Resource Management for 5G NR Integrated Access and Backhaul: a Semi-centralized Approach

Matteo Pagin, Tommaso Zugno, *Student Member, IEEE,*

Michele Polese, *Member, IEEE,* Michele Zorzi, *Fellow, IEEE*

Abstract

The next generations of mobile networks will be deployed as ultra-dense networks, to match the demand for increased capacity and the challenges that communications in the higher portion of the spectrum (i.e., the mmWave band) introduce. Ultra-dense networks, however, require pervasive, high-capacity backhaul solutions, and deploying fiber optic to all base stations is generally considered to be too expensive for network operators. The 3rd Generation Partnership Project (3GPP) has thus introduced Integrated Access and Backhaul (IAB), a wireless backhaul solution in which the access and backhaul links share the same hardware, protocol stack, and also spectrum. The multiplexing of different links in the same frequency bands, however, introduces interference and capacity sharing issues, thus calling for the introduction of advanced scheduling and coordination schemes. This paper proposes a semi-centralized resource allocation scheme for IAB networks, designed to be flexible, with low complexity, and compliant with the 3GPP IAB specifications. We develop a version of the Maximum Weighted Matching (MWM) problem that can be applied on a spanning tree that represents the IAB network and whose complexity is linear in the number of IAB-nodes. The proposed solution is compared with state-of-the-art distributed approaches through end-to-end, full-stack system-level simulations with a 3GPP-compliant channel model, protocol stack, and a diverse set of user applications. The results show how that our scheme can increase the throughput of cell-edge users up to 5 times, while decreasing the overall network congestion with an end-to-end delay reduction of up to 25 times.

Matteo Pagin, Tommaso Zugno and Michele Zorzi are with the Department of Information Engineering, University of Padova, Padova, Italy. Email: {paginmatte, zugnotom, zorzi} @dei.unipd.it

Michele Polese is with the Institute for the Wireless Internet of Things, Northeastern University, Boston, MA USA. Email: m.polese@northeastern.edu

I. INTRODUCTION

Future generations of cellular networks will provide groundbreaking network capacity, in conjunction with a significantly lower delay and ubiquitous coverage [1]. 5th generation (5G) and beyond deployments will support new mobile broadband use cases, e.g., Augmented (AR) and Virtual Reality (VR), and expand into new vertical markets, enabling an unprecedented degree of automation in industrial scenarios, Vehicle-to-Everything (V2X) communications, remote medical surgery and smart electrical grids.

To this end, the 3rd Generation Partnership Project (3GPP) has introduced various technological advancements with the specifications of the 5G Radio Access Network (RAN) and Core Network (CN), namely NR and 5G Core (5GC) [2]. In particular, NR features a user and control plane split, a flexible Orthogonal Frequency Division Multiplexing (OFDM) frame structure, and the support for millimeter wave (mmWave) communications, while the CN introduces virtualization and slicing [3].

Specifically, the mmWave band, i.e., the portion of the spectrum between 30 GHz and 300 GHz, represents the major technological enabler toward the Gbit/s capacity target. These frequencies are characterized by the availability of vast chunks of contiguous and currently unused spectrum, in stark contrast with the crowded sub-6 GHz frequencies. However, mmWaves exhibit unfavorable propagation characteristics such as high isotropic losses and a marked susceptibility to blockages and signal attenuation [4], [5]. These issues can be partially mitigated using beamforming through large antenna arrays, thanks to the small wavelengths and advances in low-power Complementary Metal-Oxide Semiconductor (CMOS) RF circuits [6]; nevertheless, their introduction alone is not enough for meeting the high service availability requirement. Therefore, mmWaves networks also need densification, to decrease the average distance between mobile terminals and base stations and improve the average Signal-to-Interference-plus-Noise Ratio (SINR). The theoretical effectiveness of this technique is well understood [7]; however, achieving dense 5G deployments is extremely challenging from a practical point of view. Specifically, providing a fiber backhaul among base stations and toward the CN is deemed economically impractical, even more so in the initial 5G deployments [8].

Recently, wireless backhaul solutions for 5G networks have emerged as a viable strategy toward cost effective, dense mmWave deployments. Notably, the 3GPP has promoted Integrated Access and Backhaul (IAB) [9], i.e., a wireless backhaul architecture which dynamically splits

the overall system bandwidth for backhaul and access purposes. IAB has been integrated in the latest release of the 3GPP NR specifications. Prior research has highlighted that IAB represents a cost-performance trade-off [8], [10], as base stations need to multiplex access and backhaul resources, and as the wireless backhaul at mmWaves is less reliable than a fiber connection. In particular, IAB networks may suffer from excessive buffering (and, consequently, high latency and low throughput) when a suboptimal partition of access and backhaul resources is selected, thus hampering the benefits that the high bandwidth mmWave links introduce [8], [11]. Therefore, it is fundamental to solve these non-trivial challenges to enable a smooth integration of IAB in 5G and beyond deployments.

A. Contributions

This article tackles the access and backhaul partitioning problem by proposing an optimal, semi-centralized resource allocation scheme for 3GPP IAB networks, based on the Maximum Weighted Matching (MWM) problem on graphs. It receives periodic Key Performance Indicator (KPI) reports from the nodes of the IAB deployment, constructs a spanning tree that represents the deployment, and uses a simplified, low-complexity version of the MWM to partition the links between access and backhaul. After a feedback step, each node can then schedule the resources at a subframe-level among the connected devices. To the best of our knowledge, this is the first MWM-based resource allocation framework for 3GPP IAB networks at mmWaves, designed with three goals, i.e., it is flexible, integrated with the 3GPP specifications, and has low complexity.

The flexibility makes it possible to easily adapt the resource allocation strategy to different requirements, use cases, and classes of traffic for 5G networks. We achieve this by developing a generic optimization algorithm, which identifies with a configurable periodicity the access and backhaul partition that optimizes a certain utility function. The selection of the utility function prioritizes the optimization of different metrics, e.g., throughput or latency, which in turn can be mapped to different classes of traffic. To achieve the second goal, i.e., the compliance with the 3GPP IAB specifications, the resource allocation framework relies only on information that can be actually exchanged and reported in a 3GPP deployment. In this regard, we also review the latest updates related to the 3GPP IAB standardization activities.

Finally, the algorithm operates with a low complexity, i.e., we propose a version of the MWM algorithm that can be applied on spanning trees with linear complexity in the number of nodes

in the network infrastructure, and demonstrate its equivalence to the generic (and more complex) MWM. Additionally, the proposed framework also relies on a feedback exchange that is linear in the number of base stations, and is thus decoupled from the number of users.

Furthermore, we evaluate the performance of the proposed scheme with an end-to-end, full-stack system-level simulation, using the ns-3 mmWave module [12] and its IAB extension [11]. This represents the first evaluation of an optimized resource allocation scheme for IAB with a simulator that is based on a 3GPP-compliant protocol stack, uses 3GPP channel models, and integrates realistic applications and transport protocols. The extended performance evaluation highlights how the proposed scheme improves the throughput of a diverse set of applications, with a 5-fold increase for the worst case users, with different packet sizes and transport protocols, while decreasing the latency and buffering at intermediate nodes by up to 25 times for the smallest packet sizes.

B. State of the Art

This section reviews relevant research on resources allocation in a multi-hop wireless network, deployed through either IAB or other wireless mesh solutions [13].

The literature adopts different approaches to model and solve the resource allocation problem. The first, discussed in [14]–[20] is based on conventional optimization techniques. Specifically, the authors of [14] present a simple and thus tractable system model and find the minimal number of Next Generation Node Bases (gNBs) featuring a wired backhaul that are needed to sustain a given traffic load. Their work is further extended in [15], which provides an analysis of the performance benefits introduced by additional, fiber-less gNBs. In [16], the mobile network is modeled as a noise-limited, k-ring deployment. Such model is then used to obtain closed-form expressions for the max-min rates achieved by User Equipments (UEs) in the network. Moreover, [17] proposes a system model which leads to an NP-hard optimization problem, even though it considers single-hop backhaul networks only, and uses deep Reinforcement Learning (RL) to reduce its computation complexity. In [18], the joint routing and resource allocation problem is tackled via a Linear Programming (LP) technique. Notably, this work assumes that data can be transmitted (received) toward (from) multiple nodes at the same time. Similarly, the authors of [19] formulate a Time Division Duplexing (TDD), multi-hop resource allocation optimization problem which leverages the directionality of mmWave antennas, albeit in the context of Wireless Personal Area Networks (WPANs). Since such problem is also NP-hard,

a sub-optimum solution is found. Finally, [20] focuses on joint link scheduling, routing and power allocation in multi-hop wireless networks. As in previous cases the obtained optimization problem is not tractable: in this instance such obstacle is overcome by studying the dual problem via an iterative approach.

The second approach relies on stochastic geometry to model IAB networks [10], [21]. Specifically, [21] determines the rate coverage probability of IAB networks and compares different access/backhaul resource partitioning strategies. Similarly, [10] provides a comparison of orthogonal and integrated resource allocation policies, although limited to single-hop wireless networks.

Another significant body of literature leverages Markov Chains (MCs) to study IAB networks; some of these works can be interpreted as a direct application of such theory [22], [23], while others [24]–[27] exploit a more complex framework. The papers which belong to the former class are based on the pioneering work of [28], which inspects the stability of generic multi-hop wireless networks and formulates a throughput-maximizing algorithm known as *back-pressure*. In particular, [22] focuses on the optimization of the timely-throughput, i.e., takes into account that packets usually have an arrival deadline. Such problem is then addressed by formulating a Markov Decision Process (MDP), leading to a distributed resource allocation algorithm. Similarly, [23] proposes an algorithm that also targets throughput optimality but, contrary to the back-pressure algorithm, manages to avoid the need for per-flow information. On the other hand, the body of literature which belongs to the latter class uses the MC-derived Network Utility Maximization (NUM) framework first introduced in [29] and [30]. Specifically, the authors of [24] focus on satisfying the Ultra-Reliable Low-Latency Communication (URLLC) Quality of Service (QoS) requirements by jointly optimizing routing and resource allocation. Then, the problem is solved using both convex optimization and RL techniques. In [25], an in-depth analysis of a mmWave, multi-hop wireless system is presented, proposing and comparing three different interference frameworks, under the assumption of a dynamic TDD system. This work is extended in [26] and [27], which consider respectively a Spatial Division Multiple Access (SDMA) and a Multi-User (MU)-Multiple Input, Multiple Output (MIMO) capable system.

Finally, only a small portion of the literature [8], [11], [31] analyzes the end-to-end performance of IAB networks. Specifically, the authors of [11] extend the ns-3 mmWave module, introducing realistic IAB functionalities which are then used to characterize the benefit of deploying wireless relays in mmWave networks. Their work is extended in [31], where path selection policies are formulated and their impact on the system performance is inspected. A

further end-to-end analysis of IAB networks is carried out in [8], providing insights into the potentials of this technology and the related open research challenges.

Concluding, the literature exhibits the presence of algorithms relying on a varying degree of assumptions on the network topology and the knowledge of system. Furthermore, most of the aforementioned studies lack an end-to-end, full-stack system-level analysis of the proposed solution. Conversely, this paper proposes a semi-centralized resource allocation scheme, which also has a low complexity, both computationally and in terms of required feedback. Moreover, we provide considerations on how our proposed solution can be implemented and deployed in standard-compliant 3GPP IAB networks, and compare such solution to the state of the art with an end-to-end, realistic performance analysis.

C. Paper structure

The remainder of this paper is organized as follows. Sec. II describes our assumptions and the system model. Then, Sec. III presents a novel scheme for resource partitioning in mmWave IAB networks, along with considerations on how it can be implemented in 3GPP NR. Finally, Sec. IV describes the performance evaluation results and Sec. V concludes this paper.

II. IAB NETWORKS

The following paragraphs identify the characteristics and constraints of mmWave IAB, according to the 3GPP design guidelines presented in [9] and the specifications of [32].

A. Network topology

In general, an IAB network is a deployment where a percentage of gNBs (i.e., the IAB-nodes) use wireless backhaul connections to connect to a few gNBs (i.e., the IAB-donors) which feature a wired connection to the core network, as can be seen in Fig. 1. Moreover, these deployments exhibit a *multi-hop* topology where a strict parent-child relation is present. The former can be represented by the IAB-donor itself or an IAB-node; the latter by either UEs or downstream IAB-nodes. In [9], no a priori limit on the number of backhaul hops is introduced. As a consequence, 3GPP argues that IAB protocols should provide sufficient flexibility with respect to the number of backhaul hops. Moreover, the Study Item (SI) on IAB [9] highlights the support for both the topologies depicted in Fig. 1a, i.e., Spanning Tree (ST) and Directed Acyclic Graph (DAG) IAB. Clearly, the former exhibits less complexity but, at the same time, poses some limits in terms of

(a) ST and DAG topologies.
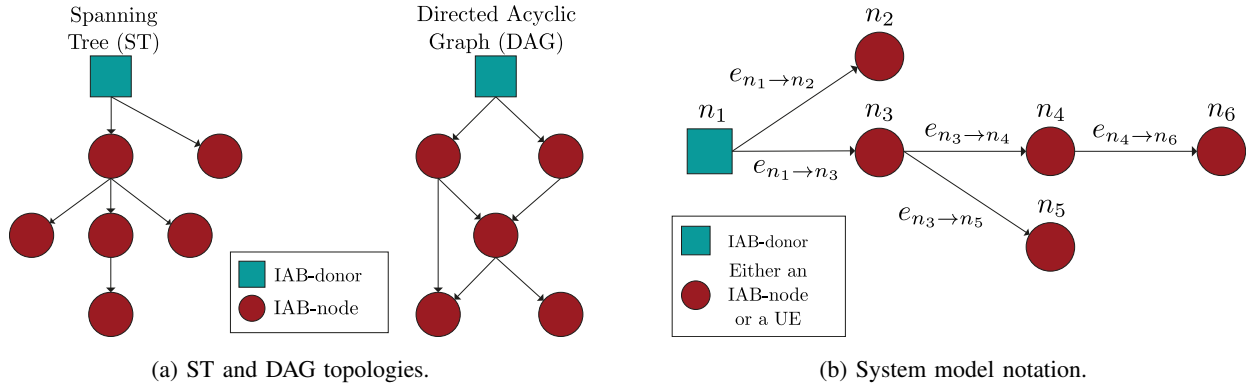
(b) System model notation.

Figure 1. Comparison of the IAB network topologies analyzed in [9] and related notation.

network performance: the possible presence of obstacles may result in a service interruption, due to the unique backhaul route established by the UEs. On the other hand, a DAG topology offers routing redundancy, which can be used not only to decrease the probability of experiencing a "topological blockage," but also for load balancing purposes.

B. Multiple access schemes and scheduling

An in-band, dynamic partitioning of the access and backhaul spectrum resources is currently preferred by 3GPP [9], [32], together with half-duplex operations of the IAB-nodes. Moreover, most of the literature suggests that 5G mmWave systems will operate in a TDD fashion [4], [33]. This choice is mainly driven by the stringent latency requirements which the next generation of mobile networks will be required to support, and by the usage of analog or hybrid beamforming. The usage of Frequency Division Duplexing (FDD), in conjunction with the presence of large chunks of bandwidth, would lead to severe resource under-utilization and make channel estimation more difficult.

Based on these considerations, the system model exhibits a TDD, Time Division Multiple Access (TDMA)-based scheduling where the access/backhaul interfaces are multiplexed in a half-duplex manner. It follows that at any given time instant, each node of the IAB network cannot be simultaneously involved in more than one transmission or reception. In particular, IAB-nodes cannot schedule time and frequency resources which are already allocated by their parent for backhaul communications which involve them. Finally, the introduction of resource coordination mechanisms and related signaling is explicitly supported in the IAB specification drafts [9], [32]. Nevertheless, these solutions must reuse as much as possible the available NR specifications and require at most minimal changes to the Rel.15 5GC and NR.

C. System model

According to these assumptions and referring to Fig. 1b., a generic IAB network can be modeled as a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where the set of nodes $\mathcal{N} \equiv \{n_1, n_2, \ldots n_{|\mathcal{N}|}\}$ comprises the IAB-donor, the various IAB-nodes and the UEs. Accordingly, the set of directed edges $\mathcal{E} \equiv \{e_1, e_2, \ldots e_{|\mathcal{E}|}\} \equiv \{e_{n_j \to n_k}\}_{j,k}$, where the edge $e_{n_j \to n_k}$ originates at the parent node n_j and terminates at the children n_k, comprises in all the active cell attachments, either of mobile terminals to a gNB or from IAB-nodes towards their parent node. Since the goal of this paper is to study backhaul/access resource partitioning policies, this generic model can be actually simplified: in fact, all the UEs connected to a given gNB can be represented by a single node in $\mathcal{G}$ without any loss of generality. Similarly, the same holds true for their links toward the serving gNB, which can then be represented by a single edge. Furthermore, this work focuses on ST topologies only. Nevertheless, the proposed framework can be easily extended to the case of a DAG IAB network: it suffices to introduce a preliminary routing step where an ST $\mathcal{G}'$ is computed from the actual DAG network $\mathcal{G}$, for example by using the strategies presented in [31]. Such process can be possibly repeated at each allocation instance, effectively removing any constraint on the network topology.

We define as *feasible schedule* any set of links $\mathcal{E}' \subseteq \mathcal{E}$ such that none of them share a common vertex, i.e., $\forall e_{n_j \to n_k} \neq e_{n_l \to n_m} \in \mathcal{E}'$ it holds that $n_j \neq n_m$ and $n_l \neq n_k$. Let then f_u be a utility *additive map*, namely, a function such that the overall utility experienced by the system when scheduling edges e_1 and e_2 satisfies $f_u(e_1, e_2) = f_u(e_1) + f_u(e_2)$. Let also $\mathcal{W} \equiv \{w_1, w_2, \ldots w_{|\mathcal{E}|}\}$ be the set of positive weights whose generic entry w_j represents the utility which is obtained when scheduling the j-th edge, namely, $w_j \equiv f_u(e_j)$. Then, the overall utility of the system is $\mathcal{U} \equiv \sum_{e_k \in \mathcal{E}'} f_u(e_k) = \sum_{e_k \in \mathcal{E}'} w_k$. The goal is to find the feasible set $\mathcal{E}'^*$ which maximizes the overall utility, i.e., $\underset{\mathcal{E}'}{\mathrm{argmax}}\,\mathcal{U}$. In computer science, this task is typically referred to as the *Maximum Weighted Matching* problem [34].

Finding the MWM of a given graph, in the general case, is not trivial from a computational point of view. In fact, the fastest known MWM algorithm for generic graphs has a complexity of $\mathcal{O}(|V||E| + |V|^2 \log |V|)$ [35], posing serious limitations to the suitability of such algorithm to 5G and beyond networks, which target a connection density of 1 million devices per km^2. However, we argue that under the aforementioned assumptions on the system model, which restrict the network to an ST topology, it is possible to design an MWM-based centralized

resource partitioning framework which exhibits linear complexity with respect to the network size and which, as a result, is able to satisfy the scalability requirements highlighted by 3GPP in [9].

III. Semi-centralized resource allocation scheme for IAB networks

This section presents an MWM algorithm for ST topologies (Sec. III-A), an efficient and MWM-based centralized resource partitioning framework for IAB networks (Sec. III-B) and some considerations about its implementation (Sec. III-C).

A. MWM for ST graphs

As the first of our contributions, we present an algorithm, hereby called `T-MWM`, which computes the MWM of an ST in linear time. In particular, `T-MWM` is a bottom-up algorithm which, upon receiving as input a weighted ST $\mathcal{G}$ described by its edge list $\mathbf{E}$ and the corresponding weight list $\mathbf{W}$, produces as output a set of active edges $\mathbf{E}^*$ which are an MWM of $\mathcal{G}$. Furthermore, $\mathbf{E}$ is from now on assumed to exhibit the following invariant: each IAB parent precedes its children in the list, hence avoiding the need for a recursion. This is automatically obtained as each IAB child connects after its parent, and is thus added to the list in a subsequent position. Nevertheless, this assumption can be easily relaxed, albeit at the cost of losing as a side-effect the bottom-up design.

As can be seen in Alg. 1, the `T-MWM` algorithm basically performs two traversals. During the first one, the utility yielded by the various nodes and their children is computed. Then, during the second traversal, this knowledge is used for computing an MWM of the network.

The correctness of this procedure can be easily proved. Consider the sub-tree of $\mathcal{G}$ whose root is represented by the generic node n_k. Let also $\mathbf{F}(n_k)$ be the utility yielded by a MWM of such sub-tree which activates a link originating from n_k, and $\mathbf{G}(n_k)$, conversely, the utility provided when the MWM contains no links which originate from such node. Then, the correctness of the first phase of the algorithm, namely the computation of the $\mathbf{F}$ and $\mathbf{G}$ vectors, follows directly from the following Lemmas.

Lemma 1. *Let n_k be an arbitrary internal node of $\mathcal{G}$ and $\{n_j\}_k$ be the set of its children. Then, any MWM of $\mathcal{G}$ must contain an edge which has as one of its vertices either n_k or an element of $\{n_j\}_k$.*

Algorithm 1 Tree-Maximum Weighted Matching

    Input: A weighted ST $\mathcal{G}$ encoded by a list $\mathbf{E}$, which associates each node in $\mathcal{G}$ to its edges, and the corresponding weights list $\mathbf{W}$.

    Output: An MWM $\mathbf{E}^*$ of $\mathcal{G}$.

1: **procedure** `T-MWM`$(\mathbf{E}, \mathbf{W})$
2:     $\mathbf{F} \leftarrow \mathbf{0}$; $\mathbf{G} \leftarrow \mathbf{0}$ ▷ Initialize the utility vectors to zero vectors
3:     $\mathbf{E}^* \leftarrow \{\}$ ▷ Initialize the set of active edges as empty
4:     **for** each node $n_k \in \mathbf{E}$ **do**   ▷ Iterate over the various nodes, in ascending order w.r.t. to their depth in $\mathcal{G}$
5:         $maxUtil \leftarrow 0$; $\mathbf{maxEdge}(n_k) \leftarrow \{\}$
6:         **for** each edge $e_{k,j} \equiv (n_k, n_j)$ **do**   ▷ Iterate over its edges
7:             $\mathbf{G}(n_k) \leftarrow \mathbf{G}(n_k) + \mathbf{F}(n_j)$
8:             $currUtil \leftarrow \mathbf{W}(e_{k,j}) + \mathbf{G}(n_j) - \mathbf{F}(n_j)$
9:             **if** $currUtil > maxUtil$ **then**
10:                $maxUtil \leftarrow currUtil$; $\mathbf{maxEdge}(n_k) \leftarrow e_{k,j}$
11:                $\mathbf{F}(n_k) \leftarrow \mathbf{G}(n_k) + maxUtil$
12:             **end if**
13:         **end for**
14:     **end for**
15:     **for** each node $n_k \in \mathbf{E}$ **do**   ▷ Iterate over the various nodes, in ascending order wrt to their depth in $\mathcal{G}$
16:         **if** $\mathbf{F}(n_k) \geq \mathbf{G}(n_k)$ **then**
17:             $\mathbf{E}^* \leftarrow \mathbf{E}^* \cup e_{k,\mathbf{maxEdge}(n_k)}$
18:             $\mathbf{F}(\mathbf{maxEdge}(n_k)) \leftarrow -1$ ▷ Ensure child does not get activated multiple times
19:         **end if**
20:     **end for**
21:     **return** $\mathbf{E}^*$
22: **end procedure**

Proof: Suppose there exists an MWM $\mathbf{E}^*$ of $\mathcal{G}$ which does not contain any such edge. Then the set $\hat{\mathbf{E}}^* \equiv \mathbf{E}^* \cup \{e_{n_k \to n_m}\}$, where $e_{n_k \to n_m}$ is the edge from n_j to its (arbitrary) child n_m is still a feasible activation set, since no edge in $\mathbf{E}^*$ shares such vertices. Furthermore, since the weights are positive we have that $f_u(\hat{\mathbf{E}}^*) \equiv f_u(\mathbf{E}^*) + \mathbf{W}(e_{n_k \to n_m}) > f_u(\mathbf{E}^*)$, which is clearly a contradiction. ∎

Lemma 2. *For any internal node n_k:*

$$
\begin{cases}
\mathbf{G}(n_k) = \sum_{\{n_j\}_k} \mathbf{F}(n_j) \\
\mathbf{F}(n_k) = \sum_{\{n_j\}_k} \mathbf{F}(n_j) + \max_{\{n_j\}_k}\{\mathbf{W}(e_{n_k \to n_j}) + \mathbf{G}(n_k) - \mathbf{F}(n_k)\}
\end{cases}
$$

where the set $\{n_j\}_k$ comprises all the children of n_k. Conversely, for leaf nodes $\mathbf{F}(n_l) \equiv \mathbf{G}(n_l) \equiv 0$.

Proof: This lemma can be proved by induction over the height h_k of the sub-tree corresponding to node n_k. The base case is $h_k = 0$, i.e., when n_k is a leaf node; in this case, trivially, both $\mathbf{F}(n_k)$ and $\mathbf{G}(n_k)$ are zero since no links exhibit n_k as parent node and the sub-tree of $\mathcal{G}$ which originates in n_k consists of n_k only, respectively.

Consider then the node n_k having a sub-tree of height $h_k > 0$. From Lemma 1 we know that either n_k or (at least) one of its children must be included in any MWM. If on the one hand we do not activate any edge which originates from n_k, then no constraints on the children's activation are introduced. Therefore, in this case the maximum utility is obtained when *all* the children are active, hence $\mathbf{G}(n_k) = \sum_{\{n_j\}_k} \mathbf{F}(n_j)$. On the other hand, if we activate an edge from n_k to one of its children n_m then no additional edges which originate from the latter can be added to $\mathbf{E}^*$. It follows that the utility obtained in this instance reads:

$$
\sum_{\{n_j \neq n_m\}_k} \mathbf{F}(n_j) + \mathbf{W}(e_{n_k \to n_m}) + \mathbf{G}(n_m)
$$

and can be rewritten as:

$$
\sum_{\{n_j\}_k} \mathbf{F}(n_j) + \mathbf{W}(e_{n_k \to n_m}) + \mathbf{G}(n_m) - \mathbf{F}(n_m)
$$

Such utility is maximized when n_m is chosen as $\underset{\{n_j\}_k}{\mathrm{argmax}} \{\mathbf{W}(e_{n_k \to n_j}) + \mathbf{G}(n_j) - \mathbf{F}(n_j)\}$, yielding:

$$
\mathbf{F}(n_k) = \sum_{\{n_j\}_k} \mathbf{F}(n_j) + \max_{\{n_j\}_k} \{\mathbf{W}(e_{n_k \to n_j}) + \mathbf{G}(n_k) - \mathbf{F}(n_k)\}
$$

∎

Finally, the validity of the last phase of `T-MWM` follows from Lemma 3.

Lemma 3. *Given an ST $\mathcal{G}$, let $\mathcal{G}_k$ be its sub-tree of root n_k. Then, an MWM of $\mathcal{G}$ can be computed by performing, in a recursive fashion and starting from the root, the following procedure:*

1) *If* $\mathbf{F}(n_k) \geq \mathbf{G}(n_k)$, *add to* $\mathbf{E}^*$ *the edge from* n_k *to* n_m, *where the latter is defined as* $n_m \equiv \underset{\{n_j\}_k}{\mathrm{argmax}} \{\mathbf{W}(e_{n_k \to n_j}) + \mathbf{G}(n_j) - \mathbf{F}(n_j)\}$. *Then, repeat recursively on all the sub-trees corresponding to* n_k*'s children* $\{n_j\}_k$ s.t. $n_j \neq n_m$ *and on the children of* n_m *itself.*

2) *If* $\mathbf{F}(n_k) < \mathbf{G}(n_k)$, *repeat recursively on all the sub-trees which exhibit the children of* n_k *as their root.*

Proof: This Lemma follows directly from the definitions of $\mathbf{F}$ and $\mathbf{G}$ and the previous Lemmas. Specifically, the above procedure always yields a feasible activation, i.e., a matching of $\mathcal{G}$. In fact, in either options we never recurse on a node which has already been activated, hence no pair of edges $\in \mathbf{E}$ can share any vertices. Furthermore, due to the properties of $\mathbf{F}$ and $\mathbf{G}$ and their validity for each sub-tree in $\mathcal{G}$, the edges of $\mathbf{E}^*$ comprise a *maximal* matching, i.e., they yield the maximum possible utility among all the feasible schedules. ∎

Regarding the computational complexity of the proposed algorithm, it can be observed that during the first phase the main loop effectively scans each edge of $\mathcal{G}$, hence exhibiting a complexity $\mathcal{O}(|E|)$. Moreover, the second phase of T-MWM has complexity $\mathcal{O}(|V|)$, since it loops through all the network nodes. Therefore, we can conclude that the overall asymptotic complexity of the algorithm is $\mathcal{O}(|V| + |E|)$, or, equivalently, $\mathcal{O}(|V|)$ since in an ST the number of edges equals $|V| - 1$.

B. Semi-centralized resource partitioning scheme

Based on the system model introduced in Sec. II, and the T-MWM algorithm, we present a generic optimization framework which partially centralizes the backhaul/access resource partitioning process, in compliance with the guidelines of [9]. The goal of this framework is to aid the distributed schedulers, adapting the number of OFDM symbols allocated to the backhaul and access interfaces to the phenomena which exhibit a sufficiently slow evolution over time, i.e., large scale fading and local congestion. This optimization is undertaken with respect to a generic additive utility function f_u. An IAB network of arbitrary size is considered, composed of a single IAB-donor, multiple IAB-nodes and a (possibly time-varying) number of UEs which connect to both types of gNBs. Furthermore, let the topology of the IAB network be pre-computed, for instance by using the policies of [31], and assume that a central controller is installed on the IAB-donor.
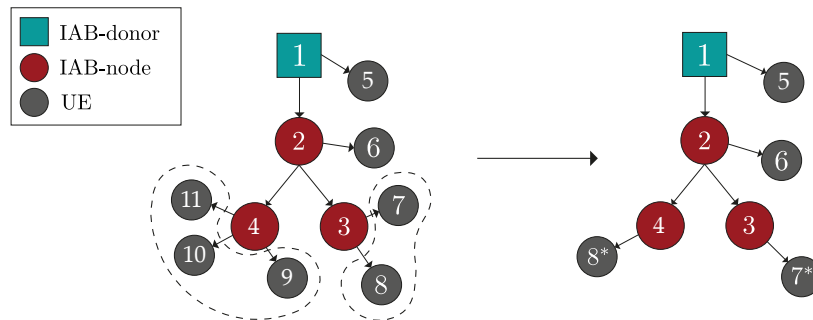
Figure 2. Creation of the IAB network graph. The original topology, exhibiting the actual cell attachments, is depicted on the left. Conversely, the reduced topology is shown on the right.

The proposed framework can be subdivided into the following phases, which are periodically repeated every T_{alloc} subframes:

1) **Initial setup**. This step consists in the computation of the simplified IAB network graph $\mathcal{G} \equiv \{\mathcal{V}, \mathcal{E}\}$. Specifically, $\mathcal{V}$ is composed of the donor, the various IAB-nodes and, possibly, of additional nodes which represent the set of UEs that are connected to a given gNB. Accordingly, $\mathcal{E}$ contains the active cell associations of the aforementioned nodes. This process, depicted in Fig. 2, must be repeated every time the IAB topology changes, i.e., whenever a new UE performs its Initial Access (IA) procedure or an IAB-node connects to a different parent due to a Radio Link Failure (RLF).

2) **Information collection**. During this phase, the various IAB-nodes send to the central controller a pre-established set of information for each of their children in $\mathcal{G}$. For instance, this feedback may consist in their congestion status and/or information regarding their channel quality. To such end, the implementation of this paper uses modified versions of pre-existing NR Release 16 Control Elements (CEs), as strongly recommended in the IAB SI [9]. However, the scheme does not actually impose any limitations in such regard.

3) **Centralized scheduling indication**. Upon reception of the feedback information, the central controller calculates the set of weights $\mathcal{W}$ accordingly. Then, an MWM of $\mathcal{G}$ is computed using the T-MWM algorithm and producing as output the activation set $\mathbf{E}^*$, which maximizes the overall utility of the system with respect to the chosen utility function. Subsequently, $\mathbf{E}^*$ is used in order to create a set of *favored* downstream nodes, i.e., of children which will be served with the highest priority by their parent, as depicted in Fig. 3. Finally, these scheduling indications are forwarded to the various IAB-nodes which act as parents in the edges of $\mathbf{E}^*$.
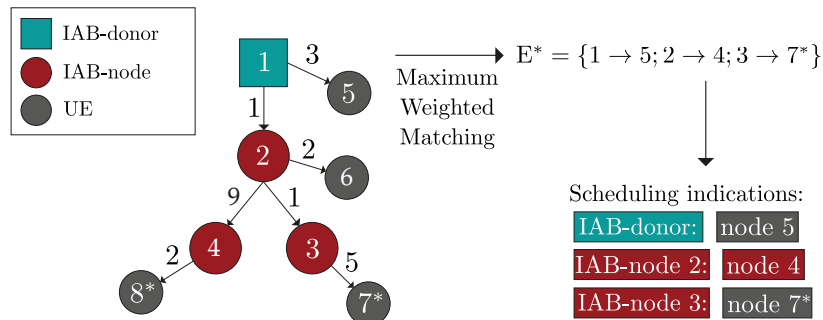
Figure 3. Computation of the MWM and of the corresponding scheduling indications.

4) **Distributed scheduling allocation**. During this phase, the various IAB-nodes make use of the indications received by the central controller, if available, in order to perform the actual scheduling (which is, therefore, predominantly distributed). Specifically, the favored nodes are served with the highest priority, while the remaining downstream nodes are scheduled if and only if the resource allocation of the former does not exhaust the available OFDM symbols.

It is important to note that since $\mathcal{G}$ contains only the IAB-nodes, the donor and at most one "representative" UE per gNB, the proposed scheme effectively performs only the backhaul/access resource partitioning in a centralized manner. On the other hand, the actual Medium Access Control (MAC)-level scheduling is still undertaken in a distributed fashion, albeit leveraging the indications produced by the central controller. The major advantages which this two-tier design exhibits, compared to a completely centralized solution, are the presence of a relatively light signaling overhead and the ability to promptly react to fast channel variations, for instance caused by small scale fading.

C. Implementation of centralized allocation schemes in mmWave IAB networks

The remainder of this section discusses how the proposed scheme can be implemented in IAB deployments, with references to how the 3GPP specifications can support it.

Basically, the resource allocation framework requires (i) a central controller, which is installed on the IAB-donor, or could be deployed in a RAN Intelligent Controller (RIC) following the O-RAN architecture [36]; and (ii) a scheduler which exchanges resource coordination information with the former and computes the weights for the resource allocation. In particular, referring to the aforementioned phases of the proposed scheme, the following implementation considerations can be made.

1) Initial setup: The setup of the various centralized mechanisms is subdivided into two sub-phases: an initial configuration, where the relevant entities are initialized, and a periodic update of the topology information. The former takes place when the IAB-nodes are first connected to the network. During this phase, the controller acquires preliminary topology information, by leveraging the configuration messages which are already exchanged during the usual Rel.16 IA procedure, generating a map which associates the depth in the IAB network to a pair of child-parent global identifiers (which from now on will be referred to as "IDs"). Since this phase takes place when no UE has performed its IA procedure yet, the exchanged topology information concerns the donor and IAB-nodes only.

Moreover, the central controller is in charge of periodically updating the topology information. In order to minimize the signaling overhead, this process does not require any additional control information: in fact, the status information which is already collected in a periodic manner can be leveraged in such regard. Specifically, the periodic info received from the various IAB-nodes, which will carry a list of ID-value pairs, is analyzed. The child-parent associations are then compared with the ones known by the controller, updating the latter whenever the two topology information happens to differ.

2) Information collection: The generation of the feedback information is performed in a distributed manner by both the IAB-nodes and the IAB-donor. To such end, the current implementation features the forwarding of information on the channel quality and buffer status, in the form of Channel Quality Informations (CQIs) and Buffer Status Reports (BSRs) respectively. This choice is driven by both the will of maximizing the re-utilization of the NR Rel.16 specifications and the goal of making use of MAC-level CEs only, hence avoiding the introduction of any constraint regarding the supported IAB-relaying architecture.

In particular, the CQI and BSR information is generated by analyzing the corresponding CEs which are received by the host gNB, and checking whether the source Radio Network Temporary Identifier (RNTI) belongs to an IAB-node or to a UE. In the first case, the corresponding ID is retrieved and an entry carrying such identifier along with its CQI/BSR value is generated. The feedback information concerning the UEs, instead, is averaged in the case of the CQIs and added up for the BSRs, to obtain a single value for each gNBs. It can be noted that both CQIs and BSRs are available to the scheduler, since the UL buffer statuses are already periodically reported by the downstream nodes via their BSRs and the DL statuses can be easily retrieved by the former, since the Radio Link Control (RLC) buffers reside on the same node as the scheduler

itself, i.e., the gNB.

Referring to the 3GPP specifications of [9], [37], [38], the buffers occupancy can then be forwarded to the IAB-donor by introducing a periodic-only BSR whose period is controlled by an ad hoc Radio Resource Control (RRC) timer. Similarly, the channel qualities can be reported by the various IAB-nodes via additional periodic CQIs which would carry only the CQI index, hence neglecting the Rank Index (RI) and Precoding Matrix Index (PMI), since this information would generate unnecessary signaling overhead. These CEs would preferably leverage pre-existing NR measurements: the main novelty would be the introduction of their periodic reporting to the IAB-donor. To such end, the 5G CQI and BSR data-structures require an additional field which carries the ID, if the chosen IAB-relaying architecture does not feature an Adaptation Layer [9]. Conversely, relaying solutions which support the latter can reuse the NR CEs and let such layer introduce an additional header.

3) Centralized scheduling indication: Periodically, the controller located at the donor makes use of the feedback received by the IAB-nodes to first compute the weights of the various network links and then generate the centralized scheduling indications. We propose the following policies to compute the weights for the MWM problem:

1) **Max Sum-Rate (MSR)**. This policy maximizes the overall Physical (PHY)-layer through-put, i.e., the utility function is

$$f_u^{\text{MSR}} \equiv \sum_{e_{i \to k} \in \mathbf{E}^*} c_{i,k},$$

and the weight assigned to the edge from node i to node k reads $w_{i,k} \equiv c_{i,k}$, where $c_{i,k}$ is the capacity of the link $e_{i \to k}$.

2) **Backlog Avoidance (BA)**. This resource partitioning strategy aims at avoiding congestion. Therefore, the system utility is:

$$f_u^{\text{BA}} \equiv \sum_{e_{i \to k} \in \mathbf{E}^*} q_{i,k},$$

where the weight $w_{i,k}$ reads $q_{i,k}$, namely, the amount of buffered data which would reach its next hop in the IAB network by crossing the link $e_{i \to k}$.

3) **Max-Rate Backlog Avoidance (MRBA)**. This represents the most balanced option among the three, since it exploits favorable channel conditions while also preventing network congestion and favoring network fairness. The weight assigned to link $e_{i \to k}$ is:

$$w_{i,k} \equiv c_{i,k} + \eta \cdot q_{i,k} \cdot \left(\frac{\mu}{\mu_{thr}} \right)^k,$$

where η, μ_{thr} and k are arbitrary parameters and μ represents the number of subframes which have elapsed since the last time edge $e_{i \to k}$ has been marked as favored.

Once the weights are computed, the controller obtains an MWM of the network via an implementation of the aforementioned `T-MWM`. This function outputs the activation set $\mathbf{E}^*$, i.e., a map associating the ID of the parent gNB to the one of its favored downstream node. Notably, this set does not necessarily contain scheduling indications for *each* IAB-node in the network: an entry corresponding to a given gNB is present if and only if such node is indeed active in the MWM. First, this map is used by the controller in order to keep track of which link has not been favored and for how long; this information may then be used to introduce a weight prediction mechanism, improving the robustness of the scheme with respect to the information collection period. Finally, these scheduling indications are forwarded to the corresponding IAB-nodes.

4) Distributed scheduling allocation: The last phase of the resource allocation procedure consists in the distributed MAC-level scheduling. Before assigning the available resources, the various schedulers check whether any indication has been received from the controller. Based on this condition, the buffer occupancy information is then split into two groups. The first contains the BSRs related to the favored RNTI (if any), with the caveat that if the latter indicates the cumulative access link, then this set contains the BSRs of all the UEs attached to the host gNB, while the other comprises the remaining control information. The resource allocation process is then undertaken twice: first considering the set of favored BSRs only, then the remainder of these CEs.

Thanks to this design, the favored link(s) is (are) scheduled with the highest priority, while the rest of the network only gets the remaining resources. In such a way, the information received by the controller is actually used as an **indication** and not as the eventual **resource allocation**. For instance, the gNBs are free to override these indications whenever the buffer of the favored child is actually empty, due to discrepancies between its actual status and the related information available to the controller. In such a way, the unavoidable delay between the information collection and the reception of the scheduling does not lead to any resource underutilization. Moreover, this is achieved with minimal changes to the state of the art schedulers, making the proposed scheme relatively easy to implement and deploy in real-world networks.

IV. Performance evaluation

We implemented the proposed resource allocation scheme in the popular open source simulator ns-3, exploiting the mmWave module [12] and its IAB extension [11], to characterize the system-level performance of the proposed solution with realistic protocol stacks, scenarios, and user applications.

The ns-3 mmWave module is based on [39] and introduces mmWave channel models, including the 3GPP channel model for 5G evaluations [40], and highly customizable PHY and MAC layer implementation, with an NR-like flexible OFDM numerology and frame structure. Additionally, the IAB module [11] models wireless relaying functionalities which mimic the specifications presented in [9]. Specifically, this module supports both single and multi-hop deployment scenarios, auto-configuration (within the network) of the IAB-nodes and a detailed 3GPP protocol stack, allowing wireless researchers to perform system-level analyses of IAB systems in ns-3.

It is of particular relevance to understand how the scheduling operations are implemented in the IAB module, since they offer not only the baseline for the proposed scheme, but also valid guidelines for real-world deployments. The current ns-3 IAB schedulers exhibit a TDMA-based multiplexing between the access and backhaul interfaces. Moreover, scheduling decisions are undertaken in a distributed manner across the IAB network, i.e., each gNB allocates the resources which its access interface offers (to both UEs and IAB-nodes) independently of the other gNBs in the network. In fact, in an IAB network these scheduling decisions are *almost* independent of one another: if a parent node schedules the backhaul interface of a downstream node, clearly the latter will be constrained in its own scheduling decisions, as it will not be allowed to allocate the time resources which have already been scheduled for backhaul transmissions by its parent. Therefore, in a tree-based, multi-hop wireless network the various gNBs need to know in advance the scheduling decisions performed by their upstream nodes: to solve this problem, the authors of the IAB module for ns-3 introduced a "*look-ahead backhaul-aware scheduling mechanism*" [11]. Such mechanism features an exchange of Downlink Control Information (DCI) between the access and backhaul interfaces: in such a way, any time resources already scheduled by the parent for backhaul communications can be marked as such by the corresponding downstream node, preventing any overlap with other transmissions. Furthermore, the *look-ahead* mechanism requires the schedulers of the various gNBs to commit to their resource allocation for a given time T at a time $T - k$, where $k - 1$ is the maximum distance (in terms of wireless hops) of
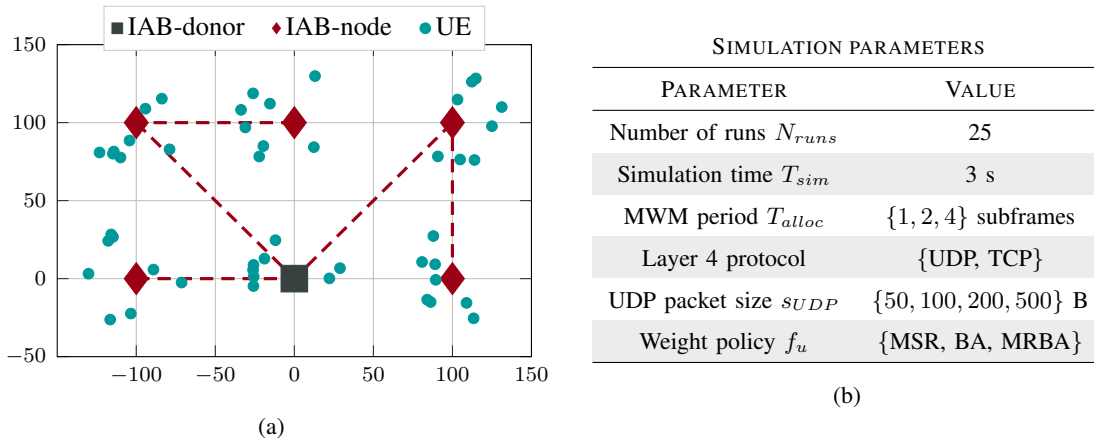
Figure 4. On the left, a realization of the simulation scenario is depicted; the dotted lines represent the cell-attachments of the IAB-nodes. On the right, a brief list of simulation parameters is provided.

any node from the donor. In such a way, the DCIs will have time to propagate across the IAB network and reach the farthest node at time $T - 1$, thus allowing its scheduler to perform the resource allocation process at least one radio subframe in advance.

A. Simulation scenario and parameters

The purpose of these simulations is to understand the performance of the proposed resource partitioning framework in the context of its target deployment, i.e., a multi-hop IAB network. As a consequence, the reference scenario consists of a dense urban deployment with a single IAB-donor and multiple IAB-nodes, as depicted in Fig. 4a. In particular, the various gNBs are distributed along an urban grid where the donor is located at the origin while the IAB-nodes are deployed along the street intersections, with a minimum inter-site distance of 100 m. The IAB-nodes attachments are computed using the so-called *HQF* policy presented in [31]; however, this choice does not introduce any loss of generality since such parameter is fixed for all the runs. A given number of UEs are deployed within the surroundings of these base stations, with an initial position which is randomly sampled from circles of radius ρ and whose centers are the various gNBs.

Both the IAB-donor and the IAB-nodes are equipped with a phased array featuring 64 antenna elements, and transmit with a power of 33 dBm; conversely UEs are equipped with 16 antenna elements and their transmission power is restricted to 23 dBm. Notably, the presence of additional antenna elements at the gNBs is a key (but reasonable) assumption, as it allows base stations to achieve a high beamforming gain. In turn, it is possible to achieve a high capacity, which is

fundamental to avoid performance bottlenecks, given the absence of a fiber backhaul. The UEs download data which originates from sources that are installed on a remote host; both the User Datagram Protocol (UDP) and the Transmission Control Protocol (TCP) are used. For the UDP simulations, the rate of the sources is varied from 4 to 40 Mbps to introduce different degrees of saturation in the network. Therefore, in these simulations only DL traffic is considered. Finally, the performance of the proposed policies is hereby compared with the baseline of [11], indicated as "Dist." by examining end-to-end throughput, latency, and a network congestion metric.

B. Throughput

The first metric which is inspected in this analysis is the end-to-end throughput at the application layer. As a consequence, only the packets which are correctly received at the uppermost layer of the destination node in the network are taken into account. In particular, for each UE and each simulation run, the long-term average throughput is computed as follows:

$$S_{k,n}^{\text{APP}} \equiv \frac{B(T_{sim}, k, n)}{T_{sim}}$$

where $B(t, j)$ is the cumulative number of bits received up to time t by the k-th UE, during the n-th simulation run. Then, the distribution of $\mathbf{S}^{\text{APP}}$, namely, the vector containing the collection of the $S_{k,n}^{\text{APP}}$ values across the different runs and UEs, is analyzed.

Figs. 5a and 5b report the Empirical Cumulative Distribution Function (ECDF) of $\mathbf{S}^{\text{APP}}$, for a UDP packet size of 100 and 500 bytes, respectively, and the policies introduced in Sec. III-C. In the former, we can notice that the introduction of the centralized framework increases by up to 15% the percentage of UEs whose throughput matches the rate of the UDP sources. Moreover, by focusing on the leftmost portion of Fig. 5a we can observe another interesting result, concerning the throughput experienced by the UEs which do not fulfill their QoS requirements. In fact, with respect to the first quartile the distributed scheduler achieves the worse performance, with 25% of the UEs obtaining a throughput smaller than 3.3 Mbps. The centralized framework significantly improves these results, even though the extent of such improvements varies quite dramatically across the different policies. Compared with the distributed case, the MSR policy achieves a higher throughput with respect to all the percentiles, albeit exhibiting the same high variance of the former. Instead, the BA and MRBA policies have a dramatic impact on the system performance, introducing a 5-fold increase of the worst case throughput coupled with a significantly lower variance.

(a) s_{UDP} = 100 B, i.e., UDP rate of 8 Mbps.

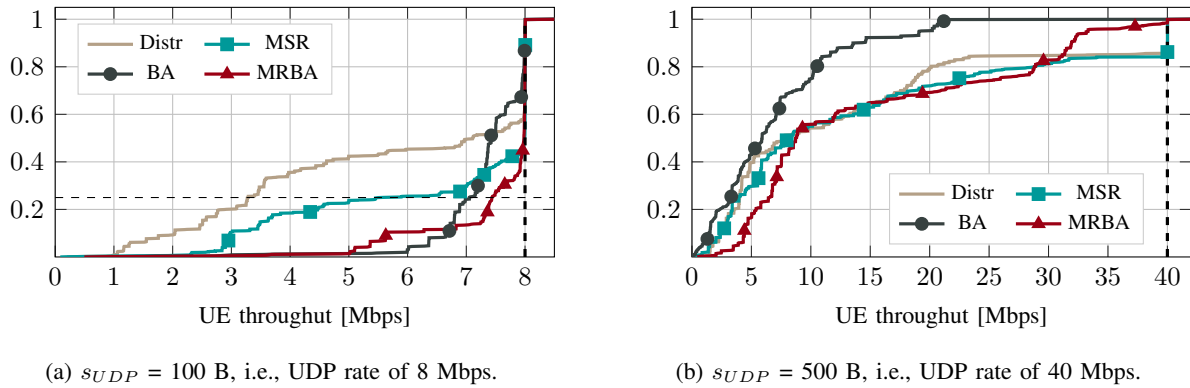(b) s_{UDP} = 500 B, i.e., UDP rate of 40 Mbps.

Figure 5. Per-UE end-to-end throughput ECDFs. The dashed line represents the rate of the UDP sources.

These results can be explained as follows: since a UDP packet size of 100 bytes does not saturate the capacity of the access links, the main performance bottleneck of this configuration is represented by the buffering of the aggregated traffic on the intermediate backhaul links. Therefore, the MSR policy provides only minimal improvements compared to the performance of the distributed scheduler, since it simply favors the links which exhibit a higher SINR. Conversely, the prioritization of the most congested links which is introduced by the other two strategies successfully tackles the former problem. In particular, the BA policy exhibits the highest worst case throughput, albeit at the cost of satisfying the QoS requirements of approximately 20% of the UEs. On the other hand, the bias towards high SINR channels introduced by the MRBA strategy has the opposite effect, improving mostly the higher percentiles but also outperforming MSR and the baseline in the lower percentiles.

By increasing the UDP packet size to 500 bytes, the network becomes noticeably saturated, as depicted by Fig. 5b; in fact, in this instance only a minority of the UEs achieves a throughput which is comparable to the source rate. With this configuration, the BA strategy achieves the worst performance, providing a significantly lower throughput across all the percentiles. On the other hand, the differences among the behavior of the remaining strategies are less evident. In particular, the MSR policy exhibits only a slight improvement over the distributed solution, albeit noticeable across the whole ECDF. The MRBA, conversely, introduces performance benefits which mostly affect the bottom percentiles only. However, with this strategy only a limited portion of the UEs achieves the target throughput of 40 Mbps. As a consequence, we can conclude that with the configuration depicted in Fig. 5b the network is approaching the capacity of the mmWave channels. Therefore, buffering phenomena are likely occurring at each intermediate IAB-node.
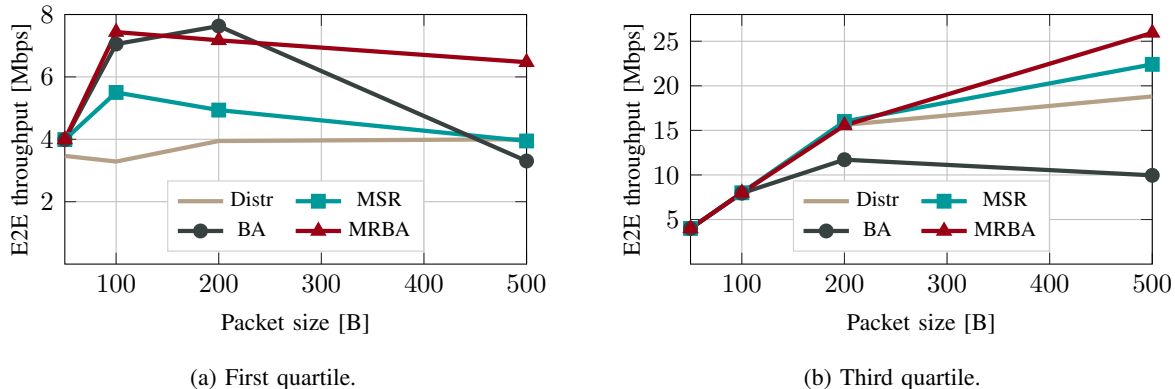
(a) First quartile.

(b) Third quartile.

Figure 6. End-to-end throughput quartiles, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.

Moreover, we can say that in a saturated network the congestion is so severe that prioritizing the bottleneck links is not enough: we also need to take into account the channel conditions and prioritize the links which not only are congested, but also have the "biggest chance" of getting rid of the buffered data due to the temporary better channel quality.

Finally, Fig. 6 presents the first and third quartiles of $\mathbf{S}^{APP}$ as a function of the UDP packet size s_{UDP}. It can be noted that, with respect to the first quartile, the MRBA outperforms all the other policies by delivering a throughput which is up to 90% higher than the one obtained by the distributed scheduler. In fact, Fig. 6b shows how MRBA achieves also the best third quartile, albeit the improvement over the distributed solution is less dramatic. Furthermore, we can observe how the positive impact of the BA strategy is inversely proportional to the saturation in the network. We can then conclude that the bias it introduces loses its effectiveness as the buffering phenomena start to affect the majority of the IAB-nodes.

C. Latency

Just like the aforementioned metric, the latency is measured end-to-end at the application layer. Thanks to this choice, the resulting delay accurately represents the system-level performance, as it includes the latency which is introduced at each hop in the IAB network.

In particular, for each packet correctly received at the uppermost layer of its final destination, the following quantity is traced:

$$D_i^{APP} \equiv \sum_{l_k \in \mathcal{E}_i} D_i^{l_k}$$

where $\mathcal{E}_i$ comprises the links in the IAB network that are crossed by the i-th packet, while the term $D_i^{l_k}$ indicates its point-to-point latency over the path link l_i. Finally, these values are
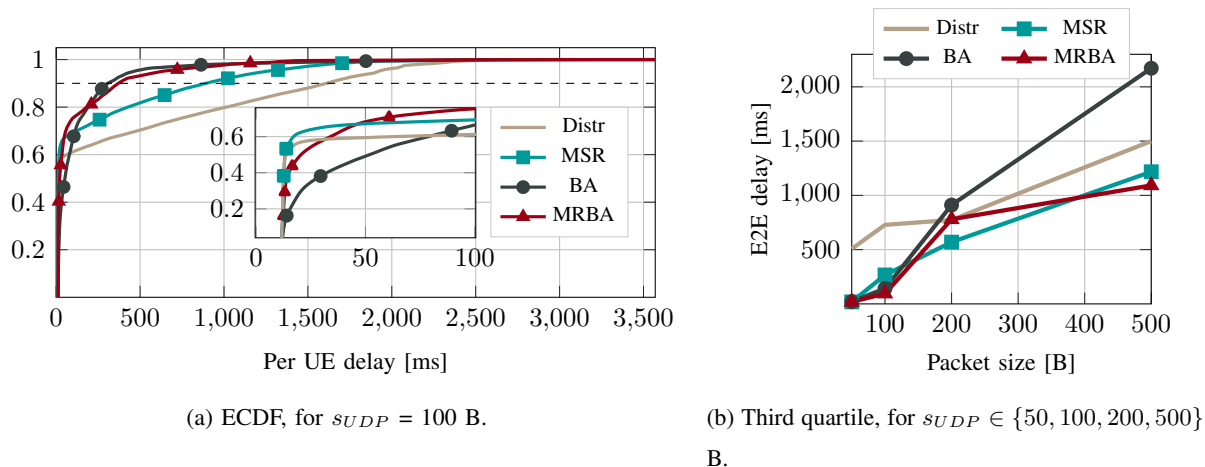
(a) ECDF, for s_{UDP} = 100 B.

(b) Third quartile, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.

Figure 7. Per-UE end-to-end delay statistics.

collected for each of the various runs into the vector $\mathbf{D}^{\mathrm{APP}}$ and its statistical properties are inspected.

Fig. 7a shows the empirical ECDF of $\mathbf{D}^{\mathrm{APP}}$ for a packet size of 100 bytes. It can be noticed that, in this case, the 90th percentile achieved by the BA and the MRBA policies are approximately 20 % smaller than the one obtained by the distributed scheduler. Moreover, these strategies manage to dramatically reduce the number of packets received with extremely high delay, i.e., in the order of seconds, showing the dramatic impact of buffering in the baseline configuration. Conversely, the MSR policy provides the best performance with respect to the best case delay only, although it still outperforms quite significantly the distributed strategy.

These trends are exacerbated by Fig. 7b, which shows the third quartile of $\mathbf{D}^{\mathrm{APP}}$ as a function of the UDP packet size s_{UDP}. In fact, we can notice that the effectiveness of the BA policy is inversely proportional to the network saturation; the opposite holds true with respect to the MSR strategy. It follows that, for UDP rates in the order of 5 to 10 Mbps, the network is mainly plagued by local congestion which causes the insurgence of buffering in some of the nodes. Conversely, as the rate of the UDP sources increases the system shifts to a capacity-limited regime, a phenomenon which explains the dominance of the MSR and MRBA policies.

D. Network congestion

The network congestion is measured by collecting, every T_{alloc} subframes, the RLC buffers status of the various nodes into the vector $\mathbf{B}^{\mathrm{RLC}}$. It must be noted that, since RLC Acknowledged Mode (AM) is used, these values will indicate data which is related to both new packets and possible retransmissions.
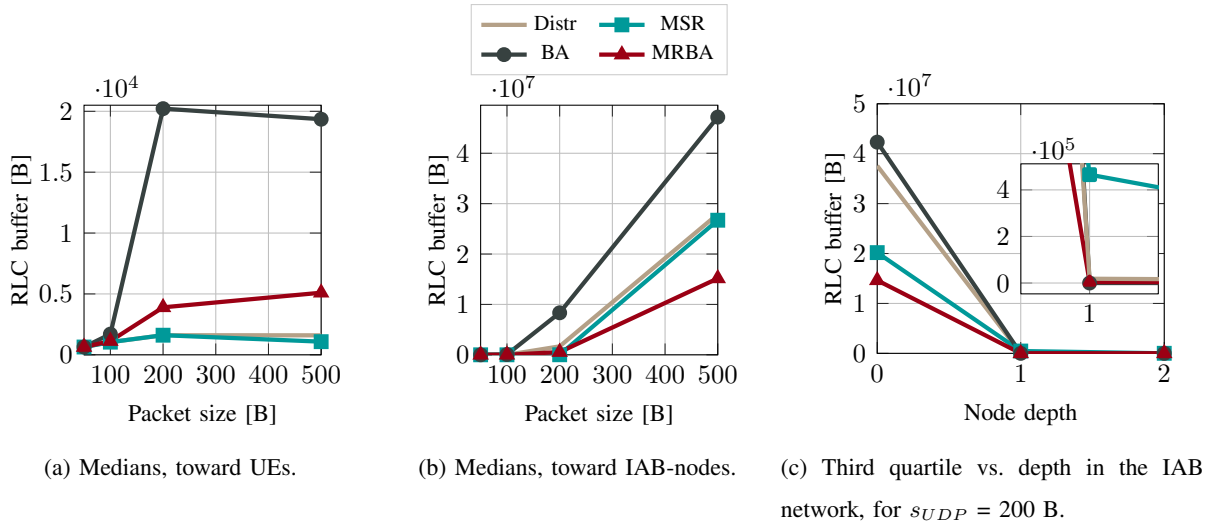
(a) Medians, toward UEs.

(b) Medians, toward IAB-nodes.

(c) Third quartile vs. depth in the IAB network, for $s_{UDP} = 200$ B.

Figure 8. Buffer occupancy statistics, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.

Figs. 8a and 8b show the median of $\mathbf{B}^{\mathrm{RLC}}$, for traffic flows whose next hop in the network is represented by either UEs or IAB-nodes respectively. The BA strategy achieves the worst performance in this metric, leading to unstable systems in the cases of $s_{UDP} = \{200, 500\}$ B. A reason for this behavior can be found in the "locality" of the BA policy criteria and the lack of influence of the past allocations on the weights. These characteristics may lead to favoring the same link in a repeated manner, hence offering little remedy to the end-to-end congestion. On the other hand, the buffer occupancy achieved by the MSR strategy depicts a system behavior which, in accordance with previous observations, is extremely similar to that of the distributed case. Interestingly, with these configurations the network congestion occurs primarily at the donor and, in general, on the backhaul links towards IAB-nodes. This phenomenon can be explained as follows: even though, on average, the channel qualities of the backhaul links experience a better SINR, the maximum number of such links which can be concurrently activated is limited, due to the TDD configuration. Therefore, the MSR policy may introduce a bias towards the access links instead, since their activation yields the highest sum capacity, despite their lower channel quality. Finally, the MRBA policy achieves the lowest amount of RLC buffering. Specifically, Fig. 8b shows that, compared to the MSR and distributed strategies, the median buffer occupancy among backhaul links is up to 60% smaller, albeit at the cost of slightly more congested UE buffers.

Finally, Fig. 8c depicts the third quartiles of $\mathbf{B}^{\mathrm{RLC}}$ as a function of the depth of the corresponding gNB in the IAB network. It is possible to notice that, regardless of the policy in use,
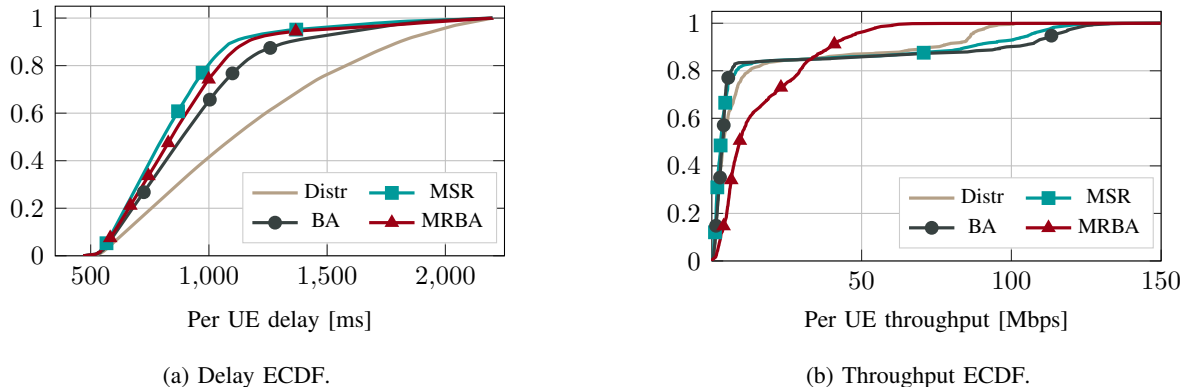
(a) Delay ECDF.

(b) Throughput ECDF.

Figure 9. End-to-end delay and throughput statistics, for TCP layer 4 protocol.

the amount of buffering at the various gNBs generally decreases as their distance to the donor increases. This follows from the fact that nodes which have a lower depth exhibit, on average, a bigger subtending tree; therefore the amount of traffic which makes use of their backhaul links is significantly higher.

E. Performance with TCP traffic

This subsection extends the aforementioned analysis by inspecting the performance of the proposed scheme in the case of TCP traffic. Specifically, a TCP full-buffer source model is used, and the various centralized resource allocation policies are compared against the distributed scheduler.

Fig. 9a shows the ECDF of the end-to-end delay experienced by the successfully received packets. Similarly to the UDP case, the distributed scheduler exhibits the worse performance in this regard. However, the behavior of the centralized policies is remarkably different. In particular, with this configuration the MSR policy provides the best results, followed quite closely by the MRBA and BA strategies. Fig. 9b, which depicts the statistics of the end-to-end throughput achieved by the various UEs, helps explain these results. The net effect of the BA and MSR policies is, approximately, a 20% increase of the peak throughput. Conversely, the MRBA strategy causes a redistribution of the achieved data rate, massively improving the lower quartiles (up to the 80-th), albeit at the expense of the maximum throughput.

Therefore, we can conclude that regardless of the specific policies used, the proposed scheme improves the system performance by limiting the insurgence of local buffering, aiding the end-to-end congestion control mechanism offered by TCP. Furthermore, it can be noted that both a
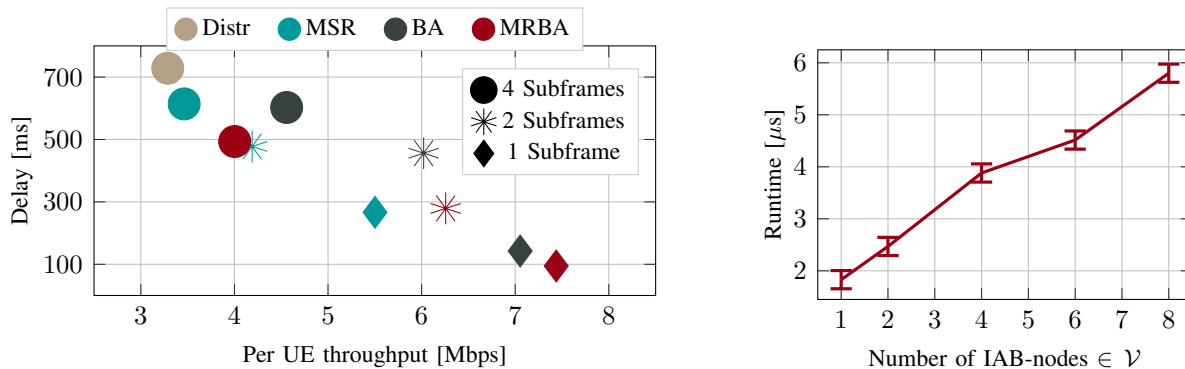
(a) Combined per UE end-to-end throughput first quartile and end-to-end delay third quartile, as a function of the centralized allocation period T_{alloc}.

(b) MWM runtime as a function of the number of IAB-nodes in the network.

Figure 10. Considerations on the formulated assumptions.

prioritization of the most congested links and of the channels featuring a higher quality results in performance benefits in the average case, although it also causes a decrease of the network fairness. On the other hand, the MRBA policy manages to optimize the backhaul/access resource partitioning, while introducing an increase in the throughput fairness at the same time.

F. Further considerations

It is of particular relevance to analyze the performance of the centralized policies when relaxing the most restrictive hypothesis, i.e., the capability of reliably exchanging feedback information in a timely manner, and to understand how restrictive such assumption actually is. To such end, Fig. 10a shows the performance of the proposed framework as a function of the centralized allocation period T_{alloc}. In particular, each of the depicted points represents the joint end-to-end throughput and delay achieved with the different configurations. As expected, in general the effectiveness of the various centralized policies progressively deteriorates as the frequency of the scheduling indications decreases. Interestingly, the BA policy exhibits the lowest performance degradation with respect to an increase of the allocation period, which suggests that this phenomenon has a slower evolution over time compared to the one exhibited by the channels quality. Nevertheless, the key takeaway is that all of the proposed allocation strategies outperform the distributed solution, across both metrics. However, the trend depicted by Fig. 10a also suggests that there exists a threshold value of T_{alloc} after which the performance of the proposed frameworks brings only marginal performance benefits.

Additionally, the running time of the MWM algorithm presented in Sec. III-A was analyzed, in order to understand whether it may partially invalidate the timely feedback assumption. Specifically, Fig. 10b presents the statistics of the various MWM execution times, obtained on a machine equipped with an i7-6700 4-core processor clocked at 3.4 GHz. The first observation which can be made is that this empirical analysis confirms the previously estimated asymptotic complexity, depicting a running time which exhibits a linear dependence on the number of gNBs in the network. Furthermore, it can be noted that the runtime of the MWM algorithm does not exceed 6 μs, even for a significant number of IAB-nodes connected to the same IAB-donor. As a consequence, we can conclude that the execution times of the centralized allocation process do not pose any threat to the timely feedback assumption, since they are reasonably smaller than the duration of the minimum centralized allocation period.

V. CONCLUSIONS

In this paper we proposed a centralized resource partitioning scheme for 5G and beyond IAB networks, coupled with a set of allocation policies. We showed that the introduction of this light resource allocation cooperation dramatically improves the end-to-end throughput and delay achieved by the system already, preventing (or at the very least limiting) the insurgence of network congestion on the backhaul links. Specifically, the MRBA policy exhibits the most promising results, offering up to a 5-fold increase in the worst-case throughput and approximately a 50% smaller worst-case latency, compared to the distributed scheduler. On the other hand, the effectiveness of the BA and MSR policies varies quite significantly across the specific system configuration and inspected metric.

We provided considerations on the implementation of a centralized resource allocation controller in real world deployments. In particular, we acknowledged that the proposed scheme relies on the assumption of IAB-nodes being capable of exchanging timely feedback information with the IAB-donor. Even though the amount of signaling data which the proposed solution requires is quite low, and its performance is quite robust with respect to an increase of the central allocation period, we argue that this remains a significant constraint. Moreover, such drawback is exacerbated by the unfavorable mmWaves propagation characteristics. As a consequence, we deem that centralized solutions, which rely on timely exchange of control information with the IAB-donor, are likely to require dedicated control channels, possibly at sub-6 GHz, in order to grant the utmost priority and reliability to the feedback information. Therefore, we can conclude

that the aforementioned framework can bring dramatic performance benefits to IAB networks, although its introduction in 5G and beyond deployments requires additional research efforts.

For this reason, as part of our future work we plan to design machine-learning algorithms which predict the network evolution at the IAB-donor. This improvement will allow us to relax the timely feedback assumption, by increasing the minimum centralized allocation period which leads to performance benefits over distributed strategies. Moreover, we foresee to implement mechanisms which adapt the parameters of the MRBA policy to the system load and configuration, and additional resource partitioning strategies. Finally, the generalization of the proposed framework to SDMA systems will be studied. The use of such multiple access scheme should significantly improve the performance of mmWave wireless backhauling by introducing the possibility of concurrently serving multiple terminals, provided that they exhibit a sufficient distance among them.

REFERENCES

[1] ITU-R, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," Sep. 2015.

[2] 3GPP, "NR; NR and NG-RAN Overall description; Stage-2," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.300, Jul. 2020, v16.2.0.

[3] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.

[4] F. Khan and Z. Pi, "mmWave mobile broadband (MMB): Unleashing the 3–300 GHz spectrum," in *34th IEEE Sarnoff Symposium*, Princeton, NJ, USA, 2011.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[6] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-Wave Communications: Physical Channel Models, Design Considerations, Antenna Constructions, and Link-Budget," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 870–913, Dec. 2017.

[7] F. Gómez-Cuba, E. Erkip, S. Rangan, and F. J. González-Castaño, "Capacity Scaling of Cellular Networks: Impact of Bandwidth, Infrastructure Density and Number of Antennas," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 652–666, Nov. 2017.

[8] 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 Communications Magazine*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[9] 3GPP, "NR; Study on integrated access and backhaul," Technical Specification (TS) 38.874, Jan. 2019, v.16.0.0.

[10] C. Saha and H. S. Dhillon, "Millimeter Wave Integrated Access and Backhaul in 5G: Performance Analysis and Design Insights," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2669–2684, Dec. 2019.

[11] M. Polese, M. Giordani, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "End-to-End Simulation of Integrated Access and Backhaul at mmWaves," in *IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Barcelona, Spain, 2018.

[12] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-End Simulation of 5G mmWave Networks," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2237–2263, Third quarter 2018.

[13] V. Gambiroza, B. Sadeghi, and E. W. Knightly, "End-to-end performance and fairness in multihop wireless backhaul networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MOBICOM)*, Philadelphia, PA, USA, 2004, pp. 287–301.

[14] M. N. Islam, S. Subramanian, and A. Sampath, "Integrated Access Backhaul in Millimeter Wave Networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, USA, 2017, pp. 1–6.

[15] M. N. Islam, N. Abedini, G. Hampel, S. Subramanian, and J. Li, "Investigation of performance in integrated access and backhaul networks," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Honolulu, HI, USA, 2018.

[16] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Max-min rates in self-backhauled millimeter wave cellular networks," *arXiv preprint arXiv:1805.01040*, 2018.

[17] W. Lei, Y. Ye, and M. Xiao, "Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 970–979, September 2020.

[18] M. E. Rasekh, D. Guo, and U. Madhow, "Interference-aware routing and spectrum allocation for millimeter wave backhaul in urban picocells," in *53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, 2015.

[19] M. Bilal, M. Kang, S. C. Shah, and S.-G. Kang, "Time-Slotted Scheduling Schemes for Multi-hop Concurrent Transmission in WPANs with Directional Antenna," *ETRI Journal*, vol. 36, no. 3, pp. 374–384, Jun. 2014.

[20] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," in *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, San Francisco, CA, USA, 2003.

[21] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

[22] R. Singh and P. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 127–142, Jan. 2019.

[23] B. Ji, C. Joo, and N. Shroff, "Throughput-optimal scheduling in multihop wireless networks without per-flow information," *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 634–647, Apr. 2013.

[24] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, and M. Latva-Aho, "Path selection and rate allocation in self-backhauled mmWave networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, 2018.

[25] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castaño, J. C. Burguillo, S. Rangan, and B. Lorenzo, "On the analysis of scheduling in dynamic duplex multihop mmWave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6028–6042, Nov. 2015.

[26] F. Gomez-Cuba and M. Zorzi, "Optimal link scheduling in millimeter wave multi-hop networks with space division multiple access," in *2016 Information Theory and Applications Workshop (ITA)*, La Jolla, CA, USA, 2016.

[27] F. Gómez-Cuba and M. Zorzi, "Optimal Link Scheduling in Millimeter Wave Multi-hop Networks with MU-MIMO radios." *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1839–1854, Mar. 2020.

[28] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," in *29th IEEE Conference on Decision and Control*, Honolulu, HI, USA, 1990.

[29] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[30] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, Apr. 1998.

[31] M. Polese, M. Giordani, A. Roy, D. Castor, and M. Zorzi, "Distributed Path Selection Strategies for Integrated Access and Backhaul at mmWaves," in *IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, 2018.

[32] 3GPP, "NR; Integrated Access and Backhaul (IAB) radio transmission and reception," Technical Specification (TS) 38.174, Jun. 2020, v0.1.0.

[33] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.

[34] B. Korte and J. Vygen, *Combinatorial Optimization*. Springer Berlin Heidelberg, 2002.

[35] H. N. Gabow, "Data structures for weighted matching and nearest common ancestors with linking," in *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, USA, 1990.

[36] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Computer Networks (COMNET)*, vol. 182, Aug. 2020.

[37] 3GPP, "NR; Medium Access Control (MAC) protocol specification," Technical Specification (TS) 38.321, Jul. 2020, v16.1.0.

[38] ——, "NR; Radio Resource Control (RRC); Protocol specification," Technical Specification (TS) 38.331, Apr. 2020, v16.0.0.

[39] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An open source product-oriented LTE network simulator based on ns-3," in *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, Miami, Florida, USA, 2011.

[40] T. Zugno, M. Polese, N. Patriciello, B. Bojović, S. Lagen, and M. Zorzi, "Implementation of a spatial channel model for ns-3," in *Proceedings of the 2020 Workshop on ns-3*, Gaithersburg, MD, USA, 2020.