IAB vs. RIS: Performance-Cost Tradeoffs in 5G/6G Systems with Multicast and Unicast Traffic in Roadside Deployments

Olga Chukhno, Dmitri Moltchanov, Gianluca Brancati, Sara Pizzi, Antonella Molinaro, and Giuseppe Araniti

***Abstract*—The introduction of millimeter wave (mmWave) and sub-terahertz (sub-THz) frequency bands in 5G and future 6G networks promises an unprecedented capacity enhancement at the air interface driven by highly directional transmissions. While this facilitates interference suppression and increased deployment density, it also presents challenges in multicast service delivery, particularly due to the use of directional antennas and environmental factors that can cause signal blockage. Technologies such as Integrated Access and Backhaul (IAB) and Reconfigurable Intelligent Surfaces (RISs) have emerged to address these challenges and reduce capital expenditure. This study comprehensively compares IAB- and RIS-based designs for cost-efficient densification in mmWave and sub-THz 5G/6G systems, focusing on both unicast and multicast traffic in roadside-type deployments. Two deployment scenarios with tunable parameters are analyzed, and optimization frameworks are formulated for each approach, accounting for propagation characteristics, radio properties, and antenna directionality. The evaluation metrics not only assess the performance of each technology (IAB and RIS) but also consider the deployment costs required to achieve equivalent performance levels. Numerical results show that both RIS- and IAB-based deployments can effectively support multicast and unicast traffic, with IAB systems demonstrating superior performance in terms of overall resource utilization up to 50% in sparse deployment scenarios. Instead, in highly dense deployment scenarios, RISs exhibit superior scalability and resource efficiency compared to IABs, achieving up to a 3 times improvement factor. Furthermore, unlike IAB deployments, the performance of RIS-based systems continuously improves as the number of RIS nodes increases. From a capital expenditure perspective, RIS deployments prove to be more cost-efficient than IAB systems, provided that the unit cost of RIS is lower.**

***Index Terms*—5G/6G, New Radio, millimeter Wave, multicast, unicast, IAB, half-duplex, RIS, deployment, resource allocation.**

Received 21 July 2025; revised 11 September 2025; accepted 5 October 2025. Date of publication 8 October 2025; date of current version 4 February 2026. This work was supported in part by the Academy of Finland through Projects "Machine Learning Methods and Algorithms for 6G Terahertz Cellular Access" (HARMONIOUS) and "Machine Learning Algorithms for Energy Efficient and QoS Aware Communications in Heterogeneous 6G mmWave/sub-THz Networks" (ML6GThz) and in part by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of "Innovation Ecosystems", building "Territorial R&D Leaders" under Directorial Decree n. 2021/3277 - Project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009. Recommended for acceptance by S. Sun. *(Corresponding author: Dmitri Moltchanov.)*

Olga Chukhno, Gianluca Brancati, Sara Pizzi, and Giuseppe Araniti are with the Mediterranea University of Reggio Calabria, 89124 Reggio Calabria, Italy, and also with the CNIT 43124 Parma, Italy (e-mail: olga.chukhno@unirc.it; gianluca.brancati@unirc.it; sara.pizzi@unirc.it; araniti@unirc.it).

Dmitri Moltchanov is with Tampere University, 33100 Tampere, Finland (e-mail: dmitri.moltchanov@tuni.fi).

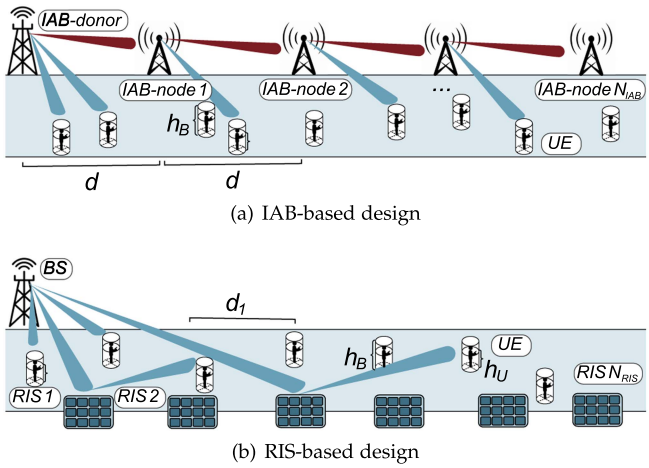
Antonella Molinaro is with the Mediterranea University of Reggio Calabria, 89124 Reggio Calabria, Italy, also with the CNIT 43124 Parma, Italy, and also with the Université Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: antonella.molinaro@unirc.it).

Digital Object Identifier 10.1109/TMC.2025.3619418

I. Introduction

THE introduction of millimeter wave (mmWave, 30-100 GHz) and sub-terahertz (sub-THz, 100-300 GHz) frequency bands in 5G and future 6G networks promises to dramatically boost the air interface capacity [1]. These frequencies also naturally require highly directional antennas to extend the coverage of individual base stations (BS) [2], facilitating efficient interference suppression [3], [4], and enabling a significant increase in deployment density to meet growing user demands. However, in addition to this great potential, the use of mmWave and sub-THz bands introduces new and unique challenges for system design. First, environmental factors such as signal blockage can cause dramatic declines in received signal strength, necessitating extremely dense deployments to ensure ubiquitous connectivity [5]. Furthermore, the deployment of directional antennas complicates service provisioning, especially for multicast services [6] whose efficiency heavily relies on the service area of a single antenna configuration [7], [8].

To address these challenges and mitigate capital expenditures (CAPEX), two viable solutions have recently emerged that enable cost-effective network densification. On the one hand, *integrated access and backhaul (IAB)*, recently standardized by the 3rd Generation Partnership Project (3GPP) [9], uses the same wireless infrastructure to provide both access and backhaul connectivity, relying on multi-hop communications [10]. An IAB donor is a BS with a wired connection to the 5G core network, while other BSs, known as IAB nodes, wirelessly connect to the donor for backhaul and simultaneously serve user equipment (UEs) with access connectivity. By employing low-cost IAB nodes as relays between the UE and the donor, IAB systems bring service points closer to users, effectively reducing the impact of blockages at low deployment costs. From a system design perspective, IAB enables the full utilization of directional communications at both IAB nodes and the IAB donor (as shown in Fig. 1(a)). On the other hand, another cost-effective densification option relies on the use of *reconfigurable intelligent surfaces*



(a) IAB-based design

(b) RIS-based design

Fig. 1. The considered reference scenarios.

(RISs), which enable the manipulation of incident beams to steer them toward users [11] (see Fig. 1(b)). By deploying multiple RISs within the service area of a single BS, this technology capitalizes on spatial diversity to avoid blockages. Although RISs do not reuse the network operator's bandwidth, the cost of a single unit is expected to be significantly lower than that of an IAB node [12], [13].

The need to support multicast traffic in 5G/6G systems further complicates the deployment choice. First, in both IAB and RIS-based deployments, managing user association at the access interfaces becomes more complex, especially when multiple steerable beams with dynamically varying half-power beamwidth (HPBW) are used. As beams become narrower (i.e., smaller HPBW), the beamforming gain increases, which is desirable for throughput; however, the coverage area becomes smaller. This leads to more frequent beam switching and finer-grained control over which users are served by which beams. When multicast groups are distributed over a wide area, the system must either use wider beams (lower gain and spectral efficiency) or split the group into multiple subgroups served by different beams, increasing resource consumption and coordination complexity.

Moreover, in IAB systems, the choice of association point is influenced by the type of traffic – unicast vs. multicast. On one hand, multi-hop traffic accumulation in the considered roadside deployment can create backhaul bottlenecks. However, multicast traffic may benefit IAB systems in this scenario, as only a single data copy needs to traverse the wireless backhaul. Additionally, IAB enables placing service points closer to users, potentially reducing resource usage at access interfaces. Conversely, the RIS topology operates with a single resource pool (without access replication, as in IAB) and significantly longer BS-RIS-UE paths, potentially increasing air-interface load. Therefore, the optimal technology choice for roadside deployments is not straightforward and warrants further investigation under realistic traffic assumptions.

In RIS-based deployments, the choice of the optimal RIS to serve a multicast group adds another layer of complexity [14], [15]. Unlike conventional active transceivers, RISs do not process or amplify signals—they only reflect them based on a configured phase profile. Therefore, the effectiveness of a RIS depends on the geometry of the environment, user mobility, and channel conditions. When multiple RISs are available, the system must select the RIS that provides the best combined reflection path (from NR BS to RIS to UEs), while ensuring that the passive beamformed signal reaches all users in the multicast group with acceptable quality. This task becomes even more difficult when users are spatially dispersed or moving, as the reflection angle must optimally align with all users simultaneously, which is not always feasible.

Despite the mentioned opportunities and challenges, no studies have yet attempted to quantify which densification approach – RIS-based or IAB-based – provides better performance and lower deployment costs for realistic mmWave/sub-THz 5G/6G systems serving a mix of multicast and unicast traffic. Such a quantification is essential because IAB and RIS differ significantly in both performance and deployment cost. For example, in a connected highway scenario, the challenge lies in maintaining uninterrupted coverage despite frequent line-of-sight obstructions. An IAB-based roadside deployment can actively relay traffic alerts when the direct link to the base station is blocked. The active nature of IAB provides robust coverage for high-speed scenarios. However, each IAB node requires power and backhaul connectivity, leading to higher infrastructure deployment costs. In contrast, a RIS-based deployment may place passive RIS panels on nearby buildings or roadside infrastructure to reflect the base station's signal around obstacles, potentially extending coverage around curves or through tunnels. While RIS offers a cost-effective, power-free solution, its effectiveness depends heavily on precise placement and reflection geometry, and it lacks the ability to amplify or regenerate signals.

This paper aims to fill the mentioned gap by comprehensively comparing IAB- and RIS-based designs for cost-effective densification in mmWave/sub-THz 5G/6G systems that handle both unicast and multicast traffic. We consider two comparable deployments with tunable parameters and develop suitable optimization frameworks for each approach, accounting for the specifics of mmWave/sub-THz propagation, including blockage effects and multi-beam directional antennas. As a metric of interest, we focus on the resource usage required to serve a given user density, and compare the deployment costs needed to achieve equivalent performance. Based on these comparisons, we provide recommendations on the appropriate use of IAB and RIS technologies for 5G/6G network densification.

The *main contributions* of our work are:

- Mathematical optimization frameworks for two 3GPP-based mmWave 5G New Radio (NR) network optimization strategies, RIS- and IAB-based, accounting for multicast and unicast traffic, propagation specifics, radio characteristics, and directional antennas.
- A comparison of these optimization strategies in terms of the minimum resources required to serve a given user density and deployment costs to achieve an equivalent performance level.
- Observations that IAB systems slightly (up to 50%) outperform RIS-based deployments in terms of resource utilization, but RIS-based systems are significantly more

cost-efficient in sparse deployment scenarios. Instead, RISs demonstrate superior scalability and resource efficiency in highly dense deployment scenarios compared to IABs, showing up to a $3\times$ improvement factor.
- Recommendations for RIS-based densification in outdoor deployments, while IAB systems are preferable in complex multi-path propagation environments and in specific use-cases, such as outdoor-to-indoor communications.

The rest of the paper is organized as follows. Section II reviews related works on IAB system design, the use of RIS in 5G/6G systems, and multicasting with directional antennas. The system model and assumptions are presented in Section III. Section IV formalizes the optimization task. Numerical results are discussed in Section V, and conclusions are provided in the final section.

II. Related Work

In this section, we review the related work, beginning with an overview of the literature on multicasting in systems that utilize directional antennas and then providing the technological specifics of IAB- and RIS-based architectures.

A. Multicasting in Systems With Directional Antennas

In recent years, several studies have focused on designing strategies for group-oriented communications in directional systems because the strong directionality required at extremely high frequencies precludes serving all clients in a multicast group with a single transmission [16].

In [17], a scalable beam grouping algorithm is designed with the objective of maximizing the throughput delivered to multicast groups. First, the access point is trained with per-beam per-client Received Signal Strength Indicator (RSSI) measurements by partially traversing a codebook tree. Then, the training information is exploited to design a scalable beam grouping algorithm that approximates the minimum multicast group data transmission time.

In [18], the trade-off between serving many users simultaneously (thus saving resources at the access point) and providing a high Signal-to-Noise Ratio (SNR) is analyzed. Specifically, the beam widths, beam directions, and retransmissions are optimized by formulating the problem under a multi-objective optimization framework. Furthermore, a heuristic algorithm for optimal multicasting with a single lobe antenna pattern has been presented by considering the delay-energy trade-off.

The possibility of exploiting device-to-device (D2D) communications in directional multicast systems has been investigated in the literature. In [19], D2D multi-hop and concurrent transmissions are jointly exploited in the proposed efficient multicast scheduling (EMS) strategy, tailored to achieve lower energy consumption in comparison with mmWave multicast performed through serial unicast transmissions. The optimal multicast scheduling problem with D2D communications is also analyzed in [20], where concurrent transmission, multicast group partition, and beam selection in a multilevel codebook are considered to minimize the total multicast transmission time.

Enabling efficient broadcast and multicast transmissions for vehicle-to-anything services is addressed in [21], wherein an improved beamformed broadcast/multicast technology that builds on adaptive and robust beam management techniques has been developed. The proposed solution is especially suitable for mmWave bands, with large antenna arrays deployed at 5G NR BS.

A non-orthogonal multiple access (NOMA) unicast-multicast system is designed in [22], where a cooperation strategy to maximize users' reliability is introduced. Similarly, a cooperative multicast scheme for mmWave systems utilizing NOMA has been introduced in [23]. Analytical expressions for the signal-to-interference-plus-noise (SINR) ratio coverage probability are then derived to evaluate the performance of the proposed heuristics. However, NOMA deployment in 5G NR mmWave is still under discussion by the 3GPP.

Stochastic geometry, together with queuing theory, is applied in [24], where a framework to estimate the NR access point parameters in the presence of multicast and unicast traffic is presented. 5G NR Mixed Mode (MM) is proposed in [25] to enable multicasting in the 5G NR Release 17. Specifically, all the modifications required to provide flexible, dynamic, and seamless switching between unicast and multicast/broadcast transmissions and traffic multiplexing under the same radio structures are discussed.

Energy efficiency aspects, which are crucial for resource-constrained device communication, are considered in [26], where the authors propose a heuristic resource management framework aimed at simultaneously minimizing energy and maximizing network throughput. Multi-beam antenna operation is analyzed in [8], where the globally optimal solution for multi-beam mmWave BS operation is presented. More recently, resource allocation for multicast services in dual mmWave and μWave base station deployments with multi-beam directional antennas has been tackled in [27]. Also, work [28] proposes a solution for efficient multicasting in mobile systems, which uses a combination of sidelink/D2D, unicast, and multicast transmissions.

An up-to-date and exhaustive treatment of performance optimization methods for 5G/6G mmWave/sub-THz systems is carried out in [7], where challenges and opportunities of 5G/6G multicast mmWave/sub-THz resource allocation specifics are thoroughly discussed.

B. IAB for 5G/6G mmWave/sub-THz Systems

The IAB architecture represents the first attempt to introduce multi-hop communications into commercial cellular systems. To reduce the cost of a single IAB node while preserving the UE functionality and standardizing the design of IAB donors and IAB nodes, the IAB node's logical structure is divided into two main components: the distributed unit (DU) and the mobile termination (MT), as shown in Fig. 2. The MT performs the physical, MAC, and radio link control (RLC) functions of a standard UE, while the DU is responsible for relaying traffic to downstream IAB nodes or to UEs. The system is centrally controlled via a control unit at the BS. To facilitate this, 3GPP

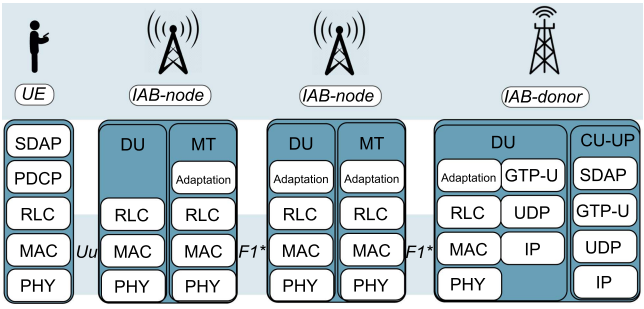


Fig. 2. IAB system and protocol architecture.

introduced the backhaul adaptation protocol (BAP), which is responsible for seamless signaling exchanges between remote IAB nodes and the BS. To simplify network design, IAB only supports tree or directed acyclic graph (DAG) topologies, avoiding the need for complex routing protocols.

While the IAB architecture is compatible with both FR1 (sub-6GHz) and FR2 (mmWave) frequency bands, one of its major goals is to enable cost-efficient network densification and mitigate blockage in mmWave/sub-THz systems. IAB [9] supports both in-band and out-of-band backhauling modes. The former utilizes the entire set of resources available at both IAB donors and IAB nodes, and is considered the primary option by 3GPP and vendors [9], [29], [30]. The use of the same bandwidth for both access and backhaul, known as in-band backhauling, requires careful transmission schedule management. It has been shown that while the simultaneous reception from access and backhaul is feasible by utilizing scheduling and/or spatial diversity mechanisms [31], simultaneous transmission and reception require either full-duplex operation and/or time-division duplex resource management. As full-duplex wireless systems are in early stages of deployment [32], 3GPP currently mandates support for half-duplex operation in IAB systems [9].

IAB is expected to be employed at up to $10-20\%$ of 5G sites [33]. Despite this, the mandatory class of IAB systems – half-duplex ones – has not received significant attention thus far. In their seminal study, the authors in [34] proposed to utilize reinforcement learning techniques to determine the set of links in an IAB network that can be simultaneously activated. As a metric of interest, the network capacity is utilized. In [35], the authors considered a half-duplex system with a multi-beam IAB donor and IAB nodes, as well as a multi-connectivity operation. By utilizing a computer simulation approach, they demonstrated that the use of a multi-beam at the IAB donor may provide load balancing between access and backhaul. These observations have been further confirmed in [36]. The common shortcoming of most of the studies performed so far is the adoption of an oversimplified representation of the service specifics at the access part of the system.

All the above-mentioned studies assumed unicast traffic. However, in practice, IAB systems should be designed to provide seamless communication to users, irrespective of their traffic type and association point. The literature on the support of multicasting in IAB systems is virtually non-existent. The authors in [7] mentioned this as a critical task for future cellular systems. In [37], the authors considered sidelink-based IAB technology to enhance multicasting support in 5G systems. However, they focused on out-of-band backhauling, which simplifies the system architecture.

C. RIS for 5G/6G mmWave/sub-THz Systems

In the realm of 5G/6G wireless systems, RIS stands out as a transformative technology that addresses the challenges of the mmWave and sub-THz frequency bands [38]. The successful deployment and functionality of an RIS depend on key operational parameters, including the presence of a direct path between the UE and the network, operating regime (far-field or near-field), frequency range (sub-6GHz or mmWave, i.e., FR1 or FR2), antenna configuration between the network and UE (SISO, SIMO, MISO, and MIMO), and the nature of the RIS itself, which can be passive, semi-active, or active [39], [40], [41].

Defined as a surface rather than a volumetric material, RIS is strategically engineered for implementation simplicity and to reduce losses while maintaining precise control over electromagnetic waves [40]. The introduction of RIS elements into the environment enables control and manipulation of wireless signals, facilitating beamforming, signal focusing, and overcoming obstacles [42]. This integration adds a new dimension to wireless network design, offering opportunities for optimized coverage, improved spectral efficiency, and interference mitigation.

More importantly, RIS can be implemented using mostly passive components, avoiding the need for high-cost active components, such as power amplifiers [39], [41]. This design choice results in low implementation costs and energy consumption, making RIS a cost-effective and energy-efficient solution for enhancing wireless communication systems, especially when compared to traditional relaying methods [43]. Passive RIS structures can be seamlessly integrated into a wireless communication environment, offering flexibility and easy deployment in various settings [41], [44].

Researchers have shown keen interest in exploring the potential of RIS to address the challenges associated with mmWave and sub-THz frequency bands. Numerous comprehensive surveys and tutorials have been conducted to consolidate the existing body of knowledge on RIS, delving into various aspects of RIS ranging from its fundamental principles to its practical applications [45], [46], [47], [48]. Moreover, delving deeper into RIS aspects, the authors have investigated several technical aspects such as optimal RIS deployment [15], reflective element configuration [49], antenna structure [50], comparative analysis of passive and active RISs [51], [52] and with respect to relaying [53], performance evaluation [54], analysis [55], and implementation [56], among many others [57].

In the context of IAB systems, the integration of RISs has garnered attention for its potential to enhance reliability and address the challenges associated with mmWave and sub-THz frequency bands. For example, in [13], the authors employed mathematical programming models aimed at full network planning optimization to generate optimal network layouts utilized

to evaluate scenarios, positions, and configurations where RISs can effectively enhance the reliability of these networks. In [12], the authors examined the resilience of optimized mmWave IAB Radio Access Network (RAN) deployments with RISs. The results demonstrated that although RISs cannot entirely replace IAB nodes, the synergy between the two technologies effectively sustains elevated levels of network resilience. Moreover, the combined deployment incurs lower costs compared to IAB-only planning. However, comparative analysis across IAB and RIS-based approaches, which could assess the costs associated with achieving equivalent performance when considering unicast and multicast traffic, is missing in the literature. In summary, no prior work compares the performance and cost of IAB and RIS coverage extension mechanisms directly for mixed unicast and multicast traffic in realistic 5G/6G mmWave/sub-THz scenarios. Existing IAB/RIS research often simplifies access specifics and overlooks multicast-specific challenges. Also, while IAB-RIS integration has been explored for reliability, a direct cost-performance comparison with purely IAB or RIS deployments for mixed traffic is missing. *This paper addresses this gap by directly comparing IAB and RIS densification strategies for mixed traffic under realistic conditions, providing novel insights and practical recommendations for cost-effective 5G/6G deployments. Our analysis will quantify resource efficiency and deployment cost trade-offs, offering a more informed understanding of each technology's strengths and weaknesses for network densification.*

TABLE I
MAIN NOTATION UTILIZED IN THE PAPER

Parameter	Definition
N_{IAB}	Number of IAB nodes
N_{RIS}	Number of RISs
N_{UE}	Number of UEs
d	2D distance between IAB deployment entities, m
d_1	2D distance between RIS deployment entities, m
e	Communication service, $e \in \mathcal{E} = \{1, \ldots, E\}$
B_e	Guaranteed bit rate delivered by the network
N	Number of planar antenna array elements
$\theta^{\pm}_{3db}$	Upper and lower 3-dB points of antenna array, rad
θ_m	Location of array maximum, rad
β	Antenna array orientation, rad
y	3D distance between UE and serving entity, m
$S(y)$	SINR at 3D-distance y, SINR
P_A	Transmit power, dBm
G_A	Antenna gain at the NR BS, IAB donor, or IAB node, dBi
G_U	Antenna gain at the UE, dBi
S_F	Shadow fading, dB
U_F	Fast fading, dB
N_0	Thermal noise power, dB/Hz
M_I	Interference margin, dB
W	Available bandwidth, MHz
$L_{dB}(y)$	Path loss in decibel scale
f_c	Carrier frequency, GHz
A, ζ	Propagation coefficients
y_1	Three-dimensional distance between RIS and UE, m
β_0	Path loss at 1 meter
N_{re}	Number of reflective elements
$p_B(y)$	Distance-dependent blockage probability
h_B	Height of blocker, m
r_B	Radius of blocker, m
h_U	Height of UEs, m
h_A	Height of NR BS, IAB donor, and IAB nodes, m
M	Number of time slots time horizon
L	Number of beams
R_b	Number of resource blocks in time slot
w_{PRB}	PRB size, MHz

III. SYSTEM MODEL AND ASSUMPTIONS

In this section, we introduce the system model and outline the modeling assumptions related to IAB- and RIS-based deployments and traffic patterns, as well as propagation, blockage, and antenna models. Then, we describe the resource allocation procedure. A notation guide is provided in Table I.

A. Deployment and Traffic Models

1) IAB Deployment: The IAB donor and IAB nodes offer connectivity to UEs and operate according to the specifications outlined in 3GPP TR 38.874 for the 5G NR system. The IAB network comprises one IAB donor and N_{IAB} IAB nodes. Only the IAB donor is assumed to have a wired connection to the 5G Core (5GC) and is responsible for providing wireless backhaul connectivity to the IAB nodes. These nodes operate in an in-band half-duplex mode, meaning that they cannot transmit data to its UEs while simultaneously receiving data over the backhaul link.

The deployment scenario considers a street environment, with the analysis focusing on the complete communication path from the IAB donor to the UE. The path comprises only the wireless access link if the UE is directly connected to the IAB donor. Conversely, if the UE is served by an IAB node, the communication path includes both the access link from the UE to the IAB node and the wireless backhaul path from the IAB donor to the IAB node, which may traverse multiple intermediate IAB nodes.

For reference, the IAB donor is located at the origin of the coordinate system, and the two-dimensional (2D) Euclidean distance between IAB network entities is denoted by d.

2) RIS Deployment: In this scenario, the NR BS is connected to the 5GC, while RISs are deployed throughout the network to reflect and manipulate the wireless channel between the NR BS and the UEs. The RIS-assisted network operates in an in-band full-duplex regime and involves a single-reflection path between the BS and the UEs, utilizing one of the N_{RIS} passive RISs.

Similar to the IAB setup, we consider a street-level environment and focus on the wireless access link from the NR BS to the UE via a reflection path involving a single RIS (i.e., NR BS $\rightarrow$ RIS $\rightarrow$ UE, see Fig. 1(b)). While IAB inherently supports multihop links due to its active nature, the passive nature of current RIS technology limits practical deployments to single reflection path.

The NR BS is located at the origin of the coordinate system. RIS 1 is positioned at a 2D distance d_1, where $d_1 \leq d$, from the NR BS. Additional RISs, if present, are arranged sequentially in a chain, each separated by a 2D distance d_1 from the adjacent RIS.

3) UE Deployment and Traffic: A total of N_{UE} UEs are distributed within the considered service area, each requesting a communication service $e \in \mathcal{E} = \{1, \ldots, E\}$, with a guaranteed bit rate (GBR) B_e to be delivered over the downlink.

The UE-to-service association is determined by user subscriptions, which may fall under either unicast or multicast categories.

Accordingly, service delivery within the RAN can occur through unicast or multicast transmission modes.

In the case of unicast service delivery, each UE requires an individual data stream, resulting in separate transmissions. In contrast, multicast transmission enables the network to serve multiple UEs with a single transmission, or through a hybrid approach combining unicast and multicast – using either sequential or parallel transmissions across different beams – based on the operator's resource allocation policy.

Importantly, for multicast traffic, only a single data replica needs to traverse the backhaul link, regardless of the number of UEs in the multicast group.

B. Antenna Model

We assume that both the NR BS and the UEs are equipped with planar antenna arrays characterized by a cone-shaped radiation pattern. Following the classical representation from [58], the HPBW of the array is defined as $\theta = 2|\theta_m - \theta_{3db}|$, where θ_{3db} is the angle at which the output power drops to -3 dB from its peak value, and θ_m denotes the angle corresponding to the main lobe maximum, given by $\theta_m = \arccos(-\beta/\pi)$, with β representing the phase excitation difference. The mean gain over the HPBW is calculated as

$$G = \frac{1}{\theta^{+}_{3db} - \theta^{-}_{3db}} \int_{\theta^{-}_{3db}}^{\theta^{+}_{3db}} \frac{\sin(N\pi\cos(\theta)/2)}{\sin(\pi\cos(\theta)/2)} d\theta, \quad (1)$$

where N is the number of antenna elements, and the upper and the lower 3-dB points are $\theta^{\pm}_{3db} = \arccos[\pm 2.782/(N\pi)]$.

We assume that the BS, IAB donor, and IAB nodes utilize either hybrid or fully digital beamforming architectures, which allow them to simultaneously sweep multiple beams. Within the beams swept by the IAB donor and IAB nodes, a dedicated directional beam is allocated for the backhaul connection between the IAB donor and the IAB node or between IAB nodes. The remaining beams are designated for access links, supporting communication between the IAB donor/node and the associated UEs.

In the case of the RIS-based network, all available beams are exclusively utilized for the wireless access link, enabling communication from the NR BS to UEs via reflection through one of the deployed RISs.

C. Propagation Models

To account for interference from other network nodes (e.g., adjacent donors, IAB nodes, or other cells), we incorporate an interference margin, M_I, into our SINR calculations. This approach is well-justified in the context of mmWave/sub-THz systems [7]. Specifically, the SINR at a three-dimensional (3D) distance y between the NR BS, IAB donor, or IAB node, and a UE is expressed as in [7]:

$$S(y) = \frac{P_A G_A G_U S_F U_F}{(N_0 W + M_I) L(y)}, \quad (2)$$

where P_A is the transmit power, G_A is the antenna gain at the NR BS, IAB donor, or IAB node, G_U is the antenna gain at the UE, S_F represents the shadow fading, U_F denotes fast fading capturing small-scale variations in received signal strength, N_0 is the thermal noise power density, W is the bandwidth, M_I is the interference margin, and $L(y)$ is the path loss expressed in linear scale.

The path loss measured in dB is determined according to [59] as

$$L_{dB}(y) = 32.4 + 21\log_{10} y + 20\log_{10} f_c, \quad (3)$$

where f_c is the carrier frequency in GHz, and y is the 3D distance between the transmitter (NR BS, IAB donor, or IAB node) and the UE. The path loss represented in (3) can be converted into a linear scale using Ay^{ζ}, where A and ζ are the propagation coefficients. To account for the Line of Sight (LoS)[1] and blocked conditions, we introduce coefficients (A_1, ζ) and (A_2, ζ) as follows:

$$A_1 = 10^{2\log_{10} f_c + 3.24}, A_2 = 10^{2\log_{10} f_c + 4.74}, \zeta = 2.1, \quad (4)$$

as 3GPP recommends $\zeta = 2.1$ for the LoS state.

In the RIS scenario, the SINR between the NR BS and the UE via a single-reflection path is given by [60]:

$$S_{RIS} = \frac{P_A G_A G_U S_F U_F \beta_0^2 N_{re}^2}{d_1^{\zeta} y_1^{\zeta} (N_0 W + M_I)}, \quad (5)$$

where β_0 is the path loss at 1 m, N_{re} is the number of RIS reflective elements, and y_1 is the distance between the RIS and the UE.

D. Blockage Model

We account for human body blockage by assuming an attenuation of 15 dB, as reported in [61]. Blockers are modeled as vertical cylinders with height h_B and radius r_B, and are spatially distributed on the Euclidean plane according to a Poisson distribution with density λ_B. Following the model in [62], the probability of blockage at a distance y between the transmitter and the receiver is given by

$$p_B(y) = 1 - \exp^{-2\lambda_B r_B \left[\sqrt{y^2 - (h_A - h_U)^2}\frac{h_B - h_U}{h_A - h_U} + r_B\right]}, \quad (6)$$

where h_A represents the height of NR BS, IAB donor, and IAB nodes, while h_B and h_U are the heights of the blocker and of UEs, respectively.

E. Radio Resource Characterization

We consider an Orthogonal Frequency Division Multiple Access (OFDMA) scheme with a slotted time structure specified in 5G NR [63]. The NR BS, IAB donor, and IAB nodes are each configured with a total bandwidth W, a carrier frequency f_c, and a specific NR numerology μ defining the size of the physical resource block (PRB), denoted by w_{PRB}.

The number of PRBs required to serve a UE depends on the modulation and coding scheme (MCS), which is selected based on the transmission mode (unicast/multicast), the channel conditions, and the association of the UE – either with the IAB

[1] This work focuses exclusively on improving Line-of-Sight (LoS) coverage in deployments along streets, roads, and highways, where direct visibility between nodes is assumed.

donor/IAB nodes (in the IAB-based system) or the NR BS/RIS (in the RIS-assisted system).

In this work, we consider a finite time horizon, T, over which radio resources are allocated. Thus, to parameterize the OFDMA scheme, let M denote the number of time slots (e.g., subframe) within T, indexed by t. Furthermore, we define L as the maximum number of beams that can be swept simultaneously by the NR BS and IAB donor. Since IAB nodes typically operate under more constrained hardware capabilities, we denote the number of beams by L_n, with $L_n \leq L$.

The maximum number of PRBs available at the NR BS or IAB donor over the time horizon is given by MLR_b, while for an IAB node, it is ML_nR_b, where R_b is the number of PRBs available per beam per time slot. It is important to note that for IAB-based systems, one beam must be reserved for wireless backhauling, thus the maximum number of unicast/multicast subgroups that can be supported simultaneously by the IAB donor and IAB nodes is $(ML-1)$ and (ML_n-1), respectively. In contrast, RIS-assisted networks do not require beam reservation for backhauling, and can therefore support up to ML unicast or multicast subgroups.

Note that for the IAB system, we assume strict resource division between uplink and downlink and focus exclusively on the downlink direction due to the nature of the considered services, e.g., multicast, which is a downlink-dominated use case. Unlike [64], to broaden the scope and applicability of the proposed comparison methodology, we do not enforce a specific TDD pattern, such as the 4:1 DDDSU frame structure (D – Downlink, S – Special, U – Uplink); instead we allow flexibility depending on the bandwidth and resource allocation for uplink traffic. In the downlink, we further assume flexible TDM allocation between backhaul and access links, while still adhering to the half-duplex constraint that prohibits simultaneous transmission and reception. Thus, depending on the uplink assumptions, our model aligns with either static or dynamic TDM as defined in [64]. Finally, our analysis is based on a scheduling window T that may differ from the transmission time interval (TTI) reflecting operator-specific choices and service requirements. For T greater than TTI, appropriate scaling is applied to ensure compatibility with half-duplex IAB operation. Applications with low latency sensitivity, such as multicast video delivery, naturally support larger values of T, which can help smooth out traffic variations and improve efficiency.

IV. Optimization Framework

In this section, we formalize the optimal resource allocation problem for serving both unicast and multicast traffic in IAB- and RIS-assisted 5G NR systems. The problem is modeled as a variable-cost, variable-size bin packing problem (BPP).

The objective is to determine: the optimal grouping of multicast UEs and the optimal association of both unicast and multicast UEs to: *(i)* the IAB donor or IAB nodes (in the IAB-based system) or *(ii)* the appropriate RIS (in the RIS-based system), while minimizing the overall service cost, defined as the ratio of occupied PRBs to total available PRBs over the entire time horizon.

We start by formulating the common representation of unicast and multicast traffic, applicable to both IAB- and RIS-system types, and organizing them into *subgroups* and *suits* to represent UE allocations. It then aims to identify the optimal way to group multicast UEs and associate all UEs (unicast and multicast) with either IAB donors/nodes or RISs – the framework is then formalized for two distinct system architectures. First, we present the RIS-based optimization model, due to its relatively simpler structure. For RIS-based networks, the focus is on single reflection paths, incorporating constraints related to multi-beam operation, power limitations, and PRB utilization. The IAB-based network formulation is more complex, additionally accounting for backhaul links, the inherent burnt PRBs due to in-band half-duplex operation of IAB nodes, and the hierarchical structure of the IAB network. Finally, operational limits and boundary conditions, such as transmit power, beam capacity, synchronization, and total PRB consumption, are formalized for both system types.

A. Unicast and Multicast Traffic Formulation

Multicasting.

Let $\mathcal{K}^m = \{1, \ldots, K^m\}$ denote the set of multicast UEs, each requiring a specific service e. These UEs are served via directional beams, where each beam covers a multicast subgroup, i.e., a subset of UEs within K^m.

There are $2^{K^m}-1$ possible non-empty subgroup combinations, each denoted as $\mathcal{K}_j^m$, where $j \in \mathcal{J}, \mathcal{J} = \{1, 2, \ldots, 2^{K^m}-1\}$. The size of a multicast subgroup j is represented as $|\mathcal{K}_j^m|$. Importantly, only UEs subscribing to the same service e can be grouped within the same multicast subgroup and served by a single directional beam.

1) Suits of Subgroups: To ensure full coverage of multicast UEs, we define suits—collections of non-overlapping multicast subgroups that together include all UEs in K^m. Each suit is denoted as $\mathcal{G}_k^m$ for $k^m = 1, 2, \ldots, |\Omega|$, where Ω is the set of all valid suits. Each UE appears exactly once in a suit, meaning

$$\bigcup_{j \in \mathcal{G}_k^m} \mathcal{K}_j^m = \mathcal{K}^m, k^m = 1, 2, \ldots, |\Omega|,$$
$$\mathcal{K}_{j_1}^m \bigcap \mathcal{K}_{j_2}^m = \emptyset, j_1 \neq j_2, \quad \forall j_1, j_2 \in \mathcal{G}_k^m. \tag{7}$$

This ensures that each multicast UE is assigned to exactly one subgroup per suit.

Unicasting: Let $\mathcal{K}^u = \{1, \ldots, K^u\}$ represent the set of unicast UEs, each served individually via a dedicated directional beam. For unicast traffic, we define suits $\mathcal{G}_k^u$, where each suit simply collects a subset of unicast UEs to be scheduled together in the optimization framework.

2) Combined Suits: To represent joint scheduling of both unicast and multicast traffic, we define combined suits $\mathcal{G}_k$, which are obtained by concatenating a multicast suit $\mathcal{G}_k^m$ and a unicast suit $\mathcal{G}_k^u$. Each combined suit $\mathcal{G}_k$ represents a candidate allocation configuration to be considered during optimization.

B. Problem Formalization: RIS Network

In an RIS-assisted 5G NR system, the focus is on the wireless access link, specifically on the single reflection paths from the NR BS to a RIS and subsequently from the RIS to the UE.

The NR BS operates with a total bandwidth W, a carrier frequency f_c, and a numerology μ, which determines the number and granularity of PRBs.

Unlike conventional systems where the BS communicates directly with UEs, the RIS-aided architecture leverages passive elements to reflect and manipulate the electromagnetic signal, thereby shaping the propagation environment to enhance coverage and signal quality. As a result, more PRBs are generally required to satisfy UEs' demands when compared to conventional direct-link NR BS systems. Indeed, this passive approach typically results in higher path loss and lower received signal strength compared to active, direct-link transmissions. Consequently, to achieve the same target data rates or service levels as in conventional NR BS systems, the network may need to compensate for the reduced signal quality by using more PRBs, particularly when multiple users are served simultaneously or when multicast traffic is involved.

1) BS Domain: In the context of a multi-beam RIS-aided NR BS where the number of simultaneously active beams is $L > 1$, the indices of the unicast and multicast subgroups within a suit $\mathcal{G}_k$ are further partitioned into subsets, each associated with a distinct beam. These subsets are denoted as $\mathcal{G}_k^l \subseteq \mathcal{G}_k$, for $l = 1, 2, \ldots, L$, where each subset $\mathcal{G}_k^l$ is scheduled on beam l.

To ensure proper scheduling and avoid overlap between beams, the following conditions must be satisfied:

$$\mathcal{G}_k = \bigcup_{l=1}^{L} \mathcal{G}_k^l,$$

$$\mathcal{G}_k^{l_1} \bigcap \mathcal{G}_k^{l_2} = \emptyset, l_1 \neq l_2, \quad \forall l_1, l_2 \in \{1, \ldots, L\}. \tag{8}$$

This formulation guarantees that: each subgroup is assigned to exactly one beam; all subgroups in the suit $\mathcal{G}_k$ are accounted for, and there is no duplication of subgroup assignments across beams.

2) Decision Variables for Subgroup Assignment: To represent the scheduling decisions over time, we define a binary decision variable, $g_j^t \in \{0, 1\}$, which indicates whether subgroup j is served during time slot t. Specifically, $g_j^t = 1$ if subgroup j is served at time slot t, while $g_j^t = 0$ otherwise.

To compactly express the subgroup scheduling status across all options in a given time slot, we define a binary indicator vector: $\mathbf{g}^t = (g_1^t, \ldots, g_{|\mathcal{J}|}^t)$ to represent the set of subgroups that are active (i.e., served) during slot t.

3) Operational Limits for Beams and Subgroups: At any given time slot $t \in \mathcal{T}$, the BS can simultaneously sweep at most L beams. Since each beam can serve one subgroup at a time, this imposes an upper limit of L subgroups that can be served concurrently. This beam scheduling constraint is expressed as:

$$\sum_{j \in \mathcal{G}k} g_j^t \leq L, \forall t \in \mathcal{T}. \tag{9}$$

In addition, we impose a constraint on the total service time for each suit $\mathcal{G}_k$. Specifically, the number of time slots during which a subgroup from beam l of suit k is scheduled must not exceed the total number of available time slots M in the scheduling horizon. This ensures that the suit is fully served within the given timeframe:

$$\sum_{j \in \mathcal{G}_k^l} \sum_{t \in \mathcal{T}} g_j^t \leq M, \quad \forall l = 1, \ldots, L, \quad \forall k = 1, \ldots, |\Omega|. \tag{10}$$

4) Transmit Power Constraints in Multi-Beam System: In a multi-beam system, the total transmit power available per antenna is constrained and must be distributed across the beams serving active subgroups. At any given time slot $t \in \mathcal{T}$, the aggregate power allocated to all scheduled subgroups must not exceed the maximum allowable transmit power $P_{\max}$. This constraint is expressed as:

$$\sum_{j \in \mathcal{G}_k} g_j^t P_j \leq P_{\max}, \forall t \in \mathcal{T}, \tag{11}$$

where P_j denotes the transmit power required to serve subgroup j, $g_j^t \in \{0, 1\}$ is the decision variable indicating whether subgroup j is scheduled at time t, and $P_{\max}$ is the maximum transmit power available at the antenna array during a single time slot. The power value P_j is computed using the propagation model described in Section III, taking into account the SINR threshold associated with the MCS selected for each subgroup [63].

5) Boundary Conditions in Resource Utilization: The number of the PRBs required to serve subgroup j through a beam emitted by the NR BS is denoted by a_j, and it depends on the session bitrate B_e, the spectral efficiency s_j corresponding to subgroup j, and the PRB size w_{PRB}:

$$a_j = \min_{r \in \mathcal{R}} \left(\frac{B_e}{s_j(r) w_{\text{PRB}}} \right), \tag{12}$$

where $\mathcal{R}$ represents the set of available RISs and the BS.

The time slot allocation for subgroup j is represented by the binary scheduling vector $\mathbf{g}_j = (g_j^1, \ldots, g_j^M)$, where the number of time slots required to allocate a_j PRBs under a per-beam capacity of R_b PRBs per slot satisfies:

$$\sum_{t \in \mathcal{T}} g_j^t = \left\lceil \frac{a_j}{R_b} \right\rceil, \; j \in \mathcal{J}. \tag{13}$$

Furthermore, to maintain beam consistency, each subgroup must be assigned the same beam throughout its service duration. This imposes the condition for any $j \in \mathcal{J}$

$$a_j \leq M R_b, \tag{14}$$

ensuring that the subgroup's PRB requirement fits within the maximum PRB budget per beam across the time horizon.

Finally, the total PRB consumption across all subgroups in suit $\mathcal{G}_k$ must not exceed the cumulative PRB capacity over all L beams and M time slots

$$\sum_{j \in \mathcal{G}_k} a_j \leq L M R_b. \tag{15}$$

6) Objective Function: The goal of the optimization framework is to minimize the resource utilization by reducing the proportion of PRBs consumed relative to the total PRB capacity available over the scheduling horizon. Specifically, the objective is to minimize the ratio of allocated PRBs to the maximum

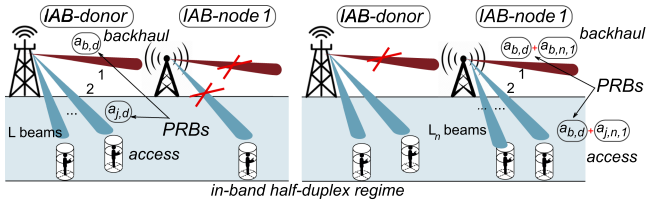


Fig. 3. Illustration of half-duplex constraint in IAB: simultaneous transmission and reception is not feasible at IAB donor and nodes.

available PRBs across all beams and time slots, subject to the constraints for subgroup coverage, beam-specific subgroup assignment, operational limits for beams and subgroups, transmit power budget, time slot assignment, and total PRB capacity across all beams:

$$\min_{k\in 1,\ldots,|\Omega|} \quad \sum_{j\in\mathcal{G}_k}\left\lceil \frac{a_j}{MLR_b}\right\rceil,$$
$$\text{s. t.} \quad (7),(8),(9),(10),(11),(13),(14),(15). \tag{16}$$

C. Problem Formalization: IAB Network

In an IAB-aided 5G NR system, the problem formulation must account for both the wireless access link (from the IAB node to the UE) and the backhaul link (from the IAB donor to the IAB node). All IAB entities—namely the donor and nodes—are configured with a common bandwidth W, carrier frequency f_c, and NR numerology μ, which determines the size and structure of the available PRBs.

A key constraint in IAB networks is their operation in in-band half-duplex mode, where an IAB node cannot simultaneously transmit to UEs and receive data from the donor over the backhaul. This restriction leads to resource underutilization: during the time slots when the node awaits data reception from the backhaul, its PRBs remain idle, effectively being "burned" despite being reserved.

Moreover, the total number of PRBs required to serve a UE through an IAB node is not limited to the access link only, but it includes also PRBs for the backhaul path (possibly through multiple hops). A schematic illustration of the IAB-based system is provided in Fig. 3.

1) IAB donor/IAB Node Domain: We define the subsuits $\mathcal{G}_k^{l_d}$ and $\mathcal{G}_k^{l_{n,i}}$ as the subgroup indices from $\mathcal{G}_k$ assigned to the beams of the IAB donor and IAB node i, respectively. Specifically, $\mathcal{G}_k^{l_d}\subseteq\mathcal{G}_k$ denotes the set of subgroups scheduled on beam $l_d=1,\ldots,L$ of the IAB donor, and $\mathcal{G}_k^{l_{n,i}}\subseteq\mathcal{G}_k$ denotes the subgroups scheduled on beam $l_n=1,\ldots,L_n$ of IAB node i, $\forall i\in\{1,\ldots,N_{IAB}\}$.

The overall set of subgroups $\mathcal{G}_k$ is then composed of:

$$\mathcal{G}_k=\left(\bigcup_{l_d=1}^{L_d}\mathcal{G}_k^{l_d}\right)\bigcup\left(\bigcup_{i=1}^{N_{IAB}}\bigcup_{l_n=1}^{L_n}\mathcal{G}_k^{l_{n,i}}\right), \tag{17}$$

ensuring that each UE is served by only one BS (either the IAB donor or a specific IAB node). To guarantee this, the following disjointness constraints must hold:

$$\mathcal{G}_k^{l_{d_1}}\bigcap\mathcal{G}_k^{l_{d_2}}=\emptyset, l_{d_1}\neq l_{d_2},\quad \forall l_{d_1},l_{d_2}\in\{1,\ldots,L\},$$
$$\mathcal{G}_k^{l_{n_1,i}}\bigcap\mathcal{G}_k^{l_{n_2,i}}=\emptyset, l_{n_1}\neq l_{n_2},\quad \forall l_{n_1},l_{n_2}\in\{1,\ldots,L_n\},$$
$$\mathcal{G}_k^{l_d}\bigcap\mathcal{G}_k^{l_{n,i}}=\emptyset,\quad \forall l_d\in\{1,\ldots,L\}, l_n\in\{1,\ldots,L_n\}. \tag{18}$$

respectively, ensuring that: no overlap occurs among subgroups assigned to different beams of the IAB donor; no overlap occurs among beams within each IAB node; and no overlap occurs between the donor and any IAB node.

2) Decision Variables for Subgroup Assignment: We define binary decision variables $g_{j,d}^{t_d}\in\{0,1\}$ and $g_{j,n,i}^{t_{n,i}}\in\{0,1\}$ to represent whether subgroup $j\in\mathcal{J}$ is served by the IAB donor or IAB node i, respectively, at a specific time slot. In particular, $g_{j,d}^{t_d}=1$, if subgroup j is served at time slot $t_d\in\mathcal{T}_d$ by a beam of the IAB donor, otherwise $g_{j,d}^{t_d}=0$. Similarly, $g_{j,n,i}^{t_{n,i}}=1$, if subgroup j is served at time $t_{n,i}\in\mathcal{T}_{n,i}$ by a beam of the IAB i, otherwise $g_{j,n,i}^{t_{n,i}}=0$.

Using these variables, we construct the following indicator matrices: $\mathbf{G}_d$ for the IAB donor, and $\mathbf{G}_{n,i}$, for each IAB node $i\in\{1,\ldots,N_{IAB}\}$. Each row $\mathbf{g}_{j,d}$ or $\mathbf{g}_{j,n,i}$ in the matrix indicates when (i.e., at which time slots) subgroup j is scheduled for service during the time horizon $\mathcal{T}_d$ or $\mathcal{T}_{n,i}$. Conversely, each column $\mathbf{g}^{t_d}$ or $\mathbf{g}^{t_{n,i}}$ specifies which subgroups are being served at a given time slot, t_d or $t_{n,i}$, by the corresponding entity.

3) Synchronization Constraint: To ensure proper synchronization of backhaul transmissions in the IAB network, we introduce the following binary variables: $g_b^{t_d}\in\{0,1\}$, which indicates whether the IAB donor is sweeping a beam for the backhaul connection to IAB node 1 at time slot $t_d\in\mathcal{T}_d$; and $g_b^{t_{n,i}}\in\{0,1\}$, which indicates whether IAB node i is sweeping a beam for the backhaul connection to IAB node $i+1$ at time slot $t_{n,i}\in\mathcal{T}_{n,i}$.

These binary indicators ensure that each backhaul transmission is scheduled synchronously with the corresponding reception window at the downstream IAB node. This constraint is essential to maintain correct timing alignment and avoid collisions or data loss due to mismatched access-backhaul scheduling in the in-band half-duplex IAB architecture.

4) Operational Limits for Beams and Subgroups: At any given time slot $t_d\in\mathcal{T}_d$ the IAB donor can sweep at most L beams, where one beam is reserved for the backhaul link, and the remaining $L-1$ beams can be allocated to serve access subgroups. Similarly, each IAB node $i=1,\ldots,N_{IAB}$ can simultaneously sweep up to L_n beams at time slots $t_{n,i}\in\mathcal{T}_{n,i}$, with one beam reserved for the backhaul connection to the next hop and the rest used for access.

These operational constraints are formalized as:

$$\sum_{j\in\mathcal{G}_k^{l_d}} g_{j,d}^{t_d}+g_b^{t_d}\leq L,\forall t_d\in\mathcal{T}_d,$$
$$\sum_{j\in\mathcal{G}_k^{l_{n,i}}} g_{j,n,i}^{t_{n,i}}+g_b^{t_{n,i}}\leq L_n,\forall t_{n,i}\in\mathcal{T}_{n,i}, i\in\{1,\ldots,N_{IAB}\}. \tag{19}$$

Moreover, within the scheduling horizon T, which may or may not align with the standard 5G NR subframe duration, the number of time slots allocated for serving all subgroups (including the backhaul) must not exceed the total number of available slots M for each beam direction. This leads to the following cumulative constraints:

$$\sum_{j\in\mathcal{G}_k^{l_d}}\sum_{t_d\in\mathcal{T}_d} g_{j,d}^{t_d}+\sum_{t_d\in\mathcal{T}_d} g_b^{t_d}\le M,$$

$$\sum_{j\in\mathcal{G}_{n,i}^{l_{n,i}}}\sum_{t_{n,i}\in\mathcal{T}_{n,i}} g_{j,n,i}^{t_{n,i}}+\sum_{t_{n,i}\in\mathcal{T}_{n,i}} g_b^{t_{n,i}}\le M,\forall i\in\{1,\ldots,N_{IAB}\}. \quad (20)$$

5) Transmit Power Limitations in Multi-Beam System: The total transmit power per antenna must remain within budgetary limits when serving any subgroup $j\in\mathcal{G}_k^{l_d}$ via the IAB donor, or $j\in\mathcal{G}_k^{l_{n,i}}$ via the IAB node i, for any time slot $t_d\in\mathcal{T}_d$ or $t_{n,i}\in\mathcal{T}_{n,i}$, respectively. The constraints are formulated as follows:

$$\left[\sum_{j\in\mathcal{G}_k^{l_d}} g_{j,d}^{t_d}+g_b^{t_d}\right]P_{j,d}\le P_{\max}^d,\forall t_d\in\mathcal{T}_d,$$

$$\left[\sum_{j\in\mathcal{G}_k^{l_{n,i}}} g_{j,n,i}^{t_{n,i}}+g_b^{t_{n,i}}\right]P_{j,n,i}\le P_{\max}^n,\forall t_{n,i}\in\mathcal{T}_{n,i},i\in\{1,\ldots,N_{IAB}\}, \quad (21)$$

where $P_{j,d}$ and $P_{j,n,i}$ represent the transmit power of the beam required to serve subgroup j via the IAB donor and IAB node i, respectively; $P_{\max}^d$ and $P_{\max}^n$ represent the maximum allowable power budget per antenna for the donor and each IAB node.

6) UE Association With IAB Donor and IAB Nodes: In our framework, the association of UEs to either the IAB donor or IAB nodes is determined by exhaustively exploring all possible groupings of UEs and their corresponding assignments to BSs, namely the IAB donor and IAB nodes.

Alternatively, the UE-to-BS association can be determined externally, using predefined criteria such as SINR, experienced throughput, latency, or packet loss. In such cases, the IAB donor may act as a central controller to dynamically assign UEs to access points based on real-time network metrics. To further enhance adaptability, load balancing techniques can be implemented to evenly distribute UEs across the network, preventing bottlenecks and ensuring efficient resource utilization. Additionally, the use of Artificial Intelligence (AI) or Machine Learning (ML) models can improve the decision-making process by predicting user mobility patterns, traffic demand, and channel conditions [65].

Our framework is designed to be flexible and supports such externally determined associations by allowing the association variables to be pre-fixed or parameterized, thereby accommodating both optimization-based and heuristic or AI-driven approaches.

7) Boundary Conditions in Resource Utilization: We define the cost of serving subgroup $j\in J$ in terms of the required number of PRBs allocated by the IAB donor and IAB nodes. These costs are expressed as $a_{j,d}$ and $a_{j,n,i}$, respectively, and are computed based on the session bitrate B_e, spectral efficiency $s_{j,d}$ or $s_{j,n,i}$, and the PRB size w_{PRB}:

$$a_{j,d}=\frac{B_e}{s_{j,d}w_{\text{PRB}}},\ a_{j,n,i}=\frac{B_e}{s_{j,n,i}w_{\text{PRB}}}+a_{b,d}+\sum_{i'=1}^{i-1}a_{b,n,i'}, \quad (22)$$

where $a_{b,d}$ represents the PRBs required for the backhaul link between the IAB donor and its immediate neighboring IAB node, and $a_{b,n,i'}$ denotes the PRBs for backhaul links between intermediate IAB nodes i' and its neighboring IAB node $i'+1$ up to node $i-1$. Due to the in-band half-duplex operation of IAB networks, IAB nodes cannot simultaneously transmit to UEs and receive data from upstream nodes. As such, the total PRB cost for any UE associated with an IAB node includes both the access and backhaul PRBs.

The time allocation for subgroup j is captured through binary vectors $\mathbf{g}_{j,d}=(g_{j,d}^1,\ldots,g_{j,d}^M)$ and $\mathbf{g}_{j,n,i}=(g_{j,n,i}^1,\ldots,g_{j,n,i}^M)$, i.e., rows of matrices $\mathbf{G}_d$ and $\mathbf{G}_{n,i}$. The elements of these vectors give the time interval duration for serving subgroup j by the IAB donor and IAB node i:

$$\sum_{t_d\in\mathcal{T}_d}(g_{j,d}^{t_d}+g_b^{t_d})=\left\lceil\frac{a_{j,d}+a_{b,d}}{R_b}\right\rceil,$$

$$\sum_{t_{n,i}\in\mathcal{T}_{n,i}}(g_{j,n,i}^{t_{n,i}}+g_b^{t_{n,i}})=\left\lceil\frac{a_{j,n,i}+a_{b,n,i}}{R_b}\right\rceil,\ \forall i\in\{1,\ldots,N_{IAB}\}. \quad (23)$$

To ensure feasibility, each beam must serve its assigned subgroup for the entire service duration. Hence, for any subgroup $j\in\mathcal{J}$, the resource usage is constrained by:

$$a_{j,d}\le MR_b,\ a_{j,n,i}\le MR_b,\forall i\in\{1,\ldots,N_{IAB}\},$$

$$\sum_{j\in\mathcal{G}_k^{l_d}} a_{j,d}+a_{b,d}\le LMR_b,$$

$$\sum_{j\in\mathcal{G}_k^{l_{n,i}}} a_{j,n,i}+a_{b,n,i}\le L_nMR_b,\forall i\in\{1,\ldots,N_{IAB}\}. \quad (24)$$

8) Objective Functions: Given the half-duplex constraint that prevents IAB donors and nodes from transmitting and receiving simultaneously, the objective is to minimize the resource consumption over the entire time horizon. The total cost is measured as the sum of PRBs used for all subgroups, served by either the IAB donor or IAB nodes, subject to constraints related to UE grouping and coverage, beam scheduling and sweep limits, transmit power budget, and slot allocations and service time:

$$\min_{k\in1,\ldots,|\Omega|}\quad\sum_{j\in\mathcal{G}_k}\left\lceil\frac{a_{j,d}}{MLR_b}+\sum_{i=1}^{N_{IAB}}\frac{a_{j,n,i}}{N_{IAB}ML_nR_b}\right\rceil,$$

$$\text{s. t.}\quad (7),(17),(18),(19),(20),(21),(23),(24). \quad (25)$$

V. Numerical Results

In this section, we present the results of our simulation campaign. We begin by outlining the evaluated system configuration

and simulation setup. Then, we analyze the performance of IAB- and RIS-based systems under varying UE densities and numbers of RIS/IAB nodes, based on relevant performance metrics. We assess and compare the deployment costs associated with RIS and IAB nodes. Finally, we present the results for scenarios with increased numbers of UEs as well as higher numbers of RIS and IAB nodes.

A. Simulation Settings and Performance Metrics

The system configuration for both RIS- and IAB-based networks is designed following the models and assumptions presented in Section III. This includes the deployment of UEs, NR BS, RISs, IAB donors, and IAB nodes, along with assumptions related to traffic generation, resource allocation, propagation environment, blockage models, and antenna settings.

Our simulations consider a scenario with $2-8$ multicast UEs, uniformly distributed across the service area, along with unicast UEs, up to a total of 16 UEs. The available system bandwidth is set to $W = 100$ MHz, and we adopt NR numerology $\mu = 3$, which corresponds to a PRB size $w_{\text{PRB}} = 1.44$ MHz, reflecting typical mmWave system parameters. We examine a standard 1ms subframe in 5G NR, which contains $M = 8$ time slots. Therefore, when we consider a simulation duration equivalent to one 1ms subframe, we are inherently limited to M=8 time slots.The carrier frequency is set to $f_c = 30$ GHz [59], and all base station and IAB donor/node are equipped with 64×4 antenna arrays (64 antenna elements in the horizontal dimension and 4in the vertical), enabling directional beamforming in both azimuth and elevation. RIS units consist of 8100 reflective elements, each with a reflection efficiency of 1 [66].

The transmit power budget is fixed at 40 dBm for the NR BS and IAB donor, and 33 dBm for IAB nodes [35], [59]. Each UE session is modeled with a fixed data rate of 10 Mbps. Shadow fading is modeled as a zero-mean Gaussian random variable with a standard deviation of $\sigma = 4$, and an interference margin of 3 dB is applied [59]. A complete list of default simulation parameters is provided in Table II.

We evaluate two operational regimes in our study: capacity-limited and outage-limited.

To establish a baseline, we first determine the maximum coverage area of a standalone NR BS or IAB donor (i.e., with no RIS or IAB relays in between). Based on our simulations, the maximum achievable cell radius in such a configuration is $R = 350$ m. This ensures that no UE in the blocked state experiences an outage, based on a minimum SNR threshold of S_{th}=-9.47 dB.

Next, to simulate an outage-limited regime, we double the area, resulting in dimensions of 25 m $\times$ 700 m. Under this condition, some UEs enter outage states due to blockages or distance-related attenuation. We then densify the environment by progressively adding 1 to 3 RISs or IAB nodes to enhance coverage and improve system performance.

To evaluate and compare the performance of IAB and RIS deployments, we use the following key metrics:

- *Average number of multicast UEs served in multicast mode*, excluding those that fall back to unicast delivery, reflecting efficiency of multicasting.

TABLE II
DEFAULT PARAMETERS FOR NUMERICAL ASSESSMENT

Parameter	Value
Number of multicast UEs, K^m	vary
Number of unicast UEs, K^u	vary
Number of IAB nodes, N_{IAB}	vary
Number of RISs, N_{RIS}	vary
Area of interest	25 m x 700 m
Operating frequency, f_c	30 GHz
Bandwidth, W	100 MHz
PRB size, w_{PRB}	1.44 MHz
Subcarrier spacing, Δ	0.06 MHz
Height of BS, h_A	10 m
Height of blocker, h_B	1.7 m
Height of UE, h_U	1.5 m
Shadow fading SDT, σ	4
SINR threshold, S_{th}	-9.47 dB
Transmit power, $P_{\max}/P^d_{\max}/P^n_{\max}$	40/40/33 dBm
Power spectral density of noise, N_0	-174 dBm/Hz
Guaranteed bit rate, B_e	10 Mbps
BS antenna array	64×4
Number of reflective elements per RIS, N_{re}	8100
Subframe duration	1 ms
Slot duration	125 μs
5G NR numerology, μ	3
Number of time slots, M	8
Number of available resource blocks, R_b	66
Number of beams, L/L_n	3/2

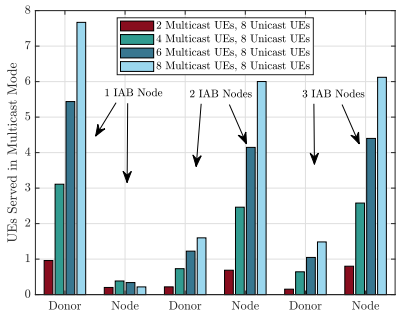
- *Average number of activated access beams*, i.e., the number of selected suits required to serve all unicast and multicast UEs, reflecting efficiency of antenna array utilization.
- *Total number of utilized PRBs* across the system, reflecting the overall resource efficiency.

B. Performance Assessment: IAB Network

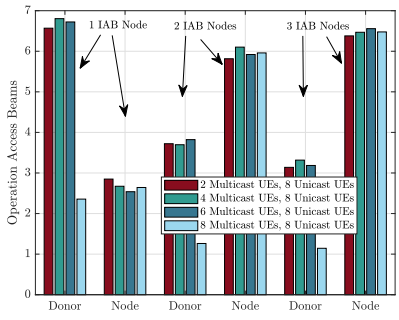
1) Impact of the Number of Multicast UEs: We begin our evaluation of the IAB network by analyzing the impact of the number of multicast UEs, keeping the number of unicast UEs fixed at 8, while varying the number of deployed IAB nodes (see Fig. 4).

The first metric of interest, shown in Fig. 4(a), is the average number of multicast UEs served in multicast mode (i.e., served in groups of at least two). This metric also indicates the required HPBW to serve multicast UEs; as larger groups generally need wider beams to simultaneously cover all users. When only one IAB node is deployed, most multicast UEs are served by the donor due to its higher transmit power and reduced dependency on backhaul links. As the number of IAB nodes increases, this trend reverses: more multicast UEs are served directly by the IAB nodes, offloading traffic from the donor.

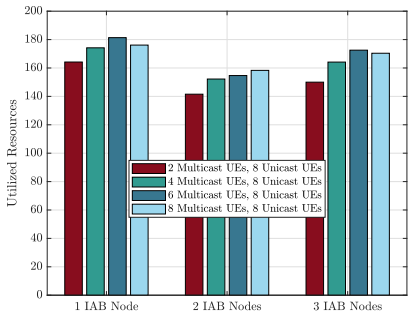
Next, Fig. 4(b) presents the average number of downlink beams activated to serve all UEs in the system. Similar to the previous metric in Fig. 4(a), the number of beams activated by the IAB donor decreases with an increasing number of deployed IAB nodes. Interestingly, the relationship between the number of multicast UEs and beam activation is not linear. Generally, the number of beams follows a parabolic trend: initially increasing with the number of multicast UEs before decreasing. However, with a single IAB node deployment for the IAB node and with a triple IAB node deployment for the IAB donor, beam usage



(a) Average number of UEs served in multicast mode



(b) Average number of access beams to serve UEs



(c) Total number of utilized resources

Fig. 4. Performance metrics with respect to **IAB** densification with 8 unicast UEs.

decreases, indicating the formation of larger multicast groups. As the quantity of nodes increases, there is a corresponding escalation in the number of beams associated with these nodes.

Finally, Fig. 4(c) illustrates the total number of PRBs utilized in the system. The lowest resource utilization occurs with an intermediate deployment of two IAB nodes, which strikes a balance between access and backhaul resource demands. With only one IAB node, higher resource consumption arises from longer downlink transmission distances and less spatial reuse. On the other hand, deploying 3 IAB nodes increases the demand on backhaul resources, which adds to overall PRB usage. Thus, a moderate number of IAB nodes, i.e., 2, provides the most resource-efficient configuration by optimizing both link distances and the load across access and backhaul.

2) Impact of the Number of Unicast UEs: Since this work considers the coexistence of unicast and multicast traffic, it is also important to examine how the number of unicast UEs affects overall system performance. To this end, we vary the number of both multicast and unicast UEs from 2 to 8, across different IAB node settings. The results are presented as 3D bar charts in Fig. 5.

Fig. 5(a) shows a predictable increase in the number of activated beams at the IAB donor as the number of unicast UEs grows. However, the rate of increase in beam count is sublinear with respect to the number of unicast UEs. This suggests that multicast groups become larger when more unicast traffic is present and demands more resources. This is likely due to the system favoring multicast grouping to save resources under higher load.

A similar trend is observed for the IAB nodes, as shown in Fig. 5(b). The key difference is that while the IAB donor reduces its beam usage with increased IAB node deployment – thanks to traffic offloading – the IAB nodes themselves activate more beams as their density increases. This reflects their growing role in directly serving unicast UEs, often using narrower HPBWs to efficiently target smaller user groups.

Finally, Fig. 5(c) evaluates total PRB usage across the system. Resource utilization increases nearly linearly with the increase in the total number of UEs, driven primarily by the rise in unicast traffic, where each UE typically demands dedicated resources. However, when varying the IAB node density, a parabolic trend emerges: the lowest resource consumption is observed with a moderate IAB deployment (e.g., 2 nodes), which offers a favorable balance between access efficiency and backhaul overhead.

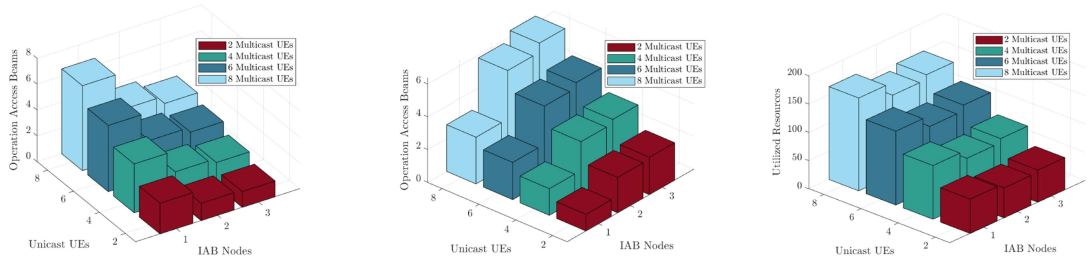
These observations highlight a key insight: while multicast grouping can improve resource efficiency, the growing demand from unicast users can quickly dominate system capacity. Therefore, careful planning of IAB node placement and density is critical to optimize performance, reduce overhead, and ensure scalability under mixed traffic conditions.

C. Performance Assessment: RIS Network

In this section, we perform a parallel analysis for RIS-based networks under equivalent configurations and compare their performance with IAB-based deployments.

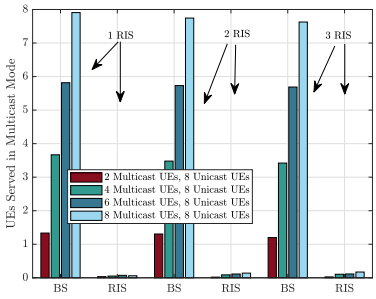
1) Impact of the Number of Multicast UEs: As with the IAB performance analysis, we begin by evaluating the impact of the number of multicast UEs on three key metrics: *(i)* the average number of UEs served in multicast mode, *(ii)* the average number of access beams activated to serve all UEs, and *(iii)* the total number of resources utilized. The results, observed under varying RIS deployment densities, are presented in Fig. 6.

Fig. 6(a) reveals a contrasting trend compared to the IAB scenario. While in IAB deployments, the donor only dominates multicast service under low IAB node number, in RIS-based networks, the BS consistently serves the majority of multicast UEs. This dominance persists even as the number of RISs increases. Notably, the absolute number of multicast UEs served by the BS in the RIS case is higher than that served by any single entity in the IAB network. This highlights the key role of RISs: enhancing the effective channel quality between the BS and UEs – especially in environments where direct line-of-sight is obstructed. Interestingly, although the total number of

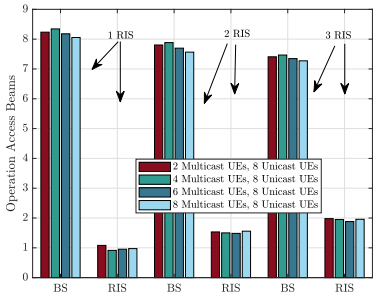


(a) Average number of access beams: donor (b) Average number of access beams: node (c) Total number of utilized resources

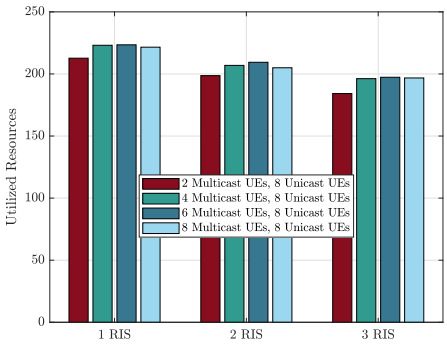
Fig. 5. Performance metrics for **IAB** deployment with respect to both IAB and unicast UE densification.



(a) Average number of UEs served in multicast mode



(b) Average number of access beams activated to serve UEs



(c) Total number of utilized resources

Fig. 6. Performance metrics with respect to **RIS** densification with 8 unicast UEs.

multicast UEs increases, the number served by RISs remains almost constant, suggesting that the BS remains the dominant transmitter in multicast delivery.

Fig. 6(b) illustrates the average number of beams activated by the BS and RISs. As expected, the BS activates more sequential beams to serve a growing number of multicast UEs. However, this relationship is nonlinear: it follows a parabolic trend, similar to what was observed with IAB donor beams under low node density. Nevertheless, the correlation is relatively weak, implying that increasing the number of multicast UEs does not significantly affect the total number of beams (or multicast suits). RISs, on the other hand, require fewer beams than the BS, with beam count decreasing as the number of multicast UEs increases – a sign that RISs are able to serve more users per beam. However, RIS beam usage increases with the number of available RISs, due to improved channel conditions that make RISs more favorable when UEs are located closer to them.

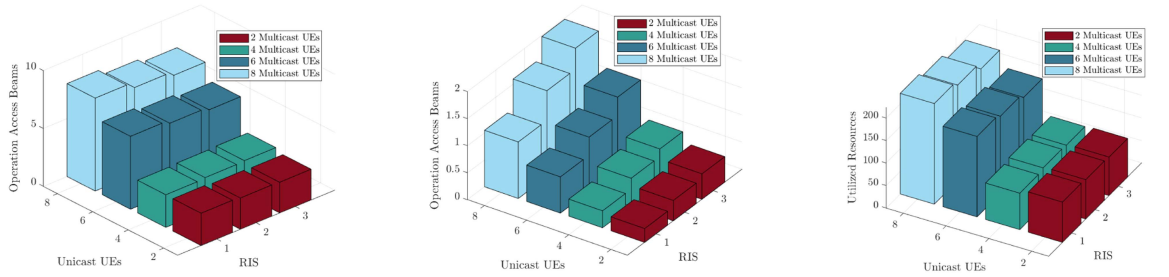
Finally, Fig. 6(c) presents the total PRB consumption required to serve all UEs. Compared to IAB networks, RIS-assisted deployments tend to consume significantly more resources. This is primarily due to longer signal paths: instead of direct communication, signals typically travel from the BS to the RIS and then to the UE, increasing path loss and overall transmission cost. Despite the improved channel quality offered by RISs, these longer effective distances lead to higher PRB demands. Moreover, unlike IAB networks, where backhaul resources are transmitted once and then reused by the nodes, RIS deployments lack data reuse capabilities. In RIS systems, each user must receive the full transmission from the BS, even when served together.

Interestingly, however, RIS densification leads to a linear reduction in resource consumption. This is because with more RISs, UEs are likely to be closer to a reflector, improving the link quality and reducing required PRBs. This observation emphasizes the positive impact of RIS density on resource efficiency, despite the inherently higher consumption compared to IAB.

2) Impact of the Number of Unicast UEs: We now assess the impact of unicast UE presence on the performance of the RIS-based network, as illustrated in Fig. 7.

Fig. 7(a) and Fig. 7(b) display a linear increase in the number of activated beams as both the number of RISs and the number of unicast UEs rise. This behavior mirrors the trends observed in Fig. 6(b), where unicast UEs play a significant role in shaping both beam activation patterns and multicast group (or suite) formation. As more unicast UEs are introduced, the system must allocate additional beams to accommodate individual connections, even when RISs are deployed.

Turning to resource utilization, Fig. 7(c) highlights a substantial increase in the total number of PRBs consumed as the number of unicast UEs grows. This confirms a key challenge in



(a) Average number of access beams: BS (b) Average number of access beams: RIS (c) Total number of utilized resources

Fig. 7. Performance metrics for RIS deployment with respect to both **RIS** and unicast UE densification.

RIS-assisted networks: the coexistence of unicast and multicast traffic. Because unicast transmissions typically require dedicated resources, their growing presence can strain the system and limit the capacity to efficiently serve multicast groups. As unicast UEs increase, more PRBs are allocated to unicast traffic, which in turn reduces the pool of available resources for multicast services. This imbalance can lead to multicast UEs being grouped into larger, less efficient multicast groups, resulting in degraded quality of service.

These observations highlight the trade-offs inherent in RIS-based network design, particularly in scenarios with mixed traffic types. To fully evaluate the suitability of RIS and IAB technologies for practical deployments, it is also important to consider not just performance metrics, but cost efficiency. Thus, in the next section, we shift focus to a deployment cost analysis for both IAB and RIS systems, offering a more comprehensive (multi-faceted) perspective on the trade-offs between resource efficiency, performance, and economic viability.

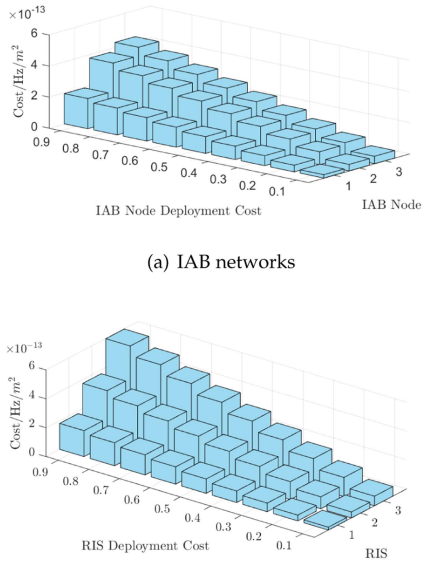
D. Deployment Cost Comparison

In this section, we analyze and compare the deployment costs of IAB and RIS systems, focusing on a unified metric: cost per Hz per square meter. For IAB networks, this cost reflects the deployment expense of each IAB entity – considering both access and backhaul links – normalized by bandwidth and coverage area. In contrast, for RIS networks, the cost is computed based solely on access-related bandwidth, as RISs do not handle backhaul communication. This unified metric enables a direct economic comparison between IAB and RIS technologies, helping network operators determine the most cost-effective solution based on specific deployment requirements and constraints.

We assess the sensitivity of this cost metric to two main factors: the number of deployed entities (i.e., IAB nodes or RISs) and the individual cost associated with each entity. These results are presented in Fig. 8.

When analyzing the deployment of a single network entity (either an IAB node or an RIS), RIS consistently emerges as the more cost-efficient option across a wide range of normalized deployment costs. This is primarily due to the simpler architecture and passive nature of RISs, which generally translates into lower per-unit costs compared to active IAB nodes.

However, this initial cost advantage of RISs diminishes – and may even reverse – as the number of deployed entities



(a) IAB networks

(b) RIS networks

Fig. 8. Densification cost analysis.

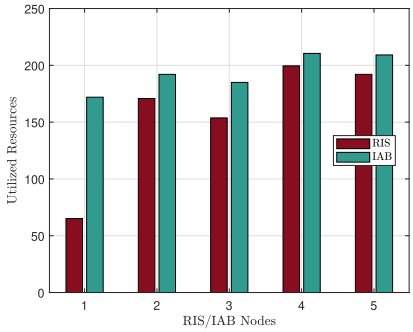
increases. Specifically, when deploying two or three entities, IAB becomes more cost-effective, depending on the normalized cost values. This suggests that while RISs offer a lower entry cost, scaling RIS deployment may be more resource-intensive in terms of cost-per-area and bandwidth efficiency. This shift in cost-effectiveness arises from the fact that each RIS contributes incrementally less to overall performance as more are deployed, while the IAB architecture benefits from more efficient resource reuse, particularly via shared backhaul transmissions.

In practical scenarios, the actual cost of deploying RIS and IAB nodes may vary substantially. RISs, with their passive nature, lower hardware complexity, and minimal power requirements, can potentially be significantly cheaper to deploy and maintain. Our analysis shows that if the cost of a single RIS is at least 0.2 points lower than that of a comparable IAB node, RIS deployments become consistently more cost-effective, regardless of the number of entities deployed. For example, if the cost to deploy one IAB node is 0.8 (normalized), and a RIS costs 0.5, the RIS solution offers a clear economic advantage.

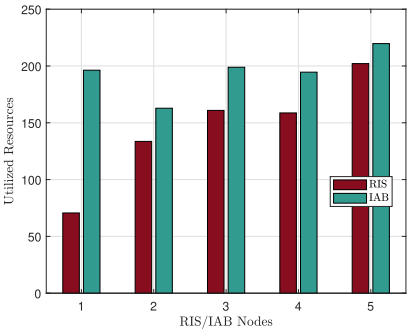
Table III summarizes the performance and cost trade-offs between selecting RIS and IAB technologies for a configuration involving 8 multicast and 8 unicast UEs.

TABLE III
TECHNOLOGY CHOICE PER CONFIGURATION WITH 8 MULTICAST AND 8 UNICAST UES

Number of nodes	Resources	Cost
1 node	IAB	RIS
2+ nodes	IAB	IAB



(a) Uniformly distributed UEs



(b) UEs distributed in clusters

Fig. 9. Total number of utilized resources with respect to **IAB/RIS** densification with 10 multicast UEs and 10 unicast UEs.

These findings highlight the potential for RISs to provide meaningful cost savings, especially when hardware and installation expenses are kept low. However, realizing this advantage depends on several practical factors that go beyond theoretical cost modeling. Further investigation is warranted to explore deployment logistics, maintenance overhead, power consumption, and long-term operational costs, all of which influence the real-world cost-efficiency of IAB and RIS technologies.

E. Densification Scenario

Finally, we present the results from a densification scenario, featuring a higher number of UEs and an increased deployment of both RISs and IAB nodes, as shown in Fig. 9. This analysis assesses the total number of utilized resources with respect to IAB/RIS densification, involving 10 multicast UEs and 10 unicast UEs. We compare two UE deployment distributions: a uniform distribution and a clustered distribution, shown in Fig. 9(a) and Fig. 9(b), respectively. In the latter, users were grouped into 5 clusters.

Recall that the total number of utilized resources demonstrated in Fig. 4 and Fig. 6 for a less densified scenario shows that IAB outperforms RIS with respect to this metric. Now, when the number of UEs increases, we observe the reverse trend. Specifically, RIS-based deployment starts to outperform IAB-based one in terms of the amount of utilized resources. For just one RIS we even observe up to a 3 times improvement factor. However, when more than a single RIS is utilized, the amount of utilized resources increases. In contrast, IAB deployments exhibit almost no changes in terms of the considered metric when the deployment is densified. This behavior is explained by the fact that multicast users tend to cluster, allowing RIS deployments to serve them efficiently as larger groups. However, in IAB deployments, these clusters are often served by multiple IAB nodes, which increases resource demands on the access interface.

VI. CONCLUSION

Network densification has become a vital strategy for operators to overcome the limitations imposed by blockage in 5G mmWave NR systems. To this end, 3GPP has proposed two cost-effective densification technologies: IAB and RIS. This paper aimed to comparatively evaluate these two solutions under realistic traffic conditions, characterized by a coexistence of unicast and multicast services at the air interface. To support this goal, we developed a unified mathematical optimization framework capable of modeling both IAB and RIS-assisted deployments. This framework enabled a rigorous comparison of the two technologies in terms of resource utilization and deployment costs.

Our numerical results indicate that RIS-assisted networks generally allow the BS to serve a greater number of multicast UEs compared to IAB-based systems. However, this advantage comes at the cost of higher resource consumption in sparse deployments, largely due to longer communication paths that involve signal reflection. However, RISs offer enhanced scalability and resource efficiency over IABs in scenarios with high deployment density, achieving an improvement factor of up to 3 times. Instead, IAB networks display a more complex relationship between node density and system performance, revealing the existence of an optimal IAB node density that balances resource efficiency and coverage.

In contrast, RIS networks benefit consistently from densification, as the overall resource usage decreases with the deployment of additional RIS units. In terms of deployment costs, RIS consistently outperforms IAB, especially when the unit cost of RIS is significantly lower than that of an IAB node. This makes RIS an attractive option for cost-sensitive deployments, particularly in scenarios where rich reflective environments are available. Summarizing, RIS-based densification is shown to be comparable to IAB in supporting both unicast and multicast services, while offering superior cost-efficiency.

While this study focuses on roadside deployments of RIS- and IAB-enhanced 5G NR systems, note that the developed framework is generalizable. Future work will explore more complex environments, such as urban grid deployments with BS and RIS/IAB nodes placed at intersections.

REFERENCES

[1] W. Jiang et al., "Terahertz communications and sensing for 6G and beyond: A comprehensive review," *IEEE Commun. Surveys Tut.*, vol. 26, no. 4, pp. 2326–2381, Fourthquarter 2024.

[2] Y. J. Guo and R. W. Ziolkowski, *Advanced Antenna Array Engineering for 6G and Beyond Wireless Communications*. Hoboken, NJ, USA: Wiley, 2021.

[3] A. Shafie, N. Yang, S. Durrani, X. Zhou, C. Han, and M. Juntti, "Coverage analysis for 3D terahertz communication systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1817–1832, Jun. 2021.

[4] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1791–1808, Mar. 2017.

[5] E. Sopin, D. Moltchanov, A. Daraseliya, Y. Koucheryavy, and Y. Gaidamaka, "User association and multi-connectivity strategies in joint terahertz and millimeter wave 6G systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 12765–12781, Dec. 2022.

[6] V. Naumov, Y. Gaidamaka, N. Yarkina, and K. Samouylov, *Matrix and Analytical Methods for Performance Analysis of Telecommunication Systems*. Berlin, Germany, Springer, 2022.

[7] N. Chukhno et al., "Models, methods, and solutions for multicasting in 5G/6G mmWave and sub-THz systems," *IEEE Commun. Surveys Tut.*, vol. 26, no. 1, pp. 119–159, Firstquarter 2024.

[8] N. Chukhno et al., "Optimal multicasting in millimeter wave 5G NR with multi-beam directional antennas," *IEEE Trans. Mobile Comput.*, vol. 22, no. 6, pp. 3572–3588, Jun. 2023.

[9] "Study on integrated access and Backhaul (release 16)," 3GPP, Sophia Antipolis Cedex, France, Tech. Rep. TR 38.874, Jan. 2019.

[10] D. Pugliese et al., "Integrating terrestrial and non-terrestrial networks via IAB technology: System-level design and evaluation," *Comput. Netw.*, vol. 253, 2024, Art. no. 110726.

[11] A. Kaushik et al., "Multi-function reconfigurable intelligent and holographic surfaces for 6G networks," *IEEE Netw.*, vol. 39, no. 1, pp. 10–13, Jan. 2025.

[12] P. Fiore, E. Moro, I. Filippini, A. Capone, and D. De Donno, "Boosting 5G mm-wave IAB reliability with reconfigurable intelligent surfaces," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2022, pp. 758–763.

[13] G. Leone, E. Moro, I. Filippini, A. Capone, and D. De Donno, "Towards reliable mmWave 6G RAN: Reconfigurable surfaces, smart repeaters, or both?," in *Proc. 20th Int. Symp. Model. Optim. Mobile, Ad hoc, Wireless Netw.*, 2022, pp. 81–88.

[14] G. Brancati, O. Chukhno, N. Chukhno, and G. Araniti, "Reconfigurable intelligent surface placement in 5G NR/6G: Optimization and performance analysis," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2022, pp. 1–6.

[15] G. Brancati, E. F. Pupo, O. Chukhno, N. Chukhno, M. Murroni, and G. Araniti, "Reconfigurable intelligent surface deployment and orientation in beyond 5G multicast networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, 2023, pp. 1–6.

[16] K. Boussetta and A.-L. Belyot, "Multirate resource sharing for unicast and multicast connections," in *Broadband Communications: Convergence of Network Technologies*. Berlin, Germany, Springer, 2000, pp. 561–570.

[17] S. Naribole and E. Knightly, "Scalable multicast in highly-directional 60-GHz WLANs," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2844–2857, Oct. 2017.

[18] A. Biason 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.

[19] Y. Niu, Y. Liu, Y. Li, X. Chen, Z. Zhong, and Z. Han, "Device-to-device communications enabled energy efficient multicast scheduling in mmWave small cells," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1093–1109, Mar. 2018.

[20] Y. Niu, L. Yu, Y. Li, Z. Zhong, and B. Ai, "Device-to-device communications enabled multicast scheduling for mmWave small cells using multi-level codebooks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2724–2738, Mar. 2019.

[21] M. Fallgren et al., "Multicast and broadcast enablers for high-performing cellular V2X systems," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 454–463, Jun. 2019.

[22] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, "NOMA-enabled cooperative unicast–multicast: Design and outage analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7870–7889, Dec. 2017.

[23] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[24] A. Samuylov et al., "Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3421–3434, May 2020.

[25] E. Garro et al., "5G mixed mode: NR multicast-broadcast services," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 390–403, Jun. 2020.

[26] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Efficient management of multicast traffic in directional mmWave networks," *IEEE Trans. Broadcast.*, vol. 67, no. 3, pp. 593–605, Sep. 2021.

[27] O. Chukhno et al., "Optimal multicasting in dual mmWave/ μ wave 5G NR deployments with multi-beam directional antennas," *IEEE Trans. Broadcast.*, vol. 69, no. 4, pp. 840–855, Dec. 2023.

[28] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Beyond complexity limits: Machine learning for sidelink-assisted mmWave multicasting in 6G," *IEEE Trans. Broadcast.*, vol. 70, no. 3, pp. 1076–1090, Sep. 2024.

[29] Nokia, "Integrated access and backhaul: Why it is essential for mmWave deployments," Accessed: May 16, 2023. [Online]. Available: https://www.nokia.com/blog/integrated-access-and-backhaul-why-it-is-essential-for-mmwave-deployments/

[30] Qualcomm, "What's in the future of 5G millimeter wave," Accessed: May 16, 2023. [Online]. Available: https://www.qualcomm.com/news/onq/2021/01/whats-future-5g-millimeter-wave

[31] Y. Sadovaya et al., "Self-interference assessment and mitigation in 3GPP IAB deployments," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[32] M. Dibaei and A. Ghaffari, "Full-duplex medium access control protocols in wireless networks: A survey," *Wireless Netw.*, vol. 26, no. 4, pp. 2825–2843, 2020.

[33] Light Reading, "Verizon to use 'integrated access backhaul' for fiber-less 5G," Accessed: May 16, 2023. [Online]. Available: https://www.lightreading.com/mobile/5G/verizon-to-use-integrated-access-backhaul-for-fiber-less-5g%20/d/d-id/754752

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

[35] Y. Sadovaya et al., "Integrated access and backhaul in millimeter-wave cellular: Benefits and challenges," *IEEE Commun. Mag.*, vol. 60, no. 9, pp. 81–86, Sep. 2022.

[36] N. Tafintsev et al., "Joint path selection and resource allocation in multi-hop mmWave-based IAB systems," in *Proc. IEEE Int. Conf. Commun.*, 2023, pp. 4194–4199.

[37] O. Chukhno et al., "Mission-critical connectivity enhanced by IAB in beyond 5G: Interplay of sidelink, directional unicasting, and multicasting," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1826–1838, 2023.

[38] R. Liu, Q. Wu, M. Di Renzo, and Y. Yuan, "A path to smart radio environments: An industrial viewpoint on reconfigurable intelligent surfaces," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 202–208, Feb. 2022.

[39] ETSI, "Reconfigurable intelligent surfaces (RIS); Use cases, deployment scenarios and requirements," GR RIS 001 V1.1.1, Apr. 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/RIS/001_099/001/01.01.01_60/gr_RIS001v010101p.pdf

[40] ETSI, "Reconfigurable intelligent surfaces (RIS); Communication models, channel models, channel estimation and evaluation methodology," GR RIS 003 V1.1.1, Jun. 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/RIS/001_099/003/01.01.01_60/gr_RIS003v010101p.pdf

[41] ETSI, "Reconfigurable intelligent surfaces (RIS); Technological challenges, architecture and impact on standardization," GR RIS 002 V1.1.1, Aug. 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/RIS/001_099/002/01.01.01_60/gr_RIS002v010101p.pdf

[42] G. C. Trichopoulos et al., "Design and evaluation of reconfigurable intelligent surfaces in real-world environment," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 462–474, 2022.

[43] N. Chukhno, A. Orsino, J. Torsner, A. Iera, and G. Araniti, "5G NR sidelink multi-hop transmission in public safety and factory automation scenarios," *IEEE Netw.*, vol. 37, no. 5, pp. 129–136, Sep. 2023.

[44] RIS Tech Alliance (RISTA), "Reconfigurable intelligent surface, white paper," v 1.0, 2023. [Online]. Available: https://www.risalliance.com/en/riswp2023_en/

[45] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 90–96, Dec. 2020.

[46] Y. Liu et al., "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tut.*, vol. 23, no. 3, pp. 1546–1577, thirdquarter 2021.

[47] C. Pan et al., "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[48] M. Mohammadi, Z. Mobini, D. Galappaththige, and C. Tellambura, "A comprehensive survey on full-duplex communication: Current solutions, future trends, and open issues," *IEEE Commun. Surveys Tut.*, vol. 25, no. 4, pp. 2190–2244, Fourthquarter 2023.

[49] A. Zappone, M. Di Renzo, X. Xi, and M. Debbah, "On the optimal number of reflecting elements for reconfigurable intelligent surfaces," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 464–468, Mar. 2021.
[50] G. Interdonato, F. Di Murro, C. D'Andrea, G. Di Gennaro, and S. Buzzi, "Approaching massive MIMO performance with reconfigurable intelligent surfaces: We do not need many antennas," 2022, *arXiv:2203.07493*.
[51] Z. Zhang et al., "Active RIS vs. passive RIS: Which will prevail in 6G?," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2023.
[52] K. Zhi, C. Pan, H. Ren, K. K. Chai, and M. Elkashlan, "Active RIS versus passive RIS: Which is superior with the same power budget?," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1150–1154, May 2022.
[53] M. Di Renzo et al., "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 798–807, 2020.
[54] X. Cao et al., "Massive access of static and mobile users via reconfigurable intelligent surfaces: Protocol design and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1253–1269, Apr. 2022.
[55] O. Chukhno, N. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Modeling reconfigurable intelligent surfaces-aided directional communications for multicast services," in *Proc. IEEE Glob. Commun. Conf.*,, 2022pp. 5850–5855.
[56] Y. Feng, Q. Hu, K. Qu, W. Yang, Y. Zheng, and K. Chen, "Reconfigurable intelligent surfaces: Design, implementation, and practical demonstration," *Electromagn. Sci.*, vol. 1, no. 2, 2023, Art. no. 0020111.
[57] N. Chukhno et al., "Are D2D and RIS in the same league? Cooperative RSSI-based localization model and performance comparison," *Ad Hoc Netw.*, vol. 175, 2025, Art. no. 103862.
[58] C. A. Balanis, *Antenna Theory: Analysis and Design*, 4th ed.Hoboken, NJ, USA: Wiley, 2016.
[59] "Technical specification group radio access network; study on channel model for frequencies from 0.5 to 100 GHz (release 18)," 3GPP, Sophia Antipolis Cedex, France, Tech. Rep., TR 38.901 V18.0.0, Mar. 2024.
[60] W. Mei, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless networks: From single-reflection to multireflection design and optimization," *Proc. IEEE*, vol. 110, no. 9, pp. 1380–1400, Sep. 2022.
[61] 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. IEEE Glob. Commun. Conf.*, 2017, pp. 1–7.
[62] M. Gapeyenko et al., "Analysis of human-body blockage in urban millimeter-wave cellular communications," in *Proc. IEEE Int. Conf. Commun.*, 2016, pp. 1–7.
[63] "NR; physical channels and modulation (release 15)," 3GPP, Sophia Antipolis Cedex, France, TR 38.211, Mar. 2025.
[64] "System performance evaluation in multi-hop IAB network," 3GPP, Sophia Antipolis Cedex, France, Tech. Rep. TSG RAN WG1 Meeting 95, Nov. 2018.
[65] N. Chukhno, S. Saafi, and S. Andreev, "ML-Aided dynamic BSR periodicity adjustment for enhanced UL scheduling in cellular systems," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 3513–3527, 2025.
[66] I. Yildirim, A. Uyrus, and E. Basar, "Modeling and analysis of reconfigurable intelligent surfaces for indoor and outdoor applications in future wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1290–1301, Feb. 2020.

Olga Chukhno received the double PhD degree from Tampere University, Finland and from the Mediterranea University of Reggio Calabria, Italy. She is currently an assistant professor with the Mediterranea University of Reggio Calabria and CNIT, Italy. Her research interests mainly include wireless communications, programmable heterogeneous networks, advanced algorithms for managing distributed services, with a particular focus on XR applications. She was the recipient of the MSCA Innovative Training Network Fellowship.

Dmitri Moltchanov received the MSc and Cand.Sc. degrees from the St. Petersburg State University of Telecommunications, Russia, in 2000 and 2003, respectively, and the PhD degree from the Tampere University of Technology, Tampere, Finland, in 2006. He is currently a University lecturer with the Laboratory of Electronics and Communications Engineering, Tampere University, Finland. He has taught more than 70 full courses on wireless and wired networking technologies, P2P/IoT systems, network modeling, and queuing theory. He has coauthored more than 200 publications on wireless communications, heterogeneous networking, IoT applications, applied queuing theory. His research interests include research and development of terahertz 6G systems, 5G NR, non-terrestrial NB-IoT systems, DECT-2020 and V2V/V2X systems.

Gianluca Brancati received the BSc degree in information engineering and the MSc degree in telecommunication engineering from the Mediterranea University of Reggio Calabria, Italy, in 2019 and 2022, respectively. He is currently working toward the PhD degree with the Mediterranea University of Reggio Calabria, Italy. He was the visiting PhD degree student with Tampere University, Finland, and Queen's University Belfast, Northern Ireland. His research interests include wireless communications, multicast communications, and reconfigurable intelligent surfaces.

Sara Pizzi received the 1st Level Laurea (*cum laude*) and 2nd Level Laurea (*cum laude*) degrees in telecommunication engineering, and the PhD degree in computer, biomedical, and telecommunication engineering from the University Mediterranea of Reggio Calabria, Italy, in 2002, 2005, and 2009, respectively. She is currently with the University Mediterranea of Reggio Calabria and CNIT, Italy. Her research interests include wireless and mobile communication systems, non-terrestrial networks, and radio resource management.

Antonella Molinaro received the Graduation degree in computer engineering from the University of Calabria, Arcavacata, Italy, in 1991, the master's degree in information technology from CEFRIEL/Polytechnic of Milano, Milan, Italy, in 1992, and the PhD degree in multimedia technologies and communications systems in 1996. She is currently a full professor with the University Mediterranea of Reggio Calabria, Italy, and CNIT, Italy, and also with Université Paris-Saclay, France. Her research interests mainly include wireless and mobile networking, vehicular networks, and future Internet.

Giuseppe Araniti received the Laurea degree and the PhD degree in electronic engineering from the University Mediterranea of Reggio Calabria, Italy, in 2000 and 2004, respectively. He is currently a full professor of telecommunications with the University Mediterranea of Reggio Calabria. His research focuses on 5G/6G networks and including personal communications, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, device-to-device (D2D), and machine-type communications (M2M/MTC).