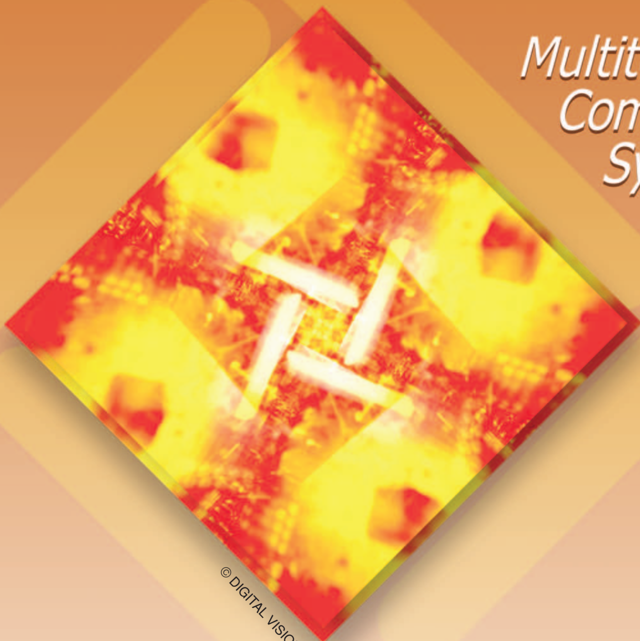
Multiterminal Communication Systems



[David Gesbert, Marios Kountouris, Robert W. Heath Jr., Chan-Byoung Chae, and Thomas Sälzer]

Shifting the MIMO Paradigm

[From single-user to multiuser communications]

The last ten years have witnessed the transition of multiple-input multiple-output (MIMO) communication from a theoretical concept to a practical technique for enhancing performance of wireless networks [1]. Point-to-point (single-user) MIMO communication promises large gains for both channel capacity and reliability, essentially via the use of space-time codes (diversity gain oriented) combined with stream multiplexed transmission (rate maximization oriented). In such a traditional single-user view of MIMO systems, the extra spatial degrees of freedom (DoF) brought by the use of multiple antennas are exploited to expand the dimensions available for signal processing and detection, thus acting mainly as a physical (PHY) layer performance booster. In this approach, the link layer protocols for multiple access (uplink and downlink) indirectly reap the performance benefits of MIMO antennas in the form of greater per-user rates or more reliable channel quality despite not requiring full awareness of the MIMO capability.

The situation with multiuser MIMO (MU-MIMO) techniques is radically different as these techniques imply the use of spatial sharing of the channel by the users, thus deeply affecting the design of the multiple access protocol. In spatial multiple access, the resulting multiuser interference is handled by the multiple antennas, which, in addition to providing per-link diversity, also give the DoF necessary for spatial separation of the users (see e.g. [1] Part IV). In practice, MU-MIMO schemes with good complexity/performance tradeoffs can be implemented to realize these ideas. On the uplink or multiple access channel (MAC), the development of MU-MIMO techniques appears as a generalization of known single-user

Digital Object Identifier 10.1109/MSP.2007.904815

1053-5888/07/$25.00©2007IEEE

MIMO (SU-MIMO) concepts to the multiuser case. As usual in information theory, the downlink or broadcast channel (BC) case is by far the most challenging one. Information theory reveals that the optimum transmit strategy for the MU-MIMO BC involves a theoretical preinterference cancellation technique known as dirty paper coding (DPC) combined with an implicit user scheduling and power loading algorithm. In that respect, the role played by seminal papers such as [2] was fundamental. In turn, several practical strategies have recently been proposed to approach the rates promised in the MU-MIMO channel involving concepts such as linear and nonlinear channel-aware precoding, channel state feedback, and multiuser receivers. A number of corresponding scheduling and user selection algorithms have also been proposed, leveraging features of different MU-MIMO strategies.

MU-MIMO techniques and performance have begun to be intensely investigated because of several key advantages over SU-MIMO communications.

- MU-MIMO schemes allow for a direct gain in multiple access capacity [proportional to the number of base station (BS) antennas] thanks to so-called multiuser multiplexing schemes.
- MU-MIMO appears more immune to most of propagation limitations plaguing SU-MIMO communications such as channel rank loss or antenna correlation. Although increased correlation still affects per-user diversity, this may not be a major issue if multiuser diversity [3] can be extracted by the scheduler instead. Additionally, line-of-sight propagation, which causes severe degradation in single-user spatial multiplexing schemes, is no longer a problem in multiuser setting.
- MU-MIMO allows the spatial multiplexing gain at the BS to be obtained without the need for multiple antenna terminals, thereby allowing the development of small and cheap terminals while intelligence and cost is kept on the infrastructure side.

The advantages above unfortunately come at a price. Perhaps the most substantial cost is due to the fact that MU-MIMO requires (although benefits from) channel state information at transmitter (CSIT) to properly serve the spatially multiplexed users. CSIT, while not essential in SU-MIMO communication channels, is of critical importance to most downlink multiuser precoding techniques. The need for CSIT feedback places a significant burden on uplink capacity in most systems, exacerbated in systems with wideband [e.g. orthogonal frequency division multiplexing (OFDM)] communication or high mobility (such as 3GPP-LTE [4], WiMax [5], etc.). Finally, another challenge related to MU-MIMO cross-layer design lies in the complexity of the scheduling procedure associated with the selection of a group of users that will be served simultaneously. Optimal scheduling involves exhaustive search whose complexity is exponential in the group size and depends on the choice of precoding, decoding, and channel state feedback technique.

Inspection of recent literature reveals several different schools of thought on the MU-MIMO downlink, each advocating a different combination of precoding, feedback, and scheduling strategies. Precoding strategies include linear minimum mean square error (MMSE) or zero-forcing (ZF) techniques and nonlinear approaches. Examples of the latter are vector perturbation, DPC techniques, and Tomlinson-Harashima precoding (THP) (a number of references are listed below). Many different feedback strategies have been suggested, including vector quantization, dimension reduction, adaptive feedback, statistical feedback, and opportunistic spatial division multiple access (SDMA). Finally, a number of scheduling disciplines have been suggested, including max-rate techniques, greedy user selection, and random user selection.

PROMISES AND CHALLENGES OF MU-MIMO NETWORKS

LESSONS LEARNED FROM MULTIUSER INFORMATION THEORY

SYSTEM AND SIGNAL MODEL

Progress in the field of multiuser information theory has been instrumental in understanding the fundamental nature and limits of the gains associated with exploiting multiple antennas in wireless networks, often also suggesting ideas for actual algorithms. We now review some aspects of MU-MIMO information theory with an eye for the key lessons learned from this field towards practical system design. A complete study of MU-MIMO information theoretic progress is beyond the scope of this article. Good references on the topic include [6] and [1, Ch. 18 and 19].

We focus on the communication between a BS or an access point equipped with N antennas, and U active terminals, where each active user k is equipped with M_k antennas. Among all terminals, the set of active users is roughly defined by the set of users simultaneously downloading or uploading packets during one given scheduling window. The length of the window is arbitrary but should not exceed the maximum latency expected by the application (likely as small as a few tens of milliseconds to several hundred milliseconds). By all means the active users over one given window will be a small subset of the connected users, themselves forming a small subset of the subscribers. We consider both the uplink and downlink but will emphasize on the challenges associated with the downlink for several reasons explained later.

In the uplink, the received signal at the BS can be written as

$$\mathbf{y} = \sum_{k=1}^{U} \mathbf{H}_k^T \mathbf{x}_k + \mathbf{n}, \tag{1}$$

where $\mathbf{x}_k$ is the $M_k \times 1$ user signal vector, possibly encompassing power-controlled, linearly combined, constellation symbols. $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$ represents the flat-fading channel matrix and $\mathbf{n}$ is the independent and identically distributed (i.i.d.), unit-variance, additive Gaussian noise vector at the BS. We assume that the receiver k has perfect and instantaneous knowledge of the channel $\mathbf{H}_k$. We focus on the flat-fading model here for the sake of exposition. Wideband models, using OFDM for example, can be accommodated by using a dependency on a frequency index. The transpose operator is simply used by convention for consistence with the downlink notation and does not presume a reciprocal link.

In the downlink illustrated in Figure 1, the received signal at the kth receiver can be written as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k \qquad \text{for} \qquad k = 1, \ldots, U, \tag{2}$$

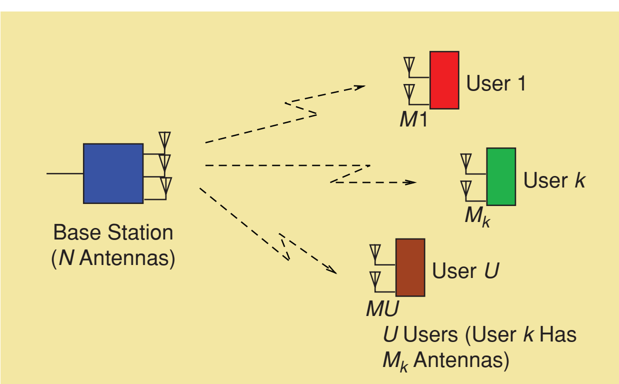
where $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$ represents the downlink channel and $\mathbf{n}_k \in \mathbb{C}^{M_k \times 1}$ is the additive Gaussian noise at receiver k. We assume that each receiver also has perfect and instantaneous knowledge of its own channel $\mathbf{H}_k$. The transmitted signal $\mathbf{x}$ is a function of the multiple users' information data, an example of which takes the superposition form $\mathbf{x} = \sum_k \mathbf{x}_k$ where $\mathbf{x}_k$ is the signal carrying, possibly nonlinearly encoded, user k's message, with covariance $\mathbf{Q}_k = \mathbb{E}(\mathbf{x}_k\mathbf{x}_k^H)$, with $\mathbb{E}(\cdot)$ the expectation operator. The power allocated to user k is therefore given by $P_k = \mathrm{Tr}(\mathbf{Q}_k)$, where Tr is the trace operator. Under a sum power constraint at the BS, the power allocation needs to maintain $\sum_k P_k \leq P$.

Assuming a unit variance for the noise, it is now known that the capacity region for a given matrix channel realization can be written as [7]:

$$\mathcal{C}_{BC} = \bigcup_{P_1,..P_U\ s.t.\ \sum_k P_k = P} \left\{ (R_1, ..R_U) \in \Re^{+U},\ R_i \leq \log_2 \frac{\det\left[\mathbf{I} + \mathbf{H}_i(\sum_{j \geq i} \mathbf{Q}_j)\mathbf{H}_i^H\right]}{\det\left[\mathbf{I} + \mathbf{H}_i(\sum_{j > i} \mathbf{Q}_j)\mathbf{H}_i^H\right]} \right\}, \tag{3}$$

where the expression should in turn be optimized over each possible user ordering. Although difficult to realize in practice, the computation of the region above is facilitated by exploiting the so-called duality results between the BC and the much simple to obtain MAC capacity region, which stipulate that the BC region can be calculated through the union of regions of the dual MAC with all uplink power allocation vectors meeting the sum power constraint P [8], [9].

The fundamental role played by the multiple antennas at either the BS or the users in expanding the channel capacity is best apprehended by examining how the sum rate (the point yielded by the maximum $\sum_k R_k$ in the capacity region) scales with the number of active users.



[FIG1] Downlink of a multiuser MIMO network. A BS communicates simultaneously with several multiple antenna terminals.

Assuming a block fading channel model and an homogeneous network where all users have the same signal-to-noise ratio (SNR), the scaling law of the sum rate capacity of MIMO Gaussian BC, denoted as $\mathcal{R}^{DPC}$ for $M_k = M$, fixed N and P, and large U is given by [10]

$$\lim_{U \to \infty} \frac{\mathbb{E}(\mathcal{R}^{DPC})}{N \log\log(UM)} = 1. \tag{4}$$

The result in (4) indicates that, with full CSIT, the system can enjoy a multiplexing gain of N, obtained by the BS sending data to N carefully selected users out of U. Since each user exhibits M independent fading coefficients, the total number of DoF for multiuser diversity is UM, thus giving the extra gain $\log\log(UM)$.

In contrast with (4), the capacity obtained in a situation where the BS is deprived from the users' channel information is reduced to (in the high SNR regime)

$$\mathbb{E}(\mathcal{R}^{NoCSIT}) \approx \min(M, N) \log SNR. \tag{5}$$

DESIGN LESSONS

Information theory highlights several fundamental aspects of MU-MIMO systems, which are in contrast much with the conventional SU-MIMO setting. First, the results above advocate for serving multiple users simultaneously in a SDMA fashion, with a suitably chosen precoding scheme at the transmitter. Although the multiplexing gain is limited by the number of transmit antennas, the number of simultaneously served users is, in principle, arbitrary. How many and which users should effectively be served with nonzero power at any given instant is the problem addressed by the resource allocation algorithm. Unlike in the single-user setting, the spatial multiplexing of different data streams can be done while users are equipped with single antenna receivers, thus enabling the capacity gains of MIMO while maintaining a low cost for user terminals. Having multiple antennas at the terminal can thus be viewed as optional equipment allowing extra diversity gain for certain users or giving the flexibility toward interference canceling and multiplexing of several data streams to such users (but reducing the number of other users served simultaneously). In addition to yielding MIMO multiplexing gains without the need for MIMO user terminals, the multiuser setup presents the advantage of being immune with respect to the possible ill-behavior of the propagation channel, which often plagues SU-MIMO communications, i.e., rank loss due to small spacing and/or the presence of strong line-of-sight component thanks to the wide physical separation between the users.

Finally, also in contrast with the conventional SU-MIMO setting, the multiplexing factor N in the downlink comes at the condition of channel knowledge at the transmitter. In the uplink this multiplexing gain is more easily extracted because the BS can be safely assumed to have uplink channel knowledge and simply implements a classical multiuser receiver to separate the contributions of the selected users in (1).

In the downlink, in the absence of CSIT, user multiplexing is generally not possible, as the BS just does not know in which direction to form spatial beams. Thus, the complete lack of channel state information (CSI) knowledge reduces the multiplexing gain to unity [11]. The exception lies in scenarios with terminal devices having enough antennas to remove costream interference at the receiver ($M_k \geq N$). In the latter case, the base may decide to either multiplex several streams to a single user or spread the streams over multiple users, achieving an equivalent multiplexing gain in both cases. This is conditioned however on the individual user channels to be full rank. Hence, the advantage of having CSIT in MU-MIMO lies in the possibility of not only serving single antenna users but also relaxing the dependence on single-user channel full rank.

[THE ADVANTAGES OF MU-MIMO TECHNIQUES AND PERFORMANCE OVER SU-MIMO COMMUNICATIONS UNFORTUNATELY COME AT A PRICE.]

MU-MIMO AND RESOURCE ALLOCATION

One of the fundamental lessons learned from information theoretic studies is that resource allocation techniques help to exploit the gains of MU-MIMO systems. From a multiuser information theoretic perspective, the capacity region boundary is achieved by serving all U active users simultaneously, where U is possibly a large number. The resource that should be allocated to each one, in the form of, e.g., P_k, is surely dependent on the instantaneous channel conditions and may vary greatly from user to user. The fact that the multiplexing gain is limited to N also suggests that the number of users effectively served with nonzero P_k at any given instant of time is directly related to the number of antennas at the BS, which is considerably less than the number of active cell users. Studies show in fact that the optimal number of users with nonzero allocated power for any given realization of the channel is upper bounded by N^2 [12]. In the remainder of the article we shall refer to this subset of users as the selected users. When restricting to linear precoding techniques such as ZF, the number of served users is directly limited by the number of DoF at the BS, N. This motivates the need to pick a good set of users, which is the aim of the resource allocation algorithm. In particular, the scheduler selects among all possible active users, for each channel realization, an optimal subgroup of terminals and respective power levels within the subgroup, so as to maximize a given performance metric. Such a metric can be the sum rate or the realization of per-user rate targets while minimizing transmit power. To maximize the sum rate, the scheduler algorithm looks for users that exhibit a compromise between a high level of instantaneous SNR (to maximize multiuser diversity [3]) and a good separability of their spatial signatures to facilitate user multiplexing. Practical and low complexity algorithms to solve the user scheduling problem are presented later in this article.

MU-MIMO SCHEMES WITH PERFECT CHANNEL KNOWLEDGE AT THE TRANSMITTER

LINEAR PRECODING

Linear precoding is a generalization of traditional SDMA, where users are assigned different precoding matrices at the transmitter. The precoders are designed jointly based on CSI of all the users based on any number of designs, including ZF and MMSE.

From a practical point of view, the relevant criteria are error probability, sum rate, signal-to-interference-plus-noise ratio (SINR), etc. The difficulty of designing capacity-optimal downlink precoding, mainly due to the coupling between power and beamforming and the user ordering, has lead to several different approaches ranging from transmit power minimization while maintaining individual SINR constraints to worst case SINR maximization under a power constraint. Duality and iterative algorithms are often used to provide solutions [13].

Consider the transmitted signal for user k given by $\mathbf{W}_k\mathbf{s}_k$, where $\mathbf{W}_k$ denotes the precoding matrix for the kth user and $\mathbf{s}_k$ is the symbol vector. We assume that service will be provided to a set of K selected users (among all active ones). Scheduling algorithms as discussed in the sequel can be applied to perform this selection across possible subsets. The received signal vector at the kth user is

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{W}_k\mathbf{s}_k + \mathbf{H}_k \sum_{l=1, l\neq k}^{K} \mathbf{W}_l\mathbf{s}_l + \mathbf{n}_k. \tag{6}$$

We assume that each user has M_k antennas and will decode the $S_k \leq M_k$ streams that constitute its data. The goal of linear precoding is to design $\{\mathbf{W}_k\}_{k=1}^{K}$ based on the channel matrix knowledge, so a given performance metric is maximized for each stream.

One of the simplest approaches for finding the precoder is to premultiply the transmitted signal by a suitably normalized ZF or MMSE inverse of the multiuser matrix channel [14], [15]. In this case, it can be assumed for simplification that $M_k = S_k = 1$. Thus, $\mathbf{H}_k = \mathbf{h}_k$ is a row vector and $\mathbf{W}_k$ (the precoding vector for the kth user) is chosen as the kth column of the right pseudoinverse (or MMSE inverse) of the composite channel $\left[\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_K^T\right]^T$. In the case when the selected users are not sufficiently separable, this approach may result in inefficient use of transmit power, causing a large rate loss with respect to the optimum sum capacity solution. This problem, however, is shown to be fixed by the scheduler when the number of active users to choose from is large enough so near-orthogonal users with good SNR conditions can be found. An additional disadvantage is that this approach does not readily extend to multiple receive antennas or streams without further degradation.

A generalization of the ZF or MMSE beamforming is to combine linear beamforming with a suitable power control policy to maximize the sum rate or realize individual SINR requirements for each user. Several approaches have been proposed, including maximizing the jointly achievable SINR margin under a total power constraint and minimizing the total transmission power while satisfying a set of SINR constraints [13]. Another generalization of ZF beamforming (ZFBF) is provided by block diagonalization (BD), which assumes $M_k = S_k \geq 1$ and $\sum_{k=1}^{K} M_k = N$. The idea is to choose $\mathbf{W}_k$ such that $\mathbf{H}_l\mathbf{W}_k = 0$, $\forall l \neq k$,thus precanceling the interference in (6) so that $\mathbf{y}_k = \mathbf{H}_k\mathbf{W}_k\mathbf{s}_k + \mathbf{n}_k$. If we define $\tilde{\mathbf{H}}_k$ as

$$\tilde{\mathbf{H}}_k = \left[\mathbf{H}_1^T \cdots \mathbf{H}_{k-1}^T \mathbf{H}_{k+1}^T \cdots \mathbf{H}_K^T\right]^T, \tag{7}$$

then any suitable $\mathbf{W}_k$ lies in the null space of $\tilde{\mathbf{H}}_k$. Let the singular value decomposition (SVD) of $\tilde{\mathbf{H}}_k$ be $\tilde{\mathbf{H}}_k = \tilde{\mathbf{U}}_k\tilde{\mathbf{D}}_k \left[\tilde{\mathbf{V}}_k^{(1)}\tilde{\mathbf{V}}_k^{(0)}\right]^H$, where $\tilde{\mathbf{U}}_k$ and $\tilde{\mathbf{D}}_k$ are the left singular vector matrix and the matrix of singular values of $\tilde{\mathbf{H}}_k$, respectively, and $\tilde{\mathbf{V}}_k^{(1)}$ and $\tilde{\mathbf{V}}_k^{(0)}$ denote the right singular matrices, each corresponding to nonzero singular values and zero singular values, respectively. Any precoder $\mathbf{W_k}$ that is a linear combination of the columns of $\tilde{\mathbf{V}}_k^{(0)}$ will satisfy the null constraint. Assuming that $\tilde{\mathbf{H}}_k$ is full rank, the transmitter requires that the number of transmit antennas is at least the sum of all users' receive antennas to satisfy the dimensionality constraint required to cancel interference for each user [16]. Under the BD constraint, $\mathbf{W}_k$ can be further optimized based on waterfilling. If excess antennas are available, eigenmode selection or antenna subset selection can be used to further improve performance [17].

A disadvantage of BD is that it requires $M_k = S_k$. This can be solved by including the receive processing in the problem formulation. For example, with a linear receive combining matrix $\mathbf{V}_k$ for user k, the received signal can be expressed as

$$\mathbf{y}_k = \mathbf{V}_k^H\mathbf{H}_k\mathbf{W}_k\mathbf{s}_k + \mathbf{V}_k^H\mathbf{H}_k \sum_{l=1,l\neq k}^{K} \mathbf{W}_l\mathbf{s}_l + \mathbf{V}_k^H\mathbf{n}_k. \tag{8}$$

The design problem then becomes selecting $\{\mathbf{W}_k, \mathbf{V}_k\}_{k=1}^K$ jointly such that $\mathbf{V}_k^H\mathbf{H}_k \sum_{l=1,l\neq k}^{K} \mathbf{W}_l = \mathbf{0}, \ \forall k$. This is difficult to solve in closed form, thus several iterative solutions have been proposed, including, e.g., [18], [19]. In such approaches, the transmitter generally computes a new effective channel for each user k using the initial receive combining vector. Using this new effective channel, the transmitter recomputes the transmit filter $\mathbf{W}_k$ to enforce a zero interference condition, and the receive filter $\mathbf{V}_k$ for each user. The algorithm repeats this process until satisfying a convergence criterion. To extend this algorithm to multiple data streams for each user, the matrix of right singular vectors is used based on the number of data streams and is used to calculate the effective channel matrix [18]–[20]. To avoid the use of extra feedback between the users and the BS, the computation of all filters (transmit and receive) normally takes place at the BS. After this computation, either the users must acquire the effective combined channel or information about the transmit filters must be sent [19].

NONLINEAR PRECODING

Linear precoding provides reasonable performance but may remain far from DPC-like precoding strategies when the available set of active users to choose from is small. Nonlinear precoding involves additional transmit signal processing to improve error rate performance. In this section, we discuss two representative methods, one based on perturbation [21], the other based on a spatial extension of THP [22].

Vector perturbation uses a modulo operation at the transmitter to perturb the transmitted signal vector to avoid the transmit power enhancement incurred by ZF methods [21]. Finding the optimal perturbation involves solving a minimum distance type problem and thus can be implemented using sphere encoding or full search-based algorithms.

Let $\mathbf{H}$ denote a $K \times N$ multiuser composite channel, assuming each user has a single receive antenna. The idea of perturbation is to find a perturbing vector $\mathbf{p}$ from an extended constellation to minimize the transmit power. The perturbation $\mathbf{p}$ is found by solving

$$\mathbf{p} = \arg \min_{\mathbf{p}' \in A\mathcal{C}\mathbb{Z}^K} \|\mathbf{G}(\mathbf{s} + \mathbf{p}')\|^2, \tag{9}$$

where $\mathbf{G}$ is a some transmit matrix such that $\mathrm{Tr}(\mathbf{G}^H\mathbf{G}) \leq P$, $\mathbf{s}$ is a modulated transmitted signal vector, and the scalar A is chosen depending on the original constellation size (e.g., $A = 2$ for QPSK), and $\mathcal{C}\mathbb{Z}^K$ is the K-dimensional complex lattice. ZF or MMSE precoder can be used for the transmit matrix $\mathbf{G}$. A set of points is used to represent symbols that are congruent to the symbol in the fundamental region. After predistortion using ZF or MMSE precoder, the resulting constellation region also becomes distorted and thus it takes more power to transmit the original point than before distortion. Among the equivalent points, if the transmitter sends the point that is the one closest to the origin to minimize transmit power, the receiver finds its equivalent image inside the fundamental constellation region using a modulo operation. This problem can be regarded as K-dimensional integer-lattice least squares problem and thus search based algorithms can be implemented. There are other methods to simplify the search based methods [23].

Several algorithms have also been proposed based on variations of THP [22], [24]. THP was originally proposed for use with Z point one-dimension pulse amplitude modulation (PAM) signal as a temporal equalization. For this constellation, THP is the same as the inverse channel filter except that an offset-free modulo $2Z$ adder is used. If the result of the summation is greater than Z, $2Z$ is subtracted until the final result is smaller than Z. Similarly, if the result of the summation is less than $-Z$, $2Z$ is added until satisfying the peak constraint. While in the original THP, a single channel is equalized with respect to time, spatial equalization is required for MIMO channels.

So far, we reviewed linear and nonlinear MU-MIMO solutions to approximate the sum capacity. In Figure 2, we compare sum capacity and achievable sum rates for DPC, coordinated beamforming [19], time sharing single-user closed loop MIMO (choosing only one user having the best channel quality and applying the SVD), and ZFBF with the dimensionality constraint [25]. In this case, no scheduling algorithm is required for DPC, coordinated beamforming, and ZFBF. We investigate scheduling issues below. Note that for the $(T, 1, T)$ scenario (i.e., the user has only one receive antenna while the BS has T transmit antennas and there are T active users in the network), there is a big gap between DPC and ZFBF, but this gap is decreased when the receivers have multiple antennas. For additional tradeoff analysis between linear and nonlinear precoding strategies, see also [26].

RESOURCE ALLOCATION TECHNIQUES HELP TO EXPLOIT THE GAINS OF MU-MIMO SYSTEMS.

In the following section, we consider the problem of choosing a subset of users for transmission in the MIMO BC. A brute-force complete search over all possible combinations of users guarantees maximizing the throughput, but the computational complexity is prohibitive when the number of users is large. Due to the complexity of the search process, both optimal and suboptimal approaches are considered. A key idea for low-complexity multiuser scheduling is that of greedy search.

OPTIMAL SCHEDULING FOR THE MU-MIMO DOWNLINK

The previous theoretical capacity results illustrate that, in general, the MIMO BC results in transmission to more than one user at a time. The problem of selecting a subset of users for transmission is a user scheduling problem, and the gain is achieved in a form of multiuser diversity. In this section we summarize some scheduling algorithms for different MU-MIMO solutions.

Linear beamforming can achieve the sum capacity when the number of active users in the system is large [10], [25], [27]. In [25], the users are equipped with only one receive antenna, and ZFBF is performed at the transmitter. Analogous to BD, this full search-based user selection algorithm can be extended to the multiple stream scenario. For simplicity, in this section, we assume that the number of receive antennas is equal to the number of data streams, where the postcoder V is not needed, and thus BD can be implemented.

Suppose $\mathcal{U} = \{1, 2, \cdots, U\}$ is the set of all users, and $\mathcal{A}_k$ one possible subset of selected users in $\mathcal{U}$. Let $\mathcal{A}$ be the set including all possible $\mathcal{A}_i$, i.e., $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots\}$. Then total achievable rate with BD is given by

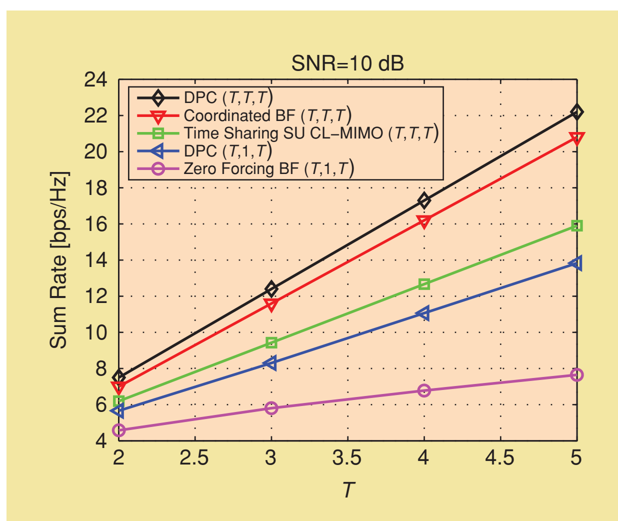
$$R_{BD|\mathcal{A}_k}(\mathrm{H}_{\mathcal{A}_k}, P, \sigma^2) = \max_{\sum_{j\in\mathcal{A}_k}\mathrm{Tr}(\mathrm{Q}_j)\le P} \sum_{j\in\mathcal{A}_k} \log\left|\mathrm{I} + \frac{\mathrm{H}_j \mathrm{W}_j \mathrm{Q}_j \mathrm{W}_j^H \mathrm{H}_j^H}{\sigma^2}\right|, \quad (10)$$

where $\mathrm{Q}_j = \mathbb{E}(\mathrm{x}_j \mathrm{x}_j)^H$ is the input covariance matrix for the user j, W_j is the precoding matrix earlier defined, and the same noise variance σ^2 is assumed at all users. Therefore, the maximum total sum rate with BD is given by $R_{BD}(\mathrm{H}_{1,\dots,U}, P, \sigma^2) = \max_{\mathcal{A}_k\in\mathcal{A}} R_{BD|\mathcal{A}_k}(\mathrm{H}_{\mathcal{A}_k}, P, \sigma^2)$. Denote $\mathcal{S}$ as the maximum number of users to be supported. For the case of BD, $\mathcal{S} \le N$. Thus, the cardinality of $\mathcal{A}$ is $\sum_{i=1}^{\mathcal{S}} C_U^i$, where C_a^b is the combination of a choosing b. Hence, it is clear that the exhaustive search over all possible combinations is computationally prohibitive when the number of users in the system is increased, and thus low-complexity user selection algorithm is desired.

GREEDY AND ITERATIVE METHODS FOR USER GROUPING

The complexity of the optimal scheduling is high, thus there has been several suboptimal algorithms that were proposed to reduce the computational complexity for user group selection [25], [27]–[29].

In the capacity-based greedy user selection algorithm, the transmitter chooses the first user with the highest channel capacity. Then, it finds the next user that provides the maximum sum rate from the remaining unselected users. The algorithm is repeated until K users are selected. Clearly, the complexity of the capacity-based greedy user selection is no more than $U \times K$ user sets, which greatly reduces the complexity compared to the exhaustive search method explained in the previous section. Note that the full search method needs to consider roughly $\mathcal{O}(U^K)$ possible user sets. The sum rate can be obtained under a number of transmit schemes, including optimal nonlinear precoders. Scheduling for the nonlinear precoders mentioned previously is an ongoing topic of research, though few results have appeared, including a greedy user selection for ZF DPC (ZFDPC), which has been proposed in [27].



[FIG2] Ergodic sum capacity and achievable sum rate as a function of the number of users, the number of transmit/receive antennas. (T_1, T_2, T_3) denotes the number of transmit antennas at the BS, the number of receive antennas at the user, and the number of active users in the network, respectively. Coordinated BF refers to the method presented in [19].

LIVING WITH PARTIAL CHANNEL KNOWLEDGE AT THE TRANSMITTER

QUANTIZATION-BASED TECHNIQUES

Quantization is the first idea that comes to mind when dealing with source compression, in this case, the random channel matrix or the corresponding precoders being the possible sources. The amount of feedback depends on the frequency of feedback (generally a fraction of the coherence time), the number of parameters being quantized, and the resolution of the quantizer. Most research focuses on reducing the number of parameters and the required resolution. The feedback problem has been solved in SU-MIMO communication systems using a concept known as limited feedback precoding [30]. The key idea of this line of research has been to quantize the precoder for a MIMO channel and not simply the channel coefficients. The challenge of extending this work to the multiuser channel is that the transmit precoder depends on the channels of the other users in the system.

CHANNEL QUALITY METRIC DESIGN IS ONE OF THE LARGELY OPEN CHALLENGES IN MU-MIMO.

Other methods for reducing feedback in MU-MIMO channels assume a single receive antenna at the mobile—extensions to multiple receive antennas is an ongoing research topic. Some of the main results on this subject are due to [31], [32], where the random codebook and Grassmannian quantization ideas are used to quantize the direction of each user's channel $\mathbf{h}_k$. The main observation in [31] is that the feedback requirements scale linearly both as the number of transmit antennas grows and as a function of the SNR (in dB), unlike the single-user case. The reason is that quantization error introduces an SINR floor since it prohibits perfect inter-user cancellation. Thus, this error must diminish for higher SNRs to allow for a balancing between the noise and the residual interference due to channel quantizing. An improvement can be obtained by quantizing the channel vector and a certain received SINR upper bound that is a function of the error between the true and quantized channel [33]. This increases the performance of the system and helps in user selection. Thresholds based on sum rate constraints on the feedback channel can also be used to reduce required feedback yet maintain capacity scaling [34].

DIMENSION REDUCTION AND PROJECTION TECHNIQUES

In addition to quantization-based approaches where the channel metric is discretized, dimension reduction techniques can be used that involve projecting the matrix channel onto one or more basis vectors known to the transmitter and receiver. In that way, the CSI matrix of size $M \times N$ is mapped into a p-dimensional vector with $1 \leq p \leq M \times N$, thus reducing the dimensionality of the CSI to p complex scalars (which in turn may be quantized). Once the projection is carried out, the receiver feeds back a metric $\varphi_k = f(\mathbf{H}_k)$, which is typically related to the square magnitude of the projected signal. Antenna selection methods fall into this category. In this case, the projection is carried out by the terminal itself. Alternatively, the projection can be the result of using a particular precoder at the BS. A good example of this approach is given by a class of algorithms using unitary precoders. We now review this approach when $M_k = 1$ and the BS serves N users. In this case, the kth user channel is a $1 \times N$ row vector denoted by $\mathbf{h}_k$. The BS designs an arbitrary unitary precoder $\mathbf{Q}$ of size $N \times N$, further scaled for power constraint. Each terminal identifies the projection of its vector channel onto the precoder by $\mathbf{h}_k \mathbf{Q}$, and reports an index and a scalar metric expressing the SINR measured under an optimal beamforming vector selection.

$$\varphi_k = \max_{1 \leq i \leq N} \frac{|\mathbf{h}_k \mathbf{q}_i|^2}{\sigma^2 + \sum_{j \neq i} |\mathbf{h}_k \mathbf{q}_j|^2}, \qquad (11)$$

where $\mathbf{q}_i$ denotes the ith column of $\mathbf{Q}$. The scheduling algorithm then consists in opportunistically assigning to each beamformer $\mathbf{q}_i$ the user that has selected it and has reported the highest SINR.

When the unitary precoder must be designed without any form of CSIT a priori, a scaled identity matrix can be used. In this case, the algorithm falls back to assigning a different selected user to each base antenna. In the small number of user case, the performance of such scheme is plagued by inter-user interference. Fortunately, interference tends to decrease as the number of users to choose in the cell becomes high.

When the dynamics of the system are limited (low mobility), the use of a fixed set of precoders may result in severe unfairness between the users. This problem can be alleviated by the randomization of the beamforming vectors. The so-called opportunistic random beamforming (ORBF) was initially proposed for single-user setting [35] and later generalized in [36]. The performance of these methods is illustrated below. The idea of [36] can be recast in the context above, assuming this time that $\mathbf{Q}$ is randomly generated at each scheduling period according to an isotropic distribution while preserving the unitary constraint. The intuition behind that scheme is that the columns $\mathbf{q}_i$, $i = 1, \ldots, N$, are like orthogonal beams, and if there are enough users in the cell, each beam will be aligned with a given user's channel while simultaneously being nearly orthogonal to the other selected users' channels. With this scheme, it is possible to spatially multiplex N users with a level of feedback given by one scalar and one index. In the case of a large number of active users, opportunistic multibeam schemes are shown to yield an optimal capacity growth of $N \log\log U$ for fixed N, which is precisely the scaling obtained with full CSIT, as shown in (4).

DEALING WITH SPARSE NETWORKS

A limitation of fixed or random opportunistic beamforming approaches is that the optimal capacity scaling emerges for a large, sometimes impractical, number of simultaneously active users in the cell. The performance degrades with a decreasing number of users (sparse networks), and this degradation is amplified when the number of transmit antennas increases, as intuition also reveals. The lack of robustness of these approaches in cases with small to moderate number of users is a serious problem that can be resolved by modifying the random beams for a better matching with the actual users' channels. This can be done at little or no extra feedback cost by one of several means. In one approach, the unitary constraint is relaxed by introducing a power control across the beams. The SINR feedback is used to adjust the power allocated to each beam [37] or simply to turn off certain beams [38], thus reducing interuser interference when the random beams are not well aligned with users' channels. In Figure 3, we compare the robustness of the single-beam ORBF [35] and multibeam ORBF [36], both with SINR feedback with respect to the number of active users in the cell. With four antennas at the BS, at 10 dB SNR, simulations suggest that at least 12 simultaneously active users are required for the multibeam gains to kick in. Whether this condition is met in practice or not is an interesting open research problem whose solution is likely to depend on the considered traffic, operational scenario, and delay constraint. With less users, the lack of CSIT destroys the benefits of user multiplexing. Interestingly, a strategy allowing for beam power control in multibeam ORBF [37] allows for a smooth transition between TDMA and SDMA regions, as shown in the figure.

Yet another approach is to exploit the second-order statistics of the channel, either in the temporal or in the spatial domain. The time domain approach consists in exploiting the natural temporal correlation of the channel to help refine the beams over time [39], [40]. In the spatial domain, statistics give information about spatial separability, which is instrumental to a proper beamforming design. Such aspect is described below.

INFORMATION THEORY HIGHLIGHTS SEVERAL FUNDAMENTAL ASPECTS OF MU-MIMO SYSTEMS.

USE OF SPATIAL STATISTICAL FEEDBACK

In practical, especially outdoor, networks, the i.i.d. channel model used so far does not hold, and each user tends to exhibit different channel statistics. The advantage of statistical CSI is its long coherence time compared with that of the fading channel. Several forms of statistical CSI are even reciprocal (i.e., holds for both uplink and downlink frequency) such as second-order correlation matrix, power of Ricean component, etc., and do not necessitate any feedback. Overall, spatial channel statistics reveal a great deal of information on the macroscopic nature of the underlying channel, including the multipath's mean angle of arrival/departure and its angular spread. More generally, a substantial amount of channel distribution information (CDI) is revealed by channel statistics, which can be used to infer knowledge on mean user separability. Clearly however, in fading channels, the CDI ought to be complemented with some form of instantaneous channel quality information (CQI) to extract multiuser diversity gain. Combining CDI and CQI can yield partial CSIT, which is very well suited to solving the scheduling stage of the MU-MIMO problem. It is an open topic for research, but some leads are presented below.

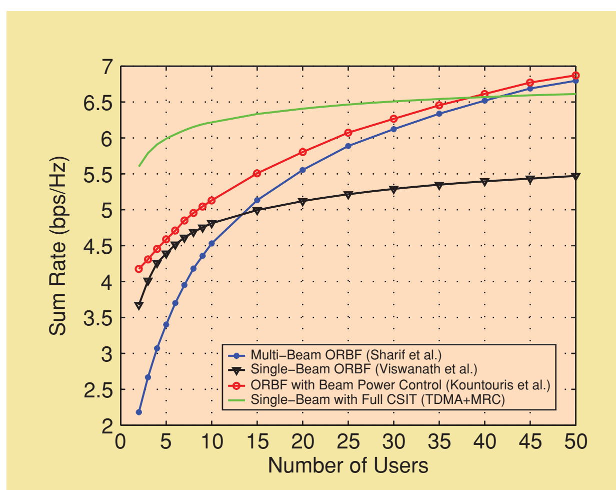
Consider the downlink of a network with single antenna mobiles, where the BS exhibits correlated transmit antennas. The channel is modeled as correlated Ricean fading, i.e., the channel vector of kth user satisfies $\mathbf{h}_k \sim \mathcal{CN}(\bar{\mathbf{h}}_k, \mathbf{R}_k)$, where $\bar{\mathbf{h}}_k \in \mathbb{C}^{1\times N}$ and $\mathbf{R}_k \in \mathbb{C}^{N\times N}$ are the mean value and transmit covariance matrix, respectively, known to the BS. A general form of CQI is

$$\gamma_k = \left\|\mathbf{h}_k \mathbf{Q}_k\right\|^2, \tag{12}$$

where $\mathbf{Q}_k \in \mathbb{C}^{N\times L}$ is a training matrix containing L orthonormal vectors $\{\mathbf{q}_{ki}\}_{i=1}^{L}$. Conditioned on the CQI feedback, a coarse estimate of the instantaneous channel realization and channel correlation at the transmitter can be calculated as the conditional expectations

$$\widehat{\mathbf{h}}_k = \mathbb{E}(\mathbf{h}_k|\gamma_k) \qquad \widehat{\mathbf{R}}_k = \mathbb{E}\left(\mathbf{h}_k^H \mathbf{h}_k|\gamma_k\right), \tag{13}$$

which can be used to provide an MMSE estimate of the instantaneous SINR [41]. Note that with $\mathbf{Q}_k = \mathbf{I}$, (12) falls back to a channel norm feedback.



[FIG3] The sum rate is compared for random beamforming schemes with SINR feedback. Multibeam (SDMA) random beamforming outperforms the single-beam (TDMA) when the number of active users is sufficient. Power control over the random beams allows for a smooth transition between TDMA and SDMA. TDMA with full CSIT outperforms partial feedback schemes for a small number of users but fails to provide multiplexing gain when this number increases.

Similarly, a maximum-likelihood (ML) estimation framework maximizing the log-likelihood function of the probability density function (pdf) of $\mathbf{h}_k$ under the scalar constraint (12) can be formulated [42]. Let $L = 1$, $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{R}_k)$ and CQI feedback $\gamma_k = |\mathbf{h}_k \mathbf{q}_k|^2$. The solution to the ML problem

$$\max_{\mathbf{h}_k} \mathbf{h}_k \mathbf{R}_k \mathbf{h}_k^H \; s.t. \; |\mathbf{h}_k \mathbf{q}_k|^2 = \gamma_k \qquad (14)$$

is given by the (dominant) generalized eigenvector associated with the largest positive generalized eigenvalue of the Hermitian matrix pair $(\mathbf{R}_k, \mathbf{q}_k \mathbf{q}_k^H)$. Once the coarse channel estimation is performed by the BS, it can be used to select up to N users according to any number of previously described performance metric based on CSIT. As a second stage, more complete CSIT may be requested by the BS only to the small set of selected users for a more accurate precoding design. The performance exceeds that of random beamforming but depends on the level of antenna correlation, i.e., angle spread σ_θ, as is shown in Figure 4. Certain techniques above are suited to specific deployments scenarios. For instance, opportunistic schemes are suited to densely populated networks. Schemes using temporal statistics are better suited to low mobility (indoor) setting, while the exploitation of spatial statistics would be more effective in outdoor cases where the elevation of the BS above the clutter decreases the angle spread of multipath and gives rise to Ricean models.

[THE IMPACT AND DESIGN OF AN OPTIMAL FORM OF CSIT UNDER FINITE RATE FEEDBACK IS STILL AN OPEN AND EXCITING PROBLEM.]

SYSTEM ISSUES

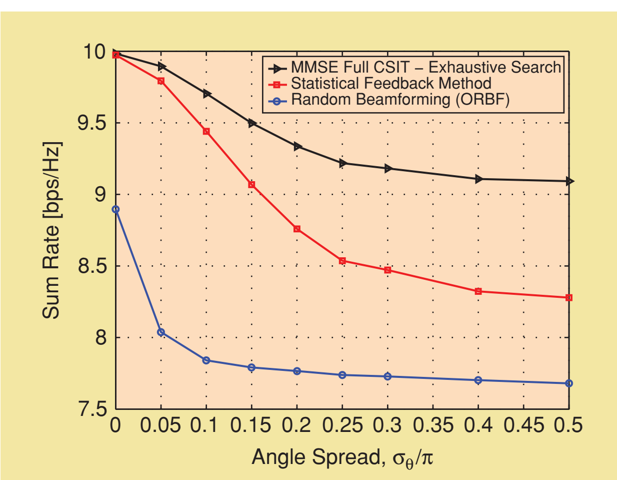
Although it is now widely recognized that MIMO techniques, in their generality, will be a key element in the evolution of broadband wireless access systems, applications of MU-MIMO solutions have yet to emerge. While spatial diversity and basic SU-MIMO techniques are available in several products and standards, adaptive antenna solutions, including MU-MIMO, are mostly considered for time division duplex (TDD) systems in low and moderate mobility where CSI can be obtained from estimation in the uplink. We believe, however, that the promise is such that these techniques will be eventually available in most systems.

Note that codebook based precoding schemes for SU- and MU-MIMO are emerging in existing and future standards [4]. MU-MIMO systems may have the potential to achieve the spectrum efficiency requirements set by operators for the next generation of mobile communication systems [43]. Practical MU-MIMO applications are still challenging however, and further studies seem needed to get a deeper understanding of the related tradeoffs and system gains (number of antennas, choice of algorithm, etc.).

When it comes to the crucial CSIT issue, one problem with designing feedback metrics is that the SINR measurement depends, among other things, on the number of other terminals being simultaneously scheduled along with the user making the measurement. Certain metrics (such as those in, e.g., [33], [36]) assume a fixed number of scheduled SDMA users. However, in practice, methods allowing fast transitions between TDMA and SDMA modes will be required. In such cases, the number of simultaneous users and the available power for each of them will generally be unknown at the terminal. Channel quality metric design in this scenario is one of the largely open challenges in MU-MIMO.

Also, opportunistic scheduling in MU-MIMO not only requires feedback for CSIT but also signaling of scheduling decisions to the terminal. The feedback and control loop in MU-MIMO introduces a nonnegligible overhead and latency in the system, which must carefully be weighed against the capacity gains expected from such techniques. Certain scenarios look promising (e.g., broadband best-effort internet access); others are more questionable, such as Voice over Internet Protocol (VoIP), where small packets are to be delivered with tight delay constraints. In addition, a poorly designed feedback channel can suffer from delays and cause the reported channel quality metrics to the transmitter to be outdated, bringing further degradations [44].

Another fundamental aspect is the impact of realistic traffic models and system loads, especially on schemes relying on high user loads (e.g., random beamforming). In recent wireless systems based on MIMO-OFDMA [5], opportunistic scheduling can be performed in up to three dimensions, namely, time, frequency, and space. Different types of traffic are likely to have different constraints with respect to the available DoF for the scheduler. For example, real-time services typically have tight delay constraints and limit the flexibility of the scheduler in the time domain. One may



[FIG4] Sum rate as a function of the angle spread σ_θ at the BS, where the number of transmit antennas is 2, the average SNR = 10 dB, and the number of active users in the cell is 50.

then wonder how many effective users are available for selection by the scheduler in each of these dimensions and how to take advantage of the different DoF to satisfy the QoS constraints for different types of traffic?

DISCUSSION

MU-MIMO networks reveal the unique opportunities arising from a joint optimization of antenna combining techniques with resource allocation protocols. Furthermore, it brings robustness with respect to multipath richness, allowing for compact antenna spacing at the BS and, crucially, yielding the diversity and multiplexing gains without the need for multiple antenna user terminals. To realize these gains, however, the BS should be informed with the user's channel coefficients, which may limit practical application to TDD or low-mobility settings. To circumvent this problem and reduce feedback load, combining MU-MIMO with opportunistic scheduling seems a promising direction. The success for this type of scheduler is strongly traffic and QoS-dependent, however. A number of complementary approaches geared toward feedback reduction were proposed, which may restore the robustness of MU-MIMO techniques with respect to a wider range of application and environments. These results and other performance studies with low feedback schemes suggest that MU-MIMO transmitters can cope with very coarse channel information. From a theoretical point of view, the impact and design of an optimal form of CSIT under finite rate feedback is still an open and exciting problem.

[MU-MIMO NETWORKS REVEAL THE OPPORTUNITIES ARISING FROM A JOINT OPTIMIZATION OF ANTENNA COMBINING TECHNIQUES WITH RESOURCE ALLOCATION PROTOCOLS.]

AUTHORS

David Gesbert (gesbert@eurecom.fr) is with Eurecom Institute, France. He obtained the Ph.D degree from ENST, France in 1997. Before joining Eurecom, he was with the Information Systems Laboratory, Stanford University, then a founding engineer of Iospan Wireless Inc, San Jose, CA, (now Intel), then with the Department of Informatics, University of Oslo. He has published 120 papers and several patents, all in the area of signal processing, communications, and wireless networks. He is a co-editor of five special issues on wireless networks and communications theory for *JSAC*, *EURASIP*, and *Wireless Communications Magazine*. He is a member of the IEEE Signal Processing for Communications Technical Committee. He co-authored three award-winning papers in 2003, 2004, and 2005. He coauthored the book *Space Time Wireless Communications: From Parameter Estimation to MIMO Systems* (Cambridge Press, 2006).

Marios Kountouris (kountour@eurecom.fr) received the diploma in electrical and computer engineering from the National Technical University of Athens, Greece in 2002 and the M.S. degree in electrical engineering from ENST Paris, France in 2004. He is currently working toward the Ph.D. degree at ENST—Eurecom Institute, France, funded by France Telecom R&D. During the summer of 2004, he interned with Samsung Advanced Institute of Technology, Korea, developing MU-MIMO precoding and scheduling techniques for 3GPP, and IEEE 802.16e systems. His research interests include precoding and resource allocation for multiuser MIMO systems and cross-layer design for wireless networks. He is a Student Member of the IEEE.

Robert W. Heath Jr. (rheath@ece.utexas.edu) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, in 1996 and 1997, respectively, and the Ph.D. from Stanford University in 2002, all in electrical engineering. From 1998 to 2001, he was with Iospan Wireless Inc, San Jose, California. In 2003, he founded MIMO Wireless Inc. Since January 2002, he has been with the Department of Electrical and Computer Engineering at The University of Texas at Austin, where he is an associate professor and a member of the Wireless Networking and Communications Group. His research interests cover a broad range of MIMO communication, including limited feedback techniques, antenna design, relaying, ad hoc networking, and scheduling algorithms as well as 60 GHz communication techniques. He is an editor for *IEEE Transactions on Communication*, a former associate editor for *IEEE Transactions on Vehicular Technology*, and a member of the Signal Processing for Communications Technical Committee of the IEEE Signal Processing Society. He is a Senior Member of the IEEE.

Chan-Byoung Chae (cbchae@ece.utexas.edu) is a Ph.D. student at the University of Texas, Austin. Prior to joining UT, he was a research engineer at the Telecommunication R&D center, Samsung Electronics from 2001 to 2005. He was a visiting research engineer at WING Lab, Aalborg University in Denmark, in 2004. He participated in the IEEE 802.16e standardization where he made several contributions and filed a number of related patents from 2004 to 2005. He was awarded the Bronze Prize in the 1996 Samsung Humantech International Paper Contest and the KSEA scholarship in 2007. His current research interests include capacity analysis and interference management in wireless mobile networks and all aspects of MIMO communications.

Thomas Sälzer (thomas.salzer@orange-ftgroup.com) is with France Telecom R&D since 2004 and currently responsible for standard activities related to mobile radio systems. He obtained degrees from the University of Stuttgart, Germany, and ENST Paris, France in 2000 as well as a Ph.D. from INSA Rennes, France in 2004. He worked for Mitsubishi Electric ITE from 2000 to 2004, where he was involved in European research projects. He holds several patents and is author and co-author of papers in the field of MIMO, OFDM, and cross-layer design.

REFERENCES

[1] H. Bölcskei, D. Gesbert, C. Papadias, and A.J. van der Veen, Eds., *Space-Time Wireless Systems: From Array Processing to MIMO Communications*, Cambridge, U.K.: Cambridge Univ. Press, 2006.

[2] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[3] R. Knopp and P. Humblet, "Information capacity and power control in single cell multi-user communications," in *Proc. IEEE Int. Conf. Commun.*, Seattle, WA, USA, June 1995, pp. 331–335.

[4] 3GPP, "Long Term Evolution, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; General description," *TS 36.201 v1.0.0*, Mar. 2007.

[5] IEEE, "Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands," IEEE Standard 802.16e-2005, Feb. 2006.

[6] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Select. Areas Commun.*, vol. 21, no. 5, pp. 684–702, June 2003.

[7] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.

[8] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple access and broadcast channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[9] P. Viswanath and D.N. Tse, "Sum capacity of the vector Gaussian channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[10] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 11–15, Jan. 2007.

[11] S.A. Jafar and A. Goldsmith, "Isotropic fading vector broadcast channels: The scalar upper bound and loss in degrees of freedom," *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 848–857, Mar. 2005.

[12] W. Yu and W. Rhee, "Degrees of freedom in wireless multiuser spatial multiplex systems with multiple antennas," *IEEE Trans. Commun.*, vol. 54, no. 10, pp. 1744–1753, Oct. 2006.

[13] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[14] 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.

[15] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.

[16] Q. Spencer, A.L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Processing*, vol. 52, no. 2, pp. 462–471, Feb. 2004.

[17] R. Chen, R.W. Heath Jr., and J.G. Andrews, "Transmit selection diversity for unitary precoded multiuser spatial multiplexing systems with linear receivers," *IEEE Trans. Signal Processing*, vol. 55, no. 3, pp. 1159–1571, Mar. 2007.

[18] Z. Pan, K.-K. Wong, and T.-S. Ng, "Generalized multiuser orthogonal space-division multiplexing," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 1969–1973, Nov. 2004.

[19] C.-B. Chae, D. Mazzarese, and R.W. Heath Jr., "Coordinated beamforming for multiuser MIMO systems with limited feedforward," in *Proc. Asilomar Conf. Sign., Syst. Computers*, Oct.–Nov. 2006, pp. 1511–1515.

[20] L. Choi and R.D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 773–786, July 2003.

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

[22] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.

[23] C. Windpassinger, R.F.H. Fischer, and J.B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2057–2060, Dec. 2004.

[24] R.F. Fischer, *Precoding and Signal Shaping for Digital Transmission*, New York: Wiley, 2002.

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

[26] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," in *Proc. Int. Zurich Sem. Commun.*, ETH Zurich, Switzerland, Feb. 2006.

[27] G. Dimić and N.D. Sidiropoulos, "On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.

[28] Z. Tu and R.S. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Commun. Lett.*, vol. 7, no. 8, pp. 370–372, Aug. 2003.

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

[30] D.J. Love, R.W. Heath Jr., W. Santipach, and M.L. Honig, "What is the value of limited feedback for MIMO channels?" *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 54–59, Oct. 2003.

[31] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[32] P. Ding, D. Love, and M. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Signal Processing*, June 2006.

[33] T. Yoo, N. Jindal, and A. Goldsmith, "Finite-rate feedback MIMO broadcast channels with a large number of users," in *Proc. IEEE Int. Symp. Inform. Theory*, Seattle, WA, USA, July 2006, pp. 3879–3891.

[34] K. Huang, R.W. Heath Jr., and J. Andrews, "Space division multiple access with a sum feedback rate constraint," *IEEE Trans. Signal Processing*, vol. 55, no. 7, pp. 3879–3891, July 2007.

[35] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[36] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channel with partial side information," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[37] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *Proc. IEEE Workshop Signal Processing Adv. Wireless Commun.*, New York, USA, June 2005, pp. 975–979.

[38] J. Wagner, Y.-C. Liang, and R. Zhang, "On the balance of multiuser diversity and spatial multiplexing gain in random beamforming," *IEEE Trans. Wireless Commun.,* 2007, to be published.

[39] M. Kountouris and D. Gesbert, "Memory-based opportunistic multi-user beamforming," in *Proc. IEEE Int. Symp. Inform. Theory*, Adelaide, Australia, Sept. 2005, pp. 1426–1430.

[40] D. Avidor, J. Ling, and C. Papadias, "Jointly opportunistic beamforming and scheduling (JOBS) for downlink packet access," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, June 2004, pp. 2959–2964.

[41] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Acquiring partial CSI for spatially selective transmission by instantaneous channel norm feedback," *IEEE Trans. Signal Processing*, Mar. 2007.

[42] M. Kountouris, D. Gesbert, and L. Pittman, "Transmit correlation-aided opportunistic beamforming and scheduling," in *Proc. Europ. Signal Processing Conf.*, Florence, Italy, Sept. 2006.

[43] "NGMN: 'Next Generation Mobile Networks Beyond HSPA and EVDO—A white paper', V3.0," [Online]. Available: http://www.ngmn-cooperation.com, Dec. 2006.

[44] M. Kobayashi, G. Caire, and D. Gesbert, "Transmit diversity vs. opportunistic beamforming in data packet mobile downlink transmission," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 151–157, Jan. 2007.

[SP]