Resource Allocation for Power Minimization in the Downlink of THP-based Spatial Multiplexing MIMO-OFDMA Systems

Marco Moretti, *Member, IEEE*, Luca Sanguinetti, *Member, IEEE*, Xiaodong Wang, *Fellow, IEEE*

Abstract

In this work, we deal with resource allocation in the downlink of spatial multiplexing MIMO-OFDMA systems. In particular, we concentrate on the problem of jointly optimizing the transmit and receive processing matrices, the channel assignment and the power allocation with the objective of minimizing the total power consumption while satisfying different quality-of-service requirements. A layered architecture is used in which users are first partitioned in different groups on the basis of their channel quality and then channel assignment and transceiver design are sequentially addressed starting from the group of users with most adverse channel conditions. The multi-user interference among users belonging to different groups is removed at the base station using a Tomlinson-Harashima pre-coder operating at user level. Numerical results are used to highlight the effectiveness of the proposed solution and to make comparisons with existing alternatives.

I. Introduction

Dynamic resource allocation in multiple-input multiple-output (MIMO) systems based on orthogonal frequency-division multiple-access (OFDMA) technologies has gained considerable research interest [1]. In most cases, subcarriers are assigned to the active users in an exclusive manner without taking advantage of the multi-user diversity offered by the spatial domain. A possible solution to exploit the spatial dimension is to make use of space-division multiple-access (SDMA) schemes, which allow the simultaneous transmission of different users over the same frequency band. The main impairment of SDMA is represented by multiple-access interference (MAI). In downlink transmissions, MAI mitigation can only be accomplished at the BS using pre-filtering techniques. The most common approach for interference mitigation is zero-forcing (ZF) linear beamforming, which relies on the idea of *pre-inverting* the channel matrix at the transmitter. Another approach is represented by the block-diagonalization ZF (BD-ZF) scheme originally proposed in [2]. Particular attention has been also devoted to dirty paper

M. Moretti and L. Sanguinetti are with the University of Pisa, Dipartimento di Ingegneria dell'Informazione, Italy ({marco.moretti,luca.sanguinetti}@iet.unipi.it). X. Wang is with the Electrical Engineering Department of Columbia University, USA (wangx@ee.columbia.edu)

coding (DPC) techniques [3] even though their implementation is still much open. A possible solution in this direction is represented by Tomlinson-Harashima precoding (THP), which can be seen as a one dimensional DPC technique [4] and has been widely used in the downlink of single-user and multi-user MIMO systems [5]–[8]. In combination with pre-filtering, another way to deal with interference in SDMA-OFDMA systems is user partitioning, which basically consists in properly selecting the set of users transmitting on the same subcarriers. As illustrated in [9], a common approach is to first group together users whose channels have low spatial cross-correlation and then to assign the subcarriers to the various groups. In [10], the authors follow a completely different approach in which the users are first divided into groups such that the spatial cross-correlations among users in different groups is low as much as possible and then subcarriers are sequentially assigned within each group.

From the above discussion, it follows that the use of SDMA schemes in MIMO-OFDMA systems makes the problem of resource allocation more challenging as it requires the joint optimization of *a*) channel assignment and user partitioning; *b*) power allocation over all active links; *c*) transmit and receive filters. To the best of our knowledge, there exists only a few works dealing with all the above problems together. In [11], the authors employ BD-ZF and Lagrange dual decomposition to derive a resource allocation scheme for minimizing the power consumption when individual user rate constraints are imposed. The main drawback of this approach is that an exhaustive search is required to find the best user allocation on each subchannel. A reduced complexity solution is illustrated in [12], in which a two-step procedure is adopted to decouple BD-ZF beamforming from subcarrier and power allocation. Although simpler than [11], it still requires an exhaustive search over a subset of users. In [13], the author exploits a layered architecture in which a user partitioning technique (resembling that discussed in [10]) is first used in conjunction with BD-ZF to partially remove multiuser interference and then carrier assignment is performed jointly with transceiver design using a linear programming (LP) formulation of the allocation problem [14].

In this work, we return to the layered architecture investigated in [13] and extend it in several directions. First, we reformulate the power minimization problem assuming that the quality-of-service (QoS) constraint of each user is given as a sum of the mean-square-errors (MSEs) over all subcarriers rather than on the sum of the achievable rates. Second, transceiver design is carried out employing a non-linear THP precoder operating at *user level* at the transmitter. Third, the choice of the user partitioning strategy is motivated by its combination with the THP precoding technique. This allows us to completely remove the multiuser interference (rather than partially removing it) and to make use of a close-to-optimal partitioning strategy. All this leads to a resource allocation scheme of affordable complexity, which is

shown by means of numerical results to outperform the solution presented in [13].

II. Problem description

We consider[1] the downlink of an OFDMA network in which a total of N subcarriers is used to communicate with K MTs, each equipped with $N_R \geq 2$ antennas[2]. The BS is endowed with $N_T > N_R$ transmit antennas. We denote by $\mathbf{s}_{n,k}$ the N_T-dimensional vector collecting the data transmitted to user k on subcarrier n and by $a_{n,k} \in \{0,1\}$ the binary allocation variable, which is equal to one if subchannel n is assigned to user k and zero otherwise. The goal of this work is to minimize the total power consumption given by

$$P_T = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{E}\left\{\mathbf{s}_{n,k}^H \mathbf{s}_{n,k}\right\} \tag{1}$$

while satisfying user QoS requirements given as a function of the sum of the MSEs over all their assigned subcarriers. To be more specific, the expression for the kth user constraint is

$$\sum_{n=1}^{N} a_{n,k} \sum_{\ell=1}^{L} \mathrm{MSE}_{n,k}(\ell) \leq \gamma_k \tag{2}$$

where L denotes the number of streams transmitted to the kth user over the nth subcarrier and $\mathrm{MSE}_{n,k}(\ell)$ denotes its corresponding MSE. The quantities $\gamma_k > 0$ are design parameters that specify different QoS requirements for each user. We assume that a maximum number of $Q = \lfloor N_T/N_R \rfloor$ users can be simultaneously allocated over each subcarrier, so that it is $\sum_{k=1}^{K} a_{n,k} \leq Q$ for each channel n. To avoid the trivial solution where a user with no allocated subcarrier consumes no power and has a zero MSE, we require that at least n_k subcarriers are assigned to each user so that it is $\sum_{n=1}^{N} a_{n,k} \geq n_k \ \forall k$.

III. Multi-user interference elimination and user partitioning

Unfortunately, solving the optimization problem described above requires an exhaustive search over all possible subcarrier allocations. Moreover, it needs also the joint optimization of the transmit and receive processing matrices for each allocation. All this makes its complexity extremely large for any practical scenario. To address this issue, we follow the approach of [10] and [13], in which the population of K users is partitioned into Q different subsets $\{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(Q)}\}$. This allows us to break the original

[1]We use $\mathbf{A} = \mathrm{blkdiag}\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}$ to represent a block diagonal matrix whereas $\mathbf{A}^{-1}$ and $\mathrm{tr}\{\mathbf{A}\}$ denote the inverse and trace of a square matrix $\mathbf{A}$. We denote $\mathbf{I}_K$ the identity matrix of order K while we use $E\{\cdot\}$ for expectation, $\|\cdot\|$ for the Euclidean norm of the enclosed vector and the superscript *, ^{T} and ^{H} for complex conjugation, transposition and Hermitian transposition. The notation $[\cdot]_{k,\ell}$ indicates the (k,ℓ)th entry of the enclosed matrix.

[2]The results can be easily extended to a more versatile system in which a different number of services is required by each MT. In this case, K would simply denote the total number of services.

problem into a sequence of Q lower-complexity optimization sub-problems, each assigning all radio resources to a subset of users. Users within the same subset are transmitted on orthogonal subcarriers and do not interfere with each other. Channels allocation is performed sequentially starting from set $\mathcal{S}^{(1)}$.

From the above discussion, it follows that, after the Q allocation sub-problems are solved, there will be Q users assigned to each subcarrier. Without loss of generality, we focus on subcarrier n. Let us denote by $\mathcal{K}_n$ the set of users assigned to n and by $\mu_n(i)$ the user in $\mathcal{S}^{(i)}$ associated to subcarrier n. To simplify the notation, in the following derivations the indexes $\mu_n(i)$ will be relabelled according to the map $\mu_n(i) \rightarrow i$. The signal $\mathbf{x}_{n,k} \in \mathbb{C}^{N_R \times 1}$ received at the kth MT over the nth subcarrier can be thus written as

$$\mathbf{x}_{n,k} = \mathbf{H}_{n,k} \sum_{i=1}^{Q} \mathbf{s}_{n,i} + \mathbf{w}_{n,k} \tag{3}$$

where $\mathbf{w}_{n,k} \in \mathbb{C}^{N_R \times 1}$ is a Gaussian vector with zero-mean and covariance matrix $\sigma^2 \mathbf{I}_{N_R}$ and $\mathbf{H}_{n,k} \in \mathbb{C}^{N_R \times N_T}$ is the channel matrix over the nth subcarrier. From (3), it follows that the interference term is given by two different contributions, namely, $\mathbf{H}_{n,k} \sum_{i=1}^{k-1} \mathbf{s}_{n,i}$ and $\mathbf{H}_{n,k} \sum_{i=k+1}^{Q} \mathbf{s}_{n,i}$. The first term represents the interference caused by the active users already allocated before the kth assignment sub-problem has been solved (i.e., users belonging to sets $\mathcal{S}^{(i)}$ with indexes $i < k$), while the second term accounts for the users with indexes $i > k$ (i.e., users which have been allocated after user k). In [13], a BD-ZF scheme is employed to remove the first term while the second one is treated as Gaussian noise. In the sequel, a THP technique operating at user level is used to remove both terms.

A. Multi-user interference elimination

The $L \leq \lfloor N_T/Q \rfloor$ symbols transmitted to the kth user over the nth subcarrier are denoted by $\{d_{n,k}(\ell);\ \ell = 1, 2, \ldots, L\}$. They belong to an M-ary quadrature-amplitude modulation (QAM) alphabet with variance $\sigma_d^2 = 2(M-1)/3$ and are stacked in the L-dimensional vector $\mathbf{d}_{n,k}$. As depicted in Fig. 1, the QL-dimensional data vector $\mathbf{d}_n = [\mathbf{d}_{n,1}^T, \mathbf{d}_{n,2}^T, \ldots, \mathbf{d}_{n,Q}^T]^T$ is pre-coded in a recursive fashion using a *strictly block* lower triangular matrix $\mathbf{B}_n \in \mathbb{C}^{QL \times QL}$ and a non-linear operator $\mathrm{MOD}_M(\cdot)$ that constrains the entries of $\mathbf{b}_{n,i} \in \mathbb{C}^{L \times 1}$ into the square region $\aleph = \{x^{(R)} + jx^{(I)} | x^{(R)}, x^{(I)} \in (-\sqrt{M}, \sqrt{M}]\}$. Denoting by $[\mathbf{B}_n]_{i,\ell} \in \mathbb{C}^{L \times L}$ the (i,ℓ)th block of $\mathbf{B}_n$, we have that $\mathbf{b}_{n,i} \in \mathbb{C}^{L \times 1}$ can be iteratively computed as [4]

$$\mathbf{b}_{n,i} = \mathbf{d}_{n,i} - \sum_{\ell=1}^{i-1} [\mathbf{B}_n]_{i,\ell}\, \mathbf{b}_{n,\ell} + \boldsymbol{\varsigma}_{n,i} \quad i = 1, 2, \ldots, Q \tag{4}$$

where $[\mathbf{B}_n]_{i,\ell} \in \mathbb{C}^{L \times L}$ is the (i,ℓ)th block of $\mathbf{B}_n$, $\boldsymbol{\varsigma}_{n,i}$ is defined as $\boldsymbol{\varsigma}_{n,i} = 2\sqrt{M}\boldsymbol{\xi}_{n,i}$ and $\boldsymbol{\xi}_{n,i} = [\xi_{n,i}(1), \xi_{n,i}(2), \ldots, \xi_{n,i}(L)]^T$ with $\xi_{n,i}(\ell)$ complex-valued quantity, whose real and imaginary parts are

suitable integers that reduce $b_{n,i}(\ell)$ to $\aleph$. The above equation indicates that the modulo operator is equivalent to adding a vector $\boldsymbol{\varsigma}_n$ to the input data $\mathbf{d}_n$. This produces the *modified* data vector $\mathbf{v}_n = \mathbf{d}_n + \boldsymbol{\varsigma}_n = [\mathbf{v}_{n,1}^T, \mathbf{v}_{n,2}^T, , \ldots, \mathbf{v}_{n,Q}^T]^T$ from which $\mathbf{b}_n$ is obtained as follows $\mathbf{b}_n = \mathbf{C}_n^{-1}\mathbf{v}_n$ where $\mathbf{C}_n \in \mathbb{C}^{LQ\times QL}$ is a block *unit-diagonal* and lower triangular matrix given by $\mathbf{C}_n = \mathbf{B}_n + \mathbf{I}_{LQ}$. The pre-coded vectors $\mathbf{b}_{n,i} \in \mathbb{C}^{L\times 1}$ are then linearly processed through the *forward* transmit matrices $\mathbf{F}_{n,i} \in \mathbb{C}^{N_T\times L}$ to produce $\mathbf{s}_{n,i} = \mathbf{F}_{n,i}\mathbf{b}_{n,i}$. The vectors $\mathbf{s}_{n,i}$ for $n = 1, 2, \ldots, N$ and $i = 1, 2, \ldots, Q$ are finally fed to the OFDMA modulator and transmitted over the channel using the N_T antennas of the BS array. As depicted in Fig. 2, at the MT the incoming waveforms are implicitly combined by the receive antennas and passed to an OFDMA demodulator whose outputs take the form in (3) with $\mathbf{s}_{n,i} = \mathbf{F}_{n,i}\mathbf{b}_{n,i}$. The complete elimination of $\mathbf{H}_{n,k}\sum_{i=k+1}^{Q}\mathbf{F}_{n,i}\mathbf{b}_{n,i}$ at the transmitter can be achieved by constraining $\mathbf{F}_{n,k}$ to lie in the null space of $\bar{\mathbf{H}}_{n,k} = [\mathbf{H}_{n,1}^T, \mathbf{H}_{n,2}^T, \ldots, \mathbf{H}_{n,k-1}^T]^T$. Accordingly, this amounts to letting $\mathbf{F}_{n,k}$ have the following structure

$$\mathbf{F}_{n,k} = \mathbf{V}_{\bar{H}_{n,k}}^{(0)}\mathbf{U}_{n,k} \tag{5}$$

where $\mathbf{U}_{n,k} \in \mathbb{C}^{[N_T-(k-1)N_R]\times L}$ is an arbitrary matrix and $\mathbf{V}_{\bar{H}_{n,k}}^{(0)} \in \mathbb{C}^{N_T\times[N_T-(k-1)N_R]}$ is a matrix whose columns form a basis for the *null space* of $\bar{\mathbf{H}}_{n,k}$ obtained from its singular value decomposition (SVD). Setting $\mathbf{F}_{n,k}$ as in (5) into (3) and stacking the received signals of all users into a single vector $\mathbf{x}_n = [\mathbf{x}_{n,1}^T\mathbf{x}_{n,2}^T\cdots\mathbf{x}_{n,Q}^T]^T$, we may write

$$\mathbf{x}_n = \mathbf{T}_n\mathbf{b}_n + \mathbf{w}_n \tag{6}$$

where $\mathbf{w}_n = [\mathbf{w}_{n,1}^T, \mathbf{w}_{n,2}^T, \ldots, \mathbf{w}_{n,Q}^T]^T$ and $\mathbf{T}_n \in \mathbb{C}^{N_RQ\times QL}$ is a block lower triangular matrix with blocks $[\mathbf{T}_n]_{k,i} \in \mathbb{C}^{N_R\times L}$ given by $[\mathbf{T}_n]_{k,i} = \mathbf{H}_{n,k}\mathbf{V}_{\bar{H}_{n,i}}^{(0)}\mathbf{U}_{n,i}$ for $k \geq i$. We are now left with the problem of removing the interference term $\mathbf{H}_{n,k}\sum_{i=1}^{k-1}\mathbf{V}_{\bar{H}_{n,i}}^{(0)}\mathbf{U}_{n,i}\mathbf{b}_{n,i}$ in (3). To this end, we decompose $\mathbf{T}_n$ in (6) as $\mathbf{T}_n = \mathbf{D}_n\mathbf{L}_n$ where $\mathbf{D}_n = \text{blkdiag}\{[\mathbf{T}_n]_{1,1}, [\mathbf{T}_n]_{2,2}, \ldots, [\mathbf{T}_n]_{Q,Q}\}$ and $\mathbf{L}_n$ is a block unit-diagonal and lower triangular matrix with

$$[\mathbf{L}_n]_{k,i} = [\mathbf{T}_n]_{k,k}^H\left([\mathbf{T}_n]_{k,k}[\mathbf{T}_n]_{k,k}^H\right)^{-1}[\mathbf{T}_n]_{k,i} \tag{7}$$

for $k > i$. Substituting $\mathbf{T}_n = \mathbf{D}_n\mathbf{L}_n$ into (6) and recalling that $\mathbf{b}_n = \mathbf{C}_n^{-1}\mathbf{v}_n$ yields $\mathbf{x}_n = \mathbf{D}_n\mathbf{L}_n\mathbf{C}_n^{-1}\mathbf{v}_n + \mathbf{w}_n$ from which setting $\mathbf{C}_n = \mathbf{L}_n$ we obtain $\mathbf{x}_n = \mathbf{D}_n\mathbf{v}_n + \mathbf{w}_n$. Recalling that $\mathbf{D}_n$ has a block-diagonal structure with blocks given by $[\mathbf{T}_n]_{k,k} = \mathbf{H}_{n,k}\mathbf{V}_{\bar{H}_{n,k}}^{(0)}\mathbf{U}_{n,k}$, it follows that the multi-user MIMO system has been decoupled into $|\mathcal{K}_n|$ parallel single-user MIMO links given by

$$\mathbf{x}_{n,k} = \mathbf{H}'_{n,k}\mathbf{U}_{n,k}\mathbf{v}_{n,k} + \mathbf{w}_{n,k} \tag{8}$$

each of which represented by the *equivalent* channel transfer matrix $\mathbf{H}'_{n,k} = \mathbf{H}_{n,k}\mathbf{V}^{(0)}_{\bar{H}_{n,k}}$. This means that each user may operate in its corresponding link independently without affecting the other active users. Henceforth, we denote by $\mathbf{H}'_{n,k} = \mathbf{\Omega}_{H'_{n,k}}\mathbf{\Lambda}^{1/2}_{H'_{n,k}}\mathbf{V}^{(1)^H}_{H'_{n,k}}$ the SVD of $\mathbf{H}'_{n,k}$. As mentioned before, the vectors $\{\mathbf{x}_{n,k}\}$ are processed by the kth mobile terminal for data recovery.

B. User partitioning

As mentioned above, MAI mitigation in SDMA-OFDMA systems is accomplished not only by precoding the users' data but also by partitioning the users and dynamically assigning the radio channels. Unfortunately, optimal grouping is a problem of combinatorial complexity whose solution can only be found through an exhaustive search. To overcome this problem, a heuristic approach widely used in the literature is to partition users on the basis of their space cross-correlations (see for example [9]). Although reasonable, this approach has still a large complexity as it requires the calculation of the cross-correlations among all users in the system over all available channels. Alternatively, in this work we exploit the fact that THP can be viewed as the transmit counterpart of the vertical Bell Labs layered space-time (V-BLAST) architecture and thus we order the users according to their channel qualities as as originally proposed in [15] and later extended to THP in [16]. In our context, the channel quality of the kth user is measured by the following quantity:

$$\pi(k) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{tr}\left(\mathbf{H}^H_{n,k}\mathbf{H}_{n,k}\right) = \frac{1}{N}\sum_{n=1}^{N}\sum_{\ell=1}^{L}\lambda_{H_{n,k}}(\ell) \tag{9}$$

where $\{\lambda_{H_{n,k}}(\ell)\}$ denote the eigenvalues of $\mathbf{H}^H_{n,k}\mathbf{H}_{n,k}$. The above quantities are used to partition users according to a *worst-first* criterion. In doing so, the users with the most attenuated channels are allocated in set $\mathcal{S}^{(1)}$ whereas the users with the best channels are grouped in $\mathcal{S}^{(Q)}$. This choice is motivated by the fact that the null-space projection in (5) progressively reduces the available spatial diversity as the group index tends to Q and the number of rows of $\bar{\mathbf{H}}_{n,k}$ increases up to $(Q-1)\,N_R$. Therefore, since power consumption is in general dominated by users with the worst channel conditions, we give those users higher priority by placing them in set $\mathcal{S}^{(1)}$. Observe that the MAI arising among users (in different sets) allocated on the same subcarriers is mitigated jointly by THP and dynamic channel assignment. With the objective of minimizing the overall required power, channel assignment will automatically couple users that tend to not interfere with each other. It is worth observing that the same ordering strategy is used in [13] following a different line of reasoning.

IV. Linear programming subcarrier assignment

Without loss of generality, we focus on the resource allocation problem over the K/Q users within the set $\mathcal{S}^{(q)}$. For notational convenience, we denote by $\mathbf{a}^{(q)}$ and $\mathbf{U}^{(q)}$ the vector and the matrix obtained

stacking the allocation variables and the precoding matrices of the users in $\mathcal{S}^{(q)}$, respectively. As before, the user indexes $\mu_n(i)$ will be relabelled according to the map $\mu_n(i) \leftarrow i$. To make the problem mathematically tractable, we assume also that the precoded symbols $\mathbf{b}_{n,k}$ are statistically independent and with the same power of user data[3], i.e., $\mathrm{E}\{\mathbf{b}_{n,k}\mathbf{b}_{n,k}^H\} = \sigma_d^2\mathbf{I}_L$. In these circumstances, using (5) it follows that the power required by the BS to transmit the signal $\mathbf{s}_{n,k}$ is given by $\mathrm{E}\{\mathbf{s}_{n,k}\mathbf{s}_{n,k}^H\} = \sigma_d^2\mathrm{tr}\{\mathbf{U}_{n,k}^H\mathbf{U}_{n,k}\}$. The optimization problem can be thus mathematically formulated as:

$$\min_{\mathbf{U}^{(q)},\mathbf{a}^{(q)}} \sum_{n=1}^{N} \sum_{k\in\mathcal{S}^{(q)}} a_{n,k}\mathrm{tr}\left\{\mathbf{U}_{n,k}^H\mathbf{U}_{n,k}\right\} \tag{10}$$

$$\text{subject to} \quad \sum_{n=1}^{N} a_{n,k}\sum_{\ell=1}^{L}\mathrm{MSE}_{n,k}(\ell) \le \gamma_k \quad k\in\mathcal{S}^{(q)} \quad \text{and} \quad \sum_{n=1}^{N} a_{n,k} \ge n_k \quad k\in\mathcal{S}^{(q)} \tag{10.1}$$

which is a mixed-integer non-linear problem and thus not convex and very difficult to solve. A possible way out is to decouple the power allocation and subcarrier assignment problems. This can be achieved by assigning n_k subcarriers to the kth user and designing the processing matrices such that the following constraint is satisfied

$$\sum_{\ell=1}^{L}\mathrm{MSE}_{n,k}(\ell) \le \frac{\gamma_k}{n_k}. \tag{11}$$

In this framework, the power is no longer an optimization variable but simply the cost of using n_k subcarriers [17]. In particular, the cost $c_{n,k}$ of using subcarrier n for user $k\in\mathcal{S}^{(q)}$ can be computed as

$$\min_{\mathbf{U}_{n,k}} \quad \mathrm{tr}\left\{\mathbf{U}_{n,k}^H\mathbf{U}_{n,k}\right\} \quad \text{subject to} \quad \sum_{\ell=1}^{L}\mathrm{MSE}_{n,k}(\ell) \le \frac{\gamma_k}{n_k}. \tag{12}$$

Once the solution of (12) is obtained, (10) can be recast as a linear integer programming (LIP) problem:

$$\min_{\mathbf{a}^{(q)}} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{S}^{(q)}} a_{n,k}c_{n,k} \tag{13}$$

$$\text{subject to} \quad \sum_{n=1}^{N} a_{n,k} = n_k \quad k\in\mathcal{S}^{(q)} \quad \text{and} \quad \sum_{k\in\mathcal{S}^{(q)}} a_{n,k} \le 1 \quad \forall n$$

where the objective function and the constraints are linear in $\{a_{n,k}\}$. In general, the solution of LIP problems can be found either performing an exhaustive search or relaxing the integrality condition on the allocation variable. In this particular case, the channel assignment in (13) has the advantage that can be modelled as a *minimum cost flow* problem and as such it is possible to show that the solution obtained by relaxing the integral condition is the optimal *integral* solution, so that very efficient solvers can be employed with no performance degradation [17].

[3]Although not rigorously true, this assumption is reasonable for large $M-$QAM constellations with size $M \ge 16$ [4].

A. Receiver design

To keep the complexity of the MTs at a tolerable level, we assume that a linear receiver is used for data recovery. As depicted in Fig. 2, vector $\mathbf{x}_{n,k}$ in (8) is first processed by $\mathbf{G}_{n,k} \in \mathbb{C}^{L \times N_R}$ to obtain

$$\mathbf{y}_{n,k} = \mathbf{G}_{n,k}\mathbf{H}'_{n,k}\mathbf{U}_{n,k}\mathbf{v}_{n,k} + \mathbf{G}_{n,k}\mathbf{w}_{n,k} \tag{14}$$

and then passed to the same modulo operator employed at the transmitter so as to remove the effect of $\boldsymbol{\varsigma}_{n,k}$. The output $\mathbf{z}_{n,k} = [z_{n,k}(1), z_{n,k}(2), \ldots, z_{n,k}(L)]^T$ is finally fed to a threshold unit which delivers an estimate of $\mathbf{d}_{n,k}$. From (14), it follows that the received samples depend on $\mathbf{G}_{n,k}$ and $\mathbf{U}_{n,k}$. The latter must be designed so as to mitigate co-channel interference while satisfying the QoS constraints. For this purpose, we adopt a ZF approach in which multi-stream interference is completely eliminated and the remaining degrees of freedom are exploited to minimize the power consumption under the constraint on the MSEs. The complete elimination of the multi-stream interference implies that

$$\mathbf{G}_{n,k}\mathbf{H}'_{n,k}\mathbf{U}_{n,k} = \mathbf{I}_L. \tag{15}$$

In these circumstances, the output $z_{n,k}(\ell)$ from the modulo operator takes the form[4] $z_{n,k}(\ell) = d_{n,k}(\ell) + n_{n,k}(\ell)$ and its corresponding MSE results given by $\mathrm{MSE}_{n,k}(\ell) = \sigma^2[\mathbf{G}_{n,k}\mathbf{G}_{n,k}^H]_{\ell,\ell}$. It can be shown that the optimal $\mathbf{G}_{n,k}$ satisfying (15) and minimizing each $\mathrm{MSE}_{n,k}(\ell)$ is the minimum norm solution of (15) [18]. The latter is found to be $\mathbf{G}_{n,k} = (\mathbf{U}_{n,k}^H\mathbf{H}'^H_{n,k}\mathbf{H}'_{n,k}\mathbf{U}_{n,k})^{-1}\mathbf{U}_{n,k}^H\mathbf{H}'^H_{n,k}$ from which it follows that $\mathrm{MSE}_{n,k}(\ell) = \sigma^2[(\mathbf{U}_{n,k}^H\mathbf{H}'^H_{n,k}\mathbf{H}'_{n,k}\mathbf{U}_{n,k})^{-1}]_{\ell,\ell}$. We now proceed with the design of the matrix $\mathbf{U}_{n,k}$, which requires to solve the following problem:

$$\min_{\{\mathbf{U}_{n,k}\}} \quad \mathrm{tr}\left\{\mathbf{U}_{n,k}^H\mathbf{U}_{n,k}\right\} \quad \text{subject to} \quad \sum_{\ell=1}^{L} \sigma^2 \left[\left(\mathbf{U}_{n,k}^H\mathbf{H}'^H_{n,k}\mathbf{H}'_{n,k}\mathbf{U}_{n,k}\right)^{-1}\right]_{\ell,\ell} \leq \frac{\gamma_k}{n_k}. \tag{16}$$

The solution can be computed as follows.

Proposition 1: The optimal $\mathbf{U}_{n,k}$ in (16) takes the form

$$\mathbf{U}_{n,k} = \mathbf{V}^{(1)}_{H'_{n,k}}\boldsymbol{\Lambda}^{1/2}_{U_{n,k}}\mathbf{S}^H_{n,k} \tag{17}$$

where $\mathbf{V}^{(1)}_{H'_{n,k}}$ is obtained from the SVD of $\mathbf{H}'_{n,k}$, $\boldsymbol{\Lambda}_{U_{n,k}}$ is diagonal and $\mathbf{S}_{n,k} \in \mathbb{C}^{L \times L}$ is a suitable unitary matrix such that $\mathrm{MSE}_{n,k}(\ell) = \epsilon_k$ for $\ell = 1, 2, \ldots, L$ with $\epsilon_k = \frac{1}{L}\frac{\gamma_k}{n_k}$. In addition, the diagonal elements of $\boldsymbol{\Lambda}_{U_{n,k}}$ are given by

$$\lambda_{U_{n,k}}(\ell) = \sqrt{\nu_{n,k}\frac{\sigma^2}{\lambda_{H'_{n,k}}(\ell)}} \quad \ell = 1, 2, \ldots, L \tag{18}$$

[4]In writing $z_{n,k}(\ell) = d_{n,k}(\ell) + n_{n,k}(\ell)$, we have neglected for simplicity the modulo-folding effect on the thermal noise. Although not rigorous, this assumption is quite reasonable for moderate values of signal-to-noise ratios (see for example the book of Robert F. H. Fisher [4] for a complete treatment of the subject).

where $\nu_{n,k}$ is such that $\sum_{\ell=1}^{L} \frac{\sigma^2}{\lambda_{U_{n,k}}(\ell)\lambda_{H'_{n,k}}(\ell)} = \frac{\gamma_k}{n_k}$.

Proof: The proof is omitted for space limitations but it can be derived using the results illustrated in [19] since the sum of the MSEs is a Schur-convex function. ■

Using the results of Proposition 1, the cost $c_{n,k}$ in (13) is eventually given by

$$c_{n,k} = \sum_{\ell=1}^{L} \lambda_{U_{n,k}}(\ell) \tag{19}$$

with $\lambda_{U_{n,k}}(\ell)$ computed as in (18).

B. Complexity analysis

All the operations required by the proposed solution are summarized in Algorithm 1 whose computational load can be assessed in terms of the number of required floating point operations (flops) as follows[5]. Observe that computing the quantities $\{\pi(k)\}$ requires $\mathcal{O}(NKN_TN_R)$ flops whereas computing the power cost $c_{n,k}$ according to (19) basically requires first to evaluate the SVDs of $\bar{\mathbf{H}}_{n,k}$ for $k = 1, 2, \ldots, K/Q$ and $n = 1, 2, \ldots, N$ and then those of $\mathbf{H}'_{n,k}$ in (8) for $k = 1, 2, \ldots, K$ and $n = 1, 2, \ldots, N$. The total number of flops required for these two operations are summarized in the second and third row of Table I. In writing these figures, we have taken into account that evaluating the SVDs of $\mathbf{H}'_{n,k}$ requires $\mathcal{O}(Q/2(Q-1)N_TN_R^2 + QN_RN_T^2)$ flops in total since $\mathcal{O}(QN_RN_T^2)$ flops are needed to compute $\mathbf{H}'_{n,k} = \mathbf{H}_{n,k}\mathbf{V}^{(0)}_{\bar{H}_{n,k}}$ whereas $\mathcal{O}(Q/2(Q-1)N_TN_R^2)$ flops are required for the SVD. Summing all the above terms it turns out that the overall complexity for computing all costs $\{c_{n,k}\}$ is approximately given by $\mathcal{O}(NKQN_RN_T^2)$. The complexity of solving (13) is an open research issue. The latest results (see for example [20] references therein) place the complexity of the assignment problem in a range between $\mathcal{O}\left(\kappa^2\right)$ and $\mathcal{O}\left(\kappa^{2.5}\right)$ with κ being the total number of nodes. In our case, the number of nodes is the sum of the number of users per single allocation problem plus the number of subcarriers, i.e., $\kappa = N + K/Q$. Since we have Q distinct subproblems to solve, the overall complexity of the LP optimization is approximately given by $\mathcal{O}\left(Q(N + K/Q)^{2.5}\right)$ flops. The computation of $\mathbf{B}_n = \mathbf{C}_n - \mathbf{I}_{LQ}$ in (4) with $\mathbf{C}_n = \mathbf{L}_n$ can be assessed as follows. Evaluating each $[\mathbf{L}_n]_{k,i}$ in (7) requires $\mathcal{O}((N_R^3 + 4LN_R^2))$ flops. Since the total number of matrices $[\mathbf{L}_n]_{k,i}$ is $Q/2(Q-1)$, it follows that $\mathcal{O}(NQ/2(Q-1)(N_R^3 + 4LN_R^2))$ flops are required to obtain all matrices $\{\mathbf{C}_n\}$ and thus all $\{\mathbf{B}_n\}$. The computational load for obtaining $\{\mathbf{F}_{n,k}\}, \{\mathbf{G}_{n,k}\}$ and $\{\mathbf{U}_{n,k}\}$ can be reasonably neglected as it basically require to put together all the unitary matrices computed above with SVDs. The processing requirements of the proposed two-layer architecture are

[5]In doing so, we make use of the following results: *i*) the multiplication of $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\mathbf{B} \in \mathbb{C}^{n\times p}$ requires $\mathcal{O}(2mnp)$ flops; *ii*) evaluating the SVD of $\mathbf{A} \in \mathbb{C}^{m\times n}$ needs $\mathcal{O}(mn^2)$ flops; *iii*) the inverse of $\mathbf{A} \in \mathbb{C}^{n\times n}$ requires $\mathcal{O}(n^3)$ flops.

summarized in Table I from which it follows that the overall number of flops is approximately given by $\mathcal{O}\left(Q(N+K/Q)^{2.5}+NKQN_RN_T^2+NQ^2N_R^3\right)$. The latter is comparable to the computational load required by the scheme illustrated in [13] as it is dominated by the computational burden required by the LP approach, especially when the number of subcarriers relatively large. However, as shown in the sequel, the proposed solution provides much better performance in terms of power reduction with respect to [13] thanks to the underlying THP scheme.

V. Numerical results

We consider a system with K uniformly distributed users in a cell of radius $R = 100$ m. The propagation channel is static, frequency-selective and modelled as a Rayleigh fading process with an exponentially decaying power delay profile. The path loss exponent is $\beta = 4$. Unless noted differently, the number of users is $K = 16$.

We compare the proposed architecture, denoted by THP Tx - Lin Rx, with three other algorithms: *a*) a ZF linear beam-former, denoted as ZF Tx, *b*) a THP scheme, denoted as THP Tx (see for example [6]), and *c*) the architecture proposed in [13] that employs linear processing at both the transmitter and the receiver (Lin Tx - Lin Rx). In details, letting $\mathbf{H}_n = [\mathbf{H}_{n,1}^T\,\mathbf{H}_{n,2}^T\cdots\mathbf{H}_{n,Q}^T]^T$ and $\mathbf{F}_n = [\mathbf{F}_{n,1}\,\mathbf{F}_{n,2}\cdots\mathbf{F}_{n,Q}]^T$, the precoding matrix for ZF Tx is $\mathbf{F}_n = \mathbf{H}_n^H(\mathbf{H}_n\mathbf{H}_n^H)^{-1}$. The THP Tx architecture is realized by setting $\mathbf{F}_n = \mathbf{Q}_n$ and $\mathbf{C}_n = \mathbf{R}_n^{-H}$ with $\mathbf{Q}_n$ and $\mathbf{R}_n$ being computed as the QR decomposition $\mathbf{H}_n^H$, i.e., $\mathbf{H}_n^H = \mathbf{Q}_n\mathbf{R}_n$. Both ZF Tx and THP Tx schemes are designed to remove the inter-stream and inter-user interference at the transmitter so that the receive filter is $\mathbf{G}_{n,k} = \mathbf{I}_L$.

We consider three different scenarios, summarised in Table II, which are designed to observe the behaviour of the proposed algorithms when the total number of available channels per user is fixed and frequency channels are progressively replaced by streams in the spatial domain. More in details, the first scenario, referred to as $S^{(1)}$, is a 2×1 MIMO system with a bandwidth $W^{(1)} = 10$ MHz and $N^{(1)} = 64$ orthogonal subchannels. The bandwidth of Scenario $S^{(2)}$ is $W^{(2)} = 5$ MHz, spanning $N^{(2)} = 32$ subchannels with a 4×2 MIMO configuration. Scenario $S^{(3)}$ transmits over a bandwidth $W^{(3)} = 2.5$ MHz with $N^{(3)} = 16$ subchannels and employs a 8×4 configuration. For each scenario we assume that the number of allocated subcarriers is $n_k^{(i)} = N^{(i)}\times Q/K$ and the total number of channels per user is $n_k^{(i)}L^{(i)} = 8$ ($i = 1,\ldots,3$; $k = 1,\ldots,K$) regardless of the scenario considered.

Figs. 3 – 5 report the total transmit power for the three scenarios as a function of the average target MSE ρ per data stream. By design, for a given value of ρ, the overall MSE is $\gamma_k^{(i)} = 8\rho$ ($i = 1,\ldots,3$; $k = 1,\ldots,K$). Results show that the gains obtained thanks to the implementation of non-linear processing progressively increase from scenario $S^{(1)}$ to $S^{(3)}$, as the spatial dimension becomes more important.

In particular, Fig. 3 shows that, with a 2×1 configuration and 64 channels, all the schemes, except ZF Tx, tend to have similar performance. The effect of resource allocation is predominant and the users transmitting on the same channel are sufficiently separated regardless of the specific architecture.

As the number of orthogonal frequency channels is reduced, the consequent diminution in frequency diversity is only partially compensated by the larger number of antennas: in facts, even if the total number of channels is the same, the spatial streams tend to be more correlated. In this case, the choice of the transceiver architecture plays a very important role since channel allocation alone is not able to fully exploit all the diversity of the the system. The results plotted in Fig. 4 show that the THP-based schemes largely outperform all other solutions.

The same trend appears in Fig. 5, where THP Tx - Lin Rx effectively exploits the spatial diversity provided by the multiple antennas. Scenario $S^{(1)}$ requires less power when compared to $S^{(2)}$ and $S^{(3)}$ as it occupies a larger bandwidth. In scenarios $S^{(2)}$ and $S^{(3)}$, the proposed scheme takes advantage of the increased spatial dimension to transmit the same amount of data employing a comparable amount of power and occupying only a fraction of the bandwidth.

Fig. 6 shows the total transmit power for an average target MSE $\rho = 0.25$ as a function of K for $S^{(1)}$ and $S^{(3)}$. For ease of representation, only the results of THP Tx, Lin Tx - Lin Rx and THP Tx - Lin Rx are reported. As before, the parameters are set so that the number of data stream per user is the same (regardless of the specific scenario). An accurate inspection of the results shows that for scenario $S^{(1)}$, the performance of the three algorithms tend to be very close for $K \geq 16$, when the resource allocation algorithm is able to fully exploit both multi-user and frequency diversity. The situation is remarkably different for scenario $S^{(3)}$ where it appears that resource allocation alone is not sufficient to completely deal with MAI. In fact, all multiuser diversity is already exploited for $K = 8$ and further increase of the number of users produce only marginal improvements. In this case, the THP Tx - LIN Rx configuration outperforms the other two schemes thanks to its capability to cancel the MAI.

VI. Conclusions

We have derived a resource allocation scheme for the downlink of SDMA-MIMO-OFDMA systems. The proposed solution relies on a layered architecture in which MAI is first removed by means of a THP technique operating at user level and then channel assignment and transceiver design are jointly addressed using a ZF-based linear programming approach that aims at minimizing the power consumption while satisfying specific QoS requirements given as the sum of the MSEs over the assigned subcarriers. The proposed approach outperforms the existing solutions, especially when the frequency diversity is small and the number of spatial modes is large.

REFERENCES

[1] J. Li, C. Botella, and T. Svensson, "Resource allocation for clustered network MIMO-OFDMA systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1 – 19, 2012.

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

[3] L.-N. Tran and E.-K. Hong, "Multiuser diversity for successive zero-forcing dirty paper coding: Greedy scheduling algorithms and asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3411 – 3416, 2010.

[4] R. F. H. Fisher, *Precoding and signal shaping for digital transmission*, Wiley, Ed. New York, 2002.

[5] V. Stankovic and M. Haardt, "Successive optimization Tomlinson-Harashima precoding (SO-THP) for multi-user MIMO systems," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*., vol. 3, 2005, pp. 1117 – 1120.

[6] Q. Zhou, H. Dai, and H. Zhang, "Joint Tomlinson-Harashima precoding and scheduling for multiuser MIMO with imperfect feedback," in *IEEE Wireless Commun. Networking Conf. (WCNC 2006)*, vol. 3, April 2006, pp. 1233 – 1238.

[7] L. Sanguinetti and M. Morelli, "Non-linear pre-coding for multiple-antenna multi-user downlink transmissions with different QoS requirements," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 852 – 856, 2007.

[8] A. A. D'Amico, "Tomlinson-Harashima precoding in MIMO systems: A unified approach to transceiver optimization based on multiplicative Schur-convexity," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3662 – 3677, Aug 2008.

[9] T. Maciel and A. Klein, "On the performance, complexity, and fairness of suboptimal resource allocation for multiuser MIMO-OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 406–419, 2010.

[10] Y. Zhang and K. Letaief, "An efficient resource-allocation scheme for spatial multiuser access in MIMO/OFDM systems," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 107–116, 2005.

[11] W. Ho and Y.-C. Liang, "Optimal resource allocation for multiuser MIMO-OFDM systems with user rate constraints," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1190 – 1203, 2009.

[12] N. Ul Hassan and M. Assaad, "Low complexity margin adaptive resource allocation in downlink MIMO-OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3365 – 3371, 2009.

[13] M. Moretti and A. Perez-Neira, "Efficient margin adaptive scheduling for MIMO-OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 278 – 287, 2013.

[14] I. Kim, I.-S. Park, and Y. H. Lee, "Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM," *IEEE Trans. Veh. Technol.*, vol. 55, no. 4, pp. 1195 – 1207, 2006.

[15] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Int. Symposium on Signals, Systems, and Electronics*, 1998, pp. 295 – 300.

[16] K. Kusume, M. Joham, W. Utschick, and G. Bauch, "Efficient Tomlinson-Harashima precoding for spatial multiplexing on flat MIMO channel," in *IEEE International Conference on Communications*, vol. 3, 2005, pp. 2021 – 2025.

[17] M. Moretti, A. Todini, A. Baiocchi, and G. Dainelli, "A layered architecture for fair resource allocation in multicellular multicarrier systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1788 – 1798, 2011.

[18] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ 08458: Prentice Hall PTR, 1993.

[19] L. Sanguinetti, A. A. D'Amico, and Y. Rong, "On the design of amplify-and-forward MIMO-OFDM relay systems with QoS requirements specified as Schur-convex functions of the MSEs," *IEEE Trans. Veh. Technol.*, vol. 52, no. 5, Jan 2013.

[20] B. Huang and T. Jebara, "Fast b-matching via sufficient selection belief propagation," in *Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

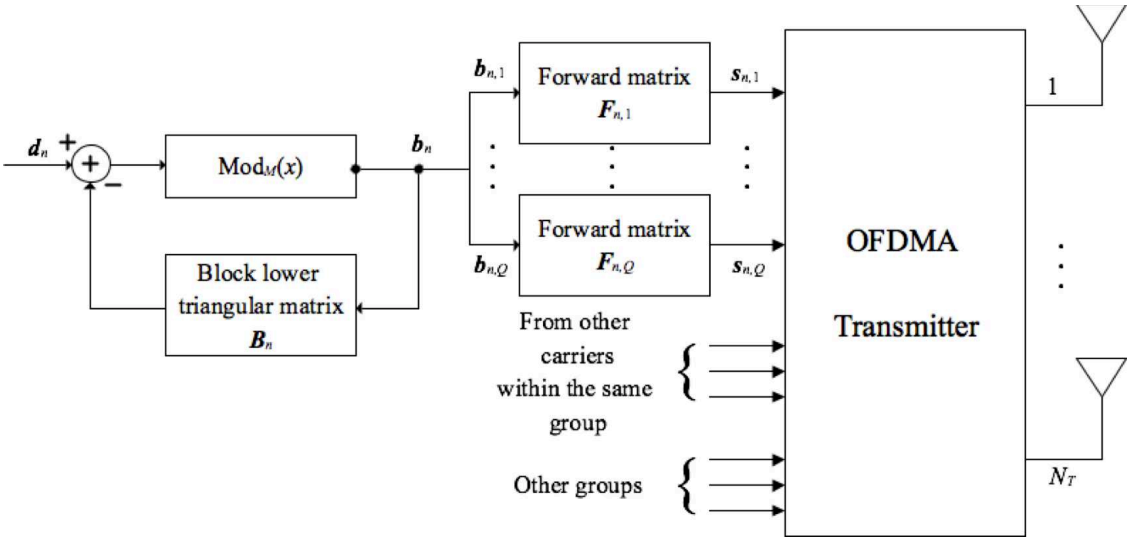

Fig. 1. Block diagram of the THP technique operating at user level.

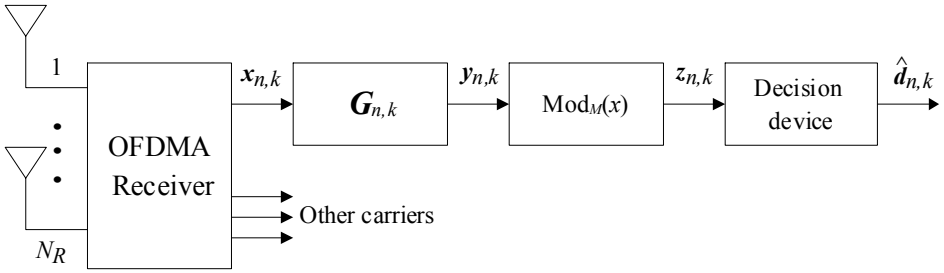

Fig. 2. Block diagram of the receiver at each MT.

TABLE I

COMPUTATIONAL LOAD

Operation	Flops
Computing quantities $\{\pi(k)\}$	$\mathcal{O}(NKN_TN_R)$
Evaluating the SVD of $\bar{\mathbf{H}}_{n,k}$	$\mathcal{O}(Q/2(Q-1)N_RN_T^2)$
Evaluating the SVD of $\mathbf{H}'_{n,k}$	$\mathcal{O}(Q/2(Q-1)N_TN_R^2+QN_RN_T^2)$
Solving the LP problem in (13)	$\mathcal{O}\left(Q(N+K/Q)^{2.5}\right)$
Computing all matrices $\{\mathbf{B}_n\}$	$\mathcal{O}(NQ/2(Q-1)(N_R^3+4LN_R^2))$

Algorithm 1 Proposed two-layer architecture

for user $k = 1$ to K **do**
  Compute $\pi(k) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{tr}\left(\mathbf{H}_{n,k}^{H} \mathbf{H}_{n,k}\right)$.
end for
Sort users according to $\pi(k)$ and group them in Q sets $\{\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(Q)}\}$.
for group $i = 1$ to Q **do**
  for user $k = 1$ to $|\mathcal{S}^{(i)}|$ **do**
    for subcarrier $n = 1$ to N **do**
      Compute the power cost $c_{n,k}$ according to (19).
    end for
  end for
  Solve the resource allocation problem in (13).
  for subcarrier $n = 1$ to N **do**
    Compute $\mathbf{B}_n = \mathbf{C}_n - \mathbf{I}_{LQ}$.
  end for
  for user $k = 1$ to $|\mathcal{S}^{(i)}|$ **do**
    for subcarrier $n = 1$ to N **do**
      Compute $\{\mathbf{F}_{n,k}, \mathbf{G}_{n,k}, \mathbf{U}_{n,k}\}$.
    end for
  end for
end for

TABLE II

SIMULATION SCENARIOS

	$S^{(1)}$	$S^{(2)}$	$S^{(3)}$
MIMO configuration	2×1	4×2	8×4
bandwidth $W^{(i)}$ (MHz)	10	5	2.5
# subcarriers $N^{(i)}$	64	32	16
# streams per subcarrier per user $L^{(i)}$	1	2	4
# subcarriers per user $n_k^{(i)}$	8	4	2

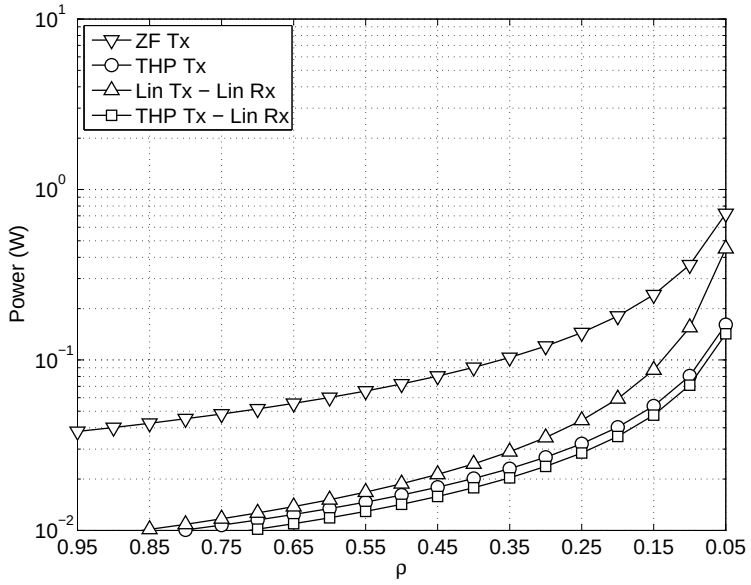

Fig. 3. Total power consumption for the investigated solutions in scenario $S^{(1)}$ for different MSEs for data stream.

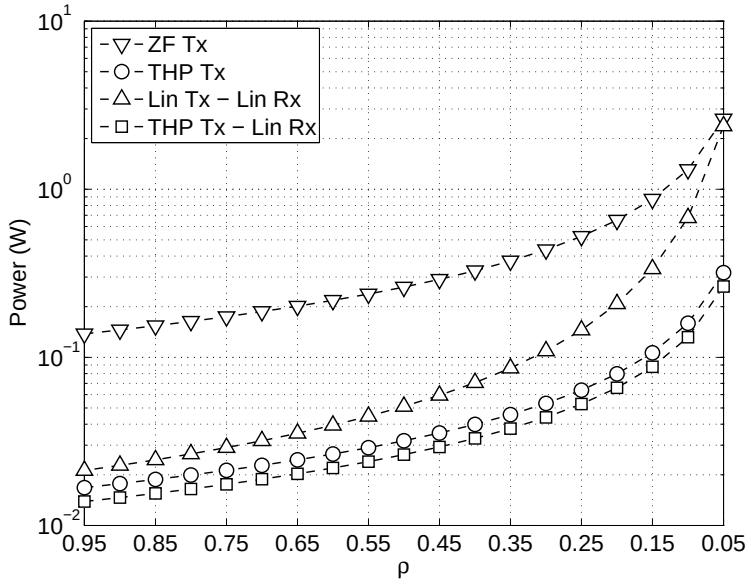

Fig. 4. Total power consumption for the investigated solutions in scenario $S^{(2)}$ for different MSEs for data stream.

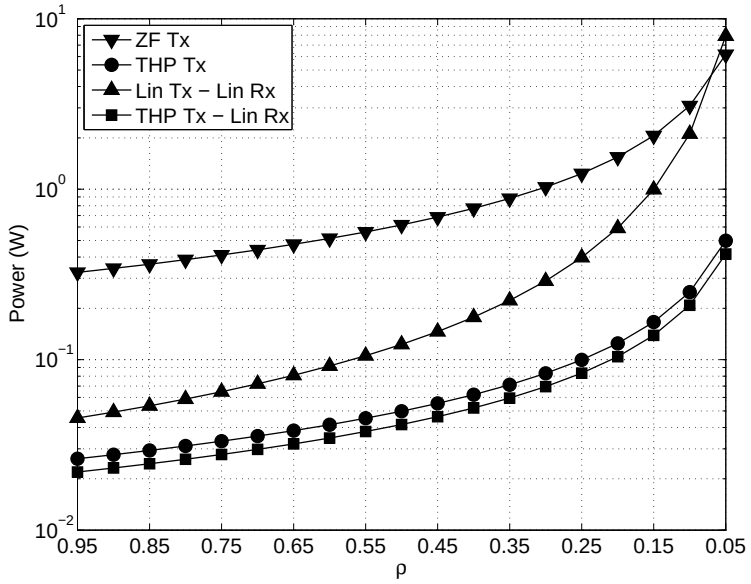

Fig. 5. Total power consumption for the investigated solutions in scenario $S^{(3)}$ for different MSEs for data stream.

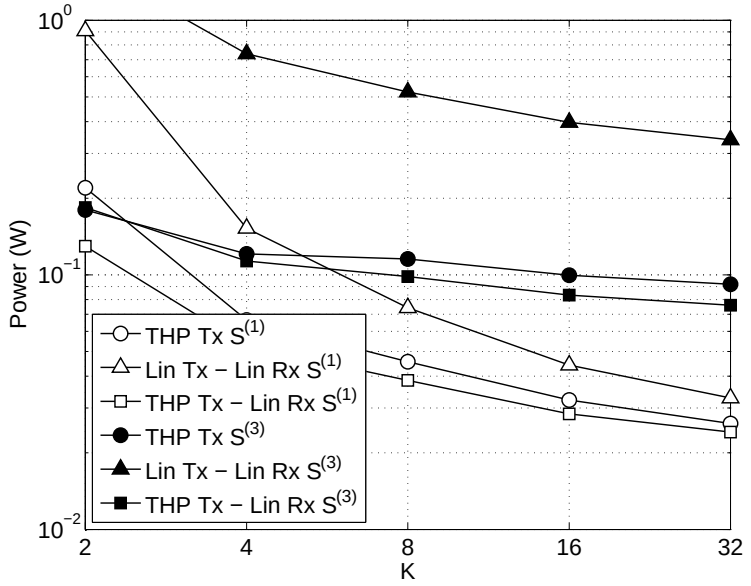

Fig. 6. Total power consumption for the investigated solutions for different number of users.