EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Multi-timescale scheme for cooperative user association and hybrid beamforming in mmWave MIMO systems

Mohammadreza Heydarian[1,2*], Didier Colle[1,2], Mario Pickavet[1,2] and Wouter Tavernier[1,2]

*Correspondence: Mohammadreza.heydarian@ugent.be

[1] Department of Information Technology, Ghent University, Ghent, Belgium
[2] IMEC, Leuven, Belgium

Abstract

Hybrid beamforming has received significant attention as a solution to the thermal issues, costs, and implementation complexities associated with fully digital mmWave extremely large MIMO (XL-MIMO) systems. The hybrid approach offers a balance between system flexibility and performance. However, in scenarios involving a large number of user equipments (UEs), even optimal beamforming techniques cannot provide satisfactory network delay and throughput if the data rate is not effectively shared among them.. Considering multiple forms of dynamism of an indoor wireless channel, we propose a multi-layer scheme to optimize resource allocation to the UEs in a cooperative multi-base station (BS) setup. We evaluated various resource allocation techniques to effectively share the bandwidth among UEs, rather than simply maximizing total downlink throughput by concentrating on well-off ones. Our findings indicate that queue length-based allocation strategies yield the lowest delay under different downlink traffics and UE/BS deployment densities. Moreover, selectively serving a subset of UEs during each channel block, rather than serving them all simultaneously, enhances network delay and throughput. To further improve resource utilization and overall performance, we introduce an extension called sub-coherence time allocation. This technique considers early downlink queue exhaustion and speculatively calculates multiple digital precoders to be activated sequentially within a channel block. Simulation results demonstrate that this approach improves delay and jitter with minimal computational overhead, achieving a 25% and 15% decrease in digital and hybrid modes, respectively.

Keywords: Hybrid beamforming, Cooperative networks, XL-MIMO, MmWave, User association, Throughput-optimal scheduling, 6 G

1 Introduction

Since its introduction two decades ago, multiple-input multiple-output (MIMO) technology has proven highly effective in channel hardening and capacity enhancement [1]. Over the years, base stations equipped with arrays of tens of antennas have been widely deployed in 4 G and 5 G networks [2]. The recent surge in wireless communication demand has necessitated the utilization of the vast bandwidth available in the millimeter-wave (mmWave) spectrum, presenting new challenges. Higher frequencies lead

to increased path loss and signal susceptibility to blockages, significantly reducing the transmission range [3]. This limitation can be mitigated by employing higher gain antennas [4]. The shorter wavelength of mmWave frequencies allows for the miniaturization of antenna units, facilitating the dense packing of antennas in configurations leading to XL-MIMO [5] (Fig. 1).

Fully digital setups, wherein each antenna is paired with a dedicated radio frequency (RF) chain, are a common approach for beamforming in MIMO systems [6]. While effective for arrays with tens of antennas, this approach is impractical for XL-MIMO systems that incorporate hundreds of antennas. The large number of RF chains required in such setups results in high energy consumption, significant heat production, substantial costs, and considerable implementation complexity [7]. To address these challenges, hybrid beamforming has been introduced. This method strikes a balance between the cost and energy efficiency of analog beamforming and the flexibility and performance of fully digital beamforming [5]. Although hybrid beamforming holds great potential for future 6 G systems, it still lacks the maturity required for widespread practical deployment.

Managing an XL-MIMO system, which consists of hundreds of UEs and multiple BSs, presents significant challenges. While beamforming techniques have received considerable attention in the literature, they alone are insufficient. Effective management requires the coordination of BSs through joint optimization techniques [11]. Due to the extensive computational complexity involved in the management and control tasks, a multi-layer algorithm is necessary to handle each sub-problem at its optimal frequency. Moreover, a more advanced strategy beyond simply maximizing the total bitrate [7] should be considered. Given these requirements, we conducted a thorough literature review and, to the best of our knowledge, found no existing study that comprehensively addresses all these design challenges. This gap serves as the primary motivation for our research.

Several studies have addressed specific aspects of XL-MIMO with a narrow focus, primarily on simple total bitrate optimization at the level of a single BS. For instance,

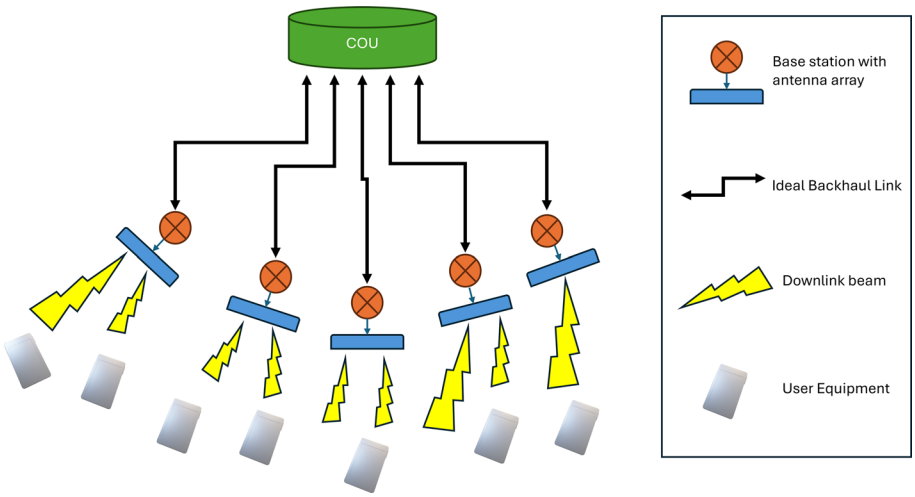

Fig. 1 The architecture of the multi-BS hybrid beamforming setup, where BSs equipped with antenna arrays are connected to a central orchestration unit via ideal backhaul links and jointly transmit downlink beams to serve the user equipment in the coverage area

[6, 12] explore antenna subset selection to maximize sum rate while minimizing energy consumption. Similarly, [13] proposes a fully analog approach for the uplink of a mmWave MIMO system, leveraging two distinct codebooks for near-field and far-field scenarios.

The concept of a multi-timescale scheme for control and management has gained considerable attention in the MIMO literature. The necessity of such a scheme for resource allocation is highlighted in [14], where a vehicular networking setup with cell-free MIMO leverages the different temporal characteristics of vehicle movement and the wireless network to develop a two-layer algorithm. Similarly, [15] proposes a two-timescale scheme for resource allocation in a MIMO system with edge computing, where the first layer manages the activation and deactivation of computation units, while the second layer handles task offloading decisions. However, these studies do not specifically address network optimization or explore the integration of a hybrid approach for practical applications.

In contrast, [7] introduces a two-timescale scheme based on the observation that, in mmWave environments, path angles exhibit lower dynamics than path gains. In the higher layer, analog beams are determined for each UE, while in the lower layer, UE selection is performed. However, this work lacks joint optimization. Although it employs a more advanced allocation strategy based on proportional fairness (PF), it does not explore alternative strategies or provide a comparative analysis of their effectiveness.

Additionally, [11] investigates a cooperative multi-BS setup in which an algorithm determines UE-BS assignments, followed by an iterative hybrid beamforming technique using fractional programming to maximize the objective function. This study considers multiple antenna array architectures, including fully connected, fixed subarray, and dynamic subarray configurations. However, its primary limitation is the reliance on total achievable bitrate as the performance metric, which does not fully capture user experience. Furthermore, hybrid beamforming is performed in its entirety during each coherence interval, despite the relatively static nature of analog beamforming, leading to unnecessary computational overhead.

Resource allocation strategies can be divided into two categories: fair scheduling and throughput-optimal scheduling [8]. The former has been explored in several works through the PF metric [7, 9]. The latter, in the form of queue length-based allocation [10], has also received considerable attention but remains underexplored in the context of hybrid MIMO.

Addressing the aforementioned gaps in the literature regarding the management challenges of XL-MIMO, this paper introduces a comprehensive multi-timescale scheme for optimizing network experience in a multi-BS hybrid beamforming setup. Inspired by [11] and [7], we decompose the overall optimization problem into multiple sub-problems with different execution intervals. Various allocation strategies, represented by different objective functions, are considered and their performance is assessed in the digital/hybrid beamforming setups.

1.1 Contribution

The major contributions of our work can be summarized as follows:

- **Novel Multi-Layer Approach to Resource allocation:** Instead of solving a single optimization problem, we propose a multi-layered approach tailored for an indoor multi-BS wireless environment. This approach involves solving sub-problems at different intervals to address various forms of channel dynamism:

  - *UE-BS Association:* With the least execution frequency, this sub-problem addresses UE movement and path-loss fluctuations.
  - *Analog Beamforming:* Executed at a higher frequency to compensate for slow fading due to minor environmental changes.
  - *UE Selection and Digital Beamforming:* With the highest execution frequency, this sub-problem manages fast fading caused by constructive and destructive interference of multipath components.

  This approach balances execution time and performance by solving each sub-problem when needed.

- **More Optimal Resource Allocation Techniques:** Instead of solely maximizing the total bitrate, we employ throughput-optimal and fair resource allocation strategies using downlink queue length and proportional fairness metrics. These strategies improve network delay and better satisfy the bandwidth requirements of all UEs.
- **Sub-Coherence Time Allocation:** Rather than using a single digital precoder for the entire channel block duration, we propose a speculative multi-precoder approach. This method considers queue length and expected achievable rates to calculate the UEs' total transmission time. If a queue is exhausted early, resources are reallocated to other UEs, ensuring higher resource utilization and lower delay.
- **A Simulation Tool for Multi-BS XL-MIMO:** To evaluate our research, we developed a modular simulator focusing on the physical layer in Python. It covers aspects such as UE mobility, multi-path mmWave channel, beam sweeping, UE selection, and zero-forcing digital beamforming. This tool can function as a stand-alone simulator or a plug-in for other network simulators like OMNeT++. It can be requested from the authors by email and is considered to be open-sourced in the future. We believe this simulator will assist other researchers by simplifying the implementation and testing of their hypotheses.

2 Methods

In this section, we delve into the technical aspects of the research. We begin by defining the notation used throughout the paper, followed by a description of the system setup, architecture, and channel model. Next, we present different resource allocation strategies and introduce the hierarchical optimization approach, which consists of three layers. Finally, we describe the simulation environment and parameters, providing the necessary theoretical foundation for the subsequent discussions (Table 1).

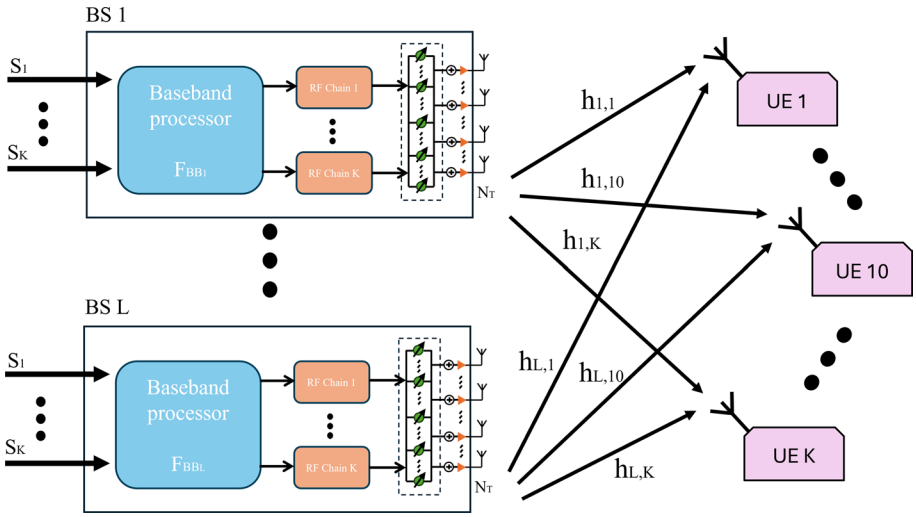

Fig. 2 Multi-BS mmWave hybrid beamforming setup with a fully connected antenna architecture, where each base station processes multiple data streams in the digital domain, then employs analog beamforming at the antenna arrays via multiple RF chains. The resulting beamformed signals from each BS provide high-capacity mmWave links to multiple users over their respective channels

Table 1 Symbols definition

Notations	Meaning
L	The number of BSs
K	The number of single-antenna UEs
P_{max}	BS power constraint
N_{RF}	The number of RF chains in each BS
N_T	The number of transmitting antennas in each BS
N_r	The number of paths in mmWave channel model
t	time
$\mathbf{F}_{RF}$	Analog precoding matrix
$\mathbf{F}_{BB}$	Digital precoding matrix
$\mathcal{W}$	Analog beamforming codebook
$\mathbf{s}$	downlink signal vector
$\mathbf{H}$	channel matrix
g	channel propagation path complex gain
ϕ	channel propagation path horizontal angle (azimuth)
θ	channel propagation path vertical angle (elevation)
λ	carrier frequency wavelength
$\mathbf{b}$	antennas array steering vector
d	distance between antenna array rows and columns
$\mathbf{x}$	BSs transmitted signal vector
$\mathbf{y}$	UEs received signal vector
n	complex additive white Gaussian noise
σ^2	noise variance
$\mathbf{A}$	UE-BS assignment matrix
w	channel bandwidth
r	UE achievable rate
q	downlink queue length

2.1 System model

2.1.1 Notation

Normal letter a denotes a scalar variable. Boldface capital $\mathbf{A}$ and small $\mathbf{a}$ letters represent matrices and vectors, respectively. Capital calligraphic letters $\mathcal{K}$ indicates finite sets, and $|\mathcal{K}|$ denotes the cardinality of the set $\mathcal{K}$. $\mathbf{I}_m$ denotes the identity matrix of size m. $\{\cdot\}^*$, $\{\cdot\}^T$ and $\{\cdot\}^H$ denote the conjugate, transpose and the conjugate transpose operators, respectively. $\|\cdot\|_F$ denote the Frobenius norm operator,and $\otimes$ symbolizes the Kronecker product. $\mathcal{CN}(\mu, \sigma^2)$ is a circularly symmetric complex Gaussian distribution with mean μ and variance σ^2.

2.1.2 Setup and architecture

We consider the downlink transmission of a multi-user MIMO (MU-MIMO) system consisting of L BSs serving K UEs. We assume UEs to have a single antenna incapable of beamforming. As shown in Fig. 2 each BS is equipped with N_{T} antennas and N_{RF} RF chains and $N_{\mathrm{T}} \gg N_{\mathrm{RF}}$. Each BS is capable of hybrid beamforming, utilizing a Baseband (BB) processor to perform digital beamforming. Each RF chain feed into several antenna elements which is used for the implementation of the analog beamforming. As discussed in [16], there exist multiple architectures for hybrid beamforming, which refer to how RF chains are connected to antenna elements. These include fully connected setup, where each RF chain can be connected to all the antennas; fixed subarray setup, where each RF chain connects to an unchanging subset of antennas; and dynamic subarray setup. Studies have shown that the fully connected hybrid architecture has a performance similar to the fully digital architecture [17]. For this reason, we have assumed a fully connected architecture for the BSs in our work..Each user has a dedicated data stream which is processed by a single RF chain. Based on [1] the multiplexing gain of a MIMO channel equals $min(K, N_{\mathrm{RF}})$ therefore N_{RF} is the maximum number of users which a BS can simultaneously serve. All the BSs are connected to a central orchestration unit (COU) serving as an orchestrator for the entire system and enabling the BSs to work cooperatively and form a system.

Each UE is assigned to and served by maximum one BS at a time which we call UE-BS association. The association between BSs and UEs is identified by UE-BS assignment matrix $\mathbf{A} \in \{0,1\}^{L\times K}$. For $l = 1, ..., L$ and $k = 1, ..., K$, if UE-k is associated with BS-l, $\mathbf{A}_{\mathrm{l,k}} = 1$; otherwise $\mathbf{A}_{\mathrm{l,k}} = 0$.

At each time t the signal for the kth user is denoted by $\mathbf{s}_k(t)$ for $k = 1, ..., K$. At each $t = T$ we assume the signal to be $\mathbb{E}(s_k(T)s_k(T)^*) = 1$ and $\mathbb{E}(s_i(T)s_j(T)^*) = 0, \forall i \neq j$. Each BS is capable of hybrid beamforming; therefore the input signal is precoded by a digital beamforming vector $\mathbf{f}_{BB_{l,k}} \in \mathbb{C}^{N_{RF}}$, and then it goes through another conversion via the analog beamforming vector $\mathbf{F}_{RF_l} \triangleq [\mathbf{f}_{RF_{l,1}}, ..., \mathbf{f}_{RF_{l,N_{RF}}}] \in \mathbb{C}^{N_T \times N_{RF}}$. Having the UE-BS assignment matrix, digital and analog beamforming matrix, the transmitted signal of BS l is calculated as

$$\mathbf{x}_l = \sum_{k=1}^{K} \mathbf{A}_{l,k}\mathbf{F}_{RF_l}\mathbf{f}_{BB_{l,k}}s_k. \tag{1}$$

Considering the power constraints of each BS l it should be ensured that

$$\sum_{k=1}^{K} \|\mathbf{A}_{l,k}\mathbf{F}_{RF_l}\mathbf{f}_{BB_{l,k}}\|_F^2 \leq P_{max,l}, \tag{2}$$

where $P_{max,l}$ is the maximum transmit power of BS-l, $l = 1, ..., L$. It is assumed that all the BSs have the same power capacity.

2.1.3 Channel model and SINR

For the channel model as in [3], we've utilized a mmWave propagation channel having N_r channel paths. Considering N_T as the number of antennas at the BS and one antenna at the UE, the channel vector $\mathbf{h}_{l,k}$ between BS-l and UE-k can be modeled as

$$\mathbf{h}_{l,k} = \sqrt{\frac{N_T}{N_r}} \sum_{n=1}^{N_r} g_n \mathbf{b}(\phi_n, \theta_n), \tag{3}$$

where g_n, ϕ_n, and θ_n are the propagation path-n's complex gain, horizontal angle and vertical angle, and $\mathbf{b} \in \mathbb{C}^{N_T}$ is the antenna array steering vector which is dependent upon the arrangement of the antenna units in the array. Although uniform linear array (ULA) has been extensively used in the literature for the simplicity, we opted for a uniform rectangular planar array (URPA) for its 2D beamforming capabilities in contrast to the limited 1D beamforming of ULA. The URPA consists of N_{T_x} rows each having N_{T_y} antennas. Each row and column is in fact a ULA with inter-element spacing distance of d_x and d_y, respectively. For simplicity and without loss of generality, we assume $d_x = d_y = \frac{\lambda}{2}$. The steering vector of an URPA in the far-field regime can be expressed as [32]

$$\mathbf{b}(\phi, \theta) = \frac{1}{\sqrt{N_T}}(\mathbf{b}_x(\phi, \theta) \otimes \mathbf{b}_y(\phi, \theta)), \tag{4}$$

where $\mathbf{b}_x(.)$ and $\mathbf{b}_y(.)$ denotes the horizontal and vertical steering vectors, respectively, such that

$$\mathbf{b}_x(\phi, \theta) = \left[1, e^{j\kappa d_x \sin\theta\cos\phi}, \ldots, e^{j(N_{T_x}-1)\kappa d_x \sin\theta\cos\phi}\right]^T, \tag{5}$$

and

$$\mathbf{b}_y(\phi, \theta) = \left[1, e^{j\kappa d_y \sin\theta\sin\phi}, \ldots, e^{j(N_{T_y}-1)\kappa d_y \sin\theta\sin\phi}\right]^T, \tag{6}$$

where $\kappa = 2\pi/\lambda$ is the wave number and λ is the carrier frequency.

Subsequently, the received signal $y_k \in \mathbb{C}$ at time t by UE-k is given by

$$y_k(t) = \sum_{l=1}^{L} \mathbf{h}_{l,k}^H(t)\mathbf{x}_l(t) + n_k(t), \tag{7}$$

where $\mathbf{h}_{l,k}$ is the channel matrix between BS-l and UE-k and $n_k(t) \in \mathbb{C}$ is the complex additive white Gaussian noise following the complex Gaussian distribution with zero mean and a variance of σ_k^2, i.e., $n_k(t) \sim \mathcal{CN}(0, \sigma_k^2)$. By substituting an instantaneous version of equation (1) in place of $\mathbf{x}_l$ we can obtain

$$y_k(t) = \sum_{l=1}^{L} \mathbf{h}_{l,k}^H(t) \sum_{k=1}^{K} \mathbf{A}_{l,k}(t)\mathbf{F}_{\mathrm{RF_l}}(t)\mathbf{f}_{\mathrm{BB_{l,k}}}(t)s_k(t) + n_k(t). \tag{8}$$

We can separate the desired signal for the UE-k from the inter-BS and intra-BS interference by

$$\begin{aligned} y_k(t) = &\sum_{l=1}^{L} \mathbf{A}_{l,k}(t)\mathbf{h}_{l,k}^H(t)\mathbf{F}_{\mathrm{RF_l}}(t)\mathbf{f}_{\mathrm{BB,l,k}}(t)s_k(t) \\ &+ \sum_{l=1}^{L} \sum_{j=1, j\neq k}^{K} \mathbf{A}_{l,j}(t)\mathbf{h}_{l,k}^H(t)\mathbf{F}_{\mathrm{RF_l}}(t)\mathbf{f}_{\mathrm{BB_{l,j}}}(t)s_j(t) + n_k(t), \end{aligned} \tag{9}$$

where the first term is the signal intended for UE-k, the second term is the signal intended for all the other UEs producing the interference, and the third term is the noise. Having separated the desired signal from the interference and noise, one can calculate the signal to noise-plus-interference (SINR) for UE-k at time t by

$$\mathrm{SINR}_k(t) = \frac{\left|\sum_{l=1}^{L} \mathbf{A}_{l,k}(t)\mathbf{h}_{l,k}^H(t)\mathbf{F}_{\mathrm{RF}_l}(t)\mathbf{f}_{\mathrm{BB}_{l,k}}(t)\right|^2}{\sum_{\substack{j=1 \\ j\neq k}}^{K} \left|\sum_{l=1}^{L} \mathbf{A}_{l,j}(t)\mathbf{h}_{l,k}^H(t)\mathbf{F}_{\mathrm{RF}_l}(t)\mathbf{f}_{\mathrm{BB}_{l,j}}(t)\right|^2 + \sigma_k^2}. \tag{10}$$

The SINR can be used in the calculation of the bitrate based on Shannon–Hartley theorem stating that the maximum achievable bitrate r for UE-k at time t over a communication channel is [18]

$$r_k(t) = w \log_2(1 + \mathrm{SINR}_k(t)), \tag{11}$$

where w is the bandwidth of the channel in hertz. The sum-rate of the UEs which is a linear combination of the UEs downlink bitrate can be calculated depending on the objective function, which will be discussed in the next section. It should be noted that we consider a block-fading channel model in which the channels remain constant within each block. Each block has a duration equal to the channel coherence time, which is the duration over which a wireless communication channel's characteristics remain relatively stable, and the channel block index is denoted by t, where $t \in \{1, 2, ..., T\}$. Similar to prior research on hybrid beamforming [11, 21] we assume a perfect knowledge of the channel state information (CSI) in the form of channel matrix for all the UEs. However, in practical scenarios, obtaining perfect CSI is challenging. Despite this, various methods have been developed to estimate CSI effectively. For instance, existing channel estimation algorithms, such as the adaptive compressed sensing technique [22], can be employed at the transmitters to estimate CSI with high accuracy.

2.2 Resource allocation and objective functions

Resource allocation in wireless communication has always been a major concern due to the shared medium, and this issue becomes more pronounced with the increasing number of UEs, particularly in MU-MIMO systems. Traditionally, the primary objective in the MIMO literature has been to maximize the sum of the spectral efficiency (SE) or the bitrate of the UEs [19, 20]. While this approach effectively increases overall network throughput, it does not ensure proper resource utilization, guarantee fair distribution of resources among UEs or account for their varying requirements, potentially leading to suboptimal user experiences. To address these limitations, two main categories of resource allocation exist: throughput-optimal and fair scheduling [8]. **Throughput-optimal scheduling** focuses on maximizing the overall network throughput, efficiently utilizing available bandwidth to achieve higher data rates. However, it often overlooks fairness, leading to some users receiving poor service. **Fair scheduling**, in contrast, aims to equitably distribute resources among users, prioritizing user satisfaction and PF metric. While this method ensures fair resource allocation, it might not always achieve the highest possible throughput due to its emphasis on equity.

To improve network performance, resource utilization, and user satisfaction, we propose a more expressive techniques. We have identified three strategies for resource allocation, each with its own corresponding objective function. Although the primary goal is to maximize the sum-rate, the definition of sum-rate varies based on the strategy used, which is reflected in the different objective functions. We will discuss these strategies and their respective objective functions in detail.

2.2.1 Queue length minimization (QLM) strategy

Under QLM, the focus is on maximizing resource utilization in various network conditions. The downlink queue length is considered as an indicator of user requirements. The aim is to allocate resources based on actual user needs, thereby reducing delays and enhancing the overall network experience [10]. The sum-rate at time t is calculated as

$$R(t) = \sum_{k=1}^{K} r_k(t) q_k(t), \tag{12}$$

where R, r and q are, respectively, the sum-rate, the maximum achievable bitrate and downlink queue length for UE-k at time t.

2.2.2 Long-term PF maximization (PFM) strategy

Long-term PFM is an approach that aims to balance the cumulative data rates of all the users in the network. The cumulative data rate for a user k at time T is defined by an exponential moving average of the data rates it observed up to that moment $\{r_k(t)\}_{t=1}^{T}$. At time t the cumulative data rate of UE-k is given by [7]

$$R_k(t) = (1-\delta) R_k(t-1) + \delta r_k(t), \tag{13}$$

where $\delta \in [0, 1]$ is a constant weight. This metric reflects the long-term service quality experienced by the users. The sum-rate under PFM at time t is calculated by [9]

$$R(t) = \sum_{k=1}^{K} \frac{r_k(t)}{R_k(t-1)}. \tag{14}$$

The goal is to allocate resources in a way that maximizes the proportional fairness of the users ensuring that each user receives a fair share of the resources.. The fairness aspect is particularly important in scenarios where users may have varying channel conditions and data requirements. By focusing on proportional fairness, the network can achieve a balanced distribution of resources, preventing conditions where some users monopolize the available bandwidth at the expense of others.

2.2.3 Queue aware PF maximization (QPFM) strategy

QPFM aims to strike a balance between QLM and PFM by incorporating elements from both strategies. This approach ensures fairness while also paying attention to the queue lengths. The objective function considers both the current queue lengths and the PF metric in an attempt to allocate resources in a manner that not only maximizes fairness but also minimizes delays caused by long queue lengths. The sum-rate for QPFM can be formulated as

$$R(t) = \sum_{k=1}^{K} \frac{r_k(t)q_k(t)}{R_k(t-1)}. \tag{15}$$

Using the sum-rate calculated by the objective function we can formulate the optimization problem in the following section.

2.3 Multi timescale schemes

The aim of the optimization is to maximize the instantaneous sum-rate under a certain objective function at each time $t \in \{1, ..., T\}$. For this purpose the instantaneous optimal BS-UE assignment matrix $\mathbf{A}$ and hybrid beamformer $\left\{\mathbf{F}_{\mathrm{RF}_l}, \mathbf{f}_{\mathrm{BB}_{l,k}}\right\} \forall l, \forall k$ must be found subject to the transmit power constraints at BSs and the unit modulus constraints of phase shifters at the antennas. Having the BS power constraints and the objective function, the instantaneous sum-rate maximization problem can be formulated as

$$\begin{aligned}
\max_{\mathbf{A}(t),\{\mathbf{F}_{\mathrm{RF}_l}(t),\mathbf{f}_{\mathrm{BB}_{l,k}}(t)\}\forall l,\forall k} \quad & R(t) \\
\text{s.t.} \quad & \sum_{l=1}^{L} \mathbf{A}_{l,k}(t) \leq 1, \quad \forall k, \\
& \sum_{k=1}^{K} \mathbf{A}_{l,k}(t) \leq N_{\mathrm{RF}}, \quad \forall l, \\
& |\mathbf{F}_{\mathrm{RF}_l}(t)_{i,j}| = 1, \quad \forall l, i, j, \\
& \sum_{k=1}^{K} \|\mathbf{A}_{l,k}(t)\mathbf{F}_{RF_l}(t)\mathbf{f}_{BB_{l,k}}(t)\|_F^2 \leq P_{\max}, \quad \forall l,
\end{aligned} \tag{16}$$

where $P_{\max}$ is the BSs maximum transmit power. The first constraint indicates that each UE can be assigned to maximum one BS at a time. The second denotes the maximum number of UEs that can be simultaneously served by a BS, which equals the number of

RF chains[16]. The third represents the fixed amplitude of phase shifters, and the last constraint specifies that the hybrid beamforming matrices should be designed such that the maximum transmit power is not exceeded. Optimizing the sum-rate maximization problem in equation (16) involves several challenges. First, the objective function $R(t)$ is complex and includes fractional and logarithmic terms. Also, the optimization problem is a mixed integer problem with $\mathbf{A}(t)$ existing in the discrete space, while $\mathbf{F}_{\mathrm{RF}}(t)$ and $\mathbf{f}_{\mathrm{BB}}(t)$ reside in the continuous space. Finally, the non-convex unit modulus constraint for the analog beamformer $\mathbf{F}_{\mathrm{RF}}(t)$ is difficult to handle. Using an exhaustive search method to design hybrid beamformers by examining all potential integer variables $\{\mathbf{A}_{l,k}\}_{\forall l,\forall k}$ leads to exponential growth in computational complexity as the number of users and base stations increases. To address this, we implement a sub-optimal solution dividing the problem into manageable sub-problems [7]. Accordingly, we suggest a multi-timescale scheme to tackle the resource optimization in a multi-BS MIMO environment. We consider three sub-problems with different intervals, namely UE-BS association, analog beamforming via beam sweeping and digital beamforming as can be seen in Fig. 3.

Initially, we employ the channel strength to determine the UE-BS assignment matrix $\mathbf{A}(t)$. Because of the hardware limitations existing in the mmWave systems [4, 5] and to make the problem more manageable, we employ a codebook for the analog beamforming. We use beam-sweeping over the codebook to find the optimal codeword for each UE [25] which determines the $\mathbf{F}_{\mathrm{RF}}(t)$. In the digital beamforming sub-problem, we employ the zero-forcing technique, which is widely adopted in both literature and industry due to its effective balance between performance and computational efficiency [23]. In an iterative process we try to select an optimal subset of UEs to be served during the next coherence time which is manifested in the digital beamformer $\mathbf{f}_{\mathrm{BB}}(t)$.

Solving all of these sub-problems in each time block would be time-consuming and impractical, so we consider the various dynamic aspects of the wireless channel to determine the frequency of each task [24]. To tackle the short-term fading, we perform digital

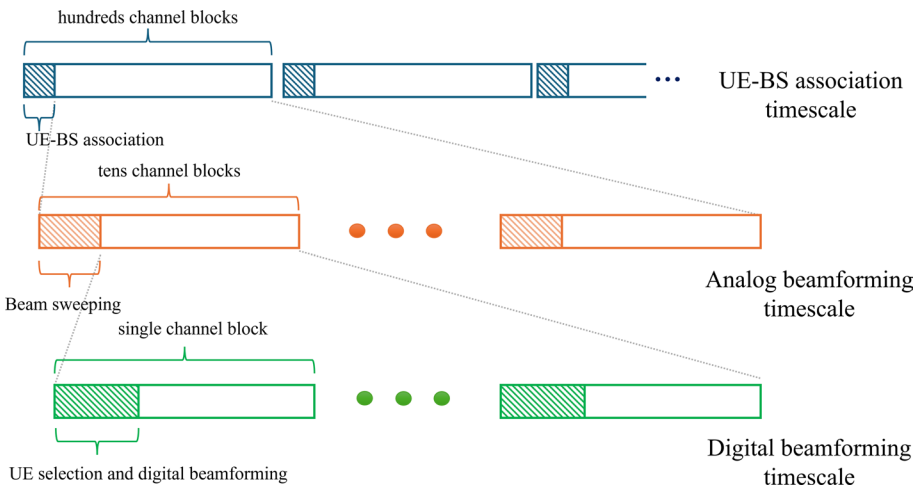

Fig. 3 Illustration of the hierarchical structure in the proposed multi-timescale beamforming scheme. Each UE-BS) association timeslot (top layer) encompasses multiple analog beamforming timeslots (middle layer), each of which is further subdivided into several digital beamforming timeslots (bottom layer). This structure enables flexible and efficient beam adaptation across different temporal granularities

beamforming in each channel block. Given the less frequently changing nature of the long-term fading, we perform the analog beamforming once within multiple blocks to account for this phenomenon. Finally, the UE-BS association addresses the fading caused by the movement of the UEs, so it is performed once after multiple analog beamforming steps. Leveraging the varying characteristics of the wireless channel, this hierarchical approach optimizes computational resources, maintains system performance, and enhances the system's ability to adapt to real-time changes, leading to a more robust and efficient network operation.

2.3.1 UE-BS assignment

Because of the directional characteristics of mmWave channels, the SINR for each UE is greatly affected by the channel gain. Therefore, similar to [11] the UE-BS assignment matrix $\mathbf{A}(t)$ is determined by optimizing the sum-channel-gain, which can be expressed as

$$\max_{\{\mathbf{A}_{l,k}(t)\}\forall l,\forall k} \quad \sum_{l=1}^{L}\sum_{k=1}^{K}\left|\mathbf{A}_{l,k}(t)\mathbf{h}_{l,k}(t)\mathbf{h}_{l,k}^{H}(t)\right|^2 \qquad \text{s.t.} \quad \sum_{l=1}^{L}\mathbf{A}_{l,k}(t)=1, \quad \forall k \qquad \sum_{k=1}^{K}\mathbf{A}_{l,k}(t)\le N_{RF}, \quad \forall l. \tag{17}$$

The first and second constraints ensure that each UE is assigned to one and only one BS and, RF chain count limit is enforced. This problem can be solved by a brute-force approach having a computational complexity of $\mathcal{O}(K^L)$ which is impractical. Therefore, similar to [26] Gale–Shapley (stable-marriage) algorithm, as shown in Algorithm 1 has been utilized with a lower complexity of $\mathcal{O}(K^2)$ to find a sub-optimal but timely feasible matching.

Initially, each BS creates a preference list ranking all the UEs based on channel gains and each UE does the same for BSs. Then, each UE proposes to the BS it prefers the most. Each BS then tentatively accepts proposals based on its ranking and available RF chains, rejecting lower-ranked UEs if necessary. The process iterates with UEs that were rejected proposing to their next preferred BS. This step repeats until all UEs are either matched to a BS or have no remaining preferred BSs to propose to. The algorithm guarantees that the final matching is stable, meaning no UE-BS pair would both prefer to be matched with each other over their current assignments.

Algorithm 1 Gale–Shapley algorithm (stable-marriage)

Input: L: Set of BSs $\{BS_1, BS_2, ..., BS_L\}$, K: Set of UEs $\{UE_1, UE_2, ..., UE_K\}$,
PreferenceList_BS: List of UEs for each BS, PreferenceList_UE: List of BSs for each UE, MaxRFChains: Max RF chains per BS
Output: Assignment matrix $A(L \times K)$
Initialize $A(L \times K)$ to 0, all UEs as unassigned, BSs with MaxRFChains;
while *unassigned UE with a preferred BS exists* **do**
  Select such a UE, say UE_k ;
  Let BS_l be the highest ranked BS in UE_k's list not yet proposed to ;
  UE_k proposes to BS_l ;
  if BS_l *has available RF chains* **then**
    Accept UE_k ;
    Update $A[BS_l, UE_k] \leftarrow 1$, decrease RF chains of BS_l by 1 ;
    Mark UE_k as assigned ;
  else
    Let UE_m be the least preferred UE currently assigned to BS_l ;
    if BS_l *prefers* UE_k *over* UE_m **then**
      Reject UE_m, update $A[BS_l, UE_m] \leftarrow 0$, mark UE_m as unassigned ;
      Accept UE_k, update $A[BS_l, UE_k] \leftarrow 1$, mark UE_k as assigned ;
    else
      Reject UE_k ;
return Assignment matrix A

2.3.2 Analog beamforming by beam sweeping

A beam sweeping technique over a far-field angular domain codebook [13] is utilized to find the optimal codeword, i.e., analog beamforming vector for each UE. The codebook, denoted by $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_i\}$, consists of the analog beamforming vectors, where $\mathbf{w}_i$ represents the i-th vector in the set $\mathcal{W}$. After associating the UEs to BSs, the corresponding analog beamforming vectors are selected from the codebook $\mathcal{W}$, i.e., $\mathbf{f}_{\mathrm{RF}_{l,k}}(t) \in \mathcal{W}, \quad \forall l, \forall k$. This process determines the analog precoder at each BS-l $\mathbf{F}_{\mathrm{RF}_l}(t)$. The corresponding optimization problem can be formulated as

$$\max_{\{\mathbf{w}_k\}, \forall k} \quad \sum_{k=1}^{K} \sum_{l=1}^{L} \left| \mathbf{A}_{l,k}(t) \mathbf{h}_{l,k}^{H}(t) \mathbf{w}_k \right|^2 \qquad \text{s.t.} \quad \mathbf{w}_k \in \mathcal{W}, \quad \forall k \tag{18}$$

This problem involves combinatorial optimization and can be solved using an exhaustive search method with a computational complexity of $\mathcal{O}(K \times |\mathcal{W}|)$, as each UE is associated with only one BS.

2.3.3 User selection

The purpose of this sub-problem is to select a subset of UEs assigned to each BS maximizing the objective function. The chosen UEs will be served in the upcoming channel block, while the others remain inactive. Digital beamforming, using the zero-forcing

(ZF) technique, will be employed to simultaneously serve the selected UEs, effectively minimizing intra-BS interference by applying an inverse filter.

Each BS, equipped with N_{RF} RF chains, can serve up to N_{RF} UEs simultaneously. Given L BSs in the network, the task is to select the best subset of N_{RF} UEs for each BS independently with an overall computational complexity of $\mathcal{O}(L \times N_{RF}!)$. This factorial growth in complexity renders any exhaustive search approach impractical. To achieve feasible computational performance while maintaining near-optimal solutions, we consider two methods inspired by [7]: the **greedy** and the **adaptive top-k** method.

The greedy method, as shown in Algorithm 2, incrementally builds the optimal subset by selecting the UE that provides the highest marginal gain at each step. Specifically, the algorithm starts with an empty set and iteratively adds the UE that maximizes the overall objective function. The process continues until no further improvement is possible or the maximum number of UEs is reached. This approach ensures computational efficiency by focusing on the most beneficial UEs at each iteration, providing a good approximation of the optimal solution.

Algorithm 2 Greedy UE Selection and Beamforming

Input: L: BSs $\{BS_1, ..., BS_L\}$, K: UEs $\{UE_1, ..., UE_K\}$, N_{RF}: RF chains per BS, A: UE-BS assignment matrix, ObjectiveFunction, H: Channel matrix
Output: SelectedUEs, F_{BB}
for *each* $BS_l \in L$ **do**
  Initialize CurrentSet as empty;
  while *size of CurrentSet* $< N_{RF}$ **do**
    BestUE $\leftarrow$ None, BestObjectiveValue $\leftarrow -\infty$;
    Improvement $\leftarrow$ False;
    for *each* $UE_k \in K$ **do**
      if $UE_k \notin$ *CurrentSet* ***and*** $A[BS_l, UE_k] = 1$ **then**
        Add UE_k to CurrentSet;
        Compute digital beamforming matrix F_{BB} using zero-forcing with CurrentSet;
        Evaluate ObjectiveFunction with CurrentSet;
        if *ObjectiveValue* $>$ *BestObjectiveValue* **then**
          BestUE $\leftarrow UE_k$;
          BestObjectiveValue $\leftarrow$ ObjectiveValue;
          Improvement $\leftarrow$ True;
        Remove UE_k from CurrentSet;
    if *not Improvement* **then**
      Break;
    Add BestUE to CurrentSet;
  SelectedUEs$[BS_l] \leftarrow$ CurrentSet;
  Compute F_{BB} for BS_l using SelectedUEs$[BS_l]$;
return SelectedUEs, F_{BB}

Despite the better performance of the greedy algorithm compared to an exhaustive search, it is still computationally expensive. Therefore, a less optimal but more performant method is also evaluated. The implementation of the adaptive top-k algorithm as can been viewed in Algorithm 3, proceeds as follows:

Initially, UEs are sorted based on their individual channel strengths, independent of other UEs' influences. Subsequently, for each k from 1 to N_{RF}, the algorithm selects the top-k UEs and applies (ZF) beamforming. Each subset's performance is evaluated, and the optimal subset among these top-k UEs is selected. This approach balances computational efficiency and performance, making it a practical alternative to the greedy algorithm in real-time applications.

Algorithm 3 Adaptive Top-k UE Selection and Beamforming

Input: L: BSs $\{BS_1, ..., BS_L\}$, K: UEs $\{UE_1, ..., UE_K\}$, N_{RF}: RF chains per BS, A: UE-BS assignment matrix, H: Channel matrix, ObjectiveFunction
Output: SelectedUEs, F_{BB}
for *each* $BS_l \in L$ **do**
  Initialize OptimalSubset as empty;
  BestObjectiveValue $\leftarrow -\infty$;
  Sort UEs in K assigned to BS_l based on their individual channel strengths using H;
  for $k \leftarrow 1$ **to** N_{RF} **do**
    CurrentSubset $\leftarrow$ top-k UEs assigned to BS_l;
    Compute digital beamforming matrix F_{BB} using zero-forcing with CurrentSubset;
    Evaluate ObjectiveFunction with CurrentSubset;
    if *ObjectiveValue* $>$ *BestObjectiveValue* **then**
      OptimalSubset $\leftarrow$ CurrentSubset;
      BestObjectiveValue $\leftarrow$ ObjectiveValue;
  SelectedUEs[BS_l] $\leftarrow$ OptimalSubset;
  Compute F_{BB} for BS_l using SelectedUEs[BS_l];
return SelectedUEs, F_{BB}

2.3.4 Digital beamforming

Suppose $\mathcal{A}_l$ denotes the set of active UEs for BS l at a given time. Let $\mathbf{H}_{l,\mathcal{A}_l} \in \mathbb{C}^{|\mathcal{A}_l| \times N_{\mathrm{RF}}}$ be the channel matrix corresponding to the active UEs, where $|\mathcal{A}_l|$ is the number of active UEs. Given the analog beamforming matrix of the active UEs $\mathbf{F}_{\mathrm{RF}_{l,\mathcal{A}_l}} \in \mathbb{C}^{N_T \times |\mathcal{A}_l|}$ for BS l, the effective channel matrix $\mathbf{H}_{\mathrm{eff}_l}$ for the active UEs can be represented as:

$$\mathbf{H}_{\mathrm{eff}_l} = \mathbf{H}_{l,\mathcal{A}_l} \mathbf{F}_{\mathrm{RF}_{\mathcal{A}_l}} \tag{19}$$

The ZF precoder $\mathbf{F}_{\mathrm{BB}_l} \in \mathbb{C}^{N_{\mathrm{RF}} \times |\mathcal{A}_l|}$ for BS l is designed to nullify the interference among the active users. This is achieved by taking the Moore–Penrose pseudo-inverse of the effective channel matrix $\mathbf{H}_{\mathrm{eff}_l}$

$$\mathbf{H}^{\dagger}_{\mathrm{eff}_l} = \mathbf{H}^{H}_{\mathrm{eff}_l} \left(\mathbf{H}_{\mathrm{eff}_l} \mathbf{H}^{H}_{\mathrm{eff}_l} \right)^{-1}. \tag{20}$$

We note that $\mathbf{H}^{\dagger}_{\mathrm{eff}_l} = [\mathbf{f}^{*}_{BB,l_1}, ..., \mathbf{f}^{*}_{BB,l_{|\mathcal{A}_l|}}]$. Similar to [28], to ensure the transmit power constraint is met and equally distribute power among the active UEs, the ZF precoder is normalized:

$$\mathbf{f}_{\mathrm{BB}_{lk}} = \sqrt{\frac{P_{\max}}{|\mathcal{A}_l|}} \cdot \frac{\mathbf{f}^*_{BB,l_k}}{\left\| \mathbf{F}_{\mathrm{RF}_{\mathcal{A}_l}} \mathbf{f}^*_{BB,l_k} \right\|_F}. \tag{21}$$

Consequently, the digital precoding matrix corresponding to the active UEs of BS-l would be in the form of $\mathbf{F}_{\mathrm{BB}_l} = [\mathbf{f}_{\mathrm{BB}_{l1}}, ..., \mathbf{f}_{\mathrm{BB}_{l|\mathcal{A}_l|}}]$.

2.4 Simulation setup

We conduct multiple experiments to assess the performance of the scheme's various layers, focusing on execution time and network key performance indicators such as throughput, delay and jitter, which is defined as the difference of consecutive packet delays of the same UE. Additionally, we evaluate the scalability of the scheme with increasing packet injection rates and varying UE/BS densities. Finally, we introduce and evaluate the sub-coherence time allocation mechanism in different scenarios. In order to simulate the mmWave channel we've adapted the model introduced in [3]. It considers path clusters among BS-UE pairs. Based on the distances, paths have probabilities of being line-of-sight (LOS), non-line-of-sight (NLOS), or too weak to be useful and in outage (OUT). Considering the sparsity of mmWave channel having a few distinct paths [3, 30] we consider a path count of $N_r = 5$. Each path has elevation $\theta \sim U(0, \pi/2)$, azimuth $\phi \sim U(0, 2\pi)$, and a gain with a complex Gaussian distribution $g \sim \mathcal{CN}(0, U_k^{r_\tau - 1} 10^{-0.1 Z_k})$. $U_k^{r_\tau - 1}$ denotes the path-k's power fraction having $U_k \sim U[0, 1]$, and $Z_k \sim \mathcal{N}(0, \zeta^2)$ is its path loss. We use $r_\tau = 2.8$ and, $\zeta = 4.0$ as reference value. We assume the noise power σ^2 to be -90 dBm and the carrier frequency to be 28 GHz [3].

We consider an indoor environment for the simulation, specifically an expo hall with dimensions of 120 m in length, 80 m in width, and 10 m in height as in Fig. 4. The BSs are uniformly distributed across the ceiling, facing downward. All BSs are connected

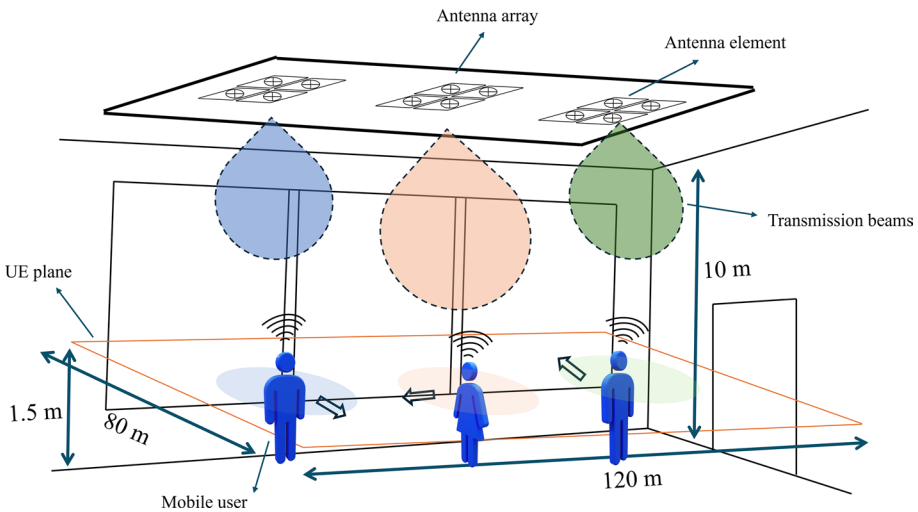

Fig. 4 Illustration of the indoor simulation environment used to evaluate the proposed scheme. Ceiling-mounted antenna arrays positioned at a height of 10 m generate directional beams to serve mobile users located on a 1.5 m elevation plane. The deployment spans an indoor area of 80 m × 120 m, enabling realistic assessment of beamforming strategies in a 3D user distribution setting

to the COU via wired connections, which manages the UE-BS association. Each BS is equipped with a URPA containing $N_T = 60$ antennas, arranged in a length of $N_{T_x} = 12$ and a width of $N_{T_y} = 5$ antennas. The BS total power constraint is $P_{\max} = 1$ Watt [31]. The antenna array can receive from an azimuth range of $(0°, 360°)$ and an elevation range of $(0°, 90°)$. The codebook $\mathcal{W}$ consists of a 16×4 grid of beams, equally spaced in both the azimuth and elevation [27].

The UEs are randomly distributed on a plane positioned 1.5 m above the ground, with a vertical variation of ± 10 cm. They move according to a Gauss-Markov 2D mobility model [29]. This model incorporates random changes in direction and speed, with a maximum speed of 5 km/h, as well as occasional stops, to realistically simulate UE indoor movement patterns for performance evaluation in wireless communication environments.

Considering the rapid changes in the channel caused by mobility and the mmWave carrier frequency, the coherence time is determined to be on the order of a few milliseconds [33]. Accordingly, we set the intervals of various sub-optimizations. The digital beamforming is done once every 1 ms, the analog beamforming is done every 10 ms, and the UE-BS assignment every 100 ms to account for the various dynamic characteristics of the wireless channel while keeping the performance of the system at an acceptable level. Parameters such as the number of UEs, BS, and RF chains per BS are adjusted based on the requirements of each scenario and experiment.

The transmission unit in this simulation is frame, each with a specific size of 2304 bytes. Each UE has a dedicated data stream with its own downlink queue in which frames are introduced. The delay of each frame is calculated as the time between frame's generation and entry to the queue and the time of leaving the queue for transmission. To model the arrival of frames into the queue, we use a Poisson distribution, commonly used for simulating random arrival processes in network traffic. The injection rate refers to the average SE assumed for each UE. Specifically, an injection rate of 1 corresponds to an average SE of 1 b/s/Hz for each UE. We set the rate parameter of the Poisson process so that the average input traffic rate for each UE matches this SE's maximum possible data rate. Unless stated otherwise, the injection rate is 1 in the experiments.

We consider two execution modes: digital (D) and hybrid (H). In the Digital mode, we employ a fully digital multi-BS beamforming setup that includes UE-BS assignment and digital beamforming, with each BS having an equal number of RF chains and antennas in the array ($N_{\mathrm{T}} = N_{\mathrm{RF}}$). In the hybrid mode, we use a hybrid beamforming setup that incorporates UE-BS assignment, analog beamforming, and digital beamforming layers, but with one order of magnitude smaller RF chains count than antennas ($N_{\mathrm{T}} >> N_{\mathrm{RF}}$). For each setup, we ran the experiment 10 times with different random seeds to ensure a diverse sample. The metrics were collected for each channel block in each run and averaged. The experiments were conducted on a system with the following specifications: two Hexa-core Intel E5645 (2.4GHz) CPUs and 24GB RAM.

3 Results and discussion

This section presents and analyzes the results of our experiments, namely: (1) Layer Impact Study, (2) Scalability Study, and (3) Sub-Coherence Time Allocation. Each set of results is examined step by step, followed by a discussion on its significance.

3.1 Layer impact study

Here, we assess the performance of different methods in each layer of the scheme and compare them to the baseline. Starting with the UE-BS assignment, we consider three methods: Gale–Shapley (GS), Random (RND), and Static (STA). The Gale–Shapley method uses the Gale–Shapley algorithm, as discussed in Algorithm 1. The Random method randomly assigns the UEs to their respective BSs, while the Static method performs the Gale–Shapley calculation to create the initial association matrix and keeps it unchanged for the duration of the experiment.

We consider both digital and hybrid modes, utilizing the Greedy algorithm (Algorithm 2) for UE selection and QPFM as the objective function. The parameters for the experiment are set as follows: $L = 4$ UEs, $N_{\mathrm{RF}} = 5$ RF chains, and $K = 20$ BSs. The experiment was conducted over 4000 channel blocks with a warm-up period of 500 blocks, so the results represent the average metrics collected from block 500 to block 4000.

Figure 5 illustrates the distribution of the network delay experienced by UEs under different assignment methods and setup modes. Clearly, the digital mode exhibits a lower delay (50% to 87% better performance in median delay) across all methods due to more effective beamforming and a higher number of RF chains. When examining the assignment methods, we observe that having a small number of BSs ($L = 4$), the random assignment yields lower delay compared to a static approach. Using the median delay as the criterion, we can see that the Gale–Shapley algorithm

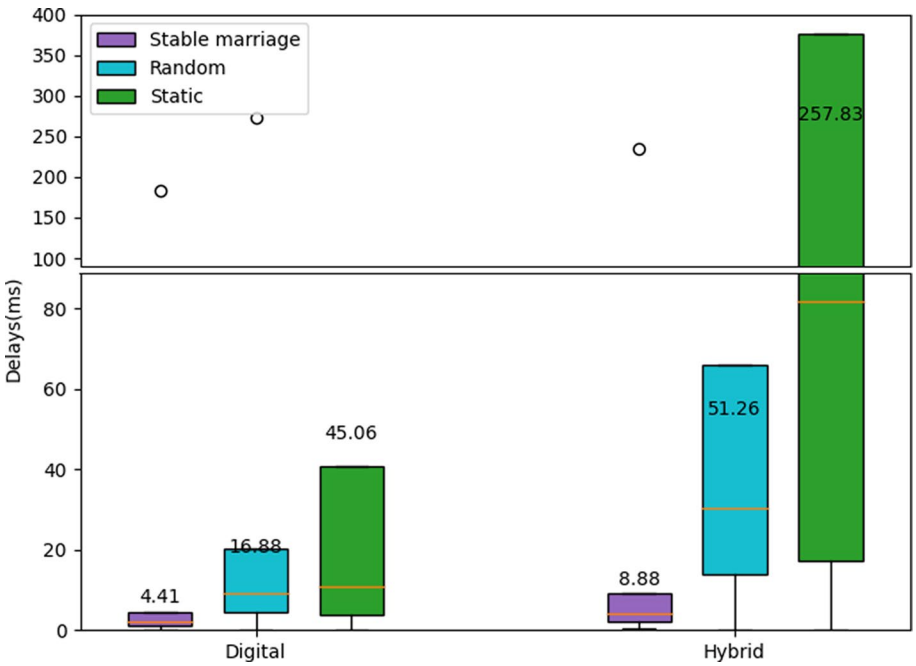

Fig. 5 Box plot comparison of delay across different transmission modes (digital and hybrid) and UE–BS assignment methods (Stable Marriage, Random, and Static). For each configuration, the box spans the interquartile range, the horizontal line within the box represents the median, and the lower whisker indicates the minimum observed delay. The upper whisker is omitted due to large deviations, with maximum values represented by circles. To enhance clarity and avoid distortion caused by extreme outliers, the vertical axis is split. The average delay for each case is annotated directly on the corresponding box

outperforms the random and static approaches by 78% and 82%, respectively, in the digital mode. This performance gain is even more pronounced in the hybrid mode, achieving improvements of 87% and 95%, respectively. Figure 6 illustrates the distribution of active UEs count. Due to similar activation criteria, a higher number of simultaneously active UEs implies better performance, which is clearly the case for the Gale–Shapley method having 15% to 25% higher active user count compared to the others. Based on these results, we can conclude that utilizing the Gale–Shapley method in the UE-BS assignment layer leads to better performance in both digital and hybrid operation modes.

The next step is to compare the performance of different UE selection algorithms: the Greedy algorithm (Algorithm 2), the Adaptive Top-k algorithm (Algorithm 3), and Serving All (SA), which keeps all the UEs assigned to a BS active all the time. We will also compare different objective functions: BM, PFM, QLM, and QPFM. Both digital and hybrid modes will be considered, using the Gale–Shapley algorithm for UE assignment. The UE, BS, RF chain count, and experiment duration remain consistent with the previous experiment. To compare the execution time, we break the total computation into four parts: UE association, beam sweeping, UE selection, and digital beamforming.

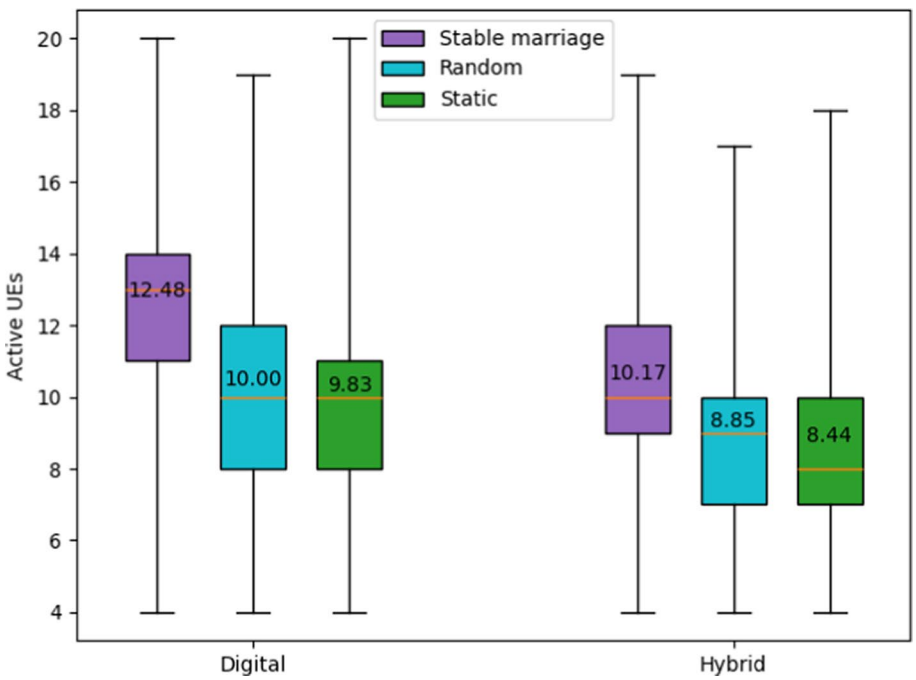

Fig. 6 Box plot showing the number of active UEs under different UE–BS assignment methods (Stable Marriage, Random, and Static) across two operation modes: digital and hybrid. Each box illustrates the interquartile range, with the horizontal orange line indicating the median. Whiskers extend to the minimum and maximum values within each group. The average number of active UEs is labeled directly on each box to highlight overall performance differences between methods

1. **UE Association** (Layer 1): This sub-problem involves all computations necessary to perform Gale–Shapley algorithm, taking into account the current BS-UE channel conditions.
2. **Beam Sweeping** (Layer 2): The aim of this sub-problem is to find the optimal analog beamformer from the codebook for each UE.
3. **UE Selection** (Layer 3): In this step, the selection algorithm and objective function are utilized to determine which UEs should remain active.
4. **Digital Beamforming** (Layer 3): Given the set of active UEs for each BS, the zero-forcing technique is applied to obtain the digital beamformer.

Please note that the execution times reported in this study are hardware dependent. Since the simulations were performed on a general-purpose machine, the observed execution times are higher than those that would be achieved on dedicated signal processing hardware. Our primary objective was to compare relative performance metrics rather than absolute real-time performance. In a practical deployment scenario utilizing specialized signal processing hardware, the execution times would be significantly reduced, providing a more accurate representation of the system's real-world performance capabilities.

An example of a more efficient hardware can be Resistive Random-Access Memory (RRAM), which allows for massively parallel, high-speed, and energy-efficient computations and is particularly attractive for MIMO precoder computation. [34] shows that using this technique, a 91× lower computation time can be achieved. Similarly, [35] reports a two-order-of-magnitude improvement in execution time using a similar approach. As in Fig. 7, we observed an execution time of 250 ns for zero-forcing in a 5×40 MIMO system, while [35] reports an execution time of 20 ns for a 16×128

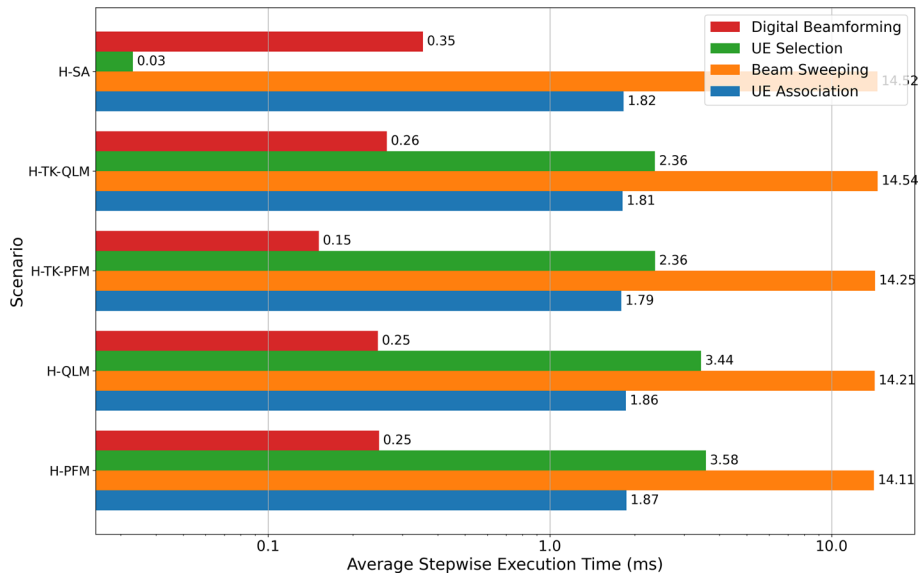

Fig. 7 Comparing the average execution time for each computational step in a single execution, without accounting for differences in execution frequency across steps. Comparing different objective functions and UE-BS selection methods in hybrid operation mode. Note that the x axis has a logarithmic scale. H: hybrid, TK: adaptive top-k, SA: serving all, PFM: proportional fairness maximization, and QLM: queue length minimization

system utilizing RRAM. This highlights the huge potential for speedup in the computation time when using specialized signal processing hardware.

Figure 7 illustrates the average execution time for a single run of each step under different UE selection methods and objective functions in hybrid mode. The two bar groups at the very bottom display the performance of the greedy algorithm, the next two upper groups represent the adaptive top-k method, and the single topmost group shows the Serve All approach. The Beam Sweeping step has the highest execution time (∼14 ms) due to the large number of BS-UE pair channel matrix assessments and its exhaustive search approach. However, its lower execution frequency (every 10 channel blocks) makes it manageable for the BSs. Regarding UE selection, the adaptive top-k method requires 30% less time than the greedy algorithm. UE association takes relatively little time (∼1.8ms) and, due to its infrequent execution (every 100 blocks), has the least computational requirement. Digital beamforming by zero-forcing has a low execution time (∼0.25ms) but, due to frequent execution, contributes significantly to the total execution time.

Figure 7 does not account for the effect of the execution frequency. To provide a more comprehensive understanding of the total computational requirements of different steps, Fig. 8 visualizes the average total execution time of each step over 1000 channel blocks. As observed, despite having a relatively low stepwise execution time, UE selection dominates the total execution time due to its high execution frequency. The Greedy approach results in 80% higher execution time compared to the Adaptive Top-K method. Additionally, beam sweeping experiences a 23% increase in execution time in the Greedy approach due to a larger number of simultaneously active UEs. In

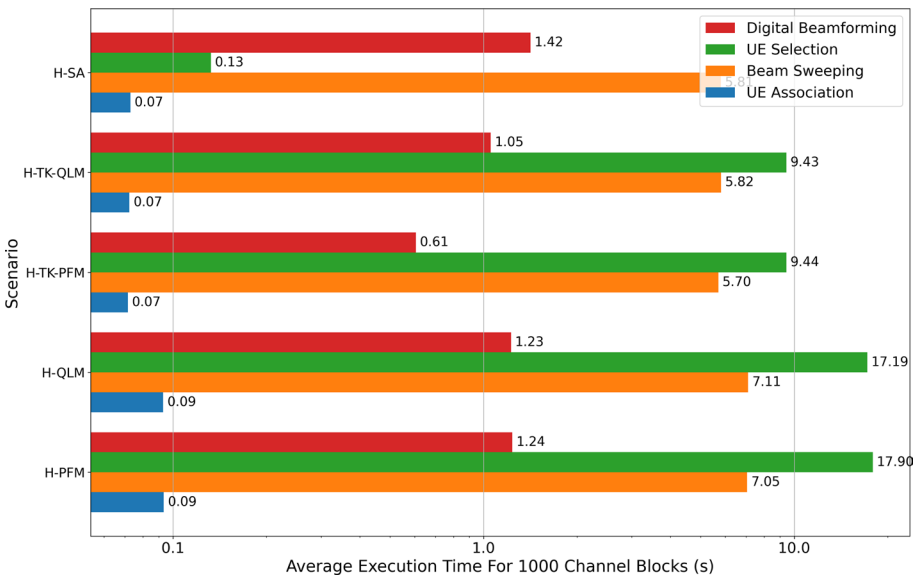

Fig. 8 Comparison of the total execution time for processing 1000 channel blocks under different objective functions and UE-BS selection methods in hybrid operation mode. This figure provides an aggregated view of execution time, capturing the cumulative computational cost of all steps involved in each scenario. Note that the x axis has a logarithmic scale. H: hybrid, TK: adaptive top-k, SA: serving all, PFM: Proportional fairness maximization and QLM: Queue length minimization

contrast, digital beamforming and UE association consistently contribute the least to the overall computation time.

A comparison of the distribution of the network delay under different objective functions and UE selection algorithms in digital mode can be found in Fig. 9. The Serve All approach results in the highest variance of delay, whereas the greedy and adaptive top-k approaches exhibit similar performance. Given the lower execution time of the adaptive top-k method (30% less time), it emerges as the better choice for UE selection in digital mode. However, this conclusion does not extend to the hybrid mode (which is not illustrated in Fig. 9 because of space limitations), where the adaptive top-k method yields 50% higher average delay compared to the greedy algorithm, rendering it unsuitable for hybrid applications. Regarding objective functions, BM consistently exhibits the highest delay variance and the poorest performance, followed by PFM, which shows moderate improvement. Queue-based objective functions, specifically QLM and QPFM, deliver the best performance across all scenarios, having the lowest variance and average delay.

3.2 Scalability study

Next, we assess the scalability and robustness of our proposed scheme under increasing traffic and UE/BS deployment density in the environment. We begin by examining the impact of increasing the traffic injection rate on different objective functions in both digital and hybrid modes. Following the previously defined injection rate parameter, we consider five rates: 0.3, 0.7, 1, 1.7 and 3. The Gale–Shapley algorithm is used for UE assignment, with $L = 4$ UEs, $N_{\mathrm{RF}} = 5$ RF chains, and $K = 20$ BSs. Each experiment runs for a duration of 3000 channel blocks, including a warm-up period of 500 blocks.

The growth of delay and jitter under increasing injection rate in different setups can be seen in Figs. 10 and 11 on a logarithmic scale. At the lowest injection rate, the average delay and jitter values in the hybrid mode are 70% to 96% higher compared to the

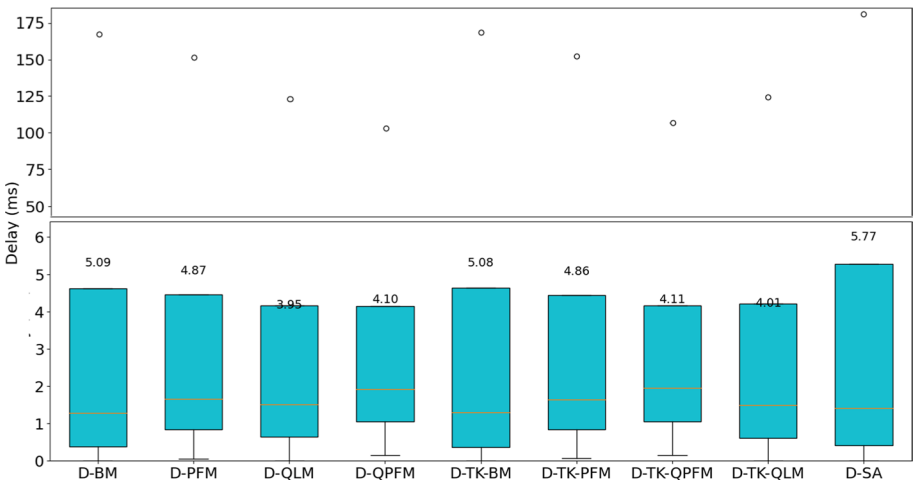

Fig. 9 Box plot of the delay under different objective functions and UE-BS selection methods in the Digital mode. Each box spans the first and third quartiles, with a horizontal line showing the median. The whiskers extend to the minimum data points that are not outliers, while maximums that are outliers are displayed as circles. Because maximums are relative very large, the vertical axis is split to maintain a clear view of the main data distribution. The average delay for each case is labeled on the corresponding box. Abbreviations: D—Digital, TK—adaptive top-k, SA—serving all, BM—bitrate maximization, PFM—proportional fairness maximization, QLM—queue length minimization, QPFM - queue-aware proportional fairness maximization

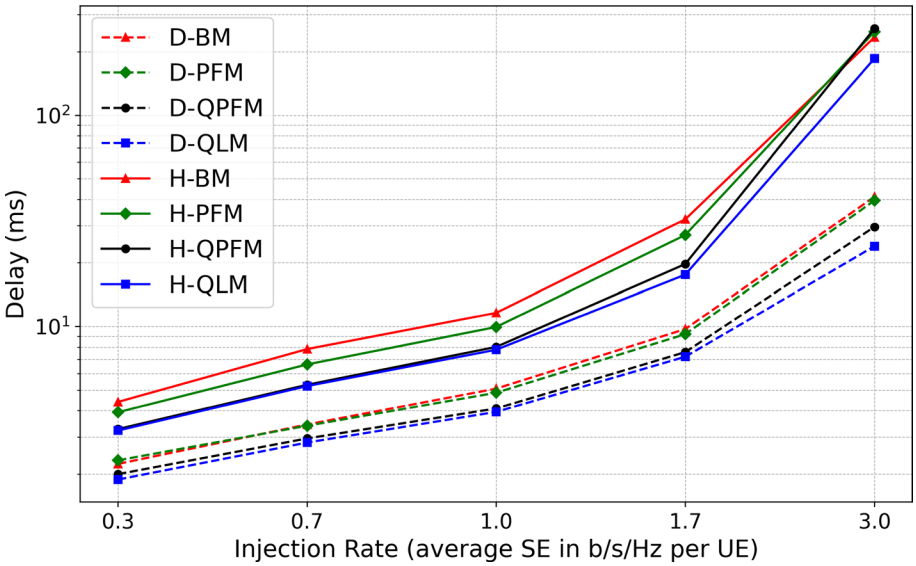

Fig. 10 Comparing the delay of different objective functions under increasing injection rate and two operation modes on a logarithmic scale. D: digital, H: hybrid, BM: bitrate maximization, PFM: proportional fairness maximization, QLM: queue length minimization and QPFM: queue-aware proportional fairness maximization

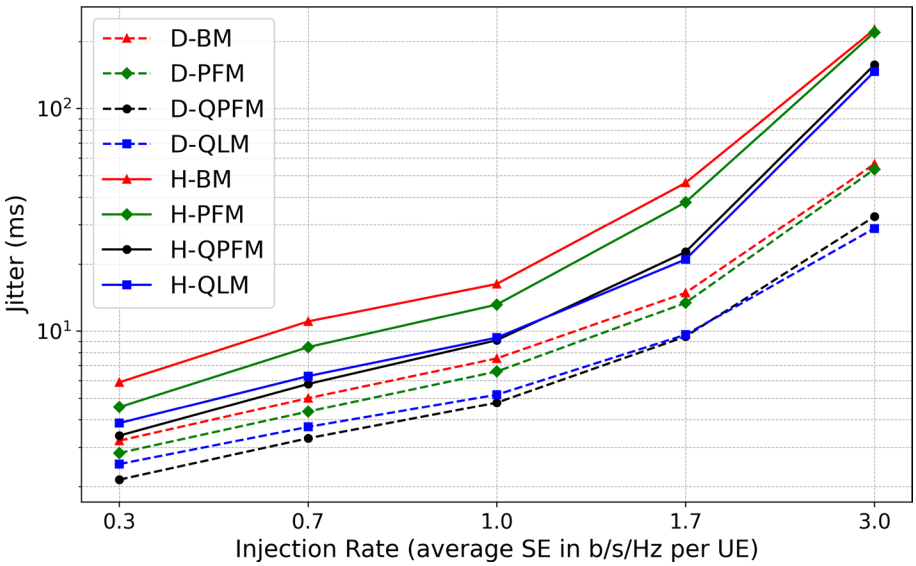

Fig. 11 Comparing the jitter of different objective functions under increasing injection rate and two operation modes on a logarithmic scale. D: digital, H: hybrid, BM: bitrate maximization, PFM: proportional fairness maximization, QLM: Queue length minimization and QPFM: queue-aware proportional fairness maximization

digital mode. This disparity becomes more pronounced as the injection rate increases, reaching 144% to 230% at the 1.7 rate. This implies better scalability and performance in digital mode Considering the objective functions, BM consistently shows the worst performance and scalability, followed by PFM. QPFM and QLM exhibit the highest performance and scalability against increasing input traffic.

We also investigated the impacts of extreme injection rates on the performance of objective functions. Jumping from the rate of 1.7 to rate of 3 in hybrid mode, we observed a significant increase in the average delay, jitter, and queue size for both QPFM (1200%) and PFM (800%), even exceeding those of BM (620%). These findings indicate that the effectiveness of proportional fairness-based approaches is limited to regular, unsaturated network operation regimes. Under extremely saturated conditions (e.g., injection rates > 2), when queues tend to grow very large, QLM proves to be a better alternative.

Next, we evaluate the impact of increased deployment density on the performance of objective functions. We consider two scenarios: an increase in UE density and an increase in BS density. In the first scenario, we maintain a constant number of BSs but increase the number of UEs and RF chains per BS, resulting in more UEs per BS. In the second scenario, both the number of BSs and UEs increase while keeping the RF chain count per BS constant. This means that although the number of UEs per BS remains unchanged, the deployment area becomes denser, leading to higher interference levels. In the UE density increase scenario, both inter-BS and intra-BS interference rise, whereas in the BS density increase scenario, only inter-BS interference worsens.

We begin by examining the impact of increased UE density in both digital and hybrid modes, employing the Gale–Shapley algorithm for UE assignment and the Greedy algorithm for UE selection. We consider $K = 4$ BSs with an increasing number of total UEs: 8, 20, 32, 64, and 128. Correspondingly, the RF chain counts are 2, 5, 8, 16, and 32. The total duration of the experiment is 2000 blocks, including a warm-up period similar to the previous experiments. Figure 12 illustrates the execution time of each step for different numbers of UEs on a logarithmic scale. As observed, UE association, beam sweeping, and digital beamforming, which have a computational complexity of $\mathcal{O}(N_{\mathrm{RF}})$, exhibit linear growth. However, the most significant growth occurs in the UE selection step, which has a complexity of $\mathcal{O}(N_{\mathrm{RF}}^2)$.

Figure 13 compares the effect of increasing the UE count on delay across different scenarios on a logarithmic scale. For the hybrid mode, the chart can be divided into two

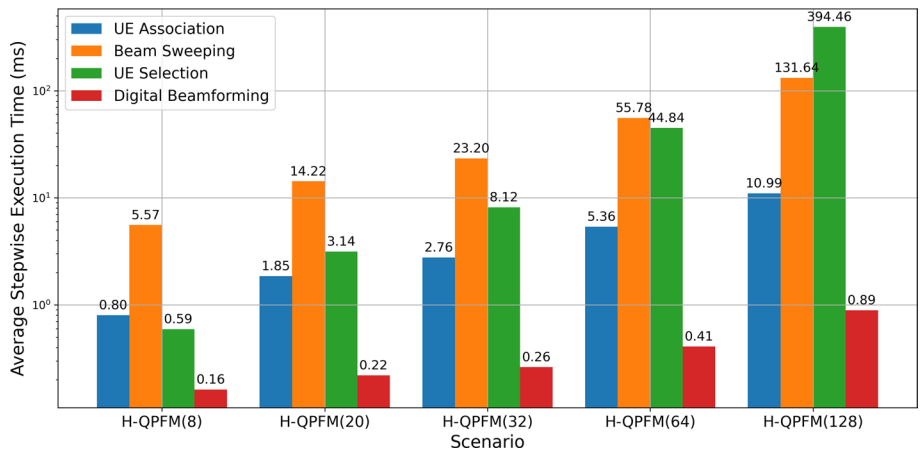

Fig. 12 Comparing the execution time of different computation step under increasing UE count on a logarithmic scale. The number of all UEs is written in the parenthesis. H: hybrid, BM: bitrate maximization, PFM: proportional fairness maximization, QLM: queue length minimization and QPFM: queue-aware proportional fairness maximization

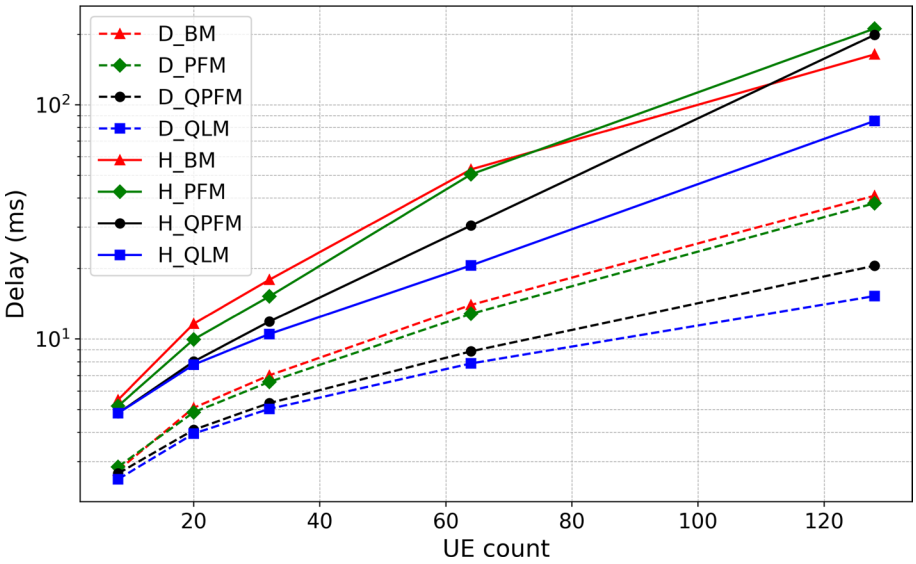

Fig. 13 Comparing the delay of different objective functions under increasing UE count on a logarithmic scale. D: digital, H: hybrid, BM: bitrate maximization, PFM: proportional fairness maximization, QLM: queue length minimization and QPFM: queue-aware proportional fairness maximization

conditions: the unsaturated regime ($L \leq 64$) and the saturated regime ($L > 64$). In the unsaturated regime, BM shows the worst performance, followed by PFM, while QPFM and QLM exhibit the lowest delay. In the saturated regime, the effectiveness of PF-based methods decreases, rendering QPFM and PFM unsuitable. The digital mode offers a higher capacity, showing a broader unsaturated range ($L \leq 128$).

For the BS density increase, we examine the hybrid mode, using the Gale–Shapley algorithm for UE assignment and the Greedy algorithm for UE selection. We consider $N_{\text{RF}} = 5$ RF chains per BS with an increasing number of BSs: 2, 4, 8, 16, and 32. Correspondingly, the UE counts are 10, 20, 40, 80, and 160. As shown in Fig. 14, in the unsaturated regime ($K < 16$), BM exhibits the worst performance, followed by PFM, QPFM, and QLM. In the saturated regime, there is a noticeable performance drop for PFM and QPFM, rendering them ineffective. It is important to note that with the increase in the number of BSs, inter-BS interference becomes dominant, causing the delay to grow noticeably under all objective functions.

3.3 Sub-coherence time allocation

Traditionally, the digital precoder is calculated at the beginning of a coherence time and remains effective throughout its duration [7, 11, 12]. In an indoor setup with high bandwidth and favorable signal power, some UEs might have exhausted their downlink queue while still considered in the zero-forcing beamformer. This situation unnecessarily limits other UEs and lowers the overall network capacity. To address this issue, we propose a new digital beamforming method called **speculative sub-coherence time allocation**. The key idea is to divide the coherence time into multiple sub-slots based on the network usage of UEs and calculate a different digital precoder for each sub-slot. As in Fig. 15, by examining the queue lengths of active UEs and their expected rates, we can estimate the time needed for their queues to be exhausted. If a UE is expected

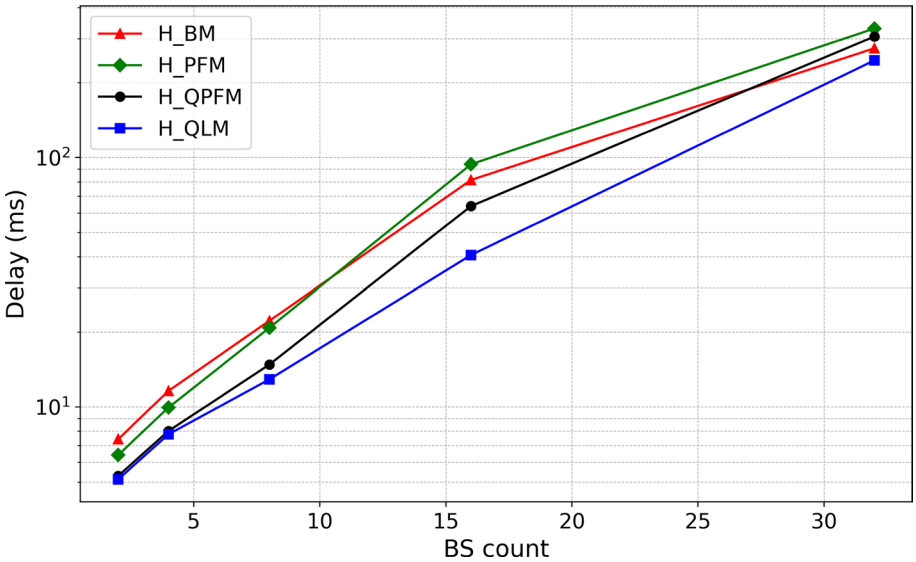

Fig. 14 Comparing the delay of different objective functions under increasing BS count on a logarithmic scale. H: hybrid, BM: bitrate maximization, PFM: proportional fairness maximization, QLM: queue length minimization and QPFM: queue-aware proportional fairness maximization

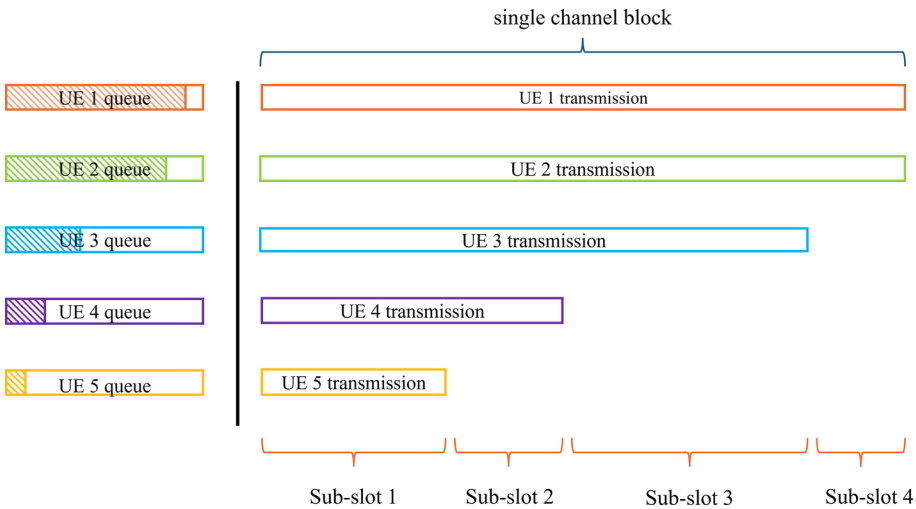

Fig. 15 Illustration of a single channel block subdivided into multiple sub-slots, each allocated according to the downlink queue length of the user. The left side shows the queue sizes for different UEs, while the right side indicates how each UE's transmission is scheduled in a sub-slot proportional to its backlog. This queue-aware scheduling enables flexible and efficient resource utilization within each channel block

to exhaust its queue before the end of the coherence time, it is removed from the active UE set for the remainder of the coherence time, and the zero-forcing precoder is recalculated without it. As illustrated in Fig. 15, at the beginning of a timeslot, we calculate the transmission time for each UE based on its queue length and expected rate. Starting from the first sub-slot, all UEs are considered, and the precoder is calculated. When a UE completes its transmission, a new sub-slot begins with one less active UE, and a new precoder is calculated. All these speculative calculations are performed at the beginning

of a timeslot, and the computed digital beamformers are activated sequentially at their designated times.

To test the effectiveness of this approach, we conducted two sets of experiments with identical setups, except one utilized multi-precoding while the other used a single precoder. We examined both hybrid and digital modes, employing the Gale–Shapley algorithm for UE assignment and the Greedy algorithm for UE selection. Only queue-based objective functions, QLM and QPFM, were considered. The layout parameters were set as follows: $L = 4$ UEs, $N_{\mathrm{RF}} = 5$ RF chains, and $K = 20$ BSs. The experiments were conducted over 1000 channel blocks with an identical warm-up period. It should be noted that this experiment assumes no retransmissions based on the HARQ mechanism.

Figure 16 compares the distribution of the network delay and demonstrates the improvement achieved with multi-precoding under different modes and objective functions. In digital mode, the average delay is reduced by 25% to 30%, accompanied by halving of the delay variance. In contrast, the hybrid mode exhibits a more modest improvement, with an average delay reduction ranging from 11% to 25%. This approach proves more effective in the fully digital mode due to the greater flexibility provided by the larger number of RF chains. Moreover, the fully queue-based QLM benefits more than QPFM. Notably, the improvement in network experience is evident in all scenarios, and coupled with the low computational overhead of the multi-precoder approach ($\sim$ 16% of the total execution time) it is a promising augmentation technique for next generation MIMO systems.

3.4 Discussion

Although system-level performance indicators such as throughput provide a broad view of overall efficiency, they do not align closely with our objective of fair and effective

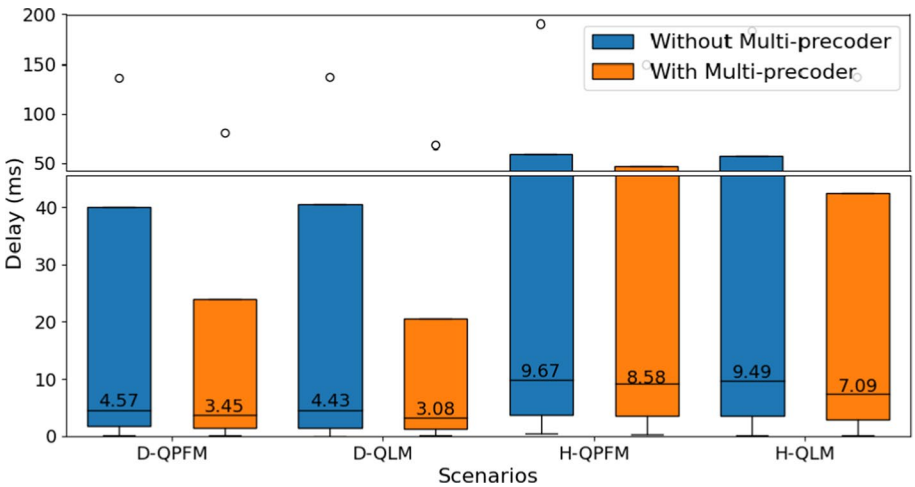

Fig. 16 Box plot of the delay under different objective functions and modes (digital/hybrid), comparing scenarios with and without sub-coherence time allocation. Each box spans the first and third quartiles, with a horizontal line for the median, and the whisker indicates the minimum. Because maximum values are outliers, they appear as a small circle. The vertical axis is split into two parts, each using a different scale to better visualize both typical and extreme values. The average delay in each scenario is printed on its bar. Abbreviations: D—digital, H—hybrid, QLM—queue length minimization, QPFM—queue-aware proportional fairness maximization

resource allocation. Instead, we prioritize metrics such as delay and jitter, which more accurately capture the impact of resource allocation on individual UEs. By emphasizing these user-centric indicators, we highlight how effectively each UE's requirements are being addressed.

A fully digital MIMO system, due to its higher RF chain count and fewer limitations, offers better overall performance, scalability, and a broader unsaturated operation region. However, in XL-MIMO setups, a purely digital configuration becomes impractical due to higher costs, increased energy consumption, and thermal issues. Therefore, a trade-off is necessary, opting for a hybrid approach. Purely queue-based objective functions (e.g., QLM) demonstrate visibly better performance and scalability compared to simple bitrate maximization (e.g., BM) or proportional fairness-based methods (e.g., PFM, QPFM). Proportional fairness is highly sensitive to network saturation and quickly loses its effectiveness, whereas a queue-based approach maintains its efficacy under varying injection rates and BS/UE densities. Hence, a queue-based objective function is recommended for designing future mmWave MIMO systems. UE assignment and active UE selection are as crucial as beamforming. Not all UEs should be served simultaneously. Effectively selecting the UEs to be served during each coherence time is as important as the technique used for beamforming. Additionally, sub-coherence time allocation can be considered for its superior resource utilization and improved performance across different setups and scenarios.

4 Conclusions

In this paper, we studied the performance of different resource allocation techniques in digital/hybrid beamforming and proposed a multi-layer scheme to enhance the network experience of UEs in a multi-BS XL-MIMO setup, considering various forms of dynamism in wireless channels. The sub-problems of UE-BS assignment, analog beamforming, active UE selection, and digital beamforming are executed at different intervals to minimize the delay and maximize the throughput for the UEs. The execution time and performance are balanced by solving each sub-problem when necessary. Additionally, we introduced an extension called sub-coherence time allocation to better utilize the limited channel resources which yields 11% to 30% lower average network delay under different operation modes. Through multiple experiments, we demonstrate the superiority of purely queue-based objective functions considering performance and scalability. They also offer a wider unsaturated operational regime. Our findings suggest that, rather than serving all UEs simultaneously, it is more effective to serve a subset of them during each coherence time.

For future work, developing the proposed scheme into a deployable protocol for multi-BS MIMO setups would be a valuable next step. Additionally, extending the scheme to support time-sensitive networking applications offers another intriguing research direction. Given that our current assumptions rely on the availability of a perfect channel matrix, future efforts should explore pilot sequencing and channel estimation techniques to address practical scenarios. A promising avenue for further research involves replacing fixed, wired-connected BSs with mobile, wireless ones. In such dynamic environments, the same beamforming techniques can be adapted not only for BS-UE connections but also for BS-BS links. To optimize the performance of

analog beamforming, effective channel budgeting techniques and trajectory prediction methods will be crucial.

Abbreviations

AWGN	Additive White Gaussian Noise
BB	Baseband
BM	Bitrate Maximization
BS	Base Station
COU	Central Orchestration Unit
CSI	Channel State Information
D	Digital (operation mode)
H	Hybrid (operation mode)
LOS	Line-of-Sight (in channel model)
MIMO	Multiple-Input Multiple-Output
mmWave	Millimeter-Wave
MU-MIMO	Multi-User MIMO
NLOS	Non-Line-of-Sight (in channel model)
OMNeT++	Objective Modular Network Testbed in C++ (a network simulation framework)
OUT	Outage (in channel model)
PF	Proportional Fairness
PFM	Proportional Fairness Maximization
QLM	Queue Length Minimization
QPFM	Queue-Aware Proportional Fairness Maximization
RF	Radio Frequency
RND	Random (UE-BS assignment method)
SA	Serving All (UE selection method)
SE	Spectral Efficiency
SINR	Signal-to-Interference-plus-Noise Ratio
GS	Gale–Shapley (UE-BS assignment method)
STA	Static (UE-BS assignment method)
TK	Adaptive Top-K (UE selection algorithm)
ULA	Uniform Linear Array
URPA	Uniform Rectangular Planar Array
UE	User Equipment
XL-MIMO	Extremely Large MIMO
ZF	Zero-Forcing

Acknowledgements
This work was supported in part by the Methusalem funding of the Flemish Government under Grant "SHAPE: Next Generation Wireless Networks" and in part by the Flemish FWO SBO S003921N VERI-END.com (Verifiable and elastic end-to-end communication infrastructures for private professional environments) project

Author Contributions
Mohammadreza Heydarian: MH came up with the research subject. He has designed the multi-timeslot scheme and its optimization steps. He has developed the simulation tool, performed the experiments and analyzed the results. MH has drafted the manuscript and designed the graphs and figures0.Didier Colle: DC supervised the research. He has contributed to the conception and design of the research work, interpretation of the results and revision of the manuscript. Wouter Tavernier: WT supervised the research. He has contributed to the conception and design of the research work, interpretation of the results and revision of the manuscript. Mario Pickavet: MP has substantially revised the manuscript.

Funding
The funders below provided the necessary funding to conduct the research, but have no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript: Methusalem funding of the Flemish Government: Grant "SHAPE: Next Generation Wireless Networks".Flemish FWO: SBO S003921N VERI-END.com (Verifiable and elastic end-to-end communication infrastructures for private professional environments) project.

Data availability
The source code of the simulation software developed during this research and the collected data from the simulation experiments which were analyzed to get the results are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 23 August 2024 Accepted: 18 June 2025
Published online: 10 July 2025

References

1. A. Goldsmith et al., Capacity limits of MIMO channels. IEEE J. Sel. Areas Commun. **21**(5), 684–702 (2003)
2. E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, T.L. Marzetta, Massive MIMO is a reality?What is next?: five promising research directions for antenna arrays. Digit. Signal Process. **94**, 3–20 (2019)
3. M.R. Akdeniz et al., Millimeter wave channel modeling and cellular capacity evaluation. IEEE J. Sel. Areas Commun. **32**(6), 1164–1179 (2014)
4. M.R. Akdeniz et al., Millimeter wave channel modeling and cellular capacity evaluation. IEEE J. Sel. Areas Commun. **32**(6), 1164–1179 (2014)
5. Minhyun, Kim, Junghoon, Lee, Junhwan, Lee. Hybrid beamforming for multi-user transmission in millimeter wave communications. 2017 International Conference on Information and Communication Technology Convergence (ICTC). IEEE (2017)
6. de Souza, J.H. Inacio et al., Quasi-distributed antenna selection for spectral efficiency maximization in subarray switching XL-MIMO systems. IEEE Trans. Veh. Technol. **70**(7), 6713–6725 (2021)
7. Kim, Junghoon, Andrews, Matthew. Learning-Based Adaptive User Selection in Millimeter Wave Hybrid Beamforming Systems. arXiv preprint arXiv:2302.08240 (2023)
8. A. Eryilmaz, R. Srikant, Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. IEEE/ACM Trans. Networking **15**(6), 1333–1344 (2007)
9. Liu, Lingjia, Nam, Young-Han, Zhang, Jianzhong. Proportional fair scheduling for multi-cell multi-user MIMO systems. 2010 44th Annual Conference on Information Sciences and Systems (CISS). IEEE (2010)
10. T. Stahlbuhk, B. Shrader, E. Modiano, Learning algorithms for minimizing queue length regret. IEEE Trans. Inf. Theory **67**(3), 1759–1781 (2021)
11. P. Ni et al., User association and hybrid beamforming designs for cooperative mmWave MIMO systems. IEEE Transactions on Signal and Information Processing over Networks **8**, 641–654 (2022)
12. M. Hanif et al., Antenna subset selection for massive MIMO systems: a trace-based sequential approach for sum rate maximization. J. Commun. Netw. **20**(2), 144–155 (2018)
13. W. Liu et al., Deep learning based beam training for extremely large-scale massive MIMO in near-field domain. IEEE Commun. Lett. **27**(1), 170–174 (2022)
14. Reza, M., Abedi, et al. Safety-Aware Age-of-Information (S-AoI) for Collision Risk Minimization in Cell-Free mMIMO Platooning Networks. IEEE Transactions on Network and Service Management (2024)
15. Abedi, M. R. et al. Low Complexity and Mobility-aware Robust Radio, Storage, Computing, and Cost Management for Cellular Vehicular Networks. IEEE Transactions on Vehicular Technology (2024)
16. A.F. Molisch et al., Hybrid beamforming for massive MIMO: A survey. IEEE Commun. Mag. **55**(9), 134–141 (2017)
17. T. Lin et al., Hybrid beamforming for millimeter wave systems using the MMSE criterion. IEEE Trans. Commun. **67**(5), 3693–3708 (2019)
18. H. Kim et al., Beamforming and power allocation designs for energy efficiency maximization in MISO distributed antenna systems. IEEE Commun. Lett. **17**(11), 2100–2103 (2013)
19. S. Xie, L. Ai, Sum?rate optimization scheme for time?varying distributed MU?MIMO systems. IET Commun. **15**(19), 2482–2491 (2021)
20. Tian, X. et al. Sum rate maximization in muti-cell muti-user networks: an inverse reinforcement learning-based approach. IEEE Wireless Communications Letters (2023)
21. H.S. Vu, K. Truong, M.T. Le, Beam division multiple access for millimeter wave massive MIMO: Hybrid zero-forcing beamforming with user selection. Int. J. Electr. Comput. Eng.(IJECE) **12**(1), 445 (2022)
22. A. Alkhateeb et al., Channel estimation and hybrid precoding for millimeter wave cellular systems. IEEE J. Select. Top. Signal Process. **8**(5), 831–846 (2014)
23. Xiang, G., et al. Linear pre-coding performance in measured very-large MIMO channels. 2011 IEEE vehicular technology conference (VTC Fall). IEEE (2011)
24. Wahab, K., Ozdemir, O., Guvenc, I. Temporal and spatial characteristics of mm wave propagation channels for UAVs. 2018 11th Global Symposium on Millimeter Waves (GSMM). IEEE (2018)
25. Mehdi, G., et al. A new codebook design for analog beamforming in millimeter-wave communication. arXiv preprint arXiv:1902.00838 (2019)
26. Hyunsoo, L., et al. Stable marriage matching for traffic-aware space-air-ground integrated networks: A gale-shapley algorithmic approach. 2022 International Conference on Information Networking (ICOIN). IEEE (2022)
27. J. Mo et al., Beam codebook design for 5G mmWave terminals. IEEE Access **7**, 98387–98404 (2019)
28. A. Alkhateeb, G. Leus, R.W. Heath, Limited feedback hybrid precoding for multi-user millimeter wave systems. IEEE Trans. Wireless Commun. **14**(11), 6481–6494 (2015)
29. Jean-Daniel Medjo Me, B., Kunz, T., St-Hilaire, M. An enhanced Gauss-Markov mobility model for simulations of unmanned aerial ad hoc networks. 2014 7th IFIP Wireless and Mobile Networking Conference (WMNC). IEEE (2014)
30. S. Rangan, T.S. Rappaport, E. Erkip, Millimeter-wave cellular wireless networks: potentials and challenges. Proc. IEEE **102**(3), 366–385 (2014)
31. S. Loyka, The capacity of Gaussian MIMO channels under total and per-antenna power constraints. IEEE Trans. Commun. **65**(3), 1035–1043 (2017)
32. Weiqiang, T., et al. Analysis of different planar antenna arrays for mmWave massive MIMO systems. 2017 IEEE 85th Vehicular Technology Conference (VTC Spring). IEEE (2017)
33. Yasaman, K., et al. Beam Coherence Time Analysis for Mobile Wideband mmWave Point-to-Point MIMO Channels. IEEE Wireless Communications Letters (2024)
34. Qunsong, Z., et al. Realizing In-Memory Baseband Processing for Ultra-Fast and Energy-Efficient 6G. IEEE Internet of Things Journal (2023)
35. P. Zuo, Z. Sun, R. Huang, Extremely-fast, energy-efficient massive MIMO precoding with analog RRAM matrix computing. IEEE Trans. Circuits Syst. II Express Briefs **70**(7), 2335–2339 (2023)

Mohammadreza Heydarian Mohammadreza Heydarian received his BSc degree in Computer Engineering from the Iran University of Science and Technology in 2016, followed by an MSc in Computer Engineering from Sharif University of Technology in 2019. Since 2022, he has been pursuing his PhD at Ghent University, where he is a member of the Fixed Internet Architectures & Optical Networks (FARON) group. His research involves multiple projects, including "SHAPE: Next Generation Wireless Networks" and high-performance computing (HPC). Currently his research focuses on the networking aspects of mmWave MIMO systems.

Didier Colle Didier Colle is senior full professor at Ghent University since 2022. He was associated professor since 2011 and full professor since 2014 at the same university and received a PhD degree in 2002 and a M. Sc. degree in electrotechnical engineering in 1997 from the same university. He is co-responsible for the research cluster on network modeling, design and evaluation (NetMoDeL) inside the IMEC IDLab research group. This research cluster deals with fixed internet architectures and optical networks, green-ict, design of network algorithms and techno-economic studies. His research is mainly conducted inside international (mainly European), national and bilateral research projects together with the industry. This research has been published in more than 500 international journal and conference articles and has resulted in more than 20 PhD degrees.

Mario Pickavet Mario Pickavet received M.Sc. and Ph.D. degrees in electrical engineering, specialized in telecommunications in 1996 and 1999, respectively. Since 2000, he is professor at Ghent University where he is teaching courses on discrete mathematics and network modeling. He is co-leading the research cluster on Network Modeling, Design and Evaluation (NetMoDeL), together with Didier Colle. His main research interests are Fixed internet architectures and optical networks, green ICT and design of network algorithms. In this context, he is currently involved in several national and international research projects. He has published more than 500 international publications, both in journals (IEEE JSAC, IEEE Comm. Mag., Journal of Lightwave Technology, Proceedings of the IEEE, ...) and in proceedings of conferences. He is co-author of the book 'Network Recovery: Protection and Restoration of Optical, SONET-SDH, IP, and MPLS.' He is holder of a bronze medal at the International Mathematical Olympiad (Sweden, 1991).

Wouter Tavernier Wouter Tavernier received his BS and MS degree in Computer Science in 2002 from Ghent University (Belgium). He joined the Internet-Based Communications Networks group (which became part of IDLab in October 2016) of Ghent University in 2006 as researcher on Carrier Ethernet. In 2012 he obtained a Ph.D. degree from the same university on reliable routing and switching. Currently, he is employed as Professor at Ghent University, where he teaches courses on computer networks. His current research interests focus on performance and resource optimization aspects of deterministic and high-performance computing networks. This work is performed in the context of European projects such as 5GPPP NGPAAS, SONATA-NFV, 5G TANGO, and Horizon Europe projects such as HEXA-X-II and OASEES. This research has been published in more than 120 scientific publications.