IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 22, 2018, accepted June 24, 2018, date of publication July 17, 2018, date of current version August 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2856867

Efficient Time-Slot Adjustment and Packet-Scheduling Algorithm for Full-Duplex Multi-Hop Relay-Assisted mmWave Networks

WENSON CHANG, (Member, IEEE), CHIEN-WEN WU, AND YI-XIN LIN
Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan

Corresponding author: Wenson Chang (wenson@ee.ncku.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Contracts 105-2221-E-006-034 and 105-2221-E-006-032.

ABSTRACT Millimeter wave (mmWave) technology is one of the innovations with the most potential for the fifth generation (5G) of wireless communication systems. With wider bandwidth and the ability to provide high-quality services, 5G systems are expected to significantly improve network performance. However, mmWave communications are vulnerable to blockage problems. Fortunately, efficient relaying and scheduling scheme can both solve the blockage problem and improve performance for the concurrent transmissions. With this paper, we aim to enhance the efficiency of concurrent transmissions for the mmWave-based networks. For the purpose of serving the transmission requests within minimal time, the full-duplex relay and adaptive transmission rates are first considered for designing the scheduling algorithm. To further enhance the scheduling flexibility, a larger scheduling space is obtained by using smaller scheduling unit. To achieve this, a smaller time slot is properly redefined so that the packet-by-packet transmissions (rather than the burst-by-burst fashion in the conventional scheme) can be carried out. Then, we propose an efficient time slot adjustment scheduling algorithm (named ETA) for multihop transmissions that incorporates full-duplex relaying, adaptive rates, and packet-by-packet transmissions. Using the simulation results, we verify the superior performance of the proposed ETA scheduling algorithm in terms of the time required to serve transmission requests.

INDEX TERMS mmWave, concurrent transmission, full-duplex relay, scheduling, WPA network.

I. INTRODUCTION

Currently, millimeter wave (mmWave) technology is one of the innovations with the most potential for the next generation of wireless communication systems. With sufficient spectrum resources and spatial reusability, versatile network architectures can be constructed based on the mmWave networks, e.g., the small-cell networks [1], [2], wireless personal area networks (WPAN) [3], [4], Ad-hoc networks [5], [6] and device-to-device communications [7]. With the high-speed mmWave networks, several high-quality and high-throughput wireless services (e.g., the high definition television and wireless gigabit alliance) can be efficiently delivered [8]–[10]. Some important mmWave-based systems have also been standardized such as IEEE 802.15.3c, 802.11ad and ECMA-387. To make it a success, however, the key is how to effectively overcome the serious blockage problem in the mmWave networks by using the multi-hop relay. Moreover, to optimize concurrent transmissions, an efficient scheduling algorithm is necessary to arrange the multiple relaying flows so that the spatial resources are fully utilized.

In the literature, numerous relaying protocols have been proposed for solving the aforementioned blockage problem [4], [5], [10]–[12]. In [5], a discrete-time two-state Markov chain was applied to characterize the state transitions between line-of-sight (LOS) links and non-line-of-sight (NLOS) links for any particular transmission pairs. Then, the shortest path was selected such that the reliability can be guaranteed. Also, all the links which satisfy signal-to-interference-plus-noise ratio (SINR) requirement can be transmitted concurrently. For the single hop scenario, a greedy and a column-generation based algorithms were proposed to maximize the instant throughput for each time slot and iteratively improve the concurrent links, respectively. For the multi-hop scenario, however, the goal is to maximize the overall throughput of all the flows by properly selecting the links for concurrent transmissions based on the estimated link states and interference conditions. In [4], two relaying protocols were proposed for mmWave-based WPAN. The first minimizes the number of relays under connectivity, bandwidth and robustness constraints. The second maximizes the achievable rate using a fixed number of relays subject to the robustness constraint. In [11], two more relaying protocols were developed to maximize the end-to-end signal-to-noise ratio (SNR) and to alleviate the effect of the blockage problem, respectively. First, the SNR distribution of a particular relaying path is analyzed. The optimal path is the one with the optimal SNR distribution; the optimal relay is the one affected by the least path loss i.e., the blockage problem. In [12], the distributed auction algorithm was utilized to solve the joint client-relaying-access point (AP) association problem for an mmWave-based network. It was found that the proper AP association and relay mechanism substantially boost high data-rate services and increase connection reliability. Nevertheless, the fairness among the clients can also be improved. In [10], to facilitate relay selection, a relay priority region (RPG) was defined to candidate relays based on optimized time-splitting for half-duplex relaying. Properly selecting relays from the RPG further maximizes the end-to-end capacity of indoor mmWave-based networks.

In addition to using a suitable relay protocols, efficiently scheduling relay flows is critical to full utilization of the spatial resources of an mmWave-based system. In [13], to maximize the instant network throughput, a two-step heuristic algorithm was proposed to solve the constrained binary integer programming problem of link scheduling. In the first step, the optimal relaying path for each data flow was determined. In the second step, the developed links are scheduled to maximize the network throughput using the minimum number of time slots. Instead of throughput maximization, serving multiple transmission requests using the minimum time slots is critical to concurrent transmissions. To this end, Niu *et al.* [1], [14] (and [2]) proposed joint relay selection and scheduling algorithms for two-hop and multi-hop relay-assisted mmWave-based systems, respectively. In [6], the relay and link selection were jointly optimized for the dual-hop mmWave networks for reducing the delivery time. Two conditions were investigated therein. With enough relays, two sub-problems of link selection and relay assignments were formulated to solve the joint problem. Whereas, without enough relays, a heuristic algorithm was proposed to solve the nonliner integer programming problem. To provide an alternative to the half-duplex relay schemes, in [15], Qin *et al.* analytically proved the advantages of the full-duplex relay approach in the wireless systems. It was concluded that using the full-duplex relay scheme can enlarge the scheduling space to alleviate mutual interference. Most notably, the end-to-end throughput can be increased by more than twofold.

Motivated by [1], [2], [14], and [15], we find that in addition to full-duplex relaying, the smaller granularity of the packet-switched scheduling unit can also expand the scheduling space. This means that it is possible to reduce the required time slots but still serve the same number of transmission requests by using the full-duplex relaying scheme together with the smaller scheduling unit. The so-called scheduling unit is the minimum time slot allocation to serve a transmission request. For an example in [1], it was shown that given a fixed transmission rate of two packets per time slot, a transmission request of six packets (named a packet burst) needs three consecutive time slots. However, if the burst of six packets can be separately scheduled using the time slot equal to the packet duration, the scheduling space (i.e., flexibility) can be increased. To clarify, the packet burst can be transmitted using six inconsecutive time slots. This is what we call the smaller granularity of the packet-switched scheduling unit.

Accordingly, we aim to improve the efficiency of concurrent transmission for mmWave networks by increasing the scheduling space. To achieve this goal, we considers adaptive transmission rates and full-duplex relaying to redesign the conventional multi-hop relaying transmission (MHRT) scheme in [1] and [2]; which only consider fixed transmission rates and half-duplex relaying. In one of our considered cases, the ability to arrange the full-duplex transmission flows alone results in reduction of the transmission time by 24.5%. Furthermore, we adjust the aforementioned granularity of the scheduling unit (i.e., using the smaller scheduling granularity by reducing the time slot duration) according to the transmission rates. In principle, the time-slot length should be adjusted according to the minimum transmission rate so that the transmission time for a single packet consumes at least one short time slot. As a result, a request of a packet burst can be scheduled packet-by-packet rather than burst-by-burst as done by the conventional MHRT scheme. In this fashion, the scheduling space can be enlarged. Note that in the conventional MHRT scheme, multiple packets should be transmitted during a single long time slot and the request of packet bursts should be scheduled burst-by-burst. Using the proposed efficient time slot adjustment scheduling algorithm (ETA) with smaller granularity in addition to adaptive transmission rates and full-duplex relaying, the transmission time is further reduced by 27.0%. The total integrated advantages of the proposed ETA scheduling algorithm enhance performance by 44.9% in total.

To complete the discussion, we consider the variant antenna beamwidth and severity of the blockage problem. Additionally, diverse combinations of feasible transmission rates are included into the performance analysis. Simulation results show that the superiority of the proposed scheme is maintained for cases with higher and more variable feasible transmission rates. These results confirm the effectiveness of the proposed ETA scheduling algorithm in mmWave networks with serious blockage problems.

This paper is organized as follows. Section II describes the system and signal models, including the considered mmWave network topology. In Section III, we describe the motivation of this paper; then, we formulate and solve the mixed

TABLE 1. Summary of existing solutions for blockage problem in the mmWave-based networks, including the proposed ETA scheduling, where “O” and “X” mean yes and no, respectively. Note that “^{1}” is for Ad-hoc networks; “^{2}” and “^{3}” are for path and relay selections.

	Centralized algorithm	Multi-flow scheduling	Full-duplex relay	Adaptive rate	Stochastic blockage	N-hop (N>2)	Time-slot adjustment
ETA	O	O	O	O	X	O	O
[1], [2]	O	O	X	X	X	O	X
[5]	X^1	O	O	X	X	O	X
[4]	O	X^2	X	X	O	O	X
[11]	X^1	X^3	X	X	O	O	X
[12]	X^1	X^3	O	X	O	X	X
[10]	O	X^3	X	X	O	O	X
[13]	O	O	O	X	O	O	X
[14]	O	O	X	X	X	X	X
[6]	O	O	X	X	O	X	X

integer nonlinear scheduling problem by developing the ETA scheduling algorithm. Section IV presents the simulation results. Section V gives concluding remarks and suggestions for future works.

A. SUMMARY OF CONTRIBUTIONS

To emphasize the innovation of the proposed ETA scheme, Table 1 summarizes existing solutions for blockage problem in the mmWave-based networks. Moreover, the contributions of this paper are listed as follows.

1 The full-duplex relay has been taken into account for expediting the multi-hop packet forwarding process while eliminating the blockage problem for the mmWave communications.

2 Moreover, to enlarge the scheduling space, the time slot is properly redesigned so that smaller scheduling granularity can be utilized; and consequently, the packet-by-packet scheduling (rather than the burst-by-burst method in the conventional MHRT scheme) can be implemented.

3 Furthermore, the mechanism of adaptive transmission rate is considered for taking the impact of interference incurred by the concurrent transmissions into account.

4 To reflect the above considerations, the scheduling algorithm should be redesigned. To this end, the mixed integer nonlinear programming (MINLP) for developing the scheduling algorithm is firstly reformulated. Then, a linearization procedure shows that the reformulated MINLP can be solved by a well designed heuristic algorithm. At last, we propose the ETA scheduling algorithm to efficiently solve the blockage problem in the mmWave communications.

II. SYSTEM AND SIGNAL MODELS

A. SYSTEM DESCRIPTION

In this paper, an mmWave network consisting of N devices is considered. All devices are equipped with electronically steerable directional antennas such that directional transmission is possible between any two devices. To lead the time-slotted operation, one of the N devices is selected to be the special user equipment (UE), i.e., piconet controller (PNC); the other $N-1$ devices are regarded as the general UE. Each UE can operate in full-duplex relay mode. Following to the same assumptions in [1]–[3], [5], [6], [16], and [17], PNC can coordinate the operations of the mmWave networks. In addition to the synchronization process, PNC takes charge of the information exchanging to attain the current network topology, transmission requests and available transmission rates. Most importantly, PNC develops routing paths and arranges the packet transmission schedules for the mmWave networks.

The request matrix $\mathbf{R} = \{r_{ij}\} \in \mathbb{C}^{N\times N}$ is constructed to indicate the request of the i-th UE to transmit r_{ij} packets toward the j-th UE. Similarly, the rate matrix $\mathbf{C} = \{c_{ij}\} \in \mathbb{C}^{N\times N}$ is constructed to indicate the attainable transmission rate for link ℓ_{ij} between the i-th and j-th UEs. According to the received SNR (denoted by γ_{ij}) and adopted modulation and coding scheme (MCS), the transmission rate c_{ij} can be set to multiple packets per time slot. Note that the destination of ℓ_{ij} (i.e. the j-th UE) feedbacks the information of c_{ij} to the source (i.e. the i-th UE) by searching the table for the combinations of the MCS and received SNR. Then, the PNC collects the information of r_{ij} and c_{ij} from the i-th UE. Based on $\mathbf{R}$ and $\mathbf{C}$ matrixes, the first mission of the PNC is to develop the most time-efficient routing path for each transmission request i.e., r_{ij}). However, because of the blockage problem, path development may not be successful. After the path development phase, the PNC can schedule every hop (i.e., link) of each path to accomplish the transmission requests using the minimum number of time slots.

B. SIGNAL REPRESENTATION

Considering the 60-GHz mmWave, the received signal power of link ℓ_{ij} at time instant t can be expressed as

$$P_{ij}(t) = k_0 P_t d_{ij}^{-\alpha} G(\theta_{ij}) \chi_{ij}(t), \quad 1 \leq t \leq T, \qquad (1)$$

where $k_0 \propto (\lambda/4\pi)^2$ is a constant coefficient; λ is the wave length; P_t is the transmission power; d_{ij} is the distance between the i-th and j-th UEs; α is the path-loss exponent; T is the time required to serve the transmission request of $\mathbf{R}$; $G(\theta_{ij})$ is the antenna gain and θ_{ij} is the incidence angle. When using mmWaves, multiple access interference (MAI) can be incurred as the radio beams of an undesired combination of transmitter [e.g., the u-th UE in (5)] and receiver [e.g., the j-th UE in (5)] are aligned. To reflect this phenomenon, the cone-plus-sphere model [18] is used to generate the radiation beam pattern as follows:

$$G(\theta) = \begin{cases} \nu \dfrac{2\pi}{\theta_m}, & |\theta| \le \theta_m \\ (1-\nu)\dfrac{2\pi}{2\pi - \theta_m}, & |\theta| > \theta_m, \end{cases} \tag{2}$$

where θ_m is the beamwidth of the mainlobe and ν is the radiation efficiency. Moreover, $\chi_{ij}(t) \in \{0, 1\}$ indicates transmission from the i-th to j-th devices. Specifically, $\chi_{ij} = 1$ when the link between the i-th and j-th devices is activated; otherwise $\chi_{ij} = 0$. Also, owing to the full-duplex relaying, it possesses the following properties:

$$\sum_{j \in N} \chi_{ij}(t) \le 1, \quad \forall \ell_{ij} \in L_i^{out}, \ 1 \le t \le T \tag{3}$$

and

$$\sum_{j \in N} \chi_{ji}(t) \le 1, \quad \forall \ell_{ji} \in L_i^{in}, \ 1 \le t \le T, \tag{4}$$

where L_i^{out} represents the outgoing links from the i-th UE, and L_i^{in} represents the incoming links to the i-th UE.

Similar to (1), the MAI $\hat{I}_{ij}$ experienced by link ℓ_{ij} can be written as

$$\hat{I}_{ij}(t) = \sum_{u \ne i, u \ne j} \rho k_0 P_t d_{uj}^{-\alpha} G(\theta_{uj}) \chi_{uj}(t), \quad 1 \le t \le T, \tag{5}$$

where ρ is the signal's cross correlation. Additionally, because of the full-duplex relaying, self-interference (SI) can be incurred. Although some SI cancellation techniques can be used to alleviate the effects of SI, a certain amount of residual SI $\tilde{I}_{ij}$ is still observed [19]–[21], i.e.,

$$\tilde{I}_{ij}(t) = \beta P_t h_L \chi_{jj}(t), \quad 1 \le t \le T, \tag{6}$$

where β indicates the performance of SI cancellation and h_L is the loop-interference gain for the full-duplex relay. Also, χ_{jj} stands for the loop transmission. By definition, $\chi_{jj} = 1$ when the j-th device is used as a full-duplex relay; otherwise $\chi_{jj} = 0$ when the half-duplex relay is used. According to [19], the channel gain of $|h_L|$ can be as small as -100 dB; consequently, $\tilde{I}_{ij}$ is negligibly small compared with $\hat{I}_{ij}(t)$. According to (1), (5) and (6), the received signal-to-interference-plus-noise ratio (SINR) $\gamma_{ij}(t)$ of link ℓ_{ij} at time t can be defined as

$$\gamma_{ij}(t) = \frac{P_{ij}(t)}{WN_0 + \hat{I}_{ij}(t) + \tilde{I}_{ij}(t)}, \quad 1 \le t \le T, \tag{7}$$

where W and N_0 represent the system bandwidth and one-sided power spectral density of additive white Gaussian noise, respectively. Moreover, the minimum required SINR to sustain a transmission rate of c_{ij} is denoted by $\gamma_{\min}(c_{ij})$.

Under the effects of MAI, link ℓ_{ij} is dropped if its $\gamma_{ij}(t)$ can not reach the minimum requirement for sustaining the lowest transmission rate (i.e., the $\hat{\gamma}_{\min}$ defined in Section III.(D)). Alternatively, its attainable transmission rate may degrade if the $\gamma_{\min}(c_{ij})$ can not be reached. Generally, the SINR requirements of the adaptive transmission rates are predefined according to the considered MCS. The effect of the adaptive transmission rate is ignored in the conventional MHRT scheme in [1] and [2].

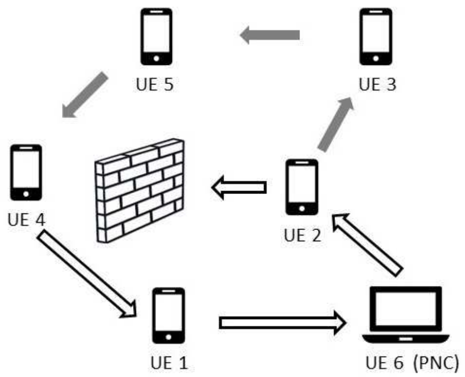


FIGURE 1. An illustrative example of the considered network topology; therein, owing to the blockage problem, the transmission request of the flow f_{24} should be routed through the path of 2 → 3 → 5 → 4.

III. EFFICIENT TIME SLOT ADJUSTMENT SCHEDULING

In this section, we firstly describe the motivation of this paper using an illustrative example. Then, we formulate the optimization problem and propose the ETA scheduling algorithm to reduce the time slots required for serving the transmission requests. Note that in the following example, the full-duplex relay is applied both in the conventional MHRT and ETA schemes for fair performance comparison. Most importantly, using the full-duplex as well as the smaller scheduling unit alone can not enhance the performance. To effectively utilize the full-duplex relay and smaller scheduling unit, the optimization problem and scheduling algorithm should be reformulated and redesigned, respectively.

A. MOTIVATION

Consider the mmWave network of six UEs as demonstrated in Fig. 1; and its corresponding rate and request matrixes are expressed as

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

and

$$\mathbf{C} = \begin{bmatrix} 0 & 2 & 6 & 2 & 4 & 2 \\ 2 & 0 & 6 & 0 & 2 & 4 \\ 6 & 6 & 0 & 2 & 6 & 2 \\ 2 & 0 & 2 & 0 & 6 & 2 \\ 4 & 2 & 6 & 6 & 0 & 2 \\ 2 & 4 & 2 & 2 & 2 & 0 \end{bmatrix}, \tag{9}$$

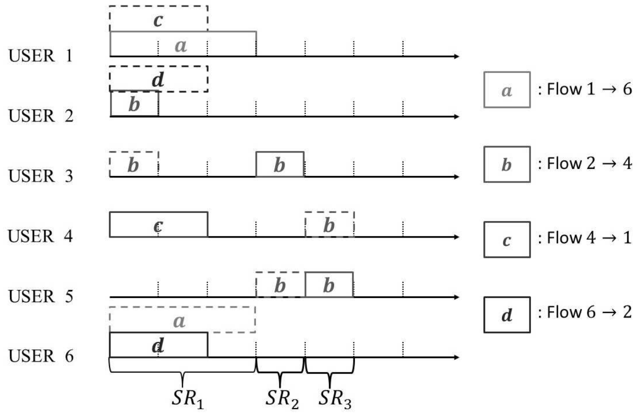


FIGURE 2. An illustrative scheduling example for the conventional MHRT scheme based on the request and rate matrixes of (8) and (9), respectively. Note that the solid-lined rectangles denote the time slots for transmitting, while the dot-lined ones are for receiving.

respectively. Observing (8), one can find that except the transmission between the 2-th and 4-th UEs (denoted by flow f_{24}), flows f_{16}, f_{41} and f_{62} can all be born via the direction transmissions. Thus, a routing path (i.e., $2 \rightarrow 3 \rightarrow 5 \rightarrow 4$) should be developed for flow f_{24}. To be clear, let $\mathcal{L}(f_{24}) = \{\ell_{23}, \ell_{35}, \ell_{54}\}$ represent the links traversed by flow f_{24}. Then, by definition, we can also have $\mathcal{L}(f_{16}) = \{\ell_{16}\}, \mathcal{L}(f_{41}) = \{\ell_{41}\}$ and $\mathcal{L}(f_{62}) = \{\ell_{62}\}$. Similarly, the time required for finishing the transmission task for each link of flow f_{24} can be defined as $\mathcal{T}(f_{24}) = \{r_{24}/c_{23}, r_{24}/c_{35}, r_{24}/c_{54}\} = \{6/6, 6/6, 6/6\} = \{1, 1, 1\}\ T_s$, where T_s is the duration of time slot. By analogy, it leads to $\mathcal{T}(f_{16}) = \{6/2\} = \{3\}\ T_s$, $\mathcal{T}(f_{41}) = \{4/2\} = \{2\}\ T_s$ and $\mathcal{T}(f_{62}) = \{8/4\} = \{2\}\ T_s$. Then, the mission of the scheduling algorithm is to serve these flows using minimum time slots.

Recall that using the conventional MHRT scheme, multiple packets can be transmitted during a single and long time slot; and the requests of a packet burst should be scheduled burst-by-burst. Therefore, flow f_{16} which takes the longest time should be firstly scheduled as illustrated in Fig. 2. Then, flows $f_{41}\, f_{62}$ and the first hop of flow f_{24} (i.e., ℓ_{23}) can share the first three consecutive time slots (i.e., t_1, t_2 and t_3) with flow f_{16}. Note that for fair comparison with the proposed ETA scheduling algorithm (as explained in the next paragraph), the full-duplex relaying scheme rather than the half-duplex scheme (as applied in [1] and [2]) is considered here. This explains the capability of sharing t_1 among flows f_{16} and f_{62}. Afterwards, the second and three hops of flow f_{24} (i.e., ℓ_{35} and ℓ_{54}, respectively) are assigned t_4 and t_5, respectively. To sum up, the first scheduling round (SR) (named $\mathcal{SR}_1$) includes links $\ell_{16}, \ell_{23}, \ell_{41}$ and ℓ_{62} (denoted by $\mathcal{SR}_1 = \{\ell_{16}, \ell_{23}, \ell_{41}, \ell_{62}\}$ for simplicity). Moreover, it leads to $\mathcal{SR}_2 = \{\ell_{35}\}$ and $\mathcal{SR}_3 = \{\ell_{54}\}$. At last, the MHRT scheme takes five T_s in total to serve the transmission request of (8).

In contrast to the MHRT scheme, as aforementioned, the proposed ETA algorithm aims to use smaller granularity of scheduling unit to enlarge the scheduling space. To this end, we redesign the duration of time slot (i.e., T'_s) according to the following rules:

$$\begin{array}{cccccc} \overbrace{}^{\text{rate}} & & \overbrace{}^{\text{required time slot}} & & \overbrace{}^{\text{packet}} & \\ 2R & \times & n_1 T'_s & = & 1 & \\ 4R & \times & n_2 T'_s & = & 1 & , \\ 6R & \times & n_3 T'_s & = & 1 & \end{array} \tag{10}$$

where $R = 1/T_s$. Making $n_1 = 6$, $n_2 = 3$ and $n_3 = 2$ can result in $T'_s = 1/(12R)$. Accordingly, 6, 3 and 2 time slots are required to transmit a packet using rates of $2R$, $4R$ and $6R$, respectively. Whereas, using the conventional MHRT scheme, a burst (consisted of several packets) should be transmitted using a single long time slot. For example,

$$\begin{array}{cccccc} \overbrace{}^{\text{rate}} & & \overbrace{}^{\text{required time slot}} & & \overbrace{}^{\text{packet}} & \\ 2R & \times & T_s & = & 2 & \\ 4R & \times & T_s & = & 4 & . \\ 6R & \times & T_s & = & 6 & \end{array} \tag{11}$$

It is known that time slot is a unit of scheduling. Thus, using shorter time slot as in (10), a packet can be scheduled to transmit using multiple units (i.e., packet-by-packet scheduling) so as to enlarge the scheduling space; whereas using the time slot in (11) results in burst-by-burst scheduling. Most importantly, the packet size will not be altered even though the duration of time slot has been adjusted; and this adjustment is only for packet transmission session through the data channel rather than the control channel. Thus, after some necessary control messages exchanging procedures, PNC can broadcast the scheduling results; and then all the UEs deliver their packets during the designated time slots based on the adjusted duration of time slot.

According to T'_s, the rate matrix $\mathbf{C}$ becomes

$$\mathbf{C} = \begin{bmatrix} 0 & 1/6 & 1/2 & 1/6 & 1/3 & 1/6 \\ 1/6 & 0 & 1/2 & 0 & 1/6 & 1/3 \\ 1/2 & 1/2 & 0 & 1/6 & 1/2 & 1/6 \\ 1/6 & 0 & 1/6 & 0 & 1/2 & 1/6 \\ 1/3 & 1/6 & 1/2 & 1/2 & 0 & 1/6 \\ 1/6 & 1/3 & 1/6 & 1/6 & 1/6 & 0 \end{bmatrix}. \tag{12}$$

That means using the lowest transmission rate (i.e., 1/6), one packet takes $6T'_s$. Therefore, the transmission task of flow f_{16} needs $6 \times 6\ T'_s$. Then, it leads to $\mathcal{T}(f_{41}) = \{4 \times 6\} = \{24\}\ T'_s$, $\mathcal{T}(f_{62}) = \{8 \times 3\} = \{24\}\ T'_s$ and $\mathcal{T}(f_{24}) = \{6 \times 2, 6 \times 2, 6 \times 2\} = \{12, 12, 12\}\ T'_s$. Now, we aim to schedule the transmission tasks "packet-by-packet" rather than "burst-by-burst" fashion applied by the conventional MHRT scheme. Consequently, the considered four flows can be scheduled as illustrated in Fig. 3. It should be noticed that it takes $42T'_s = 42/12\ T_s$ to finish request $\mathbf{R}$ in (8). Most importantly, using the smaller time slot T'_s can improve transmission efficiency by $(5 - 42/12)/5 = 30\%$.

B. PROBLEM FORMULATION

To facilitate the presentation, some terminologies are defined as follows:

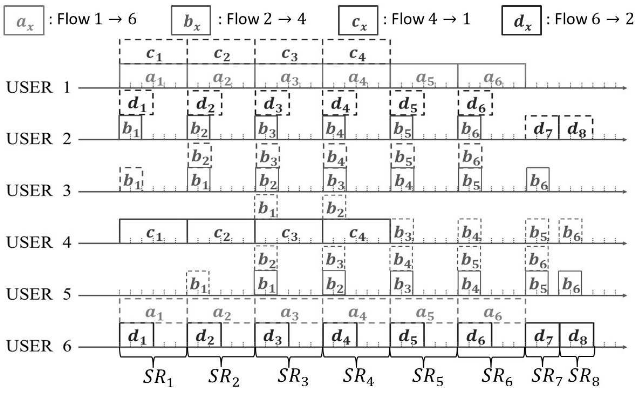

a_x : Flow 1 → 6
b_x : Flow 2 → 4
c_x : Flow 4 → 1
d_x : Flow 6 → 2
USER 1
USER 2
USER 3
USER 4
USER 5
USER 6
c_1 c_2 c_3 c_4
a_1 a_2 a_3 a_4 a_5 a_6
d_1 d_2 d_3 d_4 d_5 d_6 d_7 d_8
b_1 b_2 b_3 b_4 b_5 b_6
SR_1 SR_2 SR_3 SR_4 SR_5 SR_6 SR_7 SR_8

FIGURE 3. **An illustrative scheduling example for the proposed ETA algorithm based on the request and adjusted rate matrixes of (8) and (12), respectively. Note that the solid-lined rectangles denote the time slots for transmitting, while the dot-lined ones are for receiving.**

1. $\overline{\mathcal{SR}}$: the total number of SRs to serve all the transmission requests.
2. $\mathcal{T}(s)$: the duration of the s-th SR.
3. $\mathcal{P}(f_{uv})$: the routing path for flow f_{uv}.
4. $\mathcal{L}_{uvn}$: the n-th link of routing path $\mathcal{P}(f_{uv})$ (i.e. the n-th packet of flow f_{uv} in other words).
5. $\mathcal{L}_{pqm} \Rightarrow \mathcal{L}_{uvn}$: link $\mathcal{L}_{uvn}$ suffers the interference produced by link $\mathcal{L}_{pqm}$.
6. $|\mathcal{P}(f_{uv})|$: the measure of the number of links included in routing path $\mathcal{P}(f_{uv})$.
7. $\mathcal{C}_{uvn}$: the transmission rate of link $\mathcal{L}_{uvn}$.
8. $\mathcal{A}_{uvn}^{pqm}(s)$: if two adjacent links $\mathcal{L}_{uvn}$ and $\mathcal{L}_{pqm}$ are scheduled during the s-th SR, $\mathcal{A}_{uvn}^{pqm}(s) = 1$; otherwise $\mathcal{A}_{uvn}^{pqm}(s) = 0$. Note that any two arbitrary links which use the same devices as source or destination are named adjacent links. In the conventional scheme, however, two links are adjacent when one uses a particular device as source and another uses it as a destination; and vice versa [1], [2]. That results from the restriction of using the half-duplex relay.
9. δ_{uvn}^s : if the n-th link of routing path $\mathcal{P}(f_{uv})$ is served during the s-th SR $\delta_{uvn}^s = 1$; otherwise $\delta_{uvn}^s = 0$.

Now, the optimization problem of efficient scheduling can be formulated as follows:

$$\min \sum_{s=1}^{\overline{\mathcal{SR}}} \mathcal{T}(s) \tag{$P1$}$$

$$\text{s.t.} \sum_{s=1}^{\overline{\mathcal{SR}}} \delta_{uvn}^s = r_{uv}, \quad \forall u, v, n, \tag{13}$$

where the constraint (13) means that the transmission request of r_{uv} should be served within $\overline{\mathcal{SR}}$ SRs. Specifically, link $\mathcal{L}_{uvn}$ should be scheduled by r_{uv} times to finish the request of r_{uv}. Moreover, the traffic which can be carried during $\overline{\mathcal{SR}}$ SRs (i.e., $\mathcal{T}(s) \cdot \delta_{uvn}^s \cdot \mathcal{C}_{uvn}$) should satisfy

$$\sum_{s=1}^{\overline{\mathcal{SR}}} \mathcal{T}(s) \cdot \delta_{uvn}^s \cdot \mathcal{C}_{uvn} \geq r_{uv}, \quad \forall u, v, n, \tag{14}$$

where $\mathcal{C}_{uvn}$ is decided by (7) and it will be evaluated by Lines 15-16 of Algorithm 1 (discussed in the latter). It should be noticed that the required $\overline{SR}$ for $P1$ is unknown. Thus, it should satisfy the overall transmission requests of (13) and the transmission capacity of (14) as done in [1], [2], [5], and [13]. Since the proposed ETA scheduling algorithm can serve a request packet-by-packet, different links of the same flow can be served during the same SR. Therefore, it results in the following constraint.

$$\sum_{n=1}^{|\mathcal{P}(f_{uv})|} \delta_{uvn}^s \leq |\mathcal{P}(f_{uv})|, \quad \forall u, v, s. \tag{15}$$

In principle, the packets of any particular flow should be transmitted in sequence. That means link $\mathcal{L}_{uvn}$ should be served priori to link $\mathcal{L}_{uv(n+1)}$, which leads to

$$\sum_{s=1}^{\bar{S}} \delta_{uvn}^s \geq \sum_{s=1}^{\bar{S}} \delta_{uv(n+1)}^s, \quad \forall u, v, n, \bar{S} \in [1, \cdots, \overline{\mathcal{SR}}]. \tag{16}$$

Furthermore, the following constrain response to (3) and (4), respectively. To be clear, the collision between any two arbitrary adjacent links during the same SR should be avoided as

$$\delta_{uvn}^s + \delta_{pqm}^s \leq 1, \quad \text{if } \mathcal{A}_{uvn}^{pqm}(s) = 1, \ \forall \mathcal{L}_{uvn}, \mathcal{L}_{pqm}, s. \tag{17}$$

Recall the eighth definition in Section III(B). One should notice that the restriction of using the adjacent links simultaneously is different from that in [1] and [2]. Therein, this restriction restrains two links to share a device at the same time; whereas, here, a device can be scheduled to transmit and receive simultaneously. At last, the SINR of each link (e.g., $\mathcal{L}_{uvn}$) should reach the minimum requirement of $\gamma_{\min}(\mathcal{C}_{uvn})$ such that the transmission rate (i.e., $\mathcal{C}_{uvn}$) can be maintained, i.e.,

$$\frac{k_0 P_t d_{ij}^{-\alpha} G(\theta_{ij}) \delta_{uvn}^s}{WN_0 + \sum_{\mathcal{L}_{pqm} \Rightarrow \mathcal{L}_{uvn}} \rho k_0 P_t d_{hk}^{-\alpha} G(\theta_{hk}) \delta_{pqm}^s + \beta P_t h_L \delta_{uv(n+1)}^s} \geq \gamma_{\min}(\mathcal{C}_{uvn}) \delta_{uvn}^s, \quad \forall u, v, n, s, \tag{18}$$

where the i-th and j-th devices build the n-th link of flow f_{uv} such that $\mathcal{L}_{uvn} = \ell_{ij}$; and similarly we get $\mathcal{L}_{pqm} = \ell_{hk}$.

It should be reemphasized that using the smaller scheduling unit and full-duplex relay lead to the optimization problem different from the conventional schemes in [1] and [2]. Specifically, using the full-duplex relay results in the redefinition of adjacent link and its corresponding constraint of (17), which is different from the constraints of [1, eq. (10)] and [2, eq. (9)]. Moreover, using the smaller scheduling unit, the constraints of (13)–(15) are different from those of

[1, eqs. (6)–(8)] and [2, eqs. (6)–(7)]), respectively. In consequence, the resulted scheduling space is enlarged, which complicates the design of scheduling algorithm.

C. LINEARIZATION OF $P1$

Owing to the integral and nonlinearity of the constraints, $P1$ is indeed a typical problem of MINLP. Thus, the reformulation-linearization technique (RLT) is applied to linearize it so as to define another optimization problem $P2$; and then we propose the ETA scheduling algorithm to solve $P2$ in the following subsection as done in [1]–[6]. Note that the linearization procedure is to show that the MINLP problem can be solved by properly designing a linear programming algorithm. For the sake of readability, the linearization procedure is postponed to Appendix.

D. ETA SCHEDULING ALGORITHM

De facto, to well solve the optimization problem of $P1$ involves two phases, i.e., the path developing and scheduling phases. For fair performance comparison, Algorithm 1 of [1] and [2] is firstly applied to develop routing path for each transmission request. Secondly, to enlarge the scheduling space, time slot is redesigned according to the rules introduced in Section III, i.e., (10). Base on the developed paths and redesigned duration of time slot, the scheduling algorithm aims to serve all the transmission requests using minimum number of time slots. To specifically focus on and evaluate the performance of scheduling algorithm, the protocol design for coordinating the devices and collecting the required information (e.g., the request matrix of (8), rate matrix of (9) and locations of devices) is skipped, which is indeed beyond the scope of this paper.

Now, we design the ETA scheduling algorithm to solve the problem of $P2$ under the constraints of (13), (15), (16), (17), (21), (24) and (25). To begin with, some terminologies are defined to improve the readability of pseudo codes.

1 $\mathbb{P}$: the set of all the paths developed in the first step, which can be denoted by $\mathbb{P} = \{\mathcal{P}(f_{uv})\}\ \forall u, v$.

2 $\mathbb{L}$: the set of all the links which build $\mathcal{P}(f_{uv}) \in \mathbb{P}\ \forall u, v$. Thus, it can be denoted by $\mathbb{L} = \{\mathcal{L}_{uvn}\}\ \forall u, v, n$.

3 $\mathbb{L}_h$: the unscheduled headmost links of $\mathcal{P}(f_{uv}) \in \mathbb{P}\ \forall u, v$. Note that $\mathbb{L}_h \subseteq \mathbb{L}$.

4 $\mathbb{L}_u$: the set of the unfinished scheduled links. Note that $\mathbb{L}_u \subseteq \mathbb{L}$, and link $\mathcal{L}_{uvn}$ should be scheduled by r_{uv} times to finish its transmission task.

5 $\mathbb{L}_s$: the set of links which are scheduled to transmit during the s-th SR.

6 $\mathcal{T}_{uvn}(s)$: the number of time slots for link $\mathcal{L}_{uvn}$ to transmit one packet during the s-th SR.

7 Λ_{uvn}: the number of links which are adjacent to link $\mathcal{L}_{uvn}$.

8 $\hat{r}_{uvn}$: the remaining times for link $\mathcal{L}_{uvn}$ to be scheduled. Note that initially $\hat{r}_{uvn} = r_{uv}\ \forall u, v, n$.

9 $\hat{\gamma}_{\min}$: the minimum SINR to support the minimum transmission rate.

Algorithm 1 ETA Scheduling Algorithm

```
 1: Initialize $n = 1$ for each routing path; $s = 0$; $\hat{r}_{uvn} = r_{uv}\ \forall u, v, n$;
 2: while $\| \mathbb{L} \| > 0$ do
 3:     Set $s = s + 1$; $\mathbb{L}_s = \emptyset$; $\mathcal{T}(s) = 0$;
 4:     Obtain $\mathbb{L}_h, \mathbb{L}_u$;
 5:     $\mathbb{L}'_u = \mathbb{L}_h \cup \mathbb{L}_u$;
 6:     while $\left(\| \mathbb{L}_s \| \leq N \text{ and } \| \mathbb{L}'_u \| > 0\right)$ do
 7:         Let $\Lambda_{\min} = \min_{u,v,n} \Lambda_{uvn}\ \forall \mathcal{L}_{uvn} \in \mathbb{L}'_u$;
 8:         Let $\mathbb{L}^1 = \{\mathcal{L}_{uvn}\}$ with $\Lambda_{uvn}{=}\Lambda_{\min} \forall \mathcal{L}_{uvn} {\in} \mathbb{L}'_u$;
 9:         Let $\mathcal{T}_{\max}(s) = \max_{uvn} \mathcal{T}_{uvn}(s)\ \forall \mathcal{L}_{uvn} \in \mathbb{L}^1$;
10:         Let $\mathbb{L}^2 = \{\mathcal{L}_{uvn}\}$ with $\mathcal{T}_{uvn}(s){=}\mathcal{T}_{\max}(s) \forall \mathcal{L}_{uvn} \in \mathbb{L}^1$;
11:         Random select one link $\mathcal{L}_{uvn}$ from the set $\mathbb{L}^2$;
12:         $\mathbb{L}_s = \mathbb{L}_s \cup \mathcal{L}_{uvn}$;
13:         $\mathbb{L}_u = \mathbb{L}_u \cup \mathcal{L}_{uvn}$;
14:         if $(\mathcal{A}_{uvn}^{pqm}(s) = 0\ \forall \mathcal{L}_{pqm} \neq \mathcal{L}_{uvn} \in \mathbb{L}_s)$ then
15:             Obtain the SINR $\gamma_{ij}\ \forall \ell_{ij} \in \mathbb{L}_s$ using (7);
16:             if $\left(\gamma_{ij} < \hat{\gamma}_{\min}\ \forall \ell_{ij} \in \mathbb{L}_s\right)$ then
17:                 $\mathbb{L}_s = \mathbb{L}_s - \mathcal{L}_{uvn}$;
18:                 Go to Line 32;
19:             else
20:                 Update the rate matrix $\mathbf{C}$;
21:                 Update $\mathcal{T}_{pqm}(s)\ \forall \mathcal{L}_{pqm} \in \mathbb{L}_s$;
22:             end if
23:             $\mathcal{T}(s) = \max\left(\mathcal{T}(s), \mathcal{T}_{pqm}(s)\right)\ \forall \mathcal{L}_{pqm} \in \mathbb{L}_s$;
24:             $\hat{r}_{uvn} = \hat{r}_{uvn} - 1$;
25:             if $(\hat{r}_{uvn} = 0)$ then
26:                 $\mathbb{L} = \mathbb{L} - \mathcal{L}_{uvn}$;
27:                 $\mathbb{L}_u = \mathbb{L}_u - \mathcal{L}_{uvn}$;
28:             end if
29:         else if $(\mathcal{A}_{uvn}^{pqm}(s) = 1\ \forall \mathcal{L}_{pqm} \neq \mathcal{L}_{uvn} \in \mathbb{L}_s)$ then
30:             $\mathbb{L}_s = \mathbb{L}_s - \mathcal{L}_{uvn}$;
31:         end if
32:         $\mathbb{L}'_u = \mathbb{L}'_u - \mathcal{L}_{uvn}$;
33:     end while
34:     Output $\mathbb{L}_s$ and $\mathcal{T}(s)$;
35: end while
```

Algorithm 1 summarizes the proposed ETA scheduling algorithm. In addition to the rate and request matrixes, Line 1 initiates some necessary parameters. Specifically, Lines 4 and 16 correspond to the constraints of (16) and (18), respectively. Moreover, both of Lines 14 and 29 reflect the constraints of (15) and (17). At last, Line 25 is for constraints of (13) and (14). The principal difference between the proposed ETA scheduling algorithm and the conventional scheme from [1] and [2] lies in Lines 14-30 of the algorithm. Explicitly, Lines 14, 29 and 16-21 reflect the full-duplex relay scheme and adaptive transmission rates (as explained in the last paragraph of Section II), respectively. Also, Lines 23-27 originates from redesigning the time slots for packet-by-packet scheduling as indicated by (10) and (12).

The proposed ETA has the same order of complexity [i.e., $\mathcal{O}(N^5)$] as the conventional MHRT scheme in [1] and [2]. Specifically, in the proposed ETA scheduling algorithm, the first while loop on Line 2 has $\| \mathbb{L} \|$ iterations, which is maximally $F \times H_{max}$, where F and H_{max} denote the number of flows and the maximum number of relaying hops, respectively. Additionally, the second while loop on Line 6 has maximally N iterations; whereas both the SINR calculation procedure on Line 15 and the updating procedure on Lines 20 and 21 require $\mathbb{L}_s$ iterations. Because $\mathbb{L}_s$ is maximally N, the overall complexity becomes $\mathcal{O}(N^3 \times F) = \mathcal{O}(N^5)$. Note that in our considered scenario (similar to that in [1]), $F = N(N-1)$. To clarify, the order of complexity for the scheduling algorithm in [1] and [2] can be expressed as $\mathcal{O}(N \times F^2)$. Note that the same algorithm was used in [1] and [2]; however different scenarios were considered therein, the values of F are N and $N(N-1)$ in [1] and [2], respectively. Thus, the corresponding orders of complexity are $\mathcal{O}(N^3)$ and $\mathcal{O}(N^5)$.

We herein demonstrate the procedure in the pseudo codes using the example in Fig. 3. For simplicity, the effect of MAI is ignored, which means that SINR checking and its related processes (Lines 15-22) are not performed. In SR_1, the unscheduled headmost links of the unfinished flows include links ℓ_{16}, ℓ_{23} (the first link of f_{24}), ℓ_{41}, and ℓ_{62}. After checking the minimum adjacent links (Lines 7 and 8) and the maximum weight (Lines 9 and 10), link ℓ_{16} is selected (Line 11) and considered for transmission during SR_1 (Lines 12 and 13). Then, its candidacy is verified (Line 14) by discovering that links ℓ_{23}, ℓ_{41}, and ℓ_{62} are not adjacent to link ℓ_{16}.

The candidacy of links ℓ_{23}, ℓ_{41}, and ℓ_{62} is also iteratively verified (Line 14) in the while loop of Line 6. Therefore, links ℓ_{23}, ℓ_{41}, ℓ_{62} and ℓ_{16} can be put into SR_1. Because all scheduled links during SR_1 have unfinished requests (i.e., untransmitted packets), they should be reconsidered by SR_2. Following the same procedures of Lines 6-33, links ℓ_{16}, ℓ_{23} (the first link of f_{24}), ℓ_{35} (the second link of f_{24}), ℓ_{41}, and ℓ_{62} can then be scheduled for transmission in SR_2. Note that because of full-duplex relaying, links ℓ_{23} and ℓ_{35} can be transmitted concurrently. Finally, the routine in Lines 6-33 leads to the result described in Section III(A).

IV. SIMULATION RESULTS

As mentioned in Section III.(D), this paper focuses on the performance of scheduling algorithm. Thus, the coordinating and data collecting procedures are not taken into account. Also, after some discussions about the characteristics of the path developing algorithm, more performance comparisons will be made for investigating the effectiveness of the proposed scheduling algorithm in this section. To this end, we compare the average time $\left(\text{i.e. } \mathcal{T}_{sum} = \sum_{s=1}^{\mathcal{SR}} \mathcal{T}(s) \text{ in } P1\right)$ required to serve transmission request $\mathbf{R}$ between the proposed ETA algorithm and conventional MHRT scheme in [1] and [2], respectively, for the various network densities, maximum number of relaying hops H_{max}, traffic loads, antenna beamwidths and combinations of the transmission rates. For fair performance comparison, MHRT with full-duplex relay scheme (denoted by MHRT-FD) is also included. It should be noticed that owing to the blockage problem, the direct links may not exist for some requests. In these situations, routing paths should be constructed. However, in some stringent cases, it is not possible to successfully develop routing paths. Therefore, the required time only takes the ones with reachable routing paths into account. It should be noticed that the successful transmissions solely related to the successful developed routing paths. When a routing path can be developed, the corresponding transmission request can be successfully served. In addition, the number of successfully developed paths depends on the blockage rate and the density of network. For describing the severity of blockage problem, the blockage rate is defined as

$$BR_L = \frac{\text{number of the blocked links}}{(N^2 - N)/2}, \tag{19}$$

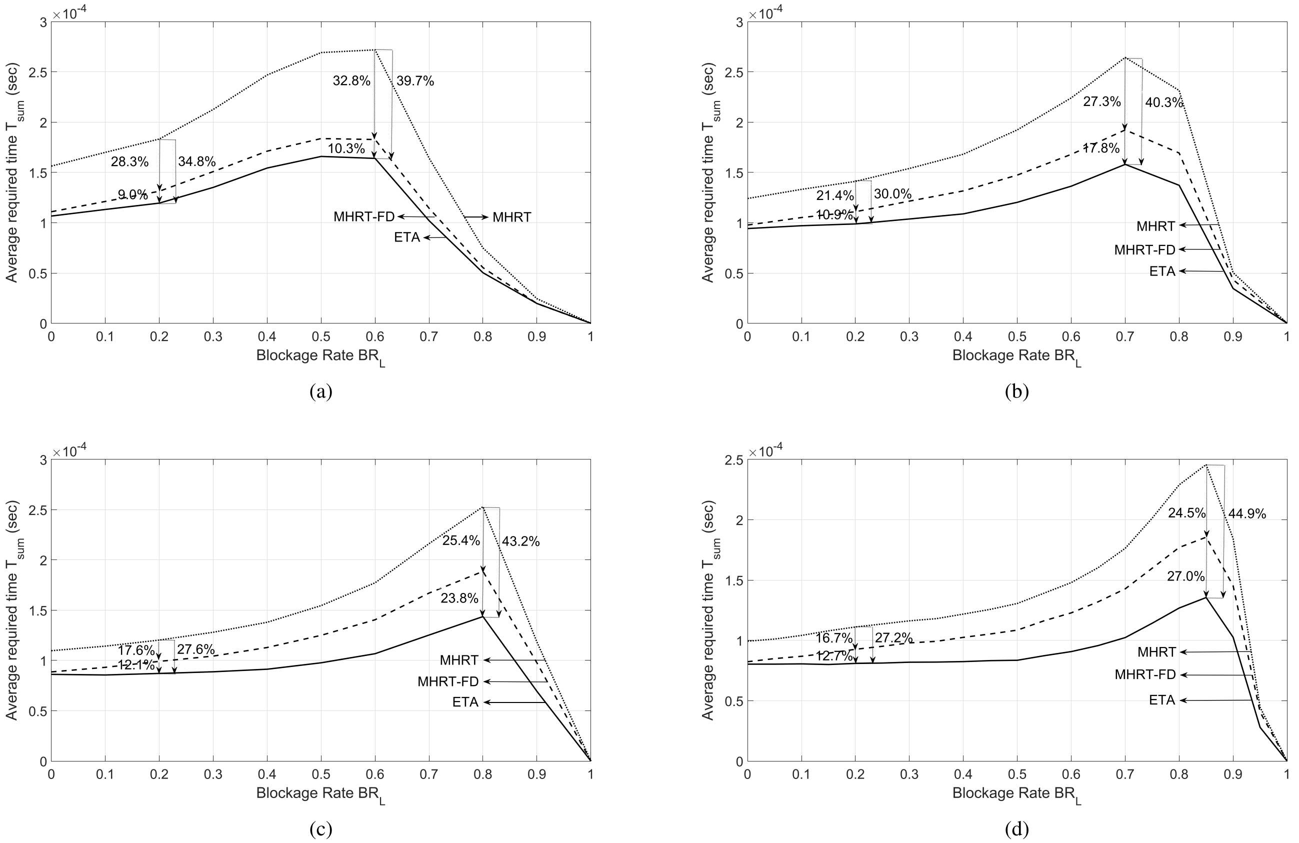
where $(N^2 - N)/2$ is the number of links between any two UEs. Note that no routing paths can be developed for the case with $BR_L = 1$.

A. SIMULATION SETUP

The mmWave network is constructed according to the system description in Section II.[1] Also, referring to the MHRT scheme in [1] and [2], N =5, 10, 15 and 20 full-duplex UEs are uniformly distributed around the square of 10 × 10 m^2. According to the distance, the combinations of the available transmission rates can be [2,4,6], [4,8,12] and [2,4,6,8,10] packets per time slot (packet/T_s), respectively. Moreover, using the proposed ETA scheme, the transmission rates can be degraded under the impact of MAI; or otherwise in some cases with stronger MAI, the link can even be dropped in the current SR and scheduled to transmit in the latter SR. It should be noted that the effect of rate degradation was ignored in the conventional MHRT scheme from [1] and [2].

To observe the MAI effect, three beamwidths are taken in account, including $\theta = 0°$, $30°$ and $60°$, respectively. Note that the one with $\theta = 0°$ corresponds to the case without considering MAI effect. In the routing path developing algorithm (i.e., Algorithm 1 of [1]), the maximum number of relaying hops H_{max} ranges from 2 to 6. Moreover, variant number of transmission requests F (i.e., the nonzero elements in request matrix $\mathbf{R}$) are considered, which increases from 5 to 45; and the volume of each transmission request is generated according to the Poisson process with arrival rate $\lambda = 6 \sim 54$ packets per flow. The time slot duration T_s is 6 μs. Table 4 summaries the simulation parameters.

[1] In [1] and [2], $N = 10$ was considered to generate $F = 10$ flows. However, in this paper, N varies from 5 to 20 to further reveal that "the proposed ETA scheduling algorithm becomes more effective for a high-density network with serious blockage problem".



(a)

(b)

(c)

(d)

FIGURE 4. Comparison of the average required time $\mathcal{T}_{sum}$ for various blockage rates BR_L with $F = 10$ flows and (a) $N = 5$, (b) $N = 10$, (c) $N = 15$ and $N = 20$ UEs, where the antenna beamwidth is $\theta = 0°$; the combination of available transmission rates is [2,4,6] packet/T_S; and the maximal number of relaying hops is $H_{max} = 4$.

B. PATH DEVELOPING PHASE

1) IMPACTS OF NETWORK DENSITY N ON ROUTING PATHS

Table 2 demonstrates the average (a) successfully developed paths and (b) corresponding required hops for $F = 10$ flows with respective to BR_L, where the numbers of UEs are $N = 5$, 10, 15 and 20, respectively. Also, the maximal number of relaying hops is $H_{\max} = 4$. Observing Table 2, it can be expected that the required hops to reach each destination increases as blockage rate BR_L increases. However, the number of required hops reduces when the blockage problem is unacceptably stringent. For example, with $N = 5$, the number of required hops reduces from 1.7 at $BR_L = 0.6$ to 1.2 at $BR_L = 0.8$. This is because when BR_L becomes larger than 0.8, only the paths which require fewer hops can be successfully developed. Also, as the blockage rate increases to $BR_L = 1$, no path can be developed.

TABLE 2. Average number of the (a) successfully developed paths and (b) corresponding required hops for $F = 10$ flows with respective to BR_L, where the numbers of UEs are $N = 5$, $N = 10$ and $N = 15$, respectively. Also, the maximal number of relaying hops is $H_{\max} = 4$.

(a)

	BR_L				
	0.2	0.4	0.6	0.8	1.0
N=5	10	9.9	8.2	2.7	0
N=10	10	10	9.9	6.6	0
N=15	10	10	10	9.0	0
N=20	10	10	10	9.7	0

(b)

	BR_L				
	0.2	0.4	0.6	0.8	1.0
N=5	1.2	1.5	1.7	1.2	—
N=10	1.2	1.5	2.0	2.3	—
N=15	1.2	1.5	2.0	2.5	—
N=20	1.2	1.5	2.0	2.6	—

2) IMPACTS OF MAXIMAL NUMBER OF RELAYING HOPS $H_{\max}$ ON ROUTING PATHS

Moreover, the capability of developing routing path also depends on the maximal number of relaying hops H_{max}. To reflect this fact, Table 3 shows average number of the (a) successfully developed paths and (b) corresponding required hops with respective to H_{max} for $F = 10$ flows, where the numbers of UEs is $N = 10$. It can be observed that with larger H_{max}, the number of successfully developed path can be increased. Specifically, more routing paths can be

TABLE 3. Average number of the (a) successfully developed paths and (b) corresponding required hops with respective to H_{max} for $F = 10$ flows, where the numbers of UEs is $N = 10$.

(a)

	BR_L				
	0.2	0.4	0.6	0.8	1.0
$H_{max} = 2$	10	10	8.7	4.2	0
$H_{max} = 3$	10	10	9.8	5.8	0
$H_{max} = 4$	10	10	9.9	6.6	0
$H_{max} = 5$	10	10	9.9	6.9	0
$H_{max} = 6$	10	10	9.9	7.0	0

(b)

	BR_L				
	0.2	0.4	0.6	0.8	1.0
$H_{max} = 2$	1.2	1.4	1.5	1.5	0
$H_{max} = 3$	1.2	1.5	1.9	2.0	0
$H_{max} = 4$	1.2	1.6	2.0	2.3	0
$H_{max} = 5$	1.2	1.6	2.1	2.4	0
$H_{max} = 6$	1.2	1.6	2.1	2.5	0

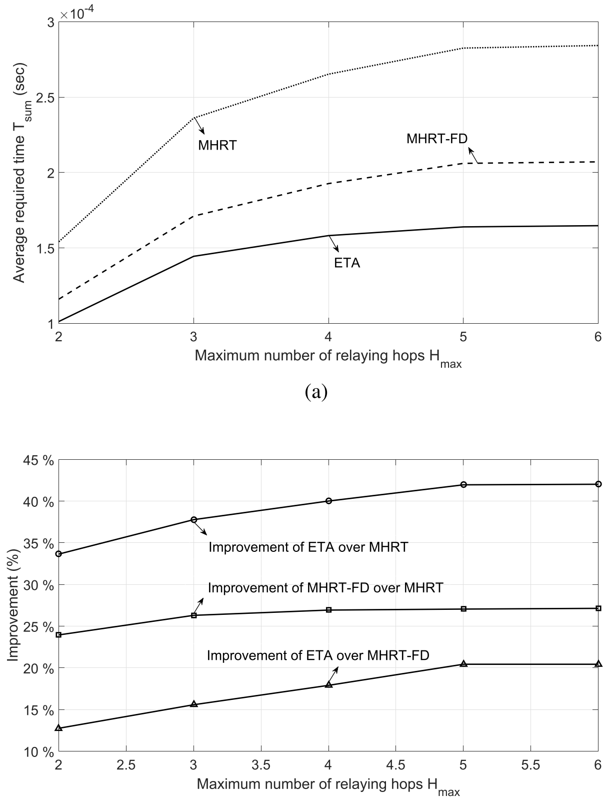
developed using more hops. For example, 2.4 hops on average are required for developing 6.9 routing paths when $BR_L = 0.8$ and $H_{max} = 5$. However, as H_{max} increases to 6, 7.0 routing paths can be developed by using 2.5 hops on average. In addition, in the stringent cases with higher BR_L, increasing H_{max} alone can not further solve the blockage problem. That means under the condition of higher BR_L, raising the network density N is necessary (as observed in Table 2).

C. SCHEDULING PHASE

1) IMPACT OF THE NETWORK DENSITY N

Fig. 4 compares the average required time $\mathcal{T}_{sum}$ for various blockage rates BR_L with $F = 10$ flows and (a) $N = 5$, (b) $N = 10$, (c) $N = 15$ and (d) $N = 20$ UEs, where the antenna beamwidth is $\theta = 0°$; the combination of available transmission rates is [2,4,6] packet/T_s; and the maximal number of relaying hops $H_{max} = 4$. The superiority of the proposed ETA algorithm can be observed in all considered cases of network density. Considering the cases of $BR_L = 0.6$ with $N = 5$, $BR_L = 0.7$ with $N = 10$, $BR_L = 0.8$ with $N = 15$, and $BR_L = 0.85$ with $N = 20$, respectively, the ability of the ETA scheme to arrange the full-duplex transmission flows alone can reduce $\mathcal{T}_{sum}$ by 32.8%, 27.3%, 25.4% and 24.5%. Moreover, the ETA scheduling itself can further reduce the corresponding $\mathcal{T}_{sum}$ by 10.3%, 17.8%, 23.8%, and 27.%. Integrating these two characteristics, the proposed ETA scheduling algorithm outperforms the conventional MHRT scheme by 39.7%, 40.3%, 43.2% 44.9% respectively.

It is noteworthy that the proposed ETA scheduling algorithm becomes more effective for a high-density network with serious blockage problem. Comparing $\mathcal{T}_{sum}$ for $BR_L = 0.2$, the improvements obtained by using the ETA scheme are remarkably larger in the aforementioned cases (i.e., the serious blockage problems with $BR_L = 0.6, 0.7, 0.8$ and 0.85, respectively). This is because under the condition of acceptably high BR_L (as explained in Table 2(b)), more hops are required to develop a routing path. More hops necessitate a larger scheduling space. With a larger scheduling space, the advantages of packet-by-packet scheduling and the ability of arranging the full-duplex transmission flows are magnified. In summary, these observations confirm the applicability of proposed ETA scheduling algorithm for the future generation of dense mmWave networks.



(a)

(b)

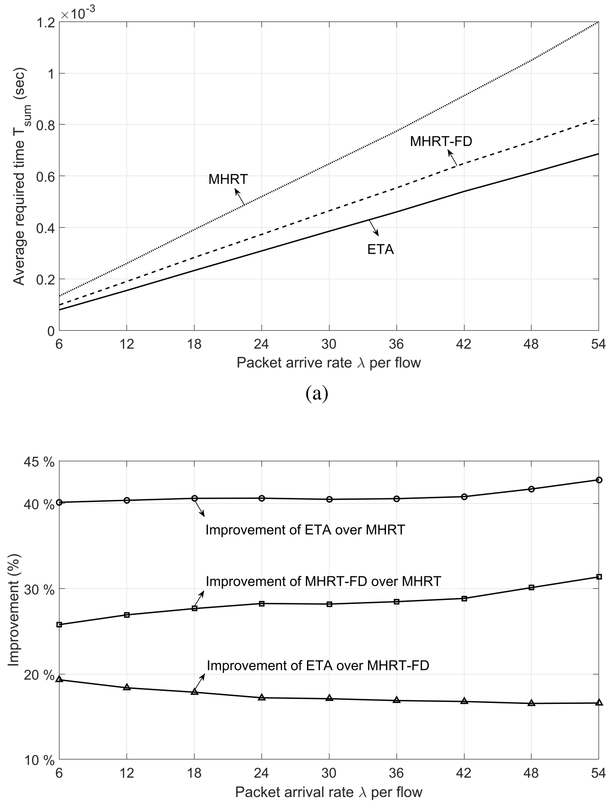
FIGURE 5. (a) comparison and (b) improvement of the average required time $\mathcal{T}_{sum}$ with respective to maximum number of relaying hops H_{max}, where the blockage rate $BR_L = 0.7$; the numbers of flows and UEs are $N = F = 10$.

2) IMPACT OF THE MAXIMAL NUMBER OF RELAYING HOPS H_{max}

Fig. 5 shows the (a) comparison and (b) improvement of the average required time $\mathcal{T}_{sum}$ with respective to maximum number of relaying hops H_{max}, where the blockage rate $BR_L = 0.7$; the numbers of flows and UEs are $N = F = 10$. Recall that using more hops can successfully developing more routing paths for the transmission requests (as explained in Table 3). Thus, it can be observed that longer $\mathcal{T}_{sum}$ is needed for serving more requests in the cases with larger H_{max}. Moreover, as H_{max} increases, the improvement of $\mathcal{T}_{sum}$ increases as shown in Fig. 5(b). This phenomenon can be explained by the larger scheduling spaces obtained by using

TABLE 4. Simulation parameters.

Parameter	Value
No. of UEs N	5,10,15,20
No. of flow F	$5 \sim 45$
Maximal no. of relaying hops $H_{\max}$	$2 \sim 6$
Square area	$10 \times 10\ m^2$
Packet size	1000 bytes
Time slot duration T_s	$6\ \mu s$
Blockage rate BR_L	0.1-1.0
Traffic model	Poisson process with $\lambda = 6 \sim 54$ packets/flow
Beamwidth of mainlobe θ_m	$0°$, $30°$, $60°$
Radiation efficiency ν	0.9
Transmission rate	[2,4,6], [2,4,6,8,10], [4,8,12] packet/T_s



(a)

(b)

FIGURE 6. (a) comparison and (b) improvement of the average required time $\mathcal{T}_{sum}$ with respective to packet arrival rate per flow λ, where the blockage rate $BR_L = 0.7$; the numbers of flows and UEs are $N = F = 10$. Also, the maximal number of relaying hops is $H_{\max} = 4$.

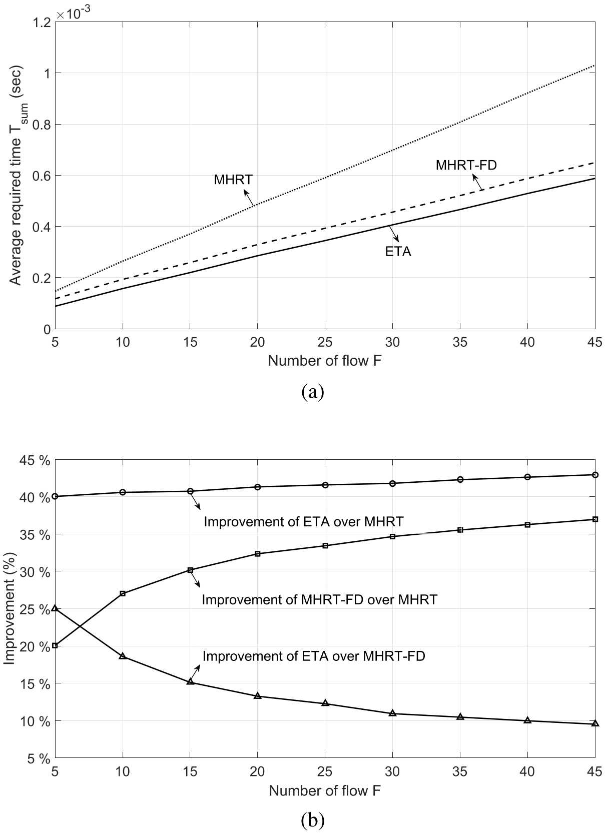
more relaying hops (as explained in Fig. 4). However, either $\mathcal{T}_{sum}$ itself or its improvement saturates with $H_{max} \geq 5$.

3) IMPACT OF THE TRAFFIC LOAD

Referring to [1] and [2], the traffic load can be defined as

$$Load = \frac{\lambda \times F}{C_0}, \tag{20}$$

where C_0 is the basic transmission rate. In our considered cases, $C_0 = 2$ packet/T_s. To observe the impacts of the traffic load on $\mathcal{T}_{sum}$, we vary the packet arrival rate per flow λ and number of flows F in Figs. 6 and 7, respectively. Firstly, as illustrated in Figs. 6(a) and 7(a), the increasing trends of $\mathcal{T}_{sum}$ with λ and F can be expected. However, the increasing rates (i.e., the slopes) are different for the MHRT, MHRT-FD and ETA schemes. To be specific, the slope of MHRT is apparently higher than that of MHRT-FD and ETA. This is because the ability of arranging full-duplex relaying can expedite the packet forwarding process, which restrains the increasing rates of $\mathcal{T}_{sum}$ for the MHRT-FD and ETA schemes. Furthermore, the ETA scheduling itself (i.e., packet-by-packet scheduling using shorter time slot) can further strengthen this effect for the cases with higher λ.



(a)

(b)

FIGURE 7. (a) comparison and (b) improvement of the average required time $\mathcal{T}_{sum}$ with respective to number of flow F, where the blockage rate $BR_L = 0.7$; the number of UEs is $N = 10$; the packet arrival rate per flow $\lambda = 12$; and the maximal number of relaying hops is $H_{\max} = 4$.

Observing Figs. 6(b) and 7(b), the ability of arranging full-duplex relaying can explain the increasing tread for the "improvement of ETA over MHRT" and "improvement of MHRT-FD over MHRT". However, the reason for the diluted "improvement of ETA over MHRT-FD" with λ and F lies in the reduced advantage of packet-by-packet scheduling for the cases with larger λ and F. Note that for Poisson distribution, the variance grows with λ, which leads to the larger differences in the number of packets within data bursts. In this situation, more requests can be served during an SR by using the conventional MHRT scheme. Therein, the duration

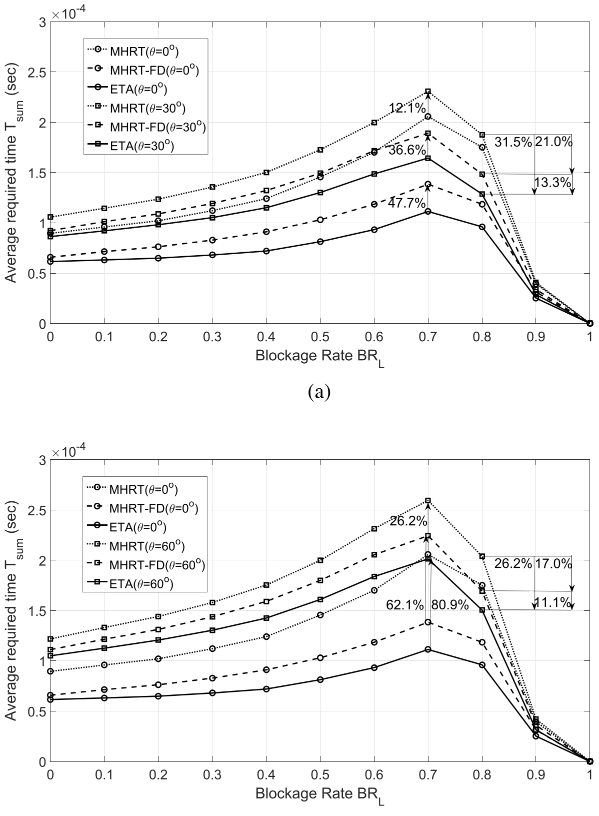


FIGURE 8. Comparison of the average required time $\mathcal{T}_{sum}$ (a) between the case with $\theta = 0^o$ and case with $\theta = 30^o$; (b) between the case with $\theta = 0^o$ and case with $\theta = 60^o$ for various BR_L and rate combination of [2,4,6] packet/T_s, where the numbers of flows and UEs are $N = F = 10$. Also, the maximal number of relaying hops is $H_{\max} = 4$.

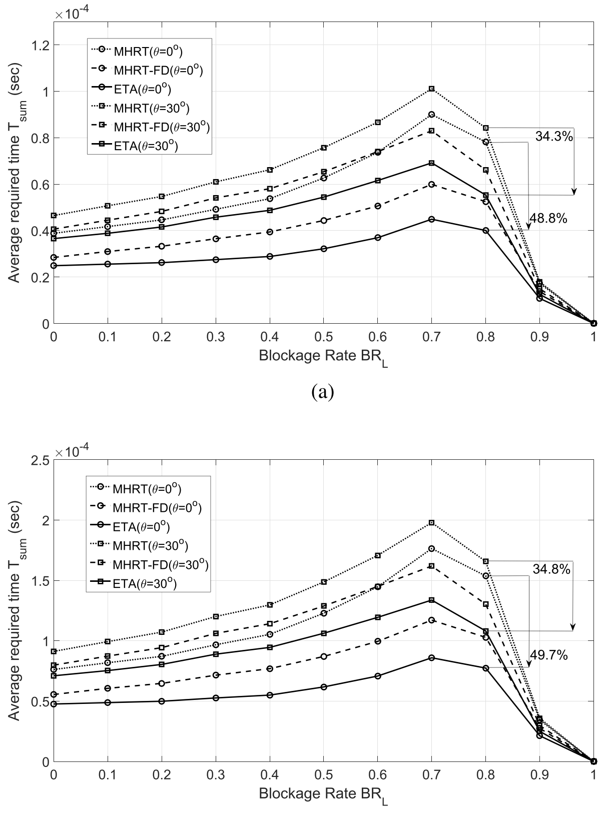


FIGURE 9. The average required time $\mathcal{T}_{sum}$ with respective to BR_L for (a) rate combination of [4,8,12] packet/T_S and (b) rate combination of [2,4,6,8,10] packet/T_S, respectively, where the antenna beamwidths are $\theta = 0°$ and $30°$; the numbers of flows and UEs are $N = F = 10$. Also, the maximal number of relaying hops is $H_{\max} = 4$.

of an SR is defined by the longest data burst scheduled for transmitting during that period. In general, a data burst shorter than the SR can be served simultaneously. Therefore, with a longer SR, more data burst can be served simultaneously such that the advantage of packet-by-packet scheduling is restrained.

Moreover, under the condition of the higher BR_L (e.g., the $BR_L = 0.7$ in Figs. 6 and 7), devices tend to be shared by more data flows. Recall that any two data flows share the same device can not be arranged for transmission during the same SR. Thus, when more devices should be shared, the advantage of packet-by-packet scheduling can be serious restrained. For example of the extreme case when all the devices are shared, only one packet can be arranged for transmission during an SR; and consequently, no gain can be obtained by using the packet-by-packet scheduling.

4) IMPACT OF THE ANTENNA BEAMWIDTH

Fig. 8 compares the average required time $\mathcal{T}_{sum}$ (a) between the cases with $\theta = 0^o$ and $\theta = 30^o$; (b) between the cases with $\theta = 0^o$ and $\theta = 60^o$ for various BR_L and rate combination of [2,4,6] packet/T_s, where the numbers of flows and UEs are $N = F = 10$. Also, the maximal number of relaying hops is $H_{\max} = 4$. It is important to find that the MAI effect becomes more significant for the cases of using the full-duplex relay and ETA scheduling. When the θ increases from 0° to 30° at $BR_L = 0.7$, the increment of $\mathcal{T}_{sum}$ for MHRT, MHRT-FD and ETA are 12.1%, 36.6% and 47.7%, respectively. Similar phenomenon can also be observed for the case with $\theta = 60^o$ in Fig. 8(b). This is because using the full-relay as well as the ETA scheduling, more links can be scheduled to transmit during an SR, which increases the chances for each link to suffer from MAI. Accordingly, the more stringent MAI level can lead to the degraded transmission rate, which detains the packet transmission. However, the remarkable performance enhancements can still be attained using the proposed scheme. Comparing with the conventional MHRT scheme at, $\mathcal{T}_{sum}$ can be reduced by 31.5% and 26.2% for the cases with $\theta = 30^o$ and 60^o, respectively.

5) IMPACT OF THE TRANSMISSION RATE

Figs. 9 shows the average required time $\mathcal{T}_{sum}$ with respective to BR_L for (a) rate combination of [4,8,12] packet/T_s and (b) rate combination of [2,4,6,8,10] packet/T_s, respectively, where the antenna beamwidths are $\theta = 0°$ and $30°$; and the numbers of flows and UEs are $N = F = 10$.

Also, the maximal number of relaying hops is $H_{\max} = 4$. Apparently, compared with the conventional MHRT scheme, the remarkable superiority of the proposed ETA scheduling algorithm can still be maintained even with higher and more chooses of feasible transmission rates. With rate combination of [4,8,12] packet/T_s, the superiority of the proposed ETA scheme over the conventional MHRT scheme can be 48.8% and 34.3% for the cases with $\theta = 0^o$ and 30^o at $BR_L = 0.8$, respectively. Also, with rate combination of [2,4,6,8,10] packet/T_s, the corresponding enhancements become 49.7% and 34.8%. Additionally, comparing Fig. 8(a) with Fig. 9(a), one can find that higher transmission rate can expedite the packet forwarding process. Specifically, the $\mathcal{T}_{sum}$ can be approximately reduced by 50% by doubling the transmission rates. The similar phenomenon can also be observed when the comparison is made between Fig. 8(a) and Fig. 9(b). In addition to the higher transmission rate, more available rates contribute to larger scheduling space. In consequence, the advantage of using the proposed ETA scheduling can be exaggerated.

V. CONCLUSION AND FUTURE WORKS

In this paper, we propose the ETA scheduling algorithm to improve the efficiency of concurrent transmissions for mmWave-based WPA networks by reducing the time required to serve transmission requests. The innovations of the proposed ETA scheduling algorithm are its arrangement of full-duplex transmission flows, rate adaptation, proper time-slot adjustment and efficient packet-by-packet scheduling strategy. Because of these characteristics, the scheduling space can be enlarged to enhance the efficiency of scheduling for concurrent transmissions. The simulation results demonstrate that the proposed ETA algorithm reduces the required service time by approximately 50% , compared with the conventional MHRT scheme in one of our considered cases. Some suggestions for future works include: (1) a power adjustment to improve the energy efficiency of the ETA scheduling algorithm and; (2) energy efficient routing path development and scheduling.

APPENDIX

In this appendix, we linearize the MINLP optimization problem $P1$. To begin with, variable κ^s_{uvn} can be newly defined to replace $\mathcal{T}(s)\cdot\delta^s_{uvn}$, i.e., $\kappa^s_{uvn} = \mathcal{T}(s)\cdot\delta^s_{uvn}$, for linearizing (14). Also, $\hat{\mathcal{T}} = \max\{\lceil\frac{r_{uv}}{\mathcal{C}_{uvn}}\rceil,\ \forall u, v, n\}$ can be defined to bound variable $\mathcal{T}(s)$ as $0 \leq \mathcal{T}(s) \leq \hat{\mathcal{T}}$. Moreover, integral δ^s_{uvn} is relaxed to be $0 \leq \delta^s_{uvn} \leq 1$. In consequence, the so-called RLT bound-factor product constraints for κ^s_{uvn} can be expressed as

$$\begin{cases} \kappa^s_{ijn} \geq 0 \\ \mathcal{T}(s) - \kappa^s_{uvn} \geq 0 \\ \hat{\mathcal{T}}\cdot\delta^s_{uvn} - \kappa^s_{uvn} \geq 0 \\ \hat{\mathcal{T}} - \mathcal{T}(s) - \hat{\mathcal{T}}\cdot\delta^s_{uvn} + \kappa^s_{uvn} \geq 0, \end{cases} \quad \forall u, v, n, s. \tag{21}$$

Moreover, in order to linearize the constraint of (18), we firstly rewrite it as

$$\begin{aligned} &\Big(k_0\, P_t d_{ij}^{-\alpha} G(\theta_{ij}) - \gamma_{\min}(\mathcal{C}_{uvn})N_0 W - \gamma_{\min}(\mathcal{C}_{uvn})\beta P_t h_L \\ &\quad \times \delta^s_{uv(n+1)}\Big)\,\delta^s_{uvn} \geq \gamma_{\min}(\mathcal{C}_{uvn}) \\ &\quad \times \sum_{\mathcal{L}_{pqm}\Rightarrow\mathcal{L}_{uvn}} \rho k_0 P_t d_{hk}^{-\alpha} G(\theta_{hk})\delta^s_{pqm}\delta^s_{uvn}, \quad \forall u, v, n, s. \end{aligned} \tag{22}$$

Then, variable $\pi^{pqm}_{uvn}(s)$ is defined so that $\pi^{pqm}_{uvn}(s) = \delta^s_{uvn}\delta^s_{pqm}$. Accordingly, the RLT bound-factor product constraints for $\pi^{pqm}_{uvn}(s)$ can be described as

$$\begin{cases} \pi^{pqm}_{uvn}(s) \geq 0 \\ \delta^s_{uvn} - \pi^{pqm}_{uvn}(s) \geq 0 \\ \delta^s_{pqm} - \pi^{pqm}_{uvn}(s) \geq 0 \\ 1 - \delta^s_{uvn} - \delta^s_{pqm} + \pi^{pqm}_{uvn}(s) \geq 0, \end{cases} \quad \forall u, v, n, s. \tag{23}$$

Note that the linear form of the constraint (18) can be obtained by substituting $\pi^{pqm}_{uvn}(s) = \delta^s_{uvn}\delta^s_{pqm}$ into it as

$$\begin{aligned} &\Big(k_0\, P_t d_{ij}^{-\alpha} G(\theta_{ij}) - \gamma_{\min}(\mathcal{C}_{uvn})N_0 W - \gamma_{\min}(\mathcal{C}_{uvn})\beta P_t h_L \\ &\quad \times \delta^s_{uv(n+1)}\Big)\,\delta^s_{uvn} \geq \gamma_{\min}(\mathcal{C}_{uvn}) \\ &\quad \times \sum_{\mathcal{L}_{pqm}\Rightarrow\mathcal{L}_{uvn}} \rho k_0 P_t d_{hk}^{-\alpha} G(\theta_{uk})\pi^{pqm}_{uvn}(s), \quad \forall u, v, n, s. \end{aligned} \tag{24}$$

At last, the optimization problem $(P1)$ can be reformulated as

$$\min \sum_{s=1}^{\overline{\mathcal{SR}}} \mathcal{T}(s), \tag{P2}$$

$$\text{s.t. the constraints of } \sum_{s=1}^{\overline{\mathcal{SR}}} (\kappa^s_{ijn}\cdot\mathcal{C}_{uvn}) \geq r_{uv}, \quad \forall u, v, n, \tag{25}$$

(15), (16), (17), (21), and (24).

REFERENCES

[1] 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.

[2] Y. Niu *et al.*, "Exploiting device-to-device communications to enhance spatial reuse for popular content downloading in directional mmWave small cells," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5538–5550, Jul. 2016.

[3] 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.

[4] G. Zheng, C. Hua, R. Zheng, and Q. Wang, "Toward robust relay placement in 60 GHz mmWave wireless personal area networks with directional antenna," *IEEE Trans. Mobile Comput.*, vol. 15, no. 3, pp. 762–773, Mar. 2016.

[5] Z. He, S. Mao, and T. S. Rappaport, "On link scheduling under blockage and interference in 60-GHz ad hoc networks," *IEEE Access*, vol. 3, pp. 1437–1449, Sep. 2015.

[6] Z. He, S. Mao, S. Kompella, and A. Swami, "On link scheduling in dual-hop 60-GHz mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11180–11192, Dec. 2017.

[7] W. Chang and J.-C. Teng, "Energy efficient relay matching with bottleneck effect elimination power adjusting for full-duplex relay assisted D2D networks using mmWave technology," *IEEE Access*, vol. 6, pp. 3300–3309, Jan. 2018.

[8] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[9] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[10] G. Yang, J. Du, and M. Xiao, "Maximum throughput path selection with random blockage for indoor 60 GHz relay networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3511–3524, Oct. 2015.

[11] S. Biswas, S. Vuppala, J. Xue, and T. Ratnarajah, "On the performance of relay aided millimeter wave networks," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 576–588, Apr. 2016.

[12] Y. Xu, H. Shokri-Ghadikolaei, and C. Fischione, "Distributed association and relaying with fairness in millimeter wave networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 7955–7970, Dec. 2016.

[13] Z. He, S. Mao, S. Kompella, and A. Swami, "Minimum time length scheduling under blockage and interference in multi-hop mmWave networks," in *Proc. Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–7.

[14] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "Blockage robust and efficient scheduling for directional mmWave WPANs," *IEEE Trans. Veh. Technol.*, vol. 64, no. 2, pp. 728–742, Feb. 2015.

[15] X. Qin *et al.*, "Impact of full duplex scheduling on end-to-end throughput in multi-hop wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 158–171, Jan. 2017.

[16] F. Yildirim and H. Liu, "A cross-layer neighbor-discovery algorithm for directional 60-GHz networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4598–4604, Oct. 2009.

[17] H. Deng and A. Sayeed, "mm-Wave MIMO channel modeling and user localization using sparse beamspace signatures," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2014, pp. 130–134.

[18] 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.

[19] Z. Wei, X. Zhu, S. Sun, Y. Huang, A. Al-Tahmeesschi, and Y. Jiang, "Energy-efficiency of millimeter-wave full-duplex relaying systems: Challenges and solutions," *IEEE Access*, vol. 4, pp. 4848–4860, Sep. 2016.

[20] Z. Wei, X. Zhu, S. Sun, and Y. Huang, "Energy-efficiency-oriented cross-layer resource allocation for multiuser full-duplex decode-and-forward indoor relay systems at 60 GHz," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3366–3379, Dec. 2016.

[21] G. Yang, M. Xiao, H. Al-Zubaidy, Y. Huang, and J. Gross, "Analysis of millimeter-wave multi-hop networks with full-duplex buffered relays," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 576–590, Feb. 2018.

WENSON CHANG (M'00) received the B.S. and M.S. degrees in electrical engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 1998 and 2000, respectively, and the Minor M.S. degree in applied mathematics and the Ph.D. degree in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2005 and 2006, respectively. In 2006, he was with the Institute of Computer and Communication Engineering, National Cheng Kung University, Taiwan, as an Assistant Professor, where he has been an Associate Professor since 2011. His current research interests include wireless communications and networks, cross-layer design and optimization, and game theory applications in the wireless communications. In 2006, he was awarded the IEEE student travel grant for ICC and the membership of the Phi Tau Phi Scholastic Honor Society. In 2014 and 2017, he was granted the Outstanding Teaching Award by the National Cheng Kung University and received the Best Paper and Honorable Mention Awards in the International Conference of COCORA and Local Conference of NST-ITCom, respectively.

CHIEN-WEN WU was born in New Taipei, Taiwan, in 1994. He received the B.S. degree from the Department of Communications Engineering, National Taipei University, in 2016, and the master's degree from the Wireless Communication and Network Laboratory, Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan, Taiwan, in 2018.

His research interests include millimeter wave technology, wireless networks, and device-to-device communications.

YI-XIN LIN was born in Taichung, Taiwan, in 1992. He received the B.S. degree from the Department of Communications Engineering, Yuan Ze University, in 2015, and the master's degree from the Wireless Communication and Network Laboratory, Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan, Taiwan, in 2017. His research interests include MAC protocols, wireless networks, and mmWave-based techniques.