System-Level Analysis of Full-Duplex Self-Backhauled Millimeter Wave Networks

Manan Gupta, Ian P. Roberts, and Jeffrey G. Andrews

Abstract

Integrated access and backhaul (IAB) facilitates cost-effective deployment of millimeter wave (mmWave) cellular networks through multihop self-backhauling. Full-duplex (FD) technology, particularly for mmWave systems, is a potential means to overcome latency and throughput challenges faced by IAB networks. We derive practical and tractable throughput and latency constraints using queueing theory and formulate a network utility maximization problem to evaluate both FD-IAB and half-duplex (HD)-IAB networks. We use this to characterize the network-level improvements seen when upgrading from conventional HD IAB nodes to FD ones by deriving closed-form expressions for (i) latency gain of FD-IAB over HD-IAB and (ii) the maximum number of hops that a HD- and FD-IAB network can support while satisfying latency and throughput targets. Extensive simulations illustrate that FD-IAB can facilitate reduced latency, higher throughput, deeper networks, and fairer service. Compared to HD-IAB, FD-IAB can improve throughput by $8\times$ and reduce latency by $4\times$ for a fourth-hop user. In fact, upgrading IAB nodes with FD capability can allow the network to support latency and throughput targets that its HD counterpart fundamentally cannot meet. The gains are more profound for users further from the donor and can be achieved even when residual self-interference is significantly above the noise floor.

I. INTRODUCTION

Noteworthy hurdles exist in the cost-effective deployment of millimeter wave (mmWave) cellular networks that can reliably supply users with high data rates and low latency—stemming largely from the severe pathloss and blockage vulnerability when communicating at such high carrier frequencies [1]–[4]. Integrated access and backhaul (IAB) is a promising means to deploy mmWave networks with the base station (BS) density necessary to deliver reliable, widespread

coverage [5]–[12]. IAB is a multihop network deployment where the majority of BSs—called *IAB nodes*—wirelessly backhaul their traffic to fiber-backhauled *donor nodes*, possibly relaying through other IAB nodes.

To meet the strict quality-of-service requirements for these modern applications, full-duplex (FD) technology provides an approach to address resource bottlenecks—potentially improving rate scaling and latency—and augments existing resource allocation solutions. Equipping IAB nodes with FD capability allows them to simultaneously transmit and receive over the same bandwidth, virtually doubling the available radio resources compared to conventional half-duplex (HD) operation. It is well known that this transceiver-level upgrade can directly translate to link-level gains, but the network-level gains are less clear. In this paper, we study the potential gains in network performance—in terms of throughput, latency, and network depth—when upgrading from HD IAB nodes to ones with FD capability, which may transcend the potential doubling of spectral efficiency.

A. Motivation, Background, and Related Work

Multihop networks have been an active area of research for a few decades [13]–[15], as they require fewer fiber connections to tessellate an area according to the given coverage criterion, for example, received signal-to-noise ratio (SNR) above a threshold. However, per-user throughput deteriorates and packet delays increase with the number of hops between the donor and a user-equipment (UE) [14], [16]. Even though IAB networks operating at mmWave frequencies benefit from larger bandwidth (offering high data rates on the backhaul) and reduced interference (from directional communication and raised integrated noise power), they are subject to the same fundamental *throughput-coverage* trade-off as conventional multihop networks. As the number of hops increase, the throughput and latency performance degrade due to packet relaying, buffering, and link multiplexing delays [6], [17].

As a result, to satisfy throughput and latency targets, efficient resource utilization at the IAB node is of utmost importance and has led to a variety of studies on route selection [9], [18], [19], link scheduling [8], [20], load balancing [21], and topology optimization [11], [22]. In [8], [9], authors propose reinforcement learning frameworks that aim to minimize end-to-end latency of packets, and [18] presents routing strategies that minimize the number of hops to improve throughput and latency. Link scheduling and power allocation solutions that leverage simulated-annealing are proposed in [20], the benefits of offloading UEs to IAB nodes is studied in [21],

and genetic algorithm based schemes for IAB node placement and non-IAB link distribution are developed in [22]. For a comprehensive survey on recent developments in IAB, please see [23].

In this work however, we investigate how FD capability can alleviate the aforementioned resource bottleneck and latency issues present in IAB networks. Recent breakthroughs in self-interference cancellation using analog [24], digital [25], and spatial [26], [27] cancellation techniques can rid a receive signal of self-interference. In [28], the authors prototype a two-hop IAB network and show that the throughput for FD-IAB is almost twice that of HD-IAB. In [25], the authors evaluate a multi-user FD-IAB network and present user selection and digital self-interference cancellation techniques to maximize the received signal power at the user. These works, however, do not explore the network-level consequences of mmWave FD in an IAB deployment. In this paper, we aim to quantify the relative gain of FD-IAB over HD-IAB in terms of the UE throughput, network depth, and latency in realistic multihop IAB deployments.

B. Contributions

Our technical contributions are summarized as follows.

A single optimization framework for HD-IAB and FD-IAB networks. We formulate an optimization problem to study *latency* and *throughput*—arguably the most important performance metrics—in a mmWave IAB network. In Section II, we model the IAB network as a Jackson network of queues, which allows us to leverage results from queueing theory to model packet delay as the sojourn time and throughput as the packet arrival rate. We use this queueing model to formulate a utility maximization with throughput and latency constraints that are both practical and tractable. Our network optimization problem is a convex program and is parameterized by the routing structure of the network, the link capacities, and scheduling restrictions on each BS, which in turn depend on whether the BS is capable of HD or FD communication. Solving the convex program returns the per-user throughput and the corresponding resource allocation scheme that maximizes a desired network utility.

Characterizing the network gain of FD-IAB over HD-IAB. In Section IV, we use the optimization framework to compare FD-IAB with HD-IAB and derive a closed-form expression for *latency gain*. We also derive the maximum number of hops that achieve a target latency and target throughput for each UE, for both FD-IAB and HD-IAB. In Section VI-C, we show numerical results for both *rate gain* and latency gain of FD-IAB over HD-IAB. The benefit of equipping the IAB nodes with FD transceivers is more subtle and powerful than the familiar

link-level gains (i.e., what is typically less than a two-fold improvement in spectral efficiency). We show that the more hops between a UE and the donor, the more the UE has to gain from a FD-IAB deployment, both in terms of latency and throughput. For example, the throughput of a UE four hops from the donor can improve by $8\times$ if IAB nodes are upgraded to FD. This many-fold increase stems from the increased scheduling opportunities that FD provides at each IAB node, which reduces multiplexing delays that are particularly significant in the context of multihop routing.

Impact of imperfect self-interference cancellation on network performance. In Section VI-C we present numerical results to quantify the effect of imperfect self-interference cancellation on the throughput performance of a FD-IAB network. We show that even if the residual self-interference is 10 dB above the noise floor, a six-fold rate improvement can be achieved by FD-IAB over its HD counterpart. Reducing the residual-interference-to-noise ratio (RINR) below 0 dB can result in 8 times rate gain for a user at the fourth hop. The FD gains saturate for RINR below -5 dB, suggesting that—from a network perspective—further self-interference cancellation is likely not worthwhile. These insights can drive physical layer design decisions for FD in IAB networks.

Notation: $\mathbf{A}_{:,i}$ denotes the i-th column of matrix $\mathbf{A}$, $\mathbf{A}_{i,:}$ denotes the i-th row of matrix $\mathbf{A}$, $\mathbf{A}^T$ denotes the transpose of a matrix, $[\mathbf{A}]_{i,j}$ denotes the (i,j)-th element of $\mathbf{A}$, and $\mathbf{A}^\dagger$ denotes the hermitian transpose of a matrix. The i-th element of vector $\mathbf{a}$ is denoted by $\mathbf{a}_i$. $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function over set $\mathcal{A}$, and $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$.

II. SYSTEM MODEL

Illustrated in Fig. 1, we consider a downlink IAB deployment with one fiber-backhauled donor node and K IAB nodes which wirelessly backhaul to the donor, possibly through multiple hops. M UEs are present in the network, each of which can either be served by the donor or an IAB node. Downlink data arrives at the donor from the network core and is delivered to the UEs, using the IAB nodes as relays as needed. In this work, we will consider and evaluate both HD- and FD-equipped IAB nodes. We assume UEs are conventional HD devices. We represent the flow of data—the route between the source (donor) and a destination (UE)—through a *routing tree*, $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of vertices (devices) such that $|\mathcal{V}| = K + M + 1$ and includes the mmWave BSs and the UEs. Henceforth, the term *device* will be used to refer to the donor node, an IAB node, or a UE. $\mathcal{E}$ denotes the set of edges in the tree $\mathcal{T}$ such that if
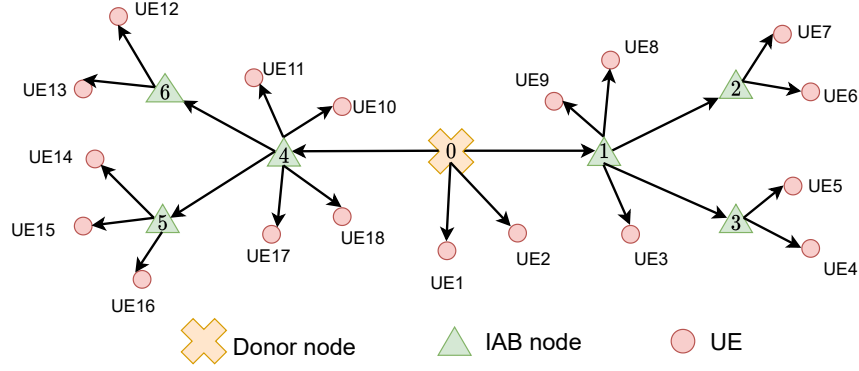
Fig. 1. A multihop routing tree with one donor node, $K = 6$ IAB nodes, and $M = 18$ UEs. Note that each IAB node has one parent but can serve multiple children.

$(u, v) \in \mathcal{E}$ then $u, v \in \mathcal{V}$ and there exists a directed edge with u as the parent and v as the child. In other words, a device v receives downlink data from a device u. Since all vertices, except for the donor, have one (and only one) parent edge, we have $|\mathcal{E}| = K + M$. Since an edge (u, v) is uniquely identified by its child v, we will index edge (u, v) as v. An illustration of an example IAB routing tree is shown in Fig. 1, where vertex 0 represents the donor and triangles denote IAB nodes. IAB1 (vertex 1) is one hop away from the donor whereas IAB2 (vertex 2) is two hops away since it communicates with the donor via IAB1. Similarly, UE1, UE3, and UE5 are one-hop, two-hop, and three-hop UEs, respectively.

Link Capacities and Scheduling: Associated to each edge $(u, v) \in \mathcal{E}$ is the capacity of the edge c_v based on its link quality; note that c_v is only indexed by v since each edge (u, v) is uniquely identified by its child (i.e., each device has only one parent). In an IAB network, simultaneously transmitting data on all edges in $\mathcal{T}$ is not possible due various hardware and design constraints, such as HD constraints and/or limited multi-user communication capabilities. A network scheduler typically decides the set of vertices that can communicate at a given time, forming the set of *active* edges. Thus, each edge is only allocated a certain fraction of the time by the scheduler to transmit data. We denote by μ_v the fraction of time allocated to edge (u, v) for data transmission, meaning $c_v \mu_v$ represents the effective long-term data rate of the edge (u, v).

Modeling the IAB Network as Network of Queues: In order to analyze the delay distribution of UEs across different hops, we model the IAB network as a network of queues. Data for each UE arrives in packets at the IAB donor following a stochastic process and must be delivered
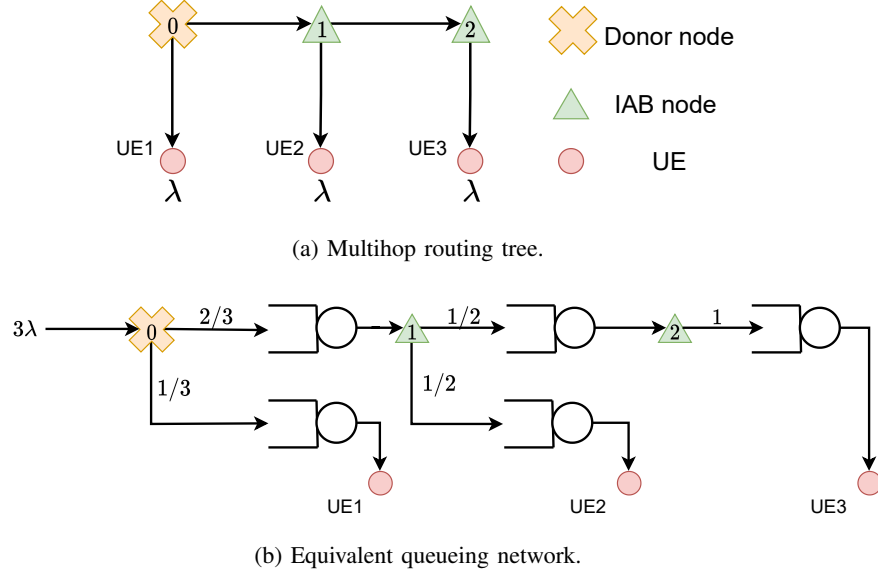
(a) Multihop routing tree.

(b) Equivalent queueing network.

Fig. 2. (a) A multihop routing tree with $K = 2$ IAB nodes and $M = 3$ UEs. The arrival rate of packets intended for each UE is λ. (b) The equivalent queueing network of (a) with traffic splitting at the donor and each IAB node. The total arrival rate into the network is 3λ.

to the destined UE along the route given by $\mathcal{T}$. We denote by $A_m(t)$ the number of packets destined for the m-th UE that arrives at the donor at time t, where $\lambda_m = \mathbb{E}[A_m(t)]$ denotes the mean of the arrival stochastic process for packets intended for the m-th UE (i.e., the *arrival rate*). We use $\boldsymbol{\lambda}$ to denote the $M \times 1$ vector of mean arrival rates λ_m for the M UEs in the network. Each edge $(u, v) \in \mathcal{T}$ maintains a queue to buffer packets that u must transmit to v and is referred to as queue (u, v).

Let $\mathbf{F}$ denote the $|\mathcal{E}| \times M$ routing matrix such that $[\mathbf{F}]_{l,m} = 1$ if traffic for the m-th UE is routed through the l-th edge and is zero otherwise. The arrival rate to each queue is given by the vector $\mathbf{F}\boldsymbol{\lambda}$. Since the effective long-term data rate of edge (u, v) is $c_v \mu_v$, the mean service time for queue (u, v) is given by $1/(c_v \mu_v)$. Modeling each edge as a queue equivalently implies that the donor and IAB nodes maintain queues to buffer packets for each of their children. A packet arriving at BS k is placed in queue (k, v) with probability $(\mathbf{F}\boldsymbol{\lambda})_v / (\mathbf{F}\boldsymbol{\lambda})_k$, as shown in Fig. 2b. This ensures that the average number of packets delivered to each UE is the same as the number of packets arriving at the donor destined for that UE.

III. FORMULATING AN IAB NETWORK DESIGN PROBLEM

With an IAB network model in place, we will now formulate a network utility maximization subject to practical delay and throughput constraints. Our goal is to use this optimization problem

to study the network performance improvement—in terms of throughput and latency—when the HD transceivers at the IAB nodes are upgraded to FD ones, for a given deployment. Our optimization problem is parameterized by the network topology and the choice of either HD- or FD-equipped IAB nodes. Specifically, we use a routing matrix $\mathbf{F}$ to describe the routes between the donor and the UEs and is readily obtained for a given routing tree. We use a capacity matrix $\mathbf{C}$ and a scheduling matrix $\mathbf{G}$ to describe the effective long-term rate of each link, both of which depend on the choice of HD or FD IAB nodes. Feasibility constraints enforce that each link meets its demand. A latency constraint ensures that a fraction of packets η are delivered to their target UE within δ units of time. We aim to find the average arrival rates $\boldsymbol{\lambda}$ and resource allocations $\boldsymbol{\mu}$ that maximize a chosen network utility. We outline these components of our design in detail as follows and subsequently assemble our network design optimization problem.

Constraint 1: Fundamental constraints on arrival rate and resource fractions. As follows, we capture the fundamental properties of the arrival rates $\boldsymbol{\lambda}$, each of which must be non-negative, and resource allocations $\boldsymbol{\mu}$, each of which must lie in the interval $[0, 1]$.

$$\boldsymbol{\lambda} \geq \mathbf{0}, \quad \mathbf{0} \leq \boldsymbol{\mu} \leq \mathbf{1} \tag{1}$$

Constraint 2: Constraint on feasible arrival rate. The signal-to-interference-plus-noise ratio (SINR) of the signal at device v when receiving from its parent u is given by SINR_v, which depends on a variety of environmental and system factors such as pathloss and transmit/receive beamforming. The capacity of the edge (u, v) is

$$c_v = W \cdot \log_2(1 + \mathrm{SINR}_v) \tag{2}$$

where W is the system bandwidth. Note that FD IAB nodes incur self-interference, making SINR the metric of interest. HD IAB nodes are effectively interference-free under noise-limited network conditions, meaning we can simply use the SNR.

Often, mmWave communication leads to noise-limited conditions due to highly directional communication, severe pathloss at mmWave frequencies, and high susceptibility to blockages [2], [29]. The severity of network interference in FD-IAB network is still an open question, but for tractability we assume it is negligible, largely thanks to the aforementioned properties of mmWave communication [4], [30]. Furthermore, the donor and other IAB nodes would likely be pole-mounted and presumably deployed to avoid directed interference on the backhaul transmissions, while UEs are likely on the ground. Together, these establish sufficient spatial isolation between

transmitting and receiving devices that happen to be nearby and scheduled simultaneously. Also, since the UEs are HD, backhaul transmissions would not suffer interference from arbitrary directions, and as observed in [31], a competent scheduling algorithm further reinforces the noise-limited behavior since it avoids scheduling interfering links in the same slot. Our simulations also suggest noise-limited behavior for both HD- and FD-IAB deployments.

We denote by $\mathbf{C}$ the $|\mathcal{E}| \times |\mathcal{E}|$ diagonal matrix for edge capacities c_v. Then, the product $\mathbf{C}\boldsymbol{\mu}$ is the vector containing the long-term average rate of each edge or, equivalently, the vector of mean service rates of each queue. For the queues to be stable, the mean service rate of each queue should be greater than its mean traffic arrival rate, leading to

$$\mathbf{C}\boldsymbol{\mu} > \mathbf{F}\boldsymbol{\lambda}. \tag{3}$$

Constraint 3: Scheduling constraints. We denote by $\mathbf{G}$ the $(K+1) \times |\mathcal{E}|$ scheduling matrix for the network, which defines the resource constraints on each BS (donor node or IAB node). More precisely, $[\mathbf{G}]_{k,v} = 1$ if BS k must allocate orthogonal time resources to edge (k, v), and zero otherwise. For example, for the routing tree in Fig. 2a the HD scheduling matrix $\mathbf{G}_{\mathrm{HD}}$ and the FD scheduling matrix $\mathbf{G}_{\mathrm{FD}}$ are given in (4). For both matrices, the second row represents IAB1. In the HD case, it must split resources between receiving from the donor, transmitting to UE1, and transmitting to IAB2. On the other hand, when IAB1 is FD-capable, it can receive while transmitting, and must no longer dedicate resources to edge $(0, 1)$.

$$\mathbf{G}_{\mathrm{HD}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{G}_{\mathrm{FD}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Constraint 4: Probabilistic latency constraint. If the arrival processes $A_m(t)$ are Poisson and the queues have exponential service times, then the queueing network in Fig. 2b is a *Jackson Network* [32, Section 9.9.1] without feedback loops, and the steady-state joint distribution of the queueing network is product-form. Both the Poisson arrival process and the exponentially varying packet size [33] are well-accepted models. As a result, we can treat each queue (u, v) as an independent $M/M/1$ queue with input arrival rate $(\mathbf{F}\boldsymbol{\lambda})_v$ and mean service rate $c_v \mu_v$ (the effective capacity from device u to device v). Consequently, if D_v denotes the random variable representing the delay experienced by a typical packet over queue (u, v), then D_v follows an exponential distribution with the following cumulative density function (CDF).

$$\mathbb{P}[D_v \leq d] = 1 - \exp\left(-(c_v \mu_v - (\mathbf{F}\boldsymbol{\lambda})_v) \cdot d\right) \tag{5}$$

Let $\mathsf{R}(s,d) = ((s,u_1),(u_1,u_2),\dots,(u_k,d))$ denote the route between source-destination pair (s,d). A route is defined as the sequence of edges in $\mathcal{T}$ traversed by a packet to go from s to d. For brevity, we will denote by $\mathsf{R}(m)$ the route between the donor and UE m. Let D_m be defined as the total delay experienced by a packet destined for the m-th UE, which is simply the sum of delays the packet incurred along its route.

$$D_m = \sum_{(u,v)\in\mathsf{R}(m)} D_v \tag{6}$$

Since the per-hop delays $\{D_v\}$ are independent exponential random variables with different means, the sum delay D_m follows a hypoexponential distribution which is non-convex and intractable. Thus, for our analysis we will use a stricter notion of delay. Let $\tilde{D}_m$ denote the maximum delay experienced on any hop in its route to UE m.

$$\tilde{D}_m = \max_{(u,v)\in\mathsf{R}(m)} D_v \tag{7}$$

It trivially holds that $D_m \leq h_m \tilde{D}_m$, where $h_m = \sum_{(u,v)} \mathbb{1}_{\{(u,v)\in\mathsf{R}(m)\}}$ denotes the number of hops between the donor and UE m. The CDF of $\tilde{D}_m$ is given by

$$\mathbb{P}\left[h_m\tilde{D}_m \leq d\right] = \prod_{(u,v)\in\mathsf{R}(s,d)} \left(1 - e^{-(c_v\mu_v - (\mathbf{F}\boldsymbol{\lambda})_v)d/h_m}\right). \tag{8}$$

If we desire that a fraction η of packets are delivered to their target UE within δ units of time, we can form a probabilistic latency constraint as $\mathbb{P}[h_m\tilde{D}_m \leq \delta] > \eta$, or equivalently $\log(\mathbb{P}[h_m\tilde{D}_m \leq \delta]) > \log(\eta)$. Mathetically, we have

$$\sum_{(u,v)\in\mathsf{R}(m)} \log\left(1 - e^{-(c_{u,v}\mu_{u,v} - (\mathbf{F}\boldsymbol{\lambda})_v)\delta/h_m}\right) \geq \log(\eta), \quad m = 1, 2, \dots, M. \tag{9}$$

Note that δ is the parameter which reflects the delay threshold of the network, whereas η sets how strict it is that the threshold is met.

Objective: Maximize network utility. As is widely used by network designers, the objective of our design is to maximize sum utility of the network. We denote by $U(\cdot)$ a non-decreasing and concave utility function. There is a considerable body of literature exploring different applications of various network utility functions such as network-wide proportional fairness [34], network-wide max-min fairness [7], network-wide logarithmic utility for balanced load distribution [35], and α-optimal user association [36]. The particular choice of $U(\cdot)$ is at the liberty of the network designer and design requirements. In our simulation we will use the logarithmic utility function,

$U(\cdot) = \log(\cdot)$, which maximizes the product of rates and achieves a healthy balance between network sum-rate and fairness, and is a common choice for network design and evaluation [32].

Optimization problem. With our constraints and objective mathematically defined, the network design problem can be formulated as follows, which aims to find the arrival rates $\boldsymbol{\lambda}$ and resource allocations $\boldsymbol{\mu}$ that maximize the sum utility of the network subject to our constraints.

$$\underset{\boldsymbol{\lambda},\boldsymbol{\mu}}{\text{maximize}} \sum_{m=1}^{M} U(\lambda_m) \tag{10a}$$

$$\text{subject to } \boldsymbol{\lambda} \geq \mathbf{0}, \ \mathbf{0} \leq \boldsymbol{\mu} \leq \mathbf{1} \tag{10b}$$

$$\mathbf{C}\boldsymbol{\mu} - \mathbf{F}\boldsymbol{\lambda} > \mathbf{0} \tag{10c}$$

$$\mathbf{G}\boldsymbol{\mu} \leq \mathbf{1} \tag{10d}$$

$$\sum_{(u,v)\in\mathsf{R}(m)} \log\left(1 - e^{-(c_{u,v}\mu_{u,v}-(\mathbf{F}\boldsymbol{\lambda})_v)\delta/h_m}\right) \geq \log(\eta), \quad m = 1, 2, \ldots, M \tag{10e}$$

The objective (10a) is to maximize the network utility with respect to $\boldsymbol{\lambda}$. Constraints (10b) are the fundamental properties of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, and constraint (10c) ensures that the queues are stable and do not accumulate to infinity. Along with the quality of the wireless environment, the capacity matrix $\mathbf{C}$ depends on whether the IAB nodes are HD- or FD-capable. Constraint (10d) represents the scheduling constraint on each BS, and the scheduling matrix $\mathbf{G}$ depends on if the IAB nodes are upgraded from HD to FD. Finally, (10e) represents the probabilistic latency constraint.

Remark 1. Having FD capability at the IAB nodes relaxes the scheduling constraints (10d) on the IAB nodes, and they can receive data from their parent while simultaneously transmitting data to one of their children. This introduces a variety of potential network gains. First, a relaxed scheduling constraint provides more scheduling opportunities to the backhaul links. This expands the support of the network, also termed its *throughput region*, and as a consequence, an FD-IAB network can support higher arrival rates per UE, denoted by the elements of $\boldsymbol{\lambda}$, while still satisfying (10e). Second, an IAB network with FD IAB nodes can support tighter delay constraints. In other words, relaxing the constraint (10d) by upgrading to FD IAB nodes, an IAB network can achieve latency targets δ that may have been infeasible for the equivalent HD-IAB network. Moreover, since delay can grow arbitrarily as a network's operating point approaches the boundary of its throughput region, a seemingly minor expansion in the throughput region can translate to a considerable reduction in latency. Third, with the expansion of the throughput region and the feasible values of δ, FD-IAB can meet throughput targets with *deeper* networks

and support more hops. This is advantageous for operators as they can then provide comparable quality-of-service with fewer fiber-connected donor nodes, reducing infrastructure costs.

IV. MINIMUM FEASIBLE DELAY THRESHOLD AND LATENCY GAIN

The optimization problem in (10) provides a framework to study these network trade-offs and meet system requirements. We now seek analytical expressions and insights using the optimization framework in Section III. Specifically, we aim to answer the following questions. *(i) What is the minimum delay threshold δ^* that an IAB network parameterized by $\mathbf{F}$, $\mathbf{G}$, $\mathbf{C}$ can support? (ii) What is the gain in δ^* for FD-IAB over HD-IAB? (iii) What is the maximum number of hops feasible for an IAB network and how does it vary with δ^*? (iv) How does δ^* vary with the minimum per-UE throughput requirement $\lambda_{\min}$?* With these questions in mind, we formulate and solve a linear program to find the minimum delay threshold, subject to the constraints from (10).

A. Minimum Feasible Delay

Latency is a key performance metric in modern networks. 5G cellular networks, for example, are designed to support end-to-end packet delays on the order of milliseconds for mission critical and tactile internet applications [37]. In (10), the delay threshold of a network is captured by δ and constraint (10e). For a network designer, it is useful to know the minimum delay a network can support and how it relates network parameters, such as its routing matrix $\mathbf{F}$, link capacities $\mathbf{C}$, and scheduling matrix $\mathbf{G}$. Mathematically, this is equivalent to finding the minimum δ such that problem (10) is feasible.

Minimizing δ without any constraints on $\boldsymbol{\lambda}$, however, could lead to solutions that are practically undesirable where UEs receive zero throughput. We address this by introducing a new constraint

$$\boldsymbol{\lambda} \geq \lambda_{\min}\mathbf{1} \tag{11}$$

to ensure a minimum average arrival rate $\lambda_{\min}$ is met at each UE, which can be tuned by network engineers. While one can solve for the minimum δ using numerical solvers, the delay constraint (10e) is non-convex when optimizing over δ, $\boldsymbol{\lambda}$, and $\boldsymbol{\mu}$, and hence intractable. This motivates us to substitute (10e) with a tighter constraint by replacing $\mathbb{P}[\tilde{D}_m h_m \leq \delta] \geq \eta$ with

$\mathbb{P}[D_v h_m \leq \delta] \geq \eta, \forall (u,v) \in \mathsf{R}(m)$. Combining all of this leads to the formulation of problem (12), which includes minimizing over δ and incorporates modified constraints (12b) and (12e).

$$\underset{\delta,\boldsymbol{\mu},\boldsymbol{\lambda}}{\text{minimize}} \; \delta \tag{12a}$$

$$\text{subject to } \boldsymbol{\lambda} \geq \lambda_{\min}\mathbf{1}, \; \mathbf{0} \leq \boldsymbol{\mu} \leq \mathbf{1}, \delta \geq 0 \tag{12b}$$

$$\mathbf{C}\boldsymbol{\mu} - \mathbf{F}\boldsymbol{\lambda} > \mathbf{0} \tag{12c}$$

$$\mathbf{G}\boldsymbol{\mu} \leq \mathbf{1} \tag{12d}$$

$$1 - e^{-(c_v \mu_v - (\mathbf{F}\boldsymbol{\lambda})_v)\delta/h_m} \geq \eta \; \forall (u,v) \in \mathsf{R}(m), \; \forall m = 1, 2, \ldots, M \tag{12e}$$

The new latency constraint (12e) ensures that the per-hop delay is less than δ/h_m, which is a stricter requirement than (10e), and h_m is as defined in (8). Rearranging the terms in (12e) and applying the change of variable $t = -\log(1-\eta)/\delta$, problem (12) can be reformulated as the following linear program.

$$\underset{t,\boldsymbol{\lambda},\boldsymbol{\mu}}{\text{maximize}} \; t \tag{13a}$$

$$\text{subject to } \boldsymbol{\lambda} \geq \lambda_{\min}\mathbf{1}, \; \mathbf{0} \leq \boldsymbol{\mu} \leq \mathbf{1}, \; t \geq 0 \tag{13b}$$

$$\mathbf{G}\boldsymbol{\mu} \leq \mathbf{1} \tag{13c}$$

$$c_v \mu_v - (\mathbf{F}\boldsymbol{\lambda})_v \geq t h_m, \quad \forall (u,v) \in \mathsf{R}(m), \; \forall m = 1, 2, \ldots, M \tag{13d}$$

Lemma 1. *The optimal arrival rate vector for* (13) *is* $\boldsymbol{\lambda}^* = \lambda_{\min}\mathbf{1}$.

Proof. Let (u', v') be the bottleneck edge such that $c_{v'}\mu_{v'}^* - (\mathbf{F}\boldsymbol{\lambda}^*)_{v'} = t^* h_{m'}$, for some UE m' and $\lambda_{m'}^* > \lambda_{\min}$, where t^* and $\boldsymbol{\mu}^*$ denote the corresponding optimal points. Since, $c_{v'}\mu_{v'}^* - \lambda_{\min}(\mathbf{F}\mathbf{1})_{v'} > c_{v'}\mu_{v'}^* - (\mathbf{F}\boldsymbol{\lambda}^*)_{v'}$, $\lambda_{\min}\mathbf{1}$ relaxes the bottleneck constraint and achieves a higher objective, which contradicts the optimality of t^*. $\square$

Remark 2. Note that the $\boldsymbol{\lambda}^* = \lambda_{\min}\mathbf{1}$ is not a unique solution. For the non-bottleneck (u,v) and m such that $c_v \mu_v^* - \lambda_{\min}(\mathbf{F}\mathbf{1})_v > t^* h_m$, any arrival rate λ_m which satisfies

$$\lambda_{\min} < \lambda_m < \frac{c_v \mu_v^* - t^* h_m}{(\mathbf{F}\mathbf{1})_v}$$

will not change the optimal t, and hence is also optimal.

Theorem 1. *Given the routing matrix* $\mathbf{F}$*, scheduling matrix* $\mathbf{G}$*, and capacity matrix* $\mathbf{C}$*, the optimal solution for* (13) *is given by*

$$t^* = \min_{k=0,1,\ldots,K} \frac{1 - \lambda_{\min}\mathbf{G}_{k,:}\mathbf{C}^{-1}\mathbf{F1}}{\mathbf{G}_{k,:}\mathbf{C}^{-1}\tilde{\mathbf{h}}}, \tag{14}$$

where $\mathbf{G}_{k,:}$ *is the k-th row of* $\mathbf{G}$ *and* $\tilde{\mathbf{h}}$ *is a* $|\mathcal{E}| \times 1$ *vector such that* $\tilde{h}_v = \max\limits_{m:(u,v)\in\mathsf{R}(m)} h_m$*. If* t^* *is less than zero, then* (13) *is infeasible and the IAB network parameterized by* $\mathbf{F}, \mathbf{G}, \mathbf{C}$ *cannot support a per-UE arrival rate of* $\lambda_{\min}$*.*

Proof. Using Lemma 1 and substituting $\boldsymbol{\mu}$ from (13d) in (13c), we get

$$\mathbf{G}_{k,:}\mathbf{C}^{-1}(t\tilde{\mathbf{h}} + \lambda_{\min}\mathbf{F1}) \leq 1, \quad \forall k = 0, 1, \ldots, K$$

$$t \leq \frac{1 - \lambda_{\min}\mathbf{G}_{k,:}\mathbf{C}^{-1}\mathbf{F1}}{\mathbf{G}_{k,:}\mathbf{C}^{-1}\tilde{\mathbf{h}}}, \quad \forall k = 0, 1, \ldots, K$$

Hence, the optimal t^* achieves the tightest inequality and is given by (14). $\qquad\square$

Remark 3. The minimum feasible delay for the IAB network is given by $\delta^* = -\log(1 - \eta)/t^*$.

B. Latency Gain

Given Theorem 1, we can compute δ^*_{HD} and δ^*_{FD} for HD and FD deployments, respectively. The *latency gain* ℓ of FD-IAB over HD-IAB can then be expressed as

$$\ell = \frac{\delta^*_{\mathrm{HD}}}{\delta^*_{\mathrm{FD}}} = \frac{t^*_{\mathrm{FD}}}{t^*_{\mathrm{HD}}} = \frac{\min\limits_{k=0,1,\cdots,K} \frac{1-\lambda_{\min}(\mathbf{G}_{\mathrm{FD}})_{k,:}\mathbf{C}^{-1}_{\mathrm{FD}}\mathbf{F1}}{(\mathbf{G}_{\mathrm{FD}})_{k,:}\mathbf{C}^{-1}_{\mathrm{FD}}\tilde{\mathbf{h}}}}{\min\limits_{k=0,1,\cdots,K} \frac{1-\lambda_{\min}(\mathbf{G}_{\mathrm{HD}})_{k,:}\mathbf{C}^{-1}_{\mathrm{HD}}\mathbf{F1}}{(\mathbf{G}_{\mathrm{HD}})_{k,:}\mathbf{C}^{-1}_{\mathrm{HD}}\tilde{\mathbf{h}}}}. \tag{15}$$

Here, $\mathbf{G}_{\mathrm{HD}}$ and $\mathbf{G}_{\mathrm{FD}}$ denote the scheduling matrices for the HD and FD deployments, respectively, and $\mathbf{C}_{\mathrm{HD}}$ and $\mathbf{C}_{\mathrm{FD}}$ denote the corresponding the capacity matrices. Depending on the quality of self-interference cancellation and the structure of the IAB routing tree $\mathcal{T}$, the bottleneck BS—which achieves the minimum in (14)—could be different for FD and HD deployments. Simplified expressions can be obtained if one considers the structure of the routing matrix $\mathbf{F}$ or particular cases of $\mathbf{C}$.

To illustrate a simple example, we refer to the IAB routing tree given in Fig. 2a to have

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix}^T \tag{16}$$
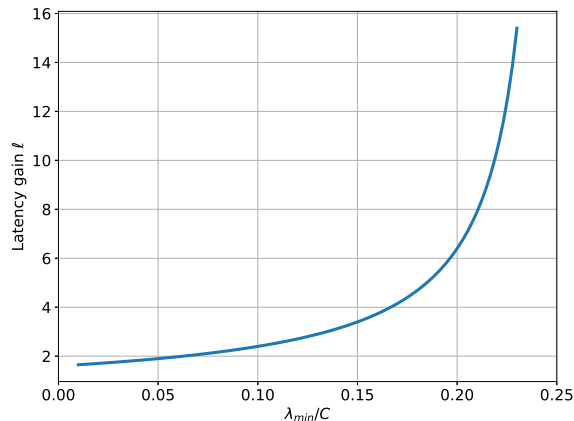
Fig. 3. Latency gain for a three-hop line network where all links have capacity C and there exists perfect self-interference cancellation at FD IAB nodes. FD-IAB can support minimum per-user arrival rates infeasible for HD-IAB and arbitrarily large latency gains can be observed.

and $\tilde{\mathbf{h}} = [1, 3, 2, 3, 3]^T$ (defined in (14)). The HD and FD scheduling matrices are as given in (4), respectively. If we assume for this example that all edges have equal capacity C and $\mathbf{C}_{\mathrm{HD}} = \mathbf{C}_{\mathrm{FD}} = C\mathbf{I}$, then using Theorem 1 we have

$$t_{\mathrm{HD}}^* = \frac{C - 4\lambda_{\min}}{8} \tag{17}$$

$$t_{\mathrm{FD}}^* = \begin{cases} \dfrac{C - 2\lambda_{\min}}{5}, & \lambda_{\min} \leq C/7 \\ \dfrac{C - 3\lambda_{\min}}{4}, & \text{else} \end{cases} \tag{18}$$

and the latency gain is given by $\ell = t_{\mathrm{FD}}^*/t_{\mathrm{HD}}^*$, which is plotted in Fig. 3. This simple but insightful example illustrates that IAB with FD nodes can support $\ell = 2.5$ times tighter latency constraints when the minimum arrival rate to each UE $\lambda_{\min}$ is 10% of the link capacity C. When $\lambda_{\min}$ is just over a fifth of C, the latency gain is $\ell = 12$ (nearly quintuples) and increases substantially for small increases in $\lambda_{\min}$ thereafter. This behavior is explained by the simple fact that FD-IAB can support higher values of $\lambda_{\min}$ which are infeasible for HD-IAB, driving the latency gain ℓ toward infinity.

V. ANALYTICAL RESULTS: LATENCY AND RATE OF A LINE NETWORK

In the interest of providing simple, intuitive, and tractable expressions that may serve as guidelines for network design, we will present a detailed analysis of a simple but practically relevant IAB deployment: the *line network*. A routing tree $\mathcal{T}$ represents a line network if no

two BSs in $\mathcal{T}$ share a parent (e.g., the IAB network shown in Fig. 2a). We also assume that $\mathbf{C}_{\mathrm{HD}} = \mathbf{C}_{\mathrm{FD}} = \mathbf{C}$, meaning FD IAB nodes are capable of perfect self-interference cancellation. In Section VI, we will show that there exists a self-interference cancellation threshold which, if achieved, is effectively as good as perfect cancellation from a network perspective. We assume that all backhaul edges have capacity R_b and all access edges have capacity R_a. We also assume that $R_b > R_a$ because BSs typically have more antennas and benefit from a higher array gain compared to UEs, along with the simple fact that IAB networks are typically designed in this fashion this since they are fundamentally backhaul-limited.

A. Latency Gain

In (15), we derived latency gain ℓ for a *general* IAB routing tree. In the following, we will derive a closed-form expression for (14) under both HD and FD deployments of a line network, along with the corresponding latency gain ℓ.

Theorem 2. *Let $\mathcal{T}$ represent a line network with K IAB nodes and one donor such that each BS supports w UEs, where all backhaul edges have capacity R_b and all access edges have capacity R_a. Then, for a feasible $\lambda_{\min}$ and under perfect self-interference cancellation, we have*

$$
t_{\mathrm{HD}}^*(\lambda_{\min}, K) = \begin{cases} \dfrac{1 - w\lambda_{\min}\left(\frac{3}{R_b} + \frac{1}{R_a}\right)}{\frac{2(K+1)}{R_b} + \frac{Kw}{R_a}}, & \lambda_{\min} \leq \dfrac{R_a}{4(K+1)\left(\frac{R_a}{R_b}\right)^2 + (2K+3)w\left(\frac{R_a}{R_b}\right) + w} \\[3em] \dfrac{1 - w\lambda_{\min}\left(\frac{2K-1}{R_b} + \frac{1}{R_a}\right)}{\frac{2(K+1)}{R_b} + \frac{2w}{R_a}}, & \text{else} \end{cases}
$$

$$(19)$$

$$
t_{\mathrm{FD}}^*(\lambda_{\min}, K) = \begin{cases} \dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_b} + \frac{1}{R_a}\right)}{\frac{K+1}{R_b} + \frac{Kw}{R_a}}, & \lambda_{\min} \leq \dfrac{R_a}{(K+1)\left(\frac{R_a}{R_b}\right)^2 + (K+1)w\left(\frac{R_a}{R_b}\right) + w} \\[3em] \dfrac{1 - w\lambda_{\min}\left(\frac{K}{R_b} + \frac{1}{R_a}\right)}{\frac{K+1}{R_b} + \frac{w}{R_a}}, & \text{else} \end{cases}.
$$

$$(20)$$

Proof. See Appendix A. □

Note that, since $\mathcal{T}$ represents a line network, $K + 1$ is also the number of hops between the donor and the furthest UE. It is interesting to note that the bottleneck BS is either the first or

the last IAB node. This can be attributed to the function $f(\cdot)$ (see Appendix A) and its property that it is either non-increasing or non-decreasing depending on whether $\lambda_{\min}$ is above or below a threshold value given in Theorem 2. On close inspection, we see that f is the ratio of the fraction of time the BS is idle and the time it takes for the BS to serve one unit of the total arriving traffic (i.e., $\lambda_{\min}\mathbf{F1}$) under the delay constraints. The idle time is an increasing function of k and a decreasing function of $\lambda_{\min}$, whereas $\lambda_{\min}\mathbf{F1}$ is an increasing function of k. This explains the peculiar property of f. One can also interpret f as the product of the service rate of the BS and the fraction of time it is idle—in some sense it is the effective service rate. The inverse of which is the average delay experienced by the arriving flow at the BS. The objective of (13) (or equivalently (12)) is to find the *bottleneck* BS with the maximum service time within the resource allocation constraints. The delay threshold δ an IAB network can support is limited by this bottleneck BS.

Corollary 1. *For the network in Theorem 2, the latency gain is given by*

$$
\ell = \begin{cases}
\dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_b} + \frac{1}{R_a}\right)}{1 - w\lambda_{\min}\left(\frac{3}{R_b} + \frac{1}{R_a}\right)} \dfrac{\frac{2(K+1)}{R_b} + \frac{Kw}{R_a}}{\frac{K+1}{R_b} + \frac{Kw}{R_a}}, & \lambda_{\min} \leq \Lambda_1 \\[4ex]
\dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_b} + \frac{1}{R_a}\right)}{1 - w\lambda_{\min}\left(\frac{2K-1}{R_b} + \frac{1}{R_a}\right)} \dfrac{\frac{2(K+1)}{R_b} + \frac{2w}{R_a}}{\frac{K+1}{R_b} + \frac{Kw}{R_a}}, & \Lambda_1 < \lambda_{\min} \leq \Lambda_2 \\[4ex]
2\dfrac{1 - w\lambda_{\min}\left(\frac{K}{R_b} + \frac{1}{R_a}\right)}{1 - w\lambda_{\min}\left(\frac{2K-1}{R_b} + \frac{1}{R_a}\right)}, & \text{else}
\end{cases} \tag{21}
$$

where

$$
\Lambda_1 = \frac{R_a}{4(K+1)\left(\frac{R_a}{R_b}\right)^2 + (2K+3)w\left(\frac{R_a}{R_b}\right) + w} \qquad \Lambda_2 = \frac{R_a}{(K+1)\left(\frac{R_a}{R_b}\right)^2 + (K+1)w\left(\frac{R_a}{R_b}\right) + w}.
$$

Proof. This is found directly by substituting t^* from Theorem 2 into $\delta^* = -\log(1-\eta)/t^*$. $\square$

B. Maximum Network Depth

In this section, we present analytical expressions for the maximum depth $K^{\max}$ of a line network that can support given latency and throughput targets. Note that the result of Theorem 2

can also be written as

$$t_{\mathrm{HD}}^*(\lambda_{\min}, K) = \begin{cases} \dfrac{1 - w\lambda_{\min}\left(\frac{3}{R_b} + \frac{1}{R_a}\right)}{\frac{2(K+1)}{R_b} + \frac{Kw}{R_a}}, & K \leq \dfrac{\frac{R_a}{\lambda_{\min}} - w - \frac{R_a}{R_b}\left(4\frac{R_a}{R_b} + 3w\right)}{2\frac{R_a}{R_b}\left(2\frac{R_a}{R_b} + w\right)} \\[4ex] \dfrac{1 - w\lambda_{\min}\left(\frac{2K-1}{R_b} + \frac{1}{R_a}\right)}{\frac{2(K+1)}{R_b} + \frac{2w}{R_a}}, & \text{else} \end{cases} \tag{22}$$

$$t_{\mathrm{FD}}^*(\lambda_{\min}, K) = \begin{cases} \dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_b} + \frac{1}{R_a}\right)}{\frac{K+1}{R_b} + \frac{Kw}{R_a}}, & K \leq \dfrac{\frac{R_a}{\lambda_{\min}} - w}{\frac{R_a}{R_b}\left(\frac{R_a}{R_b} + w\right)} - 1 \\[4ex] \dfrac{1 - w\lambda_{\min}\left(\frac{K}{R_b} + \frac{1}{R_a}\right)}{\frac{K+1}{R_b} + \frac{w}{R_a}}, & \text{else.} \end{cases} \tag{23}$$

Let $K_{\mathrm{HD}}^{\max}$ and $K_{\mathrm{FD}}^{\max}$ represent the maximum number of IAB nodes for the HD and FD line networks, respectively, that can guarantee a target delay threshold δ_{target} and minimum arrival rate $\lambda_{\min}$. For notational convenience, let κ_{HD} and κ_{FD} denote the break points from (22) and (23) as follows.

$$\kappa_{\mathrm{HD}} = \dfrac{\frac{R_a}{\lambda_{\min}} - w - \frac{R_a}{R_b}\left(4\frac{R_a}{R_b} + 3w\right)}{2\frac{R_a}{R_b}\left(2\frac{R_a}{R_b} + w\right)} \qquad \kappa_{\mathrm{FD}} = \dfrac{\frac{R_a}{\lambda_{\min}} - w}{\frac{R_a}{R_b}\left(\frac{R_a}{R_b} + w\right)} - 1 \tag{24}$$

Theorem 3. *Let $\mathcal{T}$ represent a line network where all backhaul edges have capacity R_b, all access edges have capacity R_a, and each BS supports w UEs. Suppose $K_{\mathrm{HD}}^{\max} + 1$ and $K_{\mathrm{FD}}^{\max} + 1$ are the maximum number of hops that can still meet a target delay threshold δ_{target} and a minimum arrival rate $\lambda_{\min}$ for FD-IAB and HD-IAB deployments, respectively. Then,*

$$K_{\mathrm{HD}}^{\max} = \begin{cases} \left\lfloor \dfrac{1 - w\lambda_{\min}\left(\frac{3}{R_b} + \frac{1}{R_a}\right) - \frac{2\zeta}{R_b}}{2\zeta\left(\frac{2}{R_b} + \frac{w}{R_a}\right)} \right\rfloor, & \zeta \leq t_{\mathrm{HD}}^*(\lambda_{\min}, \kappa_{\mathrm{HD}}) \\[4ex] \left\lfloor \dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_a} - \frac{1}{R_b}\right) - 2\zeta\left(\frac{1}{R_b} + \frac{w}{R_a}\right)}{\frac{2\zeta}{R_b} + \frac{2w\lambda_{\min}}{R_b}} \right\rfloor, & \text{else} \end{cases} \tag{25}$$

$$K_{\mathrm{FD}}^{\max} = \begin{cases} \left\lfloor \dfrac{1 - w\lambda_{\min}\left(\frac{1}{R_b} + \frac{1}{R_a}\right) - \frac{\zeta}{R_b}}{\zeta\left(\frac{1}{R_b} + \frac{w}{R_a}\right)} \right\rfloor, & \zeta \leq t_{\mathrm{FD}}^*(\lambda_{\min}, \kappa_{\mathrm{FD}}) \\[4ex] \left\lfloor \dfrac{1 - w\lambda_{\min}\frac{1}{R_a} - \zeta\left(\frac{1}{R_b} + \frac{w}{R_a}\right)}{\frac{\zeta}{R_b} + \frac{w\lambda_{\min}}{R_b}} \right\rfloor, & \text{else} \end{cases} \tag{26}$$

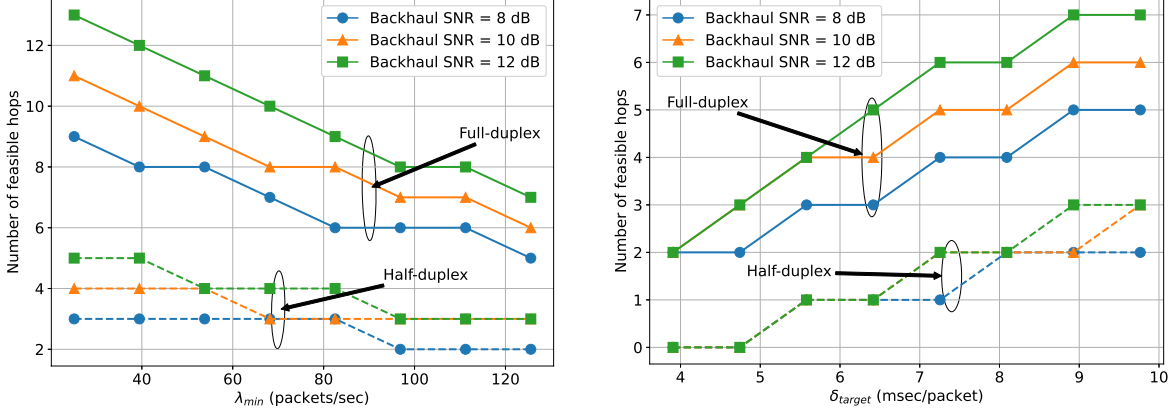*where $\zeta = -\log(1 - \eta)/\delta_{\mathrm{target}}$.*

(a) As a function of $\lambda_{\min}$ for $\delta_{\text{target}} = 10$ msec/packet.

(b) As a function of δ_{target} for $\lambda_{\min} = 125$ packets/sec.

Fig. 4. The number of feasible hops $K^{\max}$ for FD-IAB and HD-IAB as a function of (a) minimum service rate $\lambda_{\min}$ and (b) δ_{target} for various backhaul SNRs. FD-IAB can support more hops while meeting the throughput and latency constraints and the marginal return of increasing backhaul SNR is more compared to HD-IAB.

Proof. Note that t_{HD}^* and t_{FD}^* are continuous and non-increasing in $\lambda_{\min}$ and K. If $\delta^* \leq \delta_{\text{target}}$ then $t^* \geq \zeta$. Choosing the appropriate branch of the piece-wise defined t^* in (22) and (23), and solving for K by setting $t^* \geq \zeta$, we get the above results. $\square$

In Fig. 4a we present the effect of $\lambda_{\min}$ on maximum feasible network depth $K^{\max} + 1$ as computed in Theorem 3, for different backhaul SNR and access SNR of 5 dB. Theorem 3 assumes perfect self-interference cancellation and hence R_b and R_a are computed by applying the Shannon formula to SNR. We evaluate (26) and (25) for a packet size of 10 KB, bandwidth $W = 100$ MHz, and $w = 5$ UEs per BS. Solid lines denote $K_{\text{FD}}^{\max} + 1$ and dashed lines denote $K_{\text{HD}}^{\max} + 1$. The plot shows that FD-IAB can increase maximum feasible network depth by about three times for smaller throughput targets, compared to its HD counterpart. The plot also shows that a 2 dB increment in the backhaul SNR typically translates to an increment of about two hops in $K_{\text{FD}}^{\max}$. The same increment in backhaul SNR typically results in an increment of one hop in $K_{\text{HD}}^{\max}$, and sometimes no increment at all. In other words, the marginal return of improving the backhaul link quality is more for FD-IAB. Fig. 4b shows the variation of $K^{\max}$ with the latency target δ_{target} for fixed $\lambda_{\min}$ and different backhaul SNR. The plot shows similar trends as Fig. 4a and the marginal return of increasing backhaul SNR is more profound. Theorem 3 can also be used to derive a closed-form expression for *hop gain* by following steps similar to Corollory 1.

VI. SIMULATION RESULTS

In addition to the analytical results presented in the previous section, we also evaluate FD-IAB network performance—in both line and more general multihop networks—through Monte Carlo simulation and compare it against its HD counterpart.

A. Channel Model and Signal Propagation

We model mmWave channels using the Saleh-Valenzuela-based model [3]. Let N_{tx} and N_{rx} denote the number of antennas at the transmitter and receiver respectively. The $N_{\mathrm{rx}} \times N_{\mathrm{tx}}$ channel matrix $\mathbf{H}$ can be written as

$$\mathbf{H} = \sqrt{\frac{1}{N_{\mathrm{ray}} N_{\mathrm{cluster}}}} \sum_{i=1}^{N_{\mathrm{cluster}}} \sum_{j=1}^{N_{\mathrm{ray}}} h_{i,j} \mathbf{a}_{\mathrm{rx}}(\mathrm{AoA}_{i,j}) \mathbf{a}_{\mathrm{tx}}(\mathrm{AoD}_{i,j})^{\dagger} \tag{27}$$

which is simply the composition of propagating N_{cluster} clusters, each having N_{ray} rays. We denote by $h_{i,j} \sim \mathcal{CN}(0,1)$ the complex gain of the j-th ray in the i-th cluster. $\mathrm{AoA}_{i,j}$ is the angle of arrival of the ray at the receiver, and $\mathrm{AoD}_{i,j}$ is the angle of departure at the transmitter. The scalar in front of the summation handles power normalization. The vectors $\mathbf{a}_{\mathrm{tx}}(\cdot)$ and $\mathbf{a}_{\mathrm{rx}}(\cdot)$ represent the array response vectors at the transmitter and the receiver, respectively.

We assume that the network is noise-limited and interference from neighboring BSs is negligible. Hence, received signals are corrupted by additive noise and—at FD IAB nodes—by residual self-interference remaining after cancellation. Due to lack of measurements and characterization of the mmWave self-interference channel, we do not assume a particular model for the self-interference channel and our contribution does not rely on specific characteristics of a model. Instead, we evaluate our system based directly on RINR, which captures the degree of *residual* self-interference plaguing a desired receive signal after analog, digital, and/or spatial cancellation. The SINR at a FD IAB node v in such a case is expressed as

$$\mathrm{SINR}_v = \frac{\mathrm{SNR}_v}{\mathrm{RINR}_v + 1} \tag{28}$$

where RINR_v is the RINR at the receiver due to self-interference and SNR_v is the SNR the equivalent HD IAB node observes, written as

$$\mathrm{SNR}_v = \frac{P_{\mathrm{tx}} L_v \left| \mathbf{w}_{\mathrm{rx}}^{\dagger} \mathbf{H}_v \mathbf{f}_{\mathrm{tx}} \right|^2}{\sigma^2 \cdot W}. \tag{29}$$

Here, P_{tx} denotes the transmit power, L_v is the inverse pathloss, $\mathbf{H}_v$ is the channel matrix, $\mathbf{f}_{\mathrm{tx}}$, and $\mathbf{w}_{\mathrm{rx}}$ are the transmit and receive beamforming vectors, W is the system bandwidth, σ^2 is the noise power spectral density. The SNR at each UE can be expressed analogously.
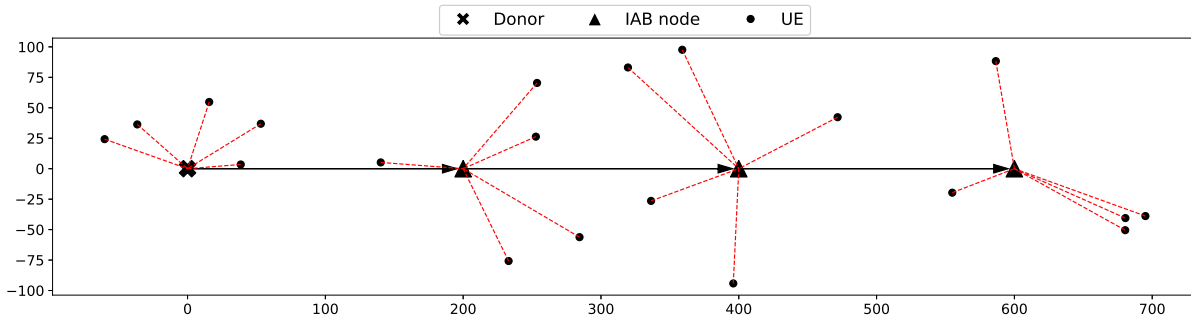
Fig. 5. A realization of the simulated line network with depth $K + 1 = 4$, having one donor and three IAB nodes. Five UEs are randomly dropped around the donor and each IAB node.

B. Network Topology and Channel Parameters

Using the proposed optimization framework, we evaluate the gain of FD-IAB over HD-IAB for various depths and topologies. We begin with line networks with depths $K + 1 = 2$, $K + 1 = 3$, and $K + 1 = 4$, and then evaluate a more general *two-child tree* network where the donor and each first-hop IAB node supports two child BSs. The IAB nodes are separated by 200 m, the BSs transmit at 30 dBm, and each BS is equipped with a uniform linear array (ULA) with $N_{BS} = 64$ antennas. Each UE has a 16-element ULA. Each BS serves five UEs which are dropped uniformly in a 100 m radius around the BS. A sample drop for one donor and three IAB nodes—a network depth of $K + 1 = 4$—line network is shown in Fig. 5 and the two-child tree network is illustrated in Fig. 11. The channels between pairs of devices are generated using (27), and **f** and **w** are drawn from a discrete fourier transform (DFT) codebook to maximize SNR (i.e., codebook-based beam alignment). For each Monte Carlo iteration, we drop a new set of UEs and generate channels such that N_{ray} is drawn uniformly from the interval $[1, 10]$ and $N_{cluster}$ is drawn from the interval $[1, 6]$ [38]. The system operates at a carrier frequency of 30 GHz with 100 MHz bandwidth. For the pathloss, we use the model adopted in the 3GPP standard [39, page 26] for urban environment (UMa), which incorporates the effects of blockage, the multi-slope nature of the pathloss, and the elevation difference between the BSs and the users, through parameters that have been fit to real-world data. The additive noise power spectral density is -174 dBm/Hz plus a 10 dB noise figure.

Packets of size 10 KB arrive at the donor according to a Poisson process like the file transfer protocol (FTP) model 3 [33]. We divide the capacity of each edge by the mean packet size and

(a) At last-hop for various network depths.

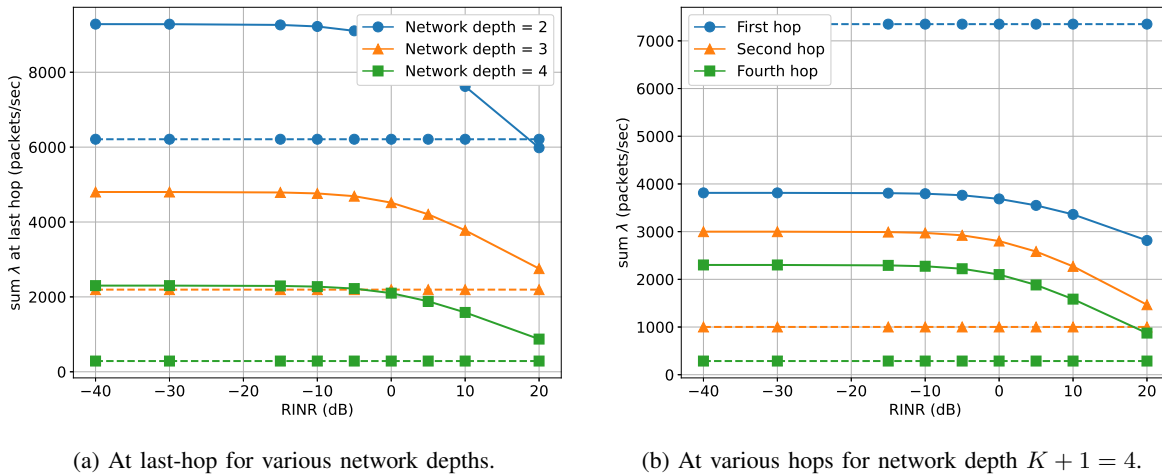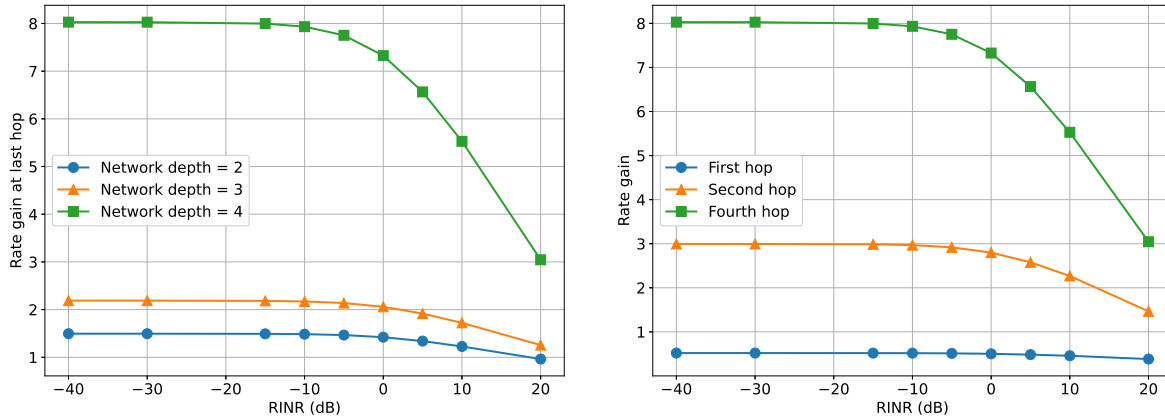(b) At various hops for network depth $K + 1 = 4$.

Fig. 6. Sum-rate as a function of RINR (a) at last-hop for various network depths and (b) at various hops in a network having depth $K + 1 = 4$, where the delay threshold is $\delta = 3.5$ msec/packet. Solid lines represent FD-IAB and dashed lines represent HD-IAB. Supporting practical data rates for users at the third and fourth hops is feasible with FD-IAB.

normalize it to packets/sec, leading to units of delay in sec/packet. Note that these results can be transformed to any packet size by appropriate scaling. We use CVXPY [40] to solve our network design for $\eta = 0.9$ (i.e., 90% of the packets must be delivered to the destination UE within the target delay threshold δ).

C. Rate and Latency Gain in Simulated Line Networks

Using (10), we will now discuss the effect of deploying FD-IAB on network throughput, under the logarithmic utility function $U(\cdot) = \log(\cdot)$. As discussed earlier, the logarithmic utility achieves a healthy balance between network sum-rate and fairness, unlike pure sum-rate maximization which would result in greedy solutions that are unfair and practically undesirable, such as allocating all the resources to a first-hop UE with the best channel.

Rate gain with imperfect self-interference cancellation. For various depths of the described line network, we solve (10) for both HD and FD modes to obtain the optimal arrival rate vectors $\boldsymbol{\lambda}_{\mathrm{HD}}$ and $\boldsymbol{\lambda}_{\mathrm{FD}}$. Comparing FD to HD, the rate gain for UE m is defined as $(\boldsymbol{\lambda}_{\mathrm{FD}})_m / (\boldsymbol{\lambda}_{\mathrm{HD}})_m$. Let $\mathcal{M}_i$ denote the set of UEs i hops from the donor. Then, the sum-rate at i-th hop is defined as $\sum_{\mathcal{M}_i} \boldsymbol{\lambda}_m$ and the rate gain is defined as $\frac{\sum_{\mathcal{M}_i} (\boldsymbol{\lambda}_{\mathrm{FD}})_m}{\sum_{\mathcal{M}_i} (\boldsymbol{\lambda}_{\mathrm{HD}})_m}$. Fig. 6a shows the last-hop sum-rate and Fig. 7a shows the rate gain at the last hop versus RINR for various network depths. Naturally, as RINR increases, sum-rate and consequently rate gain suffer due to increased residual self-interference plaguing FD operation, diminishing the resource gains of FD over HD. Even with

(a) At last-hop for various network depths.

(b) At various hops for network depth $K + 1 = 4$.

Fig. 7. From Fig. 6, the rate gain as a function of RINR (a) at last-hop for various network depths and (b) at various hops in a network having depth $K + 1 = 4$, where the delay threshold is $\delta = 3.5$ msec/packet. As RINR is decreased, the network observes diminishing gains before saturating. Appreciable gains can be seen for RINR well above 0 dB, where self-interference is as strong as noise.

self-interference that is ten times stronger than noise (i.e., $\mathrm{RINR} = 10$ dB), however, appreciable gains over HD are visible, especially for deeper networks. Reducing RINR below -5 dB does not yield meaningful rate improvements, as sum-rate saturates, which can drive physical layer FD design decisions from the perspective of a network operator.

Deeper networks benefit more from FD deployment. In Fig. 7a and Fig. 6a, we observe that an IAB network with more hops has more to gain from FD deployment. At low RINR, the last-hop gain for a network with depth two is about 1.5 whereas for a four-hop network the last-hop sum-rate improves from 300 packets/sec to more than 2000 packets/sec, an eight-fold rate improvement—far beyond the familiar potential doubling of capacity with FD. Fig. 6b and Fig. 7b present the sum-rate and rate gain across different hops for a four-hop network. With FD, users at all hops throughout the network enjoy a healthy rate, unlike in HD-IAB, where only first-hop users see a high rate (even having used a logarithmic utility). UEs deeper in the network see higher rate gains with FD, since their rates under HD-IAB are so poor.

Note that the rate gain for the first hop UE is about 0.5. This is because of the shape of the constraint set in (10) for HD-IAB. With the HD restriction, and even with the logarithmic utility, the first-hop UEs are greedily served because the third- and fourth-hop UEs can meet the delay constraint with very small arrival rates. However, a more fair operating point can be reached with FD-IAB. Thus, even though it seems to be sacrificing performance at the first-hop, in reality
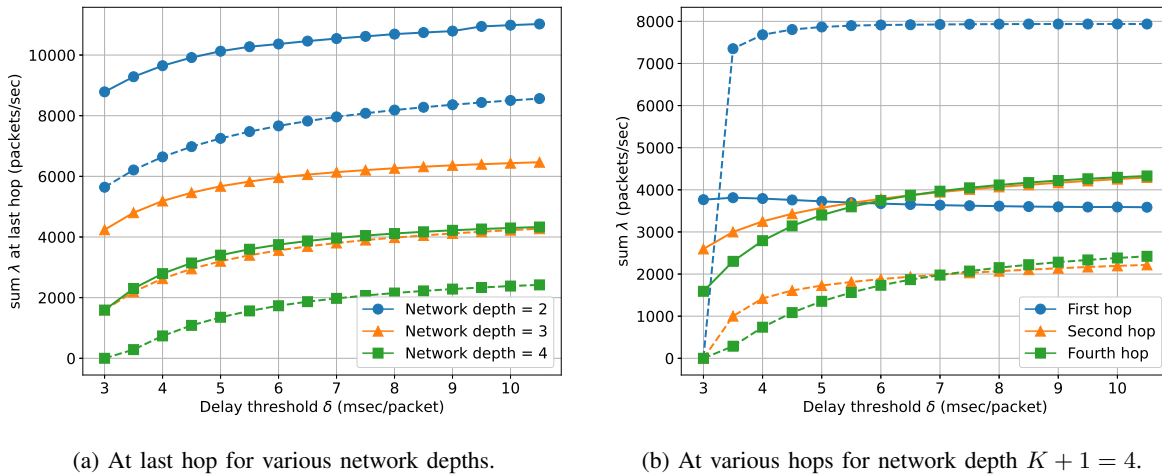
(a) At last hop for various network depths.

(b) At various hops for network depth $K + 1 = 4$.

Fig. 8. (a) Sum-rate for a last-hop user as a function of delay threshold δ. (b) Sum-rate for users at various hops as a function of delay threshold δ. Solid lines represent FD-IAB and dashed lines represent the HD-IAB, and for FD we consider $\mathrm{RINR} = -15$ dB. Supporting practical arrivals rates at third and fourth hops for strict delay thresholds is feasible with FD-IAB.

(a) At the last hop for various network depths.

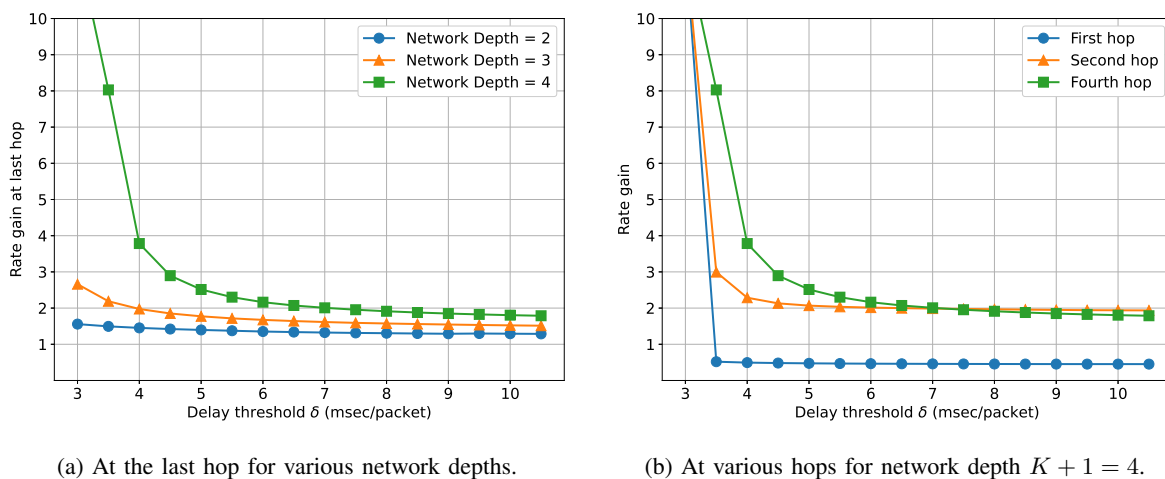(b) At various hops for network depth $K + 1 = 4$.

Fig. 9. (a) From Fig. 8a, the rate gain for a last-hop user as a function of delay threshold δ. (b) From Fig. 8b, the rate gain for users at various hops as a function of delay threshold δ. As stricter delay thresholds are enforced, FD-IAB offers increasing rate gain at the last hop over HD-IAB; very strict delay thresholds that cannot be met by HD-IAB can be met by FD-IAB.

FD-IAB improves the overall network performance, measured by the logarithmic utility.

FD-IAB allows for tighter delay constraints. In our previous results, we fixed a delay threshold δ in examining rate gain. Now we evaluate rate gain for changes in delay threshold δ. In Fig. 8a and Fig. 9a, we plot last-hop sum-rate and last-hop rate gain as a function of delay threshold δ for various network depths. In doing so, we consider an $\mathrm{RINR} = -15$ dB in light of the diminishing gains at low RINR, with the understanding that increased residual self-interference would degrade the backhaul capacities and affect the rate gain as in Fig. 7.
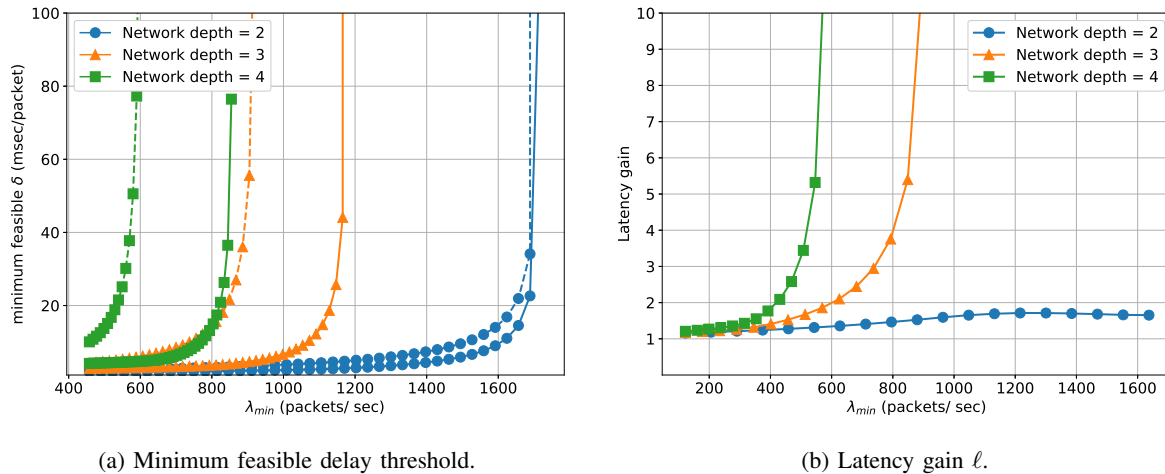
(a) Minimum feasible delay threshold.

(b) Latency gain ℓ.

Fig. 10. (a) Minimum feasible delay δ^* of FD-IAB (solid) and HD-IAB (dashed) as a function of minimum rate requirement $\lambda_{\min}$ for various network depths, where $\mathrm{RINR} = -15$ dB. (b) From (a), the corresponding latency gain ℓ of FD-IAB over HD-IAB as a function of minimum rate requirement $\lambda_{\min}$ for various network depths.

The rate improvement offered by FD-IAB is seen across network depths, with deeper networks experiencing a higher rate gain as before. By alleviating the multiplexing delay at the IAB nodes, FD-IAB improves the arrival rates at stricter delay constraints. In fact, delay constraints which were infeasible for HD-IAB can be made feasible by upgrading to FD. Note that in Fig. 9a, the gain at the fourth hop for $\delta = 3$ msec/packet is infinite, because the corresponding HD-IAB deployment is infeasible (rate of 0 packets/sec in Fig. 8a). The rate gain saturates with increasing δ, because for large delay thresholds the per-user arrival rate is limited by the backhaul capacities and not the multiplexing delay at the IAB node.

Fig. 8b and Fig. 9b show the sum-rate and rate gain across different hops for a four-hop network. The observations are consistent with the story so far: UEs further from the donor have more to benefit from FD-IAB. At very small δ, the gain tends to infinity since a delay constraint that cannot be met by HD-IAB can in fact be met with FD. Like before, we see FD-IAB is capable of delivering fairer service to UEs throughout the network, while still meeting the delay constraints across multiple hops.

FD-IAB supports higher per-UE rates. To study improvements in per-UE rate with delay constraint, we solve the linear program (13) for both FD-IAB and HD-IAB modes and employ our definition of latency gain (15). We perform simulations fixing $\mathrm{RINR} = -15$ dB and examine the $(\lambda_{\min}, \delta^*)$-pairs that can be met by HD and those that can be met by FD. Fig. 10b illustrates these results for various network depths, where curves offering higher $\lambda_{\min}$ at lower δ^* are
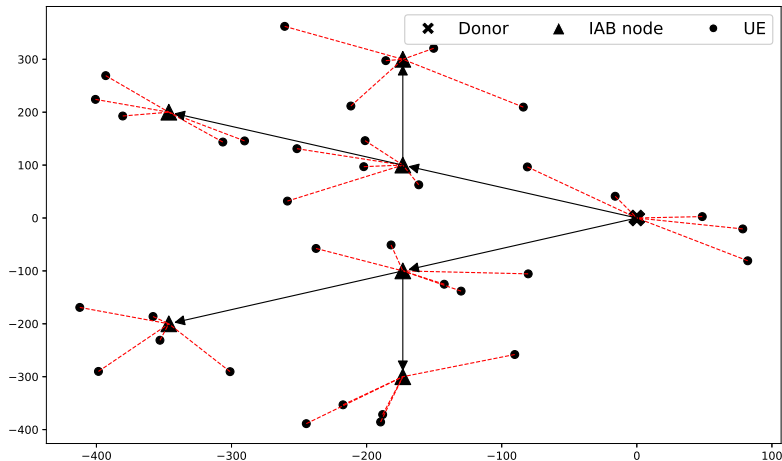
Fig. 11. A realization of the simulated two-child tree network with one donor and six IAB nodes. The donor and each IAB serves five UEs. Note that first-hop IAB nodes deliver backhaul to child IAB nodes, along with serving access.

more desirable (i.e., toward the lower right). For a two-hop network, both the HD and the FD deployments handle the latency and throughput constraints equally well. With more than two hops, however, the power of FD in this network setting is brought to light. For a given latency target, FD-IAB can deliver a minimum rate $\lambda_{\min}$ far greater than HD-IAB: for $\delta^* = 20$ msec/packet, FD offers over 1.5 times higher $\lambda_{\min}$ in a four-hop network and over 1.3 times higher $\lambda_{\min}$ in a three-hop network. Minimum rate targets $\lambda_{\min}$ that can be met by HD-IAB networks only with prohibitively high delays, can also be met by FD-IAB while simultaneously offering low latency. For instance, in a three-hop network, HD-IAB can deliver $\lambda_{\min} = 900$ packets/sec only be introducing delays well beyond what is practical; a FD-IAB network can do so while meeting a latency of approximately $\delta^* = 5$ msec/packet. As a result, latency gain can tend toward infinity.

D. Beyond Line Networks

While line networks are expected to be common for initial IAB deployments, the presented framework is in no way limited to them. To briefly illustrate this, we use our framework to study the IAB deployment in Fig. 11, which consists of one donor node, $K = 6$ IAB nodes, and five UEs per BS. We refer to this as a *two-child tree network* since the donor and each first-hop IAB node deliver backhaul to two IAB nodes. While not explicitly shown due to space constraints, the key takeaways from the discussion on line networks holds, namely: (i) rate gain saturates
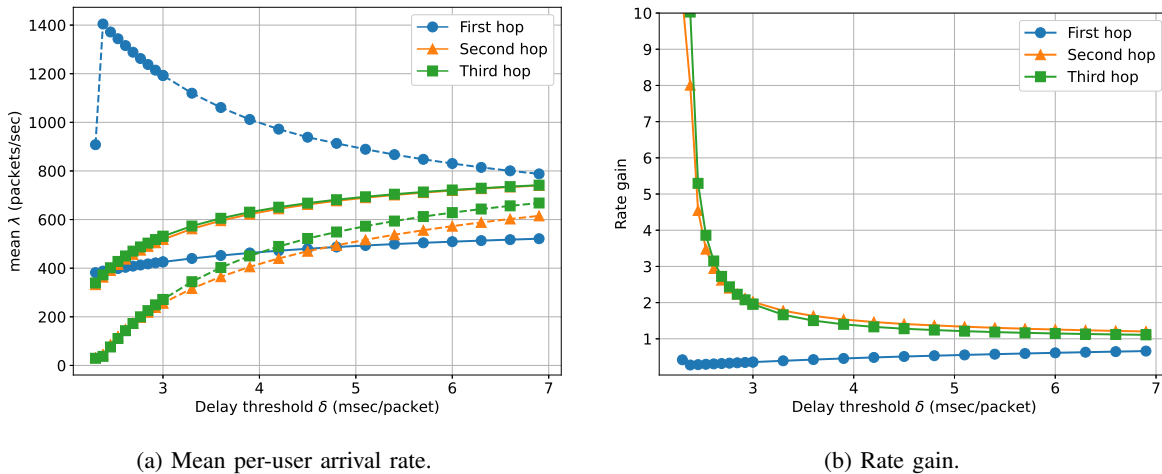
(a) Mean per-user arrival rate.

(b) Rate gain.

Fig. 12. (a) Mean per-UE rate of FD-IAB (solid) and HD-IAB (dashed), and (b) Rate gain of FD-IAB over HD-IAB, as a function of delay threshold δ at various hops in the two-child network in Fig. 11, where $\mathrm{RINR} = -15$ dB. FD-IAB alleviates the resource bottleneck due to branching at the donor and achieves a fairer rate distribution.

for RINR below around -5 dB, (ii) UEs further from the donor have more to benefit from an FD-IAB deployment, and (iii) FD-IAB can support larger $\lambda_{\min}$ for the same minimum feasible delay δ^* as was shown in Fig. 10.

Nonetheless, there are some additional insights to draw from this two-child tree network. To start, we consider Fig. 12a, which shows the mean per-user arrival rate at various hops in the network as a function delay threshold. Unlike line networks, the donor and first-hop IAB nodes now deliver backhaul to more than one IAB node. Consequently, a HD-IAB deployment cannot support strict latency requirements at the second- and third-hop with a practically viable arrival rate. This is because first-hop IAB nodes must multiplex incoming and outgoing backhaul links, significantly worsening the multiplexing delay and the radio resource bottleneck. First-hop users, however, can enjoy high rates—even under strict delay thresholds—since they are not subjected to these relaying delays. The disparity in user service is evident in Fig. 12a as the delay threshold is made stricter, with first-hop arrival rate increasing as users deeper in the network see their service degrading.

With FD-IAB, however, the resource bottleneck is alleviated and the network can serve its multihop users much more fairly and with satisfactory rates. We can see from Fig. 12a that FD-IAB can support delay thresholds infeasible for its HD counterpart, with a healthy throughput across hops. The corresponding rate gain of FD-IAB over HD-IAB is shown in Fig. 12b. FD-IAB can meet very strict delay thresholds for second- and third-hop users when HD-IAB cannot,

driving the rate gain toward infinity as δ is decreased. As δ increases, the rate gain saturates like in Fig. 9b. For first-hop users, we observe a rate gain less than one, which may seem undesirable at first glance but like Fig. 7b is justified by FD-IAB achieving a fairer distribution of arrival rates across hops.

Comparing these results of a two-child tree network to that of the simulated line network, we notice a few key differences. Comparing Fig. 12b to Fig. 9b, we see that the rate gain of the second- and third-hop users in the two-child tree network saturates to noticeably less than 2, while that for the line network saturates to nearly 2. This is difficult to explain precisely, though it is likely largely due to the fact that the donor and each first-hop IAB node must multiplex outgoing backhaul transmissions—which line networks do not suffer from. FD cannot directly alleviate these multiplexing costs, suggesting that the network may be bottlenecked by the resources consumed when juggling multiple backhual links. This motivates future work to more thoroughly study this disparity across network topologies and investigate how to best deploy IAB networks when IAB nodes are FD-equipped.

VII. CONCLUSION

We present a framework for analyzing the throughput and latency of IAB networks, which can be readily used by network engineers to evaluate and optimize IAB deployments. We use this framework to study the merits of equipping IAB nodes with FD capability, as a means to alleviate resource bottlenecks traditionally faced by IAB networks. Analytical and numerical results illustrate the network performance improvements of FD-IAB over HD-IAB and show that the former offers a multitude of benefits. FD-IAB can facilitate lower latency, higher throughput, deeper networks, and fairer service compared to conventional HD-IAB. FD upgrades can be particularly transformative for deeper networks, where users at or near the last hop—which may suffer in HD-IAB deployments—can enjoy improvements in rate and latency that dramatically enhance their quality-of-service. Furthermore, FD-IAB widens the feasible operating region of the network, allowing it to support latency and rate targets that HD-IAB fundamentally cannot meet. We show that these gains can be observed in the presence of residual self-interference that is near or even above the noise floor. These results motivate the use of FD to alleviate the obstacles in latency and rate scaling suffered by traditional multihop IAB networks.

APPENDIX A

PROOF OF THEOREM 2

Proof. Note that, $[\mathbf{G}_{\mathrm{HD}}]_{k,l} = 1$ if vertex l is the parent of k where as $[\mathbf{G}_{\mathrm{FD}}]_{k,l} = 0$. We define

$$f_{\mathrm{HD}}(k) = \frac{1 - \lambda_{\min}(\mathbf{G}_{\mathrm{HD}})_{k,:}\mathbf{C}^{-1}\mathbf{F1}}{(\mathbf{G}_{\mathrm{HD}})_{k,:}\mathbf{C}^{-1}\tilde{\mathbf{h}}} \quad f_{\mathrm{FD}}(k) = \frac{1 - \lambda_{\min}(\mathbf{G}_{\mathrm{FD}})_{k,:}\mathbf{C}^{-1}\mathbf{F1}}{(\mathbf{G}_{\mathrm{FD}})_{k,:}\mathbf{C}^{-1}\tilde{\mathbf{h}}}$$

The half-duplex case. Since we know the structure of the scheduling matrices, we have

$$f_{\mathrm{HD}}(k) = \begin{cases} \frac{1 - w\lambda_{\min}\left(\frac{1}{R_a} + \frac{K}{R_b}\right)}{\frac{K+1}{R_b} + \frac{w}{R_a}}, & k = 0 \\ \frac{1 - w\lambda_{\min}\left(\frac{1}{R_a} + \frac{2(K-k)+1}{R_b}\right)}{\frac{2(K+1)}{R_b} + \frac{w(k+1)}{R_a}}, & 1 \le k < K \\ \frac{1 - w\lambda_{\min}\left(\frac{1}{R_a} + \frac{1}{R_b}\right)}{\frac{K+1}{R_b} + \frac{w(K+1)}{R_a}}, & k = K \end{cases}.$$

$f_{\mathrm{HD}}(1)$ has a smaller numerator and larger denominator compared to $f_{\mathrm{HD}}(0)$, and therefore $f_{\mathrm{HD}}(0) \ge f_{\mathrm{HD}}(1)$. Similarly, $f_{\mathrm{HD}}(K) \ge f_{\mathrm{HD}}(K - 1)$. Hence, $k = 0$ and $k = K$ will never result in the bottleneck inequality. By observing the derivative of $f_{\mathrm{HD}}(k)$ for $1 \le k < K$,

$$\frac{df_{\mathrm{HD}}(k)}{dk} = \frac{w\lambda_{\min}\left(\frac{2w(k+1)}{R_a R_b} + \frac{4(K+1)}{R_b^2} + \frac{w(2(K-k)+1)}{R_a R_b} + \frac{w}{R_a^2}\right) - \frac{w}{R_a}}{\left(\frac{2(K+1)}{R_b} + \frac{w(k+1)}{R_a}\right)^2}, \quad 1 \le k < K.$$

we see that $f_{\mathrm{HD}}(k)$ is decreasing if $\lambda_{\min} \le \frac{R_a}{4(K+1)\left(\frac{R_a}{R_b}\right)^2 + (2K+3)w\left(\frac{R_a}{R_b}\right) + w}$ and increasing otherwise. Therefore, for the decreasing case, the bottleneck inequality is $f_{\mathrm{HD}}(K-1)$ and for the increasing case, the bottleneck inequality is $f_{\mathrm{HD}}(1)$.

The full-duplex case. Starting similarly with $f_{\mathrm{FD}}(k)$, we have

$$f_{\mathrm{FD}}(k) = \begin{cases} \frac{1 - w\lambda_{\min}\left(\frac{1}{R_a} + \frac{K}{R_b}\right)}{\frac{K+1}{R_b} + \frac{w}{R_a}}, & k = 0 \\ \frac{1 - w\lambda_{\min}\left(\frac{1}{R_a} + \frac{K-k}{R_b}\right)}{\frac{K+1}{R_b} + \frac{w(k+1)}{R_a}}, & 1 \le k < K \\ \frac{1 - \frac{w\lambda_{\min}}{R_a}}{\frac{w(K+1)}{R_a}}, & k = K \end{cases}.$$

Note that $f_{\mathrm{FD}}(K) \ge f_{\mathrm{FD}}(K - 1)$, allowing us to draw analogous conclusions from the HD case. Taking the derivative of $f_{\mathrm{FD}}(k)$ for $1 \le k < K$,

$$\frac{df_{\mathrm{FD}}(k)}{dk} = \frac{w\lambda_{\min}\left(\frac{w(k+1)}{R_a R_b} + \frac{K+1}{R_b^2} + \frac{w(K-k)}{R_a R_b} + \frac{w}{R_a^2}\right) - \frac{w}{R_a}}{\left(\frac{K+1}{R_b} + \frac{w(k+1)}{R_a}\right)^2}, \quad 1 \le k < K$$

we see that when $\lambda_{\min} < \frac{R_a}{(K+1)\left(\frac{R_a}{R_b}\right)^2 + w(K+1)\left(\frac{R_a}{R_b}\right) + w}$, $f_{\mathrm{FD}}(k)$ is decreasing (over $1 \le k < K$) and $f_{\mathrm{FD}}(0) > f_{\mathrm{FD}}(1)$. Hence, the bottleneck inequality is $f_{\mathrm{FD}}(K - 1)$. On the other hand, if $\lambda_{\min} \ge \frac{R_a}{(K+1)\left(\frac{R_a}{R_b}\right)^2 + w(K+1)\left(\frac{R_a}{R_b}\right) + w}$, then $f_{\mathrm{FD}}(k)$ is increasing for $1 \le k < K$ and $f_{\mathrm{FD}}(0) \le f_{\mathrm{FD}}(1)$. Hence, the bottleneck inequality is $f_{\mathrm{FD}}(0)$. $\square$

REFERENCES

[1] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, June 2011.

[2] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, no. 1, pp. 335–349, May 2013.

[3] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr 2016.

[4] 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.

[5] 3GPP TR 38.874, in *Study on integrated access and backhaul*, http://ftp.3gpp.org//Specs/archive/38_ series/38.874/38874-100.zip, Aug. 2020.

[6] M. Cudak, A. Ghosh, A. Ghosh, and J. G. Andrews, "Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 88–94, April 2021.

[7] M. Eslami Rasekh, D. Guo, and U. Madhow, "Joint routing and resource allocation for millimeter wave picocellular backhaul," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 783–794, Feb. 2020.

[8] M. Gupta, A. Rao, E. Visotsky, A. Ghosh, and J. G. Andrews, "Learning link schedules in self-backhauled millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8024–8038, Dec. 2020.

[9] A. Ortiz, A. Asadi, G. H. Sim, D. Steinmetzer, and M. Hollick, "Scaros: A scalable and robust self-backhauling solution for highly dynamic millimeter-wave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2685–2698, Dec 2019.

[10] 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, June 2017.

[11] M. N. Islam, S. Subramanian, and A. Sampath, "Integrated access backhaul in millimeter wave networks," in *Proc. IEEE WCNC*, April 2017.

[12] M. N. Kulkarni, J. G. Andrews, and A. Ghosh, "Performance of dynamic and static TDD in self-backhauled millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6460–6478, Oct 2017.

[13] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, "Impact of interference on multi-hop wireless network performance," *Wireless Networks*, vol. 11, no. 4, pp. 471–487, Jul 2005.

[14] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, March 2000.

[15] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3413–3424, Aug 2009.

[16] A. Zemlianov and G. de Veciana, "Capacity of ad hoc wireless networks with infrastructure support," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 3, pp. 657–667, March 2005.

[17] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5G mmWave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[18] M. Polese, M. Giordani, A. Roy, D. Castor, and M. Zorzi, "Distributed path selection strategies for integrated access and backhaul at mmWaves," in *Proc. IEEE GLOBECOM*, Dec. 2018, pp. 1–7.

[19] T. K. Vu, M. Bennis, M. Debbah, and M. Latva-Aho, "Joint path selection and rate allocation framework for 5G self-backhauled mm-wave networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 4, pp. 2431–2445, Apr 2019.

[20] F. Gómez-Cuba and M. Zorzi, "Twice simulated annealing resource allocation for mmWave multi-hop networks with interference." in *Proc. IEEE ICC*, July 2020, pp. 1–7.

[21] 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, 2019.

[22] C. Madapatha, B. Makki, A. Muhammad, E. Dahlman, M.-S. Alouini, and T. Svensson, "On topology optimization and routing in integrated access and backhaul networks: A genetic algorithm-based approach," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2273–2291, Sept. 2021.

[23] Y. Zhang, M. A. Kishk, and M.-S. Alouini, "A survey on integrated access and backhaul networks," *Frontier Commun. Netw.*, vol. 2, pp. 1–24, June 2021.

[24] I. P. Roberts, H. B. Jain, and S. Vishwanath, "Equipping millimeter-wave full-duplex with analog self-interference cancellation," in *Proc. IEEE ICC WKSHP*, July 2020, pp. 1–6.

[25] A. Bishnu, M. Holm, and T. Ratnarajah, "Performance evaluation of full-duplex IAB multi-cell and multi-user network for FR2 band," *IEEE Access*, vol. 9, pp. 72 269–72 283, May 2021.

[26] I. P. Roberts, J. G. Andrews, H. B. Jain, and S. Vishwanath, "Millimeter-wave full duplex radios: New challenges and techniques," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 36–43, Feb. 2021.

[27] Z. Xiao, P. Xia, and X.-G. Xia, "Full-duplex millimeter-wave communication," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 136–143, 2017.

[28] G. Y. Suk, S.-M. Kim, J. Kwak, S. Hur, E. Kim, and C.-B. Chae, "Full duplex integrated access and backhaul for 5G NR: Analyses and prototype measurements," *CoRR*, vol. abs/2007.03272, June 2021. [Online]. Available: https://arxiv.org/abs/2007.03272

[29] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb 2015.

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

[31] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Max-min rates in self-backhauled millimeter wave cellular networks," Aug. 2018. [Online]. Available: https://arxiv.org/abs/1805.01040

[32] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective.* New York, NY, USA: Cambridge University Press, 2014.

[33] 3GPP TR 38.889, in *Feasibility Study on Licensed-Assisted Access to Unlicensed Spectrum*, http://www.3gpp.org/ftp/Specs/archive/36_series/36.889/36889-d00.zip, July 2015.

[34] Y. Bejerano, S.-J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," *IEEE/ACM Trans. on Netw.*, vol. 15, no. 3, pp. 560–573, June 2007.

[35] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[36] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. on Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[37] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[38] I. P. Roberts, "MIMO for MATLAB: A toolbox for simulating MIMO communication systems," Nov. 2021. [Online]. Available: https://arxiv.org/abs/2111.05273

[39] 3GPP TR 38.901, in *Study on channel model for frequencies from 0.5 to 100 GHz*, http://www.3gpp.org/ftp/Specs/archive/38series/38.901/38901-e20.zip, Sept 2017.

[40] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.