

Received 11 September 2025, accepted 4 November 2025, date of publication 10 November 2025,
date of current version 14 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3630984

RESEARCH ARTICLE

User Association Strategies in 5G/6G IAB Networks With Multicast and Unicast Traffic

NATALIA YARKINA¹, DMITRI MOLTCHANOV¹, AND MIKKO VALKAMA¹, (Fellow, IEEE)

Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland

Corresponding author: Natalia Yarkina (natalia.yarkina@tuni.fi)

This work was supported by the Academy of Finland under the Projects “Enabling Mobile Terahertz Communication for 6G Cellular Networks” (EMERGENT) and “Machine Learning Algorithms for Energy Efficient and QoS Aware Communications in Heterogeneous 6G mmWave/sub-THz Networks” (ML6GThz).

ABSTRACT The increasing demand for mobile video and emerging extended/augmented reality services necessitates significant advancements in network capacity. 5G New Radio (NR) and future 6G systems address this with the introduction of high-band, above 24 GHz, spectrum (Frequency Range 2, FR2), but also resource-efficient multicast communications. Another 5G NR innovation, the Integrated Access and Backhaul (IAB) enables wireless self-backhauling and offers a promising solution for cost-efficient network densification needed for FR2. This paper examines the impact of user association in FR2 IAB deployments that concurrently support unicast and multicast traffic. We consider a multi-node IAB network implementing a practical hierarchical codebook-based beamforming and investigate two types of cell and beam selection strategies: those that aim to save backhaul resources and those that favor multicast grouping. Our extensive system-level simulation explicitly accounts for intra- and inter-cell interference, spacial multiplexing, FR2 propagation specifics such as human-body blockages, and the IAB duplexing mode. The study provides valuable insights into how various environmental and system parameters, including the share of multicast traffic, affect the IAB network capacity and the gains from multicast delivery. The results show that in half-duplex FR2 IAB with mixed traffic the most efficient user association policy combines the backhaul-saving strategy with a controlled multicast grouping and can achieve up to a 40 % increase in system capacity.

INDEX TERMS Cell association, integrated access and backhaul, self-backhauled networks, millimeter wave, 5G new radio, multicast, performance evaluation.

I. INTRODUCTION

While the adoption of smartphones keeps growing worldwide, the main contributor to the surge in mobile data traffic is video content – both short-form and long-form – which already accounts for three-thirds of the total traffic [1], [2]. Video streaming is the most popular add-on service among 5G users [2], which on the one hand generates important revenues, but on the other hand, sets out demanding requirements for network capacity. An even further increase of network throughput is needed for the expected advent of such novel services as the extended/augmented reality, which rely on extensive use of visual content [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Heng Foh¹.

In response to this challenge, the 5G New Radio (NR) access technology has incorporated the high-band millimeter-wave (mmWave) spectrum in the range 24–52.6 GHz, referred to as the Frequency Range 2 (FR2). Recently, this band has also been adopted for 6G systems [4]. Widely available, large mmWave bands are capable of offering gigabit-per-second data rates and very low latencies [5], [6]. In addition, when the same content is intended for multiple users, the capacity of a 5G network can be increased even further by means of multicast communications [7]. These can be utilized to deliver linear mobile TV, live events such as sports, video conferencing including education, or public safety content [8]. Although the 3GPP did not specify multicast support in the initial 5G NR releases, the support for resource-efficient group communications has

been thoroughly defined in more recent Releases 17 and 18 [9]. As will be discussed in more detail in Section II, this required the introduction of new delivery functions affecting multiple protocols in the 5G NR stack. Such standardization efforts have been motivated by new use cases and markets for 5G NR technologies including the fixed broadband wireless access [10], public safety use cases [11], and the competition with wired transport network operators for the broadcast content [12].

The physical properties of the mmWave signal propagation [13] greatly affect how FR2 communications are designed and deployed. First, to be efficient they require the use of large antenna arrays and advanced beam management techniques for directional transmission and spatial multiplexing, making the latter a distinctive feature of the full-fledged 5G [14]. Second, the coverage of an FR2 5G site is typically limited by the line-of-sight (LoS) distance between the transmitter and receiver [5]. As a result, in urban areas, mmWave small cells need to be located approximately 200–400 meters apart, which is 5–10 times denser than urban 4G LTE deployments [5], [6]. As a means for simple and cost-effective network densification, the 3GPP has defined the Integrated Access and Backhaul (IAB) technology [15], [16], [17] embedding self-backhauling at the system level. An IAB deployment is a 5G radio access network (RAN) of several nodes, only one of which – the IAB-donor – has a wired link to the core network (CN). The rest relay the traffic of their associated users to/from the IAB-donor over one or more wireless NR-based backhaul hops.

Directional transmissions generally aiming at a single user equipment (UE) and the inability to cover the whole cell with a single transmission make the efficient use of multicast communications in FR2 5G RAN substantially more challenging [7]. On the other hand, the shared multicast delivery can be highly advantageous in wireless backhaul of an IAB deployment [18]. To fully exploit the benefits of multicast delivery over IAB, UEs must be associated to access points taking into account not only the topology of the IAB network, but also the employed transmission mode [18], which is presently not the case [19], [20]. The interplay of such factors as the efficiency of backhaul links, half- or full-duplexing, and the proportion of multicast traffic makes devising a user association policy for FR2 IAB a complex task.

The aim of this paper is to investigate the impact of multicast traffic on user association in half- and full-duplex FR2 IAB networks. To this end, we consider an IAB deployment whose nodes utilize multi-user multiple input multiple output (MU-MIMO) communications with coordinated user pairing and scheduling. Each node is equipped with a large planar antenna array and performs beamforming using the practical hierarchical codebook-based approach. Unlike other studies addressing the user association problem, the comparison of the considered association policies is done by explicitly accounting for intra- and inter-cell interference and the specifics of multicast delivery in access and

backhaul. We assume that the user pairing and scheduling are optimized for deployment capacity maximization, formulate the resulting combinatorial optimization problem as a bin packing problem, and employ a variant of the well-known first fit decreasing heuristic [21] by adopting the signal-to-noise ratio (SNR) as the item size measure. The main metrics of interest are based on the number of users that can be supported by the system as a function of environmental, traffic and system parameters.

The main contributions of the paper are as follows.

- A system-level modeling framework is established for comparing user association strategies in FR2 IAB networks utilizing MU-MIMO and supporting multicast and unicast services.
- *System behavior takeaways:* (i) in FR2 IAB systems with MU-MIMO, multicast delivery can add 20–50 % of capacity both in half- and full-duplex implementations and more than 100 % when the share of multicast users is above 80 %, (ii) the difference in system capacity between no-blockage and heavy-blockage scenarios can reach 60 %, (iii) the use of more than 4–5 simultaneous beams does not provide any additional capacity gain in dense IAB deployments as the system becomes interference-limited, and (iv) in mixed multicast/unicast traffic conditions uniform random user layouts represent an optimistic scenario with capacity generally being 10–15 % better than in clustered layouts.
- *User association takeaways:* (i) in full-duplex IAB, the use of backhaul-aware association policies provides no performance gain compared to association based on the reference signal strength, (ii) in systems with mixed traffic, a higher degree of multicast user grouping does not lead to noticeable capacity gains when MU-MIMO is available, (iii) the use of broader beams for multicast is advantageous only when sweeping 1–2 beams simultaneously (such as in analog beamforming implementations) yielding a capacity gain of 25 % in mixed 50/50 unicast-multicast traffic and over 150–200 % in broadcast scenarios.
- The best capacity performance in a half-duplex IAB system with mixed traffic is achieved by a user association policy that combines the load-aware backhaul-saving strategy based on the achievable rate of path with multicast user grouping.

The remainder of the paper is organized as follows. We begin by providing in Section II a concise overview of multicast support, self-backhauling functionality and user association principles in 5G NR, and discuss the literature related to user association optimization in IAB. The system model is defined in Section III. User association policies are introduced in Section IV, and their comparison methodology is detailed in Section VI. Numerical results are provided and discussed in Section VII. Conclusions are drawn in the last section.

II. BACKGROUND AND RELATED WORK

This section begins with a brief introduction to multicast and IAB technologies, followed by an overview of user association in 5G NR, primarily from the perspective of 3GPP standards. Then, a concise review of the literature on user association optimization in IAB systems is provided.

A. MULTICAST COMMUNICATIONS IN 5G

The 3GPP defines the multicast communication service as the one in which the same specific content data and service are provided simultaneously to a dedicated set of UEs [22]. The inherent support for group communications in 5G systems was introduced in 3GPP Release 17 in the form of Multicast/Broadcast Services (MBS) [9]. The key difference between broadcast and multicast is that the former services are delivered to all UEs in the coverage area of the MBS service, whereas to receive the latter UEs must actively subscribe and join the corresponding session, see Fig. 1.

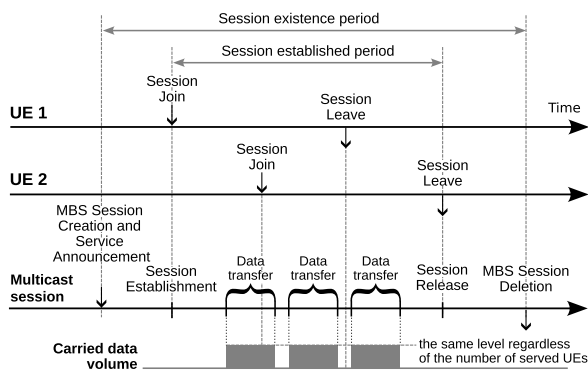


FIGURE 1. An example timeline illustrating phases and events of multicast MBS service provisioning (adapted from 3GPP TS 23.247 [22]).

Multicast traffic is delivered by means of downlink-only multicast MBS sessions, which are the MBS counterpart to protocol data unit (PDU) sessions. A multicast MBS session is characterized by the content to send, the list of UEs authorized to receive the service and optionally by a geographical area it can reach [22]. The resource-efficient *shared delivery* method is employed to transport multicast traffic over the CN. In the shared delivery, only one copy of MBS data packets is transmitted over the involved network links regardless of the number of receivers. The individual, unicast delivery method can still serve multicast traffic to legacy base stations (BSs) not supporting MBS.

In the RAN, multicast data packets can be delivered using the following transmission methods:

- point-to-multipoint (PTM), in which the BS delivers a single copy of MBS data packets to a group of UEs, and
- point-to-point (PTP), in which the BS delivers separate copies of MBS data packets independently to each UE.

MBS-enabled BSs autonomously and dynamically decide whether to use PTM, PTP or their combination based on the number of joined UEs, the MBS session quality-of-service (QoS) requirements, the channel quality and other criteria.

The same QoS requirements apply regardless of the chosen transmission method [22].

An overview of the key physical layer enhancements to support MBS and group scheduling of UEs is provided in [23]. The present state of research into multicast communications in 5G/6G mmWave/sub-THz RANs is comprehensively captured by Chukhno et al. in [7]. Although implementing the PTM delivery in FR2 NR systems is more challenging, mainly because it generally involves adapting the directions and widths of transmission beams to groups of UEs, research into the viability of multicast communications in mmWave NR has shown that they are feasible and can improve performance in many cases of interest [7], [24].

B. INTEGRATED ACCESS AND BACKHAUL

Although one-hop radio relaying was already specified by the 3GPP in LTE, back then the technology did not gain commercial traction. 5G IAB arrives in a very different context: while LTE relaying represented a mere coverage extension option, IAB is viewed as a key enabler for 5G mmWave dense deployments [6] and, along with such technologies as vehicular access and non-terrestrial networks, could facilitate the implementation of more flexible topologies in beyond-5G systems [25].

An IAB RAN uses two types of access points – IAB-donors and IAB-nodes – which appear identical to the UE. According to the standardized IAB architecture [16, Section 4.7], an IAB-donor is a gNodeB (gNB) with additional functionality for IAB support. It combines a gNB central unit (CU) and one or more gNB distributed units (DU), and performs centralized resource, topology and route management for its controlled IAB network using F1 application protocol and/or Radio Resource Control (RRC). An IAB-node implements the gNB-DU functionality to provide access connectivity to UEs and backhaul to its next-hop (children) IAB-nodes. For connecting to the DU of its parent node, which can be another IAB-node or the IAB-donor, an IAB-node implements a subset of UE functionality called the IAB mobile termination (IAB-MT), which terminates the Uu interface of the upstream (towards the donor) backhaul link. To enable packet routing over multiple hops, in the IAB backhaul the IP layer is carried over the IAB-specific Backhaul Adaptation Protocol (BAP) sub-layer [26].

IAB systems can operate in both FR1 (below 7.125 MHz) and FR2 bands and use either the same or different frequencies for access and backhauling [27]. FR2 implementations are likely to prevail due to the amount of spectrum available and high achievable throughput. Spectrum allocations at higher-frequency bands are typically unpaired, meaning that the uplink and downlink communications share a common radio channel. IAB is hence expected to primarily use time-division duplexing (TDD) imposing the half-duplex constraints on IAB network nodes so that they cannot transmit and receive simultaneously [15], [28, Chapter 5]. More

technically challenging full-duplex IAB implementations have been also extensively studied and proved more resource-efficient [29]. For a recent comprehensive overview of the IAB technology and related research see, e.g., [30].

C. USER ASSOCIATION IN 5G NR

In order to receive service in a 5G network, a UE must be assigned a suitable serving cell. This is commonly referred to as cell or user *association* [31]. A cell is uniquely specified by its NR Cell Identity (NCI), which includes the gNB identifier. Besides the primary serving cell responsible for the control plane signaling, RRC and overall anchoring the UE communication with the network, additional cells can be associated with the UE for data transmission using carrier aggregation or multi-connectivity mechanisms [16].

When a UE is switched on, it first performs the initial cell acquisition – *cell selection* [16, Section 9.2]. The UE searches the NR frequency bands and for each possible carrier frequency determines the strongest cell. It then reads the cell system information broadcast to identify its preferred network. A cell is suitable for the UE if it belongs to the preferred or equivalent network, is not barred, reserved or otherwise forbidden, and whose measured attributes satisfy the cell selection criteria as to the received signal strength and quality [16]. Once a suitable cell is found, the UE *camp*s on that cell (i.e., chooses that cell to provide available services and monitors its control channel [19]) and proceeds to *cell re-selection* to identify the best available option. At this stage, the UE may be provided with lists of exclude/allow cells, cell-specific offsets, and especially with absolute priorities as to operating frequencies, supported slices, MBS support, and other parameters. The UE makes measurements of the serving and candidate neighboring cells and ranks them. If upon evaluation a more suitable cell is found, the UE switches to that cell, but only if it has already camped on the current serving cell for more than one second [19], to avoid frequent re-selections.

User association relies on measurements continuously performed by the UE, primarily the reference-signal received power (RSRP) and the reference-signal received quality (RSRQ). Cell (re-)selection is always based on the RSRP and RSRQ measured on cell-defining synchronization signal blocks (SSB). Cell-defining SSBs carry System Information Blocks 1 (SIB1, also called Remaining Minimum System Information, RMSI) with the information required for initial access including the NCI [16]. In 5G NR, different SSBs corresponding to the same cell may be transmitted by different beams (up to 64 in FR2) that span the cell's coverage area. In multi-beam operations, the measurements from the strongest beam or the average over a given number of beams above a threshold are taken as the metrics for the cell.

Cell re-selection is performed by the UE regularly to handle mobility in RRC idle and inactive states, i.e., when there is no ongoing data transmission. The criteria and procedures for cell (re-)selection are specified in 3GPP TS

38.304 [19, Section 5.2]. For a cell to be suitable to camp on, its both RSRP and RSRQ must be greater than their respective minimum values, possibly with added offsets, signaled in SIB1. During cell re-selection, the cells operating in frequencies of equal priorities are ranked based on their RSRPs as follows. The rank of the serving cell is computed as the sum of its RSRP and the hysteresis value. The rank of a neighboring cell is its RSRP minus a possible positive or negative offset between this and the serving cell. Also, a parameter value can be specified so that any cell whose rank is within this range from the highest rank becomes a candidate for selection. In this case, the UE should choose the cell with the greatest number of beams above a given threshold among the candidate cells. An additional offset can be temporarily subtracted from a cell's RSRP and RSRQ in both criteria whenever RRC connection establishment on that cell fails [19].

When the UE is in RRC connected state, its cell association is controlled by the network via the *handover* mechanism, primarily designed to avoid service interruption or noticeable degradation due to UE mobility. Here, more flexibility is allowed as to measurements and criteria compared to cell re-selection. In connected state, besides the RSRP and RSRQ, the UE may also be measuring the signal-to-noise and interference ratio (SINR) on SSBs and all three metrics on channel state information reference signals (CSI-RS). Measurement results are reported to the network periodically or upon condition such as signal quality degradation. Upon evaluating the measurement reports, the network may decide that a handover is desirable and will instruct the UE to switch to another cell. To reduce the reaction time in case of a sudden drop in signal quality, a conditional handover (CHO) can be configured. In CHO, the handover is initiated by the UE, but on conditions and to a candidate cell chosen in advance by the network. It is worth adding that handover, redirection upon RRC release and the use of inter-frequency absolute priorities and offset parameters constitute the mechanisms of load balancing in 5G NR [16].

D. USER ASSOCIATION OPTIMIZATION FOR IAB

As seen in the previous subsection, the approach to user association has not evolved substantially from LTE to 5G NR and still largely relies on the reference signal metrics, in particular the RSRP. However, such 5G features as the use of FR2, massive MIMO, heterogeneous networks (HetNets) and self-backhauling make the conventional approach to cell association less efficient for IAB systems [32]. Furthermore, in IAB networks, user association is closely related to topology establishment, and cell selection based on the access link quality alone may result in an unbalanced topology and reduced performance [17], [20].

3GPP address cell association in IAB starting from Release 16, and extend it to mobile IAB-nodes in Release 18, however most considerations concern related signaling and not the underlying principles. IAB-MTs can only be associated to

cells for which IAB support is specifically indicated in SIB1. Association of IAB-MTs to mobile IAB-nodes is not allowed. Prioritization mechanisms can be used to favor association of users traveling in vehicles to the onboard mobile IAB-nodes. For the rest, cell association in an IAB network follows the standard 5G NR criteria and procedures.

The novel features of 5G have sparked considerable interest in rethinking user association strategies. A comprehensive survey of early studies [32] shows that first research efforts were largely concentrated on user association in HetNets with a particular focus on inter-tier load balancing [32], [33]. The employed frameworks included combinatorial optimization, game theory, stochastic geometry and Markov decision processes. At that point, bottlenecks at the wireless backhaul were identified as an emerging major issue [32], [33].

Recently, user association in multi- and single-tier self-backhauled networks has often been treated as one component of a more general problem of resource allocation optimization, coupled with transmission scheduling [34], [35], power allocation [36], [37], beam selection [38] or beamforming design [39]. Optimization targets have included the number of served UEs [34], spectrum efficiency including weighted and/or logarithmic sum rates [38], [39], and energy efficiency [36], [37]. The resulting optimization problems have been either decomposed to obtain practical suboptimal solution algorithms [34], or tackled with machine learning [38].

For instance, Wang et al. [34] considered a half-duplex IAB deployment in which multiple IAB-nodes provided mmWave access to UEs and utilized one-hop terahertz backhaul to the IAB-donor, whereas the IAB-donor provided backhaul connectivity only. A joint problem of user association and transmission scheduling was formulated to maximize the number of served UEs, i.e., the number of UEs whose data rates per scheduling period in both access and backhaul satisfied their individual requirements. Due to the complexity of the problem, the authors proposed a heuristic algorithm for obtaining a suboptimal solution. The algorithm performed user association based on the minimum ratio between the data rate required by the user and the one achievable in access. According to the reported simulation results, the proposed algorithm delivered up to 33 % gain over RSRP-based association in terms of served users. However, the algorithm inherently prioritized UEs with lower requirements, and the respective gain in system throughput was under 3 %. Similar HetNet-type deployments characterized by one-hop wireless backhaul were studied by many other researchers [36], [38].

Other recent studies considered specifically user association in multi-hop IAB and adopted a more conservative 3GPP-consistent perspective. Ranjan et al. [20] proposed several IAB-specific cell selection policies applicable to both UE and IAB-MT association. The researchers assumed the spanning-tree IAB topology and demanded that the IAB network nodes broadcast the information as to (i) the quality of the backhaul link(s) connecting them to the IAB-donor, (ii)

their hopcount to the IAB-donor for latency estimation, and (iii) their current load in terms of the number of associated IAB-nodes and UEs. The proposed policies relied upon this information and were formulated in terms of either a biased RSRP or estimated characteristics of achievable access-only or full-path data rates. In simulation, the proposed policies outperformed the RSRP rule in terms of cell-edge throughput, hopcount and load balance across the network, although the average UE throughputs could slightly degrade.

Baek et al. [40] considered a multi-node spanning-tree IAB topology operating in half-duplex TDD mode and devised a user association framework to minimize the end-to-end packet latency. The framework leveraged queueing theory and offered suboptimal algorithms for offline (simultaneous) and online (sequential) UE association. According to the reported simulation results, the proposed methods outperformed multiple benchmarks in terms of the 98th delay percentile and demonstrated a 25 % gain over the RSRP-based association. It should be noted, however, that the study only considered downstream transmission, i.e., from the IAB-donor to UEs, whereas delay-oriented control in half-duplex multi-hop IAB is particularly challenging due to the need to accommodate traffic in both directions [41].

All the studies discussed above considered unicast traffic only. Multicast communications in self-backhauled networks have hardly received the due attention from researchers so far, although both features must be supported by the 5G system [42] and can substantially impact control efficiency. Studies [43], [44] dealt with the coexistence of unicast and multicast traffic in deployments with capacity-limited backhaul, however resources of only one network tier were explicitly considered and UEs were not differentiated by their service type. Our previous work [18] explored the coexistence of unicast and multicast traffic in a one-hop half-duplex IAB deployment with the objective of evaluating potential performance gains from the shared delivery and multicast and backhaul-aware user association. Performance of a system with user association and multicast grouping optimized for energy efficiency was compared with the one serving unicast traffic only and the one with user association based on RSRP. The study showed that unicast and multicast UEs are characterized by different optimal associations in IAB, and multicast delivery permits improving the system's capacity by saving up to 40 % of resources. However, no practical multicast- and backhaul-aware association policy was proposed to fully realize the potential gains. The present paper is aimed at filling this gap.

III. SYSTEM MODEL

This section introduces the system model and modeling assumptions regarding the deployment, traffic, antennas, propagation, blockage, data rate estimation, and cell and beam selection. Following this, we present the metrics of interest. The main notations are summarized in Table 1.

TABLE 1. Main notations.

Notation	Description
$\gamma(k)$	Beam assigned to UE k
θ	Elevation (vertical) angle
λ_B	Density of human blockers
π_M	Probability for a UE to be of multicast type
φ	Azimuth (horizontal) angle
$a(k)$	Association point of UE k
B	Cell/beam selection offset (bias)
D	Requested data rate
f_{BH}	Backhaul capacity factor
f_c	Operating frequency
$G(\cdot)$	Antenna gain
g_M	Gain from multicast delivery
g_X	Gain from association policy X over the RSRP rule
K	Size of UE population, $\mathcal{K} = \{1, 2, \dots, K\}$
\hat{K}	Maximum number of served UEs
\mathcal{K}_X	Subset of UEs specified by X
\hat{L}	Maximum number of beams (layers) per AAU
\mathcal{L}_X	Subset of beams specified by X
L_A	Aggregated losses' coefficient
L_{PL}	Path loss
N_0	Thermal noise spectral density
$N_v \times N_h$	AAU UPA dimensions
\hat{P}	AAU power budget
$P_{i,j}^{Rx}$	Received power from directed transmission from i to j
R	Cell radius
r	Flat (2D) distance between transmitter and receiver
\mathcal{S}	Service area
$S_{i,j}$	SINR at j when receiving transmission directed from i
T	Planning time horizon
ΔT	Time slot duration
$u(k)$	Communication type of UE k , 1 if unicast, 0 if multicast
W	Bandwidth

A. DEPLOYMENT AND GENERAL ASSUMPTIONS

We consider the downstream communication in an IAB network deployed in a busy flat outdoor area characterized by line-of-sight (LoS) conditions, such as a city square, campus park or festival grounds. The deployment consists of an IAB-donor and two IAB-nodes, each equipped with 120-degree sectorized active antenna units (AAU). The donor and nodes are stationary and the topology of the IAB deployment is fixed, two backhaul links connecting the donor to the nodes. We focus on the sectors that provide backhaul communication and consider the service area \mathcal{S} formed by three intersecting sectors as shown in Fig. 2. For simplicity, the donor and nodes are assumed identical in terms of antennas' characteristics and power budgets.

The network operates in the mmWave spectrum with bandwidth W and utilizes orthogonal frequency-division multiple access (OFDMA) and the MU-MIMO capability, enabling resource sharing in time, frequency and spacial domains. Centralized resource allocation performed by a donor-located controller is assumed. The deployments features in-band backhauling and supports unicast and resource-efficient multicast delivery in both access and backhaul.

The antennas operate in digital or hybrid beamforming mode and can simultaneously steer up to \hat{L} beams to provide

spacial multiplexing. The actual number of simultaneous beams is limited by the AAU power budget \hat{P} , which is equally divided among active MIMO layers. A layer can serve a group of one or more UEs. Multicast traffic is delivered to the group via a PTM transmission using the same frequency resources with the modulation and coding scheme (MCS) corresponding to the lowest SINR among the multicast UEs of the group. The beam allocation is constant within one time slot, which duration ΔT corresponds to one transmission time interval (TTI).

The data destined to UEs associated with an IAB-node must be transmitted to this node by the IAB-donor over a wireless backhaul link. Backhaul communication is ensured by specific beams directed at the nodes. Shared delivery is used in backhaul for multicast traffic, i.e., only one replica of data is transmitted over the backhaul link to all multicast UEs associated with the node. When operating in half-duplex mode, IAB-nodes cannot transmit data to their UEs while receiving over backhaul.

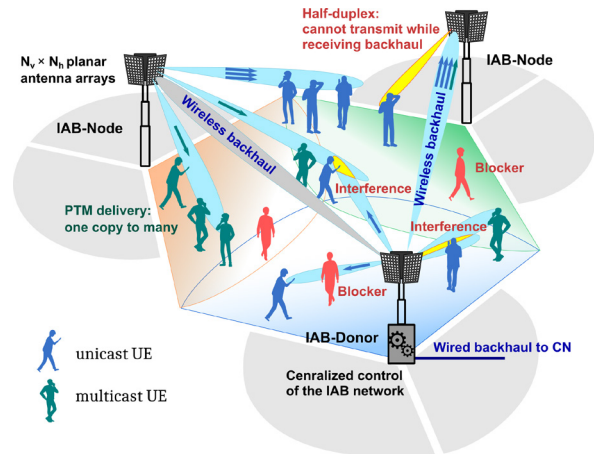


FIGURE 2. An illustration of the system model: an IAB deployment with unicast and multicast communications in half-duplex mode.

B. USERS AND TRAFFIC

A population of K UEs are independently scattered over the service area $\mathcal{S} \subset \mathbb{R}^2$, $\mathcal{K} = \{1, \dots, K\}$. We model their locations using the three spatial point processes (PP) [45], [46] illustrated in Fig. 3, namely:

- a binomial PP of K points (Fig. 3a); we recall that in this case the UEs are uniformly distributed in \mathcal{S} , and such a process can be regarded as a restriction to \mathcal{S} of a homogeneous spatial Poisson PP under the condition that the service area contains exactly K UEs [46],
- a Poisson cluster process 20-5 with cluster density $\lambda_{cl} = 20/|\mathcal{S}|$, $|\mathcal{S}|$ denoting the area of \mathcal{S} , and cluster radius $R_{cl} = 5$ m (Fig. 3b), and
- a Poisson cluster process 5-50 with cluster density $\lambda_{cl} = 5/|\mathcal{S}|$ and cluster radius $R_{cl} = 50$ m (Fig. 3c).

We assume that all UEs receive a guaranteed bit rate (GBR) service, such as an ultra-high-definition (UHD) streaming

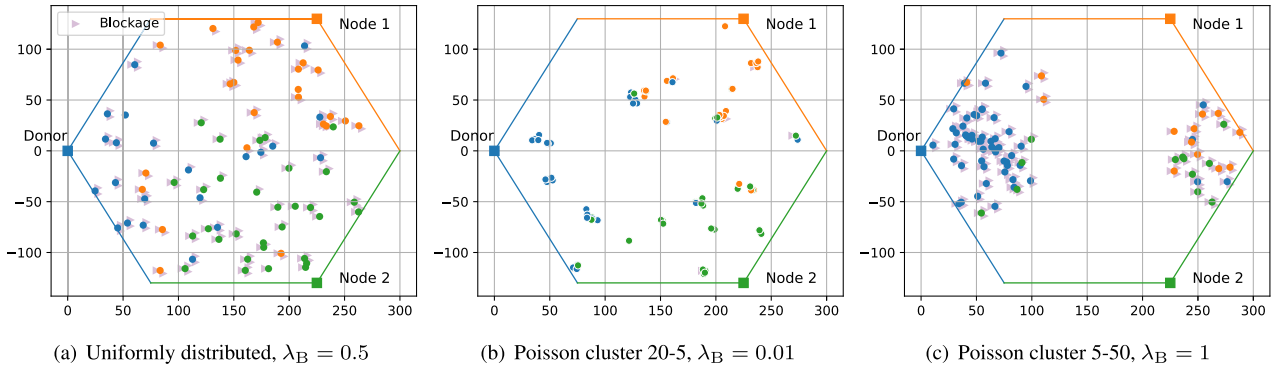


FIGURE 3. UE population samples with RSRP-based user association (indicated by color) under moderate (a), light (b) and heavy (c) blockage conditions. A UE may be associated to a non-nearest node due to blockage, shadow fading or straddling loss – the loss of beamforming gain in directions between the beams.

video, with a requested data rate D . A UE is served either via unicast or multicast communication. Let π_M represent the probability for a UE to be subscribed to the multicast service and consume the same data stream as the other multicast UEs in the area. Such a multicast service can represent, e.g., a popular live event streaming. UEs that are not subscribed to the multicast service receive their individual data streams via unicast communication. In a UE population sample, $u(k) = 1$ if UE k is unicast, and $u(k) = 0$ if it is multicast, $k = 1, \dots, K$.

C. ANTENNAS AND BEAMFORMING

The IAB-donor and nodes are equipped with AAUs utilizing $N_v \times N_h$ uniform planar arrays (UPA) whose elements are half-wavelength spaced vertically and horizontally. In line with [47, Chapter 4], we estimate the power gain pattern of an AAU’s antenna as

$$G(\theta, \varphi) = |\mathbf{a}^T(\theta, \varphi)\mathbf{w}|^2, \tag{1}$$

where \mathbf{w} represents the beamforming weight vector and $\mathbf{a}(\theta, \varphi)$ the array response vector. The entries of the latter are given, for $m = 0, \dots, N_v - 1$ and $n = 0, \dots, N_h - 1$, by

$$a_{1+m+nN_v}(\theta, \varphi) = g(\theta, \varphi)e^{j\pi m \cos \theta} e^{j\pi n \sin \theta \sin \varphi}, \tag{2}$$

where $g(\theta, \varphi)$ is the complex amplitude pattern for a single antenna element. The direction is specified by elevation and azimuth angles θ and φ in an antenna-fixed right-handed spherical coordinate system with array boresight directed towards $(\theta, \varphi) = (90^\circ, 0)$.

We assume the hierarchical codebook approach to beamforming and let the service area of each sector be covered with two grids of beams (GoB) – a grid of broader beams (GoBB) and a grid of narrow (refined) beams (GoNB) – as shown in Fig. 4. The hierarchical structure of the employed GoB is shown in Fig. 4: each broad beam has a corresponding subset of narrow beams together serving roughly the same area. Both GoBs are constructed using beamforming weights chosen among the NO_s columns of an oversampled discrete

Fourier transform (DFT) matrix \mathbf{Q} with entries [47, Chapter 6]

$$Q_{m,n} = \frac{1}{\sqrt{N}} e^{j\frac{2\pi(m-1)(n-1)}{O_s N}}, \tag{3}$$

where O_s is the oversampling factor and $N \in \{N_v, N_h\}$. For GoBB, in addition, we use a simple beam broadening method suggested in [48]. The method consists in multiplying the weight of element $n = 0, \dots, N - 1$ of an N -element uniform linear array by a factor $e^{jf_n(p,c)}$, where

$$f_n(p, c) = \left| 4\pi c \left(\frac{1}{2(N-1)} + \frac{n - \frac{N}{2}}{N-1} \right)^p \right|, \tag{4}$$

where the parameters p and c permit varying the beam shape.

We assume that the IAB-donor uses specific narrow beams directed at the IAB-nodes for backhaul communication.

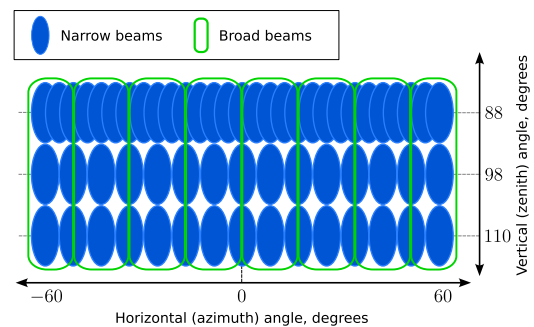
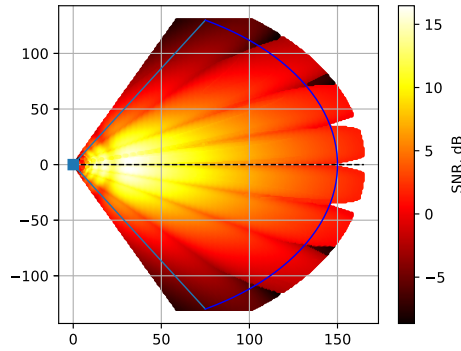


FIGURE 4. The structure of the assumed two-level GoB. Broad beams form one row of eight, narrow (refined) beams form three rows of 29-15-15. Horizontal beam steering is performed using (3). Vertically, the broad beams are directed at boresight (90°), the rows of narrow beams at 88°, 98° and 110°.

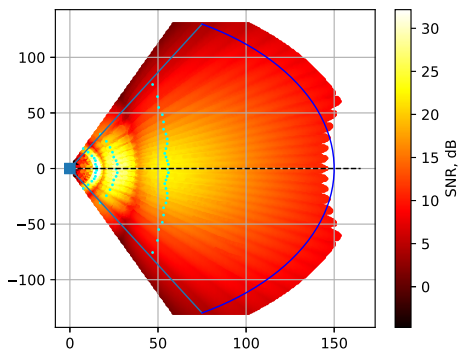
D. PROPAGATION AND BLOCKAGE MODELS

We represent propagation losses using the Urban-Micro (UMi) Street-Canyon model [49], in which the path loss for the frequency band 0.5–100 GHz can be expressed in dB as

$$L_{PL}^{[dB]} = 21 \log_{10} d + 20 \log_{10} f_c + 32.4 + 20b + \chi_{SF}, \tag{5}$$



(a) GoBB



(b) GoNB, cyan dots indicate the maximum SNR footprint for each beam

FIGURE 5. Coverage footprints of the assumed GoB in terms of SNR for a human-blocked UE at the maximum available transmitter power \hat{P} . A mechanical downtilt of 9° is assumed.

where d is the three-dimensional (3D) distance in meters between the transmitter and receiver, f_c is the carrier frequency in gigahertz, $b = 1$ or 0 according as the channel is human-blocked, and χ_{SF} is the shadow fading in decibels, normally distributed with zero mean and standard deviation σ_{SF} .

By converting the path loss (5) to a linear scale, we can write the received signal power from transmitter i to receiver j as

$$P_{i,j}^{Rx} = \frac{P_i^{Tx} G_{i,j}(j) G_{j,i}(i)}{L_{PL}(d_{i,j}, b_{i,j}) L_A}, \quad (6)$$

where P_i^{Tx} is the power emitted by i , $G_{i,j}(j)$ is the antenna gain from a directed transmission $i \rightarrow j$ in the direction of j , and $L_A^{[dB]} = L_C + N_F + L_B$ represents the aggregated losses' coefficient comprising cable losses, noise figure, and beamforming losses. The interference from a simultaneous directed transmission $i' \rightarrow j'$ at a receiver j of transmission $i \rightarrow j$ can be estimated as

$$I_{i,j}^{i',j'} = \frac{P_{i'}^{Tx} G_{i',j'}(j) G_{j,i}(i')}{L_{PL}(d_{i',j'}, b_{i',j'}) L_A}. \quad (7)$$

Here, transmitter i' can represent the same node as i (in which case we talk about intra-cell interference) or another node (inter-cell interference). Now, the SINR at receiver j can be estimated as

$$S_{i,j} = \frac{P_{i,j}^{Rx}}{N_0 W + \sum_{(i',j') \neq (i,j)} I_{i,j}^{i',j'}}, \quad (8)$$

with N_0 being the thermal noise spectral density, and the sum taken over all transmissions simultaneous to $i \rightarrow j$.

Following [50], we represent human blockers by cylinders of height $h_B = 1.7$ and base radius $R_B = 0.3$ meters and assume them scattered over the service area according to the spatial Poisson process with density λ_B per square meter. The probability of blockage at a 2D distance r between the UE and AAU is then given by [50]

$$p_B(r) = 1 - e^{-2\lambda_B R_B \left[r \frac{h_B - h_{UE}}{\Delta h} + R_B \right]}, \quad (9)$$

with $\Delta h = h_{BS} - h_{UE}$ being the height difference between the AAU and UE. We assume a human-body blockage attenuation of 20 dB [51], as specifies the next-to-last term in (5).

E. DATA RATE ESTIMATION

According to the 3GPP [52], the data rate in Mbps of a single UE can be estimated as

$$10^{-6} \sum_{j=1}^J v_L^{(j)} Q_m^{(j)} f^{(j)} R_c \frac{12 \hat{N}_{PRB}^{W(j),\mu}}{T_{smb1}(\mu)} (1 - w_{oh}^{(j)}), \quad (10)$$

where v_L is the number of multiplexed layers, Q_m the modulation order, f the scaling factor, R_c the coding rate, w_{oh} the overhead, and $\hat{N}_{PRB}^{W,\mu}$ the number of available primary resource blocks (PRB) for the given bandwidth and numerology for each of J aggregated component carriers in a band or band combination, $T_{smb1}(\mu) = 10^{-3} / (14 \times 2^\mu)$ is the symbol duration in numerology μ .

In line with this, letting $J = 1$, $v_L = 1$ and $f = 1$, we will estimate the amount of data that can be transmitted per PRB as

$$12 N_{DL} Q_m R_c (1 - w_{oh}), \quad (11)$$

where $N_{DL} \leq 14$ represents the number of symbols in the slot allocated for downlink transmission, and the parameters Q_m and R_c are specified by the applicable MCS. Because the latter is determined in function of the SINR in a manufacturer-dependent manner, with rely on [53, Table 2] for the 5G NR correspondence.

Additionally, we assume that backhaul links can be more performant than access, e.g., due to a more efficient MCS. Therefore, to evaluate the capacity of backhaul transmissions we multiply the quantity (11) by a factor $f_{BH} \geq 1$.

F. USER ASSOCIATION AND BEAM SELECTION

To receive data, UEs must be assigned a serving AAU and beam. UEs can be associated either to the IAB-donor or

to one of the IAB-nodes. The RSRP-based cell and beam selection is assumed as follows. First, each AAU sweeps its GoBB beams and the UE retains the broad beam that has provided the highest RSRP. The AAU delivering the strongest beam becomes the association point for the UE. Once the serving AAU is chosen, beam refinement occurs. The AAU associated with the UE sweeps its GoNB beams corresponding to the retained broad beam, and the one delivering the strongest signal is selected by the UE as its serving beam.

Let the nodes of the IAB network be indexed so that the IAB-donor is designated as node 0, the IAB-node located in the positive azimuth direction from the donor’s AAU as node 1, and the remaining IAB-node as node 2, see Fig. 6. Then more formally, the baseline RSRP-only association policy makes UE k associate with network node $a(k)$ that satisfies

$$a(k) = \arg \max_{i \in \{0,1,2\}} P_{i,k}^{\text{Rx}}, \quad (12)$$

where $P_{i,k}^{\text{Rx}}$ is obtained by (6) with the strongest GoBB beam, i.e., with $G_{i,k}(k) = \max_{l \in \mathcal{L}_{\text{GoBB}}} G_{i,k}^l(k)$. Hereinafter, \mathcal{L}_X denotes the set of beams in grid X .

Other considered user association and beam selection policies are detailed in Section IV. Note that we use the terms “network node”, “access point” and “association point” interchangeably, provided the meaning remains clear.

G. METRICS OF INTEREST

To evaluate and compare different user association and beam selection policies, we estimate the number of UEs \hat{K} that can be served by the deployment given the ordered UE population \mathcal{K} . UEs are served if they receive their required data rate in both access and backhaul within the considered planning horizon. More specifically, for each ordered UE sample \mathcal{K} , \hat{K} is such that the set of UEs $\{1, 2, \dots, \hat{K}\}$ can be served, whereas $\{1, 2, \dots, \hat{K}, \hat{K} + 1\}$ cannot, assuming that user pairing and scheduling is performed so as to maximize the number of served UEs.

Additionally, we use such metrics as

- the gain from association policy X compared to RSRP-based association

$$g_X = \frac{\hat{K}_X - \hat{K}_{\text{RSRP}}}{\hat{K}_{\text{RSRP}}} \times 100\%, \quad (13)$$

- the gain from multicast delivery compared to a system with unicast service only, i.e., with $\pi_M = 0$,

$$g_M = \frac{\hat{K} - \hat{K}_{\pi_M=0}}{\hat{K}_{\pi_M=0}} \times 100\%, \quad (14)$$

- the efficiency of PTM delivery, computed as the ratio of served multicast UEs to the number of multicast transmissions in access, and
- the PTM use, evaluated as the percentage of multicast UEs served using PTM delivery.

IV. BACKHAUL- AND MULTICAST-AWARE CELL AND BEAM SELECTION

In this study we describe and evaluate several strategies of backhaul- and service-type-aware user association and beam selection. The following specific features of mmWave IAB can be exploited for devising such strategies.

- 1) UEs associated with IAB-nodes necessitate backhaul transmissions, which not only introduce additional delays, but also, if voluminous, may substantially limit the capacity of IAB-nodes to provide access due to half-duplex operation.
- 2) Transmissions to multiple multicast UEs can be resource-efficient in both backhaul and access thanks to the shared and PTM delivery modes. However, for PTM delivery in access the same beam must be selected by a group of multicast UEs.
- 3) Human-body blockages resulting in important SNR drops are characteristic for communications in the mmWave spectrum. An mmWave IAB deployment must be dimensioned to provide connectivity to a UE in blockage conditions, which may result in over-provisioning for unblocked UEs, allowing for a substantially greater flexibility in their association.

In the remainder of this section we present several association strategies taking account of these considerations. The strategies are divided into backhaul-saving, aimed at reducing the workload on the backhaul links, and PTM-boosting, favoring multicast grouping. Finally, a hybrid policy combining the two approaches is proposed.

A. BACKHAUL-SAVING USER ASSOCIATION

The simplest approach to backhaul-saving user association consists in introducing a fixed offset or bias (FB) B favoring an access point with fewer hops to the CN, which is the IAB-donor in our considered deployment. This biased RSRP policy can be formulated as follows.

- 1) FB

$$a(k) = \arg \max_{i \in \{0,1,2\}} (P_{i,k}^{\text{Rx}} + B \times \mathbf{1}_{i=0}), \quad (15)$$

where $\mathbf{1}_X$ denotes the indicator function, and equals 1 if X holds and 0 otherwise.

However, because multicast UEs create less workload on the backhaul links thanks to the shared delivery, we modify the FB policy for a system with both unicast and multicast traffic. The following biased RSRP policy favors association of unicast UEs to the donor and association of multicast UEs to the IAB-nodes by means of the same fixed bias B .

- 2) FB-M

$$a(k) = \arg \max_{i \in \{0,1,2\}} \left(P_{i,k}^{\text{Rx}} + B(u(k)\mathbf{1}_{i=0} + (1 - u(k))\mathbf{1}_{i \neq 0}) \right). \quad (16)$$

Another studied pair of backhaul-reducing policies takes into account the access point workload in terms of the number of associated UEs. These are adaptations of one of the two best performing cell selection policies proposed for IAB networks by Ranjan et al. in [20], namely, the *Achievable rate of path using Scaled Minimum* (ASM).

Assume that UEs are associated sequentially according to their indices in \mathcal{K} . Let h_i denote the number of hops from network node i to the CN. In the considered deployment we have $h_0 = 0$ and $h_1 = h_2 = 1$. Let

$$C_{i,k} = \log_2 \left(1 + \frac{P_{i,k}^{\text{Rx}}}{N_0 W} \right) \quad (17)$$

be the normalized capacity of the link from i to UE k , and

$$n_{i,k} = \sum_{j < k} \mathbf{1}_{a(j)=i} + \mathbf{1}_{i=0} \sum_{m=1,2} \mathbf{1}_{\exists j < k: a(j)=m} \quad (18)$$

represent the number of receivers associated to i before association of UE k (for the IAB-donor it includes the backhaul links to IAB-nodes having associated users). Then,

a: ASM*

$$a(k) = \arg \max_{i \in \{0,1,2\}} \frac{C_{i,k}}{(1 + h_i)(1 + n_{i,k})}, \quad (19)$$

and its counterpart for a system with mixed traffic can be defined as

3) ASM-M

$$a(k) = \arg \max_{i \in \{0,1,2\}} \frac{C_{i,k}}{(1 + u(k)h_i)(1 + n_{i,k})}. \quad (20)$$

We remark that in the original ASM policy the achievable rate of path is estimated as the minimum hop capacity along the route, while $1/(1 + n_{i,k})$ is interpreted as the channel access probability for UE k . The ASM-M policy differs from ASM* in that the penalty for the extra backhaul hop is given to unicast UEs only.

B. CELL AND BEAM SELECTION TO BOOST PTM

In this subsection we propose two policies that deal with the access and, for multicast UEs, favor cell and beam selection so as to facilitate PTM grouping. The policies associate unicast UEs based on RSRP as described previously, whereas for multicast UEs, first, the strongest GoNB beam is determined in the same manner, and then an attempt to switch to a suitable beam serving other multicast UEs is performed. For each UE the set of suitable beams is restricted to those providing the GoNB RSRP no more than B dB weaker than the strongest beam.

The first policy considers all multicast UEs at once (the offline approach) and attempts to associate them using as few beams as possible. It consists of the following steps.

1) PTM-MAX

- 1: For each $k \in \mathcal{K}$ determine the strongest GoNB beam $\hat{\gamma}_k = (i, l)$, $i \in \{0, 1, 2\}$, $l \in \mathcal{L}_{\text{GoNB}}$, and the corresponding received signal strength $\hat{P}_k^{\text{Rx}} = P_{\hat{\gamma}_k, k}^{\text{Rx}}$.
- 2: Assign unicast UEs to their corresponding $\hat{\gamma}_k$.
- 3: For each multicast UEs $k \in \mathcal{K}_M = \{k \in \mathcal{K} : u(k) = 0\}$, find the set of suitable beams $\hat{\mathcal{L}}_k = \{\gamma = (i, l) : P_{\gamma, k}^{\text{Rx}} + B > \hat{P}_k^{\text{Rx}}\}$.
- 4: Let $\mathcal{L} = \cup_{k \in \mathcal{K}_M} \hat{\mathcal{L}}_k$ and find the smallest subset $\hat{\mathcal{L}} \subset \mathcal{L}$ containing a suitable beam for each multicast UE. This can be formulated as a minimum set cover problem [54], [55], in which each beam in \mathcal{L} corresponds to a subset of UEs it is suitable for, i.e., $\mathcal{K}_\gamma = \{k \in \mathcal{K}_M : \gamma \in \hat{\mathcal{L}}_k\}$, and the smallest sub-collection of such sets covering all multicast UEs must be found.
- 5: Assign multicast UE $k \in \mathcal{K}_M$ to a beam in $\hat{\mathcal{L}}$ delivering the strongest RSRP, i.e.,

$$\gamma(k) = \arg \max_{\gamma \in \hat{\mathcal{L}}} P_{\gamma, k}^{\text{Rx}}. \quad (21)$$

The second proposed policy is considerably more lightweight and handles UEs sequentially (the online approach). To associate a multicast UE k , the policy performs the following steps.

2) PTM-SEQ

- 1: Use the hierarchical approach described in Subsection III-F to determine the strongest GoNB beam $\hat{\gamma}_k = (i, l)$, $i \in \{0, 1, 2\}$, $l \in \mathcal{L}_{\text{GoNB}}$, and the corresponding received signal strength $\hat{P}_k^{\text{Rx}} = P_{\hat{\gamma}_k, k}^{\text{Rx}}$.
- 2: Iterate the previously associated multicast UEs and consider their serving beams. If for some multicast UE $m < k$ assigned to beam $\gamma(m)$ the inequality

$$P_{\gamma(m), k}^{\text{Rx}} + B > \hat{P}_k^{\text{Rx}} \quad (22)$$

holds, then assign k to $\gamma(m)$ and stop. Otherwise, assign k to beam $\hat{\gamma}_k$.

Finally, to combine the benefits of backhaul workload reduction and resource-efficient PTM delivery, we propose a hybrid sequential policy that handles UE k as follows.

3) PTM+ASM

If k is unicast, use ASM*, otherwise use PTM-seq.

V. EVALUATION FRAMEWORK

In this section we detail the modeling framework employed to compare the performance of the association strategies.

A. COORDINATE SYSTEMS AND TRANSFORMATIONS

The considered service area is depicted in Fig. 6. We adopt a global coordinate system $Oxyz$ with the origin at the IAB-donor. It is assumed that all UEs belong to the plane Oxy , referred to as the *UE plane*. The x-axis points towards the center of the considered 120-degree sector of the IAB-donor. The z-axis points to zenith and is not shown in Fig. 6.

The size of the service area, which is assumed a regular hexagon, is determined by the cell radius R . In the global coordinate system, the Cartesian coordinates of the IAB-nodes 1 and 2 are $(\frac{3}{2}R, \pm\frac{\sqrt{3}}{2}R, 0)$, and the 3D distance between any two AAUs is $\sqrt{3}R$. As AAUs are assumed located at the same height, this distance represents the inter-site distance denoted by d_{ISD} .

Each IAB-node has its associated coordinate system. These are obtained, for nodes 1 and 2, respectively, via translation of $Oxyz$ by vector $(\frac{3}{2}R, \pm\frac{\sqrt{3}}{2}R, 0)$ and rotation around z -axis by $\mp 120^\circ$. Thus, a point of the UE plane with global coordinates (r, φ) has its coordinates in the node-associated systems of the form $(|C|, \arg(C) \pm 120^\circ)$ with

$$C = re^{j\varphi} - \frac{\sqrt{3}}{2}R(\sqrt{3} \pm j). \quad (23)$$

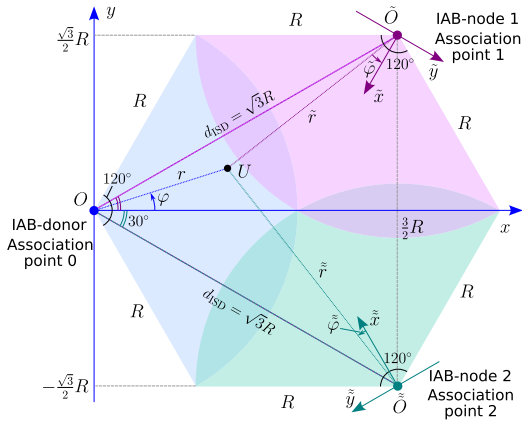


FIGURE 6. The considered service area and notation.

To model transmission beams, we introduce an antenna-fixed coordinate system $O'x'y'z'$ obtained from the global coordinate system $Oxyz$ by shifting along the positive z -axis by $\Delta h = h_{BS} - h_{UE}$ and rotating around the y -axis by an angle of β , which represents a mechanical downtilt, see Fig. 7. A UE plane point U with polar coordinates (r, φ) has, in the antenna-fixed system, spherical coordinates (r', θ', φ') , where

$$r' = \sqrt{\Delta h^2 + r^2} \quad (24)$$

corresponds to the 3D distance between the AAU and a UE placed at point U , while θ' and φ' are, respectively, the zenith and azimuth angles specifying the direction from the AAU UPA towards U . These are given by [49]

$$\begin{aligned} \theta' &= \arccos(\cos \varphi \sin \theta \sin \beta + \cos \theta \cos \beta), \\ \varphi' &= \arg(\cos \varphi \sin \theta \cos \beta - \cos \theta \sin \beta + j \sin \varphi \sin \theta), \end{aligned} \quad (25)$$

with

$$\theta = \pi - \arctan(r/\Delta h) \quad (26)$$

being the zenith angle of point U in a translated from O to O' but not rotated coordinate system (angle $\angle zO'U$ in Fig. 7).

B. ARRAY GAIN

Following [47], we let the $N_v \times N_h$ AAU UPA lay in the $O'y'z'$ -plane with the x' -axis pointing in the front direction of the array, i.e., the boresight direction corresponds to $(\theta', \varphi') = (90^\circ, 0)$, see Fig. 7. Half-wavelength spacing between adjacent elements in either direction is assumed.

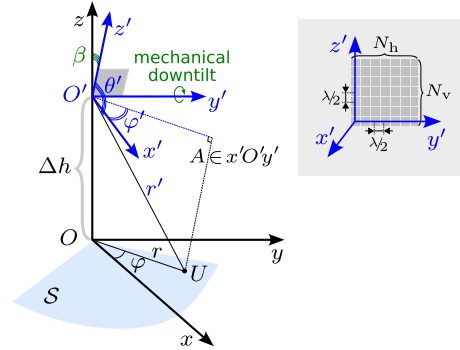


FIGURE 7. The global $Oxyz$ and the antenna-fixed $O'x'y'z'$ coordinate systems.

We compute the array gain in a direction (θ', φ') as [47, Chapter 4]

$$G(\theta', \varphi') = |AF_v(\theta', \varphi')|^2 |AF_h(\theta', \varphi')|^2 \times G_{3GPP}(\theta', \varphi'). \quad (27)$$

Here, $G_{3GPP}(\theta', \varphi')$ is the radiation power pattern of a single antenna element specified by the 3GPP [49], [56] in decibels as

$$G_{3GPP}(\theta', \varphi') = 8 - \min \left\{ \left(\min \left\{ 12 \left(\frac{\theta' - 90^\circ}{65^\circ} \right)^2, 30 \right\} + \min \left\{ 12 \left(\frac{\varphi'}{65^\circ} \right)^2, 30 \right\} \right), 30 \right\}. \quad (28)$$

The quantities $AF_v(\cdot)$ and $AF_h(\cdot)$ represent, respectively, the array factors for an N_v -element vertical and N_h -element horizontal uniform linear arrays. We assume

$$\begin{aligned} AF_v(\theta', \varphi') &= \frac{1}{\sqrt{N_v}} \sum_{m=0}^{N_v-1} e^{j(m\pi(\cos \theta' - \cos \theta'_0) + f_m(p, c))}, \\ AF_h(\theta', \varphi') &= \frac{1}{\sqrt{N_h}} \sum_{n=0}^{N_h-1} e^{j(n\pi \sin \theta' \sin \varphi' + q_n^x)}, \end{aligned} \quad (29)$$

with $f_n(p, c)$ given by (4). For backhaul transmission to the IAB-node located in direction (θ'_0, φ'_0) in the donor's antenna-fixed system, we let $c = 0$ in $f_n(p, c)$ and

$$q_n^{BH} = -n\pi \sin \theta'_0 \sin \varphi'_0, \quad (30)$$

so that $AF_v(\cdot)$ and $AF_h(\cdot)$ above correspond to the use of beamforming weights maximizing the gain in a direction (θ'_0, φ'_0) [47].

For GoB, we use (3) and let

$$q_n^{GoB} = \frac{2\pi nm}{O_s N_h} + f_n(p, c), \quad (31)$$

where $0 \leq m < O_s N_h$ specifies the beam in the grid. In our study, GoBB, Fig. 5a, is obtained with $\theta'_0 = 90^\circ$, $O_s = 1$, $m = 2i + 1$, $i = 0, \dots, 7$, and $(p, c) = (4, 5)$ to provide beam broadening. For GoNB, Fig. 5b, we let $c = 0$, $\theta'_0 \in \{88^\circ, 98^\circ, 110^\circ\}$, $O_s = 2$, $m \in \{0, \dots, 14, 18, \dots, 31\}$ for $\theta'_0 = 88^\circ$, and $O_s = 1$, $m \in \{0, \dots, 7, 9, \dots, 15\}$ for the other values of θ'_0 . We assume that the GoB beams are indexed, the set of indices denoted by \mathcal{L} .

C. SPECTRAL EFFICIENCY

Let \mathcal{K}_t^A and \mathcal{K}_t^B represent, respectively, the sets of UEs whose access and backhaul transmissions are schedule at time slot $t \in \{1, 2, \dots, T\}$. Let $\mathcal{L}_i(\mathcal{K}_t^A)$ denote the set of (different) GoB beams swept by AAU $i \in \{0, 1, 2\}$ to serve UEs in \mathcal{K}_t^A . Taking into account the backhaul beams employed by the IAB-donor to serve transmissions in \mathcal{K}_t^B , the number of simultaneous beams swept by i at t can be written as

$$L_i(t) = |\mathcal{L}_i(\mathcal{K}_t^A)| + \mathbf{1}_{i=0} \sum_{j \in \{1, 2\}} \mathbf{1}_{\exists k \in \mathcal{K}_t^B: a(k)=j} \quad (32)$$

It is assumed that the power budget of an AAU is equally partitioned among its active beams, hence, the per-beam power of AAU $i \in \{0, 1, 2\}$ at time slot t is

$$P_i^{\text{Tx}}(t) = \frac{\hat{P}}{L_i(t)}. \quad (33)$$

Let $G_{i,l}(\mathbf{x})$ denote the array gain for a beam $l \in \mathcal{L}$ swept by AAU i at a point with global coordinates \mathbf{x} . This quantity is obtained by (27) after the coordinate transformation described previously. We use (5) to estimate the path loss $L_{\text{BH},i,j}^{\text{PL}}$ of a backhaul link between AAUs i and j and the path loss $L_{i,k}^{\text{PL}}$ between AAU $i \in \{0, 1, 2\}$ and UE k . Now, the received power for the access transmission to UE $k \in \mathcal{K}_t^A$ at t can be written as

$$P_k^{\text{Rx}}(t) = \frac{P_{a(k)}^{\text{Tx}}(t) G_{a(k),\gamma(k)}(\mathbf{x}_k) G_{\text{UE}}}{L_{a(k),k}^{\text{PL}} L_A}, \quad (34)$$

where \mathbf{x}_k denotes the global coordinates of UE k , $a(k)$ its association point, and $\gamma(k)$ its serving beam.

The interference experienced by UE k includes that from simultaneous access and backhaul transmissions. The former can be written as

$$I_k^A(t) = \sum_{i \in \{0, 1, 2\}} \sum_{\substack{l \in \mathcal{L}_i(\mathcal{K}_t^A) \\ l \neq \gamma(k)}} \frac{P_i^{\text{Tx}}(t) G_{i,l}(\mathbf{x}_k) G_{\text{UE}}^{\mathbf{1}_{a(k)=i}}}{L_{i,k}^{\text{PL}} L_A}. \quad (35)$$

For the latter we have

$$I_k^B(t) = \sum_{\substack{i \in \{1, 2\} \\ \exists k \in \mathcal{K}_t^B: a(k)=i}} \frac{P_0^{\text{Tx}}(t) G_i^{\text{BH}}(\mathbf{x}_k) G_{\text{UE}}^{\mathbf{1}_{a(k)=0}}}{L_{0,k}^{\text{PL}} L_A}, \quad (36)$$

where $G_{\text{BH},i}(\mathbf{x})$ denotes the array gain for the backhaul beam directed at IAB-node i at a point with global coordinates \mathbf{x} .

For the backhaul transmission to IAB-node $i \in \{1, 2\}$, the received power, given the pairing $(\mathcal{K}_t^A, \mathcal{K}_t^B)$, is

$$P_{\text{BH},i}^{\text{Rx}}(t) = \frac{P_0^{\text{Tx}}(t) G_{\text{BH},i}^2(\mathbf{y}_i)}{L_{\text{BH},0,i}^{\text{PL}} L_A}, \quad (37)$$

and the interference from simultaneous access and backhaul transmissions, respectively, is

$$I_{\text{BH},i}^A(t) = \sum_{j \in \{0, 3-i\}} \sum_{l \in \mathcal{L}_j(\mathcal{K}_t^A)} \frac{P_j^{\text{Tx}}(t) G_{j,l}(\mathbf{y}_i) G_{\text{BH},i}^{\mathbf{1}_{j=0}}(\mathbf{y}_i)}{L_{\text{BH},j,i}^{\text{PL}} L_A}, \quad (38)$$

and

$$I_{\text{BH},i}^B(t) = \mathbf{1}_{\substack{\exists k \in \mathcal{K}_t^B: \\ a(k)=3-i}} \frac{P_0^{\text{Tx}}(t) G_{\text{BH},3-i}(\mathbf{y}_i) G_{\text{BH},i}(\mathbf{y}_i)}{L_{\text{BH},0,i}^{\text{PL}} L_A}, \quad (39)$$

where \mathbf{y}_i represents the coordinates of node i .

Having expressions for the received power and interference, we can use (8) to compute SINRs $S_k(t)$ and $S_{\text{BH},i}(t)$, respectively, for access and backhaul transmissions under pairing $(\mathcal{K}_t^A, \mathcal{K}_t^B)$ and establish the corresponding spectral efficiencies $\eta_k(t)$ and $\eta_{\text{BH},i}(t)$ as $Q_m R_c$ using, e.g., [53, Table 2].

D. USER PAIRING AND SCHEDULING

In this study, we assume that user pairing and scheduling are optimized so as to maximize the number of served UEs in the planning horizon of T time slots. A donor-associated UE is assumed served if its access transmission is allocated resources within horizon T to carry $\Delta D = DT \Delta T$ of data. In the case of multicast, these resources can be shared with other transmissions. A node-associated UE is served if within the horizon T resources are allocated for its access and backhaul transmissions, each carrying ΔD of data. For simplicity, we assume that within the horizon T no more than one time slot can be used for serving each UE in access.

For transmissions to be scheduled in the same time slot, the half-duplex (when operating in half-duplex mode), simultaneous layers and data-rate constraints must be satisfied. The *half-duplex constraint* forbids an IAB-node to receive backhaul and transmit access simultaneously. Therefore, in half-duplex mode we demand that, for any $t \in \{1, 2, \dots, T\}$ and $i = 1, 2$,

$$\mathbf{1}_{\exists k \in \mathcal{K}_t^B: a(k)=i} + \mathbf{1}_{\exists k \in \mathcal{K}_t^A: a(k)=i} < 2. \quad (40)$$

The above constraint is released in full-duplex mode.

The *simultaneous layers constraint* demands that the number of beams swept simultaneously by any AAU do not exceed \hat{L} , i.e., for $i \in \{0, 1, 2\}$,

$$L_i(t) \leq \hat{L}. \quad (41)$$

The *data-rate constraint* ensures that each transmission, backhaul and access alike, can carry at least ΔD of data in spite of intra- and inter-cell interference. Let $\mathcal{K}_{i,l}(t) \subset \mathcal{K}_t^A$ denote the set of UEs served by a GoB beam l of AAU i . The

beam group $\mathcal{K}_{i,l}(t)$ may contain multicast and unicast UEs. The number of PRBs required by the multicast UEs is

$$N_{i,l}^M(t) = \left\lceil \frac{\Delta D}{12N_{DL}(1 - w_{oh}) \min_{\substack{k \in \mathcal{K}_{i,l}(t): \\ u(k)=0}} \eta_k(t)} \right\rceil, \quad (42)$$

whereas that required by the unicast is

$$N_{i,l}^U(t) = \sum_{\substack{k \in \mathcal{K}_{i,l}(t): \\ u(k)=1}} \left\lceil \frac{\Delta D}{12N_{DL}(1 - w_{oh})\eta_k(t)} \right\rceil. \quad (43)$$

The data-rate constraint for access transmissions can hence be expressed as

$$N_{i,l}^M(t) + N_{i,l}^U(t) \leq \hat{N}_{PRB}, \quad i \in \{0, 1, 2\}, l \in \mathcal{L}_i(\mathcal{K}_t^A), \quad (44)$$

where \hat{N}_{PRB} denotes the number of available PRBs, which depends on the bandwidth and numerology.

The data-rate constraint for backhaul transmissions can be written as

$$\left\lceil \frac{\Delta D \delta(\{k \in \mathcal{K}_t^B : a(k) = i\})}{12N_{DL}(1 - w_{oh})\eta_{BH,i}(t)f_{BH}} \right\rceil \leq \hat{N}_{PRB}, \quad i \in \{1, 2\}, \quad (45)$$

where

$$\delta(\mathcal{J}) = \sum_{k \in \mathcal{J}} u(k) + \mathbf{1}_{|\mathcal{J}| > \sum_{k \in \mathcal{J}} u(k)} \quad (46)$$

is the number of data blocks of size ΔD required by a group of UEs $\mathcal{J} \subset \mathcal{K}$.

A pairing $(\mathcal{K}_t^A, \mathcal{K}_t^B)$ is *feasible* if it satisfies constraints (40), (41), (44) and (45). We say that a subset of UEs $\mathcal{J} \subset \mathcal{K}$ can be served if there exist a partitioning $\mathcal{K}_1^A, \mathcal{K}_2^A, \dots, \mathcal{K}_T^A$ of \mathcal{J} and a partitioning $\mathcal{K}_1^B, \mathcal{K}_2^B, \dots, \mathcal{K}_T^B$ of $\{k \in \mathcal{J} : a(k) \neq 0\}$ such that $(\mathcal{K}_t^A, \mathcal{K}_t^B)$ are feasible pairings for all t . Given an ordered UE population \mathcal{K} , $|\mathcal{K}| = K$, we say that at most $\hat{K} < K$ UEs can be served if the set $\{1, 2, \dots, \hat{K}\}$ can be served and the set $\{1, 2, \dots, \hat{K} + 1\}$ cannot be served.

Let $x_{k,t} = \mathbf{1}_{k \in \mathcal{K}_t^A}$ and $y_{k,t} = \mathbf{1}_{k \in \mathcal{K}_t^B}$. The problem for user pairing and scheduling optimization can then be formulated as follows.

$$\text{maximize } \hat{K} \quad (47)$$

$$\text{subject to: } \sum_{t=1}^T x_{k,t} = 1 \quad \forall k \leq \hat{K} \quad (48)$$

$$\sum_{t=1}^T y_{k,t} = 1 \quad \forall k \leq \hat{K}, a(k) \neq 0 \quad (49)$$

$$(\mathcal{K}_t^A, \mathcal{K}_t^B) \text{ is feasible } \forall t = 1, \dots, T \quad (50)$$

$$\text{over } x_{k,t}, y_{k,t} \in \{0, 1\} \quad \forall k \in \mathcal{K}, t = 1, \dots, T. \quad (51)$$

To solve the combinatorial optimization problem above, we regard it as a variant of the bin packing with conflicts [55], [57] and employ the well-known heuristics such as First Fit Decreasing (FFD) using the RSRP value as a measure of item size.

VI. NUMERICAL RESULTS

This section presents a numerical analysis of the examined system and user association policies. Following a brief overview of the simulation setup, we evaluate the influence of key parameters on system performance to better understand its behavior. Then, we numerically compare the user association and beam selection policies described in Section IV. Finally, we evaluate the use of broad GoBB beams for multicast delivery.

A. SIMULATION SETUP AND PARAMETERS

The results presented in this section are obtained using Python 3 scripts that implement the performance evaluation framework described in Section VI, with the additional libraries *NumPy* and *Random*. Unless indicated otherwise, the results are averages over 200 samples of $K = 250$ UEs.

The cluster PPs are sampled in two stages. First, the number of clusters is drawn from the Poisson distribution with mean $\lambda_{cl} \times |\mathcal{S}|$ and their centers are distributed uniformly in \mathcal{S} . Then, sequentially for UEs $k = 1, \dots, K$, a cluster is drawn randomly and the position of UE k uniformly in the chosen cluster. Thus, the numbers of UEs in the clusters follow a multinomial distribution for any first $\hat{K} \leq K$ UEs. Association and beam selection policies handle UEs in the order of their indices k .

The default values of simulation parameters are listed in Table 2.

TABLE 2. Default system parameter settings.

Notation	Value	Description
β	9°	Mechanical downtilt
λ_B	0.5 blocker/m ²	Density of blockers
μ	3	Numerology
π_M	0.5	Probability of multicast service
σ_{SF}	4 dB	Slow fading standard deviation
B	3 dB	Cell/beam selection offset (bias)
D	30 Mbps	Requested data rate
d_{ISD}	260 m	Inter-site distance
f_c	28 GHz	Operating frequency
f_{BH}	1.3	Backhaul capacity factor
G_{UE}	3.61 dBi	UE antenna gain
h_{BS}	10 m	AAU height
h_{UE}	1.5 m	UE height
\hat{L}	8	Maximum number of beams per AAU
L_C	2 dB	Cable losses
L_B	3 dB	Beamforming losses
N_0	-174 dBm/Hz	Noise power spectral density
$N_{DL}^{A,B}$	14	Number of downlink symbols per slot
N_F	7 dB	Noise figure
N_{PRB}	132	Number of available primary PRB
$N_v \times N_h$	16 × 16	AAU UPA dimensions
\hat{P}	27 dBm	AAU power budget
R	150 m	Cell radius
T	8 time steps	Planning time horizon
ΔT	0.125 ms	Time slot duration
W	200 MHz	Bandwidth
w_{oh}	0.1	Overhead

B. SYSTEM CHARACTERIZATION

The results of this subsection are obtained under RSRP-based cell and beam selection as described in Subsection III-F. Fig. 8 shows the number of served UEs \hat{K} (top row) and the

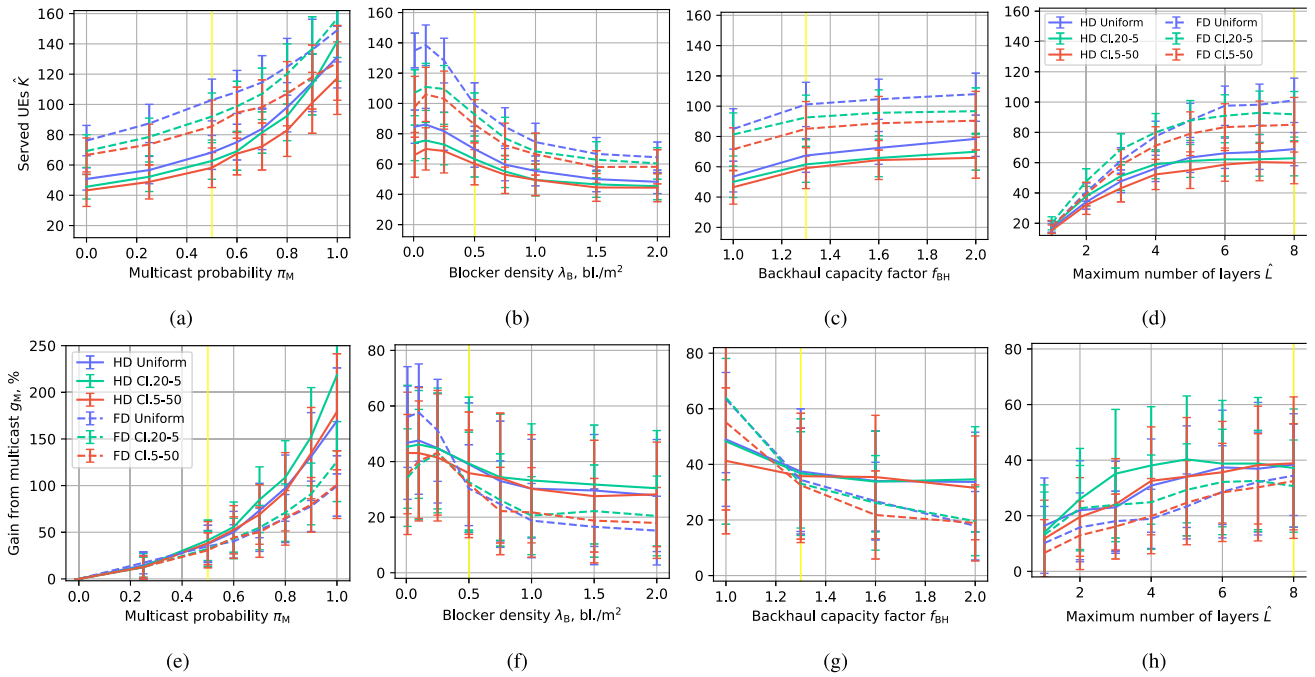


FIGURE 8. The average number of served UEs \hat{K} (top row) and the gain from multicast delivery g_M (bottom row) as functions of the multicast probability (a,e), blocker density (b,f), backhaul capacity factor (c,g) and maximum number of layers (d,h) under RSRP-based user association and beam selection for three studied UE layout scenarios: uniform (blue), Poisson cluster 20-5 (green), and Poisson cluster 5-50 (red). Solid and dashed lines show, respectively, the results for half- and full-duplex implementations. Yellow vertical lines indicate the parameter values adopted by default, namely, $\pi_M = 0.5$, $\lambda_B = 0.5$, $f_{BH} = 1.3$ and $\hat{L} = 8$.

gain from multicast g_M (bottom row) as functions of (from left to right) the multicast probability π_M , density of blockers λ_B , backhaul capacity factor f_{BH} , and the maximum number of beams swept simultaneously per AAU \hat{L} for the three UE layout scenarios presented in Subsection III-B and Fig. 3 assuming half (solid lines) and full (dashed lined) duplexing.

We observe that the capacity of the studied system in terms of the number of served UEs \hat{K} is radically affected by two parameters – the share of multicast UEs determined by π_M and the maximum allowed number of simultaneous beams \hat{L} – either of which can lead to its more than twofold variation in all scenarios, see Fig. 8(a), 8(d). The evolution is particularly visible for $0.5 < \pi_M \leq 1$ and for $1 \leq \hat{L} \leq 5$. The reasons are that the former parameter improves the PTM efficiency via increased clustering of multicast users, while the latter provides flexibility in transmission directions. We note, however, that starting from approximately $\hat{L} \sim 5$ no further improvement is observed due to the increased intra- and inter-cell interference. The variation of the backhaul capacity factor f_{BH} affects the system’s capacity to the least extent, see Fig. 8(c), although a gain of 20–25% is achieved when f_{BH} is increased from 1 to 1.3. This is because the presence of multicast traffic makes the use of backhaul connections more efficient.

The UE layout scenario affects the system’s capacity by up to 20%. The uniform layout yields the highest results in most cases except the nearly broadcast scenario where π_M approaches 1 in Fig. 8(a) and deployments with

smaller numbers of available beams, Fig. 8(d). In both cases a clustered UE layout with a large number of uniformly scattered dense UE groups permits to much better benefit from the PTM delivery, as it can be seen from Fig. 8(e) and 8(h). In all considered scenarios, except for the broadcast one, the system capacity under full duplex is consistently higher than under half duplex. However, the qualitative behavior of this metric as a function of the considered parameters is similar for both half and full duplex.

The gain from the multicast shared delivery is generally contained between 20 and 50% if multicast UEs constitute about a half of the UE population, see Fig. 8(f)–8(h), and soars to above 100% and 175%, respectively, in full and half duplex in the broadcast scenario in Fig. 8(e). Fig. 8(g) clearly shows the importance of the shared multicast delivery in backhaul when $f_{BH} = 1$ and backhaul links are less efficient. Indeed, generally, a bigger gain from multicast can be achieved if the number of served multicast UEs is large, however in Fig. 8(g) for half duplex the gain climbs to nearly 50% for $f_{BH} = 1$, where according to Fig. 8(c) the number of served UEs is 45–50, only half of which are multicast. Besides, only about 15% of the latter on average use PTM delivery, its efficiency hardly reaching $E_{PTM} = 1.1$ multicast UEs per transmission (not shown in Fig. 8), confirming that the gain stems mostly from backhaul. This is supported by the fact that the gain from multicast is generally higher in half duplex, where more system resources are consumed by backhaul transmissions.

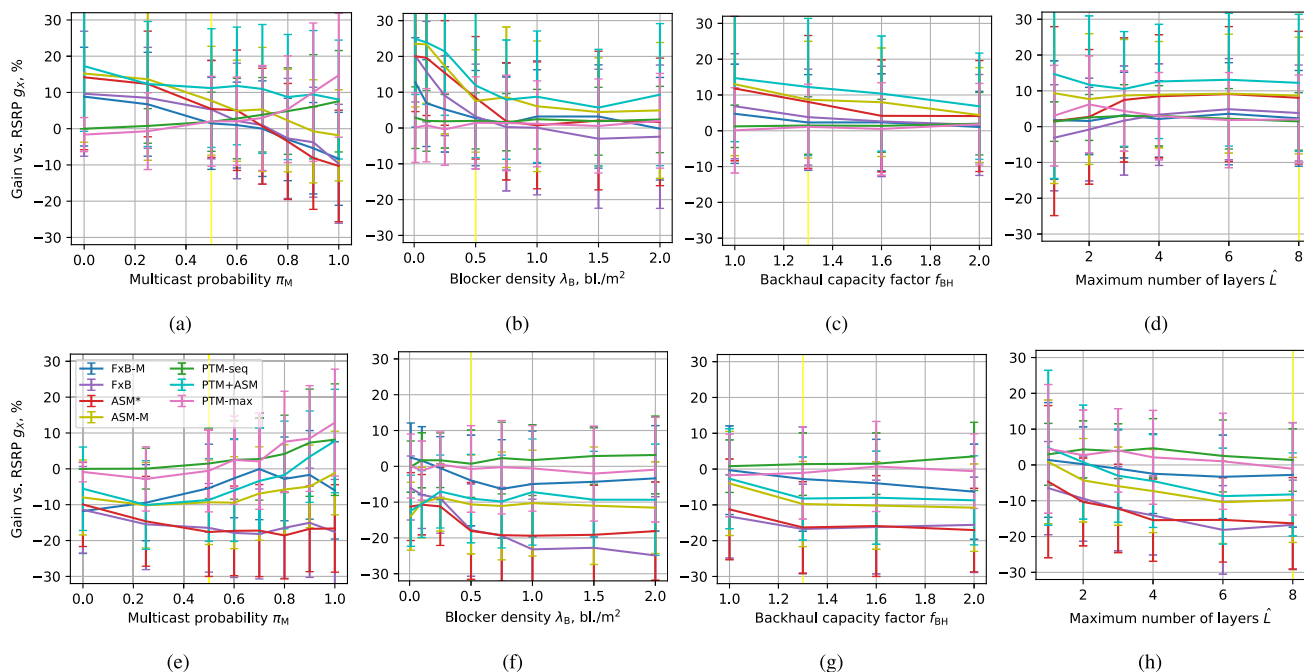


FIGURE 9. The capacity gain g_X in percentages achieved by user association and beam selection policies of Section IV compared to RSRP-based association for the uniform UE layout. The top row plots the results for half duplex, the bottom row for full duplex. Yellow vertical lines indicate the parameter values adopted by default, namely, $\pi_M = 0.5$, $\lambda_B = 0.5$, $f_{BH} = 1.3$ and $\hat{L} = 8$.

C. COMPARISON OF CELL ASSOCIATION AND BEAM SELECTION POLICIES

Figure 9 shows, as functions of parameters π_M , λ_B , f_{BH} and \hat{L} (as in Fig. 8), the capacity gains g_X achieved by the cell association and beam selection policies from Section IV vs. RSRP-based association. The results are provided for the uniform UE layout. The top row of the figure plots the results for half duplex, the bottom – for full duplex. Note that in our simulation if a UE associated by a certain policy cannot be served even if scheduled in isolation, i.e., the selected cell/beam does not provide the required signal strength, it is re-associated based on RSRP.

Fig. 9 reveals the difference between the two types of policies: the backhaul-saving, namely, *FxB* (purple), *FxB-M* (blue), *ASM** (red) and *ASM-M* (yellow), and the PTM-boosting *PTM-max* (pink) and *PTM-seq* (green). The hybrid *PTM+ASM* policy combining the two approaches is shown in cyan. It can be observed, that the considered backhaul-saving policies as well as *PTM+ASM* deliver a positive gain in terms of network capacity under half duplex, whereas under full duplex the gain in backhaul resources obtained by such policies does not compensate the loss in signal strength in access. The PTM-boosting policies, on the other hand, exhibit similar behavior for the both types of duplexing and particularly stand out in Fig. 9(a) and 9(e) as the proportion of multicast UEs grows and the system approaches the broadcast scenario. Under half duplex, the hybrid *PTM+ASM* policy successfully combines the advantages of the backhaul-saving and PTM-boosting approaches: it is the most robust to

variations of system parameters and consistently delivers the biggest gain among the studied policies, outperformed only by *PTM-max* in the broadcast scenario in Fig. 9(a).

Overall in Fig. 9, the biggest gains from user association are achieved under half duplex in none-to-light blockage conditions in Fig. 9(b) and stem primarily from saving backhaul resources for a large number of UEs in service. As it could be expected, the backhaul-saving policies perform better when the backhaul links are less efficient, see Fig. 9(c), although the achieved gain values, especially for fixed-bias policies, are not large due to smaller numbers of UEs that can be served. Among the policies of this type, the best performance is delivered by *ASM-M* (yellow).

Fig. 10 focuses on the *PTM+ASM* policy and plots the gain $g_{PTM+ASM}$ as a function of λ_B for the three studied UE layouts for the default parameter settings (solid lines) and for a scenario with $f_{BH} = 1$ (dashed lines). We observe that the hybrid policy delivers the gain of up to 50 % in none-to-light blockage conditions when the efficiency of backhaul and access links is comparable. The gain is slightly higher for the uniform UE layout, because, as Fig. 8(b) indicates, the system can handle a greater number of users, but overall the UE layout does not affect the performance of the policy.

Fig. 11 explores the performance of PTM delivery provided by *PTM+ASM* and the PTM-boosting policies. Combined with Fig. 9(a), the plots suggest that although the PTM-boosting policies consistently improve the efficiency of PTM delivery both in terms of the percentage of multicast UEs served in groups and the number of multicast UEs

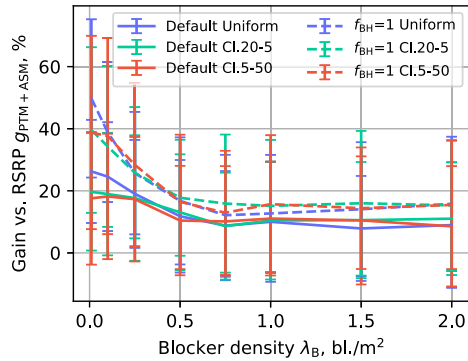


FIGURE 10. The capacity gain of PTM+ASM vs. RSRP-based association as a function of the blocker density for different UE layouts under half duplex. Solid lines show the results for the default parameter settings where $f_{BH} = 1.3$, dashed lines for a scenario in which backhaul links' efficiency is comparable to that of access, i.e. for $f_{BH} = 1$.

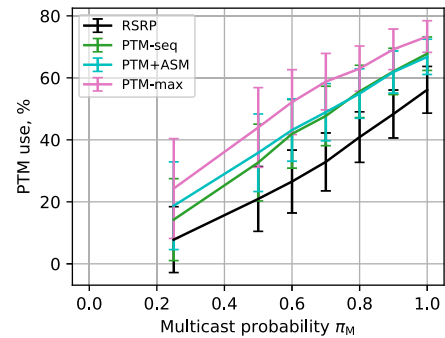
per PTM transmission, this does not necessarily result in an increase of system capacity unless the multicast UEs constitute the majority of the UE population. We also observe that PTM+ASM deliver a slightly better PTM performance compared to PTM-sec when multicast UEs constitute 25–75 % of the population. This is because PTM+ASM tries to associate unicast UEs to the donor and thus multicast UEs are more grouped at the IAB-nodes. Finally, we notice that *PTM-sec* is capable of delivering a rather high PTM efficiency, especially in the nearly broadcast scenario, in spite of its remarkably low complexity compared to *PTM-max*.

D. BROAD BEAMS FOR MULTICAST SERVICE

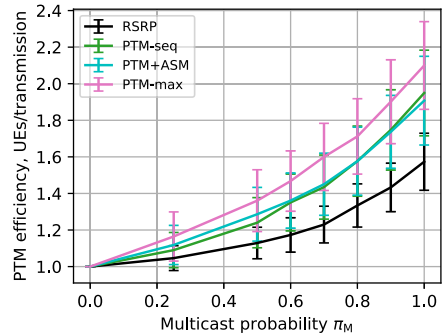
We now explore the use of the broader, GoBB beams for serving multicast users. For this, we allow the PTM-boosting policies to include GoBB beams into consideration using the same offset B from the RSRP value measured on the strongest GoBB beam. Then, at the scheduling stage, the broad beam, if selected by the policy as preferable, is considered first for serving the corresponding multicast UE.

Fig. 12 shows the metrics of interest as functions of the maximum number of simultaneous beams for the PTM-boosting policies with and without the use of GoBB beams. We observe that the use of broad beams results in clearly higher PTM use and efficiency regardless of the number of layers (Fig. 12(a), 12(b)). As Fig. 12(c) demonstrates, the policies with GoBB beams substantially outperform their counterparts employing only narrow beams in an analog beamforming scenario, i.e., in a system where only one beam can be swept by an AAU at each time step. If three or more beams can be swept simultaneously, the use of broad beams becomes inefficient.

Fig. 13 further investigates the analog beamforming scenario with $\hat{L} = 1$ and plots the metrics of interest as functions of the multicast probability π_M . RSRP-BB here indicates RSRP-based UE association GoBB-based service of multicast UEs. Whenever a UE cannot be served using its strongest GoBB beam, then a GoNB is employed. We observe



(a) PTM use as the percentage of multicast UEs served in PTM groups



(b) PTM efficiency E_{PTM} as the number of multicast UEs per transmission

FIGURE 11. Use and efficiency of PTM delivery as functions of the multicast probability for the uniform UE layout under half duplex.

in Fig. 13(a) that in the considered setting RSRP-BB outperforms the PTM-boosting policies with GoNB-only service even for mixed traffic. The PTM-boosting policies employing broad beams show excellent results outperforming RSRP-BB by a large margin. Fig. 13(b) provides additional details by plotting the resulting system capacity for different user layouts. As it could be expected, we observe that the layout with many tight clusters is particularly advantageous for broadcast in analog beamforming scenario regardless of the association method.

VII. CONCLUSION

The efficient support of multicast traffic in future 5G/6G IAB systems with directional antennas operating in MU-MIMO regime is a complex task involving several trade-offs. In our study, we developed a framework allowing to assess user associations in detail at the level of resource allocations and accurately evaluated several user association policies. The main takeaways of our study are as follows.

In an mmWave IAB system with MU-MIMO in which half the UEs receive a multicast service and consume the same data stream, the use of resource-efficient **multicast delivery** can add 20–50 % of capacity both in half- and full-duplex implementations. If the share of multicast sessions consuming the same data stream is above 80 %, multicast delivery can provide an important gain of over 100 % in terms

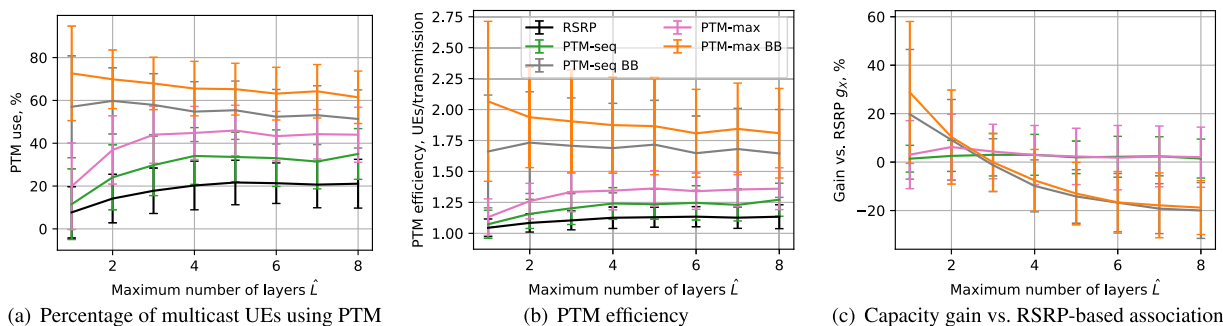
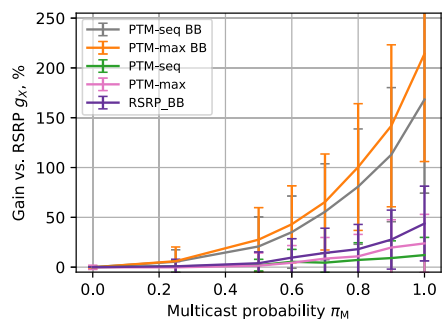
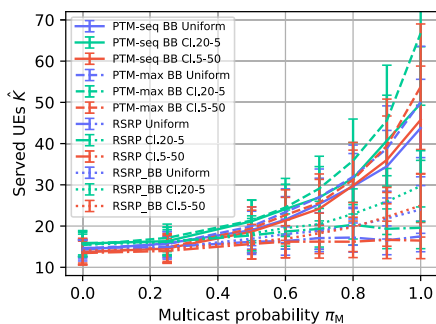


FIGURE 12. Performance measures of the PTM-boosting policies with and without the use of broad beams for PTM delivery as functions of the maximum allowed number of beams \hat{L} .



(a) Gain vs. RSRP-based association and GoNB-based service



(b) System capacity for different UE layouts

FIGURE 13. Performance metrics of PTM-boosting policies with and without the use of GoBB beams in analog beamforming scenario with $\hat{L} = 1$ as functions of the multicast probability. RSRP-BB uses RSRP-based association and preferably employs GoBB beams to serve multicast UEs.

of served users. The gain from multicast is generally bigger in half-duplex systems, which have a higher backhaul cost.

In MU-MIMO mmWave systems allowing for a large, above five, simultaneous beams per AAU, the density of **human blockers** greatly impacts the capacity of the system (providing a GBR service). In heavy blockage conditions, the capacity is observed to decrease to 60 % of its value under no-to-light blockage. Overall, when each AAU can use up to eight simultaneous beams, the uniform **UE layout** represents an optimistic scenario as to system capacity. The

capacity under cluster UE layouts is generally 10–15 % smaller. Although a user layout having numerous scattered dense clusters is generally more propitious for PTM delivery, its advantage stands out in nearly broadcast scenarios and is less marked when traffic is mixed.

Backhaul-saving user association policies, especially ASM-M, perform well in half-duplex deployments when the percentage of multicast UEs is below 50 %. They are particularly efficient in no-to-light blockage conditions, in which high channel quality allows for great flexibility in user association. In full duplex, however, such policies are observed inefficient. The performance and behavior of **PTM-boosting** policies are similar under half and full duplex. These policies are worth considering in nearly broadcast scenarios. In scenarios with mixed traffic, a higher degree of multicast user grouping, consistently provided by PTM-boosting policies, does not lead to a gain in system capacity. This can be explained by a decrease in resource allocation flexibility. On the other hand, to benefit from such flexibility, an efficient resource allocation/scheduling algorithm is needed.

More specifically, the practical low-complexity **PTM-sec** can be considered for multicast grouping in broadcast systems with codebook-based beamforming. The hybrid **ASM+PTM** policy successfully combines the backhaul saving and PTM-boosting approaches and delivers an average 15 % capacity gain in half-duplex systems with mixed traffic. A 40 % gain is reached in no-to-light blockage conditions when the efficiency of backhaul links is comparable to that of access. Overall, the policy’s performance is robust to variations of environmental and system parameters.

The use of **broader beams** for PTM service is advantageous in mixed-traffic and broadcast systems sweeping 1–2 beams simultaneously, such as analog beamforming implementations. When employed with a PTM-boosting policy, it can yield a gain of 25 % in mixed 50/50 traffic and over 150–200 % in broadcast. In systems capable of sweeping more than two simultaneous beams, the use of broader beams for multicast delivery is inefficient, again provided there is resource allocation/scheduling algorithm to exploit the multi-beam capability.

REFERENCES

- [1] *Ericsson Mobility Report*, Ericsson, Stockholm, Sweden, Nov. 2024.
- [2] S. Borole, K. Okeleke, J. Joiner, H. A. Ballon, and E. Kolta, "The mobile economy 2025," GSMA Intelligence, London, U.K., Tech. Rep., Mar. 2025.
- [3] *Massive MIMO Handbook*, 3rd ed., Ericsson, Stockholm, Sweden, 2024.
- [4] *Study on 6G Radio*, document RP-251881, 3GPP, Jun. 2025.
- [5] *5G mmWave Guide: A Resource for Operators*, GSMA, London, U.K., Feb. 2022.
- [6] M. Cudak, A. Ghosh, A. Ghosh, and J. Andrews, "Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 88–94, Apr. 2021.
- [7] N. Chukhno, O. Chukhno, D. Moltchanov, S. Pizzi, A. Gaydamaka, A. Samuylov, A. Molinaro, Y. Koucheryavy, A. Iera, and G. Araniti, "Models, methods, and solutions for multicasting in 5G/6G mmWave and sub-THz systems," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 119–159, 1st Quart., 2024.
- [8] M. A. Taga. (Mar. 2021). *5G Broadcast/Multicast: Redefining the Future of Content Delivery*. [Online]. Available: www.rohde-schwarz.com/us/solutions/broadcast-and-media/broadcast-distribution/5g-broadcast/ebook-5g-broadcast-multicast254597.html
- [9] V. K. Shrivastava, S. Baek, and Y. Baek, "5G evolution for multicast and broadcast services in 3GPP release 17," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 70–76, Sep. 2022.
- [10] H. Bolcskel, A. J. Paulraj, K. V. S. Hari, R. U. Nabar, and W. W. Lu, "Fixed broadband wireless access: State of the art, challenges, and future directions," *IEEE Commun. Mag.*, vol. 39, no. 1, pp. 100–108, Jan. 2001.
- [11] J. Li, K. K. Nagalapur, E. Stare, S. Dwivedi, S. A. Ashraf, P.-E. Eriksson, U. Engström, W.-H. Lee, and T. Lohmar, "5G new radio for public safety mission critical communications," *IEEE Commun. Standards Mag.*, vol. 6, no. 4, pp. 48–55, Dec. 2022.
- [12] *Pioneering 5G Broadcast*, 3GPP, Qualcomm, Sophia Antipolis, France, May 2021.
- [13] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.
- [14] R. M. Dreifuerst and R. W. Heath Jr., "Massive MIMO in 5G: How beamforming, codebooks, and feedback enable larger arrays," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 18–23, Dec. 2023.
- [15] *5G; NR; Integrated Access and Backhaul (IAB) Radio Transmission and Reception*, document TS 38.174, 3GPP, May 2024.
- [16] *5G; NR; NR and NG-RAN Overall Description*, document TS 38.300, 3GPP, Aug. 2024.
- [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] N. Yarkina, D. Moltchanov, A. Gaydamaka, and Y. Koucheryavy, "Coexistence of multicast and unicast services in mmWave/sub-THz self-backhauled systems: User associations and performance gains," *IEEE Trans. Veh. Technol.*, vol. 74, no. 3, pp. 4608–4624, Mar. 2025.
- [19] *5G; NR; User Equipment (UE) Procedures in Idle Mode and in RRC Inactive State*, document TS 38.304, 3GPP, Jan. 2025.
- [20] S. Ranjan, P. Chaporkar, P. Jha, and A. Karandikar, "Backhaul-aware cell selection policies in 5G IAB networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–6.
- [21] M. R. Garey and D. S. Johnson, "Approximation algorithms for bin packing problems: A survey," in *Analysis and Design of Algorithms in Combinatorial Optimization*, G. Ausiello and M. Lucertini, Eds., Cham, Switzerland: Springer, 1981, pp. 147–172.
- [22] *5G; Architectural Enhancements for 5G Multicast-Broadcast Services*, document TS 23.247, 3GPP, May 2024.
- [23] A. Prasad, B. Elmali, V. Pauli, N. Zheng, and D. Bhatoolaul, "Physical layer enhancements for high-reliability multicast and broadcast services in 5G release-17," *IEEE Commun. Standards Mag.*, vol. 8, no. 1, pp. 44–51, Mar. 2024.
- [24] A. Biazon and M. Zorzi, "Multicast via point to multipoint transmissions in directional 5G mmWave communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 88–94, Feb. 2019.
- [25] G. Wikström, J. Peisa, P. Rugeland, N. Johansson, S. Parkvall, M. Girnyk, G. Mildh, and I. L. Da Silva, "Challenges and technologies for 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.
- [26] *5G; NR; Backhaul Adaptation Protocol (BAP) Specification*, document TS 138 340, 3GPP, May 2024.
- [27] E. Dahlman, S. Parkvall, J. Peisa, and H. Tullberg, "5G evolution and beyond," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.
- [28] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, 2nd ed., Amsterdam, The Netherlands: Elsevier, 2021.
- [29] M. Gupta, I. P. Roberts, and J. G. Andrews, "System-level analysis of full-duplex self-backhauled millimeter wave networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1130–1144, Feb. 2023.
- [30] V. F. Monteiro, F. R. M. Lima, D. C. Moreira, D. A. Sousa, T. F. Maciel, B. Makki, and H. Hannu, "Paving the way toward mobile IAB: Problems, solutions and challenges," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 2347–2379, 2022.
- [31] E. Hossain, M. Rasti, and L. B. Le, "Cell association in cellular networks," in *Radio Resource Management in Wireless Networks: An Engineering Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2017, pp. 276–288.
- [32] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.
- [33] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.
- [34] L. Wang, B. Ai, Y. Niu, H. Jiang, S. Mao, Z. Zhong, and N. Wang, "Joint user association and transmission scheduling in integrated mmWave access and terahertz backhaul networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 15930–15940, Dec. 2023.
- [35] S. Gopalam, S. V. Hanly, and P. Whiting, "Distributed resource allocation and flow control algorithms for mmWave IAB networks," *IEEE/ACM Trans. Netw.*, vol. 31, no. 6, pp. 3175–3190, Dec. 2023.
- [36] F. Fang, G. Ye, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient joint user association and power allocation in a heterogeneous network," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7008–7020, Nov. 2020.
- [37] A. Mesodiakaki, F. Adelantado, L. Alonso, M. Di Renzo, and C. Verikoukis, "Energy- and spectrum-efficient user association in millimeter-wave backhaul small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1810–1821, Feb. 2017.
- [38] W. Deng, Y. Liu, M. Li, and M. Lei, "GNN-aided user association and beam selection for mmWave-integrated heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 12, no. 11, pp. 1836–1840, Nov. 2023.
- [39] G. Kwon and H. Park, "Joint user association and beamforming design for millimeter wave UDN with wireless backhaul," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2653–2668, Dec. 2019.
- [40] S. Baek, S. Choi, and S. Bahk, "DACs: User association and TDMA framing for low-latency services on integrated access and backhaul networks," *IEEE Trans. Veh. Technol.*, vol. 74, no. 5, pp. 8110–8125, May 2025.
- [41] A. Zhivtsova, V. Beschastnyi, Y. Koucheryavy, and K. Samouylov, "A survey of delay-oriented dynamic link scheduling policies for 5G/6G integrated access and backhaul systems," *IEEE Access*, vol. 12, pp. 118565–118586, 2024.
- [42] *5G; Service Requirements for the 5G System*, document TS 22.261, 3GPP, Jul. 2025.
- [43] E. Chen, M. Tao, and Y.-F. Liu, "Joint base station clustering and beamforming for non-orthogonal multicast and unicast transmission with backhaul constraints," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6265–6279, Sep. 2018.
- [44] F. Tan, P. Wu, Y.-C. Wu, and M. Xia, "Energy-efficient non-orthogonal multicast and unicast transmission of cell-free massive MIMO systems with SWIPT," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 949–968, Apr. 2021.
- [45] J. G. Andrews, R. K. Ganti, M. Haenggi, N. Jindal, and S. Weber, "A primer on spatial modeling and analysis in wireless networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 156–163, Nov. 2010.
- [46] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 3rd ed., Hoboken, NJ, USA: Wiley, 2013.
- [47] H. Asplund, J. Karlsson, F. Kronstedt, E. Larsson, D. Astely, P. von Butovitsch, T. Chapman, M. Frenne, F. Ghasemzadeh, M. Hagström, B. Hogan, and G. Jöngren, *Advanced Antenna Systems for 5G Network Deployments: Bridging the Gap Between Theory and Practice*. New York, NY, USA: Academic, 2020.

- [48] *On Forming Wide Beams*, document R1-1700772, 3GPP, Jan. 2017.
- [49] *5G; Study on Channel Model for Frequencies From 0.5 To 100 GHz*, document TR 38.901, 3GPP, May 2024.
- [50] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S.-P. Yeh, N. Himayat, S. Andreev, and Y. Koucheryavy, "Analysis of human-body blockage in urban millimeter-wave cellular communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.
- [51] G. R. MacCartney, T. S. Rappaport, and S. Rangan, "Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies," in *Proc. GLOBECOM - IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.
- [52] *5G; NR; User Equipment (UE) Radio Access Capabilities*, document TS 38.306, 3GPP, May 2022.
- [53] R. Kovalchukov, D. Moltchanov, Y. Gaidamaka, and E. Bobrikova, "An accurate approximation of resource request distributions in millimeter wave 3GPP new radio systems," in *Proc. NEW2AN*, 2019, pp. 572–585.
- [54] D. S. Johnson, "Approximation algorithms for combinatorial problems," *J. Comput. Syst. Sci.*, vol. 9, no. 3, pp. 256–278, 1974.
- [55] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 6th ed., Cham, Switzerland: Springer, 2018.
- [56] M. Rebato, L. Resteghini, C. Mazzucco, and M. Zorzi, "Study of realistic antenna patterns in 5G mmWave cellular scenarios," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
- [57] M. Gendreau, G. Laporte, and F. Semet, "Heuristics and lower bounds for the bin packing problem with conflicts," *Comput. Oper. Res.*, vol. 31, no. 3, pp. 347–358, Mar. 2004.



NATALIA YARKINA received the M.Sc. degree (Hons.) in applied mathematics and the Cand.Sc. degree in physics and mathematics from the Peoples' Friendship University of Russia, Moscow, Russia, in 2004 and 2007, respectively, and the Ph.D. degree in computing and electrical engineering from Tampere University of Technology, Finland, in 2025. Her current research interests include stochastic modeling, performance analysis, and system-level simulation of next-generation networks.



DMITRI MOLTCHANOV received the M.Sc. and Cand.Sc. degrees from St. Petersburg State University of Telecommunications, Russia, in 2000 and 2003, respectively, and the Ph.D. degree from Tampere University of Technology, Finland, in 2006. He is currently a University Lecturer with the Laboratory of Electronics and Communications Engineering, Tampere University of Technology. He has (co)-authored more than 150 publications on wireless communications, heterogeneous networking, the IoT applications, and applied queuing theory. In his career, he has taught more than 50 full courses on wireless and wired networking technologies, P2P/IoT systems, network modeling, and queuing theory. His current research interests include the research and development of 5G/6G systems, terahertz communications, ultra-reliable low-latency service, industrial IoT applications, and mission-critical V2V V2X systems.



MIKKO VALKAMA (Fellow, IEEE) received the M.Sc. (Tech.) and D.Sc. (Tech.) degrees (Hons.) in electrical engineering from Tampere University of Technology, Tampere, Finland, in 2000 and 2001, respectively. His Ph.D. Dissertation was focused on advanced I/Q signal processing for wideband receivers: models and algorithms. In 2003, he was a Visiting Postdoctoral Research Fellow with the Communications Systems and Signal Processing Institute, San Diego State University, San Diego, CA, USA. He is currently a Full Professor and the Unit Head of Electrical Engineering at Tampere University of Technology. His current research interests include radio communications, radio localization, and radio-based sensing, with particular emphasis on 5G and 6G mobile radio networks. He was a recipient of the Best Ph.D. Thesis Award of the Finnish Academy of Science and Letters for the Ph.D. Dissertation.

• • •