Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8887445, 13 pages
https://doi.org/10.1155/2021/8887445

Research Article

Two-Stage Precoding Based on Overlapping User Grouping Approach in IoT-Oriented 5G MU-MIMO Systems

Djordje B. Lukic[1,2] **Goran B. Markovic**[1] **and Dejan D. Drajic**[1,3]

[1]*School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11120 Belgrade, Serbia*
[2]*Aspire Technology Unlimited, Vladimira Popovica 6, Unit B10, 11070 Belgrade, Serbia*
[3]*Innovation Centre of School of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11120 Belgrade, Serbia*

Correspondence should be addressed to Djordje B. Lukic; djordje.lukic@aspiretechnology.com

Received 5 June 2020; Revised 10 November 2020; Accepted 17 December 2020; Published 8 January 2021

Academic Editor: Hongzhi Guo

Downlink transmission techniques for multiuser (MU) multiple-input multiple-output (MIMO) systems have been comprehensively studied during the last two decades. The well-known low complexity linear precoding schemes are currently deployed in long-term evolution (LTE) networks. However, these schemes exhibit serious shortcomings in scenarios when users' channels are strongly correlated. The nonlinear precoding schemes show better performance, but their complexity is prohibitively high for a real-time implementation. Two-stage precoding schemes, proposed in the standardization process for 5G new radio (5G NR), combine these two approaches and present a reasonable trade-off between computational complexity and performance degradation. Before applying the precoding procedure, users should be properly allocated into beamforming subgroups. Yet, the optimal solution for user selection problem requires an exhaustive search which is infeasible in practical scenarios. Suboptimal user grouping approaches have been mostly focused on capacity maximization through greedy user selection. Recently, overlapping user grouping concept was introduced. It ensures that each user is scheduled in at least one beamforming subgroup. To the best of our knowledge, the existing two-stage precoding schemes proposed in literature have not considered overlapping user grouping strategy that solves user selection, ordering, and coverage problem simultaneously. In this paper, we present a two-stage precoding technique for MU-MIMO based on the overlapping user grouping approach and assess its computational complexity and performance in IoT-oriented 5G environment. The proposed solution deploys two-stage precoding in which linear zero forcing (ZF) precoding suppresses interference between the beamforming subgroups and nonlinear Tomlinson-Harashima precoding (THP) mitigates interuser interference within subgroups. The overlapping user grouping approach enables additional capacity improvement, while ZF-THP precoding attains balance between the capacity gains and suffered computational complexity. The proposed algorithm achieves up to 45% higher MU-MIMO system capacity with lower complexity order in comparison with two-stage precoding schemes based on legacy user grouping strategies.

1. Introduction

Cellular Internet of things (IoT) has been recognized as a key enabler for digital transformation and automation of almost all industries. Before 5G New Radio (5G NR), cellular networks have been mainly designed and implemented for human-type communications. Hence, the connectivity needs of industry 4.0 can be addressed only with the implementation of massive machine type communication (mMTC) 5G NR use cases. Based on current predictions, around 5 billion cellular IoT connections are expected by 2025 [1]. Multiuser (MU) multiple-input multiple-output (MIMO) and its evolution, massive MIMO (mMIMO), have been identified as one of the most promising technologies to address the massive capacity demands in 5G networks and beyond. A combination of spatial multiplexing and transmit beamforming technique enables simultaneous transmission of independent data streams using the same radio resources and thus achieves higher throughput and spectral efficiency in MU-MIMO systems [2].

The performance of MU-MIMO system design is largely dependent on deployed user grouping method [3]. Actually,

the use of improper user grouping strategy can allocate users with highly correlated channels into the same beamforming subgroup and thus significantly reduce system capacity. In this paper, we consider practical cellular IoT scenario when the number of users K is larger than the number of transmit antennas N, i.e., $K > N$, which requires selection of G user subsets K_g. In general, in order to find the optimal subset of $K_g \leq N$ users, the complete search space of size $\sum_{k=1}^{N} \binom{K}{k}$ is required, which is prohibitively complex when the number of users becomes large [4]. Several suboptimal user grouping methods have been proposed with the aim to reduce complexity. K-means clustering is a widely used strategy for grouping of K users into the specified number of clusters, such that each user belongs to the cluster with the nearest mean [5]. However, the constraints on the cluster size cannot be imposed with K-means clustering. This presents an important disadvantage in MU-MIMO scenario since the number of users within the cluster should be less or equal to the number of base station antennas. Also, the number of clusters needs to be specified in advance and the final results are proven to be sensitive to initial parameters, while method often terminates at a local optimum [6]. In [7], Dimic and Sidiropoulos presented a suboptimal greedy user selection algorithm which iteratively selects user with the biggest contribution to the cumulative system capacity until further increase cannot be achieved. When this approach is applied, only those users characterized with the favorable channel conditions are selected, while users with less favorable channel conditions are dropped. Such behaviour can present a problem in the case of the fixed IoT endpoint devices with relatively low throughput requirements since these may be dropped in many consecutive iterations and thus not be served for a long period of time. In [8], Tian et al. introduced the concept of overlapping user grouping (OUG) based on the greedy approach (OUG-Greedy). They also demonstrated that the OUG-Greedy can achieve higher capacity than existing greedy user selection algorithms and ensure that each user will be selected in at least one beamforming subgroup. Such defined user grouping strategy takes the full advantage of the favorable propagation which represents a key property in massive MIMO systems [9]. An overlapping user grouping approach based on the spectral clustering (OUG-SC) has been also proposed in [8]. Spectral clustering method has many fundamental advantages comparing to the traditional K-means clustering. However, it also requires the number of clusters as an input. The OUG-SC algorithm has reduced computational complexity but it achieves lower throughput performance than OUG-Greedy algorithm [8]. This is due to the fact that OUG-Greedy algorithm directly optimizes sum capacity with the greedy user selection approach. On the other hand, OUG-SC algorithm uses indirect metric for channel similarity measure as a part of spectral clustering procedure [8, 10].

The joint decoding at the receiver side is not feasible in MU-MIMO system since users cannot cooperate due to their random geographic location. Hence, the successful data transmission is extremely dependent on the precoding technique deployed at the base station, i.e., the ability to simultaneously send independent signals and suppress interference between users as much as possible. When channel state information (CSI) is considered known at the transmitter side (i.e., reliably estimated), the nonlinear dirty paper coding (DPC) technique [11] can completely eliminate interuser interference and achieve the maximum MU-MIMO system capacity. The Tomlinson-Harashima precoding (THP) [12] represents the simplified version of DPC which combines symmetric modulo operation and achieves near maximum capacity performance. Another prominent nonlinear precoding technique is vector perturbation (VP) [13], which perturbs signal data vectors intended for different users in order to achieve better orthogonalization. Thus, a more reliable decoding can be achieved on the receiver side. Low complexity user grouping strategies based on VP technique were proposed to support adaptive modulation mechanism [14, 15]. In traditional VP algorithm, where the same modulation scheme is applied for all users, perturbation signal is found via closest-point lattice search which is the nondeterministic polynomial-time hard (NP-hard) problem. The lattice-reduction-aided (LR-aided) algorithm could be used to overcome this challenge. However, THP has lower complexity and it outperforms LR-aided VP in the case of the large-scale MIMO application scenario [16]. Anyhow, the computational complexity of nonlinear precoding schemes significantly increases with the number of users which complicates their practical implementation.

Conversely, the linear precoding schemes with the reduced complexity are also proposed for MU-MIMO systems, such as zero forcing (ZF) and block diagonalization (BD) [17]. These schemes are successfully deployed in long-term evolution (LTE) networks and can mitigate interuser interference by projecting signal of the intended user into the null space of all the other users. However, in the case of users with highly correlated channels, it is almost impossible to discriminate signals with the projection operation which results in high capacity loss. In order to enhance MU-MIMO system capacity and alleviate its complexity at the same time, a combination of linear and nonlinear precoding schemes, i.e., two-stage precoding scheme, is proposed in the Third Generation Partnership Project (3GPP) standardization phase for 5G NR [18, 19].

In [20], Zarei et al. proposed low-complexity two-stage H-L-THP precoding scheme which achieves performance close to the conventional THP. It was assumed that all users within the same group have identical CSI statistics. However, a concrete user grouping strategy was not considered in [20] even though it significantly contributes to the overall MU-MIMO system complexity. In [21], Trifan et al. proposed two-stage BD-THP precoding scheme based on the optimized K-means clustering with the imposed cluster size constraint and a distance metric based on the angles between users. Yet, this approach does not provide information on the channel separation between users associated with different clusters. Moreover, in this approach, user selection within the cluster is performed randomly. This can result in scheduling of users with the unsuitable mutual channel conditions and a degradation of MU-MIMO system performance.

In this paper, we propose an approach in which the existing hybrid two-stage precoding scheme is extended with the overlapping user grouping strategy. Also, the comprehensive

analysis on its computational complexity throughput and BER performance has been conducted for mMTC 5G NR use case. Instead of further modification of K-means clustering, like in [21], we here adopt the overlapping user grouping method from OUG-Greedy algorithm. This algorithm considers both user selection and user ordering in order to maximize MU-MIMO system capacity and to ensure that users with the favorable channel conditions are assigned to multiple beamforming subgroups simultaneously. Two-stage precoding technique is used afterwards to separate newly formed beamforming subgroups in the spatial domain. In the first stage, ZF scheme is used to block-diagonalize the channel matrix, i.e., to minimize the intergroup interference. In the second stage, for each subgroup, a THP scheme is used to eliminate the interuser interference. The main difference between our two-stage precoding technique and the ones proposed in [20, 21] is that calculation of precoding matrices for beamforming subgroups is done in the initial step by OUG-Greedy so that linear ZF precoder can directly use them for block diagonalization which simplifies overall beamforming procedure. It should be also noticed that application of OUG-Greedy algorithm yields to the significant capacity gain compared to the legacy user grouping when combined with two-stage precoding technique. Also, we adopted two-stage ZF-THP precoding in order to accomplish balance between the achieved capacity gains and the suffered computational complexity (i.e., in comparison to the case in which only THP is used). While the existing works on two-stage precoders based on legacy user grouping strategies compare their performance only with the performance of linear precoders or two-stage schemes with two-stage linear precoding, we here benchmark proposed algorithm against nonlinear precoders and two-stage schemes with nonlinear precoding as well. Hence, this paper also provides the comparative analysis of all precoding types combined with legacy and overlapping user grouping methods.

The rest of the paper is organized as follows. System model is introduced and user grouping problem is formulated in Section 2. In Section 3, the proposed two-stage precoding scheme based on overlapping user grouping strategy is proposed. Complexity evaluation of the proposed algorithm is carried out in Section 4. Numerical simulation results and comparative analysis with the algorithms that employ existing user grouping methods and precoding schemes are presented in Section 5. Section 6 concludes this paper and presents research directions for the future work.

2. System Model and Problem Formulation

2.1. System Model. The downlink of a single-cell MU-MIMO system is considered, in which a base station with a uniform rectangular antenna array of N antennas simultaneously transmits data to K single-antenna IoT devices (IoT users). We did not consider IoT devices equipped with multiple antennas since these are generally considered as a small and simple devices. It would not be practical to equip them with MIMO antennas because it would not provide sufficient spatial diversity between the antennas to enable effective operation. The choice of multiple antennas would demand independent RF chains per each antenna and advanced digital processing to separate the data streams. This would increase cost and complexity of IoT devices, and also increase energy consumption that is not appropriate for the battery powered devices. Channel matrix is assumed fixed during the channel coherence time and can be expressed as $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]^{\mathrm{T}} \in \mathbb{C}^{K\times N}$, where $(\bullet)^{\mathrm{T}}$ denotes matrix or vector transpose, and $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ is the channel vector between the base station and user k. As in the previous work in this area, we assume that CSI is known at the base station. Let denote y_k as the received signal at user k. The signals received by users can be written as follows:

$$\mathbf{y} = \mathbf{HBd} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^{K\times 1}$ denotes the received data for all K users in a single time slot, $\mathbf{B} \in \mathbb{C}^{N\times K}$ is the precoding matrix, $\mathbf{d} \in \mathbb{C}^{K\times 1}$ is the data vector intended for transmission to K users where $d_k \in A = \{a_{\mathrm{I}} + ja_{\mathrm{Q}} \mid a_{\mathrm{I}}, a_{\mathrm{Q}} \in \{\pm 1, \pm 3, \cdots, \pm(\sqrt{M} - 1)\}\}$ is M-QAM modulated data symbol of the kth user with modulation order M, and $\mathbf{n} \sim CN(0, \mathbf{I}_K)$ is the additive white Gaussian noise (AWGN) vector with zero mean and unit variance. The choice of this particular modulation scheme is made since the traditional THP precoder only applies for M-QAM signaling. Modified THP, which is characterized with similar complexity as traditional THP, was recently designed to support M-PSK modulations included in 5G standardization for millimeter wave communications [22]. Described system model operates on sub-6 GHz band; hence, the traditional THP precoder is sufficient for this scenario and it also simplifies receiver design. The total power of transmitted signal $\mathbf{x} = \mathbf{Bd} \in \mathbb{C}^{N\times 1}$ is constrained to $\mathbb{E}[\mathbf{xx}^{\mathrm{H}}] \leq P_T$, where $\mathbb{E}(\bullet)$ stands for the expectation operator and $(\bullet)^{\mathrm{H}}$ denotes matrix or vector Hermitian transpose. Throughout this manuscript, bold uppercase and lowercase symbols are used to denote matrices and vectors, respectively, and the normal symbols are used to represent scalars.

In many urban mMTC 5G NR use cases, IoT devices are located indoor, whereas macrobase station is located outdoor. Hence, we here consider that base station communicates with users over the spatially correlated Rayleigh channels characterized with the non-line-of-sight (NLOS) propagation [8].

In the considered scenario, base station is elevated and free of local scattering, which results in high correlation among the transmit antennas. We model spatial correlation matrix at the transmitter $\mathbf{R}_{TX} \in \mathbb{C}^{N\times N}$ using the one-ring MIMO channel scattering model shown in Figure 1, which was firstly employed by Jakes [23] and adopted in [24]. Let θ be the azimuth angle of the user located at distance S from base station and surrounded by a ring of scatterers with radius r. From Figure 1, it follows that angular spread of transmitted signal Δ can be approximated as $\Delta \approx \mathrm{arctg}(r/S)$. Spatial correlation coefficient between transmit antennas $1 \leq p, q \leq N$ is modelled as follows [24]:

$$[\mathbf{R}_{TX}]_{p,q} = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} e^{j\mathbf{g}^{\mathrm{T}}(\alpha+\theta)\left(\mathrm{u}_p - \mathrm{u}_q\right)} d\alpha, \tag{2}$$

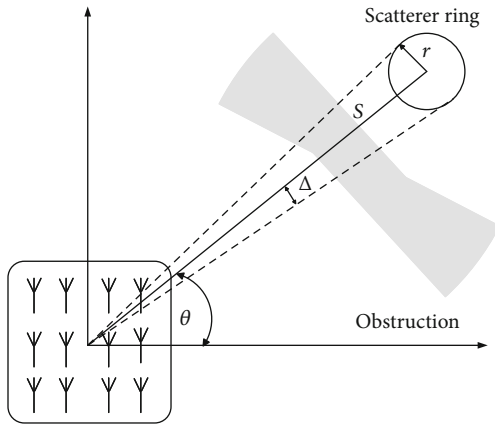

Figure 1: One-ring MIMO channel scattering model.

where $\mathbf{g}(\alpha+\theta) = -(2\pi/\lambda_c)[\cos(\alpha+\theta)\ \sin(\alpha+\theta)]^{\mathrm{T}}$ is the vector for a planar wave impinging the transmit antenna array with the angle of arrival (AoA) α, λ_c is the wavelength that corresponds to carrier frequency f_c, and $\mathbf{u}_p$, $\mathbf{u}_q \in \mathbb{R}^2$ are vectors indicating the position of base station antennas p, q in two-dimensional (2D) coordinate system.

From Equation (2), it can be verified that $\mathbf{R}_{TX}$ is a normal matrix which can be eigendecomposed as follows:

$$\mathbf{R}_{TX} = \mathbf{U}_{TX}\,\mathbf{\Sigma}_{TX}\mathbf{U}_{TX}^{\mathrm{H}}, \tag{3}$$

where $\mathbf{U}_{TX} \in \mathbb{C}^{N\times N}$ represents a unitary matrix composed of the eigenvectors of $\mathbf{R}_{TX}$ and $\mathbf{\Sigma}_{TX} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose elements are eigenvalues of $\mathbf{R}_{TX}$.

IoT devices located indoor usually experience fluctuation of the received signal power due to the obstacles on the transmission path, i.e., shadow fading. The channels of geographically proximate devices are significantly correlated when affected by the same shadowing. Spatial correlation of the channels between users $1 \le i, j \le K$ is modelled using Gudmundson's model defined in [25] and adapted for IoT networks in [26] as follows:

$$[\mathbf{R}_{RX}]_{i,j} = \frac{\sigma_s^2}{d_{\mathrm{cor}}} e^{-|d_{i,j}|/d_{\mathrm{cor}}}, \tag{4}$$

where $|d_{i,j}|$ denotes the distance between users i and j, σ_s is the standard deviation of shadow fading, and d_{cor} is the correlation distance, i.e., distance at which correlation drops to 0.5. $\mathbf{R}_{RX} \in \mathbb{C}^{K\times K}$ is also a normal matrix with eigendecomposition similar to Equation (3)

$$\mathbf{R}_{RX} = \mathbf{U}_{RX}\,\mathbf{\Sigma}_{RX}\mathbf{U}_{RX}^{\mathrm{H}}, \tag{5}$$

where unitary matrix $\mathbf{U}_{RX} \in \mathbb{C}^{K\times K}$ and diagonal matrix $\mathbf{\Sigma}_{RX} \in \mathbb{R}^{K\times K}$ include the corresponding eigenvectors and eigenvalues of $\mathbf{R}_{RX}$, respectively. We here adopted Kronecker correlation model [27], which assumes complete correlation separability between transmitter and receiver. Hence, channel matrix can be expressed as follows:

$$\mathbf{H} = \mathbf{R}_{RX}^{1/2}\mathbf{H}_{\mathrm{iid}}\mathbf{R}_{TX}^{1/2}, \tag{6}$$

where $\mathbf{H}_{\mathrm{iid}} \in \mathbb{C}^{K\times N}$ is an uncorrelated Rayleigh channel matrix whose elements are independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. Substitution of decomposed spatial correlation matrices at transmitter (Equation (3)) and receiver (Equation (5)) in Equation (6) gives the following channel matrix expression:

$$\mathbf{H} = \mathbf{U}_{RX}\mathbf{\Sigma}_{RX}^{1/2}\mathbf{H}_{\mathrm{iid}}\mathbf{\Sigma}_{TX}^{1/2}\mathbf{U}_{TX}. \tag{7}$$

2.2. User Grouping Problem Formulation. The performance of MU-MIMO system largely depends on the channel correlation among the users included in the same beamforming subgroup. Hence, the proper user grouping is necessary in order to suppress interuser interference and maximize system capacity.

Let $\mathbb{S} = \{k \mid k = 1, 2, \cdots, K\}$ denote the whole set of users clustered into G subgroups. Deterministic MIMO channel capacity for each beamforming subgroup $\mathbb{S}_g$, $g = 1, 2, \cdots, G$ is defined as [28]:

$$R_g(\mathbf{H}, \mathbf{B}) = \sum_{k\in\mathbb{S}_g} \log_2\left(1 + \beta_k \left\|\mathbf{h}_k^{\mathrm{H}}\mathbf{b}_k\right\|\right), \tag{8}$$

where $\|\bullet\|$ denotes the vector 2-norm operator. Parameters β_k symbolize the power allocation factors derived from the water-filling algorithm [29]:

$$\beta_k = \left(\frac{1}{\mu} - \frac{1}{\lambda_k}\right)^+, \tag{9}$$

where $(\bullet)^+$ is the operation defined as $(x)^+ = \max\{0, x\}$ and μ is the water level satisfying

$$\sum_{k\in\mathbb{S}_g}\left(\frac{1}{\mu} - \frac{1}{\lambda_k}\right) = P_T, \tag{10}$$

and λ_k is the effective channel gain after beamforming procedure:

$$\lambda_k = \left\|\mathbf{h}_k^{\mathrm{H}}\mathbf{b}_k\right\|, \tag{11}$$

which represents the kth eigenvalue of the effective channel matrix $\mathbf{H}_{\mathrm{eff}} = \mathbf{HB}$ [3].

Different user selections for beamforming subgroups g give different values of Equation (8). Furthermore, different user ordering within the same beamforming subgroup also yields different MU-MIMO sum capacity. In general, user grouping strategy depends on the channel matrix $\mathbf{H}$ and the transmitted signal power P_T. Thus, we define the optimal user grouping method $\mathbf{S}^*(\mathbf{H}, P_T) \triangleq \{\mathbb{S}_1, \cdots, \mathbb{S}_G\}$ as the one that maximizes MU-MIMO system capacity. The corresponding

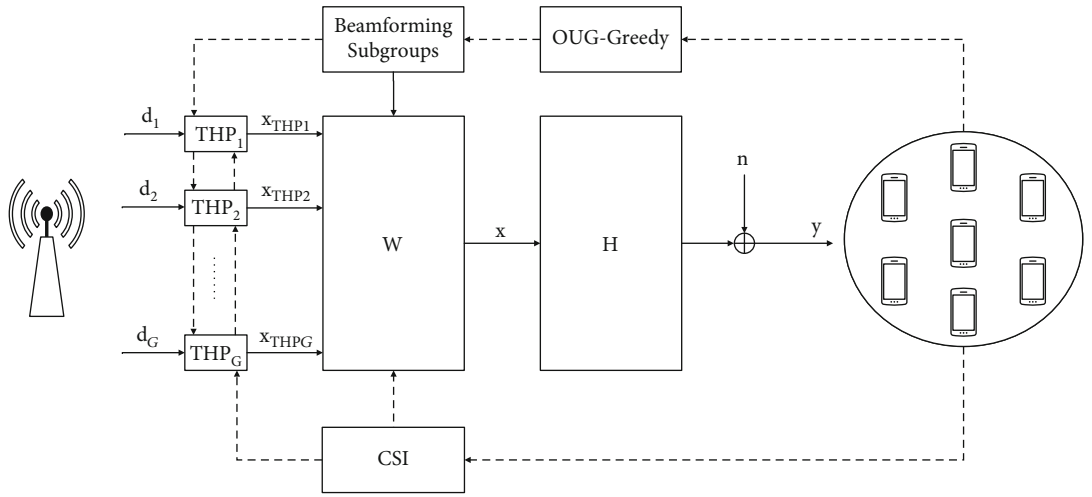

Figure 2: System model for two-stage precoding with overlapping user grouping approach.

optimal power allocation defined by $\boldsymbol{\beta}^*(\mathbf{H}, P_T) \triangleq \{\beta_1, \cdots, \beta_K\}$ gives the maximum sum capacity under the user grouping strategy $\mathbf{S}^*(\mathbf{H}, P_T)$. Putting all together, the optimal user grouping problem can be formulated as in [8]:

$$\{\mathbf{S}^*(\mathbf{H}, P_T), \boldsymbol{\beta}^*(\mathbf{H}, P_T)\} = \arg\max \sum_{g=1}^{G} R_g, \tag{12}$$

subject to $\bigcup_{g=1}^{G} \mathbb{S}_g = \mathbb{S}$ and $\sum_{k=1}^{K} \beta_k \leq P_T$. As can be seen from the previous expression, the sum capacity can be optimized with respect to the overlapping among beamforming subgroups and power allocation when solving the optimization problem (Equation (12)).

3. Two-Stage Precoding Based on Overlapping User Grouping Approach

The system model for the proposed two-stage precoding scheme based on overlapping user grouping strategy is depicted in Figure 2.

User grouping is achieved by employing the overlapping method from OUG-Greedy algorithm introduced in [8]. Let $\mathbb{S}_i$ be the set of users that have been assigned in iteration i and $\mathbb{C}_i$ be the set of remaining users that have not been selected yet. In each iteration, algorithm selects users from $\mathbb{C}_i$ in order to form the subgroup $\mathbb{S}_i$ which gives the maximum capacity defined in Equation (8) with the corresponding water-filling power allocation. This procedure is known as zero forcing with user selection (ZFS) [7] and is repeated until all users are assigned to their respective beamforming subgroups. In the next step, the searching space of subgroup $\mathbb{S}_i$ is widened to the users that have been already assigned to one of the previous subgroups. More specifically, the searching space for subgroup $\mathbb{S}_i$ obtained from ZFS algorithm is reset as follows:

$$\mathbb{C}_{i,o} = \bigcup_{j=1}^{i-1} \mathbb{S}_j, \tag{13}$$

to perform the overlapping user grouping [8]. Using the extended searching space, users with the favorable channel conditions are reselected and assigned to several beamforming subgroups at the same time. Accordingly, we obtain the set of overlapping user groups $\mathbb{S}_{1,o}, \cdots, \mathbb{S}_{G,o}$ and corresponding set of matrices $\mathbf{H}_1, \cdots, \mathbf{H}_G$ where $\mathbf{H}_g \in \mathbb{C}^{K_g \times N}$ denotes the row-reduced channel matrix which includes channel vectors of K_g users selected in beamforming subgroup $\mathbb{S}_{g,o}$.

Once users are grouped according to the OUG-Greedy algorithm, linear ZF precoding scheme is applied to suppress interference between already formed beamforming subgroups. For this purpose, precoder $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_G]$ with $\mathbf{W}_g \in \mathbb{C}^{N \times K_g}$ is designed to null off-diagonal elements of the effective ZF channel matrix:

$$\mathbf{HW} = \begin{bmatrix} \mathbf{H}_1\mathbf{W}_1 & \mathbf{H}_1\mathbf{W}_2 & \cdots \mathbf{H}_1\mathbf{W}_G \\ \mathbf{H}_2\mathbf{W}_1 & \mathbf{H}_2\mathbf{W}_2 & \cdots \mathbf{H}_2\mathbf{W}_G \\ \vdots & \vdots & \vdots \\ \mathbf{H}_G\mathbf{W}_1 & \mathbf{H}_G\mathbf{W}_2 & \cdots \mathbf{H}_G\mathbf{W}_G \end{bmatrix}. \tag{14}$$

In order to cancel intergroup interference, the effective ZF channel matrix from Equation (14) must be diagonalized, i.e., $\mathbf{H}_i\mathbf{W}_j = 0$ for every $i \neq j$. This is possible when precoding matrix for each beamforming subgroup is a Moore-Penrose pseudoinverse of the row-reduced channel matrix [30].

$$\mathbf{W}_g = \mathbf{H}_g^{\mathrm{H}} \left(\mathbf{H}_g \mathbf{H}_g^{\mathrm{H}} \right)^{-1}. \tag{15}$$

Hence, the user data in each beamforming subgroup is ideally transmitted in the null space of the channel matrix made of channel vectors related to users from all other subgroups. However, it should be noticed that it is not necessary to determine previous Equation (15) since the corresponding precoding matrices are already obtained in OUG-Greedy algorithm when calculating Equation (8). This leads to simplified ZF precoding which only includes multiplication of

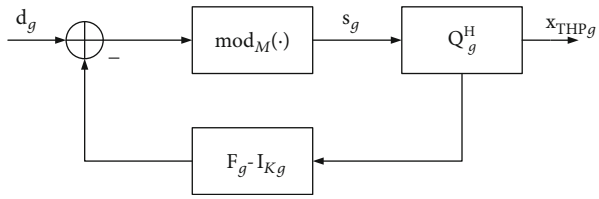

Figure 3: System model for THP scheme.

precoding matrices obtained from OUG-Greedy algorithm with the corresponding row-reduced channel matrices.

After the ZF precoding technique is performed, remaining interuser interference in each beamforming subgroup is mitigated by using the nonlinear THP precoding scheme. THP precoded signal for beamforming subgroup g is given by $\mathbf{x}_{\mathrm{THP}g} = \mathbf{Q}_g^{\mathrm{H}}\mathbf{s}_g \in \mathbb{C}^{K_g \times 1}$ as shown in Figure 3. Thus, $\mathbf{Q}_g \in \mathbb{C}^{K_g \times K_g}$ is an unitary feedforward matrix obtained from LQ decomposition of gth diagonal element of the effective channel matrix $\check{\mathbf{H}}_g = \mathbf{H}_g\mathbf{W}_g \in \mathbb{C}^{K_g \times K_g}$, and $\mathbf{s}_g \in \mathbb{C}^{K_g \times 1}$ is a data vector whose elements are calculated according to the following:

$$\left[\mathbf{s}_g\right]_{k,1} = \mathrm{mod}_M\left(\left[\mathbf{d}_g\right]_{k,1} - \sum_{l=1}^{k-1}\left[\mathbf{F}_g - \mathbf{I}_{K_g}\right]_{k,l}\left[\mathbf{s}_g\right]_{l,1}\right), \quad (16)$$

where $\mathbf{F}_g \in \mathbb{C}^{K_g \times K_g}$ represents the feedback matrix, $\mathbf{I}_{K_g} \in \mathbb{C}^{K_g \times K_g}$ denotes the identity matrix, and $\mathrm{mod}_M(\bullet)$ is a symmetric modulo function which limits transmitted power of modulated data symbols and ensures that they lie inside the Voronoi region of the original constellation and is given by the following:

$$\begin{aligned}\mathrm{mod}_M(x) = x &- 2\sqrt{M}\left\lfloor\frac{1}{2} + \Re\left\{\frac{x}{2\sqrt{M}}\right\}\right\rfloor \\ &- 2j\sqrt{M}\left\lfloor\frac{1}{2} + \Im\left\{\frac{x}{2\sqrt{M}}\right\}\right\rfloor,\end{aligned} \quad (17)$$

with $\Re\{\cdot\}$ and $\Im\{\cdot\}$ representing the real and imaginary part of complex number x. The main purpose of the feedback matrix $\mathbf{F}_g$ is to cancel the interference caused by already detected data symbols and is defined as follows:

$$\mathbf{F}_g = \mathbf{G}_g\mathbf{L}_g, \quad (18)$$

where $\mathbf{G}_g = \mathrm{diag}\left(l_{11}^{-1}, \cdots, l_{K_gK_g}^{-1}\right) \in \mathbb{C}^{K_g \times K_g}$ is the diagonal scaling matrix and $\mathbf{L}_g \in \mathbb{C}^{K_g \times K_g}$ is the lower triangular matrix derived from LQ decomposition of $\check{\mathbf{H}}_g$. Hence, THP precoding matrix $\mathbf{P}_g \in \mathbb{C}^{K_g \times K_g}$ for beamforming subgroup g can be expressed as follows:

$$\mathbf{P}_g = \mathbf{Q}_g^{\mathrm{H}}\mathbf{L}_g^{-1}\mathbf{G}_g^{-1}. \quad (19)$$

As can be seen, the proposed hybrid mechanism is based on two-stage precoding. The first stage consists of the linear precoder used to eliminate intergroup interference. To suppress interference inside every group, the nonlinear precoding is employed in the second stage. In other words, beamforming matrix $\mathbf{B}$ consists of two parts:

$$\mathbf{B} = \mathbf{WP}, \quad (20)$$

where $\mathbf{W} \in \mathbb{C}^{N \times K}$ denotes the linear ZF beamforming matrix and $\mathbf{P} \in \mathbb{C}^{K \times K}$ is the cumulative nonlinear THP beamforming matrix.

The achievable sum rate of the proposed algorithm is calculated as $\sum_{g=1}^{G} R_g$ where R_g represents channel capacity for overlapped beamforming subgroup $\mathbb{S}_{g,o}$ defined as in [28]:

$$R_g(\mathbf{H}, \mathbf{B}) = \sum_{k \in \mathbb{S}_g} \log_2\left(1 + \beta_k\left\|\mathbf{h}_k^{\mathrm{H}}\mathbf{b}_k\right\|\right), \quad (21)$$

which is equivalent to Equation (8) and derived for the case of perfect MIMO channel estimation. Previous formula was used for the performance evaluation of all existing algorithms and the proposed one in Section 5.

4. Computational Complexity Analysis

Computational complexity is an important design parameter, especially in implementation of IoT-oriented 5G systems where a massive number of IoT devices have limited battery lifetime. This section covers complexity analysis of the proposed scheme with two-stage ZF-THP precoding based on overlapping user grouping approach (marked as OUG ZF-THP algorithm). In order to achieve this, the computational complexity for deployed overlapping user grouping method and two-stage ZF-THP precoding is derived. The total computational capacity is defined as the sum of these two parts (excluding the calculations from the prior steps that can be reused in the former steps). Also, in order to compare computational complexity for the proposed and the referent algorithms, the complexity for these algorithms is given. As the referent algorithms, we here observed previously introduced OUG-Greedy grouping with the linear ZF precoding (marked as OUG-Greedy ZF algorithm) proposed in [8], two-stage BD-THP precoding based on the optimized K-means clustering (marked as K-means BD-THP algorithm) proposed in [21], and linear ZFS algorithm proposed in [7]. We here also consider scheme with THP precoding based on overlapping user grouping strategy (marked as OUG THP), a combination that was not previously observed in the literature. More on this referent scheme is given in the next section where the capacity performance analysis is presented. Since all these algorithms also comprise user grouping and the precoding part, the computational complexity is presented for the both of these parts separately, and the total complexity is given as the sum of these two (in the same way as for the proposed algorithm).

The complexity for all the observed algorithms is quantified by the number of floating-point operations (FLOPs) [30] required for multiplication (division) and addition (subtraction) of complex-valued numbers. For the sake of accuracy, we use a common assumption and count each complex-

valued multiplication as 6 FLOPs and each complex-valued addition as 2 FLOPs. Also, computational complexity required for T_c precoded data vectors is considered, where T_c represents the channel coherence time interval.

First, we consider complexity of the overlapping user grouping method. For the sake of brevity, it was assumed that each beamforming subgroup has $K_g = K/G$ users. Derivation of the effective channel gains represents the most computationally expensive operation in the overall OUG-Greedy algorithm [7]. As an alternative to Equation (11), the simplified sequential water-filling (SWF) approach for channel matrix pseudoinverse Equation (15) and channel gain Equation (11) calculation is introduced in [4] as follows:

$$\lambda_k = \left|\mathbf{h}_k \mathbf{P}_k^{\perp}\right|^2, \tag{22}$$

where $\mathbf{P}_k^{\perp} \in \mathbb{C}^{N\times N}$ is a projection matrix onto the orthogonal complement of the subspace spanned by the channels $\mathbf{h}_1, \cdots, \mathbf{h}_{k-1}$ of the currently selected users in that beamforming subgroup. Vector-matrix multiplication in Equation (21) requires $2N(4N-1)$ FLOPs [31]. In the worst-case scenario, this procedure is performed for all K users in K_g iterations. Repetition over G beamforming subgroups gives the total number of FLOPs:

$$C_{\text{OUG-Greedy}} = 2G^2K_g^2N(4N-1), \tag{23}$$

which can be formulated after substitution $K_g = K/G$ as follows:

$$C_{\text{OUG-Greedy}} = 2K^2N(4N-1). \tag{24}$$

Hence, the computational complexity of the overlapping user grouping strategy is no more than $O(K^2N^2)$ which is of the same order as ZFS with SWF mechanism [4] and two-order simpler comparing to the conventional capacity-based ZFS algorithm [32] as outlined in Table 1. In practice, the number of beamforming subgroups G will be more than 2; thus, OUG-Greedy algorithm also outperforms optimized K-means clustering method used by K -means BD-THP algorithm [21].

Next, we derive the computational complexity of two-stage ZF-THP precoding scheme. Matrix-matrix multiplication is executed in order to obtain diagonal elements of the effective channel matrix $\check{\mathbf{H}}_g$ which requires $2K_g^2(4N-1)$ FLOPs. Note that complexity of calculating Moore-Penrose pseudoinverses $\mathbf{W}_g$ has been already evaluated as part of the user grouping procedure. LQ decomposition of matrix $\check{\mathbf{H}}_g$ requires approximately $16K_g^3/3$ FLOPs [33]. Calculation of diagonal scaling matrix $\mathbf{G}_g$ requires K_g FLOPs which is used for generating feedback matrix $\mathbf{F}_g$ with complexity of $2K_g^2(4K_g-1)$ FLOPs. Subtracting identity matrix $\mathbf{I}_{K_g}$ from feedback matrix $\mathbf{F}_g$ requires K_g FLOPs. Calculating T_c data vectors $\mathbf{s}_g$ requires $4T_c(K_g^2 + K_g - 2)$ FLOPs [34]. Multiplication of $\mathbf{W}_g$ with the unitary feedforward matrix Hermitian $\mathbf{Q}_g^{\mathrm{H}}$ requires $2NK_g(4K_g-1)$ FLOPs. Previous steps are repeated G times for each beamforming subgroup.

Table 1: Computational complexity of user grouping algorithms.

User grouping algorithm	Number of FLOPs
Capacity-based ZFS	$O(KN^5)$
ZFS with SWF	$O(KN^3)$
OUG-Greedy	$O(K^2N^2)$
Optimized K-means clustering	$O(K^{G+1}G)$

Finally, $2T_cN(4K-1)$ FLOPs are needed to multiply the cumulative product $\mathbf{W}_g\mathbf{Q}_g^{\mathrm{H}}$ with data vectors $\mathbf{s}_g$ for all K users and generate T_c two-stage precoded data vectors $\mathbf{x}$. Thus, the total number of FLOPs required for two-stage ZF-THP precoding is as follows:

$$C_{\text{ZF-THP}} = \frac{40GK_g^3}{3} - 4GK_g^2 + 2GK_g - 2GK_gN + 16GK_g^2N + T_c\left(4GK_g^2 + 4GK_g + 8KN - 8G - 2N\right). \tag{25}$$

Application of the corresponding substitution $K_g = K/G$ gives more concise expression:

$$C_{\text{ZF-THP}} = \frac{2K\left(3G^2 - 3G^2N + 20K^2 - 6GK + 24GKN\right)}{3G^2} + \frac{2T_c\left(2K^2 - 4G^2 + 2GK - GN + 4GKN\right)}{G}. \tag{26}$$

As summarized in Table 2, two-stage ZF-THP precoding scheme has the same computational complexity as two-stage BD-THP scheme. The expressions for complexity of the precoding techniques summarized in Table 2 are adopted from [20] in the case of ZF precoding and THP schemes, and from [21] for BD-THP scheme.

Moreover, two-stage ZF-THP technique has the lowest complexity among conventional linear and nonlinear precoding schemes. The computational complexity required to generate one two-stage precoded data vector in the case of 32 antennas and IoT devices grouped in 4 beamforming subgroups is illustrated in Figure 4. For this choice of parameter values, presented precoding schemes have similar complexity when the number of IoT devices is less than 20. As the number of users in the cell increases, a computational complexity of ZF and THP precoders substantially escalates in comparison with ZF-THP and BD-THP. As expected, nonlinear THP scheme has the highest complexity.

Based on the previously defined computational complexity for different user grouping and precoding schemes, the total complexity for all observed algorithms is presented in Table 3. The total complexity is calculated as a sum of corresponding user grouping and precoding schemes for each

Table 2: Computational complexity of precoding schemes.

Precoding scheme	Number of FLOPs
ZF	$4K^3 + 2KN(4K-1) + K(4N-1)(K+1) + 2NT_c(4K-1) + 8K^2 + 6K$
THP	$\frac{16K^3}{3} + 2K(4KN - K + 2) + 2T_c(2K + 2K^2 + N(4K-1) - 4)$
BD-THP	$\frac{2K(3G^2 - 3G^2N + 20K^2 - 6GK + 24GKN)}{3G^2} + \frac{2T_c(2K^2 - 4G^2 + 2GK - GN + 4GKN)}{G}$
ZF-THP	$\frac{2K(3G^2 - 3G^2N + 20K^2 - 6GK + 24GKN)}{3G^2} + \frac{2T_c(2K^2 - 4G^2 + 2GK - GN + 4GKN)}{G}$

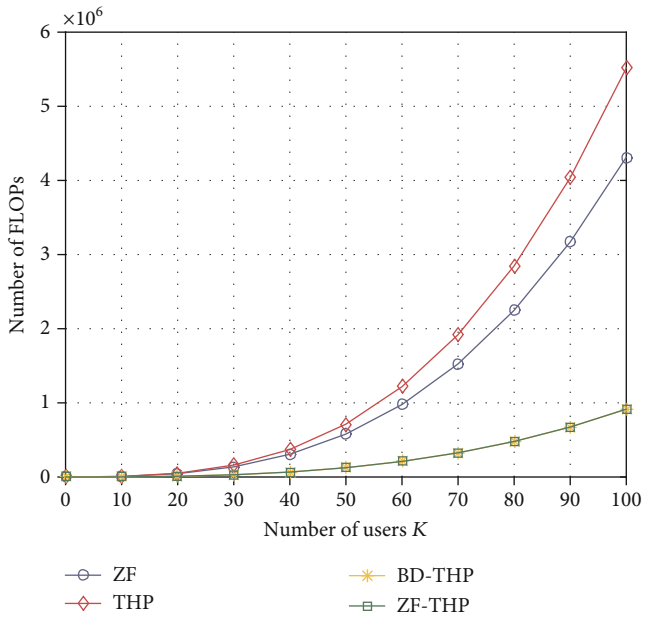

Figure 4: Computational complexity of precoding schemes.

algorithm (excluding the complexity related to calculation of beamforming matrices when ZF precoding is employed, since these were calculated as a part of user grouping procedure in ZFS, OUG ZF-THP, and OUG-Greedy ZF algorithms).

As evident in Table 3, ZFS and OUG-Greedy ZF algorithms have somewhat lower computational complexity than OUG ZF-THP algorithm. Such behaviour could be expected since these two schemes employ only linear ZF precoding with lower complexity due to the reuse of beamforming matrices already calculated as a part of user grouping procedure (similarly as for OUG ZF-THP algorithm). However, as will be presented in the next section, these schemes achieve lower overall MU-MIMO system capacity than OUG ZF-THP due to the less efficient ZF precoding in comparison to ZF-THP precoding. This is particularly evident in the case of the correlated MIMO channels (i.e., mutually dependent user channels) when linear ZF precoding cannot sucesfully mitigate all interuser interference, and more complex two-stage ZF-THP precoding achieves significantly better performance and thus enables larger capacity gains.

On the other hand, OUG ZF-THP algorithm has lower computational complexity than OUG THP and K-means BD-THP algorithms. The higher complexity of OUG THP algorithm is a consequence of using THP precoding, which possess significantly higher complexity than ZF-THP precoding. However, this more complex precoding scheme enables somewhat higher capacity gains, as will be shown in the next section. If we observe OUG ZF-THP and K-means BD-THP complexity, it is obvious in Table 2 and Figure 4 that ZF-THP and BD-THP precoding schemes require the same amount of FLOPs. However, complexity order of the optimized K-means clustering adopted in K-means BD-THP algorithm increases linearly with the number of beamforming subgroups G [21] which is not the case with overlapping user grouping method deployed in OUG ZF-THP algorithm. Hence, overall OUG ZF-THP algorithm is more computationally efficient than K-means BD-THP algorithm.

Also, it is worth mentioning that in K-means BD-THP algorithm, the computational complexity of singular value decomposition (SVD) procedure for the linear precoding part (i.e., BD procedure) is neglected since it has to be determined very infrequently from the long-term CSI. In here proposed OUG ZF-THP algorithm, calculation of beamforming matrices is done as a part of user grouping procedure which simplifies subsequent linear ZF precoding. This gives more realistic evaluation of OUG ZF-THP complexity.

5. Results and Discussion

To evaluate the performance of the proposed two-stage ZF-THP precoding based on overlapping user grouping approach (OUG ZF-THP algorithm), we compared the MU-MIMO system capacity for this algorithm with the linear OUG-Greedy ZF algorithm [8] and two-stage BD-THP precoding based on optimized K-means clustering (K-means BD-THP) [21]. First algorithm introduces the overlapping user grouping method which showed good performance in IoT-oriented MU-MIMO system. Latter one considers concrete user grouping strategy in junction with two-stage precoding scheme for the first time. Thus, these algorithms represent suitable candidates for the performance benchmarking. For the sake of completeness, we also show simulation results for ZFS [7] and THP precoding [12], combined with the overlapping user grouping strategy (OUG THP) that defines practical lower and upper bound of the MU-MIMO capacity region for this particular case,

Table 3: Computational complexity of all observed algorithms comprising the user grouping and precoding.

Algorithm	Number of FLOPs
ZFS	$O(KN^3) + 2KN(4K-1) + 2NT_c(4K-1)$
OUG-Greedy ZF	$O(K^2N^2) + 2KN(4K-1) + 2NT_c(4K-1)$
K-means BD-THP	$O(K^{G+1}G) + \frac{2K(3G^2 - 3G^2N + 20K^2 - 6GK + 24GKN)}{3G^2} + \frac{2T_c(2K^2 - 4G^2 + 2GK - GN + 4GKN)}{G}$
OUG ZF-THP	$O(K^2N^2) + \frac{2K(3G^2 - 3G^2N + 20K^2 - 6GK + 24GKN)}{3G^2} + \frac{2T_c(2K^2 - 4G^2 + 2GK - GN + 4GKN)}{G}$
OUG THP	$O(K^2N^2) + \frac{16K^3}{3} + 2K(4KN - K + 2) + 2T_c(2K + 2K^2 + N(4K-1) - 4)$

respectively. 2D MU-MIMO system environment has been created using the MATLAB software package. The MU-MIMO system capacity for all the observed algorithms was estimated according to Equation (21) defined in the Section 3. The Monte-Carlo simulation of these algorithms is performed by averaging 500 random channel realizations.

We assumed a single-cell MU-MIMO system with a base station located at the center of the cell and equipped with 128 omnidirectional antennas which represent the typical configuration of the commercial massive MIMO antenna. It simultaneously transmits data in 3.5 GHz band to 300 single-antenna IoT devices. This frequency band has been identified as a global International Mobile Telecommunications-2020 (IMT-2020) band for 5G NR deployment by International Telecommunication Union Radiocommunication Sector (ITU-R) [35]. Configuration of the planar antenna array is 8×16 (i.e., 8 antenna elements vertically and 16 antenna elements horizontally) with the aim to exploit 2D beamforming in horizontal domain. The base station antenna spacing is normalized with respect to the wavelength and set to 0.5. Data symbols are modulated with 16-QAM technique which was shown sufficient for mMTC 5G NR use cases [36]. To simulate the transmit and receive antenna correlation, we adopted Jakes' one-ring MIMO channel model [23] and Gudmundson's shadowing model [25] commonly used in cellular IoT scenarios, respectively.

Parameter values for both correlation models were taken from [8] since the same type of propagation environment was considered. The scattering objects are located around devices in radius of 30 meters [24], while correlation distance between devices is 20 meters and shadow fading varies with standard deviation of 0.4 [26]. IoT devices are uniformly distributed around base station with dedicated azimuth values at distance between 100 and 300 meters. The angular spread of the signal transmitted from the base station is derived as in [24]. This is aligned with the expected beam arrival distance in rich scattering radio environment for chosen antenna configuration and operating frequency band. An overview of the main system parameter configuration used in Monte-Carlo simulations is provided in Table 4.

In [8], it was mathematically shown that overlapping user beamforming subgroup $\mathbb{S}_{g,o}$ can select more users with a higher probability than the corresponding beamforming subgroup $\mathbb{S}_g$ and that searching space extension always results in larger capacity. We demonstrate numerically the superiority of here proposed approach that combines overlapping user grouping method with two-stage ZF-THP precoding scheme in terms of the achievable MU-MIMO system capacity.

Table 4: System parameter configuration.

Parameter	Value	Parameter	Value
θ	[-180°, 180°]	N	128
Δ	[5°, 15°]	K	300
d_{cor}	20 m	D	0.5
r	30 m	σ_s	0.4
f_c	3.5 GHz	M	16

First, we have evaluated the proposed algorithm performance in the case of the environment with uncorrelated Rayleigh fading where users' channels are mutually independent. Equivalent channel-based received signal-to-noise ratio (SNR) to throughput mapping method adopted by 3GPP [19] is used for performance evaluation. MU-MIMO system capacity comparison of the analyzed algorithms is outlined in Figure 5.

It can be seen that proposed OUG ZF-THP algorithm achieves approximately the same capacity as OUG THP and OUG-Greedy ZF algorithms. The same finding holds for the conventional ZFS and K-means BD-THP algorithms. This is due to the fact that linear precoding has almost the same performance as nonlinear precoding when users have uncorrelated channel vectors. Hence, linear precoder can efficiently suppress both intergroup and interuser interference and there is no need to use two-stage hybrid precoding mechanism. K-means BD-THP algorithm has lower throughput performance due to a lack of overlapping approach which could further populate formed subgroups using the favorable propagation property. In the case of uncorrelated MIMO channels, the capacity improvement is mainly achieved by overlapping user grouping strategy among beamforming subgroups.

Next, we consider more realistic scenario with correlated shadow fading which imposes dependency between user channels. Results in Figure 6 show that proposed OUG ZF-THP algorithm achieves significant improvement on sum capacity over the existing suboptimal approaches. We can

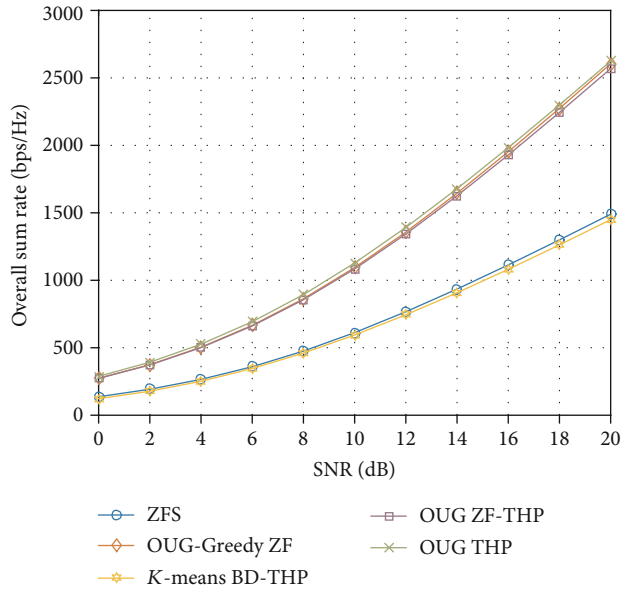

Figure 5: Sum rate comparison under uncorrelated MIMO channels.

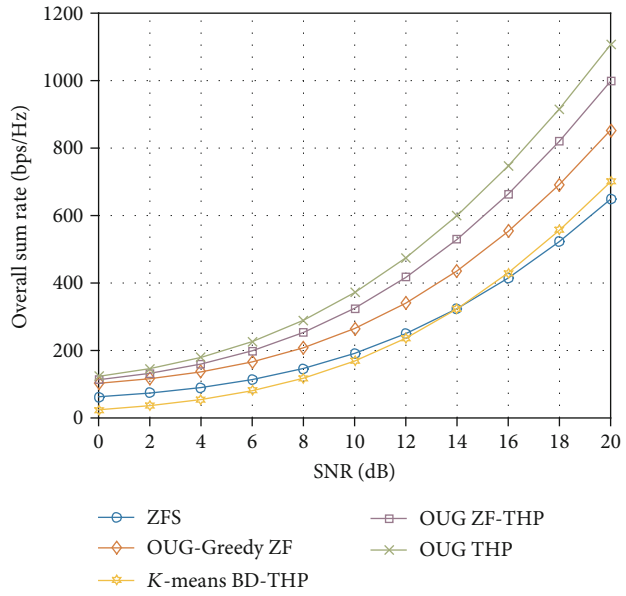

Figure 6: Sum rate comparison under correlated MIMO channels.

observe capacity increase from 10% and 30% in low SNR regime (4 dB) to 25% and 45% in high SNR regime (20 dB) comparing to OUG-Greedy ZF and K-means BD-THP algorithms, respectively. SNR reference values are chosen according to real 5G urban NLOS radio conditions at 3.5 GHz [37].

Obtained large performance gain of OUG ZF-THP algorithm is the result of the proposed combination of more advanced two-stage signal processing and overlapping among beamforming subgroups. Linear OUG-Greedy ZF algorithm achieves lower sum rate due to the correlation of user channel vectors, whereas the poor performance of K-means BD-THP algorithm comes from the random user selection within clusters. Additionally, high SNR regime improves transmission reliability and beam steering which contribute to the higher achievable system throughput. Algorithms that use overlapping user grouping based on greedy user selection methods are exploiting the favorable propagation property without generating much interuser interference which greatly enhances MU-MIMO system capacity. From Figure 6, it can also be seen that nonlinear THP precoding based on overlapping user grouping strategy (i.e., OUG THP) provides the best sum rate. However, here proposed algorithm requires significantly less FLOPs as shown in Section 4. Thus, OUG ZF-THP approach represents a good trade-off between computational complexity and MU-MIMO system performance in terms of capacity.

In order to give further performance comparison for the observed algorithms which comprise user grouping and precoding procedures, we considered the average uncoded bit error rate (BER) as the performance metric (i.e., achieved BER prior to forward error correction decoding at the receiver), where averaging is performed over a sufficient number of channel realizations and over all users. The uncoded BER is calculated as in [16], in which the upper bound for symbol error rate (SER) in the case of different precoding techniques is given, with the additional averaging realized over all users for all beamforming subgroups. The use of Gray coding for the adopted 16-QAM modulation technique is presumed for all schemes. It should be noticed that these BER values are derived in [16], under the assumption that transmitter for each user in each subgroup essentially fixes the minimum required SNR ($\gamma_{\min}$) for which it encodes data at the rate $R = \log_2(1 + \gamma_{\min})$ corresponding to the possible capacity. However, if the actual SNR value is smaller than $\gamma_{\min}$, decoding errors occur with the probability $p_{\text{outage}} = P_r(\gamma < \gamma_{\min})$ [16]. Thus, given BER represents the upper bound for the observed scenario in which the ideal CSI data is used.

The comparison of the estimated average uncoded BER for all the observed algorithms is presented in Figures 7 and 8, in the case of the environment with correlated and uncorrelated Rayleigh fading, respectively. As obviously shown in Figures 7 and 8, the algorithms which deploy more complex nonlinear precoding (i.e., K-means BD-THP, OUG ZH-THP, and OUG THP) significantly outperform those with linear ZF precoding (i.e., ZFS and OUG-Greedy). Such behaviour is expected due to more successful mitigation of interuser interference with nonlinear precoding techniques, as was already shown in the literature [20]. Also, much better BER performance is achieved for all the observed algorithms in the case of uncorrelated MIMO channels, due to the significantly lower interuser interference. The best average uncoded BER is achieved in the case of OUG THP algorithm, while proposed OUG ZF-THP algorithms have somewhat higher BER in the case of uncorrelated MIMO channels, and essentially same BER as OUG THP algorithm in the case of correlated MIMO channels.

Previous findings are summarized in Table 5 where numerical performance of all the observed algorithms in good radio conditions (i.e., 20 dB in the case of achievable sum rate and 40 dB in the case of BER analysis) is presented.

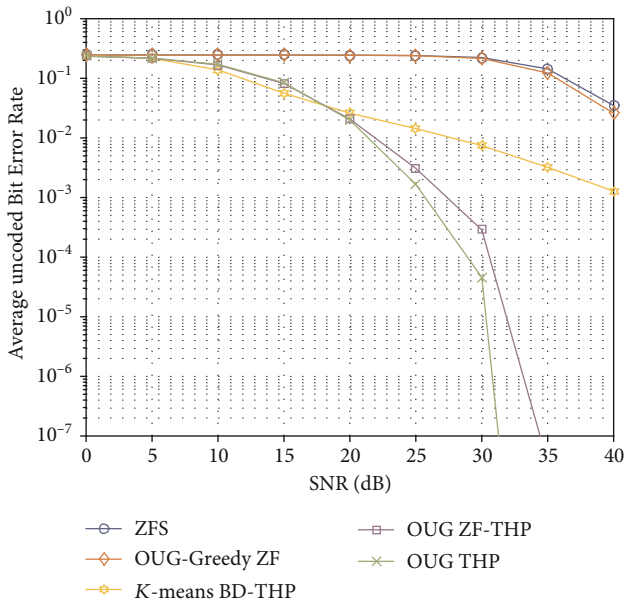

Figure 7: Average uncoded BER comparison under uncorrelated MIMO channels.

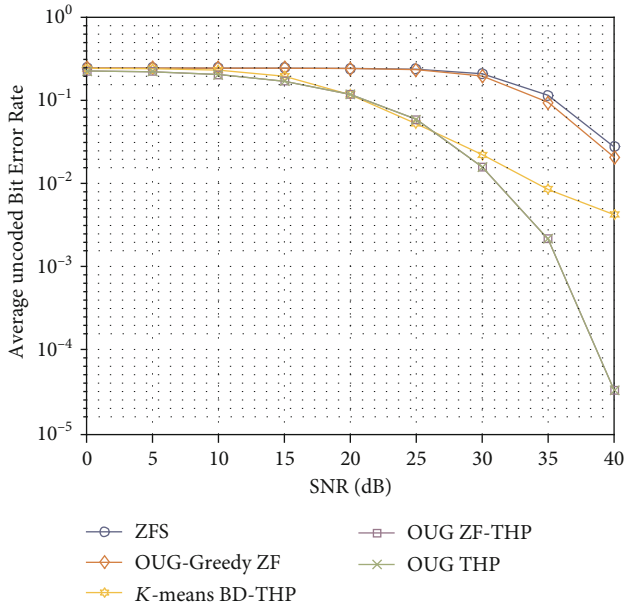

Figure 8: Average uncoded BER comparison under correlated MIMO channels.

When the number of users K increases for the same number of base station antennas N, it was shown that ZF-based user grouping strategies achieve the sum rate which approaches the capacity upper bound [4]. This comes from the fact that more combinations within the search space $\sum_{k=1}^{N}\binom{K}{k}$ are covered. Asymptotically, when user number goes to infinity, the optimal sum rate is achieved since searching among all possible combinations is done. Previous findings apply to our case as well with additional benefit from the introduction of overlapping user grouping approach. In this approach, even more users can be added in unpopulated beamforming subgroups since probability that users' channels are spatially uncorrelated increases. However, this leads to the increased number of beamforming subgroups that should be precoded, and hence, computational complexity becomes prohibitively high because of its polynomial relation with the number of users and their subgroups. To overcome this challenge, we can increase the number of base station antennas. In that way, the same number of users will be served and selected in the lower number of beamforming subgroups keeping the complexity of signal processing on reasonable level. In that case, we could exploit both beamforming and multiplexing gains. However, the increase in the number of base station antennas results with the increased hardware complexity on base station side, especially in the observed 5G NR midband where digital beamforming is envisioned. The trade-off between these two approaches is to choose reasonably a large number of base station antennas and numerous users in the cell but to keep the ratio between them relatively small. The last statement holds in the case of IoT-oriented MU-MIMO system with numerous IoT devices served in 5G cell unlike in massive MIMO case where the number of base station antennas is typically much larger than the number of users [28].

Table 5: Performance of all observed algorithms in good radio conditions.

Algorithm	Achievable sum rate (bps/Hz)	Average BER
ZFS	650	$2.9 * 10^{-2}$
OUG-Greedy ZF	825	$2.1 * 10^{-2}$
K-means BD-THP	710	$4.5 * 10^{-3}$
OUG ZF-THP	1030	$3.5 * 10^{-5}$
OUG THP	1100	$3.5 * 10^{-5}$

6. Conclusions

In this paper, we have studied user grouping and scheduling problem in IoT-oriented 5G MU-MIMO systems. We have proposed two-stage hybrid precoding scheme based on overlapping user grouping strategy for mMTC 5G NR use case. In this framework, user grouping is performed using the greedy approach that allows users with favorable channel conditions to be scheduled into the multiple beamforming subgroups simultaneously. Two-stage hybrid precoding scheme is then applied on created beamforming subgroups in order to minimize the interference in MU-MIMO system. Linear ZF precoding cancels interference among beamforming subgroups while the nonlinear THP precoding reduces remaining interference between scheduled users within each beamforming subgroup. Comparative analysis with other precoding schemes based on different user grouping methods has been presented. Numerical results demonstrate that proposed algorithm achieves much higher MU-MIMO system capacity in comparison to the existing two-stage precoding schemes based on legacy user grouping strategies, especially in large SNR regime (from 30% at 4 dB to 45% at 20 dB). Also, thorough complexity analysis has shown that despite its good

throughput performance, the proposed approach has lower computational complexity as the existing algorithms that employ user grouping methods and two-stage precoding schemes. Also, the proposed OUG ZF-THP algorithm achieves very good BER performance in the observed application scenario.

Obtained numerical results encourage further research in the area of user grouping and scheduling in 5G MU-MIMO systems. Future work will include evaluation of the proposed two-stage precoding based on overlapping user grouping approach in heterogeneous 5G network consisting of both IoT devices and legacy users with different quality of service (QoS) requirements and assessment of its performance in more realistic radio environment which imposes channel imperfections. In order to support given QoS requirements for the observed users, a deployment of adaptive modulation mechanism might be necessary. In that case, the low complexity VP precoding techniques could be observed as a promising solution, instead of here considered THP schemes.

Data Availability

The data generated from Monte-Carlo simulations to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

References

[1] A. Zaidi, A. Branneby, A. Nazari, M. Hogan, and C. Kuhlins, *Cellular IoT in the 5G Era*, Ericsson White paper, 2020.

[2] T. van Chien and E. Björnson, *Massive MIMO Communications, 5G Mobile Communications*, Springer, Basel, Switzerland, 2017.

[3] E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 1, pp. 239–284, 2017.

[4] J. Wang, D. J. Love, and M. D. Zoltowski, "User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3713–3726, 2008.

[5] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[6] N. Ganganath, C.-T. Cheng, and C. K. Tse, "Data clustering with cluster size constraints using a modified k-means algorithm," in *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 158–161, Shanghai, China, 2014.

[7] G. Dimic and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[8] R. Tian, Y. Liang, X. Tan, and T. Li, "Overlapping user grouping in IoT oriented massive MIMO systems," *IEEE Access*, vol. 5, pp. 14177–14186, 2017.

[9] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, pp. 76–80, 2014.

[10] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[11] M. H. M. Costa, "Writing on dirty paper (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[12] R. F. H. Fischer, C. Windpassinger, A. Lampe, and J. B. Huber, "Space-time transmission using Tomlinson-Harashima precoding," in *Proc. ITG SCC*, pp. 139–147, 2002.

[13] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part II: perturbation," *IEEE Transactions on Communications*, vol. 53, no. 3, pp. 537–544, 2005.

[14] R. Chen, C. Li, J. Li, and Y. Zhang, "Low complexity user grouping vector perturbation," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 189–192, 2012.

[15] A. Li and C. Masouros, "A constellation scaling approach to vector perturbation for adaptive modulation in MU-MIMO," *IEEE Wireless Communications Letters*, vol. 4, no. 3, pp. 289–292, 2015.

[16] S. Zarei, W. Gerstacker, and R. Schober, "Comparison of lattice-reduction-aided vector perturbation and Tomlinson-Harashima Precoding," in *Proc. 2019 IEEE WCNC*, Marrakech, Morocco, 2019.

[17] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[18] R1-1701678, "Non-linear precoding for downlink multiuser MIMO," Huawei, Hisilicon, 3GPP TSG RAN WG1 Meeting 88, Athens, Greece, 2017.

[19] R1-1703187, "On MU MIMO non-linear precoding in NR," Nokia, Alcatel-Lucent Shanghai Bell, 3GPP TSG RAN WG1 Meeting 88, Athens, Greece, 2017.

[20] S. Zarei, W. Gerstacker, and R. Schober, "Low complexity hybrid linear/Tomlinson–Harashima precoding for downlink large-scale MU-MIMO systems," in *Proc. 2016 IEEE Globecom Workshops*, Washington DC, USA, 2016.

[21] R. F. Trifan, A. A. Enescu, and C. Paleologu, "Hybrid MU-MIMO precoding based on K-means user clustering," *Algorithms*, vol. 12, no. 7, pp. 146–163, 2019.

[22] S. Sheikhzadeh, A. R. Forouzan, and F. Parvaresh, "Tomlinson–Harashima precoding for transmitter-side inter-symbol interference cancellation in PSK modulation," *IET Communications*, vol. 13, no. 5, pp. 610–619, 2019.

[23] W. C. Jakes, *Microwave Mobile Communications*, Wiley, New York, USA, 1974.

[24] A. Adhikary, Junyoung Nam, Jae-Young Ahn, and G. Caire, "Joint spatial division and multiplexing—the large-scale array regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[25] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, vol. 27, no. 23, pp. 2145-2146, 1991.

[26] P. Agrawal and N. Patwari, "Correlated link shadow fading in multi-hop wireless networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4024–4036, 2009.

[27] K. Yu and B. Ottersten, "Models for MIMO propagation channels: a review," *Wireless Communications and Mobile Computing*, vol. 2, no. 7, pp. 653–666, 2002.

[28] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*, Cambridge University Press, Cambridge, UK, 2016.

[29] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: a geometric approach and its application to solve generalized radio resource allocation problems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3637–3647, 2013.

[30] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

[31] R. Hunger, *Floating Point Operations in Matrix-Vector Calculus*, Technische Universität München, Associate Institute for Signal Processing, Tech. Rep. TUM-LNS-TR-05-05 Ver. 1.3, 2007.

[32] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658–3663, 2006.

[33] M. Arakawa, *Computational Workloads for Commonly Used Signal Processing Kernels*, MIT, Lincoln Laboratory, Tech. Rep. ESC-TR-2006-071, 2006.

[34] A. Garcia-Rodriguez and C. Masouros, "Power-efficient Tomlinson-Harashima precoding for the downlink of multi-user MISO systems," *IEEE Transactions on Communications*, vol. 62, no. 6, pp. 1884–1896, 2014.

[35] J. Lee, E. Tejedor, K. Ranta-aho et al., "Spectrum for 5G: global status, challenges, and enabling technologies," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 12–18, 2018.

[36] C. Bockelmann, N. Pratas, H. Nikopour et al., "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[37] J. Zhang, Z. Zheng, Y. Zhang, J. Xi, X. Zhao, and G. Gui, "3D MIMO for 5G NR: several observations from 32 to massive 256 antennas based on channel measurement," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 62–70, 2018.