again. If c_s is not A, the algorithm will add c_s to $\mathbb{N}_V$, and remove it from $\mathbb{V}$, as shown in lines 21-22. The shortest distance from A to other nodes in $\mathbb{V}$ is updated in line 23. Finally, the function returns the relay path p_* in line 24.

From Algorithm 1, we can infer that the outer while loop has no more than $2|\mathbb{V}|$ iterations, and the inner for loop has at most $|\mathbb{V}|$ iterations. Consequently, the worst-case computational complexity for Algorithm 1 is $\mathcal{O}(|\mathbb{V}|^2)$, which is achievable in practice.

Algorithm 1 Relay Path Planning Algorithm

1 **Input:** The set of REs in the cell, $\mathbb{V}$;
2 The pre-determined hop number for p_*, H_*;
3 **Function** $p_* =$ **PathPlanning** ($\mathbb{V}$, H_*, A)
4 **Initialization:** $\mathbb{N}_V = \mathbb{B}_s$, $\mathbb{P} = \emptyset$, $p_* = \emptyset$, $d_{AV} = 0$, $\mathbb{V}_0 = \mathbb{V}$;
5 **while** $A \notin \mathbb{N}_V$ and $d_{AV} \neq \infty$ **do**
6 **for** each node $i \in \mathbb{N}_V$ **do**
7 Find the relay node $v \in \mathbb{V}$ with the shortest distance to node i;
8 $r_i = d_{vi}$, $c_i = v$;
9 Find the node $s \in \mathbb{N}_V$ with the minimum r_s;
10 **if** $s == B_s$ **then**
11 $p = \{B_s \rightarrow c_s\}$, $\mathbb{P} = \mathbb{P} \cup p$;
12 **else**
13 Find the path $p_0 \in \mathbb{P}$ with $L_{p_0} = s$;
14 Obtain p by extending p_0 to $c_s, p = \{p_0 \rightarrow c_s\}$;
15 $\mathbb{P} = \mathbb{P} \cup p$;
16 **if** $c_s == A$ **then**
17 **if** $H_p == H_*$ **then**
18 $p_* = p$, $\mathbb{N}_V = \mathbb{N}_V \cup A$;
19 **else**
20 $\mathbb{P} = \emptyset$; $d_{As} = \infty$; $\mathbb{N}_V = \mathbb{B}_s$; $\mathbb{V} = \mathbb{V}_0$;
21 **else**
22 $\mathbb{N}_V = \mathbb{N}_V \cup c_s, \mathbb{V} = \mathbb{V} - c_s,$;
23 Update the shortest distance from A to other nodes in $\mathbb{V}, d_{AV}$;
24 **Return** p_*.

B. Global Time Scheduling

With Algorithm 1 solving the space path planning, now we consider the concurrent time scheduling algorithm. Since UE reaches service points by order, the scheduling is from the earliest service point to the last one. For each service point, we construct the relay path with pre-determined hops by Algorithm 1. Then multiple judgements are utilized to make sure that, the relay path could finish data cache before UE's arrival. In each time slot, concurrent transmissions of links are enabled with half-duplex assumption satisfied. If judgements are not satisfied, to accommodate the scheduling, the algorithm decreases the hop number or data amount by multiplying a shrinking parameter below 1.

To be clear, we introduce the notations first. The mobility trajectory is divided into K time slots. For each service point $u \in \mathbb{U}$, t_u, τ_u, and D_u are denoted as UE's arrival time, stay time, and received data at u. Respectively, the number of remaining available time slots before UE's arrival is denoted by $r\tau_u$. For simplicity, these variables at service points are expressed in vector form, i.e., $\mathbf{t}, \boldsymbol{\tau}, \mathbf{D}$, and $\boldsymbol{r\tau}$. Besides, to track the state of each RE and BS in K time slots, we define a binary table $\mathbf{S}_{(|\mathbb{V}|+1)\times K}$. For instance, $\mathbf{S}(1,k) = 1$ means RE 1 is scheduled or busy in the k-th time slot, and 0 means unscheduled or free. As for the path, the maximum hop number for relay paths is denoted by H_m. The hop number for the relay path from BS to s_u is denoted by H_u, and the path is denoted by p_u. The hth hop on p_u is denoted by p_u^h. For p_u^h, we denote st_u^h as the start time slot for transmission, T_u^h as the number of transmission time slots. Moreover, b_u^h is denoted as busy time slots for p_u^h (transmitter or receiver of p_u^h is busy).

The pseudo-code for the global time scheduling algorithm is presented in Algorithm 2. It iteratively schedules service points one by one according to the arrival time order. For each service point u, lines 8-9 find the nearest RE s_u, and derive

Algorithm 2 Global Time Scheduling Algorithm

1 **Input:** The set of service points $\mathbb{U}$, the set of REs $\mathbb{V}$;
2 Arrival time vector $\boldsymbol{t}$ for $\mathbb{U}$;
3 Stay time vector $\boldsymbol{\tau}$ for $\mathbb{U}$;
4 The maximum hop number H_m;
5 The shrinking parameter β;
6 **Initialization:** $\boldsymbol{r\tau} = \mathbf{0}$, $\mathbf{D} = \mathbf{0}$, $\mathbf{S} = \mathbf{0}$;
7 **while** $|\mathbb{U}| > 0$ **do**
8 Find the nearest RE s_u with the minimum t_u;
9 $D_u = R_{s_u u} \cdot \tau_u$;
10 $H_u = H_m$;
11 **while** $H_u > 1$ **do**
12 p_u=**PathPlanning** ($\mathbb{V}$, H_u, s_u);
13 **if** $p_u \neq \emptyset$ **then**
14 Obtain the rate of p_u^i, R_u^i;
15 $r\tau_u = t_u - \sum_{i=1}^{H_u} \frac{D_u}{R_u^i}$;
16 **if** $r\tau_u \geq 0$ **then**
17 $\mathbf{S}' = \mathbf{S}, h = 1$;
18 **while** $h \leq H_u$ **do**
19 Obtain start time slot for p_u^h, st_u^h;
20 Obtain transmission time slots for p_u^h, $T_u^h = \frac{D_u}{R_u^h}$;
21 Check p_u^h state with $\mathbf{S}'$, and obtain the number of busy time slots in $st_u^h \backsim (st_u^h + T_u^h)$ time slots, b_u^h;
22 **if** $b_u^h \leq r\tau_u$ **then**
23 Skip busy time slots, and schedule p_u^h in free $r\tau_u$;
24 Update $r\tau_u, \mathbf{S}'$;
25 $h = h + 1$;
26 **else**
27 Go to line 30;
28 Update $\mathbf{S} = \mathbf{S}'$;
29 Go to line 34 for the next service point;
30 $H_u = H_u - 1$;
31 $D_u = \lfloor D_u \cdot \beta \rfloor$;
32 **if** $D_u > 0$ **then**
33 Go to line 10 with less transmission data;
34 $\mathbb{U} = \mathbb{U} - u$;
35 **Return** received data D_u at each service point u.

the maximum transmission data D_u. Line 10 initializes H_u to the maximum hop number for paths. In lines 11-30, the algorithm tries to construct relay paths with as many hops as possible, and schedules transmissions for links on the relay path into appropriate time slots. Based on Algorithm 1, a relay path p_u with H_u hops is generated, and rates for hops on p_u are also obtained, as in lines 12-14. Then in line 15, without considering busy state, the algorithm calculates p_u's remaining time $r\tau_u$ before UE's arrival. If time is not enough, line 30 decreases H_u for a new path. Otherwise, in line 17, the temporary state table $\mathbf{S}'$ and count variable h are initialized for multiple judgements and links scheduling in lines 18-27.

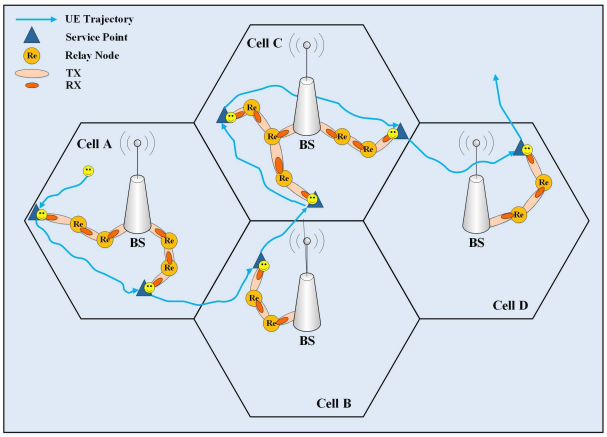


Fig. 4. An example of TESS scheme for the multi-cell scenario.

We hope p_u^h to be available during $st_u^h \backsim (st_u^h + T_u^h)$ time slots, yet the hop may be busy at some time slots by checking $\mathbf{S}'$, as shown in lines 19-21. If the algorithm can skip all busy time, and schedule p_u^h in free remaining time instead, the schedule goes to the next hop with $r\tau_u$, $\mathbf{S}'$ and h updated, as in lines 22-25. With all hops on p_u^h scheduled, the algorithm updates $\mathbf{S}$, and schedules for the next service point, as in lines 28-29. On the other hand, lines 26-27 decreases H_u if $r\tau_u$ is not enough. Furthermore, when H_m is decreased to 1, the algorithm decreases the transmission data by multiplying β, as in line 31. With less transmission data, the algorithm tries new scheduling for u. However, if D_u is decreased to 0, which indicates the scheduling for u is failed, the algorithm will start to schedule for the next service point, as in lines 32-34. For UE, received data D_u at u are returned in line 35.

For Algorithm 2, the while loop in line 7 has at most $|\mathbb{U}|$ iterations, and the while loop in line 11 has no more than H_m iterations. In the while loop, the complexity of function **PathPlanning** is $\mathcal{O}(|\mathbb{V}|^2)$ according to Algorithm 1. However, the while loop and judgements in lines 18-27 have about KH_m iterations, generally more than the function. Thus, the worst-case computational complexity for Algorithm 2, i.e., the TESS scheme, is $\mathcal{O}(K|\mathbb{U}|H_m^2)$, which can be implemented in practice.

C. Multi-Cell Scheduling

Generally, the daily movement and trajectory of users are across multiple cells with frequent handovers, which remains a significant challenge for mobility-aware transmission scheduling. To address the challenge as well as improve the robustness of TESS scheme, we extend and strengthen the algorithms in the multi-cell case.

To make it easier to follow the extension of TESS scheme, we present an example in Fig. 4. To simplify the expression, we adopt the traditional grid cellular model in the figure. For the trajectory with several service points in multiple cells, the TESS scheme first classifies service points by location, and obtains the part-trajectory in each mm-wave cell, as shown in Fig. 4. Subsequently, the relay path planning algorithm and global time scheduling algorithm above are executed in each

Algorithm 3 Multi-Cell Scheduling Algorithm

1 **Input:** The set of service points $\mathbb{U}$, the set of REs $\mathbb{V}$; The set of cells $\mathbb{C}$;
2 Arrival time vector $\boldsymbol{t}$ for $\mathbb{U}$;
3 Stay time vector $\boldsymbol{\tau}$ for $\mathbb{U}$;
4 The maximum hop number H_m;
5 The shrinking parameter β;
6 **Initialization:** $\boldsymbol{r\tau} = \mathbf{0}$, $\mathbf{D} = \mathbf{0}$, $\mathbf{S} = \mathbf{0}$, $\mathbb{U}_{c\in\mathbb{C}} = \emptyset$;
7 **for** each service point $u \in \mathbb{U}$ **do**
8 According to the location, classify u into the corresponding cell $c \in \mathbb{C}$;
9 $\mathbb{U}_c = \mathbb{U}_c \cup u$;
10 **for** each cell $c \in \mathbb{C}$ **do**
11 Call Algorithm 2 with $\mathbb{U}_c$;
12 **Return** received data D_u at each service point u.

mm-wave cell. For example, when UE is at cell A, multi-hop relay transmission paths are concurrently established in cell B, C, and D. When UE reaches the service point, the closest relay node can provide the service. Finally, the system throughput and scheduling are returned by multiple cells. Therefore, the extended TESS scheme is robust to trajectories across multiple cells.

The pseudo-code for the multi-cell scheduling algorithm is presented in Algorithm 3. We denote $\mathbb{C}$ as the set of cells, and $\mathbb{U}_c$ as the set of service points in the coverage of cell c. For each service point u, lines 7-10 classify it into the nearest mm-wave cell, and extend the service point set of the cell. Then the algorithm iteratively calls Algorithm 2 to obtain the scheduling scheme in each cell, as shown in lines 10-11. In the end, the received service data at each service point is returned. Since the multi-cell scheduling scenario utilizes the same algorithms with the single-cell scenario, the computational complexity for the multi-cell scheduling is still $\mathcal{O}(K|\mathbb{U}|H_m^2)$.

V. Performance Evaluation

In this section, we evaluate the performance of our proposed TESS scheme in realistic environments driven by human mobility trajectories, by investigating the impact of various system parameters on the system throughput. The performance of TESS scheme is evaluated in the single-cell scenario and the multi-cell scenario, respectively.

A. Evaluation Setup

For the evaluation of the single-cell scenario, we use human mobility trajectories, collected in the campus of the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. The trajectories are generated from 32 students who live in a campus dormitory, recorded by Garmin GPS 60CSx handheld receivers. In addition, these GPS devices achieve a position accuracy of better than 3 m 95% of the time [31]. 88 UE's trajectories are evaluated in a square area of $300m \times 300m$. For generality and convenience, we have the assumption that, the BS locates in the center of the area, and several static REs are deployed in the area with Poisson distribution.

On the other hand, we target the multi-cell scenario by considering realistic human mobility trajectories with real-world GPS information, which are obtained from WeChat (a Chinese social mobile application) users in Beijing. By collecting trajectories of 1000 users in 45 days, 100 typical UE's trajectories are chosen and evaluated in a $3000m \times 3000m$ square area with 100 BSs deployed on a regular grid network. Each BS locates in the center of $300m \times 300m$ square cell, and manages communications in its cell. Accordingly, several static REs are deployed in each cell, and the deployment is subject to Poisson distribution with the same expected value λ (also noted as Relay Number).

To be close to the real environment, the reference antenna model in IEEE 802.15.3c is adopted in the evaluation, exhibiting a main lobe of Gaussian form and constant level of side lobes [32]. We denote the antenna gain (dB) of by $G(\theta)$, which can be expressed as,

$$G(\theta) = \begin{cases} G_0 - 3.01 \cdot \left(\frac{2\theta}{\theta_{-3\mathrm{dB}}}\right)^2, & 0^\circ \le \theta \le \theta_{ml}/2; \\ G_{sl}, & \theta_{ml}/2 \le \theta \le 180^\circ, \end{cases} \tag{13}$$

where $\theta \in [0^\circ, 180^\circ]$ is the angle to main lobe center, $\theta_{-3\mathrm{dB}}$ is the angle of the half-power beamwidth, and θ_{ml} denotes the angle of the main lobe. The relationship between θ_{ml} and $\theta_{-3\mathrm{dB}}$ can be derived as $\theta_{ml} = 2.6 \cdot \theta_{-3\mathrm{dB}}$. On the other hand, G_0 denotes the maximum antenna gain, expressed as $G_0 = 10\log(\frac{1.6162}{\sin(\theta_{-3\mathrm{dB}}/2)})^2$. G_{sl} denotes the side lobe gain, expressed as $G_{sl} = -0.4111 \cdot \ln(\theta_{-3\mathrm{dB}}) - 10.579$. For every result, we perform 30 independent experiments with RE deployments for each trajectory, and obtain the results by the mean of trajectories. The propagation-related parameters and other related parameters are summarized in Table I. Due to similar scenarios, we adopt the channel model defined in 802.11ad TG [33] and the bandwidth from the standard of IEEE 802.11ad [34]. Especially, both P_t and θ_{-3dB} are determined by practical system conditions. Other parameter settings including β, H_m and $|\mathbb{V}|$ are based on the analysis in Section V-B.1 and Section V-C.1 for general cases.

In this evaluation, we compare our proposed TESS scheme with traditional unicast scheme, in which the BS directly transmits to UE at the service point. In addition, the system throughput defined in (6) is the main performance metric. To analyse the impact of algorithm parameters, we investigate the relationship between the system throughput and RE group size (means expected value in this paper), maximum number of hops H_m, as well as shrinking parameter β. We also look at the relationship between the system throughput and transmission power P_t, as well as half-power beamwidth θ_{-3dB} to analysis the impact of system hardware parameters.

B. Performance Comparison in the Single-Cell Scenario

1) Algorithm Parameter Analysis: To evaluate the role played by RE group sizes, we adjust the number of relays in the system and plot the system throughput in Fig. 5. Other simulation parameters are shown in Table I. From the results,

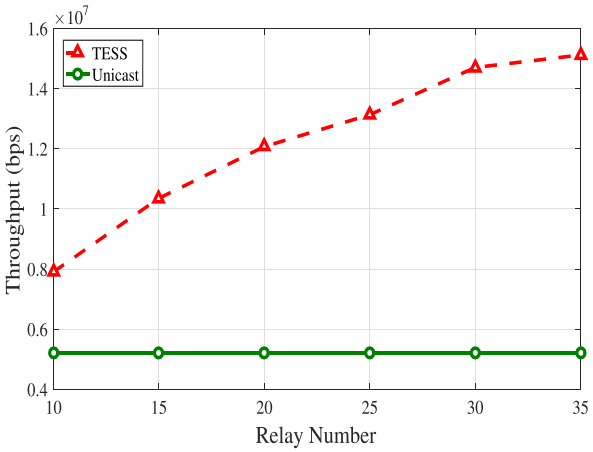


Fig. 5. The system throughput comparison of two schemes under different RE group sizes (single-cell scenario).

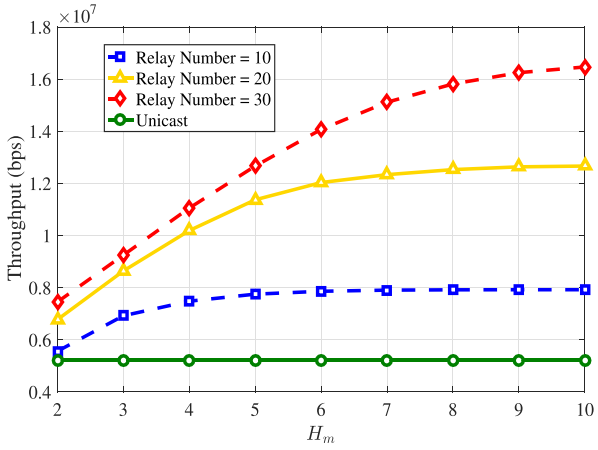


Fig. 6. The system throughput comparison of two schemes with different RE group sizes under different maximum number of hops (single-cell scenario).

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Value
Bandwidth	W	2160 MHz
Background noise	N_0	-134 dBm/MHz
Path loss exponent	γ	2
Efficiency of the transceiver design	η	0.5
Time slot duration	Δ	1 s
Transmission power	P_t	30 dBm
Half-power beamwidth	$\theta_{-3\text{dB}}$	30°
Shrinking parameter	β	0.9
Maximum number of hops	H_m	7
Relay group size (expected value)	$\lvert\mathbb{V}\rvert$	20

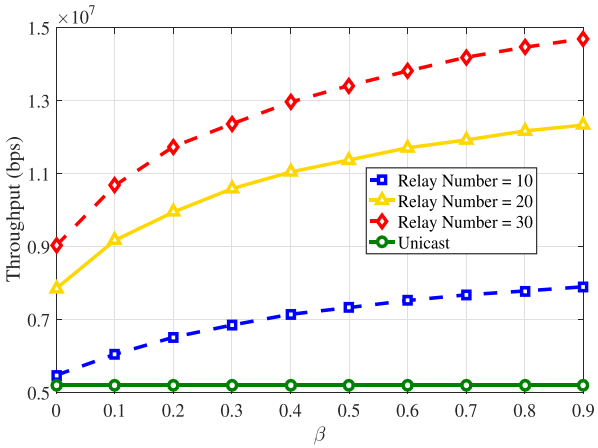


Fig. 7. The system throughput comparison of two schemes with different RE group sizes under different shrinking parameters (single-cell scenario).

we can clearly observe that, TESS improves the system throughput by about 118.2% compared with the unicast scheme when the RE group size is 20. Since relays in physical proximity are advantageous for relaying communications, more REs means shorter distance and higher system throughput. With the group size varying from 10 to 35, TESS improves the system throughput by 87.5%. However, owing to relays not enabled in the unicast scheme, its throughput does not change with the RE group size. It seems costly that such relays are deployed for only one user. Nevertheless, content popularity has been found to follow the classic Zipf's law [35], which means a small amount of popular contents are always requested by most users. For example, the data including videos cached at relays can also serve other users. According to this, TESS is a potential and promising solution for the scenario with multiple mobile users.

We study the gain of the maximum number of hops, and plot the system throughput comparison of two schemes with different RE group sizes under different H_m in Fig. 6. As the maximum number of hops increases, the system throughput of TESS increases. With more hops, more paths with larger transmission data can be scheduled for service points. Hence, higher system throughput can be achieved. We can observe that when H_m is small like 1 or 2, the gaps between different RE group sizes are not large since the choice for relaying paths is few. When H_m increases, the gaps become quite large, which demonstrates the advantages of multi-hop relaying communications. We also observe that when H_m is 5, the system throughput of TESS with 10 REs stop changing with H_m, and similar H_m for TESS with 20 REs is 7, while the system throughput of TESS with 30 REs is still increasing, which indicates a positive correlation between the maximum number of hops and RE group sizes. However, larger H_m also stands for more computational time. Thus, we should select the suitable value of H_m according to the RE group size and time restrictions. As for the unicast scheme, its throughput does not change with H_m.

Considering the impact of shrinking parameter on transmitted data, Fig. 7 exhibits the system throughput of two schemes with different RE group sizes under different β. Since β is a parameter to reduce the transmitted data D_u to $\beta \cdot D_u$ during scheduling, the transmitted data for service points may decrease a lot if β is too small. For example, we can observe that when RE group size is 20, TESS with $\beta = 0.9$ enhances the system throughput by about 57.7% compared with $\beta = 0$. Furthermore, we also observe that

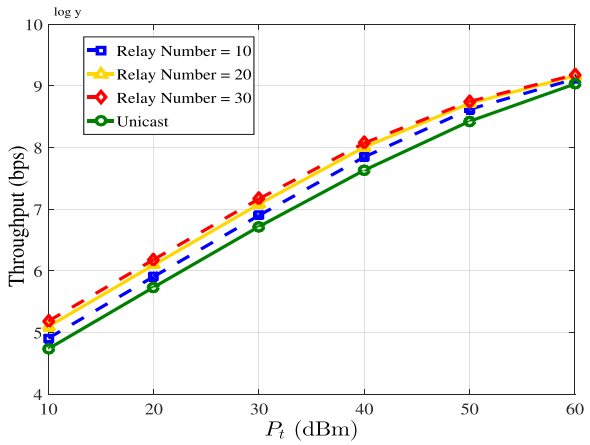


Fig. 8. The system throughput comparison of two schemes with different RE group sizes under different transmission power (single-cell scenario).

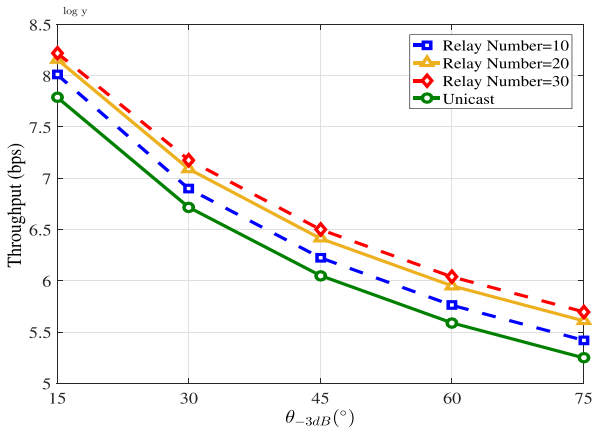


Fig. 9. The system throughput comparison of two schemes with different RE group sizes under different half-power beamwidths (single-cell scenario).

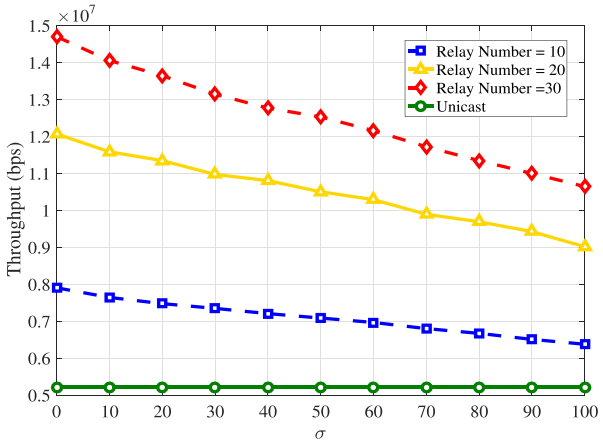


Fig. 10. The system throughput of two schemes under different prediction error (single-cell scenario).

TESS with 30 REs under $\beta = 0.3$ can achieve a little higher system throughput than TESS with 20 REs under $\beta = 0.9$. Accordingly, β should be selected according to the RE group size and system throughput requirements. Although small β decreases the system throughput significantly, when $\beta = 0$, TESS with 20 REs still improves the system throughput by 51.9% compared with the unicast scheme. The unicast scheme does not change with β for no shrinking operations.

2) System Parameter Analysis: In order to evaluate the impact of transmission power, we show the system throughput with different RE group sizes under different P_t in Fig. 8, where Y-axis is the logarithmic coordinates for better view. With the increase of the transmission power, the system throughput of different schemes increases due to higher transmission rates. Moreover, the increase trend is almost linear for P_t under 40 dBm. Specifically, the system throughput increases nearly 10 times with a tenfold increase in transmission power. Though lower transmission power means lower energy consumption for the network, it also leads to lower system throughput. Therefore, the transmission power should be selected according to the practical energy consumption and system throughput requirements. We also observe that TESS still has greater system throughput than the unicast scheme, and the larger RE group size brings larger gains.

To illustrate the impact of half-power beamwidth, we plot the system throughput of two schemes with different RE group sizes under different θ_{-3dB} in Fig. 9. We examine five cases, with θ_{-3dB} equal to 15°, 30°, 45°, 60°, and 75°, respectively. The results are shown with Y-axis using the logarithmic coordinates. Again, TESS achieves greater system throughput than the unicast scheme. Due to the high directivity of narrow beam antennas, network with smaller θ_{-3dB} (narrow beam antennas) can achieve higher antenna gains and transmission rates. For the case of 30 REs, as θ_{-3dB} narrows from 60° to 30°, the system throughput increases over tenfold. Taking this into consideration, the system throughput decreases as θ_{-3dB} increases.

From the results analysis above, we can obtain that our proposed TESS in the single-cell scenario improves the system throughput significantly compared with traditional unicast scheme under various parameters. Besides, to balance the system throughput and energy consumption as well as computational time, we should consider the factual RE group size and select appropriate system parameters. All these results and insights provide the guideline for the algorithm development and protocol design for mm-wave communications with human mobility.

3) Prediction Analysis: Since our proposed scheme relies on predicted trajectories, we present the impact of prediction error on the system throughput in Fig. 10. To evaluate the prediction error, we assume that the position prediction error $(\Delta x, \Delta y)$ is subject to two-dimensional normal distribution $N(\binom{0}{0}, \left(\begin{smallmatrix}\sigma & 0\\ 0 & \sigma\end{smallmatrix}\right))$, where Δx is the prediction error in X-axis, and Δy is for Y-axis, respectively. We can observe that the system throughput decreases globally with the increase of prediction error, which agrees with intuitions. For example, the system throughput only decreases by 21% with 10x increase in the prediction error. Since prior-known trajectories are not enabled in the unicast scheme, the baseline performance does not change with the prediction error. Despite this, TESS still outperforms the baseline with σ below 100. Furthermore, due to the spatial reuse and capacity improvement,

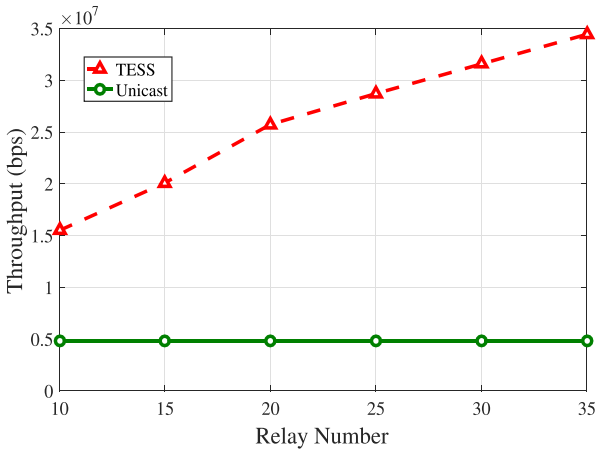


Fig. 11. The system throughput comparison of two schemes under different RE group sizes (multi-cell scenario).

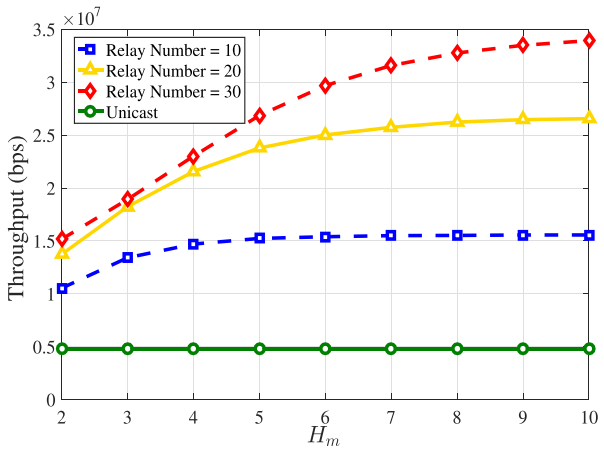


Fig. 12. The system throughput comparison of two schemes with different RE group sizes under different maximum number of hops (multi-cell scenario).

TESS with more relays are more tolerant of the prediction error. Referring to the prediction accuracy in [36] and [37], our proposed scheme is robust to the prediction error.

C. Performance Comparison in the Multi-Cell Scenario

1) Algorithm Parameter Analysis: In this subsection, we verify the effectiveness and feasibility of the TESS scheme in the multi-cell scenario. We first validate the relationship between the system throughput and the number of relays in Fig. 11. Compared with results in Fig. 5, it can be seen that the increase trend of TESS with relay number is similar. Moreover, it is worth noting that, orders of magnitude in these two figures are close, on the order of 10^7 bps. The phenomenon proves the correctness of our proposed scheme. Although the test dataset and application scenario are quite different from evaluation in Section V-B, our proposed scheme still provides a 5x increase in the system throughput when the RE group size is 20, which proves the robustness and effectiveness of TESS scheme.

The system performance versus the maximum number of hops under different RE group sizes is shown in Fig. 12, where the effectiveness of the TESS is again validated. Bigger H_m means less path loss between the BS and service points, and provides more chance to finish the global time scheduling of service points. Since the REs deployment in each cell is subject to Poisson distribution, the real number of relays is hard to exceed 15 for the case of Relay Number $= 10$ (with the probability less than 0.05). Thus, the system throughput stops changing when $H_m = 6$, for no more appropriate relays to be utilized. Similar to results in Fig. 6, there is a positive correlation between the maximum number of hops and RE group sizes in Fig. 12.

In Fig. 13, we reveal the relationship between the system throughput and the shrinking parameter β with different RE group sizes. It can be observed that, the system throughput increases with β, which is in accordance with the trend exhibited in Fig. 7. However, the increase trend in multi-cell scenario is much slower. Taking the curve of Relay Number as 20 in Fig. 7 as an example, the system throughput increases by 67% from $\beta = 0$ to $\beta = 0.9$. In contrast, the increase ratio

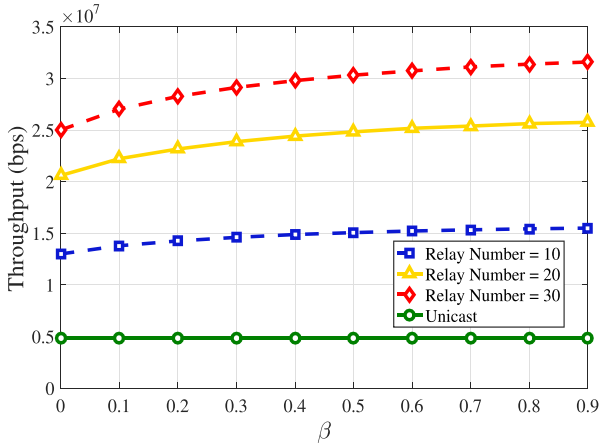


Fig. 13. The system throughput comparison of two schemes with different RE group sizes under different shrinking parameters (multi-cell scenario).

for the curve in Fig. 13 is only 25%. A possible explanation to the phenomenon is that, the number of service points in human mobility trajectories is similar, while the trajectories across multiple cells leave more time for scheduling in each cell. Especially, in the multi-cell scenario, the number of service points in each cell is rather less than it in the single-cell scenario. Therefore, the global time scheduling in each cell of multi-cell scenario is more possible to be finished with less shrinking operations. As a result, the shrinking parameter has less to do with system performance compared with the single-cell scenario. On the other hand, larger β also means more computational time, which can be obtained from Algorithm 2. Thus, selecting the suitable value of β according to its effect and time restriction should be considered in the multi-cell scenario.

2) System Parameter Analysis: To investigate the impact of transmission power on the system throughput in the multi-cell scenario, we plot their relationship with different RE group sizes in Fig. 14. The results are shown with Y-axis using the logarithmic coordinates. Again, results in the multi-cell scenario are similar with the analysis in Fig. 8. It can be seen that, the system throughput achieved by TESS with Relay Number $= 30$ and $P_t = 20$ dBm is close to the performance

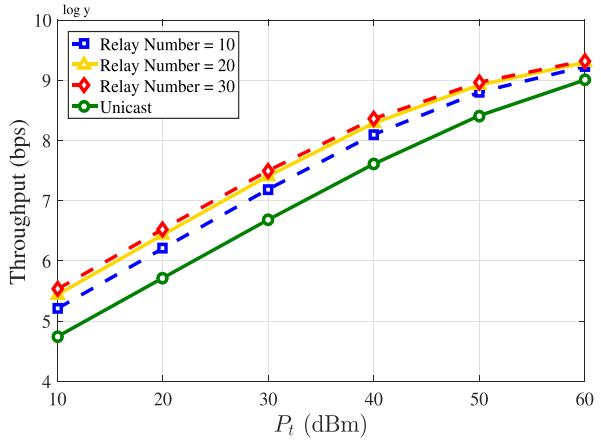


Fig. 14. The system throughput comparison of two schemes with different RE group sizes under different transmission power (multi-cell scenario).

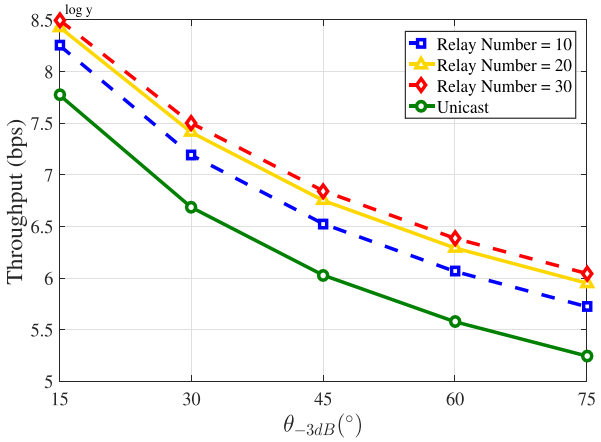


Fig. 15. The system throughput comparison of two schemes with different RE group sizes under different half-power beamwidths (multi-cell scenario).

achieved by the unicast scheme with $P_t = 30$ dBm. In this case, the benefits brought by concurrent relay transmissions can be utilized to reduce energy consumption. Furthermore, TESS can adjust the transmission power differently according to the real relay density in each cell for energy saving.

Finally, considering the multi-cell scenario, we show the system throughput comparison with different RE group sizes under different half-power beamwidths, i.e., θ_{-3dB}, in Fig. 15. The Y-axis is in the logarithmic coordinates. It is clear that the results are similar with the case in the single-cell scenario. The gap between unicast scheme and TESS scheme is quite large than it in the single-cell scenario, caused by different datasets and related parameters. From these two result figures above, we can find that system parameters (P_t, and θ_{-3dB}) have similar effects to the system performance in both single-cell scenario and multi-cell scenario.

In summary, the result analysis in Section V-B and Section V-C demonstrates our proposed scheme TESS significantly enhances the system performance compared with the unicast scheme in the multi-cell scenario. Therefore, the multi-cell scheduling provides potential solutions to mobility-aware transmissions in mmWave cells.

VI. Conclusion

In this paper, we proposed TESS for throughput efficient scheduling in mm-wave systems, which exploits multi-hop relay and concurrent transmissions based on UEs' mobility trajectories. In TESS, to overcome challenges of human mobility, we propose a relay path planning algorithm to obtain multi-hop paths, and a global time scheduling algorithm to determine the transmissions of multi-hop paths. Also, we extend TESS to the multi-cell scenario with the multi-cell scheduling algorithm. Finally, extensive performance evaluations in both the single-cell dataset and the multi-cell dataset demonstrate that TESS achieves significantly higher system throughput compared with the traditional unicast scheme.

Considering the content popularity in traffic demands, we will extend our scheme to multi-user content downloading scenarios in the future. Since the proposed scheme relies on prior-known mobility trajectories, we will also investigate the acquisition of users' mobility statistics in mm-wave systems. Furthermore, we will also investigate the delay-sensitive services for mobility-aware transmission to improve the network performance.

References

[1] (Mar. 28, 2017). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] 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.

[3] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, Nov. 2015.

[4] J. Wang *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[5] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[6] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[7] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[8] G. Shen, J. Liu, D. Wang, J. Wang, and S. Jin, "Multi-hop relay for next-generation wireless access networks," *Bell Labs Tech. J.*, vol. 13, no. 4, pp. 175–193, 2009.

[9] P. Herhold, E. Zimmermann, and G. Fettweis, "Cooperative multi-hop transmission in wireless networks," *Comput. Netw*, vol. 49, no. 13, pp. 299–324, Oct. 2005.

[10] M. Charafeddine, O. Oyman, and S. Sandhu, "System-level performance of cellular multihop relaying with multiuser scheduling," in *Proc. 47th Annu. Conf. Inf. Sci. Syst.*, Baltimore, MD, USA, Mar. 2007, pp. 631–636.

[11] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2052–2069, Oct. 2015.

[12] Y. Niu, C. Gao, Y. Li, L. Su, and D. Jin, "Exploiting multi-hop relaying to overcome blockage in directional mmwave small cells," *J. Commun. Netw.*, vol. 18, no. 3, pp. 364–374, Jun. 2016.

[13] J. Qiao, L. X. Cai, X. S. Shen, and J. W. Mark, "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.

[14] 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 ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5221–5225.

[15] 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, Jun. 2009.
[16] Y. Niu, Y. Liu, Y. Li, X. Chen, Z. Zhong, and Z. Han, "Device-to-Device communications enabled energy efficient multicast scheduling in mmWave small cells," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1093–1109, Mar. 2018.
[17] X. Ge, S. Tu, G. Mao, and C. X. Wang, "5G ultra-dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.
[18] D. Calabuig *et al.*, "Resource and mobility management in the network layer of 5G cellular ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 162–169, Jun. 2017.
[19] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmWave cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 40–47, Nov. 2016.
[20] T. Matsumura, K. Ibuka, K. Ishizu, H. Murakami, and F. Kojima, "Concept, design, and prototype of shared base station supporting millimeter waves for 5G cellular networks," in *Proc. Int. Symp. Wireless Pers. Multimedia Commun.*, Bali, Indonesia, Dec. 2017, pp. 297–303.
[21] M. Chen, Y. Hao, L. Hu, K. Huang, and V. Lau, "Green and mobility-aware caching in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8347–8361, Dec. 2017.
[22] J. Dai and D. Liu, "MAPCaching: A novel mobility aware proactive caching over C-RAN," in *Proc. IEEE PIMRC*, Montreal, QC, Canada, Oct. 2017, pp. 1–6.
[23] T. Liu, S. Zhou, and Z. N. Tsinghua, "Mobility-aware coded-caching scheme for small cell network," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.
[24] R. Wang, J. Zhang, and K. B. Letaief, "Incentive mechanism design for cache-assisted D2D communications: A mobility-aware approach," in *Proc. IEEE SPAWC*, Sapporo, Japan, Jul. 2017, pp. 1–5.
[25] D. Hunold and G. P. Fettweis, "A flexible mobility-aware transmission scheme utilizing smart antennas in a wireless cellular network," in *Proc. IEEE MoMuC*, San Diego, CA, USA, Nov. 1999, pp. 262–270.
[26] O. Yildiz and R. Sokullu, "A novel mobility aware downlink scheduling algorithm for LTE-A networks," in *Proc. ICUFN*, Milan, Italy, Jul. 2017, pp. 300–305.
[27] A. Khadka, K. Adachi, S. Sun, J. Wang, H. Zhu, and J. Wang, "Cooperative transmission strategy over users' mobility for downlink distributed antenna systems," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.
[28] 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.
[29] D. Pisingere, "Where are the hard knapsack problems?" *Comput. Oper. Res.*, vol. 32, no. 9, pp. 2271–2284, May 2005.
[30] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959.
[31] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the Levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, Jun. 2011.
[32] Q. Chen, X. Peng, J. Yang, and F. Chin, "Spatial reuse strategy in mmWave WPANs with directional antennas," in *Proc. IEEE GLOBECOM*, Anaheim, CA, USA, Dec. 2012, pp. 5392–5397.
[33] A. Maltsev, *Channel Models for 60 GHz WLAN Systems*, IEEE Standard 802.11-09-0344-07ad, Mar. 2010.
[34] *IEEE 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) Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput 60 GHz Band*, IEEE 802.11ad Standard, 2012.
[35] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, and Y. Grunenberger, "Is there a case for mobile phone content pre-staging?" in *Proc. CoNEXT*, Santa Barbara, CA, USA, Dec. 2013, pp. 321–326.
[36] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *Proc. IEEE WACV*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1186–1194.
[37] F. Altché and A. De La Fortelle, "An LSTM network for highway trajectory prediction," in *Proc. IEEE ITSC*, Yokohama, Japan, Oct. 2017, pp. 353–359.

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

Xinlei Chen received the B.E. and M.S. degrees in electrical engineering from Tsinghua University, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Department of Electric and Computer Engineering, Carnegie Mellon University, USA, in 2018. His research interests are in the areas of networking and communications, mobile embedded system, and big data.

Yong Niu (M'17) received the B.E. degree in electrical engineering from Beijing Jiaotong University, China, in 2011, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. He was a recipient of the 2018 International Union of Radio Science Young Scientist Award.

From 2014 to 2015, he visited the University of Florida, FL, USA, as a Visiting Scholar. He is currently an Associate Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His research interests are in the areas of networking and communications, including the millimeter wave communications, device-to-device communication, medium access control, and software-defined networks. He has served as Technical Program Committee Member of the IWCMC 2017, the VTC2018-Spring, the IWCMC 2018, the INFOCOM 2018, and the ICC 2018. He received the Ph.D. National Scholarship of China in 2015, the Outstanding Ph.D Graduates and Outstanding Doctoral Thesis of Tsinghua University in 2016, the Outstanding Ph.D Graduates of Beijing in 2016, and received the Outstanding Doctorate Dissertation Award from the Chinese Institute of Electronics in 2017. He was the Session Chair of the IWCMC 2017.

Bo Ai received the M.S. and Ph.D. degrees from Xidian University, China. He received a degree with the honor of Excellent Postdoctoral Research Fellow from Tsinghua University in 2007. He was a Visiting Professor at the Electrical Engineering, Stanford University, in 2015. He is currently with Beijing Jiaotong University as a Full Professor and a Ph.D. Candidate Advisor. He is the Deputy Director of the State Key Laboratory of Rail Traffic Control and Safety and the Deputy Director of International Joint Research Center. He is one of the main people responsible for the Beijing urban rail operation control system, International Science and Technology Cooperation Base, and is a backbone member of innovative engineering granted by the Chinese Ministry of Education and the State Administration of Foreign Experts Affairs.

He has authored/co-authored eight books and published over 300 academic research papers in his research areas. He holds 26 invention patents. He has been the research team leader for 26 national projects and received some important scientific research prizes. He has been notified by the Council of Canadian Academies that, based on Scopus database, that he has been listed as one of the top authors in his field all over the world. He has also been feature interviewed by IET *Electronics Letters*. His interests include the research and applications of channel measurement and channel modeling, and dedicated mobile communications for rail traffic systems.

Dr. Ai is a fellow of the Institution of Engineering and Technology and is an IEEE VTS Distinguished Lecturer. He has received many awards, such as Outstanding Youth Foundation from the National Natural Science Foundation of China, the Qiushi Outstanding Youth Award from the Hong Kong Qiushi Foundation, the New Century Talents from the Chinese Ministry of Education, the Zhan Tianyou Railway Science and Technology Award from the Chinese Ministry of Railways, and the Science and Technology New Star from the Beijing Municipal Science and Technology Commission. He was the co-chair or the session/track chair for many international conferences. He is the IEEE VTS Beijing Chapter Vice Chair and the IEEE BTS Xi'an Chapter Chair. He is currently an Editor of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS and an Editorial Committee Member of the *Wireless Personal Communications* journal. He is also the Lead Guest Editor of special issues in the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE ANTENNAS AND PROPAGATIONS LETTERS, and the *International Journal of Antennas and Propagations*.

Yong Li (M'09–SM'16) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University.

Dr. Li has served as a TPC member of several international workshops and conferences. He received the IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers Award and the Young Talent Program of China Association for Science and Technology Award. He is the general chair and a TPC chair of several international workshops and conferences. He is on the editorial board of two IEEE journals. His papers have total citations over 4100. Among them, ten are ESI highly cited papers in computer science and four received conference best paper (runner-up) awards.

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

He is currently an Associate Professor with Tsinghua University, where he is also the Vice Chair of the Department of Electronic Engineering. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture. He was a recipient of the National Scientific and Technological Innovation Prize (Second Class) in 2002.