Enhanced Multiuser Superposition Transmission through Structured Modulation

Dong Fang, Yu-Chih Huang, Giovanni Geraci, Zhiguo Ding, and Holger Claussen

***Abstract*—The 5G air interface, namely, dynamic multiple access (MA) based on multiuser superposition transmission (MUST) and orthogonal multiple access (OMA), may require complicated scheduling and heavy signaling overhead. To address these challenges, we propose a unified MA scheme for future cellular networks, which we refer to as *structured multiuser superposition transmission* (S-MUST). In S-MUST, we apply complex power allocation coefficients (CPACs) over multiuser legacy constellations to generate a composite constellation. In particular, the in-phase (I) and quadrature (Q) components of the legacy constellation of each user are separately multiplied by those of the CPACs. As such, the CPACs offer an extra degree of freedom for multiplexing users and guarantee fairness in symmetric broadcast channels. This new paradigm of superposition coding allows us to design IQ separation at the user side, which significantly reduces the decoding complexity without degrading performance. Hence, it supports low-complexity frequency-selective scheduling that does not entail dynamically switching between MUST and OMA. We further propose to quantize the CPACs into complex numbers where I and Q components of each quantized coefficient are primes, facilitating parallel interference cancellation at each user via modulo operations, last but not least, we generalize the design of S-MUST to exploit the capabilities of multiantenna base stations. The proposed S-MUST exhibits an improved user fairness with respect to conventional MUST (134% spectral efficiency enhancement) and a lower system complexity compared with dynamically alternating MUST and OMA.**

***Index Terms*—Multiuser superposition transmission, superposition coding, non-orthogonal multiple access.**

I. Introduction

The key performance indicators (KPI) of the fifth generation (5G) cellular networks include massive device connectivity, high data rates, ultra-high link reliability, and low energy consumption [3], [4]. To meet the asserted KPIs, new air interfaces which may include a new paradigm of multiple access (MA) are called for. Compared with orthogonal multiple access (OMA) schemes, non-orthogonal multiple access (NOMA) offers superior spectral efficiency with less processing overhead, and hence has received considerable attention [5]–[9]. In NOMA, lower and higher powers are allocated to nearer and farther users (UEs), respectively, enabling to schedule multiple users in the same physical resource such as time/frequency/code/space [10], [11]. At the farther user side, the desired signal is directly decoded by treating the nearer user's signal as noise. At the nearer user side, the farther user's signal is decoded, reconstructed, and subtracted from the received signal first; then, the desired signal is decoded.

Since the 87-th meeting of the 3rd Generation Partnership Project (3GPP), NOMA has been selected as a study item in long-term evolution (LTE) release 13 – termed multiuser superposition transmission (MUST) [12] – and further selected as a work item in LTE release 14. MUST is classified into 3 categories: (i) in Cat. 1, each user maps its data onto a component constellation which is adaptively allocated power levels based on the near-far geometry. The composite constellation employs non-Gray mapping; (ii) in Cat. 2, the composite constellation employs Gray mapping where the adaptive power allocation and legacy mapping of each user's data are jointly designed; (iii) in Cat. 3, the composite constellation retains the legacy uniform quadrature amplitude modulation (QAM) with Gray mapping and without adaptive power allocation for users. The nearer and farther users' data is protected by unequal error protection in terms of minimum Euclidean distances. The pros and cons of MUST Cat. 1-3 are summarized in Table I. As the nature of MUST is superposition coding, it only outperforms OMA in asymmetric Gaussian broadcast channels [13]. Indeed, in symmetric broadcast channels, MUST Cat. 1-3 struggles to provide good user fairness [14], [15]. For example, in the case of two users and in order to avoid overlapping points on the composite constellation, MUST may not be able to assign equally strong power to both users. As a result, in symmetric broadcast channels where both users should be served with equal transmission rates, the rate of one user may be higher than that of the other. In a practical design, it is the duty of the scheduler to pair users with a near-far geometry in order to retain the spectral efficiency gain of MUST [16]. However, such scheduler results in an excessively high complexity if the cellular network has a high user density and traffic demand, as the BS needs to exhaustively search through all users and pair those satisfying the condition of asymmetrical broadcast channel. While a frequency selective scheduler could allow the co-existence of OMA and MUST (dynamic MA), it would need to compare the proportional fairness (PF) metric of MUST and OMA and align the best transmission (Tx) mode across all sub-bands. Such high computational burden limits the use of MUST in massive connectivity scenarios. Motivated by the aforementioned challenges and practical issues of MUST, it is necessary to design an efficient downlink superposition transmission scheme that provides: (i) a unified air interface not requiring to dynamically switch between two MA schemes; (ii) a low complexity scheduler; and (iii) good user fairness in symmetric broadcast channels.

D. Fang was with Nokia Bell Labs, Ireland and now is with UnitedHealth Group. Y.-C. Huang is with the National Taipei University, Taiwan. G. Geraci was with Nokia Bell Labs and now is with Universitat Pompeu Fabra, Spain. Z. Ding is with University of Manchester, UK. H. Claussen is with Nokia Bell Labs, Ireland. This work was done when Dong Fang and Giovanni Geraci were in Nokia Bell Labs. The material contained in this paper has been in part presented at IEEE Globecom 2016 [1] and at IEEE VTC Spring 2017 [2].

TABLE I: Summary of pros and cons of 3 categories of MUST

	Pros	Cons
MUST Cat. 1	Amplitude-weighted superposition; high spectral efficiency	non-Gray labeled; cannot use legacy constellation at BS side
MUST Cat. 2	Amplitude-weighted superposition; high spectral efficiency,Gray labeled	cannot use legacy constellation at BS side
MUST Cat. 3	Bit-level superposition; high spectral efficiency,Gray labeled	no adaptive power allocation

This paper aims at overcoming the aforementioned limitations of MUST with a new *structured multiuser superposition transmission* (S-MUST) scheme. S-MUST employs complex power allocation coefficients (CPAC) over the in-phase (I) and quadrature (Q) components of the multiple users' legacy constellations to generate a composite constellation. As such, the CPACs offer an extra degree of freedom to guarantee user fairness in symmetric channels. The proposed S-MUST results in a unified air interface capable of replacing dynamic MA – i.e., the alternation of MUST and OMA – thus reducing the complexity of the frequency selective scheduler. We also quantize the CPACs into complex numbers where I and Q components of each CPAC are primes, enabling modulo operations based parallel interference cancellation (M-PIC) with respect to these primes, at UE side. Such M-PIC operation can be performed independently at each UE and irrespective of other users' network assistance information, such as modulation and coding scheme (MCS), power level, channel quality indicator (CQI), and precoding matrix index (PMI) – hence significantly reducing the signaling overhead. The main contributions of this paper are three-fold and can be summarized as follows:

- A new non-orthogonal multiuser superposition transmission scheme, S-MUST, is proposed.
- We provide composite constellation and mapping design for proposed S-MUST. A detection algorithm as well as the assignment of the complex power coefficients accounting for the user fairness optimization are devised.
- We design low-complexity frequency-selective scheduling and pairing algorithms for S-MUST.
- We extend the design of S-MUST to exploit the capabilities of multiantenna base stations (BSs), through a framework based on user selection, clustering, and zero forcing beamforming.

The structure of this paper is as follows: 1) the challenges of existing schemes and the motivations of our design are introduced in Section II; 2) the detailed design is shown in Section IV, including transmission and reception; power allocation; user fairness protection and scheduling; 3) the joint design of MIMO and S-MUST is discussed in Section IV; 4) the performance evaluation is provided in Section V; and 5) conclusive remarks are given in Section VI.

II. Challenges and Motivations

In this section, we present some preliminaries of conventional MUST and of dynamic MA. We then discuss the high scheduling complexity of existing schemes and the user fairness issue.

A. High scheduling complexity

In order to implement a dynamic MA scheme, a frequency-selective scheduler is required to select the best Tx mode – opportunistically alternating between MUST and OMA – for each UE across each sub-band [17], [18], based on a PF metric:

$$\mathrm{PF}_{\ell} = \sum_{\ell \in \mathcal{U}} \frac{R_{\ell}\left[t, \mathcal{U}\right]}{\bar{R}_{\ell}\left[t\right]}, \tag{1}$$

where $R_{\ell}\left[t, \mathcal{U}\right]$ denotes the instantaneous rate of UE ℓ at time t (the time index of a subframe); $\bar{R}_{\ell}\left[t\right]$ denotes the average rate of UE ℓ; and $\mathcal{U}$ is the set of UE indices. In the multiuser case, e.g., with two UEs, the PF metric, denoted by $\mathrm{PF}_{j,k}$ (where j and k are indices of paired UEs), can be calculated from the ratio of the paired UEs' instantaneous sum-rate over their average sum-rate. PMI and CQI feedback is required to evaluate the channel condition of each sub-band. The power coefficients to the farther UE, denoted by α, are determined in the MUST scheduling loop. This kind of scheduling is channel dependent, commonly used in cellular systems, and referred to frequency-selective scheduling. Instead of exploiting the frequency diversity of the channel, frequency-selective scheduling leverages the channels time and frequency selectivity to allocate valuable radio resources in an optimal manner. The main frequency selectivity scheduling operations of MUST are summarized in Algorithm 1 [17], [18]. In order to dynamically switch between MUST and OMA and retain the gain provided by MUST, the above scheduling algorithm requires to traverse all sub-bands several times to exhaustively search for the best Tx mode. The computational complexity thus increases exponentially with the number of sub-bands and UEs.

B. User fairness loss

In addition to the complexity of a frequency-selective scheduler, another drawback of MUST is the inability to guarantee user fairness in symmetric broadcast channels. In a superposition transmission scheme, user fairness is defined as the maximum rate of the weakest UE across all sets of paired UEs [14], given by:

$$\begin{aligned} &\max_{\alpha} \min_{i \in \mathcal{U}_s} R_{\ell}\left(\alpha\right) \\ &\text{s.t.} \sum_{\ell=1}^{L} \alpha_{\ell} \leq P, 0 \leq \alpha_{\ell}, \end{aligned} \tag{2}$$

where $R_\ell(\alpha)$ denotes the ℓ-th UE's rate, α denotes the power coefficient satisfying the power constraint, and $\mathcal{U}_s$ denotes the set of paired UEs.

In the following, we will discuss why standard MUST cannot guarantee user fairness by taking MUST Cat. 2 as an example. Fig. 1 illustrates the composite constellation of MUST Cat. 2., where there are far and near UEs, denoted by UE 1 and UE 2, respectively. Suppose both UEs adopt 4-ary constellation, namely, 2 bits/symbol rates. In MUST Cat. 2, the Gray mapped 16QAM is virtually treated as the superposition constellation so that each symbol on it can be treated as a superimposed symbol of both users. The first 2 bits are assigned to near UEs, marked by black. The last 2 bits are assigned to far UE, marked by red. As such, one can observe that the minimum Euclidean distance of far UE is larger than that of near UE. This indicates the far UE has higher error protection. In symmetric broadcast channels where both UEs should be served with the equal rates, in order to avoid an overlap on the composite constellation, two UEs will be assigned with different powers, i.e., different error protection in terms of minimum Euclidean distance. As a result, user fairness cannot be guaranteed, especially in low-to-moderate signal-to-noise ratio (SNR) regimes. As such, the scheduling algorithm of MUST attempts to pair users with asymmetric channels (see the condition $\mathrm{CQI}_j > \mathrm{CQI}_k$ in Algorithm 1, "UE pair selection for MUST").

Algorithm 1: Dynamic MA Frequency-Selective Scheduling [17], [18]

```
Given PMI and CQI feedback and the range of $\alpha$: $(0.025, 0.3]$;
Initialize the set of paired UEs $\mathcal{U}_s = \emptyset$;
*Single UE selection for OMA:*
for each sub-band do
    for each UE $i$ in the active UE set $\mathcal{U}$ do
        calculate $\mathrm{PF}_i$;
    end for
    $\hat{i} = \arg\max_{i \in \mathcal{U}} \{\mathrm{PF}_i\}$;
    $\mathcal{U}_s \leftarrow \mathcal{U}_s \cup \left(\hat{j}, \hat{k}\right)$
end for
*UE pair selection for MUST:*
for each sub-band do
    for each near-far UE pair $(\mathrm{UE}_j, \mathrm{UE}_k)$ in $\mathcal{U}$ do
        if $\mathrm{PMI}_j$=$\mathrm{PMI}_k$ and $\mathrm{CQI}_j > \mathrm{CQI}_k$ then;
            for all $\alpha$ do
                $\hat{\alpha} = \arg\max_{\alpha} \{\mathrm{PF}_{j,k}(\alpha)\}$;
                calculate $\mathrm{PF}_{j,k}(\hat{\alpha})$;
            end for
        else
            continue;
        end if
    end for
    $\left(\hat{j}, \hat{k}\right) = \arg\max_{i \in \mathcal{U}} \{\mathrm{PF}_{j,k}(\hat{\alpha})\}$;
    $\mathcal{U}_s \leftarrow \mathcal{U}_s \cup \left(\hat{j}, \hat{k}\right)$.
end for
```

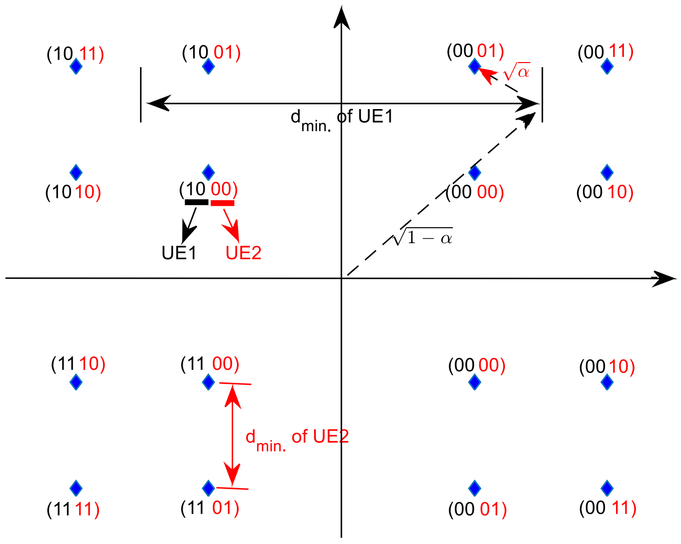


Fig. 1: Composite constellation of MUST Cat. 2 [12].

Algorithm 1: Dynamic MA Frequency-Selective Scheduling [17], [18] (continued)

```
*Tx mode selection:*
for each sub-band do
    if $\mathrm{PF}_{\hat{j},\hat{k}}(\hat{\alpha}) > \mathrm{PF}_{\hat{i}}$ then
        Tx mode=MUST;
    else
        Tx mode=OMA.
    end if
end for
*UE alignment and sub-band release:*
for each UE do
    if no. of MUST Tx. mode>no. of OMA Tx. mode then
        best Tx mode=MUST;
    else
        best Tx mode=OMA;
    end if
    Release sub-bands where selected UE is scheduled with other Tx mode than the best Tx mode. UE selected is such sub-bands must be scheduled with the best Tx mode in the next scheduling round.
end for
```

III. Detailed Design of S-MUST

In this section, we provide a detailed design for the proposed S-MUST scheme, including transmission design in subsection A, reception design in subsection B, power allocation in subsection C, user fairness in subsection D and scheduler design in subsection E. The main advantage of the proposed S-MUST over a hybrid MA scheme (OMA and MUST) is that S-MUST does not need to switch between OMA and MUST when the subband/subchannel is symmetric or not. This entails that, at the base station's side, the scheduler can be "dummy", assigning each physical resource block to multiple users without considering whether the subbands/subchannels are symmetric or not. In addition, S-MUST employs the same encoding-decoding mechanism to deal with both symmetric and asymmetric channels. In contrast, as shown in Algorithm 1, OMA+MUST requires to assign a whole piece of physical resource block to a single user when the corresponding subband/subchannel is symmetric, while assigning it to multiple users when the corresponding subband/subchannel is asymmetric. In addition, dynamic MA employs two different encoding-decoding mechanisms, one for symmetric channel, namely, OMA and the other for asymmetric channel, namely, MUST.

A. Transmission at the BS

In the proposed S-MUST, the superimposed signal can be generated from the following mapping function

$$x = \lambda \mathcal{S}(v_1, \ldots, v_L), \tag{3}$$

where $v_\ell, \ell \in \{1, ..., L\}$ denote the coded symbols of the ℓ-th UE, e.g., for a 2^m-ary modulation, $v_\ell \triangleq [v_{\ell,1}, ..., v_{\ell,m}]$ is an m-bit binary tuple where m is an integer and $v_{\ell,t}, t \in \{1, ..., m\}$ is the t-th bit of v_ℓ; λ is a scaling factor to meet the power constraint; and $\mathcal{S}$ denotes the mapping function. In the following, we will describe three categories of S-MUST, and we will employ $\mathcal{S}_{\text{Cat.1}}$, $\mathcal{S}_{\text{Cat.2}}$ and $\mathcal{S}_{\text{Cat.3}}$ to denote the corresponding mapping functions.

1) S-MUST Cat. 1: The mapping function is defined as:

$$\mathcal{S}_{\text{Cat.1}}(v_1, ..., v_L) \triangleq \sum_{\ell=1}^{L} \alpha_\ell \mathsf{I}(\mathcal{M}_\ell(v_\ell)) + j \sum_{\ell=1}^{L} \beta_\ell \mathsf{Q}(\mathcal{M}_\ell(v_\ell)), \tag{4}$$

where $\mathcal{M}_\ell(\cdot)$ denotes the legacy modulation mapper for each user, e.g., 16-QAM; $\mathsf{I}(\cdot)$ and $\mathsf{Q}(\cdot)$ represent the I and Q separation; α_ℓ and β_ℓ denote the I and Q components of the CPAC, respectively, which satisfy the power constraint

$$\mathbb{E}\left[\left|\sum_{\ell=1}^{L} \alpha_\ell \mathsf{I}(\mathcal{M}_\ell(v_\ell)) + j \sum_{\ell=1}^{L} \beta_\ell \mathsf{Q}(\mathcal{M}_\ell(v_\ell))\right|^2\right] = \sum_{\ell=1}^{L} \alpha_\ell^2 + \beta_\ell^2 \leq P. \tag{5}$$

A systematic illustration of S-MUST Cat. 1 is shown in Fig. 2, where transmission block (TB), i.e., data stream, is encoded by forward error correction (FEC) codes and then mapped into legacy constellation. The IQ separation splits I and Q data streams and formulates the mapping function as in (4).

2) S-MUST Cat. 2: The mapping function is defined as:

$$\mathcal{S}_{\text{Cat.2}}(v_1, \ldots, v_L) \triangleq \mathcal{G}(\mathcal{S}_{\text{Cat.1}}(v_1, \ldots, v_L)), \tag{6}$$

where $\mathcal{G}(\cdot)$ denotes the permutation of Gray labeling.

Before we introduce S-MUST Cat. 3, we include some algebra preliminaries as the prelude to S-MUST Cat. 3's mapping function.

Definition 1: (Square-free Integers) An integer is said to be square-free if its prime factorization contains no repeated factors.

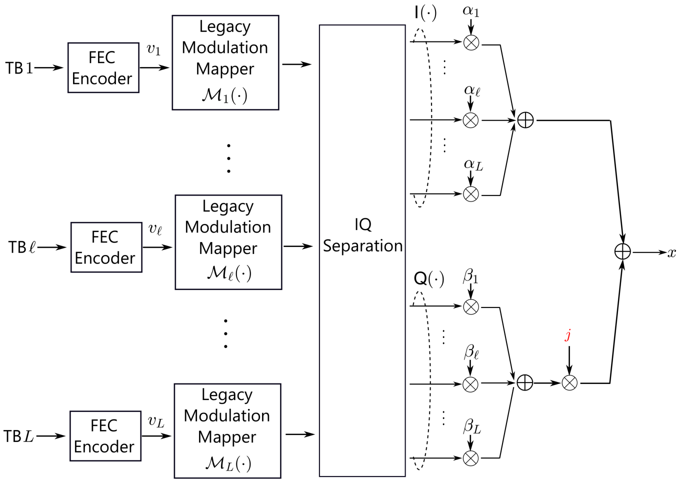


Fig. 2: Schematic illustration of the proposed S-MUST Cat. 1.

Definition 2: (Modulo operations) The notation $x \bmod a$ denotes reducing $x \in \mathbb{Z}$ modulo the integer interval $[-a, a)$. That is,

$$x \bmod a = x - b \cdot [a - (-a)],$$

where $b \in \mathbb{Z}$ is the (unique) integer such that

$$x - b \cdot [a - (-a)] \in [-a, a).$$

Lemma 1: (Chinese Remainder Theorem (CRT) [19]). Let $p_1, ..., p_n$ be relatively prime numbers. For $v_\ell \in \mathbb{Z}_{p_\ell}$, $\ell \in \{1, ..., L\}$, there exists a ring isomorphism [19]:

$$\mathcal{W}(v_1, ..., v_L) = \left(s_1 \cdot v_1 \cdot \prod_{\ell \neq 1} p_\ell + \ldots + s_L \cdot v_L \cdot \prod_{\ell \neq L} p_\ell\right) \mod \prod_\ell p_\ell, \tag{7}$$

where $s_1, \ldots, s_L \in \mathbb{Z}$ are such that

$$\mathcal{W}(v_1, \ldots, v_L) \mod p_\ell = v_\ell. \tag{8}$$

We note that $s_1, \ldots, s_L$ can be easily obtained by solving the Bezout's identity and are solely for (8) to hold. For the application to be discussed later, asking (8) may be too much as long as there exists a one-to-one mapping so that v_ℓ can be easily obtained from a simple modulo operation. One such mapping can be obtained by removing $s_1, \ldots, s_L$ to get

$$\mathcal{W}(v_1, \ldots, v_L) = \left(v_1 \cdot \prod_{\ell \neq 1} p_\ell + \ldots + v_L \cdot \prod_{\ell \neq L} p_\ell\right) \mod \prod_\ell p_\ell. \tag{9}$$

We note that the first term inside (9), $v_1 \cdot \prod_{\ell \neq 1} p_\ell$, has every primes except for p_1 as its factors (note that $v_1 \in \mathbb{Z}_{p_1}$ so cannot be a factor of p_1). Moreover, every other term in (9) has p_1 as its factor. Hence, after $\mod p_1$ operation, only $v_1 \cdot (\prod_{\ell \neq 1} p_\ell) \mod p_1$ remains. Similar reasoning shows leads to

$$\mathcal{W}(v_1, \ldots, v_L) \mod p_\ell = a_\ell \cdot v_\ell \mod p_\ell, \tag{10}$$

where $a_\ell = \prod_{\ell' \neq \ell} p_{\ell'} \mod p_\ell$ is independent of v_ℓ. We would like to emphasize that removing $s_1, \ldots, s_L$ allows us to circumvent the complexity of solving Bezout's identity at the transmitter. The price is that each receiver ℓ now has to compute a_ℓ, which can be done quite easily.

In this category, we adopt legacy 2^m-ary QAM; hence, both the I and Q components become $2^{m/2}$-ary pulse amplitude modulation (PAM). Then α_ℓ and β_ℓ in S-MUST Cat. 1 are quantized into square-free integers $\hat{\alpha}_\ell$ and $\hat{\beta}_\ell$ which can be factorize into $\hat{\alpha}_\ell = \Pi_{\ell'=1,\ell'\neq\ell}^{L} q_{\ell'}$ and $\hat{\beta}_\ell = \Pi_{\ell'=1,\ell'\neq\ell}^{L} p_{\ell'}$, respectively, where $p_\ell > 2^{m/2}$ and $q_\ell > 2^{m/2}$ for $\ell \in \{1, \ldots, L\}$. We then apply the mapping inspired by CRT in (9) to get

$$\begin{aligned}
&\mathcal{S}_{\text{Cat.3}}\left(v_1, ..., v_L\right) \triangleq \\
&\left[\sum_{\ell=1}^{L}\left(\mathsf{I}\left(\mathcal{M}_\ell\left(v_\ell\right)\right)\cdot \prod_{\ell'=1,\ell'\neq\ell}^{L} q_{\ell'}\right)\right] \bmod \prod_{\ell=1}^{L} q_\ell \\
&+j\left[\sum_{\ell=1}^{L}\left(\mathsf{Q}\left(\mathcal{M}_\ell\left(v_\ell\right)\right)\cdot \prod_{\ell'=1,\ell'\neq\ell}^{L} p_{\ell'}\right)\right] \bmod \prod_{\ell=1}^{L} p_\ell,
\end{aligned} \tag{11}$$

Based on (11), a systematic illustration of S-MUST Cat. 3 is shown in Fig. 3. The benefit of using this kind of mapping is that M-PIC is feasible at the UE side, which enjoys low system complexity and less overhead. More details are provided in the following subsection B 3).

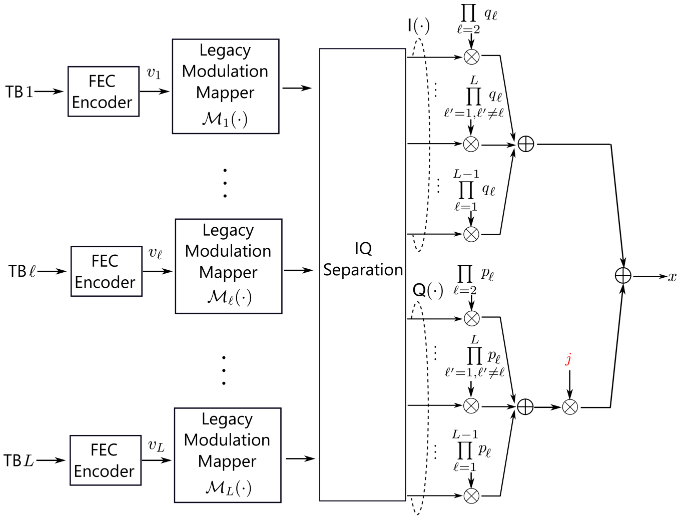

TB 1
FEC Encoder
v_1
Legacy Modulation Mapper $\mathcal{M}_1(\cdot)$
TB ℓ
FEC Encoder
v_ℓ
Legacy Modulation Mapper $\mathcal{M}_\ell(\cdot)$
TB L
FEC Encoder
v_L
Legacy Modulation Mapper $\mathcal{M}_L(\cdot)$
IQ Separation
$\mathsf{I}(\cdot)$
$\prod_{\ell=2} q_\ell$
$\prod_{\ell'=1,\ell'\neq\ell}^{L} q_\ell$
$\prod_{\ell=1}^{L-1} q_\ell$
$\mathsf{Q}(\cdot)$
$\prod_{\ell=2} p_\ell$
$\prod_{\ell'=1,\ell'\neq\ell}^{L} p_\ell$
$\prod_{\ell=1}^{L-1} p_\ell$
j
x

Fig. 3: Schematic illustration of the proposed S-MUST Cat. 3.

B. Reception at the UEs

Let us consider a BS serving L UEs. The superimposed signal at UE ℓ can be written as

$$y_\ell = h_\ell x + n_\ell, \tag{12}$$

where x is the superposition codeword transmitted by the BS with power P; h_ℓ is the channel coefficient from the BS to UE ℓ; and n_ℓ is Gaussian noise with zero mean and variance $2\sigma^2$ per complex dimension; the transmit signal-to-noise ratio at UE ℓ is given by

$$\mathsf{SNR}_\ell = \frac{P\left|h_\ell\right|}{2\sigma^2}, \forall \ell \in \{1, ..., L\}. \tag{13}$$

1) Detection for S-MUST Cat. 1 and Cat. 2: By applying channel compensation over the received signal, one can obtain

$$\tilde{y}_\ell = \frac{y_\ell}{h_\ell} = x + n_{\ell,\text{eqv}}, \tag{14}$$

where $n_{\ell,\text{eqv}} \triangleq n_\ell / h_\ell$ is the equivalent noise with variance $2\sigma^2_{\text{eqv},\ell} = 2\sigma^2 / \left|h_\ell\right|^2$.

For simplicity of notation, let $x_\ell \triangleq \mathcal{M}_\ell\left(v_\ell\right)$, $x_{\ell,\mathsf{I}} \triangleq \mathsf{I}\left(x_\ell\right)$ and $x_{\ell,\mathsf{Q}} \triangleq \mathsf{Q}\left(x_\ell\right)$, $\ell \in \{1, ..., L\}$, denote the modulated signal, and its I and Q components, respectively, and let $\tilde{y}_{\ell,\mathsf{I}} \triangleq \mathsf{I}\left(\tilde{y}_\ell\right)$ and $\tilde{y}_{\ell,\mathsf{Q}} \triangleq \mathsf{Q}\left(\tilde{y}_\ell\right)$ denote the I and Q components of the received signal, respectively, which can be obtained through IQ separation at the UE. Without loss of generality, let us assume the following channel gain ordering: $|h_1| \leq |h_2| \leq \ldots \leq |h_L|$. As such, the ℓ-th UE, $\forall \ell \in \{2, ..., L\}$ can apply successive interference cancellation (SIC) from the code level of UE 1 to its own level. Taking the detection of the I component as an example, SIC can be implemented through multistage decoding as follows:

$$\begin{aligned}
\hat{x}_{1,\mathsf{I}} &\approx \underset{x_{1,\mathsf{I}} \in \mathsf{I}(\mathcal{M}_1(\mathbb{Z}_{2^m}))}{\arg\max} \left|\tilde{y}_{1,\mathsf{I}} - x_{1,\mathsf{I}}\right|^2 \\
&\vdots \\
\hat{x}_{\ell,\mathsf{I}} &\approx \underset{x_{\ell,\mathsf{I}} \in \mathsf{I}(\mathcal{M}_\ell(\mathbb{Z}_{2^m}))}{\arg\max} \left|\tilde{y}_{\ell,\mathsf{I}} - x_{\ell,\mathsf{I}} - \sum_{\ell'=1}^{\ell-1} \hat{x}_{\ell',\mathsf{I}}\right|^2,
\end{aligned} \tag{15}$$

where $\hat{x}_{1,\mathsf{I}}, ..., \hat{x}_{\ell,\mathsf{I}}$ are the recovered I components of the modulated signals. Detection of the Q component can be performed in a similar fashion.

Based on the decoding metric in (15), the log-likelihood ratios (LLRs) for the t-th bit of $v_1, ..., v_\ell$ can be represented as follows:

$$\begin{aligned}
\mathsf{LLR}(v_{1,t}) &\approx \log \frac{\sum\limits_{v_{1,t}=0} \exp\left(\frac{1}{\sigma^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I}} - x_{1,\mathsf{I}}\right|^2\right)}{\sum\limits_{v_{1,t}=1} \exp\left(\frac{1}{\sigma^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I}} - x_{1,\mathsf{I}}\right|^2\right)}, \\
&\vdots \\
\mathsf{LLR}(v_{\ell,t}) &\approx \log \frac{\sum\limits_{v_{\ell,t}=0} \exp\left(\frac{1}{\sigma^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I}} - x_{\ell,\mathsf{I}} - \sum\limits_{\ell'=1}^{\ell-1} \hat{x}_{\ell',\mathsf{I}}\right|^2\right)}{\sum\limits_{v_{\ell,t}=1} \exp\left(\frac{1}{\sigma^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I}} - x_{\ell,\mathsf{I}} - \sum\limits_{\ell'=1}^{\ell-1} \hat{x}_{\ell',\mathsf{I}}\right|^2\right)},
\end{aligned} \tag{16}$$

where $\mathsf{LLR}(v_{\ell,t})$ is fed to the channel decoder to recover the useful signal.

2) Detection for S-MUST Cat. 3: The detection method of S-MUST Cat. 3 is different than that of S-MUST Cat. 1 and Cat. 2 and based on M-PIC. Let us take UE ℓ as an example: as illustrated in Fig. 4, due to the property of CRT described in (10), the ℓ-th code level can be *peeled off* via a modulo operation with respect to θ_ℓ, $\forall \ell \in \{1, ..., L\}$, given by

$$\begin{aligned}
\tilde{y}_{\ell,\mathsf{I},\text{mod}} &= \left[\mathsf{I}\left(\tilde{y}_\ell\right)\right] \bmod q_\ell, \\
\tilde{y}_{\ell,\mathsf{Q},\text{mod}} &= \left[\mathsf{Q}\left(\tilde{y}_\ell\right)\right] \bmod p_\ell,
\end{aligned} \tag{17}$$

where $\tilde{y}_{\ell,\mathsf{I},\text{mod}}$ and $\tilde{y}_{\ell,\mathsf{Q},\text{mod}}$ denote the I and Q components of the received signal after the modulo operation, which are fed to the following metric to calculate the bit-wise LLR:

$$\begin{aligned}\mathsf{LLR}(v_{\ell,t}) &= \log \frac{\sum\limits_{v_{\ell,t}=0} p\left(\tilde{y}_{\ell,\mathsf{I},\text{mod}}|x_{\ell,\mathsf{I}}\right)}{\sum\limits_{v_{\ell,t}=0} p\left(\tilde{y}_{\ell,\mathsf{I},\text{mod}}|x_{\ell,\mathsf{I}}\right)} \\ &= \log \frac{\sum\limits_{v_{\ell,t}=0} \exp\left(\frac{1}{\tilde{\sigma}^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I},\text{mod}} - x_{\ell,\mathsf{I}}\right|^2\right)}{\sum\limits_{v_{\ell,t}=0} \exp\left(\frac{1}{\tilde{\sigma}^2_{\text{eqv}}}\left|\tilde{y}_{\ell,\mathsf{I},\text{mod}} - x_{\ell,\mathsf{I}}\right|^2\right)},\end{aligned} \tag{18}$$

where $\tilde{\sigma}^2_{\text{eqv}}$ denotes the variance per real dimension of the noise folded through the modulo operation. In high-SNR regime, this can be approximated by the noise variance before the modulo operation. One can observe from (18) that no SIC decoding is needed, and each user only extracts its desired signals without requiring knowledge of other users' MCS, power level, CQI, and PMI. This proposed M-PIC approach thus significantly reduces the signaling overhead.

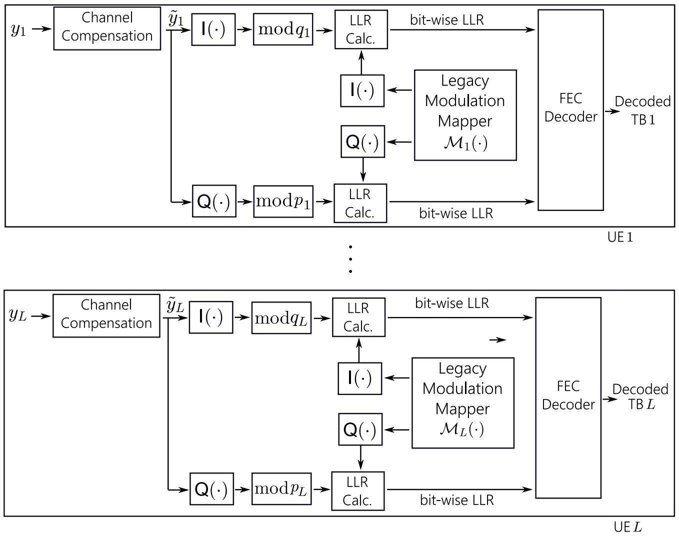


Fig. 4: M-PIC decoding for S-MUST Cat. 3.

C. Power Coefficient Selection

The selection of the CPACs is crucial for the system performance and to guarantee user fairness. Indeed, the power coefficients should be such that no overlapped points occur on the composite constellation. While the criterion for selecting the power coefficients in [20] is based on maximizing the minimum Euclidean distance, such approach cannot guarantee optimal user fairness. In contrast, we propose to select the CPACs according to the following maximum-fairness criterion:

$$\begin{aligned}&\max_{\alpha,\beta}\ \min_{\ell\in\{1,\dots,L\}} I\left(Y_\ell; X_\ell|X_{\ell-1},\dots,X_1\right) \\ &\text{s.t.}\ \sum_{\ell=1}^{L} \alpha_\ell^2 + \beta_\ell^2 \le P, \\ &\qquad \alpha_\ell \ge 0, \beta_\ell \ge 0\end{aligned} \tag{19}$$

where $I\left(Y_\ell; X_\ell|X_{\ell-1},\dots,X_1\right)$ is the mutual information between received and transmitted signals of the ℓ-th UE, given that all signals up to $\ell-1$-th have been successfully decoded. One can compute $I\left(Y_\ell; X_\ell|X_{\ell-1},\dots,X_1\right)$ through the chain rule as follows [21]

$$\begin{aligned}I\left(Y_\ell; X_\ell,\dots,X_1\right) =& I\left(Y_\ell; X_1\right) \\ &+ I\left(Y_\ell; X_2|X_1\right) \\ &+ I\left(Y_\ell; X_\ell|X_{\ell-1},\dots,X_1\right).\end{aligned} \tag{20}$$

Even though the IQ separation decoding is implemented over the I and Q components separately, the mutual information $I\left(Y_\ell; X_\ell,\dots,X_1\right)$ takes both the I and Q components into account such that the two degrees of freedom can be jointly exploited.

As SNR/CQI is the feedback usually adopted in current standardization, one can employ a look-up-table based on the broadcasted SNR/CQIs – as in equation (17) – to select the appropriate q_ℓ. Similar to the method of creating a look-up table (LUT) to select the appropriate modulation and coding scheme in LTE as a function of the SNR, the optimized pair $(\tilde{\alpha}_\ell, \tilde{\beta}_\ell)$ can be stored in an L-dimensional LUT at the BS, where each cell corresponds to an unique vector $[\mathsf{SNR}_1,\dots,\mathsf{SNR}_L]$. Given the feedback $\mathsf{SNR}_\ell, \forall\ell \in \{1,\dots,l\}$, a BS can select the optimal $(\tilde{\alpha}_\ell, \tilde{\beta}_\ell)$ pair to perform S-MUST transmissions as illustrated in Fig. 5. As a lightweight solution, in S-MUST Cat. 3 the product of primes can be quantized from the optimized $(\tilde{\alpha}_\ell, \tilde{\beta}_\ell)$ pair. One can adopt the PFA algorithm to find the distinct primes $\tilde{q}_\ell$ and $\tilde{p}_\ell$ on the I and Q components, respectively. Said computations can be performed offline, and one can construct a similar L-dimensional LUT to obtain the pair (p_ℓ, q_ℓ) based on the feedback $[\mathsf{SNR}_1,\dots,\mathsf{SNR}_L]$.

D. User Fairness in Symmetric Channels

The proposed S-MUST is able to multiplex users via superposition transmission without sacrificing user fairness even when they experience similar channel conditions. An example is given as follows, whereas numerical results will be provided in Section V.

Example: Let us consider two UEs, UE 1 and UE 2, both experiencing similar channel conditions, i.e., $\mathsf{SNR}_1 \approx \mathsf{SNR}_2$, and let us assume that both UEs adopt QPSK.Here is an example: the channel gains are sampled from Rayleigh fading symmetric broadcast channel so that S-MUST gets the CPACs $\alpha_1 = 2.3$, $\alpha_2 = 3.11$, $\beta_1 = 3.01$, and $\beta_2 = 2.18$ using (19) to construct S-MUST Cat. 1 and Cat. 2. Then one can quantize said CPACs into $q_1 = 2$, $q_2 = 3$, $p_1 = 3$, and $p_2 = 2$, obtaining a composite constellation for S-MUST Cat. 3 as the one illustrated in Fig. 6. As S-MUST Cat. 3 adopts IQ separation and M-PIC detection, let $d_{\min,\mathsf{I},\ell}$ and $d_{\min,\mathsf{Q},\ell}$, $\ell \in \{1,2\}$ denote the minimum Euclidean distances of UE ℓ on the I and Q components of the composite constellation, respectively. From Fig. 6, we can observe that both UEs have equal error protection in items of Euclidean distances and hence user fairness is guaranteed.

Remark: In the above Example, due to the legacy QPSK constellation mapper, the I component of the composite constellation can be alternatively represented as $\{-5,-1,1,5\}$

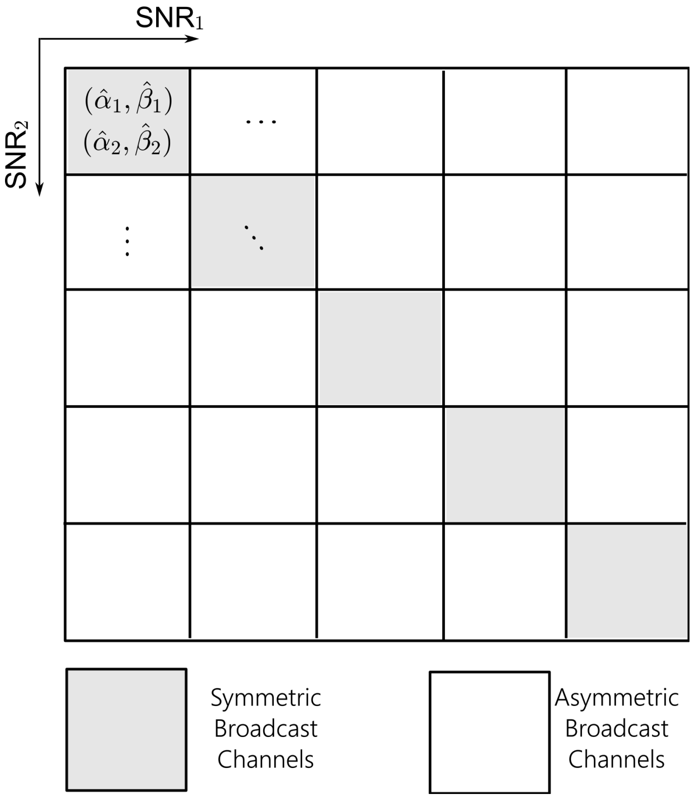

SNR$_1$
SNR$_2$
$(\hat{\alpha}_1, \hat{\beta}_1)$
$(\hat{\alpha}_2, \hat{\beta}_2)$
$\cdots$
$\vdots$
$\ddots$
Symmetric Broadcast Channels
Asymmetric Broadcast Channels

Fig. 5: Example of LUT for power coefficient selection in the case of two users.

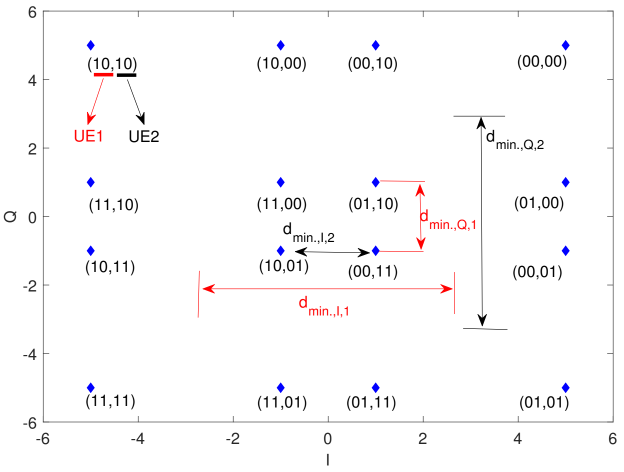


Fig. 6: Example of composite constellation when both UEs adopt QPSK in a symmetric broadcast channel.

and { (1,1), (1,0), (0,1), (0,0) }, where in the latter representation the first and second bits of each pair correspond to UE 1 and UE 2, respectively. Simply applying the modulo operation $\bmod 3$ – as per Definition 10 – at UE 1 over $\{-5, -1, 1, 5\}$ has the desired effect of canceling out the component of UE 2.

E. Proposed Scheduling Algorithm for S-MUST

Unlike dynamic MA, which opportunistically switches between MUST and OMA, S-MUST is able to provide a unified downlink MA air interface. The latter can significantly reduce the complexity of the PF scheduling operations compared to dynamic MA. Our proposed scheduling algorithm for S-MUST is provided in Algorithm 2.

Algorithm 2: Proposed Scheduling Algorithm for S-MUST (Two UEs)

```
Given the PMI and CQI feedback and the 2-D LUT for optimal CPACs $\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right]$;
*UE pair selection for S-MUST:*
**for** each sub-band **do**
    **for** each UE pair $(\text{UE}_j, \text{UE}_k)$ in $\mathcal{U}$ **do**
        **if** $\text{PMI}_j$=$\text{PMI}_k$ **then**;
            $\text{SNR}_j = \frac{P \cdot \text{CQI}_j}{2\sigma^2}$; $\text{SNR}_k = \frac{P \cdot \text{CQI}_k}{2\sigma^2}$;
            $\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right] = \text{LUT}\left(\text{SNR}_j, \text{SNR}_k\right)$;
            calculate the $\text{PF}_{j,k}\left(\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right]\right)$;
        **else**
            continue;
        **end if**
    **end for**
    $\left(\hat{j}, \hat{k}\right) = \arg\max_{j,k \in \mathcal{U}}\{\text{PF}_{j,k}\left(\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right]\right)\}$;
    $\mathcal{U}_s \leftarrow \mathcal{U}_s \cup \left(\hat{j}, \hat{k}\right)$.
**end for**
```

IV. Design of MIMO-based S-MUST

In this section, we discuss the joint design of MIMO and S-MUST, including the system model, user clustering, and beamforming design, as shown in subsections A, B, and C, respectively.

We extend the design of S-MUST to multi-antenna BSs, where the spatial degrees of freedom at each BS can be exploited to create several transmission beams, each carrying signals intended to multiple UEs. Such MIMO-based S-MUST design allows an N_t-antenna transmitter to serve $N_\text{c} \cdot M$ users on the same PRB. The proposed solution is based on user selection and clustering, zero forcing (ZF) beamforming, and S-MUST encoding/decoding, as illustrated by the flow chart in Fig. 7. The remainder of this section will provide a detailed description for each of these building blocks.

A. System Model for MIMO-based S-MUST

We consider multiuser MISO downlink, where the BS is equipped with N_t antennas, and K is the total number of single-antenna users in a cell. Knowledge of the channels

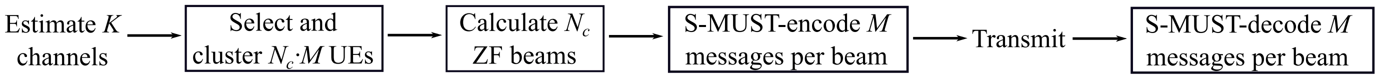

Estimate K channels
Select and cluster $N_c{\cdot}M$ UEs
Calculate N_c ZF beams
S-MUST-encode M messages per beam
Transmit
S-MUST-decode M messages per beam

Fig. 7: Flow chart of the proposed joint design of MIMO and S-MUST.

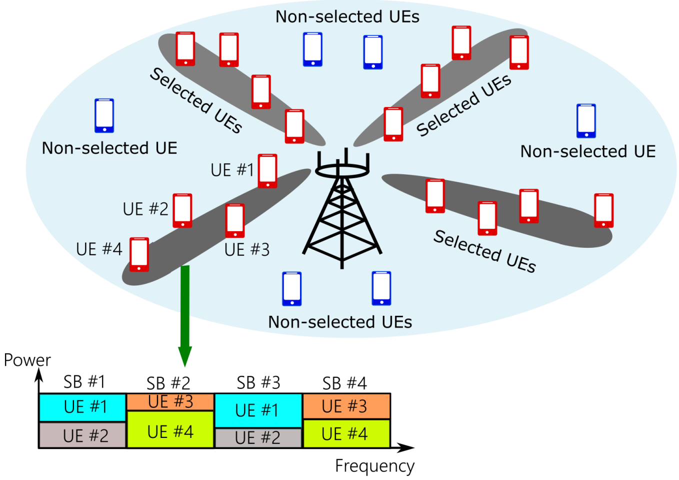


Fig. 8: Frequency-selective user selection, clustering, and beamforming in MIMO-based S-MUST.

to all K users is assumed to be available at the BS. This can be acquired through orthogonal uplink pilot symbols (for TDD systems) or downlink pilot symbols followed by uplink channel feedback (for FDD systems)[22].

Let $N_{\mathrm{c}} \cdot M = |\mathcal{U}|$ be the number of users scheduled for simultaneous transmission, which are divided into N_{c} groups or *clusters*, each containing $M \geq 2$ users. We denote by $\mathbf{y}_{n,m}$ the signal received by the m-th user in the n-th cluster, $m = 1, \ldots, M$, $n = 1, \ldots, N_{\mathrm{c}}$, given by

$$y_{n,m} = \left(\mathbf{h}^{\mathrm{H}}_{n,m}\mathbf{w}_n\right) x_n + \underbrace{\sum_{j\neq n}\left(\mathbf{h}^{\mathrm{H}}_{n,m}\mathbf{w}_j\right) x_j}_{e_{n,m}} + z_{n,m}, \qquad (21)$$

where $(\cdot)^{\mathrm{H}}$ denotes conjugate transpose; x_j is the superposition codeword transmitted to the j-th cluster, generated from (3); $\mathbf{h}^{\mathrm{H}}_{n,m}$ and $z_{n,m}$ respectively denote the channel vector between the transmitter and the m-th user in the n-th cluster and the corresponding thermal noise, and $e_{n,m}$ denotes the inter-cluster interference.

The inter-cluster interference $e_{n,m}$ can be reduced by employing linear precoding combined with an efficient user selection and clustering algorithm. Moreover, we note that x_n obtained from (3) is the sum of M signals transmitted simultaneously on the same spatial dimension $\mathbf{w}_n$. Therefore, signals intended to users lying in the same cluster create mutual interference. This intra-cluster interference can be removed through an interference cancellation scheme.

B. User Clustering

The proposed clustering algorithm selects $N_{\mathrm{c}} \cdot M$ users out of the K available ones, and groups them into N_{c} clusters of M users each. Fig. 8 provides an example for the case of $K = 20$ available (blue and red) UEs, $N_{\mathrm{t}} = 4$ transmit antennas and clusters, and $M = 3$ selected (red) UEs per cluster. The algorithm ensures two conditions: (i) that users within the same cluster experience highly correlated channels, and (ii) that users lying in different clusters experience highly uncorrelated channels. The former ensures that all users within a j-th cluster receive a strong component of the signal beam intended for that cluster. The latter aims at reducing the inter-cluster interference when paired with ZF precoding. It should be noted that (ii) improves the performance of ZF beamforming with respect to the case when inter-cluster channel correlation is not controlled [23].

More specifically, the proposed clustering algorithm consists of two phases. In the first phase, one user is selected for each of the N_{c} clusters, ensuring that the channels $\mathbf{h}_{n,1}$, $n = 1, \ldots, N$ of these users, denoted as the *cluster heads*, have significant orthogonal components.[1] In the second phase of the proposed clustering algorithm, $M - 1$ additional users are selected for each cluster, such that all channels $\mathbf{h}_{n,m}$, $m = 1, \ldots, M$ of users that lie in the same n-th cluster are highly correlated. The two phases of the proposed clustering algorithm are provided in Algorithm 3. Once users have been arranged in clusters, a scheduling algorithm can be employed to obtain the CPACs and generate the superposition transmission. Such procedure is provided in Algorithm 4 and works similarly to what is described in Section III for the case of single-antenna BSs.

C. Zero Forcing Beamforming

After UEs have been selected and clustered, each BS adopts ZF beamforming for the simultaneous transmission of signals to different clusters. Zero forcing beamforming is of particular interest because it is a linear scheme with low-complexity implementation, and because it can control the amount of interference across clusters [24]–[26]. In our proposed MIMO-based S-MUST design, each BS stacks up the N_{c} channels to the selected cluster heads in the following matrix

$$\mathbf{H} = [\mathbf{h}^{\mathrm{T}}_{1,1}, \ldots, \mathbf{h}^{\mathrm{T}}_{N_{\mathrm{c}},1}] \qquad (22)$$

and calculates the beamforming vectors $\mathbf{w}_n$, $n = 1, \ldots, N_{\mathrm{c}}$, as follows

$$\mathbf{w}_n = \frac{1}{\sqrt{\gamma}}\mathbf{h}^{\mathrm{H}}_{n,1}\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}, \qquad (23)$$

where $(\cdot)^{\mathrm{T}}$ denotes transpose and $\gamma = \mathrm{tr}\{\mathbf{H}^{\mathrm{H}}\mathbf{H}(\mathbf{H}\mathbf{H}^{\mathrm{H}})^{-2}\}$ is a power normalization constant. We note that under ZF beamforming the following condition holds

$$\mathbf{h}_{n,1}\mathbf{w}_j = 0 \quad \forall j \neq n, \qquad (24)$$

[1] Some of the operations performed in this phase are similar to the ones in [23] for orthogonal multiuser transmission. However, it should be noted that the algorithm in [23] may fail to find suitable cluster configurations when the numbers K and N are comparable. Another issue with the algorithm in [23] is that it employs an orthogonality threshold whose optimal value is unknown and depends on the system parameters. The two issues above do not occur with the proposed clustering algorithm.

Algorithm 3: Proposed Clustering Algorithm for MIMO-based S-MUST (First Phase)

```
1: initialize $\mathcal{T}_1 = \{1, \ldots, K\}$
2: initialize $i = 1$
3: for each user $k \in \mathcal{T}_1$ do
4:     estimate channels $\mathbf{g}_k$
5: end for
6: for each user $k \in \mathcal{T}_i$ do
7:     calculate $\tilde{\mathbf{g}}_k$, the component of $\mathbf{g}_k$ orthogonal to the
       subspace spanned by $\{\mathbf{h}_{1,1}, \ldots, \mathbf{h}_{i-1,1}\}$
           $$\tilde{\mathbf{g}}_k = \mathbf{g}_k - \sum_{j=1}^{i-1} \mathbf{h}_{j,1} \frac{\mathbf{h}_{j,1}^{\mathrm{H}} \mathbf{g}_k}{\|\mathbf{h}_{j,1}\|^2}$$
       (when $i = 1$, this implies $\tilde{\mathbf{g}}_k = \mathbf{g}_k$)
8: end for
9: select the first user for the $i$-th cluster as
           $$\pi(i) = \underset{k \in \mathcal{T}_i}{\operatorname{argmax}} \|\tilde{\mathbf{g}}_k\|$$
10: $\mathcal{U}_i = \{\pi(i)\}$
11: $\mathcal{T}_{i+1} = \mathcal{T}_i \backslash \{\pi(i)\}$
12: $\mathbf{h}_{i,1} = \mathbf{g}_{\pi(i)}$
13: $i \leftarrow i + 1$
14: if $\mathcal{T}_{i+1}$ is nonempty and $i \leq N_{\mathrm{c}}$ then
15:     go to line 6
16: else
17:     the first phase is completed, go to line 19
18: end if
```

Algorithm 3: Proposed Clustering Algorithm for MIMO-based S-MUST (Second Phase)

```
19: reconsider all remaining users
           $$\mathcal{T} = \{1, \ldots, K\} \backslash \cup_{j=1}^{N_{\mathrm{c}}} \mathcal{U}_j$$
20: initialize $m = 2$
21: for each user $k \in \mathcal{T}$ do
22:     for $n = 1, \ldots, N_{\mathrm{c}}$ do
23:         calculate $\bar{g}_{k,n}$, the correlation between $\mathbf{g}_k$ and $\mathbf{h}_{n,1}$
               $$\bar{g}_{k,n} = \frac{|\mathbf{h}_{n,1}^{\mathrm{H}} \mathbf{g}_k|}{\|\mathbf{h}_{n,1}\| \|\mathbf{g}_k\|}$$
24:     end for
25: end for
26: for $n = 1, \ldots, N_{\mathrm{c}}$ do
27:     select the most correlated user as
               $$\pi_n(m) = \underset{k \in \mathcal{T}}{\operatorname{argmax}} |\bar{g}_{k,n}|$$
28:     $\mathcal{U}_n \leftarrow \mathcal{U}_n \cup \{\pi_n(m)\}$
29:     $\mathbf{h}_{n,m} = \mathbf{g}_{\pi_n(m)}$
30:     $\mathcal{T} = \mathcal{T} \backslash \pi_n(m)$
31: end for
32: $m \leftarrow m + 1$
33: if $m \leq M$ then
34:     go to line 26
35: else
36:     the second phase is completed, go to line 1
37: end if
```

therefore cluster heads do not receive any inter-cluster interference. However, all remaining users in each cluster do receive inter-cluster interference, since

$$\mathbf{h}_{n,m} \mathbf{w}_j \neq 0 \quad \text{if } m \neq 1, \tag{25}$$

and such interference is treated as noise and dealt with by the S-MUST decoder.

D. Encoding and Decoding

On each beam formed by the ZF precoder, superposition transmission and reception is performed according to the S-MUST encoding and decoding procedures described in Section III.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed S-MUST scheme. A detailed list of the simulation parameters is provided in Table II.

A. Performance of S-MUST

In what follows, QPSK or 16-QAM are adopted as the component constellation for each user, which form 16-QAM or 256-QAM composite constellations, respectively.

Fig. 9 compares the user fairness of several schemes in symmetric broadcast channels in terms of minimum bits per channel use (BPCU) – i.e., those of the worst user – versus

Algorithm 4: Proposed Scheduling Algorithm for MIMO-based S-MUST

```
1: for $n = 1, \ldots, N_{\mathrm{c}}$ do
2:     for each sub-band do
3:         for each UE pair $(\mathrm{UE}_j, \mathrm{UE}_k)$ in $\mathcal{U}_n$ do
4:             $\mathsf{SNR}_j = \frac{P \cdot \mathrm{CQI}_j}{2\sigma^2}$; $\mathsf{SNR}_k = \frac{P \cdot \mathrm{CQI}_k}{2\sigma^2}$;
5:             $\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right] = \mathsf{LUT}\left(\mathsf{SNR}_j, \mathsf{SNR}_k\right)$;
6:             calculate the $\mathrm{PF}_{j,k}\left(\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right]\right)$;
7:         end for
8:         $\left(\hat{j}, \hat{k}\right) = \underset{j,k \in \mathcal{U}_n}{\arg\max} \{\mathrm{PF}_{j,k}\left(\left[(\hat{\alpha}_1, \hat{\alpha}_2), \left(\hat{\beta}_1, \hat{\beta}_2\right)\right]\right)\}$;
9:         $\mathcal{U}_{s,n} \leftarrow \mathcal{U}_{s,n} \cup \left(\hat{j}, \hat{k}\right)$.
10:    end for
11: end for
```

TABLE II: Simulation parameters.

Cellular Layout	Hexagonal, wrapped around
Topology	7 sites (no sectorization)
Bandwidth	10 Mhz
Tx Antenna No.	1 or 2 (omni-directional)
Rx Antenna No.	1 (omni-directional)
No. of UEs per cell	150 (full-buffer traffic model)
BS inter-site distance	500 m
BS Tx power	46 dBm
Thermal noise density	-174 dBm/Hz
Rx noise figure	5 dB
Path loss model	$128.1 + 37.6 \log_{10}(D)$, D in km
Fast fading	i.i.d. Rayleigh fading

SNR, where $h_1 = h_2 = 1$ and each user adopts QPSK modulation. We used SIC for S-MUST Cat. 1 and 2 and MUST Cat. 1-3 to generate the results in Fig 9, and M-PIC for S-MUST Cat. 3. We can observe that: (i) S-MUST outperforms MUST Cat. 1 in regimes of moderate SNR, exhibiting a 4.3 dB enhancement; (ii) S-MUST Cat. 2 outperforms MUST Cat. 2 in low-SNR regime with a 3.8 dB enhancement; (iii) S-MUST outperforms MUST Cat. 3 in low-SNR regime with a 4.2 dB enhancement; (iv) S-MUST Cat. 1 and 2 achieve nearly equal performance while S-MUST cat.1 is slightly worse; (v) S-MUST Cat. 3 performance is worse than OMA and all SIC-based schemes, as M-PIC is a sub-optimal decoder, while it enjoys lower complexity and less overhead; and (vi) S-MUST almost achieves the same user fairness as OMA, i.e., equal user rates.

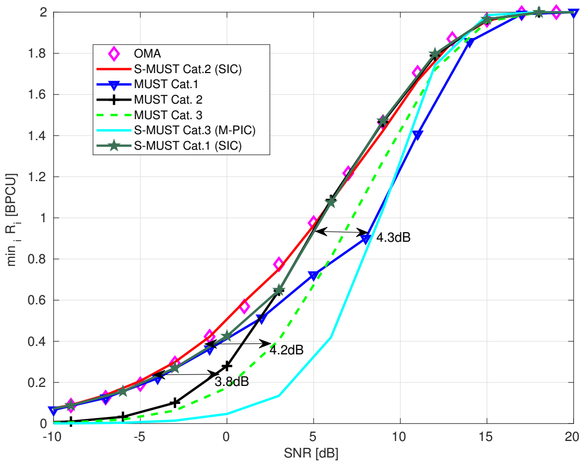


Fig. 9: User fairness comparison in symmetrical broadcast channels under QPSK.

Fig. 10 provides a similar performance comparison for the case where each user adopts a 16-QAM modulation. Similar observations can be made: (i) S-MUST Cat. 2 outperforms MUST Cat. 1 in moderate-SNR regime with a 7.1 dB enhancement; (ii) S-MUST Cat. 2 outperforms MUST Cat. 2 in low-SNR regime with a 6.3 dB enhancement; (iii) S-MUST Cat. 2 outperforms MUST Cat. 3 in low-SNR regime with a 9.2 dB enhancement; (iv) S-MUST Cat. 1 and 2 achieve nearly equal performance; (v) S-MUST cat. 3 performance is worse than OMA and all SIC based schemes as M-PIC is sub-optimal decoder while it enjoys the lowest complexity and less overhead; (vi) S-MUST Cat. 1 and 2 almost achieve the same user fairness as OMA.

B. Performance of MIMO-based S-MUST

In the following, we evaluate the performance of the proposed MIMO-based S-MUST design (S-MUST Cat. 2), by comparing it to MIMO-based designs of conventional MUST, and to a MIMO-based dynamic MA approach where OMA and MUST are opportunistically alternated. In what follows, each BS is equipped with 2 antennas, each UE is equipped with a single antenna, and 4 UEs share the same PRB. Fig. 11 and Fig. 12 show the cumulative distribution function (CDF) of the minimum rate – i.e. that of the worst user –, where QPSK or 16-QAM are adopted as the component constellation for each user, respectively, yielding 16-QAM or 256-QAM composite constellations. In both cases, one can observe that MIMO-based S-MUST achieves almost equal fairness performance as the one of dynamic MA. Moreover, MIMO-based S-MUST outperforms MIMO-MUST Cat. 1, Cat. 2, and Cat. 3 across the whole rate region. In particular, for the 5%-worst rate (bottom-left region of the curves, representing the cell edge), MIMO-based S-MUST can provide a two-to-three-fold rate gain.

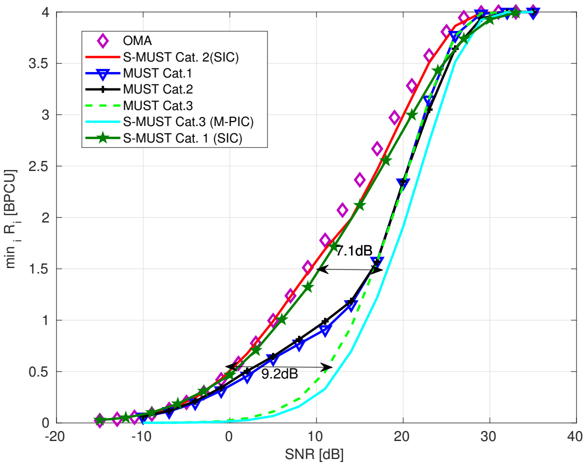


Fig. 10: User fairness comparison in symmetrical broadcast channels under 16-QAM.

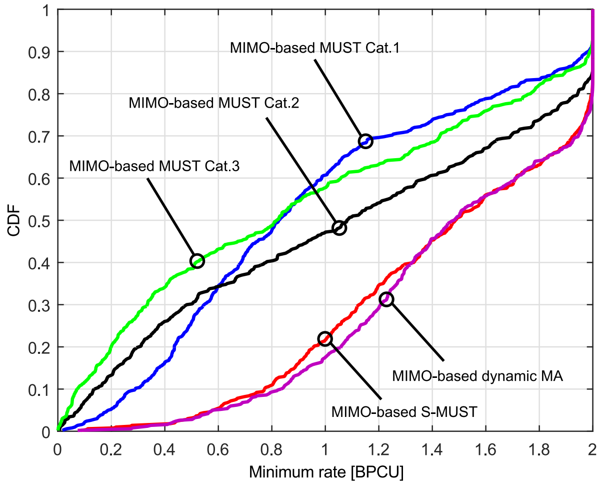


Fig. 11: User fairness comparison in symmetrical multi-antenna broadcast channels under QPSK.

VI. Conclusion

We proposed a new downlink multiuser superposition transmission scheme for future 5G cellular networks, which we denoted *structured multiuser superposition transmission* (S-

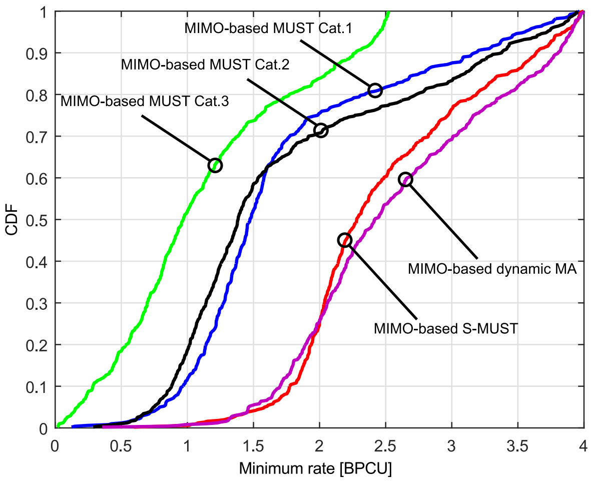


Fig. 12: User fairness comparison in symmetrical multi-antenna broadcast channels under 16-QAM.

MUST). In S-MUST, we apply complex power allocation coefficients (CPACs) over users' legacy constellations to generate a composite constellation. Said CPACs offer an extra degree of freedom for multiplexing users while ensuring that fairness is guaranteed even for symmetric broadcast channels. The newly proposed paradigm of superposition coding provides a unified multiple access air interface, and allows simple parallel decoding based on IQ separation, CPAC quantization, and modulo operations. We also devised suitable scheduling operations for S-MUST, and designed a MIMO-based version of S-MUST for multi-antenna BSs. We demonstrated that the proposed S-MUST design achieves better user fairness compared with conventional MUST, while exhibiting lower complexity compared to dynamic MA.

References

[1] D. Fang, Y.-C. Huang, Z. Ding, G. Geraci, S.-L. Shieh, and H. Claussen, "Lattice partition multiple access: A new method of downlink non-orthogonal multiuser transmissions," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Dec. 2016, pp. 1–6.

[2] G. Geraci, D. Fang, and H. Claussen, "A new method of MIMO-based non-orthogonal multiuser downlink transmission," in *Proc. IEEE Veh. Tech. Conference (VTC)*, June 2017, pp. 1–6.

[3] Nokia Networks, "Ten key rules of 5G deployment - Enabling 1 Tbit/s/km^2 in 2030," *white paper*, Apr. 2015.

[4] Ericsson, "5G radio access - Capabilities and technologies," *white paper*, Apr. 2016.

[5] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Tech. Conference (VTC)*, June 2013, pp. 1–5.

[6] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Int. Sym. on Intelligent Signal Processing and Commun. Systems (ISPACS)*, Nov. 2013, pp. 770–774.

[7] B. Kim, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[8] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Letters*, vol. 4, no. 4, pp. 405–408, Aug. 2015.

[9] X. Liu and X. Wang, "Efficient antenna selection and user scheduling in 5G massive MIMO-NOMA system," in *Proc. IEEE Veh. Tech. Conference (VTC)*, May 2016, pp. 1–5.

[10] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[11] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[12] 3GPP Technical Report 36.859, "Study on downlink multiuser superposition transmission (MUST) for LTE (release 13)," Nov. 2015.

[13] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[14] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[15] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Letters*, vol. 20, no. 7, pp. 1465–1468, July 2016.

[16] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for noma downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[17] R1-156107, "System-level evaluation results of MUST," in *3GPP TSG RAN WG1 Meeting #82bis*, Oct. 2015.

[18] R1-155931, "System-level evaluation results for downlink multiuser superposition schemes," in *3GPP TSG RAN WG1 Meeting #82bis*, Oct. 2015.

[19] T. W. Hungerford, *Algebra*. Springer, 1974.

[20] R1-164977, "Selection of power ratios and modulation for MUST PDSCH," in *3GPP TSG RAN WG1 Meeting #85*, May 2016.

[21] U. Wachsmann, R. F. H. Fischer, and J. B. Huber, "Multilevel codes: theoretical concepts and practical design rules," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1361–1391, Jul. 1999.

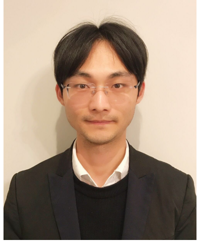
[22] L. Galati Giordano, L. Campanalonga, D. López-Pérez, A. Garcia Rodriguez, G. Geraci, P. Baracca, and M. Magarini, "Uplink sounding reference signal coordination to combat pilot contamination in 5G massive MIMO," in *Proc. IEEE Wireless Commun. Networking Conference (WCNC),* Apr. 2018. Available as *arXiv:1712.06890.*

[23] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[24] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication - Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[25] G. Geraci, R. Couillet, J. Yuan, M. Debbah, and I. B. Collings, "Large system analysis of linear precoding in MISO broadcast channels with confidential messages," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1660–1671, Sept. 2013.

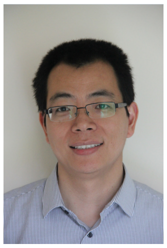
[26] H. H. Yang, G. Geraci, T. Q. S. Quek, and J. G. Andrews, "Cell-edge-aware precoding for downlink massive MIMO cellular networks," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3344–3358, July 2017.



Dong Fang received double B.Sc. degrees from Beijing University of Posts and Telecommunications, China and Queen Mary University of London, UK, in 2010, with the first class honour. He received the Ph.D. degree from the University of York, in 2014. He had been a research associate at the same university from Nov. 2013 to Apr. 2015. He was a research engineer at Nokia Bell Labs from Apr. 2015 to Dec. 2017. He has been a senior data scientist in UnitedHealth Group since Dec. 2017,.

Yu-Chih Huang (M14) received his Ph.D. degree in electrical and computer engineering from Texas A&M University (TAMU) in 2013. He was a post-doctoral research associate at TAMU from 2013 to 2015. Since February 2015, he has been with the Department of Communication Engineering, National Taipei University, Taiwan, where he is currently an Associate Professor. In 2012, he spent the summer as a research intern in Bell Labs, Alcatel-Lucent. His research interests include network information theory, lattice theory, coding theory, and wireless communications. He received the 2018 IEEE Information Theory Society Taipei Chapter and IEEE Communications Society Taipei/Tainan Chapters Best Paper Award for Young Scholars.

Giovanni Geraci is a ”Junior Leader Fellow” at Universitat Pompeu Fabra (Spain), where he leads an externally funded research project on UAV Communications. He earned a Ph.D. degree from the University of New South Wales (Australia) in 2014. Moreover, he gained industrial innovation experience at Nokia Bell Labs (Ireland), where he was a Research Scientist in 2016-2018. His background also features appointments at the Singapore University of Technology and Design (Singapore) in 2014-2015, the University of Texas at Austin (USA) in 2013, Supelec (France) in 2012, and Alcatel-Lucent (Italy) in 2009. He has co-authored over 50 publications attracting more than 1000 citations, and is co-inventor of a dozen pending patents on wireless communications and networking. Giovanni is deeply involved in the research community, serving as an editor for the IEEE Transactions on Wireless Communications and IEEE Communications Letters, and as a workshop or special session co-organizer at IEEE Globecom17, Asilomar18, and IEEE ICC19. He is also a frequent speaker and his contributions include a workshop keynote at IEEE PIMRC18 and tutorials at IEEE WCNC18, IEEE ICC18, and IEEE Globecom18.



Zhiguo Ding (S’03-M’05) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005. From Jul. 2005 to Apr. 2018, he was working in Queen’s University Belfast, Imperial College, Newcastle University and Lancaster University. Since Apr. 2018, he has been with the University of Manchester as a Professor in Communications. From Oct. 2012 to Sept. 2018, he has also been an academic visitor in Princeton University. Dr Ding’ research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Technology*, and *Journal of Wireless Communications and Mobile Computing*, and was an Editor for *IEEE Wireless Communication Letters*, *IEEE Communication Letters* from 2013 to 2016. He received the best paper award in IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, the IEEE Jack Neubauer Memorial Award 2018 and the IEEE Best Signal Processing Letter Award 2018.

Holger Claussen (SM’10) received the Ph.D. degree in digital communications from the University of Edinburgh, U.K., in 2004. He is the Leader of the Indoor Networks Research Department, Nokia Bell Labs, Ireland, and Nokia Bell Labs, USA. He and his team are innovating in all areas related to future Evolution, deployment, and operation of wireless networks to enable exponential growth in data traffic. He has authored 1 book, over 125 publications, and 120 filed patent families.