Optimal and Approximation Algorithms for Joint Routing and Scheduling in Millimeter-Wave Cellular Networks

Dingwen Yuan, Hsuan-Yin Lin, *Senior Member, IEEE*, Jörg Widmer, *Fellow, IEEE*, and Matthias Hollick, *Member, IEEE*

Abstract—Millimeter-wave (mmWave) communication is a promising technology to cope with the exponential increase in 5G data traffic. Such networks typically require a very dense deployment of base stations. A subset of those, so-called macro base stations, feature high-bandwidth connection to the core network, while relay base stations are connected wirelessly. To reduce cost and increase flexibility, wireless backhauling is needed to connect both macro to relay as well as relay to relay base stations. The characteristics of mmWave communication mandates new paradigms for routing and scheduling. The paper investigates scheduling algorithms under different interference models. To showcase the scheduling methods, we study the maximum throughput fair scheduling problem. Yet the proposed algorithms can be easily extended to other problems. For a full-duplex network under the no interference model, we propose an efficient polynomial-time scheduling method, the *schedule-oriented optimization*. Further, we prove that the problem is NP-hard if we assume pairwise link interference model or half-duplex radios. Fractional weighted coloring based approximation algorithms are proposed for these NP-hard cases. Moreover, the approximation algorithm parallel data stream scheduling is proposed for the case of half-duplex network under the no interference model. It has better approximation ratio than the fractional weighted coloring based algorithms and even attains the optimal solution for the special case of uniform orthogonal backhaul networks.

Index Terms—Millimeter-wave, 5G, backhaul, max-min fairness, full-duplex, half-duplex, matching, coloring, conflict graph, pairwise link interference, single user spatial multiplexing.

I. INTRODUCTION

5G AND beyond cellular systems are embracing millimeter wave (mmWave) communication in the 10-300 GHz band where abundant bandwidth is available to achieve Gbps data rate. One of the main challenges for mmWave systems is the high propagation loss. Although it can be partially compensated by directional antennas [1], [2], the effective
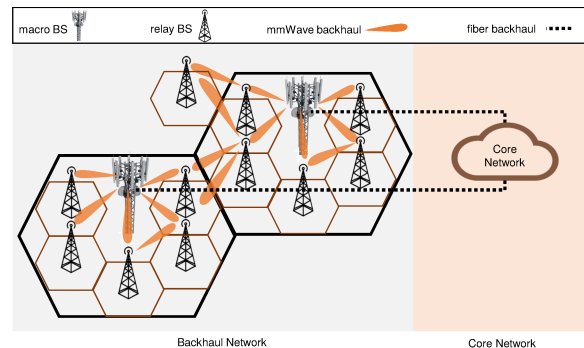
Fig. 1. mmWave self-backhauling setup.

communication range of a mmWave base station (BS) is around 100 meters for typical use cases. Thus, base station deployment density in 5G will be significantly higher than in 4G [3], [4]. This leads to high infrastructure cost for the operators. In fact, besides the cost of site lease, backhaul link provisioning is the main contributor to this expense because the mmWave access network may require multi-Gbps backhaul links to the core network. Currently, such a high data rate can only be accommodated by fiber-optic links which have high installation costs and are inflexible for reconfiguration.

Recent studies show that mmWave self-backhauling is a cost-effective alternative to wired backhauling [3], [5]. This approach is particularly interesting in an NG-RAN (Next Generation Radio Access Network) where one or more *5G base stations* (gNBs) have fiber backhaul to the core network and act as gateways for the other gNBs [6]. We refer to the gateway gNBs as *macro BSs* and the other gNBs as *relay BSs*. Fig. 1 illustrates such a setup in which each relay BS can be reached by at least one macro BS directly or via other relay BSs. Moreover, the directionality of mmWave communication reduces or removes the wireless backhaul interference and allows simultaneous scheduling of multiple links over the same channel as long as their antenna beams do not overlap. However, the number of simultaneous data streams a base station can handle is limited by the number of its radio frequency (RF) chains. Furthermore, a base station may only support half-duplex communication, i.e. it *can not* work as a transmitter and a receiver at the same time.

As of now, much of the research on mmWave communication has been dedicated to issues that the mobile users (UEs) face in the access networks. How to maximize performance such as throughput and energy efficiency in mmWave backhaul networks has received less attention. Transmission scheduling that incorporates the possibility of multi-hop routing is a most crucial research question to be addressed.

A naive scheduling which lets a macro BS serve all the relay BSs in its macrocell in a round robin fashion is neither practical nor efficient. If a relay BS' link to a macro BS is weaker than its links to other nearby relay BSs (which in turn have high-capacity links to a macro BS), a schedule allowing multi-hop routing can be more favorable since it alleviates the bottleneck at the macro BSs. At the same time, the limited interference at mmWave frequencies makes it efficient to maximize spatial reuse and operate as many data streams simultaneously as possible. The goal of this paper is to design a scheduler that exploits these characteristics to optimize mmWave backhaul efficiency for full-duplex and half-duplex radios assuming pairwise link interference or not, and realistic or maximum single-user spatial multiplexing.

The paper is organized as follows: related work is discussed in §II. The system model is described in §III. The *maximum throughput fair scheduling* (MTFS) serves as a concrete problem for presenting our methods for backhaul scheduling. In §IV, we present a polynomial time optimal algorithm for the full-duplex MTFS problem assuming no interference (NI) between links. In §V, we show that the MTFS problem is NP-hard for both cases of pairwise link interference (PI) model and half-duplex radios. Yet it is solvable in polynomial time for a special case—the so-called *uniform orthogonal backhaul network*. In §VI, we propose two general approximation algorithms based on fractional weighted coloring for the NP-hard cases. For half-duplex radios under the NI model, we propose the approximation algorithm—parallel data stream scheduling (PDS) in §VII, which has better a performance bound than the fractional weighted coloring based algorithms. It provides the optimal scheduling for the case of uniform orthogonal backhaul networks. §VIII elaborates on the extension of the scheduling algorithms to more general scenarios. §IX demonstrates the efficiency and effectiveness of the algorithms through numerical evaluations. Finally, §X concludes the paper.

II. Related Work

Our previous paper [7] discussed the optimal full-duplex scheduling of mmWave backhaul networks and its approximation algorithm assuming zero interference between any pair of links and the linearity of link capacity in terms of RF chains. This paper is a continuation of the work, which includes both full-duplex and half-duplex scheduling of mmWave backhaul networks for pairwise link interference and realistic single-user spatial-multiplexing model. A few works studying mmWave network scheduling [8]–[12] share the assumption that the traffic demand is measured in discrete units of timeslots or packets. The resulting optimization problems are all formulated as mixed integer programs (MIP). As MIPs are in general NP-hard, optimal solutions can only be computed for very small networks. For practical use, they all rely on heuristics, which are based for example on greedy edge coloring [8], [10] or finding the maximum independent set in a graph [9], [11], [12]. Furthermore, [8], [9], [11], [12] assume that routing is pre-determined, which does not fully exploit the freedom given by a reconfigurable backhaul, and may thus limit performance.

In contrast, we relax the constraint of in-order flow scheduling (i.e., if needed, packets are queued for a short time) which does not harm the long-term throughput, and allows a timeslot to be of any length (a following step can be used to discretize the length). Based on these assumptions, we propose polynomial time optimal and approximation scheduling algorithms. We show by simulation that they are practical for mmWave cellular networks. Moreover, the scheduling takes QoS optimization goals or QoS requirements as input and finds an efficient routing automatically. The first attempt to solve the problem of joint routing and scheduling in a network with Edmonds' matching formulation goes back to [13]. Hajek's scheduling algorithm differs from ours in that it minimizes schedule length. Furthermore, we use a one-step *schedule-oriented* approach while they first compute the optimal link time and then compute the minimum length schedule given the link time. The matching-based approach is also used in [14] to study the problem of throughput-optimal routing and scheduling in a wireless directed acyclic graph (DAG) with a time-varying connectivity, to which a throughput-optimal dynamic broadcast policy is proposed.

Two of our approximation algorithms (F^3WC) are adapted from [15] which proposed scheduling algorithms for multihop wireless networks based on fractional weighted vertex coloring of conflict graphs. We show the generality of F^3WC algorithms by demonstrating that all constraints of half-duplex, RF chain number and pairwise link interference can be modeled in a conflict graph.

While current mmWave radios only support half-duplex operation, full-duplex is feasible through proper analog and digital cancellation [16] and is likely to be used in the future. More details about the design of a full-duplex mmWave radio are discussed in [17]. In addition, we assume that multiple data streams can be transmitted concurrently between a single pair or multiple pairs of BSs. This assumption is backed by the feasibility of single-user and multi-user MIMO in mmWave communication [18]–[20], exploiting beamforming, multipath, spatial and polarization diversity.

Existing research works on mmWave backhaul scheduling favor centralized solutions for various optimization goals, including throughput [7], [10], [11], delay [21], energy consumption [12], makespan [8], [22], wireless bandwidth [23] and flows with satisfied QoS requirements [9]. To achieve these optimization goals, most of the works perform link scheduling to maximize spatial reuse while some control routing [7], [8], [10], [22], [23], transmission power [11], [12], [23] and bandwidth allocation [23]. 3GPP is currently performing a study on Integrated Access and Backhaul (IAB) [24] which describes mechanisms for sharing radio resources between access and backhaul links. Yet designing a high-performance IAB is still an open problem [25], [26]. Our algorithms can be equally applied to IAB networks because of our graph-based scheduling method, thus contributing to this emerging field.

III. System Model

The system model considers a backhaul network which consists of one or more macro BSs and multiple relay BSs. The macro BSs act as the backhaul gateways for the relay BSs. We assume that any two macro BSs are connected by fiber optics with infinite capacity (data rate). Thus, data can be exchanged between two macro BSs with zero delay.[1]

A node (BS) is equipped with multiple RF chains. Two nodes may have a different number of RF chains. We assume analog or hybrid beamforming which allows up to the same number of *simultaneous data streams* incident to a node as its number of RF chains. In addition, the maximum number of

---

[1]We can equally support non-zero communication delay between macro BSs. This can be done by adding wired links of fixed capacity between macro BSs. These links do not interfere with each other or with the wireless links. For presentation clarity, we assume that wired links introduce zero delay.
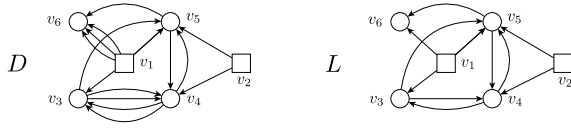
Fig. 2. Example of *directed network* and *link network*. □ and ○ represent a macro and a relay BS, respectively.

Fig. 3. Two interfering links $l^{(1)}$ and $l^{(2)}$. An angle between two dashed rays starting from a node is the beamwidth of that node.

simultaneous data streams supported by a link is determined by the spatial diversity of that link (see. §III-A). These data streams may have different capacities. Finally, the network is either *full-duplex* or *half-duplex*, meaning all nodes are equipped with full-duplex or half-duplex radios, respectively.

We use the convention that *arc* and *edge* refer to *directed* and *undirected edge* of a graph. The notations (u, v) and $\{u, v\}$ are used to represent an arc from u to v and an edge between u and v, respectively. To model the backhaul network, we start with a *directed network* D, which is an *arc weighted directed multigraph*.[2] In this paper, we use the notation $V(G)$ and $E(G)$ to refer to the vertex set and arc (edge) set of a directed (or undirected) graph G. Thus, $V(D)$ and $E(D)$ are the vertex set and arc set of D, representing the set of all BSs and the set of all possible data streams. Let $B(D)$ and $M(D)$ be the set of macro BS vertices and the set of relay BS vertices, respectively. Let L be the *link network* of D. It is a subgraph of D which is constructed by replacing each set of parallel arcs by a single arc. Let $l = (u, v) \in E(L)$ denote the link from node u to node v and $d(l)$ the maximum number of simultaneous data streams allowed by l. Then there are $d(l)$ parallel arcs from u to v in D, which are labeled as $l_1 \ldots l_{d(l)}$. Let $c : E(D) \mapsto \mathbb{Q}_+$ be the *capacity* function defined for each arc (data stream) where $\mathbb{Q}_+$ is the set of non-negative rational numbers. The value $c(e)$ is the capacity of a data stream e. Fig. 2 illustrates a toy example for a directed network and the corresponding link network.

In addition, we define an *RF chain number* function $r : V(D) \mapsto \mathbb{N}$ where $\mathbb{N}$ is the set of natural numbers $\{1, 2, \ldots\}$. $r(u)$ is the number of RF chains of node u. We denote the total number of RF chains of D by $r(D)$, i.e., $r(D) = \sum_{v \in V(D)} r(v)$. Scheduling is made based on the directed network D.

A. Single User Spatial Multiplexing Model

Parallel data streams from a transmitter u to a receiver $v \neq u$ may be scheduled simultaneously if both u and v use multiple RF chains. This is referred to as *single user spatial multiplexing* (SU-SM) [18]. In this paper, we assume that every RF chain of a node v is assigned the same transmission power $p_{tx}(v)$. $p_{tx}(u)$ and $p_{tx}(v)$ may be different if $u \neq v$. This is a reasonable assumption for a multi-transceiver node that has multiple RF chains, each with their own amplifiers and phased arrays (see [27] for an example MIMO system). We investigate two models for SU-SM:
  1) REAL-SU-SM: This is a general and realistic model. A link $l = (u, v) \in E(L)$ supports at most $d(l)$ data streams where $d(l) \leq \min(r(u), r(v))$ and is determined by the spatial diversity of link l such as the *channel matrix rank*. Since only the total capacity of the data streams is relevant in our scheduling problem, we can equivalently model their capacities such that $c(l_1) \geq \cdots \geq c(l_{d(l)}) > 0$. This means that we have

---

[2]A directed multigraph allows parallel arcs with the same head node and the same tail node, but disallows loops.
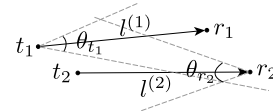
capacity $c(l_1)$ if only one data stream is active. With each additional data stream we have a more marginal increase in total capacity. Using $k \leq d(l)$ RF chains at both ends, we get a total capacity of $\sum_{i=1}^{k} c(l_i)$.
  2) MAX-SU-SM: This model achieves the maximum possible capacity of SU-SM, where each link provides sufficient spatial diversity. It is a special case of the REAL-SU-SM model with $d(l) = \min(r(u), r(v))$ and $c(l_1) = \cdots = c(l_{d(l)}) > 0$. In this model, capacity is also defined for a link such that $c(l) = c(l_1)$.

B. RF Chain Number Constraint and Half-Duplex Constraint

The constraint due to the limited number of RF chains is that given a set of simultaneously scheduled data streams $E \in E(D)$, for each node v, the number of arcs in E that are incident to v is at most $r(v)$. Furthermore, a schedule can be either *full-duplex* (FD) or *half-duplex* (HD). FD allows incoming and outgoing arcs of a vertex to be scheduled simultaneously, whereas HD does not. We will prove in the paper that the duplex state is a key parameter for scheduling. It determines whether the computational complexity is polynomial time or NP-hard.

C. Interference Model

Compared to lower frequency omnidirectional communication, mmWave frequencies significantly reduce the interference between concurrent links due to the short communication range and the directionality brought by beamforming. We will study and compare two interference models in this paper:
  1) NI: no interference between concurrent links.
  2) PI: pairwise link interference.

The NI model (no interference assumption) is reasonable for the situation of narrow beams and sparse deployments. The PI model specifies that given two links $l^{(1)}, l^{(2)} \in E(L)$ sharing no common vertex, the function $\text{intf}(l^{(1)}, l^{(2)}) = 1$ if there is interference between $l^{(1)}$ and $l^{(2)}$, i.e., they cannot be scheduled simultaneously; otherwise $\text{intf}(l^{(1)}, l^{(2)}) = 0$. As shown in Fig. 3, we say that $l^{(1)}$ *interferes* $l^{(2)}$ if (1) r_2 and t_1 are within the beamwidths of each other, and (2) $\text{SINR} = \frac{p_{r_2}^{t_2}}{p_{r_2}^{t_1} + N_0} < \tau$, where p_r^t is the received power at receiver r due to transmitter t; N_0 is the noise power and τ is a SINR threshold. $\text{intf}(l^{(1)}, l^{(2)}) = 1$ if $l^{(1)}$ interferes $l^{(2)}$ or $l^{(2)}$ interferes $l^{(1)}$. Otherwise, $\text{intf}(l^{(1)}, l^{(2)}) = 0$. For the case that two links have vertices in common, we assume that advanced signal processing (MIMO precoding and combining) can be performed at the common vertices (transmitters or receivers) to remove the interference.

According to [28], the *protocol model* is a sufficiently accurate interference model for directional mmwave communication. Due to the high gain of directional antennas, interference from a single node is typically either negligible or destructive. Obviously, the protocol model with an arbitrary interference range can be translated into pairwise link interference. This justifies that the PI model is reasonably accurate and the NI model is optimistic.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

4

IEEE/ACM TRANSACTIONS ON NETWORKING

D. Optimization Goal

Our paper focuses on downlink communication. With simple adaptation, the proposed algorithms can also be applied to an uplink or a joint uplink and downlink optimization. We will explain this in §VIII, after presenting the algorithms. For downlink scheduling, the only data sources are the macro BSs, and we can thus remove all incoming arcs to macro BSs in D. We can observe in Fig. 2 that there are many ways to schedule downlink communication among the macro and relay BSs. Our goal is to find a schedule that is of unit length and is optimal with respect to a QoS metric, while satisfying given QoS requirements and the constraints on simultaneous transmissions (number of RF chains at a node, radio duplexity, spatial multiplexing model and interference model). In practice, the unit time length corresponds to the duration of a radio frame.

IV. OPTIMAL MAXIMUM THROUGHPUT FAIR SCHEDULING FOR A FULL-DUPLEX NETWORK UNDER THE NI MODEL

This section assumes a full-duplex backhaul network under the NI (No Interference) model. Given these assumptions, we provide a polynomial-time optimal algorithm for the maximum throughput fair scheduling problem (MTFS).

The goal of the MTFS problem is to maximize the downlink *network throughput* under the condition that the *max-min fairness* [29], [30] in throughput is achieved at the relay BSs.

Definition 1 (Maximum Throughput Fair Schedule):
Given a directed network D and a unit time schedule S, let $\boldsymbol{h}^S = [h_v^S | v \in M(D)]$ be the throughput vector *of S, where h_v^S denotes the throughput of a relay BS v, that is the total amount of data entering v minus that leaving v when S is applied.*
(i) A feasible unit time schedule S_f is said to satisfy the max-min fairness criterion if $\min_{v \in M(D)} h_v^{S_f} \geq \min_{v \in M(D)} h_v^S$ for any feasible unit time schedule S. The value $\min_{v \in M(D)} h_v^{S_f}$ is called the max-min throughput.
(ii) A feasible unit time schedule S^* is an optimal solution of the MTFS problem if S^* satisfies the max-min fairness criterion in (i) and the network throughput $\sum_{v \in M(D)} h_v^{S^*}$ is maximum.

In the following, we present our generally applicable optimization method—*schedule-oriented optimization*.

A. Schedule-Oriented Optimization

To characterize the scheduled data streams in each timeslot, we need the definition of *simple b-matching* of a graph [31].

Definition 2: Let G be an undirected graph with numbers $b : V(G) \mapsto \mathbb{N}$ and weights $c : E(G) \mapsto \mathbb{R}$, then a simple b-matching in G is a function $f : E(G) \mapsto \{0, 1\}$ and $\sum_{e \in \delta(v)} f(e) \leq b(v)$ for all $v \in V(G)$ where $\delta(v)$ is the set of edges incident to v. A maximum weight simple *b-matching f is a simple b-matching whose weight $\sum_{e \in E(G)} c(e)f(e)$ is maximum.*

Let $b = [r(v) | v \in V(D)]$, then it is obvious the arc set scheduled in a timeslot is a simple b-matching of D. On the other hand, a simple b-matching of D is an arc set that can be scheduled simultaneously in a timeslot. The schedule-oriented optimization solves a linear optimization problem, the solution to which is exactly the optimal schedule. For the mathematical formulation, we construct the *node-matching matrix*.

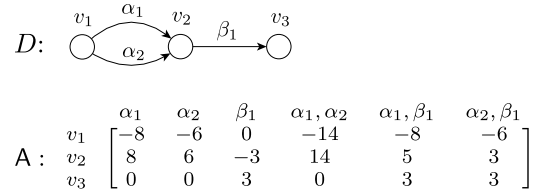*Definition 3 (Node-Matching Matrix): Given a directed network D, suppose the number of all possible simple*

Fig. 4. Node-b-matching matrix A. Every node in the backhaul network has 2 RF chains. Let two links $\alpha = (v_1, v_2)$ and $\beta = (v_2, v_3)$. The maximum number of data streams of each link is $d(\alpha) = 2$, $d(\beta) = 1$. The arc capacities in the directed network D are: $c(\alpha_1) = 8$, $c(\alpha_2) = 6$, $c(\beta_1) = 3$.

b-matchings of D is K where $b = [r(v) | v \in V(D)]$. Then the node-matching matrix $\mathsf{A} = [a_{i,j}]$ is a $|V(D)| \times K$ matrix. Each element $a_{i,j}$ is equal to the sum capacity of all arcs in the j-th simple b-matching that enter the i-th vertex of D minus the sum capacity of all arcs in the j-th simple b-matching that leave the i-th vertex of D.

As we will see, the node-matching matrix helps to formulate the throughput constraints at each relay BS. Fig. 4 gives an example of node-matching matrix for a directed network.

Let A be the node-matching matrix of D. We define A^M as a submatrix of A, which consists of the rows related to relay BSs. As the set of arcs scheduled in each timeslot of a schedule must be a simple b-matching in D, we define $\boldsymbol{t}^S$ as a $K \times 1$ *length vector*, each element of which is the length of a potential timeslot corresponding to a simple b-matching. Let the minimum throughput among all relay BSs be θ. Then we can solve the MTFS problem in two steps: (i) maximizing θ; the solution θ^* is the max-min throughput, and (ii) computing the optimal schedule S^* that offers the highest network throughput subject to the constraint $\theta \geq \theta^*$.

Linear programs for MTFS. The linear program to maximize θ in step (i) is

$$\max \theta \tag{1a}$$
$$\text{s.t. } \mathsf{A}^M \boldsymbol{t}^S \geq \mathbf{1}\theta \tag{1b}$$
$$\mathbf{1}^{\mathsf{T}} \boldsymbol{t}^S = 1 \text{ and } \boldsymbol{t}^S \geq \mathbf{0}, \tag{1c}$$

where $\mathbf{1}$ and $\mathbf{0}$ represent the all-one and all-zero column vectors. The superscript '$^{\mathsf{T}}$' denotes the vector transpose. (1b) is the constraint that the throughput at each relay BS should be at least θ. (1c) is the constraint that the schedule should be of unit length. The feasibility of the schedule is implicitly guaranteed by the formulation in terms of simple b-matchings.

After we have obtained the solution θ^* from (1a), we can formulate the linear program that maximizes the network throughput, under the condition that each relay BS has throughput at least θ^*:

$$\max \boldsymbol{c}^{\mathsf{T}} \boldsymbol{t}^S \tag{2a}$$
$$\text{s.t. } \mathsf{A}^M \boldsymbol{t}^S \geq \mathbf{1}\theta^* \tag{2b}$$
$$\mathbf{1}^{\mathsf{T}} \boldsymbol{t}^S = 1 \text{ and } \boldsymbol{t}^S \geq \mathbf{0}, \tag{2c}$$

Here, $\boldsymbol{c}$ is the *capacity vector* whose element c_j is the cumulative capacity of all macro-BS-to-relay-BS data streams in the j-th simple b-matching M_j, i.e., $c_j = \sum_{\{e | e \in M_j, \text{tail}(e) \in B(D)\}} c(e)$, where $B(D)$ is the set of all macro BSs.

The difficulty in solving (1a) and (2a) is due to the huge number of elements in $\boldsymbol{t}^S$ (equal to the number of simple b-matchings of D, which is exponential in $|V(D)|$). Yet, we show that it is unnecessary to enumerate all of them, and both (1a) and (2a) can be solved in polynomial time.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS 5

Theorem 1: Under the assumption of a full-duplex backhaul network and the NI model, the MTFS problem can be solved in polynomial time with the ellipsoid algorithm.

Proof: See Appendix A for the proof. □

Although polynomial, in practice the ellipsoid algorithm [32] almost always runs much slower than the *simplex algorithm*. Therefore, we propose algorithms based on the *revised simplex algorithm* [33] which does not require the generation of all columns of A^M. Conceptually, the algorithms first create a feasible schedule. In each iteration, to improve the optimization goal, we replace a timeslot in the schedule by another simple b-matching (a set of simultaneous data streams) while keeping the schedule feasible, until the optimum is reached. The optimum is guaranteed to be reachable due to the correctness of the simplex algorithm in solving linear programs. The maximum weight simple b-matching algorithm [34] is used to choose a better simple b-matching (column) to enter the schedule (basis).

B. Solving the MTFS Problem

To optimize θ, we need an initial basic feasible solution to (1a). Suppose that each relay BS is reachable from at least one macro BS by following a sequence of arcs in D. Let D_1 be a subgraph of D such that only the first data stream of each link (the first arc of each set of parallel arcs in D) is kept in D_1. We add a root vertex v_r to D_1 and add an arc from v_r to each macro BS vertex in D_1. We perform a *breadth-first-search* (BFS) in D_1 starting from v_r. The result is a tree T spanning v_r and all BSs. Removing v_r from T, we get a forest T' that has exactly $|M(D)|$ arcs. The initial schedule S_0 is constructed as follows: S_0 has $|M(D)|$ timeslots, each of which contains a different arc in T'. Moreover, it is required that the throughput of every relay BS is the same and the schedule takes unit time. This initial solution is unique. We convert the linear program (1a) to the standard form (3a) by introducing $|M(D)|$ surplus variables s_i.

$$\min \ \boldsymbol{f}^\mathsf{T}\boldsymbol{x} \tag{3a}$$
$$\text{s.t. } \mathsf{U}\boldsymbol{x} = \boldsymbol{g} \quad \text{and } \boldsymbol{x} \geq \boldsymbol{0}, \tag{3b}$$

where $\mathsf{U} \triangleq [\mathsf{U}^1|\mathsf{U}^2|\mathsf{U}^3] \triangleq \begin{bmatrix} \mathsf{A}^M & -\mathbf{1} & -\mathsf{I} \\ \mathbf{1}^\mathsf{T} & 0 & \mathbf{0}^\mathsf{T} \end{bmatrix}$, $\boldsymbol{f}^\mathsf{T} = [\mathbf{0}^\mathsf{T} | -1 | \mathbf{0}^\mathsf{T}]$, $\boldsymbol{x}^\mathsf{T} \triangleq [(\boldsymbol{t}^S)^\mathsf{T} | \theta | \boldsymbol{s}^\mathsf{T}]$, and $\boldsymbol{g}^\mathsf{T} \triangleq [\mathbf{0}^\mathsf{T} | 1]$. Alg. 1 shows the computation of the max-min throughput θ.

The basis B is a square matrix that consists of $|M(D)|+1$ columns from U. $\boldsymbol{f}_\mathsf{B}$ are the elements of $\boldsymbol{f}$ corresponding to B. Lines 4, 5 and 6 compute the minimum reduced cost of a column in the matrices $\mathsf{U}^1, \mathsf{U}^2$ and U^3 respectively. To decrease $-\theta$, we need to find a column of U, $\boldsymbol{u}_k$ that has negative reduced cost $f_k - \boldsymbol{p}^\mathsf{T}\boldsymbol{u}_k < 0$ to enter the basis, according to the simplex algorithm. In each iteration of the algorithm, we find the column $\boldsymbol{u}$ in U that produces minimum reduced cost η. If $\eta \geq 0$, then no columns can be used to decrease $-\theta$, thus we have reached the optimum.

Let the max-min throughput be θ^* and the related basis be B_{θ^*}. To directly use B_{θ^*} as the initial basis for the solution (2a), we add an artificial scalar variable $y \geq 0$ to (2a) and replace the constraint $\mathsf{A}^M \boldsymbol{t}^S \geq \mathbf{1}\theta^*$ with $\mathsf{A}^M \boldsymbol{t}^S - \mathbf{1}y \geq \mathbf{1}\theta^*$. Since θ^* is the max-min throughput, the feasible y must be 0. Hence, the optimal solution to (2a) is unaffected. Again, we convert (2a) into the standard form of (3a), which is solvable with the revised simplex algorithm.

Algorithm 1 Compute the Max-Min Throughput θ^*

---

1 Set the basis B according to the initial schedule S_0;
2 **while** *True* **do**
3   Compute the dual variable $\boldsymbol{p}^\mathsf{T} = \boldsymbol{f}_\mathsf{B}^\mathsf{T} \mathsf{B}^{-1}$;
4   Set weight $w(e)$ to each arc $e = (v_i, v_j)_l \in E(D)$ where

$$w(e) \triangleq \begin{cases} c(e)(p_j - p_i) & \text{if } v_i \in M(D) \\ c(e)p_j & \text{otherwise.} \end{cases}$$

   Do max weight simple b-matching on D and let the max weight be z. Compute $\eta_1 = -z - p_{|M(D)|+1}$;
5   Compute $\eta_2 = -1 + \sum_{k=1}^{|M(D)|} p_k$;
6   Compute $\eta_3 = \min_{1\leq k \leq |M(D)|} p_k$;
7   Compute $\eta = \min(\eta_1, \eta_2, \eta_3)$ and let the corresponding column be $\boldsymbol{u} \in \mathsf{U}$;
8   **if** $\eta \geq 0$ **then**
9   **return** $\theta^* = \theta$ and $\mathsf{B}_{\theta^*} = \mathsf{B}$;
10   **else**
11   Update B by replacing a column of B with $\boldsymbol{u}$ according to the simplex algorithm;
12   **end**
13 **end**

In the standard form, U remains unchanged. We redefine $\boldsymbol{f}^\mathsf{T} \triangleq [-\boldsymbol{c}^\mathsf{T} | 0 | \mathbf{0}^\mathsf{T}]$, $\boldsymbol{x}^\mathsf{T} \triangleq [(\boldsymbol{t}^S)^\mathsf{T} | y | \boldsymbol{s}^\mathsf{T}]$, and $\boldsymbol{g}^\mathsf{T} \triangleq [\mathbf{1}^\mathsf{T}\theta^* | 1]$. The optimization algorithm is similar to Alg. 1 and is outlined in Alg. 2. Since the basis B is a square matrix of dimension $|M(D)| + 1$, it follows that the optimal schedule S^* contains no more than $|M(D)| + 1$ timeslots. Additionally, since the links of a flow from a macro BS to a destination relay BS may not be scheduled in sequential order, some transmission opportunities of the flow in the first few frames may be wasted. Therefore, maximum throughput is achieved in the long-term.

C. Reducing Simple b-Matching to Matching

To do maximum weight simple b-matching, we can either use dedicated algorithms such as [34], or reduce it to a matching (equivalent to simple 1-matching) problem [35], for which highly efficient algorithms and implementations are available. In this work, we use the state-of-the-art C++ implementation for *maximum weight perfect matching* on a general graph [36], where a perfect matching matches all vertices of a graph. Thus, we need to reduce a maximum weight simple b-matching problem to a maximum weight perfect matching problem.

1) Reduction for the Case of MAX-SU-SM: We use the reduction by Tutte [37]. Given a *strict graph*[3] G whose edges have positive weights $\boldsymbol{w}$ and whose vertices have numbers $b = \boldsymbol{r} \triangleq [r(v)|v \in V(G)]$, we create a graph G' as follows. Each vertex $v \in V(G)$ is mapped to $r(v)$ vertices $v^{(1)} \ldots v^{(r(v))}$. Each edge $e = \{u, v\} \in E(G)$ with weight $w(e)$, is mapped to $r(u) \cdot r(v)$ edges $\{\{u^{(i)}, v^{(j)}\}|\forall i, \forall j\}$ in G', each of which is assigned the weight $w(e)$. Then a graph G'' is created by duplicating G' and connecting each pair of symmetric vertices by an edge of zero weight. An example of the reduction is shown in Fig. 5. It is obvious that a maximum

---

[3]A graph is *strict* if it has no loops or parallel edges between any pair of vertices.

Algorithm 2 Solving the MTFS Problem

---

1 Set the basis $\mathsf{B} = \mathsf{B}_{\theta*}$;

2 **while** *True* **do**

3   Compute the dual variable $\boldsymbol{p}^{\mathsf{T}} = \boldsymbol{f}_{\mathsf{B}}^{\mathsf{T}}\mathsf{B}^{-1}$;

4   Set weight $w(e)$ to each arc $e = (v_i, v_j)_l \in E(D)$ where

$$w(e) \triangleq \begin{cases} c(e)(p_j - p_i) & \text{if } v_i \in M(D) \\ c(e)(p_j + 1) & \text{otherwise.} \end{cases}$$

    Do max weight simple b-matching on D and let the max weight be z. Compute $\eta_1 = -z - p_{|M(D)|+1}$;

5   Compute $\eta_2 = \sum_{k=1}^{|M(D)|} p_k$;

6   Compute $\eta_3 = \min_{1 \le k \le |M(D)|} p_k$;

7   Compute $\eta = \min(\eta_1, \eta_2, \eta_3)$ and let the corresponding column be $\boldsymbol{u} \in \mathsf{U}$;

8   **if** $\eta \ge 0$ **then**

9   **return** the optimal schedule S^* corresponding to B;

10   **else**

11   Update B by replacing a column of B with $\boldsymbol{u}$;

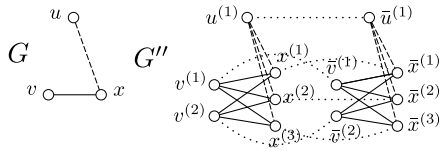**12**   **end**

13 **end**

Fig. 5. Reduction for the case of MAX-SU-SM. u, v, x have 1, 2, 3 RF chains, respectively. The edges of the same style have the same weight. Dotted edges have zero weights.

weight b-matching in G can be deduced from a maximum weight perfect matching in G''. An edge $\{u^{(i)}, v^{(j)}\}$ in a matching is mapped to the edge $\{u, v\}$ in a b-matching.

2) Reduction for the Case of REAL-SU-SM: We adapt the reduction in [38] for this case. Given a multigraph G whose edges have positive weights $\boldsymbol{w}$ and whose vertices have numbers $b = \boldsymbol{r} \triangleq [r(v)|v \in V(G)]$, we assume that $r(v) \le \deg(v)$ for any $v \in V(G)$ where $\deg(v)$ is the degree of v; otherwise we set $r(v) = \deg(v)$. We create a graph G' as follows. For each vertex v in G, we create $r(v)$ vertices in G', labeled as $v^{(1)} \ldots v^{(r(v))}$. These vertices are called *outer vertices*. For each edge $e = \{u, v\}$ in G, we add two *inner vertices* e_u, e_v to G', and $r(u) + r(v) + 1$ edges of weight $w(e)$, which are $\{u^{(i)}, e_u\}$, $\forall i$; $\{e_u, e_v\}$ and $\{e_v, v^{(j)}\}$, $\forall j$. Then G'' is created by duplicating G' and connecting each pair of symmetric outer vertices by an edge of zero weight. An example of the reduction is shown in Fig. 6. It is obvious that a maximum weight simple b-matching in G can be deduced from a maximum weight perfect matching in G''. An edge $\{u^{(i)}, e_u\}$ in a matching is mapped to the edge e in a simple b-matching.

V. Further Complexity Results on MTFS Scheduling

In this section, we will develop further computational complexity results on MTFS scheduling, for different duplexity modes, interference models and single-user spatial multiplexing (SU-SM) models.
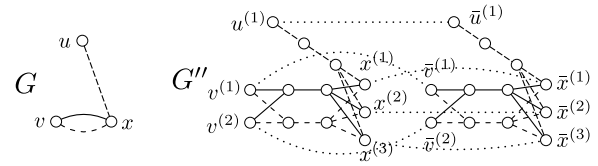

Fig. 6. Reduction for the case of REAL-SU-SM. u, v, x have 1, 2, 3 RF chains, respectively. The edges of the same style have the same weight. Dotted edges have zero weights.

A. MTFS Under the Pairwise Link Interference Model Is NP-Hard

For both full-duplex and half-duplex scheduling, the MTFS problem is NP-hard if we assume an arbitrary pairwise link interference (PI) model.

Theorem 2: The MTFS problem is NP-hard under the pairwise link interference model for both half-duplex and full-duplex backhaul networks.

Proof: The proof can be done by relating the MTFS problem to computing the fractional chromatic number. See Appendix B for details. □

B. Half-Duplex MTFS Is NP-Hard

Different from the polynomial-time solvable problem of full-duplex MTFS under the NI model, half-duplex MTFS is NP-hard, which will be proved in the following. Then we will show a special case that allows a polynomial-time optimal solution to the half-duplex MTFS problem. Since it is proved in §V-A that half-duplex MTFS is NP-hard under the PI model, we assume the NI model in this subsection.

1) Linear Programs for Half-Duplex MTFS Problem: Compared to the full-duplex case, the half-duplex scheduling has the additional half-duplex constraint—a node cannot work as transmitter and receiver simultaneously. Therefore, the matching-based optimization method that works successfully for the full-duplex scheduling cannot be applied directly. For the half-duplex case, the data streams scheduled in each timeslot must be a *half-duplex subgraph* $J \subseteq D$ which is defined as follows.

Definition 4 (Half-Duplex Subgraph): A half-duplex subgraph of a directed network D is a subgraph $J \subseteq D$ such that (i) J is a simple b-matching of D with $b = [r(v)|v \in V(D)]$ and (ii) J is a directed bipartite graph, i.e., $V(J)$ can be divided into two disjoint sets V_1, V_2 where each arc of $E(J)$ has the head in V_2 and the tail in V_1.

The constraint *(i)* is due to the number of RF chains. The constraint *(ii)* reflects the half-duplex property, because the active nodes in a timeslot can be divided into the sender set V_1 and the receiver set V_2, where a data stream only goes from a sender to a receiver. Analogous to the node-matching matrix, we define the *node-hd-subgraph matrix* for the formulation of the half-duplex MTFS problem.

Definition 5 (Node-hd-Subgraph Matrix): Given a directed network D, suppose that the number of all half-duplex subgraphs of D is K. Then the node-hd-subgraph matrix $\mathsf{L} = [l_{i,j}]$ is a $|V(D)| \times K$ matrix. Denote the i-th vertex of D as v_i, which is related to the i-th row of L. Denote the j-th half-duplex subgraph of D as J_j, which is related to the j-th column of L. Each element $l_{i,j}$ is equal to the sum capacity of all arcs in J_j that enter v_i minus the sum capacity of all arcs in J_j that leave v_i.

Fig. 7 gives an example of the node-hd-subgraph matrix for a directed network. Similar to the definition of A^M in §IV-A,
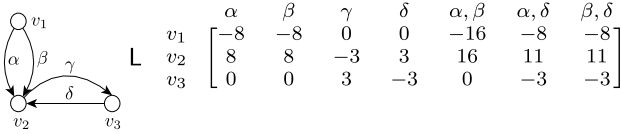
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS 7

Fig. 7. Node-hd-subgraph matrix L. v_1, v_2 and v_3 have 2, 2 and 1 RF chain respectively. The arc capacities are: $c(\alpha) = c(\beta) = 8$, $c(\gamma) = c(\delta) = 3$.

L^M is the submatrix of L that only consists of the rows related to relay BSs. The linear program formulation of the half-duplex MTFS problem is the same as (1a) and (2a) except that A^M is replaced by L^M. Yet, different from the full-duplex MTFS problem, the half-duplex MTFS problem is NP-hard. The intuitive reason is that the solution of these two problems requires computing a maximum weight simple b-matching and a maximum weight half-duplex subgraph, respectively. The first can be done in polynomial time while the second is NP-hard. We will give a formal proof in the following.

2) The Half-Duplex MTFS Problem Is NP-Hard:

Definition 6 (Maximum Weight Half-Duplex Subgraph (MWHS) Problem): Given a directed network D and a weight function $w(e)$ defined for each arc e, find a half-duplex subgraph $J \subseteq D$ such that $\sum_{e \in E(J)} w(e)$ is maximum.

Since the maximum weight simple b-matching problem can be solved in polynomial time, so can the full-duplex MTFS problem (see §IV). Analogously, if the MWHS problem could be solved in polynomial time, so could be the half-duplex MTFS problem. Unfortunately, this is not the case. The MWHS problem is NP-hard for a directed network even if it is a directed acyclic graph (DAG). Furthermore, by extending the technique for proving NP-hardness of MWHS, we can prove that the half-duplex MTFS problem is also NP-hard.

Lemma 1: The MWHS problem is NP-hard for a directed network that is a DAG.

Proof: The proof is done by reduction from the SAT problem. For further details, see Appendix C. □

Theorem 3: The half-duplex MTFS problem is NP-hard for a general directed network.

Proof: See Appendix D for the proof. □

In summary, assuming the NI model, the optimal scheduling problem of a mmWave backhaul networks can be solved in polynomial time when all BSs are full-duplex. In contrast, the problem is NP-hard when all BSs are half-duplex, i.e., it is impossible to obtain an optimal schedule in polynomial time.

C. Special Case: Half-Duplex MTFS Is Solvable in Polynomial Time

We now study a special case of the half-duplex MTFS that is solvable in polynomial time. We refer to such backhaul networks as *uniform orthogonal backhaul networks*.

Definition 7 (Uniform Orthogonal Backhaul Network): A backhaul network that is represented by the directed network D that satisfies the following conditions:
1) *There is no interference (NI model) between any pair of links.*
2) *The MAX-SU-SM model is assumed for single-user spatial multiplexing.*
3) *Each relay BS has the same number of RF chains, i.e., $r(v) \equiv r^M \in \mathbb{N}, \forall v \in M(D)$. In addition, any macro BS has the RF chain number that is a multiple of r^M, i.e., for each $i = 1 \ldots |B(D)|$ and $n_i \in B(D)$, $r(n_i) = k_i \cdot r^M$, for some $k_i \in \mathbb{N}$.*

Assuming the NI model, half-duplex MTFS is solvable in polynomial time if every node has a *single RF chain*,

because the half-duplex constraint is automatically satisfied if every node can serve only one data stream. Thus, the optimal schedule can be obtained with the optimal matching-based algorithm. Next we will prove that half-duplex MTFS is also solvable in polynomial time for uniform orthogonal backhaul networks and provide an optimal algorithm. First we look at the case that each node has the same number of RF chains R. In this case, the directed network D is a multi-digraph, each arc of which belongs to a set of R equivalent (same head, tail and capacity) parallel arcs.

Theorem 4: Given a uniform orthogonal backhaul network D, each node of which has the same number of RF chains: $r(v) \equiv R, \forall v$, the half-duplex MTFS problem can be solved in polynomial time as follows:
1) *Compute the optimal schedule S with the matching-based optimal MTFS algorithm on D's link network L where the capacity of an arc in L is the same as that of an arc in D with the same head and tail, assuming each node has one RF chain.*
2) *The optimal schedule S^* consists of R copies of S running in parallel.*

Proof: See Appendix E for the proof. □

We can now relax the condition that all nodes in D have the same number of RF chains.

Corollary 1: Given a uniform orthogonal backhaul network D, the optimal solution to the half-duplex MTFS problem can be solved in polynomial time as follows. Assume that each relay BS has the same number of RF chains, $r(v) \equiv r^M \in \mathbb{N}, \forall v \in M(D)$. In addition, any macro BS has an RF chain number that is a multiple of r^M, i.e., for each $i = 1 \ldots |B(D)|$ and $n_i \in B(D)$, $r(n_i) = k_i \cdot r^M$ with $k_i \in \mathbb{N}$.
1) *Replace each macro BS vertex n_i by k_i vertices $n_i^{(1)} \ldots n_i^{(k_i)}$, each of which has r^M RF chains. The connection of $n_i^{(j)}$ to the relay BSs is the same as that of n_i (same number of arcs with the same head and capacity). Let the resulting directed network be D'.*
2) *The optimal half-duplex MTFS schedule for D is equivalent to that for D', which is obtained with the algorithm in Theorem 4.*

Proof: The optimal half-duplex schedule for D is equivalent to the optimal one for D', since a half-duplex schedule for D can be translated into one for D' and vice versa. Moreover D' satisfies the condition of a uniform orthogonal backhaul network. □

VI. APPROXIMATION ALGORITHMS BASED ON FRACTIONAL WEIGHTED COLORING

As explained in §V, the MTFS problem is NP-hard if there is pairwise link interference or the backhaul network is half-duplex. For both cases, we must rely on approximation algorithms. Two such algorithms F³WC-FAO and F³WC-LSLO are based on the method *fractional weighted vertex coloring* of *conflict graphs*, as proposed by Wan [15]. The MTFS problem can be transformed into a fractional weighted vertex coloring problem since we can embody all four types of constraints in a conflict graph: 1) pairwise link interference, 2) number of data streams for a link restricted by spatial diversity, 3) number of data streams incident to a node restricted by the number of RF chains, and 4) half-duplex.

A. Conflict Graph

Conflict graph is a powerful tool for modeling scheduling constraints. It is an undirected simple graph (it has neither
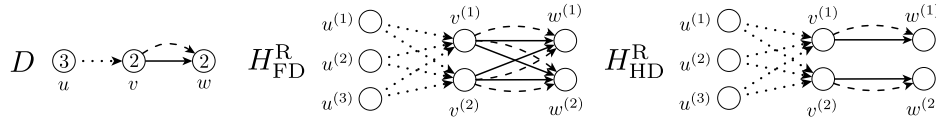
Fig. 8. Transforming a directed network D into an expanded network $H_{\mathrm{FD}}^{\mathrm{R}}$ and $H_{\mathrm{HD}}^{\mathrm{R}}$ (full-duplex/half-duplex and REAL-SU-SM model). The number inside a node is its number of RF chains. The arcs of the same style have the same capacity.

Fig. 9. Transforming a directed network D into an expanded network $H_{\mathrm{FD}}^{\mathrm{M}}$ and $H_{\mathrm{HD}}^{\mathrm{M}}$ (full-duplex/half-duplex and the MAX-SU-SM model). The number inside a node is its number of RF chains. The arcs of the same style have the same capacity.
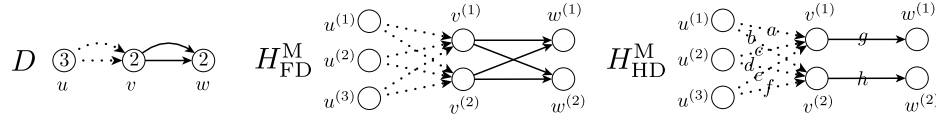
loops nor parallel edges) denoted by C, in which each vertex represents a data stream and each edge represents that two data streams cannot be scheduled simultaneously. The conflict graph is derived from the *expanded network* H, a directed graph that explicitly models the RF chains. H is the collective notation of four variants: $H_{\mathrm{FD}}^{\mathrm{R}}, H_{\mathrm{FD}}^{\mathrm{M}}, H_{\mathrm{HD}}^{\mathrm{R}}$ and $H_{\mathrm{HD}}^{\mathrm{M}}$, depending on the modeling. Each vertex of C is one-to-one mapped to each arc of H, $V(C) = E(H)$.

1) Full-Duplex Scheduling: Let us first consider the most general backhaul network which is subjected to the PI model and the REAL-SU-SM model. Given a backhaul network D, we assume that the RF chain number $r(v) \leq \deg(v), \forall v \in V(D)$. Otherwise, we set $r(v) = \deg(v)$ as the extra RF chains are redundant. The expanded network $H_{\mathrm{FD}}^{\mathrm{R}}$ is created by mapping each vertex $v \in V(D)$ into vertices $v^{(1)} \dots v^{(r(v))}$. Each arc $l_i = (u,v)_i \in E(D)$ (the i-th data stream from u to v) is mapped to $r(u) \cdot r(v)$ arcs $\{(u^{(j)}, v^{(k)})_i | \forall j, k\}$, each with capacity $c(l_i)$. We define the *expanded arcs of a link* (u,v) in L as $X\big((u,v)\big) = \{(u^{(j)}, v^{(k)})_i | \forall i, j, k\}$. We define the *expanded arcs of a data stream* $(u,v)_i$ in D as $X\big((u,v)_i\big) = \{(u^{(j)}, v^{(k)})_i | \forall j, k\}$. An example for the expanded network is shown in the middle of Fig. 8.

The conflict graph C is constructed as follows. Let a *complete graph* $K(V)$ be an undirected graph such that there is an edge between each pair of vertices in V. The edges of C are constructed by first adding the union of the edge sets of a number of complete graphs. They are 1) the ones formed by the arcs incident to each vertex in $H_{\mathrm{FD}}^{\mathrm{R}}$, $K(\delta_{H_{\mathrm{FD}}^{\mathrm{R}}}(v)), \forall v$, and 2) the ones formed by the expanded arcs of each data stream $X(e), \forall e \in E(D)$. Then we add the edges representing pairwise link interference. For each pair of interfering links in L, say l and l', we add to C the edges $\big\{\{e,e'\}|e \in X(l), e' \in X(l')\big\}$.

If the MAX-SU-SM model is assumed instead of the REAL-SU-SM model, then the expanded network $H_{\mathrm{FD}}^{\mathrm{M}}$ contains fewer arcs than $H_{\mathrm{FD}}^{\mathrm{R}}$. Again each vertex $v \in V(D)$ is mapped into $r(v)$ vertices. If there is a link $(u,v) \in E(L)$ of capacity c, then D contains $\min(r(u), r(v))$ arcs (data streams) from u to v with the same capacity c. The link is mapped into $r(u) \cdot r(v)$ arcs $\{(u^{(j)}, v^{(k)})|\forall j, k\}$ in $H_{\mathrm{FD}}^{\mathrm{M}}$, all having capacity c. The *expanded arcs of a link* (u,v) in L are defined as $X\big((u,v)\big) = \{(u^{(j)}, v^{(k)})|\forall j, k\}$. An example for the expanded network is shown in the middle of Fig. 9.

We first add the edges of the complete graphs formed by the arcs incident to each vertex in $H_{\mathrm{FD}}^{\mathrm{M}}$. Then for each pair of interfering links in L, say l and l', we add to C the edges $\big\{\{e,e'\}|e \in X(l), e' \in X(l')\big\}$.

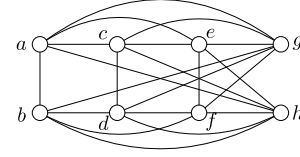*2) Half-Duplex Scheduling:* Again let us first consider the most general backhaul network subject to the

Fig. 10. The conflict graph for $H_{\mathrm{HD}}^{\mathrm{M}}$ in Fig. 9.

PI model and the REAL-SU-SM model. The expanded network $H_{\mathrm{HD}}^{\mathrm{R}}$ is more sparse than the full-duplex counterpart $H_{\mathrm{FD}}^{\mathrm{R}}$. Given a backhaul network D, we set $r(v) = \min\big(\max\big(\deg_-(v), \deg_+(v)\big), r(v)\big)$ where $\deg_-(v)$ and $\deg_+(v)$ are the number of incoming and outgoing arcs of v in D. The reason is that a higher number of RF chains is unnecessary. $H_{\mathrm{HD}}^{\mathrm{R}}$ is created by first mapping the vertices in D the same way as before. Then each arc $l_i = (u,v)_i \in E(D)$ is mapped as follows. If $r(u) \neq r(v)$, l_i is mapped to $r(u) \cdot r(v)$ arcs $\{(u^{(j)}, v^{(k)})_i | \forall j, k\}$, each with capacity $c(l_i)$ the same way as for $H_{\mathrm{FD}}^{\mathrm{R}}$. Otherwise, $r(u) = r(v)$, l_i is mapped to $r(u)$ arcs $\{(u^{(j)}, v^{(j)})_i | \forall j\}$, each with capacity $c(l_i)$. An example for the expanded network is shown in the right side of Fig. 8.

For a vertex v in D, we denote $\delta_{H_{\mathrm{HD}}^{\mathrm{R}}}^-(v)$ and $\delta_{H_{\mathrm{HD}}^{\mathrm{R}}}^+(v)$ as the arcs in $H_{\mathrm{HD}}^{\mathrm{R}}$ that enter or leave the vertices $v^{(j)}$ for all j, respectively. The conflict graph C is first constructed with the method for $H_{\mathrm{FD}}^{\mathrm{R}}$. Then we add to C the edges $\big\{\{e,e'\}|e \in \delta_{H_{\mathrm{HD}}^{\mathrm{R}}}^-(v), e' \in \delta_{H_{\mathrm{HD}}^{\mathrm{R}}}^+(v), \forall v \in V(D)\big\}$. These edges model the half-duplex constraint.

If the MAX-SU-SM model is assumed, then the expanded network $H_{\mathrm{HD}}^{\mathrm{M}}$ is even more sparse than $H_{\mathrm{HD}}^{\mathrm{R}}$. Each link $(u,v) \in E(L)$ with capacity c is mapped as follows. If $r(u) \neq r(v)$, (u,v) is mapped to $r(u) \cdot r(v)$ arcs $\{(u^{(j)}, v^{(k)})|\forall j, k\}$, each with capacity c. Otherwise, it is mapped to $r(u)$ arcs $\{(u^{(j)}, v^{(j)})|\forall j\}$, each with capacity c. An example for the expanded network is shown in the right side of Fig. 9.

The conflict graph C is first constructed with the method for $H_{\mathrm{FD}}^{\mathrm{M}}$. Then we add to C the edges $\big\{\{e,e'\}|e \in \delta_{H_{\mathrm{HD}}^{\mathrm{M}}}^-(v), e' \in \delta_{H_{\mathrm{HD}}^{\mathrm{M}}}^+(v), \forall v \in V(D)\big\}$. The conflict graph for $H_{\mathrm{HD}}^{\mathrm{M}}$ in Fig. 9 is shown in Fig. 10.

A sparse expanded network leads to a conflict graph with fewer vertices and hence shorter execution time for the algorithms. We will prove in the following why the sparse expanded networks $H_{\mathrm{HD}}^{\mathrm{R}}$ and $H_{\mathrm{HD}}^{\mathrm{M}}$ can be used for half-duplex scheduling.

Theorem 5: Given a directed network D, then any half-duplex subgraph of D can be represented by a matching in $H_{\mathrm{HD}}^{\mathrm{M}}$.

Proof: See Appendix F for the proof. □

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS 9

Algorithm 3 First-Fit Fractional Weighted Coloring

Input : C, $\boldsymbol{t} \in \mathbb{R}_+^{V(C)}$, and an ordering of $V(C)$.
Output: A fractional weighted coloring Π of $(C, \boldsymbol{t})$.
1 $\Pi \leftarrow \emptyset$;
2 $U \leftarrow \{v \in V(C) | t_v > 0\}$;
3 **while** $U \neq \emptyset$ **do**
4    $I \leftarrow$ the first-fit MIS (maximal independent set) of U;
5    $\lambda \leftarrow \min_{v \in I} t_v$;
6   add (I, λ) to Π;
7   **for** *each* $v \in I$ **do**
8       $t_v \leftarrow t_v - \lambda$;
9   **if** $t_v = 0$ **then**
10   remove v from U;
11   **end**
12   **end**
13 **end**
14 output Π;

B. General Procedure of Fractional Weighted Coloring Based Approximation Algorithms

A fractional weighted coloring based approximation algorithm consists of three steps: (i) computing the *data stream time vector* $\boldsymbol{t} = [t_e | e \in E(H)] = [t_v | v \in V(C)]$, (ii) sorting the vertices $V(C)$ and performing F^{3}WC with the given ordering, and (iii) scaling the schedule. We use the results of [15] and adapt two approximation algorithms based on fractional weighted coloring. The difference of the two algorithms lies in the linear programs for computing the link time vector $\boldsymbol{t}$ and the ordering of $V(C)$ for coloring. The coloring step uses the so-called *first-fit fractional weighted coloring* (F^{3}WC) algorithm from [15], listed in Alg. 3.

How to compute $\boldsymbol{t}$ depends on the specific algorithm. The *minimum makespan* scheduling for $\boldsymbol{t}$ is the same as the *minimum fractional weighted coloring* of $(C, \boldsymbol{t})$. The latter is defined as a set of $K \in \mathbb{N}$ pairs (I_i, λ_i) where each I_i is an independent set (a set of nonadjacent vertices) of C and $\lambda_i \in \mathbb{R}_+$ for $1 \leq i \leq K$ satisfying that $\sum_{1 \leq i \leq K, v \in I_i} \lambda_i = t_v, \forall v \in V(C)$ and the sum $\sum_{i=1}^K \lambda_i$ is the minimum. But the problem of finding a minimum fractional weighted coloring is NP-hard [39]. Let P be the *independence polytope* of C, i.e., the convex hull of the incidence vectors of the independent sets of C. Then any point in P corresponds to a feasible unit time schedule. The minimum fractional weighted coloring problem can be expressed as a linear program with the help of P.

We assume that an algorithm provides a γ-approximate ($\gamma > 1$) independent polytope Q°, i.e., $Q^\circ \subseteq P \subseteq \gamma Q^\circ$. Specifically, we have two options—F^{3}WC-FAO or F^{3}WC-LSLO with $\gamma = \alpha^*, Q^\circ = Q$ and $\gamma = 2\beta^*, Q^\circ = Q'$, respectively (the definition of these variables will be clear in the following). Step (i) is to solve the following two linear programs.

$$\max \ \theta \tag{4a}$$

$$\text{s.t.} \sum_{e \in \delta_H^-(U(v))} c(e)t_e - \sum_{e \in \delta_H^+(U(v))} c(e)t_e \geq \theta \quad \forall v \in M(D) \tag{4b}$$

$$\boldsymbol{t} \in Q^\circ, \tag{4c}$$

where $U(v) = \{v^{(i)} \in V(H) | \forall i\}$.

With the max-min throughput solution θ, we go on to compute $\boldsymbol{t}$ for the maximum network throughput.

$$\max \sum_{e \in \{\delta_H^+(U(v)) | v \in B(D)\}} c(e)t_e$$

$$\text{s.t. (4b) and (4c).}$$

Step (ii) is to sort $V(C)$ with the given method and then to perform the F^{3}WC algorithm (Alg. 3) with the computed $\boldsymbol{t}$ and vertex ordering. Since, it is guaranteed by step (i) and (ii) that the schedule length after performing the F^{3}WC algorithm is no more than one, we perform the last step to scale the schedule length to exactly unit time. The goal is to improve performance by fully utilizing the available time resource.

C. Fixed and Arbitrary Ordering (F^{3}WC-FAO)

Assume that $\langle v_1 \ldots v_n \rangle$ is an arbitrary but fixed ordering of $V(C)$ where $n = |V(C)|$. We denote $v_i < v_j$ if $i < j$. Let V_i be the set of vertices of v_i and all its *smaller* neighbors (neighbors in $\{v_1 \ldots v_{i-1}\}$). Define the *inductive independence polytope* Q of C by the ordering $\langle v_1 \ldots v_n \rangle$ as

$$Q \triangleq \left\{ \boldsymbol{t} \in \mathbb{R}_+^{V(C)} \middle| \max_{1 \leq i \leq n} t(V_i) \leq 1 \right\}, \tag{5}$$

where $t(V_i) = \sum_{v \in V_i} t_v$. Q is an approximation of the independence polytope P.

D. Largest Surplus Last Ordering (F^{3}WC-LSLO)

The largest surplus last ordering of $V(C)$ is done by first transforming the undirected graph C into a directed graph C^d by imposing a certain orientation on each edge. We specify the following orientation.

Suppose that the vertices of the directed network D have an ordering. That is, given two different vertices $w, w' \in V(D)$, if w comes before w' in the ordering, we denote $w < w'$. Given two different vertices $u^{(i)}$ and $v^{(j)}$ of the expanded network H, we denote $u^{(i)} < v^{(j)}$ if and only if $u < v$ or ($u = v$ and $i < j$). For the MAX-SU-SM model, given two different vertices $(u^{(i)}, v^{(j)})$ and $(s^{(k)}, t^{(l)}) \in V(C^d)$, $(u^{(i)}, v^{(j)}) < (s^{(k)}, t^{(l)})$, if and only if $u^{(i)} < s^{(k)}$ or ($u^{(i)} = s^{(k)}$ and $v^{(j)} < t^{(l)}$). For the REAL-SU-SM model, given two different vertices $(u^{(i)}, v^{(j)})_m$ and $(s^{(k)}, t^{(l)})_n \in V(C^d)$, $(u^{(i)}, v^{(j)})_m < (s^{(k)}, t^{(l)})_n$ if and only if $(u^{(i)}, v^{(j)}) < (s^{(k)}, t^{(l)})$ or $((u^{(i)}, v^{(j)}) = (s^{(k)}, t^{(l)})$ and $m < n)$.

The orientation is chosen according to the following rules for each edge in C. Note, the subscripts m, n are taken as empty for the MAX-SU-SM model.
1) An edge between $(u^{(i)}, v^{(j)})_m$ and a vertex of the form $(v^{(k)}, x^{(l)})_n$ such that $u \neq x$ has the orientation from the first to the second.
2) Otherwise, an edge between two vertices has the orientation from the small one to the large one.

Let D' be a digraph. For a vertex $u \in V(D')$, let $N^{in}(u)$ denote the set of in-neighbors of u in D', and let $N^{in}[u]$ denote $\{u\} \cup N^{in}(u)$. $N^{out}(u)$ and $N^{out}[u]$ are defined correspondingly. For any $\boldsymbol{t} \in \mathbb{R}_+^{V(D')}$, the *surplus* of a vertex u is defined as $t(N^{in}(u)) - t(N^{out}(u))$. The largest surplus last ordering is constructed as follows. Let $\boldsymbol{t} \in \mathbb{R}_+^{V(C^d)}$. Initialize D' to C^d. For $i = n$ down to 1, let v_i be a vertex of the largest surplus in $(D', \boldsymbol{t})$ and then delete v_i from D' and the element t_{v_i} from $\boldsymbol{t}$. The ordering of $\langle v_1 \ldots v_n \rangle$ is the largest surplus

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10

IEEE/ACM TRANSACTIONS ON NETWORKING

last ordering of $(C^d, \boldsymbol{t})$. The *independence polytope* Q' of C^d is defined as

$$Q' = \left\{ \boldsymbol{t} \in \mathbb{R}_+^{V(C^d)} \middle| \max_{u \in V(C^d)} t(N^{in}[u]) \leq 1/2 \right\}, \quad (6)$$

which is another approximation of the independence polytope P.

E. Approximation Ratios in Terms of Max-Min Throughput

The following theorem presents the worst-case approximation ratios in terms of max-min throughput of the two algorithms F^{3}WC-FAO and F^{3}WC-LSLO.

Theorem 6: The algorithms F^{3}WC-FAO and F^{3}WC-LSLO solve the MTFS problem by producing a unit-time schedule. They achieve a max-min throughput $\theta' \geq \theta^*/\alpha^*$ and $\theta' \geq \theta^*/(2\beta^*)$ respectively, where θ^* is the optimum and

- *for the case of a full-duplex network, PI and REAL-SU-SM model:*

$$\alpha^* \leq \max\left(1, \max_{l \in E(L)} \Big(\sum_{l' | \text{intf}(l',l)=1} d(l') \Big)\right) + 2$$

$$\beta^* \leq \max\left(1, \max_{l \in E(L)} \Big(\sum_{l' | l' < l, \text{intf}(l',l)=1} d(l') \Big)\right) + 2,$$

- *for the case of a full-duplex network, PI and MAX-SU-SM model:*

$$\alpha^* \leq \max_{l \in E(L)} \Big(\sum_{l' | \text{intf}(l',l)=1} d(l') \Big) + 2$$

$$\beta^* \leq \max_{l \in E(L)} \Big(\sum_{l' | l' < l, \text{intf}(l',l)=1} d(l') \Big) + 2,$$

- *for the case of a half-duplex network and PI model:*

$$\alpha^* \leq \max_{l \in E(L)} \Big(r(l) + \sum_{l' | \text{intf}(l',l)=1} d(l') \Big)$$

$$\beta^* \leq \max_{l=(u,v) \in E(L)} \Big(r(u) + \sum_{l' | l' < l, \text{intf}(l',l)=1} d(l') \Big) + 1$$

where $r(l) = r(u) + r(v)$.
Proof: See Appendix G for the proof. □

VII. APPROXIMATION ALGORITHM OF PARALLEL DATA STREAM SCHEDULING

This section proposes an effective approximation algorithm for half-duplex MTFS scheduling under the NI model. The PDS (Parallel Data Stream Scheduling) approximation algorithm (listed in Alg. 4) extends the optimal half-duplex MTFS algorithm in §V-C to cover the situation that an optimal MTFS schedule cannot be found in polynomial time. It is based on the idea that the parallel data streams between a pair of BSs are always scheduled simultaneously. An example of the graph transformation step (Line 1 to 3) of the PDS algorithm is shown in Fig. 11.

Theorem 7: Suppose that the optimal max-min throughput of the half-duplex MTFS problem on a directed network D is θ^* under the NI model. Let the max-min throughput obtained with the PDS algorithm be θ and $r_{min} = \min_{u \in V(D)} r_D(u)$ be the minimum RF chain number of any BS. Let $r_{max}^M = \max_{u \in M(D)} r_D(u)$ be the maximum RF chain number of any relay BS, and $d_{min} = \min_{l \in E(L)} d(l)$ be the minimum data stream number of any link. We have $\theta \geq \theta^*/\gamma^*$, where

---

Algorithm 4 PDS Algorithm

1 Create a network D^e based on the directed network D. D^e copies the relay BS vertices and the arcs between them from D while keeping the values of RF chain number and capacity unchanged. Let the minimum data stream number of any link in the link network L be $d_{min} = \min_{l \in E(L)} d(l)$. Each macro BS vertex v in D is mapped into $s(v)$ macro BS vertices $v^{(1)} \dots v^{(s(v))}$ in D^e where

$$s(v) \triangleq \lfloor r_D(v)/d_{min} \rfloor. \quad (7)$$

The RF chain number of each macro BS vertex $v^{(i)}$ is defined as

$$r_{D^e}(v^{(i)}) \triangleq \begin{cases} d_{min} & \text{if } i < s(v) \\ m(v) = r_D(v) - [s(v)-1]d_{min} & \text{otherwise.} \end{cases} \quad (8)$$

For each i, create $d = \min(r_{D^e}(v^{(i)}), r_{D^e}(w), d(v,w))$ arcs $(v^{(i)}, w)_j$ in D^e such that $c_{D^e}((v^{(i)}, w)_j) = c_D((v, w)_j), \forall j = 1 \dots d$, for each neighbor w of v in D;

2 Make a copy D^s of D^e and replace each set of parallel arc with a single arc. For each arc (u, v) in D^s, define the capacity function associated with D^s as $c_{D^s}((u,v)) \triangleq \sum_j c_{D^e}((u,v)_j)$. Define the RF chain number function associated with D^s as $r_{D^s}(u) \triangleq 1, \forall u \in V(D^s)$;

3 Compute the optimal MTFS schedule S for D^s with the capacity c_{D^s} and RF chain number r_{D^s} using the method in §IV-B;

4 Create the final schedule S^* based on S, by mapping the activation of an arc in D^s into the simultaneous activation of parallel arcs in D;
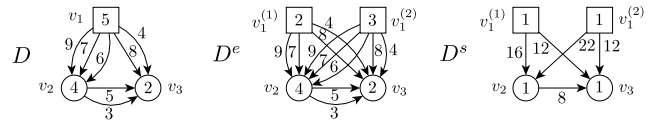

Fig. 11. Example of the graph transformations in PDS. □ and ○ represents macro BS and relay BS, respectively. The number inside a node v is $r_G(v)$ and the number next to an arc e is $c_G(e)$ where G is the related graph.

- $\gamma^* = \max(r_{max}^M, \max_{v \in B(D)} m(v)) \leq \max(r_{max}^M, 2d_{min} - 1)$, *where $m(v)$ is defined in (8), if the REAL-SU-SM model is assumed;*
- $\gamma^* = \dfrac{\max(r_{max}^M, \max_{v \in B(D)} m(v))}{r_{min}} \leq \dfrac{\max(r_{max}^M, 2r_{min}-1)}{r_{min}}$, *if the MAX-SU-SM model is assumed.*

Proof: See Appendix H for the proof. □

Let us consider some special cases of the MAX-SU-SM model. If each relay BS in D has the same RF chain number r^M and each macro BS has an RF chain number that is a multiple of r^M, then the PDS algorithm attains the optimal MTFS schedule. On the other hand, if each relay BS has r^M RF chains and any macro BS has at least r^M RF chains, then PDS has a worst-case performance ratio of $1/2$ for the max-min throughput.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS 11

Corollary 2: Assume the NI model and the MAX-SU-SM model. Given a directed network D, assume that each relay BS has the same number of RF chains r^M and any macro BS has at least r^M RF chains. The PDS algorithm achieves the max-min throughput $\theta > \theta^*/2$ where θ^* is the optimum for the half-duplex MTFS problem.

Proof: According to Theorem 7, $\theta \geq \frac{r^M}{2r^M - 1}\theta^* > \theta^*/2$. $\square$

In summary, under the NI model, the three algorithms: PDS, F³WC-FAO and F³WC-LSLO are respectively $\frac{1}{\max(r^M_{max}, \max_{v \in B(D)} m(v))} \geq \frac{1}{r_{max}}$, $1/\max_{l \in E(L)}(r(l))$ and $1/(2r_{max} + 2)$-approximate algorithms for the half-duplex MTFS problem, where $r_{max} = \max_{v \in V(D)} r_D(v)$. A ρ-approximate ($\rho \leq 1$) algorithm achieves a max-min throughput θ that is at least ρ times that of the optimal value θ^*, $\theta \geq \rho\theta^*$. Theoretically, PDS has the best performance and F³WC-LSLO has the worst.

VIII. EXTENSION TO INTEGRATED ACCESS AND BACKHAUL

To date, 3GPP is investigating the standardization of Integrated Access and Backhaul (IAB) for mmWave cellular networks [24]. Yet designing a high-performance IAB network is still an open problem [25], [26]. This paper offers joint routing and scheduling algorithms for IAB networks with optimal or guaranteed QoS. Both the optimal algorithm and approximation algorithms proposed in this paper can readily be applied to the scenario of integrated backhaul and access (IAB) networks. Due to the graph-based network modeling, our approach is applicable to both IAB networks and backhaul networks. However, the runtime efficiency may be an issue, if the IAB network includes numerous user equipments (UE).

The proposed algorithms in this paper solve a downlink optimization problem. Yet with slight modification, they can solve an uplink or a joint uplink and downlink optimization problem. By doing so, the optimal algorithms still retain their optimality while the approximation algorithms keep their approximation ratios. A joint uplink and downlink optimization may use the resources better than two separate optimizations. Conceptually, every algorithm in this paper has two parts, the routing part and the data stream conflict resolving part, which are performed either sequentially or intertwined. The routing part is a linear program that finds an efficient routing scheme for arbitrary throughput requirements on sources and destinations. So it naturally supports an uplink or a joint uplink and downlink optimization. The data stream conflict resolving part uses either the matching technique for the optimal algorithms or the conflict graph technique for the approximation algorithms.

In addition, our algorithms can be extended to solve other problems than MTFS. These include problems that can be formulated as a linear program whose variables are the active time of data streams and QoS metrics. For example, we can optimize for the constraint that each relay BS has a minimum throughput requirement. Another example is to optimize the energy consumption as it can be translated into the minimization of total transmission time in a schedule. We do not further elaborate on them as the extension is straightforward.

IX. NUMERICAL EVALUATION

In this section, we evaluate the proposed optimal and approximation algorithms for the MTFS problem in terms of max-min throughput, network throughput and execution time.
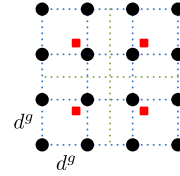

Fig. 12. An example backhaul network consisting of 4×4 relay BSs and 2×2 macro BSs.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Distance between 2 grid lines, d^g	80 m
Carrier frequency, f	28 GHz
Path loss parameters[4] α, β, σ in $PL(d) = \alpha + 10\beta \log_{10} d + \xi$	LOS: $\alpha = 61.4, \beta = 2, \sigma = 5.8$ NLOS: $\alpha = 72, \beta = 2.92, \sigma = 8.7$
Transmission power, p_{tx}	30 dB
Directivity gain, g_x	30 dB
Bandwidth, b	1 GHz
Noise $N_0 = kT_0 + F + 10\log_{10} b$	$kT_0 = -174$ dBm/Hz, $F = 4$ dB
Min SINR threshold for reception, τ	5 dB
Number of data streams, K	$K \sim \max\{\text{Poisson}(\lambda), 1\}, \lambda = 1.8$
Beamwidth, ϕ	$\phi = 20°$
Correlation coefficient in the exponential correlation matrix, r	$r = 0.9$

A. Evaluation Setting

We simulate an mmWave backhaul network, which consists of $n \times n$ relay BSs and $j \times k$ macro BSs. The relay BSs are placed on the intersections of n horizontal and n vertical grid lines. The distance between two neighboring grid lines is d^g. The grid plane is divided into $j \times k$ equal rectangles and a macro BS is placed at each rectangle center (see Fig. 12 for an example). We assume channel reciprocity in the simulation. The capacity of a link is computed with the formula of Shannon capacity. This is the value if one RF chain is used to serve the link on both ends. The received power is given by $p_{rx} = p_{tx} + g_x - PL$ where p_{tx} is the transmission power, g_x is the directivity gain and PL is the path loss. We assume a carrier frequency of 28 GHz. The channel state of a link is simulated according to the statistical model derived from the real-world measurement [40]. There are three possible channel states—LOS (line-of sight), NLOS (non line-of-sight) or outage. We only keep the links that are in LOS or NLOS state and have an SNR higher than 5 dB. The simulation parameters are listed in Tab. I.

For the PI model, we simulate the pairwise link interference according to the model in §III-C. As illustrated in Fig. 3, the 4 links $(t_1, r_1), (t_1, r_2), (t_2, r_1), (t_2, r_2)$ are assumed to be independent.

To simulate the REAL-SU-SM model, we assume that the maximum number of data streams supported by a link is Poisson distributed with the mean value 1.8 (Tab. I), following the empirical model of [40]. The total capacity of a link increases sublinearly to the number of data streams and is simulated according to the exponential correlation matrix model in [41] by choosing the correlation coefficient $r = 0.9$. A comparison of the total capacity of parallel data streams for the REAL-SU-SM and MAX-SU-SM models is shown in Fig. 13.

The proposed algorithms are implemented in MATLAB, except that we use the C++ program Blossom V for
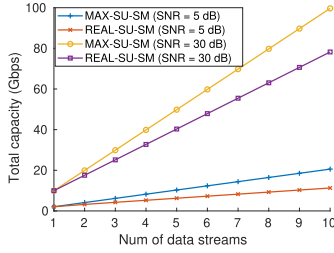
Fig. 13. Comparison of the total capacity of parallel data streams for the REAL-SU-SM and MAX-SU-SM models.

TABLE II
THE RATE OF θ^* OF MAX-SU-SM TO THAT OF REAL-SU-SM

Avg. rate	$r^M = 1$	$r^M = 2$	$r^M = 3$	$r^M = 4$	$r^M = 5$
$r^B = 1$	1.00	1.00	1.00	1.01	1.00
$r^B = 2$	1.00	1.01	1.03	1.06	1.08
$r^B = 3$	1.00	1.01	1.05	1.08	1.10
$r^B = 4$	1.00	1.02	1.08	1.11	1.14
$r^B = 5$	1.00	1.03	1.10	1.14	1.17

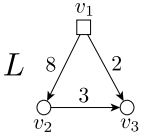| algo	OPT-FD-MTFS	OPT-HD-MTFS	
slot	#1	#1	#2
sched	$v_1 \rightarrow v_2$ $v_1 \rightarrow v_3$ $v_2 \rightarrow v_3$	$v_1 \rightarrow v_2$ $v_1 \rightarrow v_2$	$v_2 \rightarrow v_3$ $v_2 \rightarrow v_3$
time	1	0.4286	0.5714

Fig. 14. v_1 is the macro BS, and v_2 and v_3 are relay BSs. Each node has 2 RF chains. We assume NI and MAX-SU-SM models. The optimal max-min throughput for full-duplex MTFS and half-duplex MTFS problems are 5 and 3.43 respectively.

minimum cost perfect matching [36] and Gurobi [42] for linear programming.

We evaluate the optimal algorithms OPT-FD-MTFS and OPT-HD-MTFS as well as three approximation algorithms— F³WC-FAO, F³WC-LSLO, and PDS, for 10 backhaul networks with 10×10 relay BSs. OPT-FD-MTFS works for full-duplex scheduling under the NI model while OPT-HD-MTFS works for half-duplex scheduling of uniform orthogonal backhaul networks. F³WC-FAO and F³WC-LSLO are generally applicable for any combination of half-duplex/full-duplex, NI/PI model and MAX-SU-SM/REAL-SU-SM model while PDS only works for half-duplex scheduling under the NI model.

We place 1, 2×1 or 2×2 macro BSs in each network. The macro BSs and the relay BSs have the same number of RF chains r^B and r^M respectively, while r^B and r^M range from 1 to 5.

B. Optimal Algorithms

Both full-duplex and half-duplex optimal schedules can be computed efficiently for uniform orthogonal backhaul networks. Surprisingly, for such networks, the max-min throughput of both OPT-HD-MTFS and OPT-FD-MTFS schedules are usually the same. We believe that the close performance of max-min throughput for both half-duplex and full-duplex scheduling is due to the good connectivity of the backhaul network which allows plenty of scheduling possibilities. A simple network in Fig. 14 shows that the performance gap can be large.

The max-min throughput θ^* of OPT-FD-MTFS for the REAL-SU-SM model is shown in Fig. 15 for various number of macro BSs and RF chains. Generally, θ^* increases with the number of RF chains of relay BS (r^M) and of macro BS (r^B), as well as the number of macro BSs. When the number of macro BSs and r^B are fixed, θ^* gradually saturates despite the increase of r^M. In such cases, the bottleneck is at the links between macro BSs and relay BSs. To achieve higher performance in θ^*, we need to increase all three variables. Yet, adding macro-BSs would be very costly. Adding more RF chains to each macro-BS while increasing the relay BSs that are neighbors to these macro BSs seems like a more cost-effective approach. Moreover, the average throughput per relay

BS is from 1x to 1.96x of the max-min throughput. This shows that in a dense network, we can achieve a rather equal distribution of throughput among relay BSs. As expected, θ^* of OPT-FD-MTFS for the MAX-SU-SM model is greater than or equal to that of the REAL-SU-SM model. The difference increases with r^M and r^B (Tab. II), which shows that multiple RF chains are especially beneficial to a rich multi-path channel.

The distributions of execution time of OPT-FD-MTFS and OPT-HD-MTFS for uniform orthogonal backhaul networks are shown in Fig. 16. OPT-HD-MTFS achieves almost the same performance in max-min throughput and network throughput as OPT-FD-MTFS, yet it runs much faster than the latter, by shortening the execution time by 27% on average and by 79% in the best case. The reason is due to the step of merging RF chains in the OPT-HD-MTFS (same as PDS) algorithm which leads to a smaller (in terms of vertices and arcs) graph on which matching is performed. Recall that in general cases, the HD-MTFS problem is NP-hard. In addition, we observe from Fig. 16 that the execution time increases with the number of macro BSs for both algorithms. In addition, the execution time of OPT-FD-MTFS also increases with the number of RF chains at BSs due to the growth of the graph for matching.

C. Full-Duplex Approximation Algorithm

If there is mutual interference between links in a backhaul network, we cannot use the optimal full-duplex MTFS scheduling algorithm. However, two fractional weighted coloring based approximation algorithms proposed in §VI can be applied. Because the MTFS problem is NP-hard under the PI model, we use the performance of OPT-FD-MTFS as an upper bound. Fig. 17 shows the results for the REAL-SU-SM model. We observe that mmWave backhaul networks are noise-limited instead of interference-limited. On average, there are 611 directional links in an evaluated backhaul network, among which only 21 pairs of links are interfering, although we choose a relatively large beamwidth of 20°. Despite considering the interference, both algorithms achieve on average more than 70% of the optimal max-min throughput for the ideal interference-free case. In general, F³WC-FAO outperforms F³WC-LSLO in terms of max-min throughput. Besides, the theoretical approximation ratios of Theorem 6 significantly underestimate the actual performance of the F³WC algorithms. The results for the MAX-SU-SM model are omitted as they are similar.

The execution time of the F³WC algorithms and OPT-FD-MTFS are shown for two SU-SM models in Fig. 18. In general, it takes OPT-FD-MTFS less than 100 seconds to schedule a backhaul network with 100 relay BSs and the execution time even decreases with r^M. Thus, it is practical to compute the optimal schedule for full-duplex backhauls if interference can be ignored. The approximation algorithms are more efficient than OPT-FD-MTFS when r^M is small. Yet the execution time goes up quickly with r^M, especially for the REAL-SU-SM model. The reason is due to the large number of vertices in
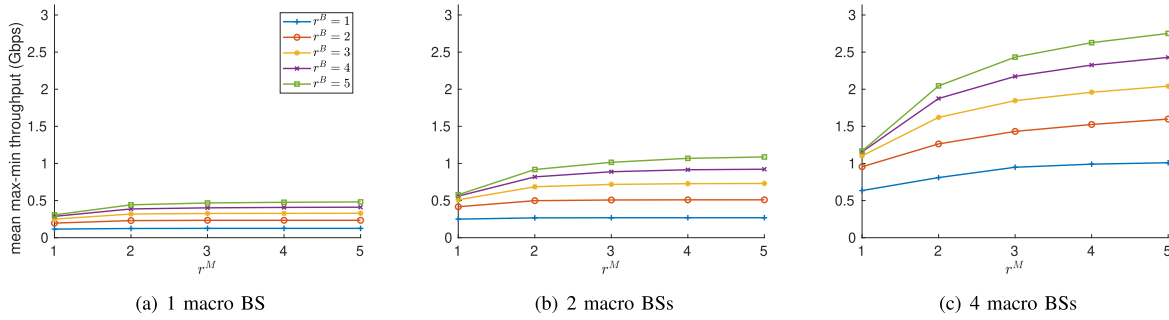
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS 13

(a) 1 macro BS   (b) 2 macro BSs   (c) 4 macro BSs

Fig. 15. Max-min throughput of the OPT-FD-MTFS algorithm for the REAL-SU-SM model and for different number of RF chains.

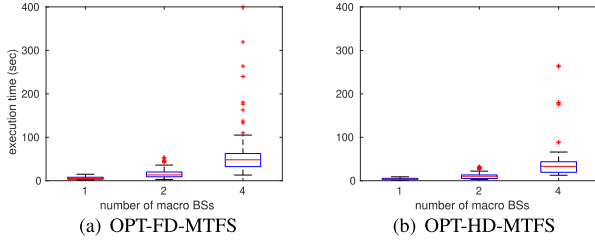(a) OPT-FD-MTFS   (b) OPT-HD-MTFS

Fig. 16. Execution time of optimal algorithms for uniform orthogonal backhaul networks.
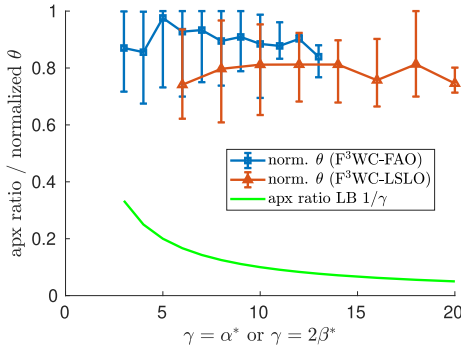

(a) PI and REAL-SU-SM

Fig. 17. Max-min throughput of full-duplex approximation algorithms for REAL-SU-SM and PI model normalized to that of OPT-FD-MTFS, and the lower bounds of the approximation ratios. Median, 5% and 95% percentiles are shown in the errorbars.

(b) NI and REAL-SU-SM

Fig. 19. Max-min throughput of half-duplex approximation algorithms normalized to that of OPT-FD-MTFS, and the lower bounds of the approximation ratios. Median, 5% and 95% percentiles are shown in the errorbars.
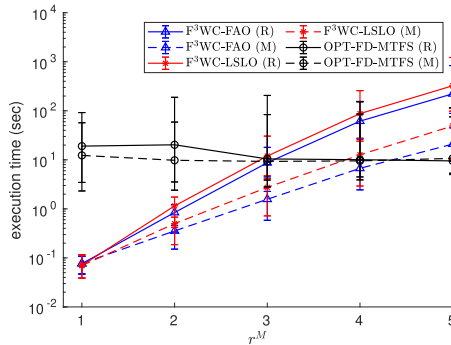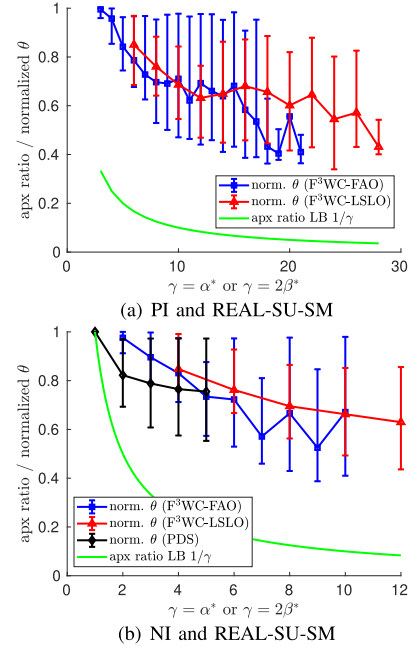

Fig. 18. Execution time comparison of approximation algorithms and OPT-FD-MTFS for full-duplex scheduling. Median, 5% and 95% percentiles are shown in the errorbars. (R) and (M) stands for REAL-SU-SM and MAX-SU-SM, respectively.

long time to solve. For future work it would be interesting to investigate how to shrink the conflict graph, in order to improve the runtime.

D. Half-Duplex Approximation Algorithms

F^{3}WC-FAO, F^{3}WC-LSLO and PDS are 3 approximation algorithms for half-duplex MTFS scheduling. The first two work for all cases while PDS only works for the NI model. We show in §IX-B that the optimal max-min throughput of the half-duplex MTFS problem is the same or very close to that of full-duplex MTFS for uniform orthogonal backhaul networks. Therefore, we use the max-min throughput of OPT-FD-MTFS as the reference for the evaluation of half-duplex approximation algorithms. Fig. 19(a) and 19(b) show the results for the PI and NI models assuming the REAL-SU-SM model. All three algorithms attain far better performance than the theoretical lower bounds. The two F^{3}WC algorithms have similar performance. Under the NI model, PDS has the best max-min throughput, being higher than 80% on average. The performance of PDS is even better for the MAX-SU-SM model. For example, it is guaranteed to reach the optimal when a backhaul network is uniform orthogonal.

Fig. 20 displays the time efficiency of the three approximation algorithms. They are all relatively efficient, requiring

the conflict graph $|V(C)|$ which is equal to the number of arcs in the expanded network H (see §VI-A). A F^{3}WC algorithm needs to solve a linear program of $|V(C)| + 1$ variables. For example, with $r^M = 5$ and the REAL-SU-SM model, the linear program has about 30,000 variables, which takes a
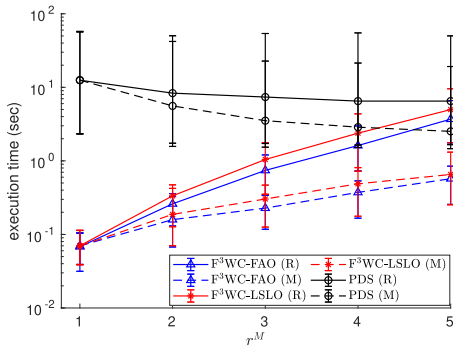
Fig. 20. Execution time of the approximation algorithms for half-duplex scheduling. Median, 5% and 95% percentiles are shown in the errorbars. (R) and (M) stands for REAL-SU-SM and MAX-SU-SM, respectively.

no more than two minutes. In comparison, F^3WC algorithms run faster because we use the property that a directed network can be sparsely expanded under the condition of half-duplex scheduling, which leads to a small conflict graph. We again observe the trend that the execution time of F^3WC goes up with r^M while that of PDS goes down. In addition, the execution time of the REAL-SU-SM model is larger than that of the MAX-SU-SM model. This is due to a larger conflict graph for F^3WC and an increase in time for maximum weight matching for PDS.

In summary, the evaluation shows that a mmWave backhaul network is generally noise-limited even for a relatively large beamwidth of $20°$. The optimal max-min throughput in practical backhaul networks is quite similar for both full-duplex and half-duplex scheduling. PDS is an ideal approximation algorithm for half-duplex scheduling under the NI model as it achieves near optimal performance within practical time. Finally, the two F^3WC algorithms have similar max-min throughput. They are competitive in execution time for small backhaul networks with a small number of RF chains and half-duplex scheduling.

X. CONCLUSION

In this article, we studied the scheduling of mmWave backhaul networks assuming a general system model of multiple macro BSs, relay BSs and RF chains as well as interference between links and realistic single-user spatial multiplexing. Under the assumption of full-duplex radios and interference-free links, we found an optimal joint routing and scheduling method—*schedule-oriented optimization* based on matching theory. It can solve any problem formulated as a linear program whose variables are data stream activation durations and QoS metrics. The method is demonstrated to be efficient in practice, capable of solving the maximum throughput fair scheduling (MTFS) problem within a few minutes for a backhaul network of 4 macro BSs, 100 relay BSs and 5 RF chains at each node. However, for the more realistic assumption of half-duplex radios or pairwise link interference, we proved that the MTFS problem is NP-hard. Subsequently, the paper proposed a number of approximation algorithms with provable performance bounds for the MTFS problem. The PDS algorithm works for half-duplex scheduling under the NI (no interference) model. It achieves the optimal performance for uniform orthogonal backhaul networks and about 80% of the optimum for general backhaul networks. The F^3WC algorithms adapted to our problem are more general than PDS

as they support any combination of full-duplex/half-duplex, REAL-SU-SM/MAX-SU-SM model and PI/NI model. Their performance is in general more than half of the optimum. In summary, the paper presents optimal and approximation algorithms that are highly practical for scheduling mmWave cellular networks.

REFERENCES

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

[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] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2196–2211, Oct. 2015.

[4] A. Ghosh *et al.*, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.

[5] C. Dehos, J. L. González, A. D. Domenico, D. Kténas, and L. Dussopt, "Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 88–95, Sep. 2014.

[6] *5G; NG-RAN; Architecture Description*, 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 3GPP, document 38.401, version 15.2.0, Jul. 2018.

[7] D. Yuan, H.-Y. Lin, J. Widmer, and M. Hollick, "Optimal joint routing and scheduling in millimeter-wave cellular networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 1205–1213.

[8] 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.

[9] Y. Zhu, Y. Niu, J. Li, D. O. Wu, Y. Li, and D. Jin, "QoS-aware scheduling for small cell millimeter wave mesh backhaul," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[10] W. Feng, Y. Li, D. Jin, L. Su, and S. Chen, "Millimetre-wave backhaul for 5G networks: Challenges and solutions," *Sensors*, vol. 16, no. 6, p. 892, Jun. 2016.

[11] Y. Li, J. Luo, W. Xu, N. Vucic, E. Pateromichelakis, and G. Caire, "A joint scheduling and resource allocation scheme for millimeter wave heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[12] Y. Niu *et al.*, "Energy-efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2674–2687, Mar. 2017.

[13] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 910–917, Sep. 1988.

[14] A. Sinha, L. Tassiulas, and E. Modiano, "Throughput-optimal broadcast in wireless networks with dynamic topology," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1203–1216, May 2019.

[15] P.-J. Wan, "Multiflows in multihop wireless networks," in *Proc. 10th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, 2009, pp. 85–94.

[16] L. Li, K. Josiam, and R. Taori, "Feasibility study on full-duplex wireless millimeter-wave systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2769–2773.

[17] T. Dinc, "Architectures, antennas and circuits for millimeter-wave wireless full-duplex applications," Ph.D. dissertation, Graduate School Arts Sci., Columbia Univ., New York, NY, USA, 2018.

[18] S. Sun, T. Rappaport, R. Heath, A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

[19] V. Raghavan *et al.*, "Millimeter-wave MIMO prototype: Measurements and experimental results," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 202–209, Jan. 2018.

[20] Q. Xue, X. Fang, and C.-X. Wang, "Beamspace SU-MIMO for future millimeter wave wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1564–1575, Jul. 2017.

[21] J. Kilpi, K. Seppanen, T. Suihko, J. Paananen, D. T. Chen, and P. Wainio, "Link scheduling for mmWave WMN backhaul," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

YUAN *et al.*: OPTIMAL AND APPROXIMATION ALGORITHMS FOR JOINT ROUTING AND SCHEDULING IN mmWAVE CELLULAR NETWORKS · · · · · · · · · · · 15

[22] E. Arribas *et al.*, "Optimizing mmWave wireless backhaul scheduling," *IEEE Trans. Mobile Comput.*, early access, Jun. 25, 2019, doi: 10.1109/TMC.2019.2924884.

[23] J. Du, E. Onaran, D. Chizhik, S. Venkatesan, and R. A. Valenzuela, "Gbps user rates using mmWave relayed backhaul with high-gain antennas," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1363–1372, Jun. 2017.

[24] *NR; Study on Integrated Access and Backhaul*, 3rd Generation Partnership Project (3GPP), 3GPP, document 38.874, release 16, Dec. 2018.

[25] M. Polese *et al.*, "Integrated access and backhaul in 5G mmWave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[26] C. Saha and H. S. Dhillon, "Millimeter wave integrated access and backhaul in 5G: Performance analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2669–2684, Dec. 2019.

[27] R. Zhao, T. Woodford, T. Wei, K. Qian, and X. Zhang, "M-cube: A millimeter-wave massive MIMO software radio," in *Proc. ACM MobiCom*, 2020, p. 15.

[28] H. Shokri-Ghadikolaei, C. Fischione, and E. Modiano, "On the accuracy of interference models in wireless communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[29] J. Tang, G. Xue, and W. Zhang, "Maximum throughput and fair bandwidth allocation in multi-channel wireless mesh networks," in *Proc. 25th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2006, pp. 1–10.

[30] L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *Proc. 21st Annu. Joint Conf. IEEE Comput. Commun. Societies*, Jun. 2002, pp. 763–772.

[31] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 5th ed. Berlin, Germany: Springer, 2012.

[32] L. G. Khachiyan, "Polynomial algorithms in linear programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 1, pp. 53–72, Jan. 1980.

[33] G. Dantzig, A. Orden, and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints," *Pacific J. Math.*, vol. 5, no. 2, pp. 183–195, Jun. 1955.

[34] H. N. Gabow, "An efficient reduction technique for degree-constrained subgraph and bidirected network flow problems," in *Proc. 15th Annu. ACM Symp. Theory Comput. (STOC)*, 1983, pp. 448–456.

[35] J. Edmonds, "Maximum matching and a polyhedron with 0, 1-vertices," *J. Res. Nat. Bureau Standards B*, vol. 69B, pp. 125–130, Jun. 1965.

[36] V. Kolmogorov, "Blossom V: A new implementation of a minimum cost perfect matching algorithm," *Math. Program. Comput.*, vol. 1, no. 1, pp. 43–67, Jul. 2009.

[37] W. T. Tutte, "A short proof of the factor theorem for finite graphs," *Can. J. Math.*, vol. 6, pp. 347–352, 1954.

[38] A. Schrijver, *Combinatorial Optimization–Polyhedra and Efficiency*. Berlin, Germany: Springer, 2003.

[39] M. Grötschel, L. Lovász, and A. Schrijver, "The ellipsoid method and its consequences in combinatorial optimization," *Combinatorica*, vol. 1, no. 2, pp. 169–197, Jun. 1981.

[40] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[41] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.

[42] Gurobi Optimization, LLC. (2019). *Gurobi Optimizer Reference Manual, Version to 9.0.1*. Accessed: Apr. 7, 2020. [Online]. Available: http://www.gurobi.com

Dingwen Yuan received the Bachelor of Engineering degree from the Department of Automation, Shanghai Jiao Tong University, in 2002, and the Master of Science and Ph.D. degrees from the Technische Universität Darmstadt in 2010 and 2016, respectively. He is currently a Post-Doctoral Researcher with the SEEMOO Lab, Technische Universität Darmstadt. His research interests include the quality-of-service guarantee in wireless sensor networks (WSNs) and the scheduling and optimization of millimeter-wave networks.

Hsuan-Yin Lin (Senior Member, IEEE) received the B.S. degree from National Tsing Hua University (NTHU), Taiwan, in 2007, and the M.S. and Ph.D. degrees from National Chiao Tung University (NCTU), Taiwan, in 2008 and 2013, respectively. He was a Visiting Scholar with Universitat Pompeu Fabra, Barcelona, Spain, and TU Darmstadt, Germany. He holds a post-doctoral position at Simula UiB, Norway, where he is currently a Research Scientist. His current interests include privacy-preserving technologies, coding in distributed storage systems, finite blocklength information theory, scheduling in millimeter-wave cellular networks, and wireless sensor networks. He is also a Guest Editor of *Entropy*: Special Issue on Information-Theoretic Approach to Privacy and Security.

Jörg Widmer (Fellow, IEEE) held positions at DOCOMO Euro-Labs, Munich, Germany, and EPFL, Switzerland. He is currently a Research Professor and the Research Director of the IMDEA Networks Institute, Madrid, Spain. He has authored more than 150 conference and journal articles and three IETF RFCs. He holds 13 patents. His research focuses on wireless networks, ranging from extremely high-frequency millimeter-wave communication and MAC layer design to mobile network architectures. He received the ERC Consolidator Grant, the Friedrich Wilhelm Bessel Research Award of the Alexander von Humboldt Foundation, the Mercator Fellowship of the German Research Foundation, the Spanish Ramon y Cajal Grant, and eight best paper awards. He is a Distinguished Member of the ACM.

Matthias Hollick (Member, IEEE) received the Ph.D. degree from the Technical University of Darmstadt (TU Darmstadt), Germany, in 2004. He has been researching and teaching at TU Darmstadt, the Universidad Carlos III de Madrid, and the University of Illinois at Urbana–Champaign. He is currently heading the Secure Mobile Networking Lab, Computer Science Department, TU Darmstadt. His research focus is on resilient, secure, privacy preserving, and quality-of-service-aware communication for mobile and wireless systems and networks.