# Clustering-Based Downlink Scheduling of IRS-Assisted Communications With Reconfiguration Constraints

Alberto Rech, *Member, IEEE*, Matteo Pagin, *Student Member, IEEE*, Leonardo Badia, *Senior Member, IEEE*, Stefano Tomasin, *Senior Member, IEEE*, Marco Giordani, *Member, IEEE*, Jonathan Gambini, Michele Zorzi, *Fellow, IEEE*

***Abstract*—Intelligent reflecting surfaces (IRSs) are being widely investigated as a potential low-cost and energy-efficient alternative to active relays for improving coverage in next-generation cellular networks. However, technical constraints in the configuration of IRSs should be taken into account in the design of scheduling solutions and the assessment of their performance. To this end, we examine an IRS-assisted time division multiple access (TDMA) cellular network where the reconfiguration of the IRS incurs a communication cost; thus, we aim at limiting the number of reconfigurations over time. Along these lines, we propose a clustering-based heuristic scheduling scheme that maximizes the cell sum capacity, subject to a fixed number of reconfigurations within a TDMA frame. First, the best configuration of each user equipment (UE), in terms of joint beamforming and optimal IRS configuration, is determined using an iterative algorithm. Then, we propose different clustering techniques to divide the UEs into subsets sharing the same suboptimal IRS configuration, derived through distance- and capacity-based algorithms. Finally, UEs within the same cluster are scheduled accordingly. We provide extensive numerical results for different propagation scenarios, IRS sizes, and phase shifters quantization constraints, showing the effectiveness of our approach in supporting multi-user IRS systems with practical constraints.**

***Index Terms*—Intelligent Reflecting Surfaces (IRS); millimeter wave (mmWave) communication; multiple access; scheduling; clustering; optimization.**

## I. Introduction

The ever-increasing growth of mobile traffic has called both academia and industry to identify and develop solutions for extending the radio spectrum beyond the crowded sub-6 GHz bands. As a result, the use of millimeter wave (mmWave) band for cellular communications has been included in the latest releases of the 3rd Generation Partnership Project (3GPP) standard, namely fifth-generation (5G) New Radio (NR) [2]. Moreover, Terahertz (THz) frequencies are being investigated as enablers for sixth-generation (6G) networks as well [3].

However, transmissions in the mmWave and THz bands are subject to challenging propagation conditions, mainly due to severe path loss and susceptibility to blockages [4]. To mitigate these limitations, the research community has explored solutions to improve network coverage, for example using integrated access and backhaul (IAB) nodes, as also approved by the 3GPP as part of 5G new radio (NR) specifications for Rel-16 [5]. In particular, IAB allows base stations, or next generation Node Bases (gNBs) in 5G NR parlance, to establish wireless (rather than traditional fiber-like) backhaul links, possibly through multiple hops, to a donor, thus reducing deployment costs [6]. Still, IAB involves complex signal processing and saturation of the available resources and may be costly and energy-consuming for network operators.

In light of this, intelligent reflecting surfaces (IRSs) are being investigated as solutions to overcome the harsh propagation conditions shown by mmWave and THz bands in a cost- and energy-efficient manner [7]. IRSs are meta-surfaces, whose radiating elements can *passively* tune the phase shift of impinging signals to favorably alter an electromagnetic field towards an intended destination. They can be configured to beamform the reflected signal virtually in any direction, hence acting as a relay to improve the signal quality without an active (power-consuming) amplification [8].

### A. Prior Work

Despite the substantial research hype, most recent studies on IRSs rely on strong assumptions that do not match real-world deployments. Specifically, a significant body of literature is based on the assumption that IRSs establish an ideal (i.e., fiber-like) control channel with the base station [9]–[12]. Instead, actual deployments are expected to feature a wireless, i.e., error-prone, IRS control, possibly implemented with low-cost technologies [13], [14]. This introduces constraints on the IRS reconfiguration period, which needs to be synchronized with the base station to beamform the signal towards the user equipment (UE) served during the specific time slot [7], a similar research problem to scheduling in cellular networks.

In this perspective, IRS-assisted downlink scheduling solutions have been widely studied in different domains, each

Alberto Rech, Matteo Pagin, Leonardo Badia, Stefano Tomasin, Marco Giordani, and Michele Zorzi are with the Department of Information Engineering, University of Padova, Italy. Jonathan Gambini is with the Milan Research Center, HUAWEI, Italy. Alberto Rech is also with Paris Research Center, HUAWEI, France. Corresponding author: A. Rech (email: alberto.rech@huawei.com).

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 marzo 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU Project code MUR: PE_0000001 Concession Decree No. 1549 of 11/10/2022 adopted by the Italian Ministry of University and Research, CUP C93C22005250001 Project title RESearch and innovation on future Telecommunications systems and networks, to make Italy more smART (RESTART). Part of this paper has been presented at the IEEE Wireless Communications and Networking Conference (WCNC), [1].

with its own theoretical constraints. For example, in orthogonal frequency-division multiple access (OFDMA) user scheduling, all the users scheduled in a given time slot must be served using the same reflection coefficients, due to the lack of frequency selective beamforming capabilities at the IRS. In this context, dynamic optimization schemes, wherein the IRS configurations are adjusted at each time slot, have been studied in [15], [16]. The authors of [17] consider a two-user downlink transmission problem in an IRS-assisted scenario over fading channels, and compare the results of different basic orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) schemes. It is found that, while NOMA is the best solution, by exploiting IRS reconfiguration in each slot of the fading block, time division multiple access (TDMA) outperforms frequency division multiple access (FDMA), and its performance is similar to that of NOMA. A hybrid TDMA-NOMA approach, instead, was investigated in an uplink scenario in [18], [19], in the context of a wireless-powered network, where users are grouped based on their channel gains. Then, UEs within the same group transmit in a non-orthogonal fashion, while different groups are assigned to different time slots. Moreover, a user scheduling algorithm based on graph neural networks, able to jointly optimize the IRS configuration and the gNB beamforming in downlink, was recently presented in [20]. Similarly, the authors of [21]–[23] evaluated the performance of several non-orthogonal downlink scheduling methods, such as rate-splitting multiple access (RSMA). Finally, IRSs with energy harvesting capabilities are considered in [24]. In this work, the authors propose a trade-off between the system sum capacity and the IRS energetic self-sustainability, with the goal of achieving coverage flexibility and low deployment costs.

Still, most of the literature poses little to no reconfiguration constraints for the IRS. However, early IRS control circuitry prototypes, which have low power consumption (i.e., a few hundreds of mW), have a non-negligible phase-shifts reconfiguration time [25], [26], thus posing additional constraints in the system design. For example, the prototypes in [27] and [28] have a reconfiguration time of a few tens of ms, even though architectures based on field programmable gate array (FPGA) such as in [29] promise to achieve much lower configuration times, i.e., in the order of tens of microseconds. Still, the overhead (in terms of time) increases as the number of IRS elements increases, as investigated in [30]–[32]. In any case, a constraint on the number of reconfigurations (and relative period) is desirable to ensure system synchronization and minimize the IRS downtime during reconfiguration. In this regard, it is of interest to (*i*) investigate the level of performance degradation experienced by IRS-assisted systems when considering practical constraints, including limitations in the number of reconfigurations, and (*ii*) design algorithms that can mitigate these constraints. The limitation on the number of IRS reconfigurations in a given time frame has been initially studied in [25], where the authors evaluate the capacity of both OMA and NOMA schemes of a 2-user IRS-assisted single input single output (SISO) system under Rayleigh fading conditions. Still, additional research efforts is required to fully characterize the impact of IRS reconfigurations constraints on the network.

### *B. Contributions*

In this paper, we propose a TDMA scheduling policy for downlink cellular transmissions based on clustering algorithms, to maximize the sum capacity in IRS-assisted network deployments with practical constraints. Our main contributions are summarized as follows:

- We account for practical IRS limitations by considering a fixed maximum number of reconfigurations of IRS reflecting elements within a time frame, thus setting a simple constraint on the overhead entailed by the control of the IRS.
- We formalize an optimization problem to determine the optimal IRS configurations to maximize the sum capacity while satisfying the reconfiguration per frame constraint. Then, we convert the sum capacity problem into a clustering problem. The latter determines sets of UEs that can be served with the same (possibly suboptimal) IRS configuration while minimizing the related capacity loss.
- We design, as an alternative to typical clustering algorithms based on distance measures, a new class of algorithms which we denote as *capacity-based clustering*. These algorithms adjust the cluster configuration taking into account the sum capacity and the user fairness. Specifically, we propose three clustering algorithms: capacity-weighted clustering (CWC), which favors users experiencing the best channel conditions, one-shot capacity-based clustering (OSCBC), which represents a low-complexity alternative to the former, and inverse capacity-weighted clustering (ICWC), which promotes fairness among the cluster UEs.
- We compare via simulation the performance of distance- and capacity-based clustering in different IRS-assisted scenarios. Extensive numerical results show that scheduling based on clustering can reduce by up to 50% the number of IRS reconfigurations, thus promoting communication efficiency at the expense of a slightly lower sum capacity.

With respect to [1], we introduce new capacity-based clustering strategies to improve fairness and provide more extensive numerical results to demonstrate the scalability of the proposed solutions as a function of the density of UEs and the IRS size. Moreover, we evaluate the performance of the proposed scheduling strategies considering realistic IRS network constraints, including the quantization of phase shifts, and for different channel propagation conditions. In this sense, we provide additional results in terms of the computational complexity of the proposed distance- and capacity-based clustering algorithms, as well as in terms of fairness. Finally, we remark that due to the novelty of the considered scenario, to the best of our knowledge the effectiveness of our solution cannot be directly compared with any works in the literature. Indeed, the most similar works [16], [17], [25], [30] exhibit substantial differences in the considered contexts. More specifically:

- In comparison to [16], which assumes the IRS configurations to be fixed, our work focuses on optimizing user

scheduling in conjunction with IRS configurations in a dynamic manner.

- [17], [25] analyze a basic scenario where a single-antenna gNB serves single-antenna UEs, while we consider multiple antennas at both the base station and UEs, and an OFDMA multiuser access scheme.
- While [25] characterizes the capacity region for $K > 1$ UEs, the computational complexity of the proposed scheme limits its applicability to $K = 2$ UEs. In contrast, our solution supports more realistic scenarios, where $K \gg 1$.
- The authors of [30] use the position estimate of a moving UE to minimize the IRS reconfiguration overhead while guaranteeing a minimum signal-to-noise-ratio (SNR) in the single-UE context. In our work, on the other hand, we consider a multi-user scenario and reduce the number of IRS configurations by clustering UEs in the channel state information (CSI) domain.

These fundamental differences in system setup and assumptions prevent a meaningful simulation-based comparison.

### C. Organization and Notation

The rest of the paper is organized as follows. In Section II, we introduce the system model. In Section III, we present the sum capacity optimization problem. In Section IV, we describe the scheduling framework, while in Sections V and VI we present distance-based and capacity-based clustering algorithms, respectively. In Section VII, we show numerical results and compare the different scheduling and clustering solutions. Finally, Section VIII draws the main conclusions.

Scalars are denoted by italic letters; vectors and matrices by boldface lowercase and uppercase letters, respectively; sets are denoted by calligraphic uppercase letters. $\mathrm{diag}(\boldsymbol{a})$ indicates a square diagonal matrix with the elements of $\boldsymbol{a}$ on the principal diagonal, and $\mathrm{vec}(\boldsymbol{A})$ denotes the vectorization operator, staking the columns of matrix $\boldsymbol{A}$ into a column vector. $\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\dagger}$ denote the transpose and the conjugate transpose of matrix $\boldsymbol{A}$, respectively. $[\boldsymbol{A}]_{k\ell}$ denotes the scalar value in the $k$-th row and $\ell$-th column of matrix $\boldsymbol{A}$, while $[\boldsymbol{a}]_k$ denotes the $k$-th element of vector $\boldsymbol{a}$. The imaginary unit is denoted as $j = \sqrt{-1}$, and $\angle a$ denotes the phase of $a \in \mathbb{C}$. The operator $\diamond$ denotes the Khatri-Rao product. Finally, $\mathbb{E}[\cdot]$ denotes statistical expectation.

## II. System Model

We consider downlink data transmissions for the multi-user multiple input multiple output (MIMO) communication system shown in Fig. 1, wherein the transmission from the gNB to the $K$ UEs is assisted by an IRS. The gNB and the UEs are equipped with $N_{\mathrm{g}}$ and $N_{\mathrm{U}}$ antennas, respectively. We assume that the direct link between the gNB and the UEs is unavailable due to blockage. As a consequence, the gNB transmits signals to the UEs by exploiting the virtual link offered by the IRS. In this context, the IRS configuration is managed by the gNB through the IRS controller, by exploiting a dedicated link between the gNB and the IRS, thus with no additional communication overhead in the gNB-UE link. Time is divided into frames of $K$ slots, and each UE is served exactly once in a frame in a TDMA fashion, which ensures there is no co-channel interference as UEs are separated in the time domain. In our scenario we expect that the gNB-IRS channel has rank one, i.e., a single dominant path, which effectively prevents multi-stream transmissions and spatial multiplexing. However, when higher-rank channels are available, either multi-stream transmission to each UE or spatial multiplexing can be considered. For the former case, the proposed solution applies straightforwardly. Moreover, our approach can to be suitably modified to accommodate for the latter scenario. A detailed investigation of this point is left for future work. We assume that UEs are either static or moving slowly, which is the most typical application scenario for IRS-assisted networks. Under such conditions, the channel coherence time is in the order of 10 ms [33, Fig. 5]. Considering that perfect CSI of all UEs is acquired at the gNB at the beginning of each frame (a realistic assumption that does not affect the proposed scheduling framework for IRS communication), it is reasonable to conclude that the channel remains constant throughout the whole time frame. Here, we assume that the CSI is available for any IRS configuration.

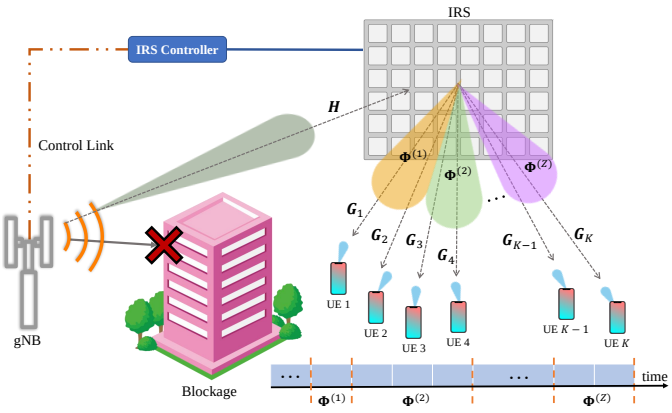


Fig. 1. Downlink TDMA scheduling for multi-user IRS-aided systems.

### A. IRS Model

Each of the $N_{\mathrm{I}}$ elements of the IRS acts independently as an omnidirectional antenna unit that reflects the impinging electromagnetic field by introducing a tunable phase shift on the baseband-equivalent signal. We denote as $\phi_n = e^{j\theta_n}$ the reflection coefficient of the $n$-th IRS element, where $\theta_n \in \mathcal{P}_\theta$ is the induced phase shift, and $\mathcal{P}_\theta$ is the set of possible phase shifts. Recent works argue that continuous phase shifts are hardly implementable in practice [34]. Therefore, we consider both continuous and quantized phase shifts. While in the former case the set of phase shifts is $\mathcal{P}_\theta = [-\pi, \pi)$, in the latter we have $\mathcal{P}_\theta = \left\{0, \frac{2\pi}{2^b}, \ldots, \frac{2\pi(2^b-1)}{2^b}\right\}$ where $b > 0$ is the number of bits employed to quantize the phase shifts.

We denote with $\boldsymbol{H} \in \mathbb{C}^{N_{\mathrm{I}} \times N_{\mathrm{g}}}$ the channel matrix between the IRS and the gNB, and with $\boldsymbol{G}_k \in \mathbb{C}^{N_{\mathrm{U}} \times N_{\mathrm{I}}}$ the channel matrix of the link between the IRS and UE $k$, respectively. We

consider single-stream transmissions,[1] with $\boldsymbol{w}_k \in \mathbb{C}^{N_\mathrm{g}\times 1}$ and $\boldsymbol{v}_k \in \mathbb{C}^{N_\mathrm{U}\times 1}$ defined as the beamforming vectors at the gNB and UE $k$, respectively. Let $x_k$ be the single-stream signal transmitted by the gNB to UE $k$; the received post-processing signal can be expressed as

$$z_k = \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k \boldsymbol{\Phi} \boldsymbol{H} \boldsymbol{w}_k x_k + \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{n}_k, \tag{1}$$

where $\boldsymbol{n}_k \in \mathbb{C}^{N_\mathrm{U}\times 1}$ represents the circularly symmetric complex Gaussian noise vector with entries having zero mean and variance $\sigma_n^2$, while $\boldsymbol{\Phi} \in \mathbb{C}^{N_\mathrm{I}\times N_\mathrm{I}}$ is the IRS configuration, i.e., a diagonal matrix defined as $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_{N_\mathrm{I}})$. Note that different, and specific, IRS configurations can be adopted for different UEs. Accordingly, in the rest of the paper we let $\boldsymbol{\Phi}_k$ be the IRS configuration adopted when UE $k$ is served.

The SNR at UE $k$ under IRS configuration $\boldsymbol{\Phi}_k$ is

$$\Gamma_k(\boldsymbol{\Phi}_k) = \frac{|\boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H} \boldsymbol{w}_k|^2 \sigma_x^2}{|\boldsymbol{v}_k|^2 \sigma_n^2}, \tag{2}$$

where $\sigma_x^2$ is the power of the transmitted signal. To maximize the SNR of a given UE, a specific IRS configuration should be adopted, tailored to the UE position in the cell and the channel conditions. However, the goal of this paper is to limit the number of IRS reconfigurations to comply with realistic overhead constraints, as well as to improve the communication efficiency, and algorithms seeking to comply with these requirements will be presented in Section IV.

## III. Sum Capacity Optimization Problem

We impose a constraint on the number of IRS reconfigurations per time frame, with the goal of either limiting the reconfiguration,[2] or accounting for practical limitations that might arise in realistic deployments. On the downside, achieving this objective usually leads to SNR degradation as suboptimal IRS configurations might be adopted for some UEs. To mitigate this effect, we formulate a constrained optimization problem on the average cell sum capacity. Specifically, we assume the following conditions:

1. at most $Z$ IRS reconfigurations can occur per time frame;
2. the gNB serves $K$ UEs by partitioning them into $Z$ disjoint subsets $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, $Z \leq K$;
3. for each UE in $\mathcal{U}_z$, the same IRS configuration $\boldsymbol{\Phi}^{(z)}$ is used, i.e., $\boldsymbol{\Phi}_k = \boldsymbol{\Phi}^{(z)}, \forall\, k \in \mathcal{U}_z, \forall\, 1 \leq z \leq Z$.

Then, the achievable rate of UE $k \in \mathcal{U}_z$ is

$$R_k(\boldsymbol{\Phi}^{(z)}) = \log_2\left(1 + \Gamma_k\big(\boldsymbol{\Phi}^{(z)}\big)\right), \tag{3}$$

where $\Gamma_k(\boldsymbol{\Phi}^{(z)})$ is the SNR experienced by the $k$-th UE while configuration $\boldsymbol{\Phi}^{(z)}$ is adopted at the IRS, i.e., the configuration shared by all UEs belonging to subset $\mathcal{U}_z$.

[1]The assumption of single-stream transmissions is justified by the rank of the cascade channel matrix, which is likely equal to one. This conclusion comes from the considerations reported in [35]–[37], and has been verified numerically for the considered setup.

[2]We remark that the gNB typically communicates a (possibly new) IRS configuration in each transmission time interval (TTI). The reconfiguration constraint introduced in the proposed IRS scheduling framework is able to reduce this overhead by a factor $Z/K \leq 1$.

Let $\mathcal{I} = \{\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \ldots, \boldsymbol{\Phi}^{(Z)}\}$ be the set of IRS configurations corresponding to subsets $\mathcal{U}_1, \ldots, \mathcal{U}_Z$. The system sum capacity within a time frame is defined as

$$C(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}) = B \sum_{z=1}^{Z} \sum_{k\in\mathcal{U}_z} R_k\big(\boldsymbol{\Phi}^{(z)}\big), \tag{4}$$

where $B$ is the transmission bandwidth. The optimization problem is then formulated as

$$\max_{\mathcal{U}_1,\ldots,\mathcal{U}_Z,\mathcal{I}} \; C(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}), \tag{5a}$$

$$\text{s.t.} \quad \angle\big[\boldsymbol{\Phi}^{(z)}\big]_{n,n} \in \mathcal{P}_\theta, \quad \forall\, n, z. \tag{5b}$$

Problem (5) determines the optimal grouping strategy for the UEs subsets $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, and assigns the best IRS configuration accordingly. Therefore, (5) is both continuous (i.e., the optimization of the IRS configuration) and combinatorial (i.e., the grouping of the UEs), and can be thus classified as a mixed integer nonlinear programming (MINLP) problem. Moreover, the following theorem holds.

**Theorem 1.** *The sum capacity maximization problem* (5) *is NP-complete.*

*Proof:* First, we observe that the problem falls within the general NP class. This is because if (5) is solved to find $\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}$, both the sum capacity and the phase-shift constraints (5b) could be verified in polynomial time. To prove that the problem in NP-complete, we set $\mathcal{I}$, and consider the simplified problem

$$\max_{\mathcal{U}_1,\ldots,\mathcal{U}_Z} \quad C(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}). \tag{6}$$

This problem can be viewed as a multi-knapsack problem with different clusters $\mathcal{U}_1, \ldots, \mathcal{U}_Z$ as knapsacks, and the goal is to maximize the total system capacity. This is known to be NP-hard, as it is a generalization of the classic knapsack problem. The original sum capacity maximization problem (5), where we consider the additional degrees of freedom of the IRS configurations, remains NP-hard, thus making the problem NP-complete. ■

Given the inherent problem complexity, we adopt heuristic clustering algorithms to obtain approximated, though close-to-optimal, solutions, as described in Section IV.

## IV. Heuristic Sum Capacity Maximization

In this section, we provide heuristic solutions to (5). First, we present two clustering-based approaches to identify and group UEs with a similar optimal IRS configuration. Then, we solve the scheduling problem on the identified clusters with a TDMA approach [38]. We compute the UEs clusters by first estimating the optimal individual IRS configurations, denoted as $\boldsymbol{\Phi}_k^*$, $1 \leq k \leq K$, i.e., the IRS configurations leading to the maximum capacity for each UE $k$, as described in Section IV-A. These configurations would solve (5) for $Z = K$, as in this case all UEs are served in a TDMA fashion and with their optimal IRS configuration. The phase coefficients of the optimal IRS configuration matrices are then chosen as the initial points of a procedure leveraging clustering algorithms in the $N_\mathrm{I}$-dimensional space, as explained in Section IV-B.

### A. Optimal Individual IRS Configurations

In MIMO systems, both the gNB and the UEs adopt properly tuned beamformers to match the signal transmissions and receptions to the spatial direction providing the highest channel gain [7]. For the optimization of the IRS configuration of each individual UE, we adopt a procedure similar to that presented in [32], focusing on single-stream transmissions and, without loss of generality, on UE $k$.

For a given IRS configuration, the optimal beamforming vectors $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ coincide with the singular vectors corresponding to the highest singular value of the wireless channel matrix. In particular, we calculate the singular value decomposition (SVD) of the overall cascade channel matrix

$$\boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^{\dagger}, \tag{7}$$

where the right and left singular vectors of $\boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H}$ are the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$, and the corresponding singular values are the diagonal entries of $\boldsymbol{\Sigma}$.

In our formulation, the IRS configuration $\boldsymbol{\Phi}_k$ is one of the optimization variables. Indeed, given $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$, we can solve

$$\boldsymbol{\Phi}_k^* = \underset{\boldsymbol{\Phi}_k}{\operatorname{argmax}} \; R_k(\boldsymbol{\Phi}_k), \tag{8a}$$

$$\text{s.t.} \quad \angle[\boldsymbol{\Phi}_k]_{n,n} \in \mathcal{P}_\theta, \quad \forall n, \tag{8b}$$

where $R_k$ is the achievable rate of user $k$, $1 \leq k \leq K$, according to (3).

The derivation of the optimal IRS configuration of user $k$ requires the alignment of the channel phase coefficients. According to [39], the cascade channel can be expressed as

$$\begin{aligned} \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H} \boldsymbol{w}_k &= \operatorname{vec}\left(\boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H} \boldsymbol{w}_k\right) \\ &= \left(\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{H}^{\mathrm{T}} \diamond \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k\right) \operatorname{diag}(\boldsymbol{\Phi}_k) \end{aligned} \tag{9}$$

where $\diamond$ denotes the Khatri-Rao product operator and $\operatorname{diag}(\boldsymbol{\Phi}_k)$ return the column vector with all the elements in the diagonal of $\boldsymbol{\Phi}_k$. Then, it is sufficient to observe that the SNR is maximized when the phase shifts introduced by the IRS are aligned with the phase shifts accumulated along the various paths, i.e.,

$$\theta_{k,n} = -(\angle \left[\left(\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{H}^{\mathrm{T}} \diamond \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k\right)\right]_n), \quad \forall n. \tag{10}$$

Note that, in general, we need to know the estimated phase shift of each component resulting from the Khatri-Rao product in (10), rather than the exact phase coefficients of $\boldsymbol{H}$ and $\boldsymbol{G}_k$. Moreover, as pointed out in [39], for structured channel models adopted at mmWaves, where multipath scattering is sparse and propagation is often dominated by strong specular components, the estimation of the separated channel matrices $\boldsymbol{H}$ and $\boldsymbol{G}_k$ can be simply accommodated by optimizing a limited number of parameters.

Taking into account the possible quantization, the optimal phase shifts are given by

$$\angle[\boldsymbol{\Phi}_k^*]_{n,n} \leftarrow \underset{\psi \in \mathcal{P}_\theta}{\operatorname{argmin}} \left(\angle e^{j(\theta_{k,n} - \psi)}\right), \quad \forall n. \tag{11}$$

To overcome the interdependence between optimal IRS configurations and beamforming vectors, we propose an iterative alternate optimization approach. We first estimate the optimal beamforming vectors for a given IRS configuration using (7).

**Algorithm 1** Iterative Alternate IRS Optimization

**Require:** $\boldsymbol{G}_k, \boldsymbol{H}$
**Ensure:** $\boldsymbol{\Phi}_k^*$
$t \leftarrow 0$
$\boldsymbol{v}_k, \boldsymbol{w}_k \leftarrow \mathbf{1}$
**repeat**
  $\theta_{k,n} \leftarrow -(\angle \left[\left(\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{H}^{\mathrm{T}} \diamond \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k\right)\right]_n), \quad \forall n$
  $\angle[\boldsymbol{\Phi}_{k,t}]_{n,n} \leftarrow \operatorname{argmin}_{\psi \in \mathcal{P}_\theta}(\angle e^{j(\theta_{k,n} - \psi)})$
  $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}^{\dagger} \leftarrow$ SVD of $\boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{G}_k \boldsymbol{\Phi}_k \boldsymbol{H} \boldsymbol{w}_k$
  $\boldsymbol{v}_k \leftarrow$ column of $\boldsymbol{V}$ corresponding to the largest singular value
  $\boldsymbol{w}_k \leftarrow$ column of $\boldsymbol{U}$ corresponding to the largest singular value
  $t \leftarrow t + 1$
**until** $|R_k(\boldsymbol{\Phi}_{k,t}) - R_k(\boldsymbol{\Phi}_{k,t-1})| < \nu$
$\boldsymbol{\Phi}_k^* \leftarrow \boldsymbol{\Phi}_{k,t}$

Then, we plug the derived beamformers into (8a), and obtain the corresponding optimal IRS configuration. We repeat this two-step procedure until convergence, which, for practical purposes, is assumed to be reached when the difference between the achievable rates $R_k$, $\forall k$, in two consecutive iterations is lower than a tolerance $\nu > 0$. This procedure is summarized in Algorithm 1, where $t$ is the iteration index. The number of iterations grows with the numbers of antennas and IRS phase shifters. However, from preliminary simulations, and based on the set of parameters we considered (see Section VII), convergence is typically reached in less than 10 iterations.

### B. Clustering-based TDMA Scheduling

For an approximated but close-to-optimal solution to (5), we resort to a clustering-based approach. Our proposed clustering algorithms estimate both the subsets of UEs $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, and the relative set of IRS configurations $\mathcal{I}$. We operate on the *phase vector space*, i.e., the points to be clustered are identified by the IRS phase shifts vector $[\angle\phi_1, \ldots, \angle\phi_{N_{\mathrm{I}}}]^{\mathrm{T}} = [\theta_1, \ldots, \theta_{N_{\mathrm{I}}}]^{\mathrm{T}}$, which maps each IRS configuration $\boldsymbol{\Phi}$ to a point in $[-\pi, \pi)^{N_{\mathrm{I}}}$. In case of quantized phase shifts, the phase vector space is a lattice in the continuous space $[-\pi, \pi)^{N_{\mathrm{I}}}$.

The general clustering-based procedure works as follows:

- *Step 1:* find $\boldsymbol{\Phi}_k^*$, $\forall k$, i.e., the optimal individual IRS configurations for each UE as in Section IV-A;
- *Step 2:* build UE subsets $\mathcal{U}_z$, $z = 1, \ldots, Z$, by using a clustering algorithm, according to Sections V and VI;
- *Step 3:* assign $\boldsymbol{\Phi}^{(z)}$ to all UEs $\in \mathcal{U}_z$.

The core idea of this procedure is to use clustering algorithms to group UEs, and assign the respective IRS configurations, which are mapped to the *centroid* of the cluster. In the case of quantized phase shifts, once the clustering procedure is performed, clusters may share the same centroid and be merged. Therefore, $Z$ represents the *maximum number of clusters*, not the effective number. Moreover, we remark that the procedure above does not rely on the assumption of perfect CSI, as the grouping strategy (Step 2) and the individual optimization step (Step 1) are performed independently. Nevertheless, in the case of imperfect CSI, the estimated

individual optimal configurations may differ from the actual optimal configurations, leading to a suboptimal grouping.

In the following, we propose different techniques to build the clusters based on either a distance metric (Section V) or the achievable rate (Section VI).

## V. Distance-Based Clustering Algorithms

The class of distance-based clustering contains methods that group data points based on their similarity or dissimilarity according to a distance metric. This approach has several advantages, including the efficiency in handling large datasets, and the flexibility to adapt to many different scenarios of interest. However, distance-based clustering can be sensitive to the choice of the distance metric (which depends on the nature of the data and the clustering problem), and the initialization values. Moreover, in our specific case, it does not take into account the achievable rate, which is not directly related to the distance among the points in the phase vector space.

Since the scalar field is the range $[-\pi, \pi)$, the adopted distance has to take into account the circularity of data. However, the convergence to a local minimum for most of the clustering algorithms is guaranteed only if the points to be clustered belong to a Euclidean space. For distance-based algorithms, we thus define the bijective mapping function $f : \mathcal{P}_{\theta}^{N_{\mathrm{I}}} \rightarrow \mathbb{R}^{2N_{\mathrm{I}}}$ as

$$\begin{aligned} f(\boldsymbol{\theta}) &= f([\theta_1, \ldots, \theta_{N_{\mathrm{I}}}]) \\ &= [\cos(\theta_1), \sin(\theta_1), \ldots, \cos(\theta_{N_{\mathrm{I}}}), \sin(\theta_{N_{\mathrm{I}}})], \end{aligned} \tag{12}$$

and then define the pairwise distance between two generic IRS configurations $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as

$$\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = ||f(\boldsymbol{\alpha}) - f(\boldsymbol{\beta})||, \tag{13}$$

i.e., the Euclidean distance between the mapping on the unit $N_{\mathrm{I}}$-sphere of their respective phase vectors. In the following, with a slight abuse of notation, $f(\boldsymbol{\Phi})$ maps the phases of the complex entries in the diagonal of $\boldsymbol{\Phi}$ as in (12), and $\delta(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)$ denotes the pairwise distance between the phases of the elements in the diagonal of matrices $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$. The sum of squared distances is defined as

$$J(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}) = \sum_{z=1}^{Z} \sum_{k \in \mathcal{U}_z} \delta\left(\boldsymbol{\Phi}_k^*, \boldsymbol{\Phi}^{(z)}\right)^2, \tag{14}$$

and the distance-based clustering schemes are used to solve the following problem:

$$\min_{\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}} \quad J(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I}), \qquad \text{s.t.} \quad \text{(5b)}. \tag{15}$$

We consider and compare some of the most popular distance-based clustering algorithms, namely, K-means, agglomerative hierarchical clustering, and K-medoids.

*K-means.* K-means (KM) clustering [40] aims at finding $Z$ disjoint clusters minimizing the within-cluster squared Euclidean distances. Here, we consider the generalized Lloyd algorithm [41], which randomly selects $Z$ points in the space of phase vectors as the initial centroids. In our setup, to ensure optimal performance when $Z = K$, we force the algorithm initialization to a random selection among the phase vectors of the optimal individual IRS configurations derived in Section IV-A. Then, in the *assignment step* KM assigns each data point to the closest centroid, according to the specified distance metric. In the subsequent *update step*, the set of centroids is re-computed as the average of the data points that belong to each cluster. These steps are repeated until either convergence or a maximum number of iterations $I_{\max}^{\mathrm{KM}}$ is reached.

*Agglomerative hierarchical clustering.* The agglomerative hierarchical clustering (HC) [42] partitions a set of data points into disjoint clusters by iteratively merging points into clusters until a target number of partitions is met. In our setup, clusters are initialized as the optimal phase vectors, which thus act as the respective centroids. Then, the average distance between all pairs of data points in any pair of clusters is evaluated. The closest pair of clusters are merged into a new single cluster, whose centroid is computed as the mean of its data points. The procedure is repeated until the number of clusters is $Z$.

*K-medoids.* K-medoids (KMed) [43] is a clustering technique similar to KM, but instead of the mean of the data points within each cluster, it uses the medoid, i.e., the data point that is closest to the center of the cluster. In our setup, we consider the partition around medoids (PAM) method [44], which starts by randomly selecting $Z$ medoids among the optimal phase vectors and assigns each point to the cluster with the closest medoid. In each iteration, the algorithm evaluates potential *swaps* of medoids with non-medoids. A swap is accepted only if it results in a lower value of the sum of the squared distances to all other data points within the same cluster. The algorithm continues until the medoids no longer change.

**Theorem 2.** *The proposed distance-based clustering techniques converge to a local minimum of* (14)*.*

*Proof:* The proof directly derives from the well-known results of clustering with Euclidean distance. The exact proofs for each of the considered algorithms under distance metric (13) are reported in the Appendix. ■

## VI. Capacity-Based Clustering Algorithms

The distance-based clustering techniques presented in Section V do not directly take into account the actual capacity achievable by the UEs, which is a crucial factor for the sum capacity maximization (5). Thus, in the following we propose original capacity-based clustering algorithms that go beyond the state of the art, namely CWC (Section VI-A), OSCBC (Section VI-B), and ICWC (Section VI-C).

### A. Capacity-Weighted Clustering (CWC)

Similarly to distance-based clustering, also CWC proceeds iteratively. However, the stopping condition is based on the variation of the sum capacity of each cluster, rather than on the distance between the centroids. In this approach, the clustering algorithm itself weighs the UEs based on their achievable capacity, so that the parameters of the resulting clusters are closer to those preferred by the UEs with higher rates, thus promoting the maximization of the sum capacity.

Let $\mathbf{\Phi}_i^{(z)}$ be the IRS configuration of cluster $\mathcal{U}_{z,i}$ at iteration $i$. UEs are initially sorted in decreasing order of achievable rate. The algorithm then selects the $Z$ UEs providing the highest achievable capacity based on the expression in (3) with their optimal IRS configurations. Without loss of generality, we let $z = 1, \ldots, Z$ be the index of those UEs, and set $\mathbf{\Phi}_1^{(z)} = \mathbf{\Phi}_z^*$, $\forall\ 1 \leq z \leq Z$, as the centroids of the initial clusters $\mathcal{U}_{1,0}, \ldots, \mathcal{U}_{Z,0}$. In the following, for simplicity, we denote with $z_{k,i}$ the cluster such that $k \in \mathcal{U}_{z,i}$. Each UE $k > Z$ is assigned to the cluster whose centroid provides the lowest rate difference with respect to its ideal configuration. Let $R_k(\mathbf{\Phi}_k^*)$ be the maximum achievable rate of UE $k$, obtained from the solution of problem (1). UE $k$ is assigned to cluster

$$z_{k,i} = \underset{z}{\operatorname{argmin}}[R_k(\mathbf{\Phi}_k^*) - R_k(\mathbf{\Phi}_i^{(z)})], \tag{16}$$

where $R_k(\mathbf{\Phi}_i^{(z)})$ is the rate achieved by UE $k$ adopting the IRS configuration of cluster $z$ at iteration $i$. Note that, despite being always non-negative, the rate difference in (16) cannot be considered a distance metric as, in general, it does not satisfy the triangle inequality. However, we prove that, as the distance from the optimal configuration increases, the corresponding rate decreases, thus supporting the use of the rate difference as a clustering criterion.

**Theorem 3.** *Given the optimal IRS configuration $\mathbf{\Phi}_k^*$, the rate $R_k(\mathbf{\Phi})$ is monotonically decreasing with respect to the magnitude of any phase shifts error $\epsilon$.*

*Proof:* Let $\epsilon \in [-\pi, \pi]$ be an arbitrary error phase shift, and consider the configuration $\mathbf{\Phi}_k^\epsilon = \mathbf{\Phi}_k^* \boldsymbol{E}$, where $\boldsymbol{E} = \operatorname{diag}(e^{j\epsilon}, 1, \ldots, 1)$, i.e., the suboptimal configuration where only the first IRS element is affected by the error $\epsilon$. Assuming, without loss of generality, that $N_\mathrm{g} = N_k = 1$ and $\sigma_x = \sigma_n = 1$, the rate $R_k(\mathbf{\Phi}_k^\epsilon)$ is proportional to $\Gamma_k(\mathbf{\Phi}_k^\epsilon)$ when using configuration $\mathbf{\Phi}_k^\epsilon$. The SNR $\Gamma_k(\mathbf{\Phi}_k^\epsilon)$ can be written as

$$\Gamma_k(\mathbf{\Phi}_k^\epsilon) = |\boldsymbol{g}_k \mathbf{\Phi}_k^* \boldsymbol{E} \boldsymbol{h}|^2 \tag{17}$$

$$= \left|[\boldsymbol{g}_k]_1 [\mathbf{\Phi}_k^* \boldsymbol{E}]_{1,1} [\boldsymbol{h}]_1 + A\right|^2, \tag{18}$$

where $A = \sum_{n=2}^{N_\mathrm{I}} [\boldsymbol{g}_k]_n [\mathbf{\Phi}_k^*]_{n,n} [\boldsymbol{h}]_n$. Since $\mathbf{\Phi}_k^*$ is the optimal configuration, it satisfies (10). It follows that $A \in \mathbb{R}^+$, so (17) can be further manipulated into

$$\Gamma_k(\mathbf{\Phi}_k^\epsilon) = |[\boldsymbol{g}_k]_1||[\boldsymbol{h}]_1| e^{j\epsilon} + A|^2 \tag{19}$$

$$= A^2 + (|[\boldsymbol{g}_k]_1||[\boldsymbol{h}]_1|)^2 + 2A|[\boldsymbol{g}_k]_1||[\boldsymbol{h}]_1| \cos(\epsilon). \tag{20}$$

Finally, we evaluate the sign of the derivative of $\Gamma_k(\mathbf{\Phi}_k^\epsilon)$ with respect to the error $\epsilon$ as

$$\frac{\partial \Gamma_k(\mathbf{\Phi}_k^\epsilon)}{\partial \epsilon} = -2A|[\boldsymbol{g}_k]_1||[\boldsymbol{h}]_1| \sin(\epsilon), \tag{21}$$

and observe that $\Gamma_k(\mathbf{\Phi}_k^\epsilon)$, and therefore $R_k(\mathbf{\Phi}_k^\epsilon)$, is strictly decreasing for $0 < |\epsilon| \leq \pi$. ∎

After all the remaining UEs have been assigned to the corresponding clusters, the coordinates of the centroids are updated. At iteration $i+1$, the new IRS configuration (centroid) of cluster $\mathcal{U}_{z,i+1}$ is computed as the average of the data points in the cluster, weighted by their achievable rate, i.e.,

$$\mathbf{\Phi}_{i+1}^{(z)} = f^{-1}\left(\frac{\sum_{k \in \mathcal{U}_z} f(\mathbf{\Phi}_k^*) R_k(\mathbf{\Phi}_k^*)}{\sum_{k \in \mathcal{U}_z} R_k(\mathbf{\Phi}_k^*)}\right). \tag{22}$$

Also, in the case of phase shift quantization, an additional approximation step must be performed as

$$\angle[\mathbf{\Phi}_{i+1}^{(z)}]_{n,n} \leftarrow \underset{\psi \in \mathcal{P}_\theta}{\operatorname{argmin}} \left(\angle e^{j(\angle[\mathbf{\Phi}_{i+1}^{(z)}]_{n,n} - \psi)}\right), \quad \forall n. \tag{23}$$

This two-step procedure is repeated until convergence, which is reached when the rate difference between two consecutive iterations is lower than the sum capacity tolerance $\mu > 0$.

The rationale behind the algorithm is that, based on the initial centroid assignment, the UEs experiencing the best channel conditions, i.e., those dominating the system sum capacity, are initially served with their optimal (individual) IRS configurations. Even after the adjustment of the clusters, these UEs will always get the largest weight coefficient within the cluster. The remaining UEs, instead, will be penalized by the configuration constraints, but their impact on the sum capacity will be limited. The whole workflow of the CWC procedure is summarized in Algorithm 2.

**Algorithm 2** CWC Algorithm

**Require:** $Z$, $\boldsymbol{H}$, $\boldsymbol{G}_k$, $\forall k$
**Ensure:** $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, $\mathcal{I}$

```
Compute Φ*_k, ∀k with the procedure of Algorithm 1
Sort the UEs in decreasing order of R_k(Φ*_k)
Select the Z UEs providing the highest R_k(Φ*_k),
Set Φ^(z)_1 = Φ*_k, z = 1, ..., Z as the initial centroids.
repeat
    for each UE k do
        z_{k,i} ← argmin_z R_k(Φ*_k) − R_k(Φ^(z)_i)
    end for
    for each cluster z do
        Compute Φ^(z)_{i+1} as per (22), (23)
    end for
    i ← i + 1
until |Σ_{k∈U_z} R_k(Φ^(z)_i) − Σ_{k∈U_z} R_k(Φ^(z)_{i−1})| < μ
Assign Φ^(z) to all k ∈ U_z.
```

### *B. One-Shot Capacity-Based Clustering (OSCBC)*

The main drawback of CWC presented in Section VI-A is that it requires solving problem (16) at each iteration, relative to all the UEs in each cluster. Considering massive MIMO systems, the CWC procedure could become exceedingly complex, as it requires the SVD computation of extremely large matrices. Therefore, we propose another lower-complexity clustering algorithm, denoted as OSCBC.

As in CWC, also in OSCBC: (*i*) the UEs are sorted in decreasing order of achievable rate; (*ii*) the $Z$ IRS configurations of the $Z$ UEs experiencing the highest rates are chosen as initial centroids for the clusters; and (*iii*) the remaining UEs are assigned to the closest centroid in terms of circular distance, as per (13). Then, compared to CWC, instead of

TABLE I
COMPUTATIONAL COMPLEXITY OF DISTANCE-BASED VS. CAPACITY-BASED CLUSTERING.

| Clustering algorithm | Computational complexity |
|---|---|
| KM (Lloyd) | $O(IZKN_{\mathrm{I}})$ |
| KMed (PAM) | $O(Z^3K^2N_{\mathrm{I}})$ |
| HC | $O(K^3N_{\mathrm{I}})$ |
| CWC/ICWC | $O(IZKN_{\mathrm{g}}N_{\mathrm{I}}^2)$ |
| OSCBC | $O(Z(K-Z)N_{\mathrm{I}})$ |

recomputing the coordinates of the centroids at each iteration until convergence, the algorithm stops right after the initial association. Therefore, with OSCBC the computed centroids are the optimal configurations relative to the $Z$ UEs achieving the highest individual rate, which provides suboptimal (non-optimized) performance for the rest of the UEs in the clusters.

### C. Inverse Capacity-Weighted Clustering (ICWC)

The CWC algorithm is designed to optimize the capacity of the UEs experiencing the best channel conditions and is unfair to the other UEs in the system, which may use suboptimal IRS configurations. Therefore, we propose an additional variation of CWC, named ICWC, with the goal of achieving higher fairness among the UEs in the system. In ICWC, while the cluster association principle of (16) is preserved, the initial condition is reversed. Specifically: (*i*) UEs are sorted in increasing order of achievable rate; (*ii*) the initial configurations of the clusters $\boldsymbol{\Phi}_1^{(z)} = \boldsymbol{\Phi}_z^*$, $z = 1, \ldots, Z$, are based on the optimal configurations of the UEs with the worst channel conditions. The remaining $k > Z$ UEs are associated as in (16). Then, at iteration $i$, the IRS configuration is updated as

$$\boldsymbol{\Phi}_{i+1}^{(z)} = f^{-1}\left(\frac{\sum_{k\in\mathcal{U}_z}\boldsymbol{\Phi}_k^* R_k^{-1}(\boldsymbol{\Phi}_k^*)}{\sum_{k\in\mathcal{U}_z} R_k^{-1}(\boldsymbol{\Phi}_k^*)}\right), \tag{24}$$

and the discretization step (23) is performed (if needed). As in CWC, convergence is achieved if the rate difference between two consecutive iterations is lower than the tolerance $\mu$. While ICWC obtains lower sum capacity than CWC, it can provide significant improvements in terms of fairness, especially from the perspective of the UEs with the worst channel conditions.

### D. Computational Complexity

The computational complexity is evaluated as the number of iterations required for the clustering algorithms to: (*i*) obtain the optimal IRS configuration of each UE; (*ii*) partition the UEs into disjoint subsets, or clusters, based on distance or capacity metrics; and (*iii*) for each cluster, find the best IRS configuration to serve the corresponding UEs.

Specifically, at each iteration, the main source of complexity is the computation of the overall cascade channel matrix $\boldsymbol{G}_k\boldsymbol{\Phi}_k\boldsymbol{H}$, which has complexity $O\big(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}}\big)$. Additionally, in the case of quantized IRS phase shifts, after obtaining the optimal beamformers, the optimal phase shifts for the IRS are obtained through an exhaustive search over the set of possible phase shifts $\mathcal{P}_\theta$, yielding a complexity $O(2^bN_{\mathrm{I}})$.

Notice that different clustering algorithms, in general, require a different number of iterations $I$ to reach convergence, thus possibly introducing practical limitations. Moreover, the complexity introduced in each iteration depends on the clustering algorithm itself. In Table I and in the following text we characterize the computational complexity of each of the clustering algorithms presented in Sections V and VI.

*Distance-based clustering.* These algorithms do not require specific initialization. For KM, based on the Lloyd implementation in [41], each iteration involves calculating the distances between data points and centroids. As a result, the computational complexity is influenced by the number of iterations required for convergence, the number of data points, the number of clusters, and the dimensionality of data, resulting in an overall complexity $O(IZKN_{\mathrm{I}})$. KMed can be solved with the PAM algorithm [44], so the computational complexity is $O(Z^3K^2N_{\mathrm{I}})$ due to the pairwise distance computations between data points and medoids. Finally, the computational complexity of the agglomerate HC is primarily determined by the computation of pairwise distances among all data points, resulting in a total complexity $O(K^3N_{\mathrm{I}})$ [45].

*Capacity-based clustering.* The complexity of the OSCBC algorithm is dominated by the centroid assignment upon initialization, which has complexity $O\big(Z(K-Z)N_{\mathrm{I}}\big)$. Instead, for the CWC and ICWC algorithms, the complexity is $O(IZKN_{\mathrm{g}}N_{\mathrm{I}}^2)$, as demonstrated in the following theorem.

**Theorem 4.** *The time complexity of CWC and ICWC scales quadratically with $N_{\mathrm{I}}$ as $O(IZKN_{\mathrm{g}}N_{\mathrm{I}}^2)$.*

*Proof:* Capacity-based clustering requires an initialization stage where the algorithm selects the $Z$ UEs providing the highest (or lowest) $R_k\big(\boldsymbol{\Phi}_k^*\big)$, resulting in a complexity of $O(K\log K)$ due to the sorting of $K$ scalars. In the subsequent iterations:

1) Both CWC and ICWC compute the rate difference between each UE and the $Z$ centroids. The complexity of computing $R_k(\boldsymbol{\Phi}_i^{(z)})$ can be dominated either by the matrix multiplication in (2), or by the SVD for the single stream beamforming which require, respectively, $O\big(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}}\big)$ and $O\big(N_{\mathrm{g}}N_{\mathrm{U}}\min(N_{\mathrm{g}}, N_{\mathrm{U}})\big)$ operations for each UE and each centroid.
2) The computation of the centroids as per (22)-(24) requires $N_{\mathrm{I}} + 1$ scalar operations per UE, which has negligible complexity with respect to the rate computation.

In typical IRS-assisted systems, $N_{\mathrm{I}} \gg N_{\mathrm{g}} > N_{\mathrm{U}}$. Therefore, the complexity at each iteration is dominated by the channel matrix product, and the overall algorithm complexity is $O(IZKN_{\mathrm{g}}N_{\mathrm{I}}^2)$. ■

## VII. NUMERICAL RESULTS

After presenting our various simulation scenarios and evaluation metrics in Sections VII-A and VII-B, respectively, we assess in Section VII-C the scheduling performance of an IRS-assisted network with practical constraints.

TABLE II
SIMULATION PARAMETERS.

| Parameter | Value |
|---|---|
| Carrier frequency | 28 GHz |
| Total bandwidth ($B$) | 100 MHz |
| Noise power spectral density | $-174$ dBm/Hz |
| Number of UEs ($K$) | 100 |
| gNB antenna array ($N_{\mathrm{g}}$) | 8H×8V |
| gNB transmit power | 33 dBm |
| UE antenna array ($N_{\mathrm{U}}$) | 2H×1V |
| IRS elements ($N_{\mathrm{I}}$) | {10H×20V, 20H×40V, 40H×80V, 60H×120V} |
| Phase shift quant. bits ($b$) | {unquantized, 1-bit, 2-bits} |
| LoS probability ($p_{\mathrm{LoS}}$) | Eq. (26) |
| Individual rate opt. tolerance ($\nu$) | $10^{-6}$ [bit/s/Hz] |
| KM max. iterations ($I_{\max}^{\mathrm{KM}}$) | 50 |
| CWC/ICWC rate tolerance ($\mu$) | $10^{-3}$ [bit/s] |

### A. Simulation Parameters

Our simulation parameters are reported in Table II.

*Scenario.* All devices are assumed to lie on a 2D plane, and we consider an urban microcell (UMi) scenario, according to the 3GPP nomenclature [46], with the gNB placed at the center. According to the 3GPP specifications, the coverage area of the gNB is characterized by an average radius of 167 m and is assumed to lie in the positive $x$-axis region.

We assume that $K = 100$ UEs are randomly deployed according to a uniform distribution within the cell area, to be served in downlink by the gNB, assisted by an IRS at coordinates $(75, 100)$ m. The gNB is equipped with a uniform planar array (UPA) with 8H×8V antennas (i.e., $N_{\mathrm{g}} = 64$), and the UEs with uniform linear arrays (ULAs) of 2H×1V antennas (i.e., $N_{\mathrm{U}} = 2$). For the IRS, if not otherwise specified, we adopt a 40H×80V reflective panel ($N_{\mathrm{I}} = 3200$).

*Channel and Frame Structure.* The system operates at a carrier frequency of 28 GHz (that is in the lower part of the mmWave bands), the transmission power at the gNB is set to 33 dBm, the noise power spectral density at the receivers is $-174$ dBm/Hz, and the total system bandwidth is 100 MHz. We consider the fourth numerology of the NR frame structure [47], wherein each 10 ms frame is split into 160 slots. With this assumption, as already pointed out in Section II, channels can be considered constant over the entire frame duration. We consider the 3GPP TR 38.901 spatial channel model [46], which supports a wide range of frequencies, from 0.5 to 100 GHz (and including therefore our carrier frequency of 28 GHz), and can be integrated with realistic beamforming models. As such, channel matrices, and multipath fading, are computed based on the superposition of $N$ different clusters, each of which consists of $M$ rays that arrive (depart) to (from) the antenna arrays with specific angles and powers. Based on [46], and using the simplifications proposed in [48], the generic entry $[\boldsymbol{A}]_{pq}$ of the channel matrix can then be computed as:

$$
\begin{aligned}
[\boldsymbol{A}]_{pq} = \gamma \sum_{n=1}^{N} \sqrt{\frac{P_n}{M}} \sum_{m=1}^{M} & \overline{\mathbf{F}}_{rx}\left(\theta_{n,m}^{A}, \phi_{n,m}^{A}\right) \\
& \times \begin{bmatrix} e^{j\Phi_{n,m}^{\theta,\theta}} & \sqrt{K_{n,m}^{-1}} e^{j\Phi_{n,m}^{\theta,\phi}} \\ \sqrt{K_{n,m}^{-1}} e^{j\Phi_{n,m}^{\phi,\theta}} & e^{j\Phi_{n,m}^{\phi,\phi}} \end{bmatrix} \\
& \times \overline{\mathbf{F}}_{tx}\left(\theta_{n,m}^{D}, \phi_{n,m}^{D}\right) \\
& \times e^{j\overline{\mathbf{k}}_{rx,n,m}^{\mathrm{T}} \overline{\mathbf{d}}_{rx,p}} e^{j\overline{\mathbf{k}}_{tx,n,m}^{\mathrm{T}} \overline{\mathbf{d}}_{tx,q}},
\end{aligned}
\tag{25}
$$

where $\gamma$ is the large-scale fading coefficient (LSFC) of the considered link, which incorporates the path loss and shadowing terms. For a complete description of the remaining terms appearing in (25) we refer the interested reader to [48]. Specifically, while the gNB and the IRS can be assumed to operate in line-of-sight (LoS), the path loss between a generic UE $k$ and the IRS is modeled based on the following channel conditions:

- *non-line-of-sight (NLoS)*: UE $k$ is in NLoS with the IRS;
- *deterministic LoS (LoS)*: UE $k$ is in LoS with the IRS;
- *probabilistic LoS (LoS)*: the IRS-UE $k$ link is in LoS$\vee$NLoS with respective probabilities $p_k^{\mathrm{LoS}}(d_k) \vee 1 - p_k^{\mathrm{LoS}}(d_k)$, with

$$
p_k^{\mathrm{LoS}}(d_k) = \begin{cases} 1 & \text{if} \quad d_k \leq 18, \\ \frac{18}{d_k} + \left(1 - \frac{18}{d_k}\right) e^{-\frac{d_k}{36}} & \text{if} \quad d_k > 18, \end{cases}
\tag{26}
$$

where $d_k$ is the distance (in m) between the IRS and UE $k$. In the considered UMi scenario, and based on 3GPP specifications [46], the average LoS probability in (26) is 0.35.

For each wireless link, based on the presence of the LoS component, the path loss is then derived according to [46, Table 7.4.1-1], with shadowing standard deviation set to $\sigma_{\mathrm{SF}} = 0$. For the optimal individual IRS configuration (Section IV-A), we set $\nu = 10^{-6}$ [bit/s/Hz].

*Clustering algorithms.* In the following subsections, we present extensive simulation results to compare the performance of distance-based (KM, HC, KMed) vs. capacity-based (CWC, OSCBC, ICWC) clustering algorithms to perform scheduling in an IRS system with reconfiguration constraints. The KM clustering has been implemented with the Lloyd algorithm [41] with a maximum number of $I_{\max}^{\mathrm{KM}} = 50$ iterations. Instead, for both CWC and ICWC, we set $\mu = 10^{-3}$ [bit/s].

As an upper bound to the system performance, we also consider an "unclustered" scheduling, wherein we assume that all UEs are served with their optimal IRS configuration. This scheduling clearly violates the constraint on the maximum numbers of reconfiguration per frame, but can be regarded as the limit case when $Z = K$, i.e., all UEs belong to a cluster with cardinality one. As such, it is a suitable approach for benchmarking the performance of more practical schemes.

### B. Performance Metrics

The performance of the proposed clustering-based scheduling techniques is evaluated in terms of average sum capacity and fairness, as a function of the numbers of both clusters and UEs, under different channel conditions, IRS dimensions, and degrees of quantization for the phase shifts.

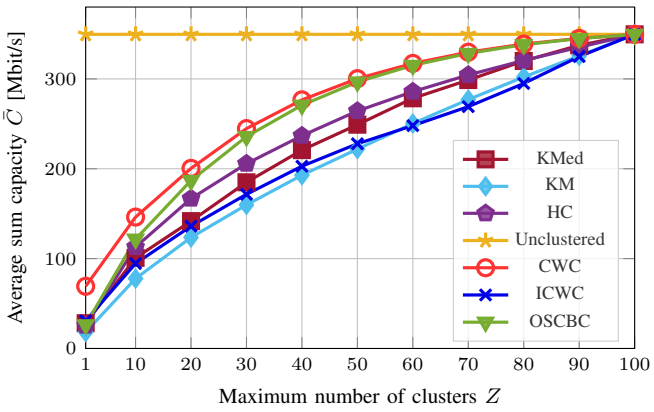


Fig. 2. Average sum capacity as a function of the maximum number of clusters $Z$, for an unquantized 40H×80V IRS, and considering a pLoS channel for the IRS-UEs links.

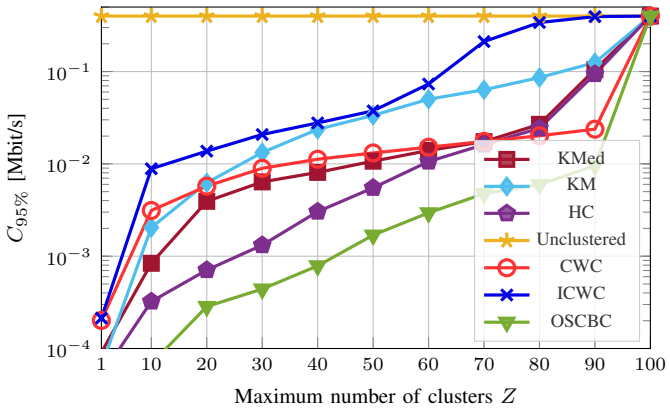


Fig. 3. 95% percentile of the user capacity as a function of the maximum number of clusters $Z$, for an unquantized 40H×80V IRS, and considering a pLoS channel for the IRS-UEs links.

*Average sum capacity.* It is derived from (4) as

$$\bar{C} = \frac{1}{K}\mathbb{E}\left[C(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I})\right], \tag{27}$$

where the expectation is computed across the different channel realizations. Moreover, as each UE is served in its specific slot, we average over the TDMA frame length, dividing the empirical expectation by the number of UEs (slots) $K$.

*Fairness.* We consider the 95% percentile of the achieved individual user capacity, computed as

$$C_{95\%} = \frac{B}{K}\inf\{x : \mathrm{CDF}(x) \geq 0.95\}, \tag{28}$$

where $\mathrm{CDF}(\cdot)$ is the empirical cumulative distribution function of $R_k(\mathbf{\Phi}^{(z)})$, $\forall k, z$. Notice that the 95% percentile of the user capacity is a practical and meaningful way to evaluate fairness, as it measures the performance of the majority of the UEs, excluding only the top 5%.

### C. Scheduling Performance

In this section, we compare the IRS scheduling performance considering distance-based vs. capacity-based clustering, and as a function of different channel conditions, reconfiguration constraints, and degrees of quantization of the phase shifts.

*Impact of the clustering algorithm.* First, Fig. 2 displays the average sum capacity $\bar{C}$ per slot as a function of the number of clusters $Z$, for unquantized IRS phase shifts, and considering a pLoS channel for the IRS-UEs links. It is evident that all the scheduling policies perform better whenever $Z$ increases, and converge to the "unclustered" policy when $Z = K$. In fact, increasing the number of clusters corresponds to a smaller intra-cluster average distance, which eventually becomes zero when $Z = K$. Among the considered clustering policies, CWC and OSCBC provide the highest sum capacity, as they are designed to maximize $\bar{C}$, and choose the IRS configurations of the UEs that achieve the highest rate. Instead, distance-based clustering achieves worse performance as it does not exploit the knowledge of the rate achievable with different IRS configurations when building the clusters. As expected, ICWC is designed to promote fairness, thus performs worse than both CWC and OSCBC in terms of sum capacity; still, it achieves similar performance as distance-based clustering. Finally, the gap between CWC and OSCBC is almost negligible: this implies that a single iteration in the clustering process is enough to achieve good sum capacity, while also promoting lower computational complexity as reported in Table I, which demonstrates the good scalability of the proposed techniques.

Fig. 3 compares the fairness performance of the different clustering algorithms, measured as the 95% percentile of the average sum capacity $C_{95\%}$, as a function of the maximum number of clusters $Z$ in pLoS conditions. Our results identify ICWC as the best clustering approach in terms of fairness, which comes at the cost of a lower sum capacity, as shown in Fig. 2. Therefore, there exists a trade-off between the achievable sum capacity and fairness. We also observe that OSCBC achieves very low fairness, as the UEs with worst channel conditions are forced to aggregate to the strongest UEs, thus via a suboptimal IRS configuration. On the other hand, we see that CWC is more than acceptable in terms of fairness, and achieves comparable performance than most of the distance-based clustering algorithms. Furthermore, $C_{95\%}$ increases as $Z$ increases, and eventually approaches the "unclustered" baseline for $Z = K$. This is due to the fact that the LoS probability in the pLoS scenario increases with the number of clusters, i.e., as the inter-cluster distance becomes smaller, which permits to experience better channel conditions, thus a higher capacity, even for the worst UEs. Finally, despite performing worse than their capacity-based counterparts, the distance-based methods are a viable alternative for constrained IRS control nodes thanks to their lower computational complexity. In such cases, HC is to be preferred for sum-capacity maximization, while KM is the best alternative to capacity-based algorithms when the 95% percentile of the average sum capacity represents the metric of interest.

*Impact of the channel.* From the above results, we concluded that distance-based clustering provides lower sum capacity and fairness compared to capacity-based scheduling, so the rest of our simulation campaign has been focused on the latter. Figs. 4 and 5 display the average sum capacity and the 95% percentile, respectively, for CWC, ICWC, and OSCBC

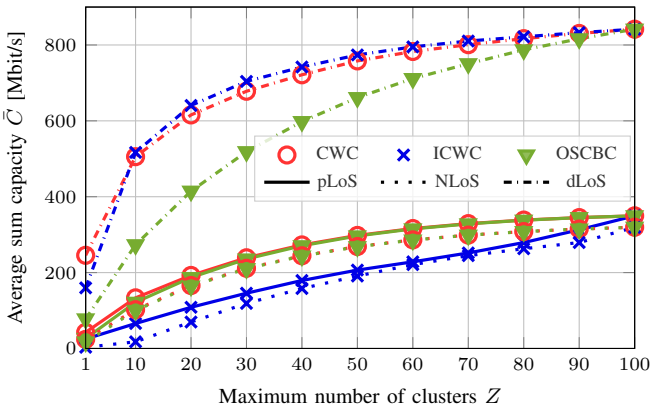


Fig. 4. Average sum capacity as a function of the maximum number of clusters $Z$, for $N_{\mathrm{I}} = 3200$, unquantized phase shifts, and for different channel conditions.

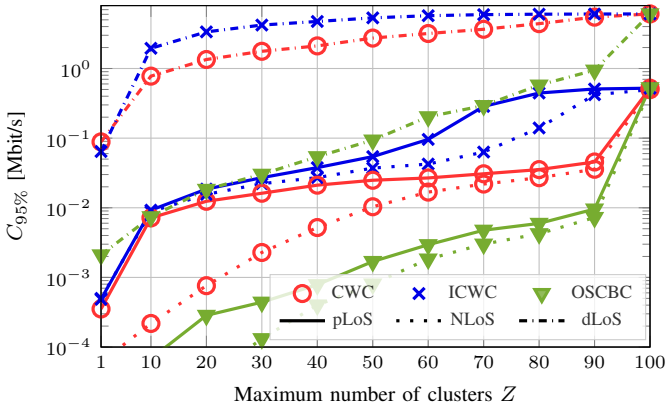


Fig. 5. 95% percentile of the user capacity as a function of the maximum number of clusters $Z$, for $N_{\mathrm{I}} = 3200$, unquantized phase shifts, and for different channel conditions.

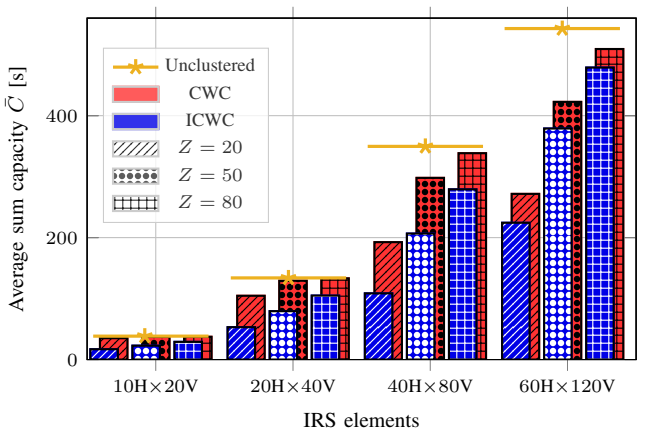


Fig. 6. Average sum capacity for CWC and ICWC as a function of the number of reflecting elements at the IRS, for unquantized phase shifts, and considering a pLoS channel for the IRS-UEs links.

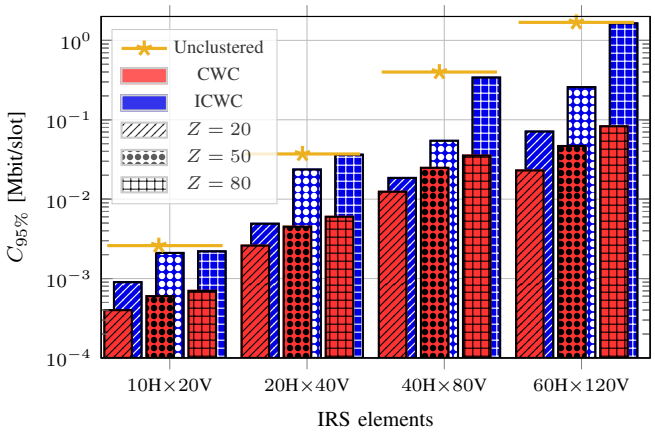


Fig. 7. 95% percentile of the user capacity for CWC and ICWC as a function of the number of reflecting elements at the IRS, for unquantized phase shifts, and considering a pLoS channel for the IRS-UEs links.

in different channel conditions. First, we observe that in the dLoS scenario, where UEs are in LoS with the IRS, the sum capacity is up to 2.6 (2.4) times higher than in the NLoS (pLoS) scenario for $Z = K$. This is mainly due to the fact that NLoS links experience (*i*) a higher path loss, and (*ii*) the lack of a dominant multipath component, thus of a clear steering direction for the IRS beam, which deteriorates the link quality. In particular, in the pLoS scenario the LoS probability decreases exponentially with the distance, therefore, the UEs that are far from the IRS typically operate in NLoS. For similar reasons, both CWC and ICWC in the dLoS scenario start to reach stability in terms of capacity with a relatively lower number of clusters than in the pLoS and NLoS scenarios.

As expected, OSCBC performs worse than its competitors, and the gap is even more significant in the dLoS scenario (around $-30\%$ in terms of sum capacity). The bad performance of OSCBC compared to CWC and ICWC is confirmed also in terms of fairness, as illustrated in Fig. 5 (see, in particular, the zoom for $50 \leq Z \leq 90$).

Finally, even though ICWC is not explicitly designed to maximize the sum capacity, it shows similar performance (if not even slightly better) as CWC in the dLoS scenario. The rationale behind this behavior is not clear and deserves more investigation. Most likely, it is related to the fact that, in the dLoS scenario, all UEs have similar channel conditions, which permits ICWC to choose, on average, a good IRS configuration even among the worst UEs in the clusters.

*Impact of the IRS configuration.* Figs. 6 and 7 show the impact of the number of IRS radiating elements on the system performance when considering the CWC and ICWC clustering algorithms. As expected, both fairness (measured in terms of the 95% percentile of the average sum capacity) and sum capacity increase as the IRS is larger and operates with more reflecting elements, regardless of the number of clusters. For example, we observe that CWC is able to approach the optimal sum capacity with as few as 20 clusters for small-sized IRS, i.e., with 10H×20V or 20H×40V arrays. The same trends are shown also in Fig. 7 in terms of fairness. Still, notice that $\bar{C}$ is below 100 Mbps, which is not compatible with the requirement of most 5G applications when the IRS is made of fewer than 200 elements, which justifies the use of larger IRS panels [12].

Nevertheless, we still observe that the number of reflecting elements has an impact on the number of clusters that are

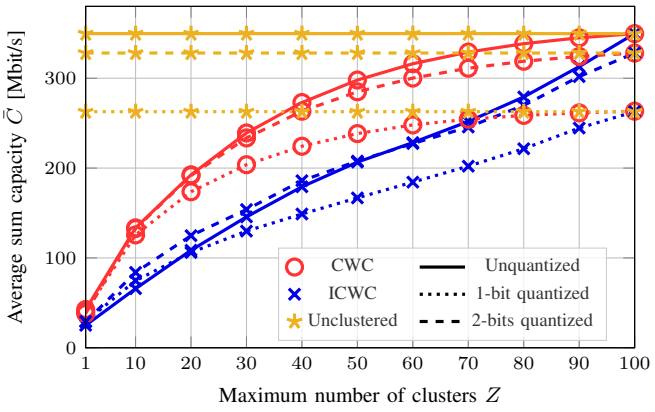


Fig. 8. Average sum capacity as a function of the maximum number of clusters $Z$, for $N_{\mathrm{I}} = 3200$ and for different degrees of quantization of the phase shifts, and considering a pLoS channel for the IRS-UEs links.

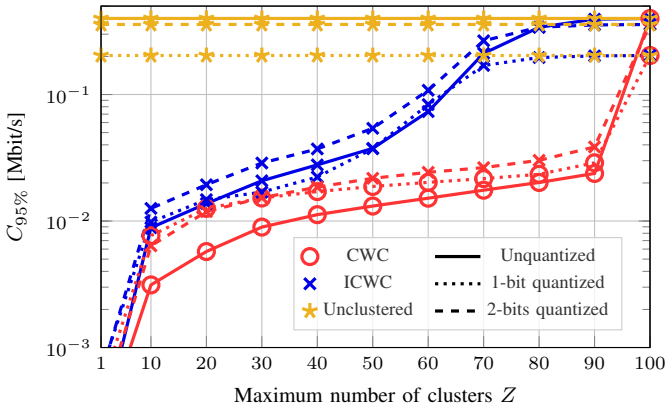


Fig. 9. 95% percentile of the user capacity as a function of the maximum number of clusters $Z$, for $N_{\mathrm{I}} = 3200$ and for different degrees of quantization of the phase shifts, and considering a pLoS channel for the IRS-UEs links.

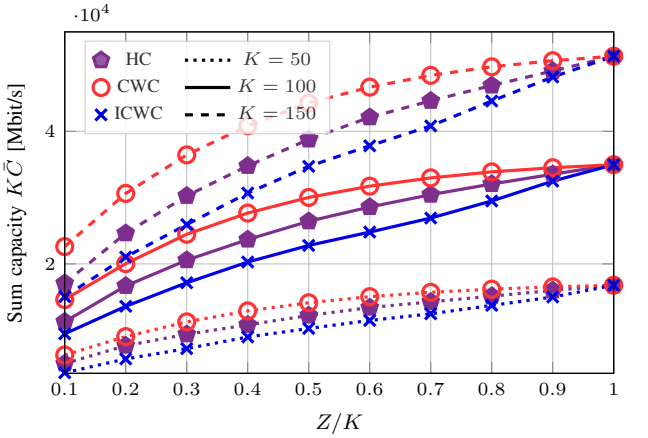


Fig. 10. Sum capacity as a function of the maximum number of clusters over the number of UEs $Z/K$, for different values of $K$, for an unquantized 40H×80V IRS, $K = \{50, 100, 150\}$, and considering a pLoS channel for the IRS-UEs links. For readability, the results are shown without averaging over the TDMA frame length.

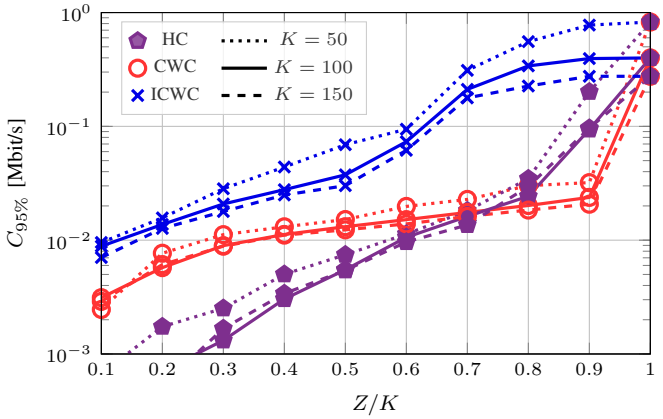


Fig. 11. 95% of the user capacity as a function of the maximum number of clusters over the number of UEs $Z/K$, for different values of $K$, for an unquantized 40H×80V IRS, $K = \{50, 100, 150\}$ UEs, and considering a pLoS channel for the IRS-UEs links.

needed to provide maximum performance. Indeed, the number of possible IRS configurations increases as we consider larger IRS antennas. In turn, this decreases the likelihood of UEs having the same (or similar) ideal configurations, and therefore, it increases the probability of being associated with increasingly suboptimal centroids if the number of clusters is small. However, if the number of phase shifters is large, the suboptimality is mitigated by the increasing number of reconfigurations. Typically, the IRS reconfiguration cost is proportional to the number of radiating elements, and therefore the specific reconfiguration cost may be different for different IRSs in general. A detailed quantitative analysis of this issue would need to go into the specifics of the various IRS architectures, which goes beyond the scope of the present paper, and will be considered in our future work.

*Impact of quantization.* Figs. 8 and 9 display the average sum capacity and the 95% percentile, respectively, as a function of the maximum number of clusters $Z$ for CWC and ICWC, and of the number of quantization bits $b$ of the phase shifts. Notice that energy and hardware constraints pose a limit to $b$ [49], which implies restricting the infinite set of possible IRS configurations to a finite set of cardinality $2^{bN_{\mathrm{I}}}$. Moreover, the quantization constraint affects the beamforming capabilities of the IRS [10], with negative implications for the resulting achievable sum capacity. In [1], results were obtained considering that the quantization was performed only at the end of the clustering procedure. Here, instead, we assume that the quantization of the phase shifts is taken into account from the initial optimization stage. The results reveal that the use of non-ideal phase shifters leads to a 30% degradation in the sum capacity when using $b = 1$ at the IRS, while the performance is close to the unquantized baseline if more quantization bits are used. Furthermore, it is shown that the gap between quantized and the unquantized performance increases with $Z$. As a result, 1-bit quantization is sufficient to guarantee a performance comparable to the unquantized case with a small number of clusters, while more quantization bits are needed to achieve higher capacity. In any case, we can conclude that our proposed capacity-based clustering algorithms are robust to phase-shift quantization.

*Scalability.* Finally, we prove the scalability performance of

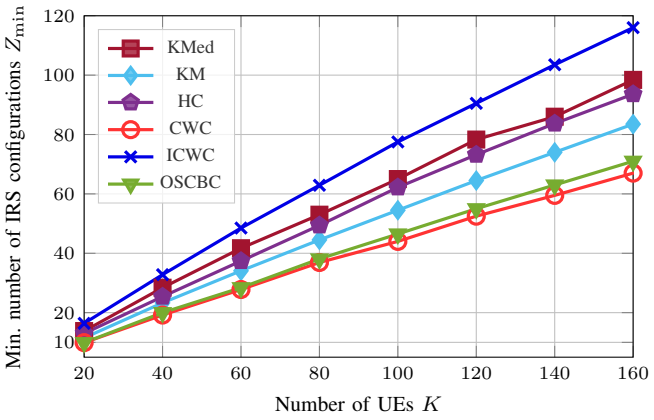


Fig. 12. Minimum number of IRS configurations (clusters) $Z_{\min}$ to achieve 80% of the maximum achievable sum capacity, for an unquantized 40H×80V IRS, and considering a pLoS channel for the IRS-UEs links.

the proposed clustering algorithms. To do so, we first show the performance of capacity-based clustering as a function of the number of UEs in Figs 10 and 11. In particular, we compare CWC and ICWC with HC as a function of the ratio $K/N$, for an unquantized 40H×80V IRS, $K = \{50, 100, 150\}$ UEs, and considering a pLoS channel for the IRS-UEs links. The results are in line with the plots in Figs 2 and 3, which demonstrates the scalability of the proposed clustering techniques for different numbers of UEs. Finally, Fig. 12 depicts the average minimum number of IRS configurations $Z_{\min}$ needed to achieve 80% of the maximum achievable sum capacity ("unclustered" baseline) as a function of the number of UEs $K$ in the system. Notably, we observe that CWC and OSCBC are confirmed to be the best algorithms to optimize the sum capacity, even for a limited number of IRS configurations. For example, both solutions achieve 80% of the maximum sum capacity with less than half the number of configurations than in the "unclustered" deployment. Moreover, we recognize the same trends as in the previous results. Specifically, capacity-based clustering outperforms distance-based clustering and requires fewer IRS reconfigurations to maximize the sum capacity (up to $-37\%$ considering CWC vs. KMed). Furthermore, the gap increases as the number of UEs increases.

## VIII. Conclusions

We considered a MIMO cellular network, in which a gNB serving multiple UEs is assisted by an IRS acting as a relay. Notably, we considered practical constraints on the IRS reconfiguration period. We studied a TDMA scheduling for downlink transmissions and formulated an optimization problem to maximize the average sum capacity, subject to a fixed number of IRS reconfigurations per time frame. We first discussed an iterative algorithm to obtain the optimal IRS configuration of each UE. Then, we proposed clustering-based scheduling algorithms, which group UEs with similar (ideal) IRS configurations based on either a distance metric or the achievable capacity, to mitigate the performance degradation due to the constraint in the number of possible reconfigurations. Different clustering algorithms were numerically evaluated in terms of computational complexity, sum capacity, and fairness under different channel conditions, as a function of the size of the IRS size and the number of users, and with or without quantization of phase shifts. The results showed that capacity-based clustering outperforms distance-based clustering, and can achieve up to 85% of the sum capacity obtained in an ideal deployment (with no reconfiguration constraints), reducing by 50% the number of IRS reconfigurations.

## Appendix
## Proof of Theorem 2

*Proof for KM (LLoyd):* In the assignment step, each UE $k$ is assigned to the cluster $z$ that minimizes the squared distance $\delta\left(\mathbf{\Phi}_k^*, \mathbf{\Phi}^{(z)}\right)^2$. This guarantees that the total sum $J(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I})$ does not increase. Then, in the update step, $\mathbf{\Phi}^{(z)}$ is recalculated as the average $\mathbf{\Phi}_k^*$ within each cluster, so as to minimize the intra-cluster sum of squared distances $\sum_{k\in\mathcal{U}_z} \delta\left(\mathbf{\Phi}_k^*, \mathbf{\Phi}^{(z)}\right)^2$, for all $z$. Therefore, the conditions of [50, Lemma 5] are satisfied, which ensures the convergence to a local minimum. Notice that [50] does not specify the number of iterations needed to reach convergence, which could be large in the case of highly dimensional spaces. Therefore, in practice, we limit the maximum number of iterations to $I_{\max}^{\mathrm{KM}}$. ■

*Proof for Agglomerative HC:* At each step, clusters are merged to minimize the increase of the total intra-cluster sum of squared distances. This is equivalent to choosing the merged cluster that results in the smallest increase of $J(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I})$. Then, as in the update step of KM, the average of the data points minimizes $\sum_{k\in\mathcal{U}_z} \delta\left(\mathbf{\Phi}_k^*, \mathbf{\Phi}^{(z)}\right)^2$, for all $z$. Once the number of clusters $Z$ reaches the desired value, convergence to a local minimum is reached. ■

*Proof for KMed (PAM):* Since, at each iteration, a swap is performed only when it leads to a lower value of the intra-cluster sum of squares, $J(\mathcal{U}_1, \ldots, \mathcal{U}_Z, \mathcal{I})$ does not increase over different iterations. Given the finite number of data points and possible configurations, the algorithm is guaranteed to converge to a configuration where no swap can further decrease the objective function, thus reaching a local minimum. ■

## References


[1] A. Rech *et al.*, "Downlink TDMA scheduling for IRS-aided communications with block-static constraints," in *Proc. IEEE WCNC Wkshps*, 2023.

[2] 3GPP, "5G; NR; Base Station (BS) radio transmission and reception," TS 38.104 (Rel. 16), 2022.

[3] F. Tariq, M. R. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[4] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[5] 3GPP, "5G; NR; integrated access and backhaul (IAB) radio transmission and reception," TS 38.174 (Rel. 17), 2022.

[6] M. Polese *et al.*, "Integrated access and backhaul in 5G mmWave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[7] R. Flamini *et al.*, "Towards a heterogeneous smart electromagnetic environment for millimeter-wave communications: An industrial viewpoint," *IEEE Trans. Antennas Propag.*, vol. 70, no. 10, pp. 8898–8910, Oct. 2022.

[8] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

[9] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[10] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[11] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[12] M. Pagin *et al.*, "End-to-end simulation of 5G networks assisted by IRS and AF relays," in *Proc. IEEE MedComNet*, 2022.

[13] 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, Jan. 2022.

[14] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Realizing wireless communication through software-defined hypersurface environments," in *Proc. IEEE WoWMoM*, 2018.

[15] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[16] J. Lee, J. Choi, and J. Kang, "Harmony search-based optimization for multi-RISs MU-MISO OFDMA systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 257–261, Feb. 2023.

[17] Y. Guo, Z. Qin, Y. Liu, and N. Al-Dhahir, "Intelligent reflecting surface aided multiple access over fading channels," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 2015–2027, Mar. 2021.

[18] D. Zhang, Q. Wu, M. Cui, G. Zhang, and D. Niyato, "Throughput maximization for IRS-assisted wireless powered hybrid NOMA and TDMA," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1944–1948, Sep. 2021.

[19] H. Al-Obiedollah, H. A. B. Salameh, K. Cumanan, Z. Ding, and O. A. Dobre, "Self-sustainable multi-IRS-aided wireless powered hybrid TDMA-NOMA system," *IEEE Access*, vol. 11, pp. 57 428–57 436, Jun. 2023.

[20] Z. Zhang, T. Jiang, and W. Yu, "Learning based user scheduling in reconfigurable intelligent surface assisted multiuser downlink," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1026–1039, May 2022.

[21] A. Bansal, K. Singh, B. Clerckx, C.-P. Li, and M.-S. Alouini, "Rate-splitting multiple access for intelligent reflecting surface aided multi-user communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9217–9229, Sep. 2021.

[22] H. Fu, S. Feng, and D. W. Kwan Ng, "Resource allocation design for IRS-aided downlink MU-MISO RSMA systems," in *Proc. IEEE ICC Wkshps*, 2021.

[23] B. Zhuo *et al.*, "Partial non-orthogonal multiple access: A new perspective for RIS-aided downlink," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2395–2399, Nov. 2022.

[24] S. Hu, Z. Wei, Y. Cai, C. Liu, D. W. K. Ng, and J. Yuan, "Robust and secure sum-rate maximization for multiuser MISO downlink systems with self-sustainable IRS," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 7032–7049, Oct. 2021.

[25] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

[26] ETSI, "Reconfigurable Intelligent Surfaces (RIS); Communication Models, Channel Models, Channel Estimation and Evaluation Methodology," ETSI GR RIS 003 V1.1.1, 2023.

[27] M. Rossanese, P. Mursia, A. Garcia-Saavedra, V. Sciancalepore, A. Asadi, and X. Costa-Perez, "Designing, building, and characterizing RF switch-based Reconfigurable Intelligent Surfaces," in *Proc. ACM WiNTECH*, 2022.

[28] G. C. Alexandropoulos *et al.*, "RIS-enabled smart wireless environments: Deployment scenarios, network architecture, bandwidth and area of influence," *EURASIP J. on Wirel. Commun. and Netw.*, vol. 2023, no. 1, p. 103, Oct. 2023.

[29] L. Yezhen, R. Yongli, Y. Fan, X. Shenheng, and Z. Jiannian, "A novel 28 GHz phased array antenna for 5G mobile communications," *ZTE Communications*, vol. 18, no. 3, pp. 20–25, 2020.

[30] V. Jamali, G. C. Alexandropoulos, R. Schober, and H. V. Poor, "Low-to-zero-overhead IRS reconfiguration: Decoupling illumination and channel estimation," *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 932–936, Apr. 2022.

[31] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.

[32] X. Qian, M. Di Renzo, V. Sciancalepore, and X. Costa-Pérez, "Joint optimization of reconfigurable intelligent surfaces and dynamic metasurface antennas for massive MIMO communications," in *Proc. IEEE SAM workshop*, 2022.

[33] Y. Lu, M. Koivisto, J. Talvitie, M. Valkama, and E. S. Lohan, "Positioning-aided 3D beamforming for enhanced communications in mmWave mobile networks," *IEEE Access*, vol. 8, pp. 55 513–55 525, 2020.

[34] X. Tan, Z. Sun, D. Koutsonikolas, and J. M. Jornet, "Enabling indoor mobile millimeter-wave networks based on smart reflect-arrays," in *Proc. IEEE INFOCOM*, 2018.

[35] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[36] J. Rains, A. Tukmanov, Q. Abbasi, and M. Imran, "RIS-enhanced MIMO channels in urban environments: Experimental insights," in *Proc. EuCAP*, 2024.

[37] T. S. Rappaport, S. Sun, and M. Shafi, "Investigation and comparison of 3GPP and NYUSIM channel models for 5G wireless communications," in *Proc. IEEE VTC-Fall*, 2017.

[38] L. Anchora, L. Badia, E. Karipidis, and M. Zorzi, "Capacity gains due to orthogonal spectrum sharing in multi-operator LTE cellular networks," in *Proc. IEEE ISWCS*, 2012.

[39] A. L. Swindlehurst, G. Zhou, R. Liu, C. Pan, and M. Li, "Channel estimation with reconfigurable intelligent surfaces—a general framework," *Proc. IEEE*, vol. 110, no. 9, pp. 1312–1338, Sep. 2022.

[40] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 321–352.

[41] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[42] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Rev.: Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[43] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," in *Rep. Fac. Math. Inf.*, vol. 87, no. 3, 1987.

[44] M. Van der Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Stat. Computat. Simul.*, vol. 73, no. 8, pp. 575–584, 2003.

[45] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, pp. 165–193, Jun. 2015.

[46] 3GPP, "5G; Study on channel model for frequencies from 0.5 to 100 GHz," TS 38.901 (Rel. 16), 2020.

[47] ——, "5G; NR; Physical channels and modulation," TS 38.211 (Rel. 16), 2020.

[48] T. Zugno, M. Polese, N. Patriciello, B. Bojović, S. Lagen, and M. Zorzi, "Implementation of a spatial channel model for ns-3," in *Proc. ACM WNS3*, 2020.

[49] M. Rivera, M. Chegini, W. Jaafar, S. Alfattani, and H. Yanikomeroglu, "Optimization of quantized phase shifts for reconfigurable smart surfaces assisted communications," in *Proc. IEEE CCNC*, 2022.

[50] M. Sabin and R. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inf. Theory*, vol. 32, no. 2, pp. 148–155, Feb. 1986.


**Alberto Rech** (*Member, IEEE*) received the B.Sc., M.Sc., and Ph.D. degrees in information engineering from the University of Padova, Italy, in 2018, 2020, and 2024, respectively. From September 2019 to August 2020, he was a double degree student at the National Taiwan University, Taipei, Taiwan. From October 2022 to April 2023 he was a Research Intern at the Advanced Wireless Technology Lab, Huawei, Paris, France, where he now serves as a Research Engineer. His research interests concern the physical layer of radio access networks, including signal processing, algorithm design, and machine learning approaches for multi-user wireless networks. He was the recipient of the Best Paper Award at IEEE WCNC Workshops 2023 and ICUFN 2023.

**Matteo Pagin** (*Student Member, IEEE*) received the B.Sc. and M.Sc. degrees in telecommunication engineering from the University of Padova in 2018 and 2020, respectively, where he is currently pursuing his Ph.D. degree. From October 2020 to September 2021, he was a Postgraduate Researcher with the Department of Information Engineering. He is with Keysight Technologies, Aalborg Research Center, Denmark. He has collaborated with several institutions, including Northeastern University, CTTC, NYU, TU Darmstadt, Orange, Viasat and Huawei. His research focuses on the design and evaluation of protocols for next-generation cellular networks (5G and beyond). He was awarded the Best Paper Award at IEEE MedComNet 2020 and IEEE WCNC Workshops 2023.

**Leonardo Badia** (*Senior Member, IEEE*) received the Ph.D. in information engineering from the University of Ferrara, Italy, in 2004. From 2002 to 2003, he was at the RST Labs (currently, Wireless@KTH), Royal Institute of Technology, Sweden. In 2006, he joined the IMT Institute for Advanced Studies, Lucca, Italy. In 2011, he moved to the University of Padua, Italy, where he is currently Associate Professor. His research interests include mathematical analysis of wireless networks, cross-layer optimization, and applications of game theory to telecommunication problems. He published more than 250 scientific papers and is an active referee of scientific journals and TPC member for conferences in the broad areas of communications and networking.

**Stefano Tomasin** (*Senior Member, IEEE*) received the Ph.D. degree from the University of Padova, Italy, in 2003. During his studies he did internships with IBM Research (Switzerland) and Philips Research (Netherlands). He joined the University of Padova where he has been Assistant Professor (2005-2015), Associate Professor (2016-2022), and Full Professor (since 2022). He was visiting faculty at Qualcomm, San Diego (CA) in 2004, the Polytechnic University in Brooklyn (NY) in 2007 and the Mathematical and Algorithmic Sciences Laboratory of Huawei in Paris (France) in 2015. His current research interests include physical layer security, security of global navigation satellite systems, signal processing for wireless communications, synchronization, and scheduling of communication resources. He is a senior member of IEEE since 2011 (member since 1999) and a member of EURASIP since 2011. He is or has been an Editor of the IEEE Transactions on Vehicular Technologies (2011-2016), of the IEEE Transactions on Signal Processing (2017-2020), of the EURASIP Journal of Wireless Communications and Networking (since 2011) and of the IEEE Transactions on Information Forensics and Security (since 2020). He serves also as a Deputy Editor-in-Chief of the IEEE Transactions on Information Forensics and Security since January 2023.

**Marco Giordani** *(Member, IEEE)* is an Associate Professor at the Department of Information Engineering (DEI) of the University of Padova. During his Ph.D., he visited New York University (NYU), USA, and TOYOTA Infotechnology Center, Mountain View, CA, USA. He co-authored 70+ published articles in the area of wireless networks, three of which have received Best Paper Awards. He is a recipient of several awards, including the 2018 IEEE Daniel E. Noble Fellowship Award from the IEEE Vehicular Technology Society, the 2021 IEEE ComSoc Outstanding Young Researcher Award for EMEA, and the Francesco Carassa Prize from the Italian Telecommunications and Information Theory Group (GTTI). Marco serves as Editor for the IEEE Transactions of Wireless Communications. He is the Director of the PhD Summer School of Information Engineering (SSIE). His research interests include design and validation of protocols for next-generation (5G/6G) wireless networks.

**Jonathan Gambini** received the M.Sc. degree (summa cum laude) and the Ph.D. degree in Telecommunications Engineering from Politecnico di Milano (Italy) in 2007 and 2010, respectively. In 2007 he was Visiting Researcher at CWCSPR, New Jersey Institute of Technology (NJ, USA). In 2011 he was Research Associate at Dipartimento di Elettronica e Informazione of Politecnico di Milano. Since 2012 he has been with Huawei Technologies, where he now holds the position of Research Expert. His research interests concern algorithmic and information-theoretic aspects of wireless communications, machine learning and statistical signal processing. He is author and co-author of several articles on peer-reviewed journals and conferences, and co-inventor of many industrial patents.

**Michele Zorzi** (*Fellow, IEEE*) received the Laurea and Ph.D. degrees in electrical engineering from the University of Padova, Padua, Italy, in 1990 and 1994, respectively. From 1992 to 1993, he was on leave with the University of California at San Diego (UCSD), USA. In 1993, he joined the Faculty of Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy. After spending three years with the Center for Wireless Communications at UCSD, in 1998 he joined the School of Engineering, University of Ferrara, Italy, where he became a Professor in 2000. Since November 2003, he has been with the Department of Information Engineering, University of Padova. His current research interests include performance evaluation in mobile communications systems, wireless sensor networks, the Internet of Things, cognitive communications and networking, millimeter-wave and terahertz communications, vehicular networks, non-terrestrial networks, and underwater communications and networks. He has served as a Member-at-Large for the Board of Governors of the IEEE Communications Society from 2009 to 2011 and from 2021 to 2023. He received several awards from the IEEE Communications Society, including the Best Tutorial Paper Award in 2008 and 2019, the Education Award in 2016, the Stephen O. Rice Best Paper Award in 2018, and the Joseph LoCicero Award for Exemplary Service to Publications in 2020. He was the Editor-in-Chief of IEEE Wireless Communications Magazine from 2003 to 2005, IEEE TRANSACTIONS ON COMMUNICATIONS from 2008 to 2011, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING from 2014 to 2018. He was the Director of Education from 2014 to 2015 and the Director of Journals from 2020 to 2021.