A NOVEL TREE-BASED SCHEDULING ALGORITHM FOR THE DOWNLINK OF MULTI-USER MIMO SYSTEMS WITH ZF BEAMFORMING

Martin Fuchs, Giovanni Del Galdo, and Martin Haardt

Ilmenau University of Technology, Communications Research Laboratory
P.O. Box 10 05 65, 98684 Ilmenau, Germany
{martin.fuchs, giovanni.delgaldo, martin.haardt}@tu-ilmenau.de
http://www.tu-ilmenau.de/crl

ABSTRACT

Spatial multiplexing in the downlink of wireless multiple antenna communications promises high gains in system throughput. However, spatially correlated users and a limited number of antennas at the base station motivates the need for a scheduling algorithm which efficiently arranges users into groups to be served in different time or frequency slots. In this paper we propose a novel tree-based scheduling algorithm which successfully solves this problem achieving a close to optimum grouping strategy. The algorithm has been tested with zero forcing beamforming techniques and is based on a new metric for the user performance considering the effect of other users present in the same group analyzing their spatial features.

1. INTRODUCTION

The use of Space-Division Multiple Access (SDMA) on the downlink of a multi-user MIMO wireless communications system can offer a substantial gain in system throughput. The gain is due to the fact that more than one user can be served using the same resource in time and frequency. When choosing a zero forcing (ZF) beamforming technique and assuming perfect channel knowledge at the transmitter, we guarantee that the data streams sent to different users will not interfere with each other. The goal of a scheduling algorithm is to arrange the users in groups such that all users belonging to a group can be efficiently multiplexed in space, while the different groups are served in different time or frequency slots. As the size of a group grows, i.e., as more users are added to a group, the SDMA gain grows. At the same time, however, it will be increasingly difficult to find efficient modulation matrices which fulfill the zero interference constraint. These two factors play against each other, making the choice of the best grouping allocation non trivial.

The novel grouping algorithm that we propose in this paper is capable of organizing an arbitrary number of users into groups based on their spatial properties. Different optimization criteria can be employed, such as maximizing the system sum capacity or guaranteeing a minimum data rate for all users.

The authors gratefully acknowledge the partial support of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under contract no. HA 2239/1-1.

2. DATA MODEL

Let M_{T} be the number of antennas at the transmitter, i.e., the base station, and M_{R} the total number of antennas for the K mobiles on the downlink of a multi-user MIMO system. Assuming flat fading and considering one time snapshot, we can write the received vector $\boldsymbol{y}$ at the M_{R} receive antennas as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{M}\boldsymbol{d} + \boldsymbol{n} \ \in \mathbb{C}^{M_{\mathrm{R}} \times 1}, \tag{1}$$

where the users' individual channel matrices are stacked in $\boldsymbol{H} = \left[\boldsymbol{H}_1^T \cdots \boldsymbol{H}_K^T\right]^T \in \mathbb{C}^{M_{\mathrm{R}} \times M_{\mathrm{T}}}$ such that $\boldsymbol{H}_i$ is the channel matrix for the i-th user. The column vector $\boldsymbol{d} = \left[\boldsymbol{d}_1^T \cdots \boldsymbol{d}_K^T\right]^T$ contains the users' data vectors. Assuming that $r = \mathrm{rank}\{\boldsymbol{H}\}$, the matrix $\boldsymbol{M} \in \mathbb{C}^{M_{\mathrm{T}} \times r}$ is a modulation matrix containing the individual users' modulation matrices $\boldsymbol{M}_i$ such that $\boldsymbol{M} = [\boldsymbol{M}_1 \cdots \boldsymbol{M}_K]$. The elements of the noise vector $\boldsymbol{n}$ are i.i.d. complex Gaussian. Let the matrix $\tilde{\boldsymbol{H}}_i$ contain the channels of all users except for the i-th, stacked as follows:

$$\tilde{\boldsymbol{H}}_i = \left[\ \boldsymbol{H}_1^T \ \cdots \ \boldsymbol{H}_{i-1}^T \ \ \boldsymbol{H}_{i+1}^T \ \cdots \ \boldsymbol{H}_K^T \ \right]^T. \tag{2}$$

A zero forcing beamforming technique yields a modulation matrix $\boldsymbol{M}_i$ whose columns lie in the common null space of all other users, i.e., in the left null space of $\tilde{\boldsymbol{H}}_i$. As a consequence, the data streams directed to the i-th user do not interfere with those sent to other users, although, depending on the beamforming chosen, they might interfere with each other, as for the block diagonalization (BD) [1].

3. THE TREE-BASED SCHEDULING ALGORITHM

The aim of a scheduling algorithm is to allocate users in groups such that all users belonging to the same group can be spatially multiplexed via the beamforming algorithm, while different groups are served in different time or frequency slots. The larger the group the greater will be the gain achieved by the SDMA (Space Division Multiple Access) scheme. Unfortunately, at the same time, the zero interference constraint will increasingly impede the selection of efficient modulation matrices.

The algorithm is based on a metric function $\eta_i^{(j,k,\ldots,m)} \geq 0$ that measures the efficiency of the transmission to the i-th user when grouped together with users $j, k, \ldots, m$. We assume that the maximum of $\eta_i^{(j,k,\ldots,m)}$ is achieved when the users' signal spaces are (spatially) orthogonal and $\eta_i^{(j,k,\ldots,m)} = 0$ if user i has the same signal space as all other users in the group. The metric function

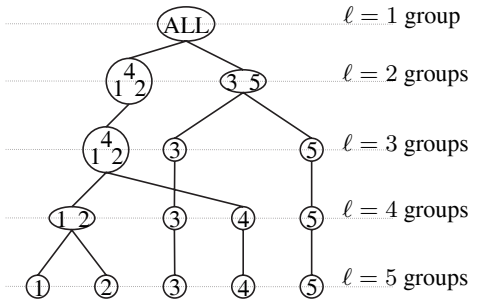

Fig. 1. An example for the tree-based scheduling algorithm: possible groupings of $K = 5$ users are generated based on a performance metric.

can be an estimate of the channel capacity, of the SNR, or any quantitative measure of the user's performance. Section 4 proposes such a metric for Zero Forcing beamforming techniques.

The algorithm consists of building a tree like the one depicted in Figure 1. Let ℓ denote the level of the tree, so that in the exemplary tree we have $\ell = 1, 2, \ldots, 5$. At the lowest level, corresponding to $\ell = 5$, we have 5 groups, one for each user. Each level ℓ of the tree represents an assignment of users into ℓ groups, as indicated by the connecting lines. For instance, for $\ell = 2$ we have two groups, the first being composed by users 1, 2, and 4. The remaining users form the second group. The tree can be constructed following either a *bottom up* or a *top-down* approach.

In the *bottom up* strategy, we start from the roots of the tree. The tree is built with an iterative procedure. At each step we have one group less. This is achieved by merging two groups of the previous level. In order to decide which of the $\binom{\ell}{2}$ possible pairs of groups to join we can follow different criteria. We can choose the pair which penalizes its members the least in terms of their η (in order to introduce a notion of quality of service). Alternatively, we choose the pair which has the best average $\overline{\eta}$ over the number of users in the group. For instance, in Figure 1, the $\ell = 4$ level is derived by merging the first two users. Following the best average criterion it means that $\frac{1}{2}(\eta_1^{(2)} + \eta_2^{(1)})$ was the maximum compared to all possible combinations of $\frac{1}{2}(\eta_j^{(i)} + \eta_i^{(j)})$ for $i \neq j$. Similarly, for $\ell = 3$, we merge the group composed by 1 and 2 with the group of user 4. Therefore the maximum, among all combinations, is achieved for $\frac{1}{3}(\eta_1^{(2,4)} + \eta_2^{(1,4)} + \eta_4^{(1,2)})$.

Alternatively the tree can be obtained by starting from $\ell = 1$, i.e., with all users in one group, and proceeding *top-down*. To come to the next lower level, we split the group with the worst average performance into two smaller groups. We could also opt to divide the group containing the user with the worst performance (by comparing the minimum η_i of all groups). We choose the partitioning which yields the best two new groups. Once more we make the choice based on the group performance.

The problem of finding the best number of groups to be used in the scheduling is often worked around in other scheduling methods such as in [2], where the number of groups is chosen to be equal to the number of highly correlated users simply based on a threshold decision. In [3] the problem is treated by imposing lower and upper bounds on the number of users per group to limit the number of combinations to be tested, thus excluding possible optimum group allocations. Our algorithm allows us to solve this problem, when the metric $\eta_i^{(\cdots)}$ is an estimate of the channel capacity or achievable user rate. In this case it is possible to choose the number of groups to be equal to the level ℓ that has the best predicted overall system capacity $\hat{C}_{\mathrm{sys},\ell}$

$$\hat{C}_{\mathrm{sys},\ell} = \frac{1}{\ell} \sum_{i=1}^{K} \eta_i^{(\cdots)}, \tag{3}$$

where $\frac{1}{\ell}$ accounts for the SDMA gain, and the i-th user's performance $\eta_i^{(\cdots)}$ is computed with respect to the users present in his group, denoted by the superscript $^{(\cdots)}$. Note that as the groups become larger, i.e., for smaller values of ℓ, the SDMA gain grows. However, at the same time, the individual performance $\eta_i^{(\cdots)}$ will presumably decrease, since the zero forcing constraint becomes increasingly stringent.

Alternatively, if the metric $\eta_i^{(\cdots)}$ is an estimate of the i-th user's SNR, then we can estimate the average SNR for the ℓ-th level, $\hat{\rho}_{\mathrm{sys},\ell}$, as

$$\hat{\rho}_{\mathrm{sys},\ell} = \frac{1}{K} \sum_{i=1}^{K} \eta_i^{(\cdots)}, \tag{4}$$

and choose the highest level which achieves a target average SNR, which corresponds to minimizing the transmit power.

In *real world applications*, a scheduling algorithm should avoid unnecessary changes in the group assignment to keep the signaling overhead low. Furthermore, it must be capable of dynamically adding and removing users. We can achieve both using the following procedure: the tree is calculated once at the beginning using either the bottom-up or top-down method and the best solution is stored. When the next channel estimate is available, we recalculate the metrics for the current level and build one level above and one below. In this way, at any given time, only three levels of the tree must be computed and the tree will adapt itself dynamically to the new system conditions. To add a user, we create a new group for it at the current tree level and let the tree evolve as previously explained. The same is done when one mobile leaves the system.

4. ESTABLISHMENT OF A NEW METRIC

In this section we propose a metric $\eta_i^{(j,k,\ldots,m)}$ that provides a quantitative measure of the efficiency of the transmission to the i-th user when this user is spatially multiplexed in the same time or frequency slot with users $j, k, \ldots,$ and m. Our goal is to avoid the calculation of the beamforming vectors during the scheduling process, otherwise required in other grouping schemes, such as in [4], where the exact user's receive SINR is used as a metric.

In Section 2 it has been illustrated that any ZF beamforming leads to a modulation matrix $\boldsymbol{M}_i$ which lies in the null space of $\tilde{\boldsymbol{H}}_i$, which we will denote as $\tilde{\mathcal{N}}_i$. In other words, the i-th modulation matrix is forced to excite only that part of the signal space of $\boldsymbol{H}_i$, denoted by $\mathring{\boldsymbol{H}}_i$, which lies in $\tilde{\mathcal{N}}_i$. The channel capacity of the i-th user C_i, assuming perfect channel knowledge at both link ends under the zero-interference constraint is

$$C_i = \log_2 \left(\det \left(\boldsymbol{I} + \frac{P}{\sigma_n^2 \cdot g} \cdot \mathring{\boldsymbol{H}}_i \, \mathring{\boldsymbol{V}}_i \, \boldsymbol{\Gamma}_i \, \mathring{\boldsymbol{V}}_i^H \, \mathring{\boldsymbol{H}}_i^H \right) \right), \tag{5}$$

where we assume that P is the total power available at the transmitter and $\frac{P}{g}$ is the power available to each of the g users present in the group. The power of the noise is σ_n^2 and the matrix $\mathring{\boldsymbol{V}}_i$ contains an orthonormal basis spanning the row space of $\mathring{\boldsymbol{H}}_i$. The diagonal

matrix $\mathbf{\Gamma}_i$ contains the power loading coefficients γ_i which are obtained from the water pouring algorithm so that $\sum \gamma_i = 1$ as, for instance, in [5]. The matrix $\boldsymbol{I}$ is an identity matrix of proper size. The SNR of the i-th user, ρ_i under the same assumptions is

$$\rho_i = \frac{P \cdot \left\| \mathring{\boldsymbol{H}}_i \right\|_F^2}{\sigma_n^2 \cdot g}, \tag{6}$$

where $\| \cdot \|_F^2$ is the Frobenius norm squared.

We can choose either the channel capacity C_i or the SNR ρ_i as our metric $\eta_i^{(j,k,\ldots,m)}$. However their exact computation would be prohibitive from the computational complexity point of view. In fact, the equivalent channel $\mathring{\boldsymbol{H}}_i$ is computed by projecting $\boldsymbol{H}_i$ into $\tilde{\mathcal{N}}_i$ via the projection matrix $\tilde{\boldsymbol{P}}_i^{(0)}$

$$\begin{aligned} \mathring{\boldsymbol{H}}_i &= \boldsymbol{H}_i \cdot \tilde{\boldsymbol{P}}_i^{(0)}, \quad \text{with} \\ \tilde{\boldsymbol{P}}_i^{(0)} &= \tilde{\boldsymbol{V}}_i^{(0)} \cdot \left[\tilde{\boldsymbol{V}}_i^{(0)} \right]^H, \end{aligned} \tag{7}$$

where the columns of the matrix $\tilde{\boldsymbol{V}}_i^{(0)}$ are the vectors of an orthonormal basis spanning the left null space of $\tilde{\boldsymbol{H}}_i$. The more users are in the group, the smaller will be $\tilde{\mathcal{N}}_i$ thus reducing the power in $\mathring{\boldsymbol{H}}_i$. The computation of $\tilde{\boldsymbol{P}}_i^{(0)}$ requires a Singular Value Decomposition (SVD) for each user in each potential group considered by the scheduling algorithm. In order to drastically reduce the computational complexity we propose an approximation to compute the projection matrix $\tilde{\boldsymbol{P}}_i^{(0)}$.

Assume that $\tilde{\mathcal{N}}_i$ is the common null space of users $j, k, \ldots, m$. Thus, $\tilde{\mathcal{N}}_i$ is the intersection of the individual null spaces

$$\tilde{\mathcal{N}}_i = \mathcal{N}_j \cap \mathcal{N}_k \cdots \cap \mathcal{N}_m, \tag{8}$$

where $\mathcal{N}_j$ is the left null space obtained from the j-th channel. Unfortunately, computing $\tilde{\mathcal{N}}_i$ from the individual null spaces still requires several SVD's. However, the projection matrix into the common null space $\tilde{\boldsymbol{P}}_i^{(0)}$ can be approximated with $\hat{\tilde{\boldsymbol{P}}}_i^{(0)}$, obtained by n repeated projections onto the individual null spaces $\mathcal{N}_j$

$$\tilde{\boldsymbol{P}}_i^{(0)} \approx \hat{\tilde{\boldsymbol{P}}}_i^{(0)} = \left(\boldsymbol{P}_j^{(0)} \cdot \boldsymbol{P}_k^{(0)} \cdot \ldots \cdot \boldsymbol{P}_m^{(0)} \right)^n, \tag{9}$$

where the projection matrix $\boldsymbol{P}_i^{(0)}$ on the i-th user's null space can be efficiently computed from the orthonormal basis $\boldsymbol{V}_i^{(1)}$ of the user's signal space with:

$$\boldsymbol{P}_i^{(0)} = \boldsymbol{I} - \boldsymbol{V}_i^{(1)} (\boldsymbol{V}_i^{(1)})^H. \tag{10}$$

The approximation converges strongly to the exact solution, as described in [6].

The approximation in equation (9) still requires K SVD's, i.e., one for each of the K users, assuming that the mobiles employ more than one antenna each. In order to reduce the complexity of finding the complete basis $\boldsymbol{V}_i^{(1)}$ further, we use a rank one approximation $\hat{\boldsymbol{v}}_i^{(1)}$, as proposed in [7]. We obtain it from the pivoted QR decomposition of $\boldsymbol{H}_i^T$, so that $\boldsymbol{H}_i^T \boldsymbol{\Pi} = \boldsymbol{Q}\boldsymbol{R}$. The matrix $\boldsymbol{\Pi}$ permutes the columns of $\boldsymbol{H}_i^T$ such that the columns of $\boldsymbol{Q}$ are sorted by their norm. The latter form a basis for the column space of $\boldsymbol{H}_i^T$. By using an iterative QR algorithm such as Stewart's SPQR [8], we can already stop after the first column of $\boldsymbol{Q}$ is found. The SPQR uses one additional column of the matrix in each iteration to obtain one more column of $\boldsymbol{Q}$. As a result, $\hat{\boldsymbol{v}}_i$ is simply the normalized column of $\boldsymbol{H}_i^T$ with the highest norm.

The approximation introduced in equation (9) further reduces the computational complexity of building the tree. In fact, in higher levels of the tree we can reuse previously calculated projection matrices, since the same terms reappear.

The approximation allows us to efficiently estimate the power contained in the projected channel $\mathring{\boldsymbol{H}}_i$, and from this quantity we can easily estimate the SNR ρ_i as:

$$\hat{\mathring{\boldsymbol{H}}}_i = \boldsymbol{H}_i \cdot \hat{\tilde{\boldsymbol{P}}}_i^{(0)} \approx \mathring{\boldsymbol{H}}_i \tag{11}$$

$$\hat{\rho}_i = \frac{P \cdot \left\| \hat{\mathring{\boldsymbol{H}}}_i \right\|_F^2}{\sigma_n^2 \cdot g} \approx \rho_i. \tag{12}$$

In the scheduling algorithm, when considering the SNR as a metric for the user's performance we simply assign $\eta_i^{(\cdots)} = \hat{\rho}_i$. Knowing the SNR is not enough to calculate the channel capacity in a MIMO system. In fact, depending on how the power is distributed among the eigenvalues of $\mathring{\boldsymbol{H}}_i \cdot \mathring{\boldsymbol{H}}_i^H$ we obtain a different channel capacity. In order to estimate the capacity without computing equation (5), we can assume a fixed eigenvalue spread. For Line Of Sight (LOS) scenarios the channels will have a low rank and an uneven distribution is most probable. We then estimate the capacity assuming that all the power is condensed in one eigenvalue. The estimate for the channel capacity becomes:

$$\hat{C}_i^{\mathrm{LOS}} = \log_2 \left(1 + \frac{P \cdot \left\| \hat{\mathring{\boldsymbol{H}}}_i \right\|_F^2}{\sigma_n^2 \cdot g} \right) = \log_2 \left(1 + \hat{\rho}_i \right). \tag{13}$$

For a Non LOS case we choose the upper bound of the channel capacity, which corresponds to an equal distribution of power among the eigenvalues

$$\hat{C}_i^{\mathrm{NLOS}} = M_{\mathrm{R}} \cdot \log_2 \left(1 + \frac{\hat{\rho}_i}{M_{\mathrm{R},i}^2} \right), \tag{14}$$

where $M_{\mathrm{R},i}$ is the number of antennas at the i-th mobile. In the latter case, each of the $M_{\mathrm{R},i}$ eigenvalue is assumed to have a power equal to $\left\| \hat{\mathring{\boldsymbol{H}}}_i \right\|_F^2 / M_{\mathrm{R},i}$, while the power loading allocates via the waterpouring algorithm $1/M_{\mathrm{R},i}$-th of the power to each mode. According to the channel to be treated we can choose for the metric either $\eta_i^{(\cdots)} = \hat{C}_i^{\mathrm{NLOS}}$ or $\eta_i^{(\cdots)} = \hat{C}_i^{\mathrm{LOS}}$.

5. SIMULATION RESULTS

The effectiveness of our methods is shown in Figure 3 with the help of simulation results based on two different channel models. The uncorrelated channel, which we denote as $\boldsymbol{H}_w$, is composed of i.i.d complex Gaussian random variables with zero mean and unit variance. The correlated channel is generated with the *IlmProp*, a geometry based channel model developed at Ilmenau University of Technology [9]. Both channels are computed for a 12×12 system consisting of 6 users with 2 antennas, $\lambda/2$ spaced, each. The scenario generated with the IlmProp is shown in Figure 2. The base station, mounting a Uniform Circular Array (UCA) with 12 antennas, is placed in an environment with multiple clusters of scatterers. The orientation of the mobile arrays is approximately orthogonal to their trajectory. All users but the first one

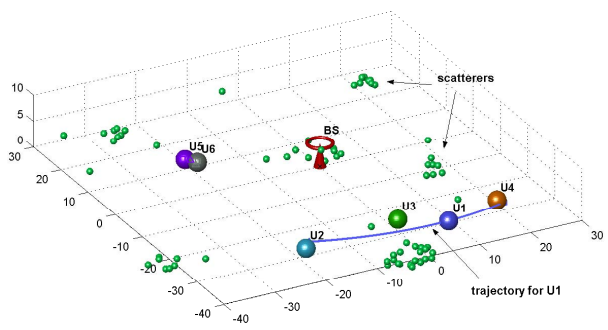

Fig. 2. Multi-User scenario generated with the IlmProp. The 6 users move around a base station mounting a 12 element Uniform Circular Array. The mobiles have Uniform Linear Arrays with two antennas each.

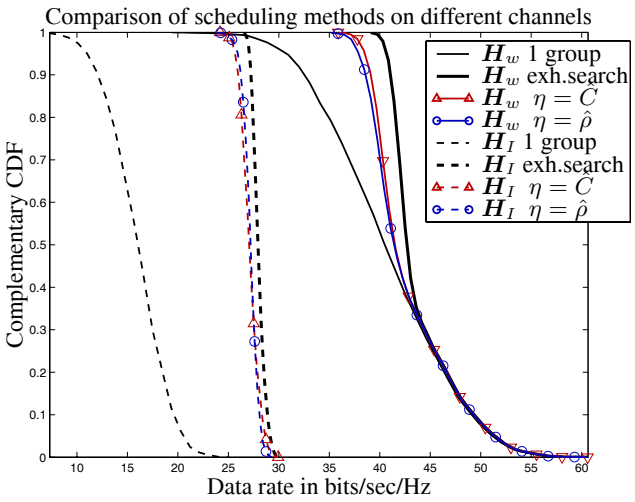

Fig. 3. CCDFs of data rates for a system with six users, simulated using the IlmProp channel ($\boldsymbol{H}_I$) as well as i.i.d. Gaussian channels ($\boldsymbol{H}_w$). The curves include the ones corresponding to the tree-based scheduling algorithm taking the capacity estimate and the SNR estimate as metrics, respectively. For comparison the curves for the 1 group solution and for the exhaustive search are plotted.

(U1) move only a few meters during the duration of the simulation. The first user starts from a position very close to the second user (U2) and travels towards U3.

Figure 3 depicts the Complementary Cumulative Distribution Functions (CCDF) of the channel capacities computed as in equation (5), i.e, assuming the zero-interference constraint. The one group solution achieves the worst capacity. An exhaustive search over all possible grouping strategies is used to show the best achievable rate. We plot for both channels the capacity curves obtained following the proposed tree-based scheduling algorithm. For one curve we use the estimate of the SNR as metric, for the other the estimate of the capacity. The level is chosen based on the performance obtained with the block diagonalization algorithm. The plots show the significant gains obtained with a proper scheduling algorithm, which are considerably greater in the case of the IlmProp channel. The reason is that the users in the IlmProp scenario have strongly correlated channels, due to their position. When allocated in one group only they will greatly interfere and thus the modulation matrices chosen by the BD algorithm will not be efficient. During our investigations we could establish that using the approximation technique of repeated projections does not significantly change the resulting performance, even if only first order projections are employed.

6. CONCLUSIONS

We introduce a novel algorithm to schedule the mobile stations in the downlink of a MIMO system for zero forcing - space division multiple access. The proposed tree-based algorithm is able to reach a close to optimum scheme using either an estimate of the user's capacity or an estimate of the user's SNR. Computational efficient techniques to compute these estimates have been shown. The scheduling is capable of grouping an arbitrary number of users for several optimization criteria, such as maximizing the system sum capacity or guaranteeing a minimum average SNR minimizing the transmit power.

7. REFERENCES

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

[2] D. Bartolome, A. Pascual-Iserte, and A. I. Perez-Neira, "Spatial scheduling algorithms for wireless systems," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China*, May 2003.

[3] Q. H. Spencer, *Transmission Strategies for Wireless Multi-user, Multiple-Input, Multiple-Output Communication Channels*, Ph.D. thesis, Brigham Young University, March 2004, http://contentdm.lib.byu.edu/ETD/image/etd378.pdf.

[4] H. Yin and H. Liu, "Performance of space-division multiple-access (SDMA) with scheduling," *IEEE Transactions on Wireless Communications*, vol. 1, no. 2, pp. 611–618, October 2002.

[5] A. Paulraj, R. Nabar, and D. Gore, *Introduction to space-time wireless communications*, Cambridge University Press, Cambridge, 2003.

[6] I. Halperin, "The product of projection operators," *Acta. Sci. Math. (Szeged)*, vol. 23, pp. 96–99, 1962.

[7] G. Del Galdo and M. Haardt, "Comparison of zero-forcing methods for downlink spatial multiplexing in realistic multi-user MIMO channels," in *Proc. IEEE Vehicular Technology Conference 2004-Spring, Milan, Italy*, May 2004.

[8] M.W. Berry, S.A. Pulatova, and G.W. Stewart, "Computing sparse reduced-rank approximations to sparse matrices," *Univ. of Tennessee Dep. of Computer Science Technical Reports*, vol. 525, 2004.

[9] G. Del Galdo, M. Haardt, and C. Schneider, "Geometry-based channel modelling of MIMO channels in comparison with channel sounder measurements," *Advances in Radio Science - Kleinheubacher Berichte*, pp. 117–126, October 2003.