From Single User to Multiuser Communications: Shifting the MIMO Paradigm

David Gesbert*, Marios Kountouris†, Robert W. Heath Jr.‡, Chan-Byoung Chae‡, Thomas Sälzer†

*Mobile Communications Department, Eurecom Institute, Sophia Antipolis, France, Email: gesbert@eurecom.fr

† France Telecom Research and Development, Issy-Les-Moulineaux, France, Email: {marios.kountouris,thomas.salzer}@orange-ftgroup.com

‡ Dept. of Electrical and Computer Engineering, The University of Texas, Austin, USA, Email: {rheath,cbchae}@ece.utexas.edu

Abstract

In multiuser MIMO networks, the spatial degrees of freedom offered by multiple antennas can be advantageously exploited to enhance the system capacity, by scheduling multiple users to simultaneously share the spatial channel. This entails a fundamental paradigm shift from single user communications, since multiuser systems can experience substantial benefit from channel state information at the transmitter and, at the same time, require more complex scheduling strategies and transceiver methodologies. This paper reviews multiuser MIMO communication from an algorithmic perspective, discussing performance gains, tradeoffs, and practical considerations. Several approaches including non-linear and linear channel-aware precoding are reviewed, along with more practical limited feedback schemes that require only partial channel state information. The interaction between precoding and scheduling is discussed. Several promising strategies for limited multiuser feedback design are looked at, some of which are inspired from the single user MIMO precoding scenario while others are fully specific to the multiuser setting.

I. INTRODUCTION

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 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 recent development of cross-layer techniques, aimed at the joint design of the PHY layer's modulation and link layer's multiple access protocols has begun to shatter this view. This is especially true in MIMO networking where the positive role played by the spatial dimension on multiple access and scheduling is now being recognized, replacing the simplistic view of MIMO as a pure PHY technology. A better understanding of the impact of MIMO antennas on multiuser communications is, by large, due to progress in the field of multiuser information theory [2]. Fundamental recent results in this area have hinted at how deeply connected PHY layer modulation/coding and link layer resource allocation and scheduling can be, at least when having overall optimum system design as objective. One interesting example of this is the conflict and degradation that may arise from certain uncoordinated designs at the PHY and link layer when both layers attempt to extract diversity (e.g. use of channel-hardening [3] single-user space-time codes at the PHY combined with multiuser diversity scheduling at the link layer).

Multiuser MIMO (MU-MIMO) information theory advocates for the use of spatial sharing of the channel by the users. Such a multiple access protocol implies an extra hardware cost (antennas and filters) but does not involve any bandwidth expansion, unlike say time-division (TDMA) or code-division (CDMA) multiple access[1]. In spatial multiple access, the resulting multiuser interference is handled by the multiple antennas which in addition to providing per-

[1]Classical multiple access protocols such as TDMA, CDMA, can be used on top of spatial multiple access.

link diversity also give the degrees of freedom 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 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 broadcast channel involves a theoretical pre-interference 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 [4] 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 non-linear 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.

Multiuser MIMO techniques and performance have begun to be intensely investigated because of several key advantages over single user 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 single user 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 [5] 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 base station 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

single user 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. OFDM) communication or high mobility (such as 3GPP-LTE [6], WiMax [7], 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 non-linear approaches. Examples of the latter are vector perturbation, DPC techniques and Tomlinson-Harashima precoding (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.

Paper organization and contributions: The goal of this article is to provide a unified view of the state-of-the-art in MU-MIMO communication, with particular emphasis on the fundamental differences with single user MIMO communication as well as on the cross-layer implications of MU-MIMO. We give an overview of some of the key promises and challenges of multiuser MIMO communications for use in tomorrow's high efficiency cellular networks, focusing on the more interesting MU-MIMO downlink. We focus on the paradigm shift occurring for MIMO techniques when transitioning from a traditional single user, PHY layer modulation/coding design approach to a multiuser, cross-layer design view of wireless communications. We first emphasize lessons learned from multiuser information theory in terms of i) capacity bounds and ii) multiple access/resource allocation design for MIMO networks. Promising signal processing techniques are reviewed and their complexity/performance tradeoffs are discussed. Different algorithms for dealing with channel state information feedback are discussed in detail including reciprocity, quantization, and opportunistic approaches. Connections with user scheduling techniques are highlighted and some joint MIMO transmission and scheduling procedures are presented. Per-

formance plots are shown for the most promising combination of precoding, feedback, and scheduling strategies. Finally we briefly describe system issues pertaining to MU-MIMO.

II. Promises and challenges of multiuser MIMO networks

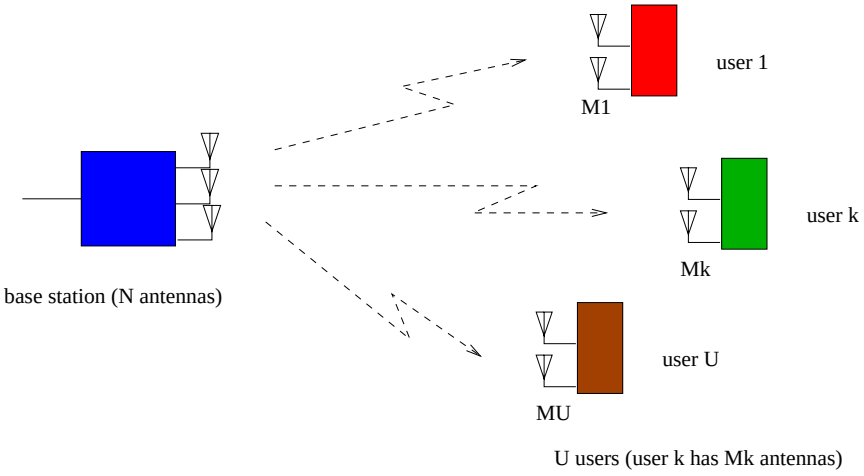

Fig. 1. Downlink of a multiuser MIMO network: A BS communicates simultaneously with several multiple antenna terminals.

A. Lessons learned from multiuser information theory

As often with important discoveries within the field of communications, the initial impulse responsible for attracting today the attention of a wide research community toward multiple antenna multiuser systems has not been of theoretical nature. Instead, current MU-MIMO ideas can be seen as the heirs to a long series of engineering advances started back in 1970s and 1980s in the area of antenna array-based communications. In fact it has been known for over three decades that making use of antenna arrays could enable the simultaneous communications with multiple users solely separated from their spatial signatures. This concept early on was labeled as SDMA, and is very closely related to that of today's MIMO spatial multiplexing, which can be interpreted as multiplexing the data streams of "virtual" users.

Nonetheless, 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 multiuser MIMO information theory, with an eye for the key lessons learned from

this field towards practical system design. A complete study of multiuser MIMO information theoretic progress is beyond the scope of this paper. Good references on the topic include [8], and [1], Chap. 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 ms to several hundred ms). 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[2] and $\mathbf{n}$ is the 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 $\mathbf{H}_k$.

In the downlink, illustrated in Fig.1, the received signal at the k-th receiver can be written as

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

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

$$\mathbf{x} = \sum_{k} \mathbf{x}_k \tag{3}$$

[2]We focus on the flat-fading model here for the sake of exposition. Wideband models, using e.g. OFDM, 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.

where $\mathbf{x}_k$ is the signal carrying, possibly non-linearly 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 $P_k = \text{Tr}(\mathbf{Q}_k)$, where Tr is the trace operator. Under a sum power constraint at the BS, the power allocation needs to maintain $\sum_k P_k \leq P$.

In contrast to single user systems where the capacity is a single number, the capacity of a multiuser system with U users is characterized by a U-dimensional rate region, where each point is a vector of achievable rates by all the U users simultaneously. Although the characterization of general broadcast capacity region is a long standing problem (unlike that in the MAC/uplink), substantial progress has been made for Gaussian MIMO channels. Despite not being degraded, the Gaussian MIMO BC channel offers significant structure that can be exploited to characterize its capacity region. Considering full CSIT, the particular role played by the dirty paper coding (DPC) in achieving points in the region was revealed by the seminal work [4]. 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 [9]:

$$\mathcal{C}_{BC} = \bigcup_{P_1,..P_U \;\; s.t. \;\; \sum_k P_k = P} \left\{ (R_1, ..R_U) \in \Re^{+U}, R_i \leq \log_2 \frac{\det\left[\mathbf{I} + \mathbf{H}_i (\sum_{j \geq i} \mathbf{Q}_j) \mathbf{H}_i^H\right]}{\det\left[\mathbf{I} + \mathbf{H}_i (\sum_{j > i} \mathbf{Q}_j) \mathbf{H}_i^H\right]} \right\} \quad (4)$$

where the above 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 simpler-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 [10], [11].

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 region) scales with the number of active users.

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 [12]

$$\lim_{U \to \infty} \frac{\mathbb{E}(\mathcal{R}^{DPC})}{N \log\log(UM)} = 1. \quad (5)$$

The result in (5) 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 degrees of freedom for multiuser diversity is UM, thus giving the extra gain $\log\log(UM)$.

In contrast with (5), the capacity obtained in a situation where the BS is deprived from the users' channel information is reduced to (in the high SNR regime)

$$\mathbb{E}(\mathcal{R}^{NoCSIT}) \approx \min(M,N)\log SNR. \tag{6}$$

1) Design lessons: Information theory highlights several fundamental aspects of multiuser MIMO systems, which come in much contrast with the conventional single user 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 non zero power at any given instant of time 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 single user MIMO communications, i.e. rank loss due to small spacing and/or the presence of strong line of sight component. In the multiuser case, the full rank of the global channel matrix is almost surely guaranteed thanks to the wide physical separation between the users.

Finally, also in contrast with the conventional single user 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. The exception lies in scenarios with terminal devices having enough antennas to remove co-stream 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.

Providing CSIT at the base poses serious challenges in practical settings where the channel information needs be conveyed via a limited feedback channel in the uplink. The often unrealistic assumption of close to perfect CSIT, as well as the considerable capacity gap between full and no CSIT, have motivated research work on schemes employing partial CSIT. Partial CSIT refers to any possible form of incomplete information on the channel, obtained by any of several means detailed later. Fortunately, work like [13] demonstrates that the optimal capacity scaling of capacity for the MIMO Gaussian BC, i.e. $N \log(SNR \log U)$ assuming U single antenna users, can be achieved for $U \to \infty$ even though the transmitter has only partial channel knowledge. Finding the information theoretic optimum strategy for exploiting partial channel knowledge at the transmitter is still an open and intriguing question, despite the many proposals in the literature, some of them being presented later in this paper.

2) 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 multiuser 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 non-zero 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 non zero allocated power for any given realization of the channel is upper bounded by N^2 [14]. In the remainder of the paper 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 degrees of freedom 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 [5]) *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 paper.

III. MU-MIMO SCHEMES WITH PERFECT CHANNEL KNOWLEDGE AT THE TRANSMITTER

We now turn to signal processing approaches to the MU-MIMO transmission problem. We choose to emphasize the downlink as it offers the most interesting challenges to a system designer. In the uplink, the signal model in (1) is clearly reminiscent of a classical multiuser detection problem [15] as far as receiver design is concerned and is not addressed here further.

As mentioned earlier, the maximum sum rate in the broadcast channel can be achieved by DPC at the BS [9]. The key idea of DPC is to pre-cancel interference at the transmitter using perfect CSI and complete knowledge of the transmitted signals. DPC, while theoretically optimal, is an information theoretic concept that has proven to be difficult to implement in practice. In this section we expand our study to a wider range of schemes, also relying on full CSIT, yet allowing a compromise between complexity and performance. We summarize several practical transmission techniques using either linear or non-linear precoding [16]–[19].

A. 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 and sum rate, maximizing 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 in order to provide solutions [20].

To explain the concept of linear precoding, consider the scenario where this time $\mathbf{s}_k$ and $\mathbf{n}_k$ denote the k-th transmit symbol vector (for beamforming scenario, $\mathbf{s}_k$ is a scalar symbol), and the additive white Gaussian noise vector. The actual transmitted signal vector for user k is then given by $\mathbf{W}_k\mathbf{s}_k$, where $\mathbf{W}_k$ denotes the precoding matrix for the k-th user. 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 k-th user is

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{W}_k\mathbf{s}_k + \mathbf{H}_k \sum_{l=1, l\neq k}^{K} \mathbf{W}_l\mathbf{s}_l + \mathbf{n}_k \tag{7}$$

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 [21], [22]. 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 k-th user) is chosen as the k-th column of the right pseudo-inverse (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 set to maximize the sum rate or realize individual signal-to-

interference-plus-noise ratio (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 [20]. The authors in [20] showed that the global optimum of beamforming can be obtained from solving a dual uplink problem and proposed a rapidly converging iterative algorithm to reduce crosstalk. Note that this approach does not maximize the achievable sum rate due to the SINR constraints, but does allow inclusion of different users' rate requirements in the problem formulation.

Another generalization of ZF beamforming 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 (7) 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

$$\tilde{\mathbf{H}}_k = \left[\begin{array}{cccccc} \mathbf{H}_1^T & \cdots & \mathbf{H}_{k-1}^T & \mathbf{H}_{k+1}^T & \cdots & \mathbf{H}_K^T \end{array} \right]^T \tag{8}$$

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

$$\tilde{\mathbf{H}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{D}}_k \left[\begin{array}{cc} \tilde{\mathbf{V}}_k^{(1)} & \tilde{\mathbf{V}}_k^{(0)} \end{array} \right]^H, \tag{9}$$

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 non-zero 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 [18]. 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 [23].

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

$$\mathbf{y}_k = \mathbf{V}_k^H \mathbf{H}_k \mathbf{W}_k \mathbf{s}_k + \mathbf{V}_k^H \mathbf{H}_k \sum_{l=1, l \neq k}^{K} \mathbf{W}_l \mathbf{s}_l + \mathbf{V}_k^H \mathbf{n}_k \tag{10}$$

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 cannot be solved in closed form, thus several iterative solutions have been proposed, including e.g. [17], [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 [16], [17], [19]. 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.

B. Non-linear 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. Non-linear precoding involves additional transmit signal processing to improve error rate performance. In this section, we discuss two representative methods, one based on perturbation [24], the other based on a spatial extension of Tomlinson-Harashima precoding (THP) [25].

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 [24]. 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

$$\mathbf{p} = \arg \min_{\mathbf{p}' \in A\mathcal{C}\mathbb{Z}^K} \|\mathbf{G}(\mathbf{s} + \mathbf{p}')\|^2 \tag{11}$$

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 which are congruent to the symbol in the fundamental region. After pre-distortion 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 [26].

Several algorithms have also been proposed based on variations of Tomlinson-Harashima precoding (THP) [25], [27]. THP was originally proposed for use with an 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 non-linear multiuser MIMO solutions to approximate the sum capacity. In Fig. 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 zero-forcing beamforming (ZFBF) with the dimensionality constraint [28]. In this case, no scheduling algorithm is required for DPC, coordinated beamforming and ZFBF. We will investigate scheduling issues in Section IV. 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 the big gap between DPC and ZFBF but this gap is decreased when the receivers have multiple antennas. For additional tradeoff analysis between linear and non-linear precoding stategies, the reader may also see [29].

IV. User scheduling in MU-MIMO networks

In this 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

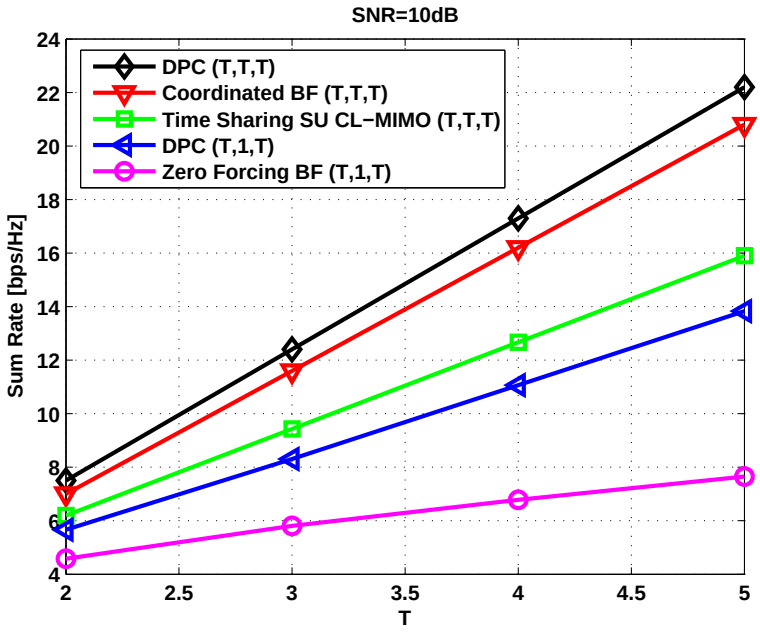

Fig. 2. Ergodic sum capacity and achievable sum rate as a function of the number of users, the number of transmit/receive antennas. (T_1, T_2, T_3) denotes the number of transmit antennas at the BS, the number of receive antennas at the user, and the number of active users in the network, respectively. Coordinated BF refers to the method presented in [19].

to maximize 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*.

A. Optimal scheduling for the MU-MIMO downlink

The theoretical capacity results in Section II 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 multiuser MIMO solutions.

It is known that linear beamforming can achieve the sum capacity when the number of active users in the system is large [12], [28], [30]. In [28], 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 $\mathbf{V}$ is not needed in this case, 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

$$R_{BD|\mathcal{A}_k}(\mathbf{H}_{\mathcal{A}_k}, P, \sigma^2) = \max_{\sum_{j\in\mathcal{A}_k} \mathrm{Tr}(\mathbf{Q}_j)\leq P} \sum_{j\in\mathcal{A}_k} \log \left| \mathbf{I} + \frac{\mathbf{H}_j \mathbf{W}_j \mathbf{Q}_j \mathbf{W}_j^H \mathbf{H}_j^H}{\sigma^2} \right| \tag{12}$$

where $\mathbf{Q}_j = \mathbb{E}(\mathbf{x}_j \mathbf{x}_j)^H$ is the input covariance matrix for the user j, $\mathbf{W}_j$ 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

$$R_{BD}(\mathbf{H}_{1,\cdots,U}, P, \sigma^2) = \max_{\mathcal{A}_k\in\mathcal{A}} R_{BD|\mathcal{A}_k}(\mathbf{H}_{\mathcal{A}_k}, P, \sigma^2). \tag{13}$$

Denote $\mathcal{S}$ as the maximum number of users to be supported. For the case of BD, $\mathcal{S} \leq N$. 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.

B. 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, among which [28], [30]–[32].

In the capacity-based greedy user selection algorithm, the transmitter chooses the single 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 among others optimal non-linear precoders. Scheduling for the non-linear precoders mentioned

in Section III-B is an ongoing topic of research, though a few results have appeared including a greedy user selection for zero-forcing dirty paper coding (ZFDPC) in [4], which has been proposed in [30].

Apart from the sum rate under optimal precoding, several other metrics may be considered for maximization at each step of the greedy scheduling algorithms, e.g. the multiuser MMSE, or for very low complexity the Frobenius norm of the multiuser channel [32]. The idea of the Frobenius norm-based user selection algorithm is to greedily select the set of users such that the sum of the effective channel energy of those selected users is as large as possible. In the case when the network has rate requirements, the selection of user groups aims at minimizing the power required to achieve the desired rate targets.

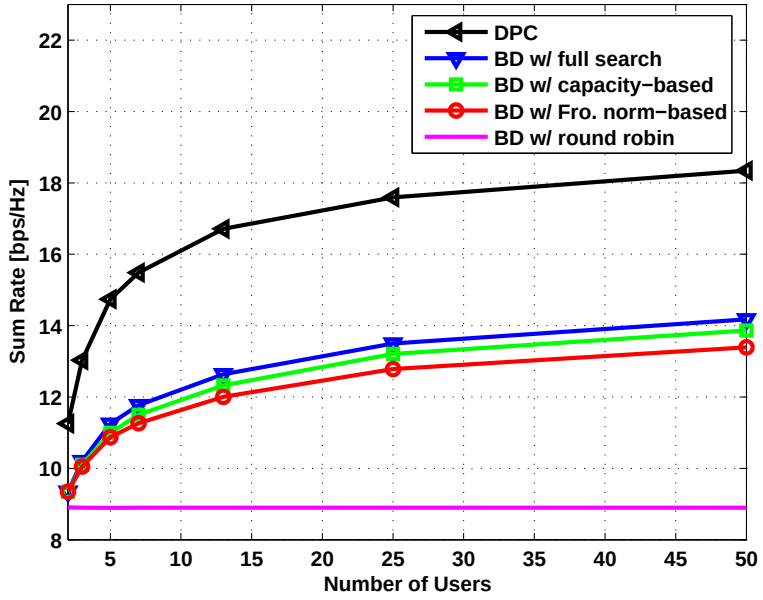

Fig. 3. Sum rate as a function of the number of users, where the number of transmit antennas is 4, the number of receive antennas is 2, and the maximum number of selected users is 2.

Fig. 3 shows the sum rate versus the number of users for $N = 4$ and $M = 2$. For simplicity, we only considered BD. The iterative based suboptimal solutions achieve about 95% compared to full search based method and show significant gain against round robin scheduler. Although DPC achieves higher sum capacity than BD approaches, BD still achieves significant part of the sum capacity requiring though only linear processing both at the BS and the user.

V. Living with partial channel knowledge at the transmitter

One of the fundamental paradigm shifts associated with multiuser MIMO is the push of intelligence and much of the signal processing complexity load from the terminals to the BS. This emphasis on the processing at the BS is accompanied, however, with the requirement that the BS be informed of the channel coefficients of all active users in the cell, prior to user group selection.

In frequency-division duplex (FDD) systems and time-division duplex (TDD) systems without calibration, often the only way for the BS to acquire channel state information from each user is through a feedback control channel. For example, control channels are used for power control or adaptive modulation. Since the bandwidth required by the feedback control channel is overhead that counts against the overall spectral efficiency of the system, and it grows in proportion to the number of active users, there is a substantial interest in compressing the CSI and using it in both the scheduling and the beam design algorithms.

Interestingly, although it was shown that the multiplexing gain disappears in the absence of any CSIT [33], recent findings suggest that the BS can live with limited channel information at the transmitter and still achieve a significant fraction of the capacity promised by the full CSIT case, although the issues of *optimally designed* limited feedback for MU-MIMO transmission techniques is still much open. MU-MIMO transmission design with limited CSIT has in fact evolved in a topic of research in its own right and many possible strategies can be pointed out. The reader is pointed to [34] for a complete state-of-the-art in this area. A few selected approaches are briefly exposed here.

One first key idea is based on splitting the feedback between the scheduling and the final beam design (or "user serving") stages, thus taking profit from the fact the numbers users to be served at each scheduling slot is much less than the number of cell users.

In [35] it is proposed to reduce feedback during the scheduling phase, which can be performed using rough channel estimates, while the stage of serving the scheduled group of K users is accomplished with near-perfect feedback as this concerns only a very small number of users compared with the number of cell users. Feedback reduction during the scheduling stage can be obtained via use of threshold-based pre-selection [36] combined with any of the approaches described below. Thresholding can be practically implemented using opportunistic feedback [37],

where users who exceed the threshold compete on a random access feedback channel. The optimal way of splitting the feedback load across the scheduling stage and the beam design stage is an interesting open problem, although design rules exploiting known rate scaling laws and bounds give promising results [38].

A. 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 single user MIMO communication systems using a concept known as limited feedback precoding [39]. 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.

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 [40], [41], 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 [40] 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 in order 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 [42]. 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 [43].

B. 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 an 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 k-th 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:

$$\varphi_k = \max_{1 \leq i \leq N} \frac{|\mathbf{h}_k \mathbf{q}_i|^2}{\sigma^2 + \sum_{j \neq i} |\mathbf{h}_k \mathbf{q}_j|^2} \tag{14}$$

where $\mathbf{q}_i$ denotes the i-th column of $\mathbf{Q}$. The scheduling algorithm then consists in opportunistically assigning to each beamformer $\mathbf{q}_i$ the user which 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 [44] and later generalized in [13]. The

performance of these methods is illustrated below. The idea of [13] 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 multi-beam 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 (5).

C. Living with sparse networks

A limitation of fixed or random opportunistic beamforming approaches is that the optimal capacity scaling emerges for large, sometimes impractical, number of simultaneously active users in the cell. The performance degrades with 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 [35] or simply to turn off certain beams [45], thus reducing inter-user interference when the random beams are not well aligned with users' channels. In Fig. 4, we compare the robustness of the single-beam ORBF [44] and multi-beam ORBF [13], both with SINR feedback, with respect to the number of active users in the cell. With four antennas at the BS, at 10dB SNR, simulations suggest that at least 12 simultaneously active users are required for the multi-beam 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 multi-beam ORBF [35] allows for a smooth transition between TDMA and SDMA regions, as shown in the figure.

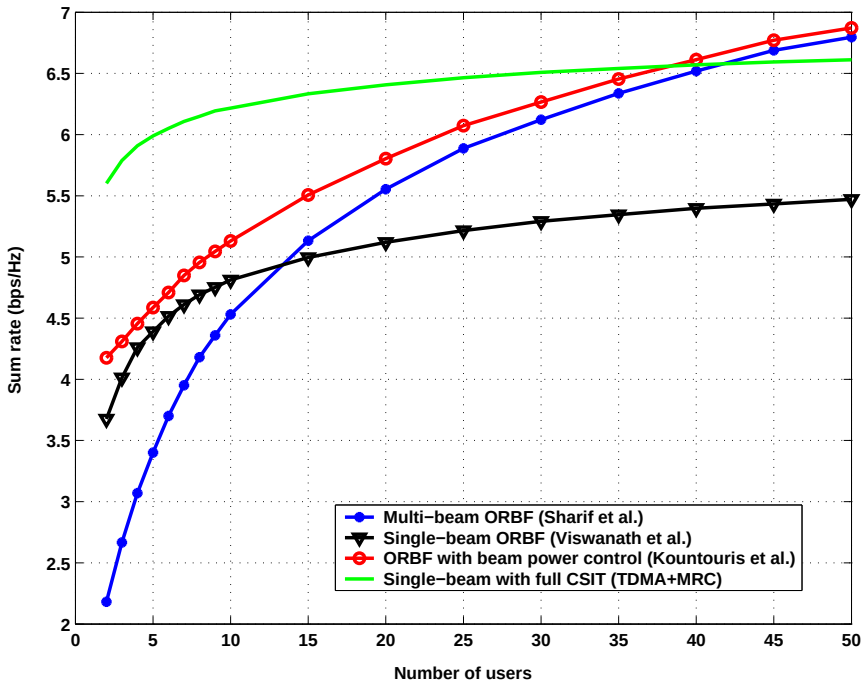

Fig. 4. The sum rate is compared for random beamforming schemes with SINR feedback. Multi-beam (SDMA) random beamforming outperforms the single beam (TDMA) when the number of active users is sufficient. Power control over the random beams allows for a smooth transition between TDMA and SDMA. TDMA with full CSIT outperforms partial feedback schemes for small number of users but fails to provide multiplexing gain when this number increases.

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 [46], [47]. In the spatial domain, statistics give information about spatial separability, which is instrumental to a proper beamforming design. Such aspect is described below.

D. 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) in order 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 k-th 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

$$\gamma_k = \|\mathbf{h}_k \mathbf{Q}_k\|^2 \tag{15}$$

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

$$\widehat{\mathbf{h}}_k = \mathbb{E}(\mathbf{h}_k|\gamma_k) \qquad\qquad \widehat{\mathbf{R}}_k = \mathbb{E}(\mathbf{h}_k^H \mathbf{h}_k|\gamma_k) \tag{16}$$

which can be used to provide an MMSE estimate of the instantaneous SINR [48]. Note that with $\mathbf{Q}_k = \mathbf{I}$, equation (15) falls back to a channel norm feedback.

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 (15) can be formulated [49]. Let $L = 1$ and $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{R}_k)$ and CQI feedback $\gamma_k = |\mathbf{h}_k \mathbf{q}_k|^2$. The solution to the ML problem

$$\begin{aligned} \max_{\mathbf{h}_k} \quad & \mathbf{h}_k \mathbf{R}_k \mathbf{h}_k^H \\ s.t. \quad & |\mathbf{h}_k \mathbf{q}_k|^2 = \gamma_k \end{aligned} \tag{17}$$

is given by

$$\widehat{\mathbf{h}}_k = \arg\max_{\mathbf{h}_k} \frac{\mathbf{h}_k \mathbf{R}_k \mathbf{h}_k^H}{\mathbf{h}_k (\mathbf{q}_k \mathbf{q}_k^H) \mathbf{h}_k^H} \tag{18}$$

which corresponds to 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 Fig.5.

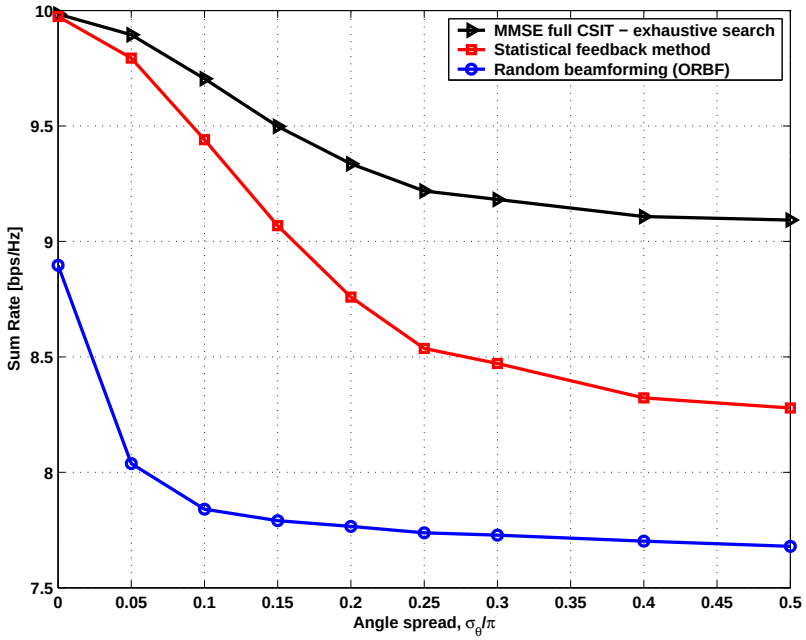

Fig. 5. Sum rate as a function of the angle spread σ_θ at the BS, where the number of transmit antennas is 2, the average SNR = 10dB, and the number of active users in the cell is 50.

Clearly the different approaches presented here for feedback reduction represent independent strategies, which could be combined together for more effectiveness. Certain techniques will be 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.

VI. 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 multiuser MIMO

solutions have yet to emerge. While spatial diversity and basic single user MIMO techniques are available in several products and standards, adaptive antenna solutions including MU-MIMO are mostly considered for TDD systems in low and moderate mobility where channel state information can be obtained from estimation in the uplink.

Note also that codebook based precoding schemes for SU- and MU-MIMO are emerging in existing and future standards [6]. MU-MIMO systems may have the potential to achieve the spectrum efficiency requirements set by operators for the next generation of mobile communication systems [50]. Practical MU-MIMO applications are still challenging however and further studies seem needed in order 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 others, on the number of other terminals being simultaneously scheduled along with the user making the measurement. Certain metrics (such as those in e.g [13], [42]) 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 multiuser MIMO.

Also, opportunistic scheduling in multiuser 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 non-negligible 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 [51].

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 [7], 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 degrees of freedom 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 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 degrees of freedom to satisfy the QoS constraints for different types of traffic?

VII. Discussion

MU-MIMO networks reveal the unique opportunities arising from a joint optimization of antenna combining techniques with resource allocation protocols (power control, user scheduling). MU-MIMO approaches are expected to provide significant multiplexing (on the order of the number of antennas used at the transmitter) and diversity gains while resolving some of the issues associated with conventional single user MIMO. Namely, 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 to 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.

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 Univ. Press, 2006.

[2] A. El Gamal and T.M. Cover, "Multiple user information theory," *Proc. IEEE*, vol. 68, no. 12, pp. 1466–1483, Dec. 1980.

[3] B. Hochwald, T. Marzetta, and V. Tarokh, "Multi-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Info. Th.*, vol. 50, no. 9, pp. 1893–1909, Sept. 2004.

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

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

[6] 3GPP, "Long Term Evolution, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; General description," *TS 36.201 v1.0.0*, March 2007.

[7] 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 Std 802.16e-2005*, Febr. 2006.

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

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

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

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

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

[13] ——, "On the capacity of MIMO broadcast channel with partial side information," *IEEE Trans. Info. Th.*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

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

[15] S. Verdú, *Multiuser Detection*. Cambridge, UK: Cambridge University Press, 1998.

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

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

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

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

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

[21] 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. Comm.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

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

[23] 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. Sig. Proc.*, vol. 55, no. 3, pp. 1159–1571, March 2007.

[24] 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. Comm.*, vol. 53, no. 3, pp. 537–544, March 2005.

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

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

[27] R. F. Fischer, *Precoding and Signal Shaping for Digital Transmission*, John Wiley and Sons, Inc., 2002.

[28] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Jour. Select. Areas in Comm.*, vol. 24, no. 3, pp. 528–541, March 2006.

[29] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," in *Proc. of Int. Zurich Sem. on Comm. (IZS'06)*, ETH Zurich, Switzerland, Febr. 2006.

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

[31] 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.

[32] 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. Sig. Proc.*, vol. 54, no. 9, pp. 3658–3663, Sept. 2006.

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

[34] R. Heath, V. Lau, D. Love, D. Gesbert, B. Rao and M. Andrews (Eds.), "Exploiting limited feedback in tomorrow's wireless communications networks," *Special issue of IEEE Journal on Selected Areas in Communications*, to appear 2008.

[35] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *Proc. IEEE Workshop on Sig. Proc. Adv. in Wir. Comm. (SPAWC) - Full version in preparation, available upon request*, New York, USA, June 2005, pp. 975–979.

[36] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?" in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Paris, France, June 2004, pp. 234–238.

[37] T. Tang and R. W. Heath Jr., "Opportunistic feedback for downlink multiuser diversity," *IEEE Communications Letters*, vol. 9, no. 10, pp. 948–950, 2005.

[38] R. Zakhour and D. Gesbert, "A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information," in *Proc. IEEE Int. Symp. on Pers., Ind. and Mob. Radio Comm. (PIMRC)*, Athens, Greece, Sept. 2007.

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

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

[41] P. Ding, D. Love, and M. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Sig. Proc.*, accepted for publication, June 2006.

[42] T. Yoo, N. Jindal, and A. Goldsmith, "Finite-rate feedback MIMO broadcast channels with a large number of users," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Seattle, WA, USA, July 2006.

[43] K. Huang, R. W. Heath Jr., and J. Andrews, "Space division multiple access with a sum feedback rate constraint," accepted for publication in the IEEE Trans. Sig. Proc.

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

[45] J. Wagner, Y.-C. Liang, and R. Zhang, “On the balance of multiuser diversity and spatial multiplexing gain in random beamforming,” *to appear in IEEE Trans. Wireless Comm.*, 2007.

[46] M. Kountouris and D. Gesbert, “Memory-based opportunistic multi-user beamforming,” in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Adelaide, Australia, Sept. 2005, pp. 1426–1430.

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

[48] D. Hammarwall, M. Bengtsson, and B. Ottersten, “Acquiring partial CSI for spatially selective transmission by instantaneous channel norm feedback,” *IEEE Trans. Sig. Proc.*, accepted for publication, Jan. 2007.

[49] M. Kountouris, D. Gesbert, and L. Pittman, “Transmit correlation-aided opportunistic beamforming and scheduling,” in *Proc. of Europ. Sig. Proc. Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[50] “NGMN: ‘Next Generation Mobile Networks Beyond HSPA and EVDO - A white paper’, V3.0,” *Available at http://www.ngmn-cooperation.com*, December 2006.

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