An Overview on Resource Allocation Techniques for Multi-User MIMO Systems

Eduardo Castañeda, *Member, IEEE,* Adão Silva, *Member, IEEE,* Atílio Gameiro, and Marios Kountouris, *Senior Member, IEEE*

***Abstract*—Remarkable research activities and major advances have been occurred over the past decade in multiuser multiple-input multiple-output (MU-MIMO) systems. Several transmission technologies and precoding techniques have been developed in order to exploit the spatial dimension so that simultaneous transmission of independent data streams reuse the same radio resources. The achievable performance of such techniques heavily depends on the channel characteristics of the selected users, the amount of channel knowledge, and how efficiently interference is mitigated. In systems where the total number of receivers is larger than the number of total transmit antennas, user selection becomes a key approach to benefit from multiuser diversity and achieve full multiplexing gain. The overall performance of MU-MIMO systems is a complex joint multi-objective optimization problem since many variables and parameters have to be optimized, including the number of users, the number of antennas, spatial signaling, rate and power allocation, and transmission technique. The objective of this literature survey is to provide a comprehensive overview of the various methodologies used to approach the aforementioned joint optimization task in the downlink of MU-MIMO communication systems.**

***Index Terms*—Downlink transmission, multi-user MIMO, precoding, resource allocation, spatial multiplexing, user scheduling.**

I. Introduction

FUTURE wireless systems require fundamental and crisp understanding of design principles and control mechanisms to efficiently manage network resources. Resource allocation policies lie at the heart of wireless communication networks, since they aim at guaranteeing the required Quality of Service (QoS) at the user level, while ensuring efficient and optimized operation at the network level to maximize operators' revenue. Resource allocation management in wireless communications may include a wide spectrum of network functionalities, such as scheduling, transmission rate control, power control, bandwidth reservation, call admission control, transmitter assignment, and handover [1], [2], [3]. In this survey, a resource allocation policy is defined by the following components: *i*) a multiple access technique and a scheduling component that distributes resources among users subject to individual QoS requirements; *ii*) a signaling strategy that allows simultaneous transmission of independent data streams to the scheduled users; and *iii*) rate allocation and power control that guarantee QoS and harness potential interference. Fig. 1 illustrates these components and the interconnection between them. The figure highlights the fact that each function of the resource allocation strategy can be performed in either optimal or suboptimal way, which is elaborated upon below.

The multiple access schemes can be classified as orthogonal or non-orthogonal. The former is a conventional scheme that assigns radio resources, e.g., code, sub-carrier, or time slot, to one user per transmission interval. The main characteristic of orthogonal multiple access schemes is their reliability, since there is no need to deal with co-resource interference. The resource allocation policy can optimize with reasonable complexity several performance metrics, such as throughput, fairness, and QoS [4]. The multiplexing gain, i.e., the number of scheduled users, is limited by the number of available radio resources in the system. In non-orthogonal multiple access, a set of users concurrently superimpose their transmissions over the same radio resource, and potentially interfere with each other. In this scheme the co-resource interference can be mitigated by signal processing and transmission techniques implemented at the transmitter and/or receiver sides. Such techniques exploit different resource domains, e.g., power, code, or spatial domain, and a combination of them are envisaged to cope with the high data rate demands and system efficiency expected in the next generation of wireless networks [5], [6].

Hereinafter, we focus on multiple access schemes based on multi-antenna transceivers operating at the spatial domain, i.e., multiple-input multiple-output (MIMO). MIMO communication, where a multi-antenna base station (BS) or access point (AP) transmits one or many data streams to one or multiple user equipments simultaneously, is a key technology to provide high throughput in broadband wireless communication systems. MIMO systems have evolved from a fundamental research concept to real-world deployment, and they have been integrated in state-of-the-art wireless network standards [7], [8], [9], e.g., IEEE 802.11n, 802.11ac WLAN, 802.16e (Mobile WiMAX), 802.16m (WiMAX), 802.20 (MBWA), 802.22 (WRAN), 3GPP long-term evolution (LTE) and LTE-Advanced (E-UTRA). Resource allocation is particularly challenging in wireless communication systems mainly due to the wireless medium variability and channel randomness, which renders the overall performance location-dependent and time-varying [10]. Nevertheless, high spectral efficiency and multiplexing gains can be attained in MIMO systems since multiple data streams can be conveyed to independent users.

E. Castañeda, A. Silva, and A. Gameiro are with the Department of Electronics, Telecommunications and Informatics, and the Portuguese Institute of Telecommunications (IT), Aveiro University, Aveiro, 3810-193, Portugal (e-mail: ecastaneda@av.it.pt; asilva@av.it.pt; amg@ua.pt).

M. Kountouris is with the Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies Co. Ltd. (e-mail: marios.kountouris@huawei.com).

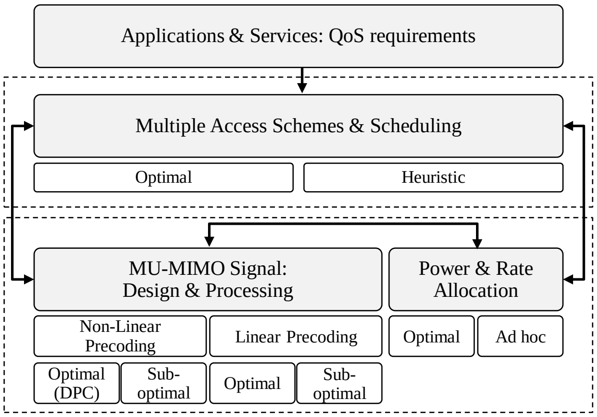


Fig. 1. Components of the resource allocation policy in MU-MIMO systems

By exploiting the spatial degrees of freedom (DoF) offered by multiple antennas we can avoid system resource wastage [11]. Multiuser (MU) MIMO systems have been extensively investigated over the last years from both theoretical and practical perspective. In a recent evolution of MU-MIMO technology, known as massive MIMO or large-scale MIMO [12], [13], few hundreds of antennas are employed at the BS to send simultaneously different data streams to tens of users. Massive MIMO has been identified as one of the promising air interface technologies to address the massive capacity requirement required demanded by 5G networks [14], [15].

The downlink transmission is particularly challenging in MU-MIMO scenarios because the geographic location of the receivers is random and joint detection cannot be performed. The main goal is to convey independent data streams to a set of properly selected users, attaining spatial multiplexing gain offered by MU-MIMO. However, determining such a users set is a very challenging task, which depends on all elements of the resource allocation strategy, e.g., individual QoS requirements, signaling schemes, rate allocation and power control strategies implemented at the transmitter. MIMO systems allow for a plethora of mighty signal processing techniques that enhance the system performance by exploiting a multi-dimensional pool of resources. This pool is composed of resources with different nature, e.g., signal spaces, transmission powers, time slots, sub-carriers, codes, and users. Efficient allocation designs over such a large set of resources implies that a tradeoff between optimality and feasibility. On the one hand, optimality can be reached by solving optimization problems over a set of integer and continuous variables, which may be a thoroughly complex task. Feasibility, on the other hand, implies that suboptimal resource allocation takes place by relaxing and reformulating optimization problems whose solutions can be found by practical and reliable algorithms.

A. Contributions of the Survey

There exists a very rich literature on MIMO communications, and this paper complements it by providing a classification of different aspects of MU-MIMO systems and resource allocation schemes. Users with independent channels provide a new sort of diversity to enhance the overall performance. However, in contrast to systems where each user accesses a dedicated (orthogonal) resource [4], [10], [16], [17], in MU-MIMO systems the additional diversity is realized when several users access the same resource simultaneously.

Accounting for multiple antennas at both ends of the radio link allows spatial steering of independent signals using precoding schemes, which results in the coexistence of many data streams conveyed to the concurrent users. Some of the main contributions of this survey are the description and classification of linear and non-linear precoding schemes, considering the amount of channel information available at the transmitter, the network scenario (e.g. single-cell or multi-cell), and the antenna settings. Each precoding scheme relies on different characteristics of the MU-MIMO channels to fully exploit the spatial domain. The paper provides a comprehensive classification of metrics that quantify the spatial compatibility, which can be used to select users and improve the precoding performance.

The spectral efficiency, error rates, fairness, and QoS are common criteria to assess performance in the MU-MIMO literature. Optimizing each one of these metrics requires specific problem and constrains formulations. The type of precoding, antenna configuration, and upper layer demands can be taken into account to design robust resource allocation algorithms, i.e. cross layer designs. Other contributions of this paper are the description and classification of different optimization criteria and general constraints used to characterize MU-MIMO system. The proposed classification incorporates the antenna configuration, the amount of channel information available at the transmitter, and upper layer requirements.

Early surveys on MU-MIMO have pointed out that resource allocation can be opportunistically enhanced by tracking the instantaneous channel fluctuations for scenarios with a single transmitter [11], [18]. However, in recent years, a large number of techniques have been developed for very diverse and heterogeneous MIMO scenarios. The paper presents a classification of state-of-the-art scheduling algorithms for MU-MIMO scenarios for single and multiple transmitter scenarios. We consider the channel state information, the objective functions optimized by the scheduler, the degree of cooperation/coordination between transmitters, and the power allocation techniques.

The goal of this survey is not to describe in detail the theory behind precoding design, rate allocation, power control or user scheduling, but rather to use their fundamental principles to get insight on the interplay among them. Our aim is to describe state-of-the-art processing techniques for MU-MIMO, point out practical challenges, and present general guidelines to design efficient resource allocation algorithms. The material favors broad intuition over detailed mathematical formulations, which are left to the references. Although the list of references is certainly not intended to be exhaustive, the cited works and the references therein may serve as a starting point for readers aiming to go beyond a tutorial.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

B. Organization

The paper is organized as follows. In Section II we present the basic ideas behind MIMO wireless communications, introduce MU-MIMO systems, and discuss the main challenges. In Section III we introduce the most commonly studied MU-MIMO channel and system models, their characteristics and conventional assumptions. Section IV is devoted to signal design and precoding schemes under different conditions of channel information. In Section V we introduce the most common metrics of spatial compatibility, which are used to categorize users and reduce the scheduling complexity. Section VI presents a classification of optimization criteria and describes the usual constraints considered in MU-MIMO systems. In Section VII we propose a classification of the several techniques to address the user scheduling problem. The specific characteristics, limitations and use cases for each technique are discussed. Section VIII is focused on scheduling algorithms with partial channel information at the transmitter. We categorize the existing approaches and present guidelines to minimize complexity and improve efficiency. In Section IX we present the most common power allocation schemes and discuss their role in MU-MIMO systems. Finally, we conclude the paper in Section X. The reader can find a list of technical terms and abbreviations summarized in Table VIII.

We adopt the following notation: matrices and vectors are set in upper and lower boldface, respectively. $(\cdot)^T$, $(\cdot)^H$, $|\cdot|$, $\|\cdot\|_p$ denote the transpose, the Hermitian transpose, the absolute value, and the p-norm, respectively. $rank(\mathbf{A})$, $null(\mathbf{A})$ denote the rank and null space of matrix $\mathbf{A}$. $Span(\mathbf{A})$ and $Span(\mathbf{A})^{\perp}$ denote the subspace and orthogonal subspace spanned by the columns of matrix $\mathbf{A}$. Calligraphic letters, e.g. $\mathcal{G}$, denote sets, and $|\mathcal{G}|$ denotes cardinality. $\mathbb{R}_+$ is the set of nonnegative real numbers and $\mathbb{C}^{N\times M}$ is the space of $N \times M$ matrices. $\mathcal{CN}(\mathbf{a}, \mathbf{A})$ is the complex Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{A}$. $\mathbb{E}[\cdot]$ denotes expectation.

II. Preliminaries

A. Multiple Antenna Systems

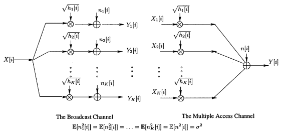
A MIMO system employs multiple antennas at the transmitter (M) and receiver (N) sides to improve communication performance. The seminal works [19], [20] provide a mathematical motivation behind multiple antenna processing and communications. Theoretical analysis has shown that the spectral efficiency, i.e., the amount of error-free bits per second per Hertz (bps/Hz), follows the scaling low $\min(M, N)$, without increasing the power or bandwidth requirements. The signal processing techniques in multi-antenna systems can be classified as *spatial diversity techniques* and *spatial multiplexing techniques* [21].

Spatial diversity techniques (see [21] and references therein), provide transmission reliability and minimize error rates. This is attained by transforming a fading wireless channel into an additive white Gaussian noise (AWGN)-like channel, i.e., one can mitigate signal degradation due to fading [11]. The probability that multiple statistically independent channels experience simultaneously deep fading gets very low as the number of independent paths increases. The spatial diversity techniques can be applied at both transmission and reception sides of the link. Transmit diversity schemes include space diversity, polarization diversity, time diversity, frequency diversity, and angle diversity. Examples of receive diversity schemes are selection combining, maximum ratio combining (MRC), and equal gain combining [22], [23].

B. Multiuser MIMO

This paper focuses on spatial multiplexing techniques, which exploit the DoF provided by MIMO. Spatial multiplexing is tightly related to multiuser communications and smart antennas processing [22]. In multiuser systems, spatial multiplexing gains can be attained by steering signals toward specific receivers, such that the power to intended users is boosted. Simultaneously, co-channel interference to unintended users can be partially or completely suppressed.

In MU-MIMO systems, the available resources (power, bandwidth, antennas, codes, or time slots) must be assigned among K active users. There are two kinds of multiuser channels: the downlink channel, also known as *broadcast channel* (BC), where a single transmitter sends different messages to many receivers; and the uplink channel, also called *multiple access channel* (MAC), where many transmitters communicate with a single receiver. There are several *explicit* differences between BC and MAC. In the former, the transmitted signal is a combination of the signals intended for all co-scheduled users, subject to total transmit power, P, c contrast, in the MAC channel, the signal from is affected by other co-scheduled users, subjec power constraints, i.e. P_k, [22]. There exis connection between BC and MAC, known as *duality*, which establishes the relationship between the capacity regions of both access channels [22], [23]. The BC-MAC duality has beer fundamental to define optimal policies for power allocation signaling, and QoS guaranteeing in MU-MIMO systems, see [24], [25], [26], [27]. The capacity regions include operative point where transmission to multiple users do not interfere with each other. Every transmission is performed over orthogonal signaling dimensions, which is a signal separation called duplexing [23]. This operation is performed by allocating communications across different time slots, known as time-division-duplex (TDD), or across separated frequency bands, known as frequency-division-duplex (FDD).

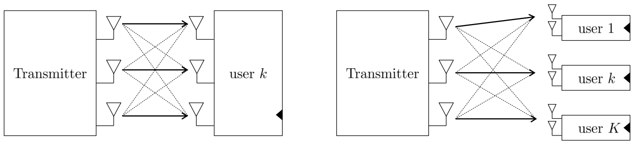


$$\text{BC}: \quad Y_j[i] = \sqrt{h_j[i]}X[i] + n_j[i]$$

$$\text{MAC}: \quad Y[i] = \sum_{j=1}^{K} \sqrt{h_j[i]}X_j[i] + n[i]$$

In the literature of MU-MIMO, two types of diversity are studied: *spatial multiplexing diversity* and *multiuser diversity* (MUDiv). The former is a consequence of the independent fading across MIMO links of different users. This means that independent data streams can be transmitted over parallel spatial channels, increasing the system capacity [28]. The latter arises when users that are geographically far apart have channels that fade independently at any point in time. Such independent fading processes can be exploited so that users with specific channel conditions are simultaneously scheduled [29]. There are two modes of transmission in MIMO systems, see Fig. 2: single user (SU) and multiuser (MU) mode. The SU-MIMO mode improves the performance of a single user, allocating one or many data streams in the same radio resource.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials



(a) SU-MIMO

(b) MU-MIMO

Fig. 2. (a) Single-user (SU) and (b) Multi-user (MU) MIMO modes

In the MU-MIMO mode, different data streams are sent to different users such that a performance metric is optimized, e.g., the average sum rate.

Selecting between SU- or MU-MIMO transmission modes depends on the accuracy of the channel state information at the transmitter (CSIT), the amount of allowed interference, the target rate per user, the number of user, the signal-to-noise ratio (SNR) regime, and the achievable capacity in each mode [30]. Nonetheless, by assuming sufficient CSIT knowledge, MU-MIMO processing techniques provide several performance gains [18]: multiple antennas attain diversity gain, which improves bit error rates (BER); directivity gains realized by MUDiv, since the spatial signatures of the users are uncorrelated, which mitigates inter-user interference (IUI); immunity to propagation limitations in SU-MIMO, such as rank loss or antenna correlation; and multiplexing gains that scale, at most, with the minimum number of deployed antennas.

C. The need of User Scheduling

In MU-MIMO BC systems, the overall performance depends on how efficiently the resource allocation algorithms manage the hyper-dimensional pool of resources (carriers, time slots, codes, power, antennas, users, etc.). Consider a system with a transmitter equipped with M antennas, and let $\bar{\mathcal{K}} = \{1, 2, 3, \ldots, K\}$ be the set of all active users, illustrated in Fig. 3. To qualitatively determine the objectives of scheduling, we provide the following definitions:

Definition 1. *Quality of Service*. We say that QoS defines a set of prescribed network-/user-based performance targets (e.g., peak rates, error rates, average delays, or queue stability), that can be measured, improved, and guaranteed for a specific upper layer application.

Definition 2. *User scheduling*. We say that a set of radio resources (e.g. time slot, codes, sub-channels, powers, etc.), has been assigned to a group of scheduled users, $\mathcal{K} \subseteq \bar{\mathcal{K}}$, so that a global performance metric is optimized subject to power and QoS constraints. Moreover, each user $k \in \mathcal{K}$, achieves non-zero rate with successful information reception.

Consider that each user $k \in \bar{\mathcal{K}}$, is equipped with N_k antennas. By having more receive than transmit antennas ($M < \sum_k N_k$), one can solve a selection problem to achieve MUDiv gains in fading fluctuating channels [11], [18], [31], [32], [33]. A fundamental task in resource management is to select a subset of users $\mathcal{K} \subseteq \bar{\mathcal{K}}$, and assign resource to it, so that a given performance metric is optimized. For the sake of illustration, consider a single-transmitter scenario, and let us formulate a general user scheduling problem for a single resource (sub-carrier, time slot, or code) as follows:

$$\text{maximize} \quad \sum_{k=1}^{K} \xi_{\pi(k)} U_{\pi(k)} \tag{1}$$

$$\text{subject to} \quad \sum_{k=1}^{K} \xi_k p_k \leq P \tag{1a}$$

$$0 \leq \xi_k p_k \leq P_k \quad \forall k \in \bar{\mathcal{K}} \tag{1b}$$

$$\sum_{k=1}^{K} \xi_k \leq c_p \tag{1c}$$

$$\xi_k \in \{0, 1\} \quad \forall k \in \bar{\mathcal{K}} \tag{1d}$$

Our goal is to maximize the sum of the utility functions, $U_{\pi(k)}, \forall k$, which depends on several parameters: the multiuser MIMO channel, $\mathbf{H}_k$, the allocated power, p_k, the individual data queues, q_k, and the encoding order, $\pi(\cdot)$. The QoS requirements can be included in the definition of $U_{\pi(k)}$, as individual weights [cf. Section VI]. Equations (1a) and (1b) define total and individual power constraints, which are set according to the scenario [cf. Section III]. The term $\xi_{\pi(k)}$ is a binary variable with value equal to 1 if the $\pi(k)$-th user is scheduled and 0 otherwise. The set of selected user is given by $\mathcal{K} = \{k \in \bar{\mathcal{K}} : \xi_k = 1\}$. The system operates in MU-MIMO mode if $1 < |\mathcal{K}| \leq c_p$, where c_p in (1c) denotes the maximum number of users or transmitted data streams, that can be sent over M antennas. If the solution of (1) only exists for $|\mathcal{K}| = 1$, the system operates in SU-MIMO mode. In such a case, the optimization problem can be formulated to attain MUDiv, multiplexing (high rate), and diversity (high reliability) gains, see [22], [34], [35]. Depending on the CSIT, the type of signaling design and coding applied to the data, theoretical analysis show that the number of users with optimal nonzero power is upper bounded[1] as $|\mathcal{K}| \leq c_p \leq M^2$, [26]. In practical systems, multiplexing gain can be scaled up to $|\mathcal{K}| \leq M$, by means of linear signal processing [cf. Section IV-B].

The mathematical formulation in (1) resembles a knapsack or subset-sum problem [37], [38], which is known to be non-polynomial time complete (NP-C). Although the users are fixed items that must be chosen to construct $\mathcal{K}$, their associated utility functions change according to the channel conditions and the resource allocation of the co-selected users. This implies that the optimization variables are, in general, globally coupled. Finding the optimal set $\mathcal{K}$, is a combinatorial problem due to the binary variables ξ_k, and the encoding order $\pi(\cdot)$. Moreover, depending on U_k, problem (1) might deal with non-convex functions on the multiple parameters, e.g. K, M, N_k, etc. The feasibility of (1) relies on the constraints and processing, e.g. the precoding schemes, the power allocation, the CSIT accuracy, [18], [cf. Section VI]. The scheduling problem can be solved optimally by exhaustively searching (ExS)

[1] The upper bound is tight for small values of M and it becomes loose as the number of transmit antennas grow large. Numerical results comparing the upper bound of $|\mathcal{K}|$ for several coding techniques can be found in [36].

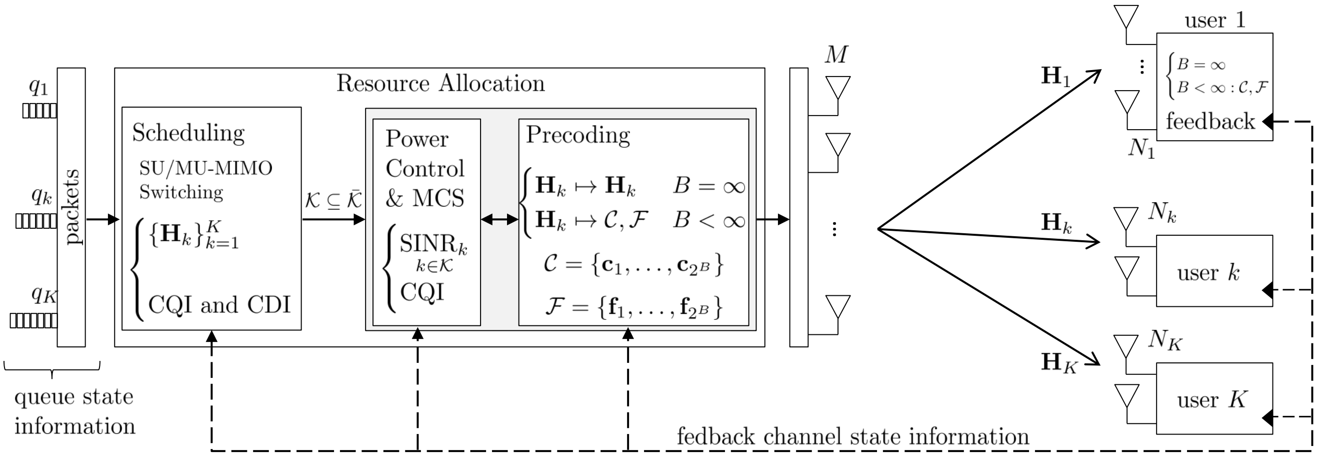

q_1
q_k
q_K
packets
queue state information
Resource Allocation
Scheduling
SU/MU-MIMO Switching
$\{\mathbf{H}_k\}_{k=1}^{K}$
CQI and CDI
$\mathcal{K} \subseteq \bar{\mathcal{K}}$
Power Control & MCS
SINR_k $k \in \mathcal{K}$
CQI
Precoding
$\mathbf{H}_k \mapsto \mathbf{H}_k \quad B = \infty$
$\mathbf{H}_k \mapsto \mathcal{C}, \mathcal{F} \quad B < \infty$
$\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{2^B}\}$
$\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_{2^B}\}$
M
$\mathbf{H}_1$
$\mathbf{H}_k$
$\mathbf{H}_K$
user 1
$B = \infty$
$B < \infty : \mathcal{C}, \mathcal{F}$
feedback
N_1
N_k
user k
N_K
user K
fedback channel state information

Fig. 3. Processing blocks and input signals in a MU-MIMO scenario. A transmitter with M antennas serves K multiple antenna receivers.

over all possible set sizes and user permutations. However, the computational complexity of ExS is prohibitively high, even for small values of K [11]. Furthermore, problem (1) can be modified to include additional dimensions, such as multiple carriers for OFDMA systems (e.g. [39], [40]) or codes for CDMA systems (e.g. [41]).

D. The need of CSI availability

Channel knowledge at the transmitter can be modeled taking into account instantaneous or statistical information, e.g. variance, covariance, angles of arrival/departure, and dominant path in line-of-sight [42]. There are two main strategies used to obtain CSI, reciprocity and feedback, which provide different feedback requirements and robustness to CSI errors. The former, known as open-loop feedback, uses uplink channel information to define the downlink channel in the next transmission interval. It is suitable for TDD since transmit directions are in identical frequencies, and the channel can be reversed. The latter, known as closed-loop feedback, requires sending the downlink channel to the transmitter using dedicated pilots, and is commonly used in FDD. The majority of the papers reviewed in Section VII assume closed-loop with full or limited channel feedback, and we refer the reader to [42], [43], [44] and references therein for further discussion on CSI acquisition and its impact on system performance.

To perform multiple antenna processing, interference mitigation, user scheduling, efficient power allocation, and to profit from MUDiv, knowledge of CSIT is compulsory. The complete lack of CSIT reduces the multiplexing gain to one, and cannot use MUDiv for boosting the achievable capacity [31], [45], [46]. In such scenarios, the optimum resource allocation and transmission schemes are performed over orthogonal dimensions [47]. In the literature of MU-MIMO, a large number of works assume full CSI (error-free), at both the receiver (CSIR) and transmitter (CSIT) sides. In practical systems, a strong downlink pilot channel, provided by the transmitter, is available to the users, hence the CSIR estimation error is negligible relative to that of the CSIT [48]. For simplicity, it is widely assumed that CSIR is perfectly known at the mobile terminals. In cellular MU-MIMO systems, channel estimation relies on having orthogonal pilots allocated to different users. The orthogonality is guaranteed for the users within the same cell, but not for those scattered across different cells. The number of BS antennas and bandwidth constraints may not allow orthogonal pilots for each user in the system, resulting in pilot contamination [49]. Under universal frequency reuse, the pilots can be drastically polluted by users at adjacent cells, when the serving BS performs channel estimation [50].

Achieving full CSIT (ideal noiseless and delay-free feedback) is highly challenging in practice. Feeding back the CSI requires rates that grow rapidly with the transmit power and the number of antennas [51]. However, by assuming full CSIT, it is possible to derive upper bounds on the performance of different signal processing techniques and scheduling algorithms. The information-theoretic and numerical results using full CSIT provide useful insights regarding the system performance bounds (e.g., [32], [52], [53]). Resource allocation strategies that optimize spectral efficiency, fairness, power consumption, and error probability can be designed to characterize optimal operating points [25], [27], [54]. Analytical results for full CSIT reveal the role of each parameter in the system, e.g., number of deployed transmit and receive antennas, number of active users, SNR regime, etc.

If channel knowledge is obtained via partial (rate-limited) feedback, the information available at the transmitter has finite resolution, resulting in quantization errors. *Partial CSIT* is comprised of two quantities: channel quality information (CQI) and channel direction information (CDI) [18]. The CQI measures the achievable SINR, the channel magnitude, or any other function of the link quality. The CDI is the quantized version of the original channel direction, which is determined using codebooks [cf. Section IV-C]. The transmitter uses both indicators for scheduling [cf. Section VIII], and the CQI is particularly used for power control, link adaptation, and interference management [55]. The CSI feedback interval

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

TABLE I
SUMMARY OF THE TYPE OF CSI AT THE TRANSMITTER FOR MISO ($N = 1$) AND MIMO ($N > 1$) CONFIGURATIONS

	MISO	MIMO
Full CSIT	[33], [36], [41], [47], [59], [60], [61], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96]	[39], [40], [97], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127]
Statistical	[66], [96], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139]	[58], [114], [117], [140], [141], [142]
Outdated	[30], [41], [48], [56], [57], [81], [96], [128], [143]	[117]
Correlated	[58], [96], [128], [134], [144], [145], [146]	[113], [117], [118], [122], [140], [144], [145], [147], [148], [149], [150], [151]
Partial CSIT	[30], [33], [51], [55], [57], [76], [133], [152], [153], [154], [155], [156], [157], [158], [159], [160], [161], [162], [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], [174]	[104], [140], [142], [147], [175], [176], [177], [178], [179], [180], [181], [182], [183], [184]

highly depends on the users' mobility,[2] and even for short-range communications (e.g., WiFi), immediate feedback is needed to achieve and maintain good performance [8]. By considering high mobility and limited feedback rates, one cannot rely on instantaneous or full CSI. In such cases, the transmitters perform resource allocation based on *statistical* CSI, which vary over larger time scales than the instantaneous CSI. The statistics for the downlink and uplink are reciprocal in both FDD and TDD, which can be used to perform resource allocation, see [58] and references therein. Table I summarizes the different types of CSIT in MU-MIMO systems, and the antenna configuration at the receivers. Partial CSIT refers to quantized channel information, which will be elaborated upon Sections IV-C and VIII.

III. MU-MIMO CHANNEL AND SYSTEM MODELS

The signal processing and scheduling algorithms described in the following sections have been developed and studied for single-hop MU-MIMO scenarios. We have classified the scenarios in two groups, see Fig. 4: single transmitter and multiple transmitters scenarios.

The implemented resource allocation strategies, user scheduling, and signal processing techniques depend on the number of coordinated transmitters, the number of antennas (M and N), the number of users (K), the SNR regime, and the CSIT accuracy. The system optimization relies on close-loop (e.g. [21], [42]) or open-loop (e.g. [82]) feedback, to achieve spatial multiplexing gains, multiuser diversity gains, and to combat interference. In cellular systems, there are two main sources of interference [185]: other active devices in the same co-channel and same cell, i.e., intra-cell or IUI; and from transmissions in other cells, i.e., inter-cell interference (ICI). The techniques to mitigate IUI and ICI depend on the type of scenario and the optimization criterion. There exist a number of scenarios where the interference cannot be reduced, see [51], [186], whose characteristics are described in the following definition:

Definition 3. *Interference-limited system.* An MU-MIMO system is said to be interference limited if the performance metric saturates (ceiling effect) with the transmit SNR. This might occur due to CSIT inaccuracy, highly correlated multiuser channels (IUI), and irreducible ICI.

A. Scenarios with a single transmitter

The objective of MU-MIMO processing is to accommodate many users per resource. Therefore, resource allocation strategies are commonly analyzed at the basic resource unit, e.g., code, single-carrier, time-slot, or frequency-time resource block. This can be done regardless the global system model (single-carrier, OFDM, or CDMA), since the same resource allocation strategy is applied over all resources, e.g., [41], [47], [60], [85], [114], [126], [142], [155], [157], [187], [188]. We adopt a signal model using the most general approach in the reviewed references. Consider a scenario where the transmitter is equipped with M antennas, and K active users are equipped with N antennas. Let $\mathbf{H}_k \in \mathbb{C}^{N\times M}$, be the discrete-time complex baseband MIMO channel of the k-th user for a given carrier. The received signal can be expressed as:

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{n}_k \tag{2}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$, is the joint transmitted signal for all users. The MIMO channel is usually assumed to be ergodic, i.e., it evolves over time and frequency in an independent and identically distributed (i.i.d.) manner. The channel is commonly modeled as *Rayleigh fading*, which is suitable for non line-of-sight communications. The complete spatial statistics can be described by the second-order moments of

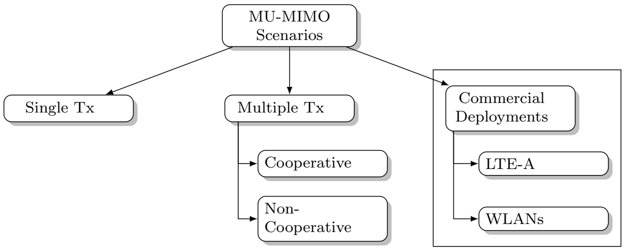


Fig. 4. Classification of MU-MIMO Scenarios and two examples of commercial technologies

[2] The authors in [56] showed that mobility defines the best reliable transmission strategy for capacity maximization, i.e., space-time coding or space division multiplexing. The results in [57] defined acceptable mobility ranges for MU-MIMO scenarios.

the channel [146]. Define the channel covariance matrix as $\mathbf{\Sigma}_k = \mathbb{E}[\mathbf{H}_k^H \mathbf{H}_k]$, which depends on the antenna configuration, propagation environment, scattering conditions and mobility. The channel can be decomposed as $\mathbf{H}_k = \sqrt{\mathbf{\Sigma}_k}\mathbf{H}_{iid}$, where $\mathbf{H}_{iid}$ has i.i.d. entries with distribution $\mathcal{CN}(0, 1)$.

Assuming spatially uncorrelated Rayleigh fading channels, i.e., $\mathbf{\Sigma}_k = \gamma\mathbf{I}$, $\forall k$ and some $\gamma > 0$, is the most common practice in the literature [189]. Physically, this implies rich scattering environments with sufficient antenna spacing at both ends of the radio link [19]. Under these conditions, the fading paths between the multi-antenna transmitter and receiver become independent. The MU-MIMO channels are modeled as narrow band, experiencing frequency-flat (constant) fading, where there is no inter-symbol interference [21]. A common simplification used in OFDM broadband systems, is to assume multiple flat-fading sub-channels [10], [187]. The received signal per sub-channel can be modeled as (2), avoiding frequency selectivity [190]. More realistic channel models for broadband MIMO systems and their performance analysis were proposed in [191]. The Rayleigh model can be thought as a particular case of the asymmetric *Ricean fading* channel model. In this case, the entries of $\mathbf{H}_{iid}$ have non-zero mean and there exists a dominant line-of-sight (LoS) component that increases the average SNR [23]. Such a model is less common in MU-MIMO systems operating at microwave (i.e., sub-6 GHz bands), since the channels become more static and the benefits of MUDiv vanish with the magnitude of the LoS paths [136]. In recent years, wireless communication over the millimeter wave (mmWave) frequency range (30-300 GHz) has proved to be a feasible and reliable technology with a central role to play in 5G [192], [193], [194]. The mmWave technology rely on directional antennas to overcome propagation loss, penetration loss, and rain fading. High directivity implies that Ricean fading channels can be used to characterize both LoS and non-LoS components present in the mmWave channels [194], [195], [196], [197].

Several authors model the MIMO channel such that the correlation at transmit and receive antennas is distinguishable, and $\mathbf{\Sigma}_k \neq \gamma\mathbf{I}$, $\forall k$, see references in Table I. There are two approaches to model and analyze performance under MIMO correlation, the *jointly correlated model* [58], [145] and the simplified *Kronecker model* [198], [199]. The former assumes separability between transmit and receive eigen-directions, and characterizes their mutual dependence. The latter assumes complete correlation separability between the transmitter and receiver arrays, see [149], [199], [200], [201] and references therein.[3] We refer to [191] for a comprehensive analysis of Rayleigh and Rician correlated MU-MIMO channels. Note that in some MIMO propagation scenarios with uncorrelated antennas, the MIMO capacity can be low as compared to the SISO one due to the *keyhole* or *pinhole* effect. This is related to environments where rich scattering around the transmitter and receiver leads to low correlation of the signals, while other propagation effects, like diffraction or waveguiding, lead to a rank reduction of the transfer function matrix [203], [204].

The MIMO channels $\mathbf{H}_k$ $\forall k$, may also include large-scale fading effects due to shadowing and path loss [22], [23]. Depending on the type of access technique (OFDMA, CDMA, or TDMA), the channel model can take into account multipath components, correlation, Doppler spread, and angular properties [47]. Another common assumption to avoid frequency dependency (particularly in low mobility scenarios), is to account for *block-fading channels* [189]. This means that the CSI is constant (within a coherence time duration) for a block of consecutive channel uses before changing independently for the next block.

The noise $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$, is usually modeled by i.i.d. normalized entries according to a circular normal (complex) Gaussian distribution with zero mean and unit variance [43]. The transmitted signal, $\mathbf{x}$, can be defined according to the encoding applied over the user data, the number of spatial streams per user, and the power allocation. If linear precoding is used [cf. Section IV], the transmitted signal is defined as

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{W}_k \mathbf{d}_k \tag{3}$$

where $\mathbf{W}_k \in \mathbb{C}^{M\times d_k}$, is the precoding matrix, $\mathbf{d}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_k})$ is the data signal, and d_k is the number of multiplexed data streams of user k. In single-transmitter MU-MIMO scenarios with signal models defined by (2) and (3), the ICI is negligible or assumed to be part of the additive background noise. Therefore, IUI is the main performance limiting factor, which can be addressed by precoding the user data, [cf. Section IV].

The active users might experience similar average long-term channel gains (large-scale fading or path loss) and SNR regimes. For some practical cellular systems, assuming homogeneous users is valid if open-loop power control is used to compensate for cell-interior and cell-edge path losses. Therefore, the resultant effective multiuser channels have quasi-identical variances [55], [82]. The user distribution affects the fading statistics of $\mathbf{H}_k$ $\forall k$, and the general design of resource allocation algorithms [198].

A possible single-transmitter MU-MIMO scenario arises in satellite communications. However, due to the characteristics of the satellite channels, marginal MIMO gains can be realized. The absence of scatterers in the satellite vicinity yields a Rician-type channel with a strong line-of-sight component, turning off the capabilities of MIMO processing. Due to the large coverage area in satellite communications, the users have heterogeneous long-term channel gains, which directly affects the resource allocation decisions. Regardless of the limited literature about MU-MIMO satellite communications, recent works show promising results and discussions about how intensive frequency reuse, user scheduling, and multibeam signal processing will be implemented in next generation broadband satellite systems [79], [154], [205].

[3] Experimental results in [202] and theoretical analysis in [201] have shown that the conventional Kronecker model may not be well suited for MU-MIMO scenarios resulting in misleading estimates for the capacity of realistic scattering environments. This occurs due to the sparsity of correlated channel matrices, and the fact that the parameters of the Kronecker model change with time and position. The model can be used in scenarios with particular conditions on the local scattering at the transmitter and receiver sides [198].

B. Scenarios with multiple transmitters

In this scenarios, the channel models and assumptions aforementioned are applied. Yet, some additional considerations are made so that signaling and connections between network entities can be modeled. Deploying several transmitters across a geographic area can provide reliable communication for heterogeneous mobile terminals (different path losses or SNR regimes relative to each transmitter). This kind of infrastructure based systems include cellular and wireless local area networks (WLAN). The resource allocation and access control can be performed based on CSI and knowledge of the interference structure [206]. If the transmitters are allowed to cooperate, e.g., through a central processing unit (CU), IUI can be mitigated using CSIT for signal design [cf. Sections II-D and IV]. Global knowledge or estimation of the interference can be used to avoid poor spectral efficiency or inaccurate assignment of radio resources.

This paper focuses on scenarios where roaming (mobility) and reuse of resources are central management tasks. The premise behind cellular communications, is to exploit the power falloff with distance of signal propagation to reuse the same channel at spatially-separated locations. This means that the serving area is divided in non overlapping cells. Any cell site within a neighborhood cannot use the same frequency channel, which makes the same reused frequency channels sufficiently far apart [207], [208]. In traditional cellular systems, a given user belongs to only one cell at a time and resource allocation is performed unilaterally by its serving BS (non-cooperative approach in Fig. 4). Each transmitter serves its own set of users, transmission parameters are adjusted in a selfish manner by measuring ICI (simple interference-awareness), and there is no information exchange between BSs [208]. If frequency reuse is employed, the BS can make autonomous resource allocation decisions and be sure that no uncoordinated ICI appears within the cell [54]. However, in many practical systems, universal frequency reuse is applied, which means that neighboring cells can access the same frequencies and time-slots simultaneously. This might increase the ICI and potentially degrade performance [209]. The mitigation of ICI is a fundamental problem since the transmit strategy chosen by one BS will affect the reception quality of the users served by adjacent BSs.

A cluster of BSs can coordinate the resource allocation, scheduling decisions, and ICI mitigation techniques (cooperative approach in Fig. 4). Dynamic clustering is an ongoing research topic (see [17], [54], [207], [208] and references therein), which promises to meet the requirements established in the third generation partnership project (3GPP) standards [209]. Different forms of ICI control have been proposed over the last years. Extensions of space multiple access techniques for multi-cell systems have received several names, *coordinated multi-point* (CoMP) [209], [210], [211], *network MIMO* [107], or *joint signal transmission/processing* (JT) [208]. These techniques exploit the spatial dimensions, serving multiple users (specially cell edge users [54]), while mitigating ICI of clustered BSs.[4] For these approaches, a cluster can be treated as a super cell, for which mathematical models from the single-cell scenario can be applied straightforwardly, e.g. [171], [210], [212].

If user data is shared among BSs, the use of *proactive* interference mitigation within a cluster can take place. This implies that coordinated BSs do not separately design their physical (PHY) and media access control layer parameters. Instead, the BSs coordinate their coding and decoding, exploiting knowledge of global data and CSI [207]. However, to guarantee large performance gains for these systems, several conditions must be met [54], [208]: global CSI and data sharing availability, which scales up requirements for channel estimation, backhaul capacity, and cooperation; coherent joint transmission and accurate synchronization; and centralized resource allocation algorithms, which may be infeasible in terms of computation load and scalability.

There is another approach of multi-cell cooperation, coined as *coordinated scheduling* (CS) with *coordinated beamforming* (CBF), which is a form of coordinated transmission for interference mitigation [208], [209]. CS/CBF refers to the partial or total sharing of CSI between BSs to estimate spatial signaling, power allocation, and scheduling without sharing data or performing signal-level synchronization [207]. CBF implies that each BS has a disjoint set of users to serve, but selects transmit strategies jointly with all other BSs to reduce ICI. In this approach, exchange of user data is not necessary, but control information and CSI can be exchanged to simultaneously transmit to a particular set of users [211]. CBF is more suitable than JT for practical implementations, since it requires less information exchange. Nevertheless, CSI acquisition, control signaling, and coordinated scheduling are challenging tasks due to limited feedback bandwidth and finite capacity backhaul [17].

C. Commercial Deployments

This paper covers two wireless technologies, whose specifications already support MU-MIMO communications: LTE-Advanced for cellular networks and IEEE 802.11ac for wireless local area networks (WLAN).

- *LTE-Advanced*: This is the 3GPP cellular system standard for 4G and beyond communications [209]. Several capabilities have been added to the LTE standard to increase capacity demands and integrated a large number of features in the access network. Among these attributes, the ones related to MIMO processing are the most relevant in the context of this paper: enhanced downlink MIMO, multi-point and coordinated transmission schemes, and multi-antenna enhancements. Due to the fact that LTE is a cellular technology, most of the studied deployments in the literature lie in the category of multiple transmitters scenarios, see Fig. 4.

[4] Theoretical analysis in [186] shows that in the high SNR regime, the achievable system capacity is fundamentally interference-limited due to the out-of-cluster interference. This occurs regardless the level of coordination and cooperation between clustered BSs. However, coordinated scheduling and user clustering can provide means to improve spectral efficiency and mitigate interference at high SNR.

MU-MIMO communication has been incorporated in LTE with the following maximum values: four users in MU-MIMO configuration, two layers (spatial streams) per user, four simultaneous layers, and robust CSI tracking. Practical antenna deployments at the transmitter use dual-polarized arrays, and the expected number of co-scheduled users for such configurations will be two for most cases [142], [213]. LTE provides a mechanism to improve performance by switching between MU or SU-MIMO mode on a per sub-frame basis, based on CSI, traffic type, and data loads. The goal of dynamic mode switching is to balance the spectral efficiency of cell edge and average cell users. This can be achieved, for instance, by using transmit diversity for users at the cell edge, or implementing spatial multiplexing for cell center users [214].

• *IEEE 802.11ac*: This WLAN standard supports multiuser downlink transmission, and the number of simultaneous data streams is limited by the number of antennas at the transmitter. In MU-MIMO mode, it is possible to simultaneously transmit up to 8 independent data streams and up to 4 users [8], [9]. Spatial multiplexing is achieved using different modulation and coding schemes per stream. MU-MIMO transmission prevents the user equipment with less antennas to limit the achievable capacity of other multiple antenna users, which generates rate gains for all receivers. A unique *compressed explicit feedback* protocol, based on channel sounding sequences, guarantees interoperability and is used to estimate CSI and define the steering matrices (beamforming). Other methods for channel estimation are described in [215] and references therein. Compared to cellular networks, WLANs usually have fewer users moving at lower speeds, the APs are less powerful that the BSs and the network topology is ad hoc. Although multiple APs could be deployed, most of the works in the literature focus on MU-MIMO systems with a single transmitter. Nonetheless, joint transmission from several APs to different mobile users is feasible in WLANs, which requires coordinated power control, distributed CSI tracking, as well as synchronization in time, frequency, and phase. The authors in [216] have shown that distributed MIMO can be achieved by enhancing the physical layer for coordinated transmission, and by implementing time-critical functions for the media access control layer. It is likely that in the next generation of 802.11 standard, coordination schemes between APs will be adopted to enable MU-MIMO communications [215].

IV. MU-MIMO PRECODING

The spatial dimension provided by the multi-antenna transceivers can be used to create independent channelization schemes. In this way, the transmitter serves different users simultaneously over the same time slot and frequency band, which is known as space-division multiple access (SDMA) [18], [21], [22], [217]. The spatial steering of independent signals consists of manipulating their amplitude and phases (the concept of *beamforming* in classic array signal processing), in order to add them up constructively in desired directions and destructively in the undesired ones [21], [54]. By jointly encoding all (co-resource) signals using channel information, it is possible to increase the signal-to-interference-plus-noise ratio (SINR) at the intended receiver and mitigate interference for non-intended receivers.

In the literature of MU-MIMO systems, the term beamforming refers to the signal steering by means of beams to achieve SDMA. The term *precoding* is used to denote the scaling and rotation of the set of beams, so that, their power and spatial properties are modified according to a specific goal. Hereinafter, we use the term precoding[5] to describe the signal processing (i.e., beam vector/matrix computation, scaling, rotation, and projection), applied to the independent signals prior to transmission. In this section we describe the most common precoding techniques used in MU-MIMO scenarios, and their characteristics according to CSIT. Table II summarizes the precoding schemes used in the surveyed literature, as well as their associated methods for user selection, which will be elaborated upon in the following sections. An important performance metric determined by the precoders $\mathbf{W}_k$, $\forall k$, related to the delivered energy through the MU-MIMO channels, is provided in the following definition.

Definition 4. *Effective channel gain*. It is the magnitude of the channel projected onto its associated precoding weight. Let $\mathbf{H}_k^{(eff)} = \mathbf{H}_k \mathbf{W}_k$, be the *effective channel* after spatial steering, thus, the effective channel gain is given by: *i*) $|\mathbf{H}_k^{(eff)}|^2$ for MISO scenarios; *ii*) for MIMO scenarios, it is given by a function of the eigenvalues of $\mathbf{H}_k^{(eff)}$: $\det\left(\mathbf{H}_k^{(eff)}(\mathbf{H}_k^{(eff)})^H\right)$ or $\|\mathbf{H}_k^{(eff)}\|_F^2$.

A. Non-linear Precoding with full CSIT

From an information-theoretic perspective, the optimal transmit strategy for the MU-MIMO BC is *dirty paper coding* (DPC) [218], and theoretical results showed that such strategy achieves the entire BC capacity region [52], [53]. The principle behind this optimum coding technique is that the transmitter knows the interference for each user. Therefore, interference can be pre-subtracted (from the information theoretical standpoint) before transmission, which yields the capacity of an interference free channel. DPC is a non-linear process that requires successive encoding and decoding, whose performance depends on the particular sequential order $\pi(\cdot)$ assigned to the co-scheduled users [219]. Although the implementation complexity of DPC for practical systems is prohibitively high, it establishes the fundamental capacity limits for MU-MIMO broadcast channels [46], [53].

Suboptimal yet more practical non-linear precoding schemes have been proposed as an alternative to DPC [36]. The error rate and interference can be minimized at the symbol level by the Tomlinson-Harashima precoding (THP),[6] which is not limited by the number of transmit or receive antennas [22]. By modifying or perturbing the characteristics of the transmitted signal, the power consumption can be minimized

[5] Some authors denote as precoding all processing techniques over the transmitted signals, which achieve multiplexing or diversity gains, i.e., both space-time coding and beamforming [143].

[6] The application of THP in MU-MIMO has been of particular interest in recent research on multibeam satellite communications [205].

TABLE II
PRECODING SCHEMES AND THEIR ASSOCIATED SCHEDULING METHODS

Method	Utility-based	CSI-Mapping	Metaheuristic (stochastic)	Classic Optimization	Exhaustive Search
DPC	[36]	-	[101]	-	-
THP	[36]	-	-	-	-
VP	[36]	[68]	-	-	-
MRT	[57], [152]	[80]	[70]	-	-
ZFBF	[36], [39], [57], [59], [60], [62], [67], [78], [85], [90], [95], [110], [111], [116], [135], [138], [143], [152], [155], [159], [162], [175], [178]	[33], [41], [55], [62], [63], [69], [71], [72], [73], [74], [80], [84], [86], [89], [91], [96], [97], [99], [106], [116], [117], [121], [129], [130], [131], [135], [139], [155], [158], [159], [161], [163], [164], [166], [170], [171], [175], [177], [181], [182]	[70], [78], [79]	[39]	[39], [82], [84], [95], [173]
ZFDP	[59], [102]	[61], [64], [65], [115], [119], [121]	-	-	-
SZF	-	[154]	[123]	-	-
CIZF	-	[75]	-	-	-
BD	[98], [100], [107], [113], [118], [122], [125], [126], [138], [141]	[66], [98], [103], [104], [113], [114], [117], [119], [120], [125], [129], [130], [131]	[124]	[39], [40]	-
MMSE/SLNR	[57], [88], [105], [108], [128]	[48], [134], [154], [168], [76], [89], [130]	[79]	-	-
Adaptive	[36], [47], [81], [87], [109], [137]	[83]	-	[77], [93], [94], [132]	-
Codebook-based	[30], [133], [142], [157], [162], [184]	[33], [66], [134], [144], [147], [156], [160], [165], [167], [168], [169], [172], [174], [176], [179], [180], [183]	-	-	-

(compared to traditional channel inversion filtering), using the non-linear vector perturbation (VP) scheme [220], [221]. This technique requires a multidimensional integer-lattice least squares optimization, whose solution can be found by several approaches.

B. Linear Precoding with full CSIT

Linear precoding is a generalization of traditional SDMA [18], which matches the signal on both ends of the radio link. This is attained by decoupling the input data into orthogonal spatial beams and allocating power according to CSIT [42]. The precoders weights are vectors or matrices jointly designed at the transmitter, according to several parameters: the type of CSIT, the coding order, the performance metric (e.g., mean-square error (MSE), error probability, power consumption, or achievable SNR), and the system constraints (e.g., power and QoS). The optimal precoder technique (linear filtering in the spatial domain), would be able to balance between signal power maximization and interference power minimization [222], [223]. Precoding design subject to general constraints can be performed using standard optimization techniques (see [224] for a comprehensive survey) or by heuristic approaches, e.g., maximizing the signal-to-leakage-plus-noise ration (SLNR[7]) [54], [226]. However, determining the optimal precoders is an NP-hard problem for many performance metrics [212], [227], whose evaluation is performed by computationally demanding algorithms [24], [217], [228]. Therefore, many works in the literature focus on more suboptimal, yet practical, schemes that can achieve spatial multiplexing gains with low computational complexity. A large number of linear precoding techniques have been developed, for which having more transmit than receive antennas, i.e., $M \geq N$, is a condition required in most cases. There are some particular precoding schemes (e.g. SLNR) that can be implemented if $M < N$, but the system becomes interference limited in the moderate and high signal-to-noise (SNR) regimes, [cf. Definition 3]. Therefore, power control and recursive adaptation of the precoders are mandatory to operate in those SNR regimes [226].

In a MISO antenna configuration, the matched filter or maximum ratio transmission (MRT) precoding maximizes the

[7] SLNR is also known in the literature as the transmit Wiener filter, transmit MMSE beamforming, or virtual SINR beamforming [225].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

signal power at the intended users. This is performed by projecting the data symbol onto the beamforming vector given by the spatial direction of the intended channel [54]. A similar precoding scheme for MIMO scenarios is the singular value decomposition (SVD) beamforming [22], which uses the eigenvectors of the channel as beamforming weights. The zero-forcing beamforming[8] (ZFBF) [52], also known as channel inversion precoding [229], completely suppresses IUI in MISO scenarios. This technique is based on prefiltering the transmit signal vector by means of the Moore-Penrose inverse [230]. An extension of ZFBF for MIMO scenarios is the block diagonalization (BD) precoding, where multiple data streams can be transmitted per user [231].

In MISO scenarios, the regularized channel inversion (often coined as MMSE precoding) enhances ZF, taking into account the noise variance to improve performance in the low SNR regime [229]. An extension of MMSE for the general MIMO scenario is the regularized BD [232]. Zero-forcing dirty paper (ZFDP) coding [52], is a technique designed for MISO settings. For a given user k, ZFDP suppresses the interference coming from the next encoded users $\{k+1,\ldots,K\}$, combining QR decomposition [230] and DPC. Extensions of ZFDP for the MIMO scenarios were defined in [231], [233] (an iterative SVD method), and [121] (combining ZF, DPC, and eigen-beamforming). Successive zero-forcing (SZF) was proposed in [233], for MIMO settings. SZF partially suppresses IUI, by encoding users similar to ZFDP, but DPC is not applied in the encoding process. The generalization of precoding schemes based on ZF for multiple antenna receivers is not trivial. This is because applying MISO decompositions methods to the MIMO channel is equivalent to treating each receive antenna as an independent user. This process does not completely exploit the multiplexing and diversity gains of MIMO systems [233].

The aforementioned precoding schemes can be classified as user-level precoders, i.e., independent codewords intended to different users are transmitted simultaneously. There is another class of precoding schemes, where simultaneous transmitted symbols are addressed to different users [234]. This class of symbol-level precoding, e.g. constructive interference zero forcing (CIZF) precoding [234], [235], has been developed for MISO settings. CIZF constructively correlates the interference among the spatial streams, rather than decorrelating them completely as in the case of user-level precoding schemes.

C. Precoding with partial CSIT

In the literature of limited feedback systems, CSIT acquisition relies on collections of predefined codewords (vector or matrix weights) or codebooks, that are known a priori at the transmitter and receiver sides. The codewords can be deterministic or randomly constructed, which defines the type of signal processing applied to achieve spatial multiplexing and interference mitigation [43]:

• *Channel quantization and precoding*: The codebook $\mathcal{C}=\{\mathbf{c}_1,\ldots,\mathbf{c}_b,\ldots,\mathbf{c}_{2^B}\}$, is used by user k to quantize its channel direction with B bits. This means that each user feeds back the index b, of the most co-linear codeword to its channel, [cf. Section V-B]. This is illustrated in Fig. 6, where the user k would report the index b, related to the cone where $\mathbf{H}_k$ has been clustered. The precoding weights are usually computed using ZFBF over the quantized channels. Due to quantization errors, the signals cannot be perfectly orthogonalized, and the sum-rate reaches a ceiling as the SNR regime increases [51]. In other words, resource allocation is performed over non-orthogonal spatial directions.

The optimal codebook design has not been fully solved in the literature. However, if the channel is assumed to have spatially i.i.d. entries, ($\mathbf{\Sigma}_k=\mathbf{I}$, and homogeneous long-term channel gains, $\forall k$), off-line designs of $\mathcal{C}$ can be realized using different approaches, e.g., the Grassmannian design [236], random vector quantization (VQ) [51], quantization cell approximation (QCA) [237], and other techniques described in [43], [177]. This sort of isotropically distributed codebooks achieve acceptable performance, since they mirror the statistical properties of the eigen-directions of $\mathbf{H}_k$, $\forall k$. For correlated channels ($\mathbf{\Sigma}_k\neq\mathbf{I}$), non-uniform or skewed codebooks must be constructed, taking into account the statistical characteristics of the dominant eigen-directions of $\mathbf{\Sigma}_k$, $\forall k$, see [140], [146], [150], [151].

• *Random beamforming (RBF)* [31]: The available precoding vectors, $\mathcal{F}=\{\mathbf{f}_1,\ldots,\mathbf{f}_b,\ldots,\mathbf{f}_{2^B}\}$, are constructed at the transmitter according to a known distribution,[9] or by methods that yield random orthonormal basis in $\mathbb{C}^{M\times 1}$, [165]. The transmitter sends pilot symbols to different user through the beams in $\mathcal{F}$, and each user feeds back the index b, of its best beam [cf. Section VIII]. RBF is an extension of the opportunistic beamforming scheme in [29], and attempts to sustain multiuser diversity over fading channels with partial CSIT. Random changes of amplitude and phase at each transmit antenna result in channels experiencing accelerated fluctuations, which enhances the MIMO processing gains. This approach is only effective when the number of users is large,[10] $K\gg M$, and the number of antennas at the transmitter M is moderate [165]. In some scenarios the performance can be enhanced by power control, or employing statistical channel information to compensate poor MUDiv.

We hasten to say that defining the optimum B, either for $\mathcal{C}$ or $\mathcal{F}$, is not a trivial optimization problem. In practice, only a tradeoff between feed back load and performance should be sought. Under mild conditions, B can be minimized for certain codebook designs while guaranteeing diversity gains [239]. Large codebooks provide accurate CSIT at a price of factors, large feedback signaling overhead and memory requirements at the receiver, which increase exponentially with B [164].

[8] Precoding based on ZFBF obtains a multiplexing gain of M and can asymptotically achieve the DPC performance when $K\rightarrow\infty$ [63].

[9] The goal is to generate 2^B codewords that are i.i.d., according to the stationary distribution of the unquantized beamforming vector [43].

[10] RBF asymptotically achieves the performance of downlink MIMO systems with full CSIT as $K\rightarrow\infty$. However, it does not achieve the full multiplexing gain at high SNR [46], and suffers from large quantization error as M grows [238].

D. Precoding in Multiple Transmitters Scenarios

Recent research work have focused on transmission cooperation and coordination in multi-cellular and heterogeneous networks. The signal processing techniques performed at the transmitters depend on several aspects [207], [208]: shared or non-shared user data; global or local CSIT; full or partial CSI; level of synchronization; per-transmitter power, backhaul, and delay constraints; full or limited coordination; heterogeneous SNR regime across the receivers; and the number of clustered transmitters.

In multi-cell scenarios, assuming full CSIT and shared user data across the transmitter, the precoders can be computed by techniques described in Section IV-B. However, the power should be constrained on a per-transmitter basis, instead of global power allocation as in single-transmitter scenarios [240]. The precoding design can be oriented to optimize different objective functions, e.g. power minimization or sum-rate maximization. Optimizing such functions deals, in general, with non-convex problems, whose solutions are non-linear precoders [208]. Precoding weights designed for common utility functions, e.g. weighted-sum-rate maximization or mean-square-error (MSE) minimization, can be realized via standard optimization techniques and sophisticated high-performance algorithms, see [224], [241] and references therein.

However, a more pragmatic approach for MU-MIMO, is given by linear precoding schemes (e.g. ZFBF, SLNR, or BD), which can be extended for multi-cell systems and providing reliable and low-complexity solutions [107], [242]. In *network MIMO* scenarios, multi-cell BD precoding can remove ICI for the clustered cells using centralized processing at a central controller (full coordination) [107]. In MISO multi-cell scenarios, linear precoding can be performed in a distributed fashion or with partial coordination. There are two fundamental schemes [54]: MRT (a competitive or egoistic scheme) and inter-cell interference cancellation (ICIC is a cooperative or altruistic scheme). For the MRT scheme, the transmitters ignore the ICI that they cause to unintended receivers in their vicinity. The goal is to maximize the received signal power of local users (effective in non-cooperative systems). The ICIC is a ZF-based scheme, which means that the transmitters design their precoders so that no interference is cause to non-intended receivers (effective in cooperative systems). The design of ICIC/ZF precoding is subject to constraints in the number of transmit and receive antennas, and its effectiveness highly depends on the CSIT across the transmitters. A close-to-optimal[11] distributed precoding scheme (for an arbitrary SNR), is attained by balancing MRT and ICIC, whose mathematical formulation is discussed in [54], [225]. If the system operates under a limited feedback constraint, the CSIT is acquired using the principles described in Section IV-C. Precoding design extensions from single to multi-cell scenarios are not straightforward. For instance, in multi-cell cooperative systems, the codebooks sizes may significantly increase [207], and depending on the delay tolerance and codebook granularity, the clustered transmitters should switch between MRT and ICIC schemes [243].

Several multi-cell CBF designs attempt to suppress IUI and ICI simultaneously with minimum coordination or control signaling between transmitters. These schemes rely on instantaneous or statistical CSI, and can be classified as [13]: *hierarchical* and *coupled* precoders. The hierarchical approach is implemented in systems where each transmitter suppresses ICI using ZF. This is attained through the sequential construction of outer and inner precoders that reduce ICI and IUI, respectively. The ICI cancellation is achieved by aligning interference subspaces at the receivers, whilst the IUI is suppressed at the transmitters using linear precoding and local CSI, e.g., [128], [130], [138], [244], [245], [246]. The coupled or nested-structure approach, is implemented in interference limited system, [cf. Definition 3]. The precoding design (beamforming weights and power allocation) optimizes an objective function subject to a set of QoS constraints, [cf. Section VI]. This kind of optimization problems has been extensively studied in the literature, covering power allocation [27] and beamforming design [224]. Under certain conditions, two specific problem formulations admit global optimal solutions via standard optimization in multi-cellular scenarios [1], [247], [248]: *i) power minimization subject to individual SINR constraints*. The CBF approach in [249] mitigates ICI by iteratively adjusting beamforming weights and transmit powers according to the experienced interference per user. *ii) the maximization of minimum SINR subject to power constraints*. The authors in [250] tackled this problem using the Perron-Frobenius theory [1], [27]. The joint beamforming and power allocation design is reformulated as an eigenvalue-eigenvector optimization problem, whose optimal solution can be found by geometrically-fast convergent algorithms.

E. Precoding in LTE-Advanced

In the LTE specification, CSI acquisition is called *implicit feedback* and relies on the following parameters [7], [214]: rank indicator (RI), which defines the number of data streams recommended for SU-MIMO transmission; precoding matrix index (PMI), which is the index of the best precoding matrix in the codebook; and channel quality indicator (CQI), which contains information of the channel quality corresponding to the reported RI and PMI [209]. The RI and PMI indexes provide the CDI of the MIMO channels, while the CQI indicates the strength of the corresponding spatial direction. The precoding matrices are defined at the BSs using different approaches [17], [213]: codebook-based precoding or non-codebook-based precoding (arbitrary precoder selection based on RI). The standard also supports a dual codebook structure in MU-MIMO mode. This increases codebook granularity, suppresses IUI more efficiently, and enhances the overall performance. The idea is that one codebook tracks the wideband and long-term channel characteristics, while the other tracks the frequency-selective and short-term channel variations [213]. In this way, the transmitter has more flexibility and accuracy when designing the precoding matrix. Another supported scheme is the *Per Unitary basis stream User and Rate Control* (PU2RC), which

[11] The optimal precoding design maximizes the received signal powers at low SNRs, minimizes the interference leakage at high SNRs, and balances between these conflicting goals at moderate SNRs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

allows multiple orthonormal bases per codebook, increasing quantization granularity [30], [156]. Although PU2RC uses deterministic codebooks, random codebooks can be used to simplify theoretical analysis, e.g. [30], [31], [51], [156]. The optimum number of bases in the codebook depends on the number of active users K, and should be optimized to maximize MUDiv and multiplexing gains [156].

LTE standard supports interference mitigation based on linear precoding schemes, such as SLNR and ZFBF with quantized channels [211]. The precoders can be dynamically recalculated after each CSI update, i.e., tracking channel variations. Linear precoding based on quantized channels outperforms codebook-based precoding (RBF or PU2RC), when then number of active users is small, i.e., *sparse networks*, $K \approx M$ [18]. The reverse holds as MUDiv increases, $K \gg M$, [156].

F. Precoding in 802.11ac

Sounding frames (null data packet) for MU-MIMO beamforming were introduced in 802.11n for channel estimation purposes. The transmitter sends known symbol patters from each antenna, allowing the receiver to sequentially construct the channel matrix, which is compressed and sent back to the transmitter [141]. The method employed to calculate the steering matrix is implementation and vendor specific relying on explicit CSI feedback, and is not defined by the 802.11ac standard [215], [251], [252]. One popular technique to construct the steering or precoding matrix is through SVD precoding [9], [141]. Since knowledge of CSIT is mandatory and feedback rates are limited, the receivers represent their estimated precoding matrix with orthogonal columns using Givens rotations [230]. Then, the set of calculated parameters (angles) at the receivers are adjusted, quantized, and fed back to the transmitter. In general, the final precoding matrix calculated by the transmitter will be different from the weights reported by the users due to the orthogonalization process [9].

Another practical approach is to compute the steering matrix using linear precoding schemes, e.g. ZFBF or MMSE [8]. In [126], the authors used BD and regularized BD with geometric-mean decomposition for a MU-MIMO scenario. The designs' goal is to balance the achievable SNR across the users, meeting the requirements of the standard. If the number of transmit antennas is larger than the total number of receive antennas, the additional DoF can be used to efficiently nullify inter-stream interference [8].

G. Discussion and Future Directions

The objective of precoding is to achieve spatial multiplexing, enhance link reliability and improve coverage in MIMO systems. Every physically realizable precoding design depends on the objective function, the CSIT accuracy, the number of transmitters involved, and more recently, the hardware characteristics [193], [197], [253]. The schemes described in previous subsections can be classified as *full digital precoders*, where the signal processing happens in the baseband at sub-6 GHz bands. However, these schemes cannot be directly implemented in state-of-the-art transceiver architectures with many antenna elements, neither upon higher frequency bands, i.e., mmWave [193].

Massive MIMO [12] differentiates itself from classical MU-MIMO by the fact that the number of antennas at the transmitter is larger compared to the number of served users. In conventional digital precoding, [cf. Section IV-B], each antenna element requires a radio frequency (RF) chain, i.e., signal mixer and analog-to-digital converter (ADC). In massive MIMO, although the number of antennas and RF chains is much larger than in conventional MU-MIMO, hundreds of low-cost amplifiers with low output power are used to replace the high power amplifier used in the latter. In order to keep the hardware cost and the circuit power consumption low, cost-effective and power-efficient hardware components are employed, e.g. low resolution ADC. And yet, this results in hardware impairments, especially low resolution quantization, that may affect the system performance. However, recent results and implementations show that the effect of hardware impairments, similarly to the effect of noise and interference, is averaged out due to the excess number of antennas [253]. Furthermore, the effect of quantization and AD conversion with very low number of bits can be taken into account in the precoding and signal design, providing schemes with promising spectral efficiency performance. In TDD massive MIMO, the major limiting factors were considered to be pilot contamination and channel reciprocity/calibration. However, both issues are now well studied and understood, and efficient transceiver designs compensating for pilot contamination and imperfect calibration are available. The major bottleneck for practical implementations of TDD-based massive MIMO remains the amount of training required. High peak rates have been demonstrated in downlink massive MIMO (in TDD and sub-6 GHz bands) with linear and non-linear precoding by several companies. Nevertheless, the amount of uplink training required can reduce the net throughput by at least half. A major challenge for massive MIMO is its successful deployment in FDD systems, in which CSI should be obtained by feedback. Efficient approaches for channel representation and CSI quantization and compression are necessary for viable implementations. Codebook-based approaches would require a relatively high number of bits for channel quantization and feedback, which not only will reduce the net throughput but also is not supported by current standards. Various new approaches for precoding and feedback in FDD massive MIMO are expected in the near future.

The current trend of using frequencies above 6 GHz for broadband wireless communications put an extra stress to massive MIMO systems. Current MIMO transceiver architectures may not be cost-effective and realistic in mmWave frequencies due to extremely high cost and power consumption. Different transceiver architectures have been recently proposed to address the hardware limitations. These new schemes require the joint optimization of precoding weights in the digital and analog domains, the so called *hybrid precoding* [193], [197], [254], [255]. In the digital domain, the low-dimensional precoding weights are computed using microprocessors. In the analog domain the RF precoders are implemented by phase shifters and variable gain amplifiers [95]. The main goal of hybrid precoding schemes is to achieve the performance of full digital precoders, but with a reduced number of RF chains [256]. The performance gap between the full digital and hybrid precoders depends on the spatial load at the transmitter, i.e., the ratio between the number of active data streams over the number of RF chains, M_{RF}, [188]. Notice that the number of co-scheduled users per resource is limited by M_{RF} in hybrid transceiver architectures, [cf. Section VII-E]. The optimal design of hybrid precoders has not been fully understood, and due to the power and amplitude constraints the sum-rate optimization problem becomes non-convex [257]. Therefore, ongoing research is focused on designing sub-optimal, yet efficient and practical, architectures that improve the joint performance of digital and analog precoders [188], [193], [197], [256], [258]. Although hybrid precoding provides a compromise between system performance and hardware complexity, it still remains challenging to implement reliable and cost-effect analog beamforming schemes at mmWave. Despite the complexity reduction using hybrid precoding, the performance gains using fully digital beamforming remain attractive. Hardware complexity may not be the major issue with digital beamforming at mmWave; significant challenges will arise in signal compression and CPRI protocols [259], which will require innovative solutions.

V. Spatial Compatibility Metrics

Let $\mathcal{K} \subseteq \bar{\mathcal{K}}$ and $\mathcal{K}' \subseteq \bar{\mathcal{K}}$ be two sets of user with at least one non-common user, where $\mathbf{H}(\mathcal{K}) = \{\mathbf{H}_i\}_{i\in\mathcal{K}}$ and $\mathbf{H}(\mathcal{K}') = \{\mathbf{H}_j\}_{j\in\mathcal{K}'}$ are their associated channels. A metric for spatial compatibility is a function of the CSIT that maps the spatial properties of the multiuser MIMO channels to a positive scalar value quantifying how efficiently such channels can be separated in space [72], i.e., $f(\mathbf{H}(\mathcal{K})) : \mathbb{C}^{|\mathcal{K}|N\times M} \mapsto \mathbb{R}_+$. Consider the two subsets $\mathcal{K}$ and $\mathcal{K}'$, a metric of spatial compatibility can be used to estimate the achievable performance of their associated multiuser channels, e.g., having $f(\mathbf{H}(\mathcal{K})) > f(\mathbf{H}(\mathcal{K}'))$ may imply that $\mathcal{K}$ is the set that achieves the maximum capacity. The mapping function $f(\cdot)$ can be used for scheduling purposes, depending on its definition and other system parameters (e.g. SNR regime and M), which will be discussed in Section VII-A2. Below we provide two definitions related to the design of user scheduling policies.

Definition 5. *Spatially compatible users*. A feasible set of users $\mathcal{K}$, is spatially compatible if the multiuser MIMO channels, $\mathbf{H}(\mathcal{K})$, can be separated in the spatial domain by means of beamforming/precoding.

Definition 6. *User grouping*. It is the task of forming a subset of users $\mathcal{K}$, according a compatibility criterion, e.g. spatial separability in Definition 5, to maximize the resource allocation and scheduling efficiency.

User grouping can be the initial step in a MU-MIMO scheduling algorithm since the characteristics of the joint channels dictate the transmission reliability and the resource allocation feasibility [cf. Definition 7]. In MU-MIMO scenarios, there exists a correspondence between the precoding capability to reduce IUI and the user grouping technique. The performance achieved by a precoder scheme is determined by the characteristics of the selected multiuser channels, i.e., providing *spatially compatible users* to the precoding processing block (see Fig. 3), is fundamental to guarantee high attainable SINRs at the receivers. It is worth mentioning that the vast majority of scheduling algorithms in the literature focuses on constructing sets of users with orthogonal or semi-orthogonal MIMO channels. Yet, for some particular signal designs, the best scheduling strategy is to group users whose channels are parallel or semi-parallel, see [75], [91]. This can also be the case in scheduling for non-orthogonal multiple access (NOMA)-MIMO schemes [6], [260], where the notion of spatial compatibility may be revised. The spatial compatibility metrics are used in the MU-MIMO literature to pair users and optimize the performance of a particular utility function. They are also used to quantize the channels in systems with limited feedback rates.

A. Null Space Projection

One of the objectives of MU-MIMO technology is to multiplex independent data streams to different users, which implies that only a subset of the transmitted data symbols are useful for each co-scheduled user. In such a scenario, a fundamental problem is to mitigate IUI, i.e., suppress the information intended to other receivers. There exist several signal processing techniques that can achieve such a goal, e.g., linear precoding or interference alignment. The effectiveness of such techniques rely upon the characteristics of the subspaces spanned by the MIMO channels, i.e, IUI is a function of the overlapped interference subspaces [261].

Consider a set of user $\mathcal{K}$ with K users, let $\tilde{\mathbf{H}}_k = [\mathbf{H}_1^T,\ldots,\mathbf{H}_{k-1}^T,\mathbf{H}_{k+1}^T,\ldots,\mathbf{H}_K^T]^T$, be the aggregated interference matrix of the user k such that $M > \max_k \ \text{rank}(\tilde{\mathbf{H}}_k)$, which is a necessary condition to suppress IUI [117]. Define $\mathcal{V}_k = \text{Span}(\tilde{\mathbf{H}}_k)$, as the subspace spanned by the channels of the subset of users $\mathcal{K}\setminus\{k\}$, and let $\mathcal{V}_k^{\perp} = \text{Span}(\tilde{\mathbf{H}}_k)^{\perp}$, be its orthogonal complement subspace. In other words, $\mathcal{V}_k^{\perp}$ spans the null space of $\tilde{\mathbf{H}}_k$, i.e., $null(\tilde{\mathbf{H}}_k) = \{\mathbf{x}\in\mathbb{C}^{M\times 1} : \tilde{\mathbf{H}}_k\mathbf{x} = \mathbf{0}\}$. The channel of the k-th user can be expressed as the sum of two vectors $\mathbf{H}_k = \mathbf{H}_k^{(\|)} + \mathbf{H}_k^{(\perp)}$, each one representing the projection of $\mathbf{H}_k$ onto the subspaces $\mathcal{V}_k$ and $\mathcal{V}_k^{\perp}$ respectively, as illustrated in Fig. 5.

For all user-level ZF-based precoding schemes described in Section IV-B, $\mathcal{V}_k^{\perp} = \bigcup_{i=1,i\neq k}^{K} \text{Span}(\mathbf{H}_i)^{\perp}$, i.e., $\mathcal{V}_k^{\perp}$ contains all overlapped interference subspaces of channel $\mathbf{H}_k$. The component $\mathbf{H}_k^{(\|)}$ is related to the signal degradation of the user k due to channel correlation, whereas $\mathbf{H}_k^{(\perp)}$ defines the *zero-forcing direction*, i.e. the spatial direction that is free of IUI. The squared magnitude of $\mathbf{H}_k^{(\perp)}$ is known as the null space projection (NSP), and directly computes the effective channel gain obtained by ZF precoding [52].

A common approach in the literature is to define the mapping $f(\mathbf{H}(\mathcal{K}))$ as a function of the exact or approximated *effective channel gains* [cf. Definition 4]. In other words, the spatial compatibility metric has to consider the precoding

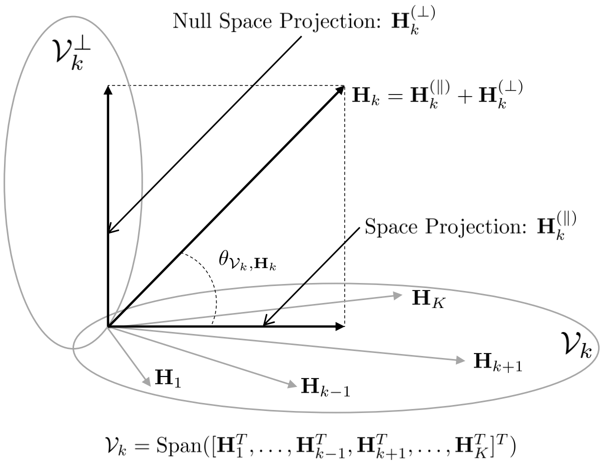

Null Space Projection: $\mathbf{H}_k^{(\perp)}$
$\mathcal{V}_k^{\perp}$
$\mathbf{H}_k = \mathbf{H}_k^{(\|)} + \mathbf{H}_k^{(\perp)}$
Space Projection: $\mathbf{H}_k^{(\|)}$
$\theta_{\mathcal{V}_k,\mathbf{H}_k}$
$\mathbf{H}_K$
$\mathcal{V}_k$
$\mathbf{H}_{k+1}$
$\mathbf{H}_1$
$\mathbf{H}_{k-1}$
$\mathcal{V}_k = \mathrm{Span}([\mathbf{H}_1^T, \ldots, \mathbf{H}_{k-1}^T, \mathbf{H}_{k+1}^T, \ldots, \mathbf{H}_K^T]^T)$

Fig. 5. Decomposition of the channel $\mathbf{H}_k$ given by its projection onto the subspaces $\mathcal{V}_k$ and $\mathcal{V}_k^{\perp}$.

scheme, the SNR regime, and the available spatial DoF. Analytical results in [39], [71], [187] have shown that the set of users $\mathcal{K}$ that maximizes the product of their effective channel gains[12] is also the set that maximizes the sum-rate in the high SNR regime. The NSP has been extensively used for user grouping and scheduling purposes whenever ZF-based precoding is implemented [cf. Section VII-A]. Note that a set of spatially compatible users would have large NSP components, which means that their associated channels are semi-orthogonal.

The computation of NSP for the general MU-MIMO scenario where $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$ $\forall k$, can be performed by several approaches, namely SVD [99], [102], [230], [263], [264], orthogonal projection matrix [52], [114], [263], [264], Gram-Schmidt orthogonalization (GSO) [63], [98], [119], [230], [263], QR decomposition [122], products of the partial correlation coefficients [262], [264], and ratio of determinants [68], [262]. Several works find sets of spatially compatible users by computing an approximation of the NSP (e.g. [54], [89], [97], [114], [126], [135], [170]), which yields efficient scheduling designs that require low computational complexity. Moreover, recent works ([265] and references therein), have proposed efficient methods to compute and track null spaces for a set of co-scheduled users, which is fundamental to mitigate IUI for ZF-based precoding in single and multiple transmitter scenarios.

B. Spatial Clustering

Spatial clustering refers to the association of a channel matrix to a spatial subspace. This technique has been used in the literature to perform user/channel grouping or scheduling in compliance with the precoding scheme and CSIT availability. In feedback limited scenarios, spatial clustering is also used to quantize the channel or precoder weights [cf. Section IV-C]. Let $\measuredangle(\mathbf{H}_i, \mathbf{H}_j)$ denote the angle between i.i.d. channels $\mathbf{H}_i, \mathbf{H}_j \in \mathbb{C}^{1\times M}$, and define the normalized inner product, also known as *coefficient of correlation*, as [230]:

$$\cos(\measuredangle(\mathbf{H}_i, \mathbf{H}_j)) = \frac{|\mathbf{H}_i \mathbf{H}_j^H|}{\|\mathbf{H}_i\|\|\mathbf{H}_j\|}, \tag{4}$$

where $\cos(\measuredangle(\mathbf{H}_i, \mathbf{H}_j)) \in [0, 1]$, and $\measuredangle(\mathbf{H}_i, \mathbf{H}_j) = \frac{\pi}{2}$, implies that the channels are spatially uncorrelated or orthogonal. The coefficient of correlation indicates how efficiently the transmitter can serve user i without affecting user j, and vice versa. For particular MU-MISO scenarios, in which the set of scheduled users is subject to the cardinality constraint $|\mathcal{K}| = 2$, the NSP is a function of $\sin^2(\measuredangle(\mathbf{H}_i, \mathbf{H}_j))$, which can be evaluated from (4) simplifying the user pairing and scheduling design [33]. Several works focused on sum-rate analysis (e.g., [33], [80], [266]) limit the cardinality to $|\mathcal{K}| = 2$ for simplicity and tractability. Other works analyzed scenarios with more practical constraints and considered two or four users, which are common values of K in standardized systems, see [8], [160], [213]. The metric (4) has been extensively used to define policies for user grouping. In MU-MIMO systems a set of user $\mathcal{K}$ is called ϵ-orthogonal if $\cos(\measuredangle(\mathbf{H}_i, \mathbf{H}_j)) < \epsilon$ for every $i \neq j$ with $i, j \in \mathcal{K}$ [74], [84], [97]. The users can be grouped into disjoint sets according to a desired threshold ϵ, and the semi-orthogonal sets can be scheduled over independent carriers (e.g. [41], [73], [116]), codes (e.g. [41]), or time slots (e.g. [63]). The optimum value of ϵ depends on the deployment parameters (K and M) and is usually calculated through simulations, since for $M > 2$ it is very hard or impossible to find the optimal ϵ in closed-form. Nevertheless, there exist closed-form expressions to compute the ergodic capacity as a function of ϵ for MU-MISO scenarios where the transmitter has two antennas $M = 2$ and the users have homogeneous [33, Ch. 7] or heterogeneous [80] large-scale fading gains.

Consider that a codebook, $\mathcal{F}$, is used to define the CDI of the MIMO channels, i.e., assume partial CSIT, [cf. Section IV-C]. In such a scenario, spatial clustering can be performed for a given parameter θ and the codeword $\mathbf{f}_i \in \mathbb{C}^{1\times M}$. Define the hyperslab $\mathfrak{F}_i(\theta)$ as [66]:

$$\mathfrak{F}_i(\theta) = \left\{\mathbf{H}_k \in \mathbb{C}^{1\times M}, k \in \mathcal{K} : \cos(\measuredangle(\mathbf{H}_k, \mathbf{f}_i)) \leq \cos(\theta)\right\}.$$

The hyperslab defines a vector subspace whose elements attain a spatial correlation not greater than $\cos(\theta)$, w.r.t. the codeword $\mathbf{f}_i$, as illustrated in Fig. 6. The parameter $\cos(\theta)$ is set to guarantee a target ϵ-orthogonality. The generalization of the hyperslab clustering[13] for MIMO settings is straightforward, i.e. $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$, $\mathbf{f}_i \in \mathbb{C}^{M\times M}$, and the coefficient of correlation between matrices can be defined as in [119], [122]. The parameter ϕ shown in Fig. 6 can be adjusted to fix the maximum co-linearity between cones [184].

Several scheduling algorithms based on spatial clustering (e.g., [48], [62], [63], [66], [68], [69], [117], [121], [122], [159], [166], [181], [266]), can achieve MUDiv gains and improve the overall performance by adjusting the parameter θ (or threshold ϵ) according to the number of competing users

[12] Some works in the literature (e.g. [84], [173], [262]) optimize the sum of effective channel gains instead of their product, which yields similar performance for the high SNR and large number of users, i.e. $K \gg M$.

[13] Some authors (e.g., [48], [84]) define $\mathfrak{F}_i(\theta)$ as a function of two parameters, θ and the minimum acceptable channel magnitude. In this way only sets of spatially compatible and strong channels can be constructed.

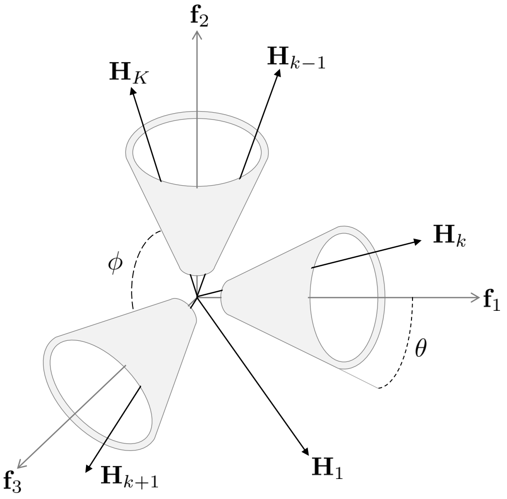


Fig. 6. The channels $\{\mathbf{H}_j\}_{j=1}^{K}$ can be clustered within cones given a set $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ formed either by orthonormal basis or codebook words.

[63], the SNR regime, the large-scale fading gain, and the precoding scheme [80]. In [48], it was shown that the optimal cardinality of the user set grouped based on spatial clustering is a function of K, θ, and M. The statistical properties of (4) have been extensively studied in the literature cf. [51], [63], [83], [84], [112], [237], [267], and such properties depend on the type of CSIT (full or partial), the MIMO channel distribution, and the system parameters (e.g. M and B).

C. Compatibility between Subspaces

In heterogeneous MU-MIMO scenarios where each user is equipped with N_k antennas $\forall k$, the signal spaces span several dimension and mutual interference between user can be measured either in the angular or in the subspace domains [117], [120]. Moreover, metric (4) cannot be used directly in heterogeneous MU-MIMO scenarios to measure spatial correlation between channels of different dimensions, thus, several alternative metrics have been proposed in the literature. Consider a set of users $\mathcal{K}$ and its MU-MIMO channel $\mathbf{H}(\mathcal{K})$; the spatial compatibility can be measured as a function of the corresponding eigenvalues:

i) The *orthogonality defect* [97] is a metric derived from Hadamard's inequality [268] that measures how close a basis is to orthogonal. It quantifies the energy degradation of the channel matrix due to the correlation between all its column vectors. User grouping algorithms based on this metric have been developed for homogeneous MIMO systems, e.g. [102], [262].

ii) The determinant geometrically represents the volume of the parallelepiped defined by the column vectors of the channel matrix. Larger determinant values implies that the column vectors of a matrix are more orthogonal [230]. The *matrix volume ratio* used in [89], [103] measures the volume reduction of $\mathbf{H}(\mathcal{K})$ w.r.t. $\mathbf{H}(\mathcal{K} + \{k'\})$, where $k' \notin \mathcal{K}$ is a candidate user attempting to be grouped. Different approximations of such a metric can be done using the arithmetic-geometric mean inequality over the squared singular values of $\mathbf{H}(\mathcal{K})$, e.g., [144], [262].

iii) The *geometrical angle* (see [104], [120] and references therein), is a metric similar to the matrix volume ratio that measures the spatial compatibility as a function of the all possible correlation coefficients [cf. (4)] between the basis of two subspaces. It can be computed from the eigenvalues of two MIMO matrices $\mathbf{H}_i \in \mathbb{C}^{N_i \times M}$, $\mathbf{H}_j \in \mathbb{C}^{N_j \times M}$, whose subspaces have different dimensions, i.e., $N_i \neq N_j$.

iv) If the elements of a multiuser channel matrix are highly correlated, the matrix is said to be ill-conditioned, i.e., it is close to singular and cannot be inverted. In numerical analysis, the *condition number* [230] quantifies whether a matrix is well- or ill-conditioned, and is computed as the ratio between the maximum and minimum eigenvalues. This metric is used to measure how the eigenvalues of $\mathbf{H}(\mathcal{K})$ spread out due to spatial correlation. The ratio between the condition numbers of two MIMO channels can be used to quantify their spatial distance and compatibility [202].

For a given user $k \in \mathcal{K}$, the spatial compatibility between $\mathbf{H}_k$ and $\tilde{\mathbf{H}}_k$, can be measured by a function of $\theta_{\mathcal{V}_k \mathbf{H}_k}$, which is the angle between $\mathcal{V}_k$ and Span$(\mathbf{H}_k)$, see Fig. 5. The following metrics assess the spatial compatibility based on geometrical properties of the multiuser channels:

i) The *principal angle between subspaces* [104], [117], [120] measures the relative orientation of the basis of $\mathcal{V}_{\mathbf{H}_k} = \text{Span}(\mathbf{H}_k)$ regarding the basis of $\mathcal{V}_{\tilde{\mathbf{H}}_k} = \text{Span}(\tilde{\mathbf{H}}_k)$. Given the bases of $\mathcal{V}_{\mathbf{H}_k}$ and $\mathcal{V}_{\tilde{\mathbf{H}}_k}$, the principal angle is associated with the largest coefficient of correlation [cf. (4)] between the bases of both subspaces.

ii) The *chordal distance* is extensively used in limited feedback systems for codebook design [236], [269], user grouping and scheduling [100], [120], [138], [175]. This metric can be computed either from the principal angles between subspaces or from the projection matrices of $\mathcal{V}_{\mathbf{H}_k}$ and $\mathcal{V}_{\tilde{\mathbf{H}}_k}$ in heterogeneous MU-MIMO scenarios, i.e., $N_i \neq N_j$ $\forall i, j \in \mathcal{K}$ with $i \neq j$.

iii) The *subspace collinearity* [120], [202] quantifies how similar the subspaces spanned by two channel matrices are, following the rationale behind (4). Given $\mathbf{H}_i, \mathbf{H}_j \in \mathbb{C}^{N \times M}$ the subspace collinearity of the matrices compares the singular values and the spatial alignment of their associated singular vectors.

iv) Other metrics measuring the distance between subspaces are the *weighted likelihood similarity measure*, the *subspace projection measure*, and the *Fubini-Study similarity metric*. These metrics have been recently propounded in [138] and used for user grouping based on statistical CSI, i.e., $\mathbf{\Sigma}_k$ $\forall k$.

D. Discussion

It is worth mentioning that neither the principal angles, nor the chordal distance can fully measure spatial compatibility in heterogeneous MU-MIMO scenarios. This is due to the fact that these metrics take into account the smallest dimension between two subspaces, potentially neglecting useful spatial correlation information [120]. Moreover, metrics that only evaluate the spatial separation between hyperplanes or the eigenvalue dispersion of $\mathbf{H}(\mathcal{K})$ neglect the degradation of the MIMO channel magnitude due to the interference subspace. Such metrics fail at maximizing the capacity since they do not evaluate or approximate the effective channel gain [100]. As the number of active users grows, the set of users that maximizes a given spatial compatibility metric may diverge from the set that maximizes the capacity, either in their elements, cardinalities, or both [120], [262].

Authors in [103] pointed out that user grouping should be a function of the spatial correlation between $\mathbf{H}_k$ and $\tilde{\mathbf{H}}_k$, and also consider the inner correlation of each multi-antenna user, i.e., the magnitudes of the eigenvectors of $\mathbf{H}_k$ $\forall k$. According to [42], [150], [198], the precoding performance is primarily affected by the correlation between transmit antennas, whereas receive antenna correlation has marginal or no impact on the precoding design. Nonetheless, the effects of receive antenna correlation has not been fully studied in the user selection literature and conclusions are usually drawn based on specific correlation models. The knowledge of statistical CSI, i.e., $\mathbb{E}[\mathbf{H}_k^H \mathbf{H}_k]$, may be assumed in scenarios with practical constraints such as limited feedback rates [cf. Section IV-C]. If the transmitter has knowledge of statistical CSI, the dominant eigen-directions of the channel covariance matrix ($\mathbf{\Sigma}_k$, $\forall k$) can be used as metrics to identify compatible users, e.g., [58], [129], [131], [134], [135], [138], [270]. Spatial clustering based on $\mathbf{\Sigma}_k$, $\forall k$, has been recently proposed to identify users with similar channel statistics in massive MIMO settings, see [13], [129], [138].

VI. System Optimization Criteria

The optimization criteria determines the optimal resource allocation strategy [198], and can be classified in two groups according to the objective function and constraints [187]. *i*) *PHY layer* based criteria, where a objective function, $U(\cdot)$, must be optimized and channel information is the only input to the resource allocation algorithms. *ii*) *Cross layer* based criteria, where optimization of $U(\cdot)$, takes into account QoS requirements (defined by upper layers) and channel information, see Fig. 3. This section presents an overview and classification of the objective functions (criteria), and their associated constraints in the MU-MIMO literature. A summary of the content and general organization of this section are presented in Fig. 7 and Table III. Two relevant concepts in multiuser system optimization are defined below.

Definition 7. *Resource allocation feasibility*. For given a set of users $\mathcal{K}$, a resource allocation strategy is called feasible if it fulfills all individual and global constraints (e.g. power and QoS), implementing precoding, power control, or a combination of both.

Definition 8. *Feasible set of users*. Given the set of all competing users $\bar{\mathcal{K}}$, the subset $\mathcal{K} \subseteq \bar{\mathcal{K}}$, is called feasible if there exist precoding weights and powers, $\forall k \in \mathcal{K}$, such that $U(\cdot)$ has a solution meeting individual QoS and power constraints.

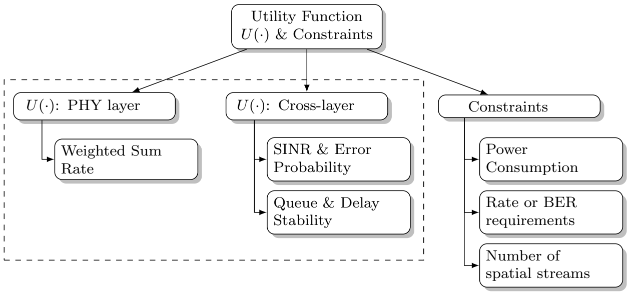


Fig. 7. Utility Functions and constraints for PHY layer and cross layer based optimization in MU-MIMO systems

A. Weighted Sum Rate (WSR) Maximization

The system optimization is achieved by maximizing the sum of individual utility functions, $U_k(\cdot)$, $\forall k \in \mathcal{K}$, subject to a set of power constraints and set cardinality ($|\mathcal{K}|$). Global optimization based on concave and differentiable utility functions is desirable, since efficient numerical methods can be applied to guarantee optimality. A common family of functions studied in the literature is defined by the α-*fair* utility function [224], [271], [272], which can be used to characterize the objective of the resource allocation strategies.

The α-*fair* function is mathematically expressed as [272]:

$$U\left(\{R_k\}_{k=1}^K\right) = \begin{cases} \sum_{k=1}^{K} \log(R_k), & \text{if } \alpha = 1 \quad \text{(5a)} \\ \sum_{k=1}^{K} \frac{(R_k)^{1-\alpha}}{1-\alpha}, & \text{otherwise} \quad \text{(5b)} \end{cases}$$

where R_k denotes the achievable transmission rate of the k-th user. Changing the values of α yields different priorities to the users, and can be used to define performance tradeoffs (e.g. throughput-fairness). For instance, $\alpha = 1$, yields maximum fairness, $\alpha = 0$, generates the sum-rate utility, and $\alpha \to \infty$ defines the minimum-rate function. These functions are, in general, non-concave,[14] and several methods have been developed to optimize them subject to power and QoS constraints, see [224] for an in-depth review.

A more common formulation of the weighted sum rate in MU-MIMO scenarios is given by the following expression:

$$U\left(\{R_k\}_{k=1}^K\right) = \sum_{k=1}^{K} \omega_k R_k \quad (6)$$

where ω_k is a non-negative time-varying weight, which defines the priority of the k-th user. This expression is related to the

[14] For instance, non-concave utility functions may arise in the context of real-time applications with hard QoS constraints [273].

TABLE III
SUMMARY OF SYSTEM OPTIMIZATION CRITERIA FOR SCHEDULING IN MISO ($N = 1$) AND MIMO ($N > 1$) CONFIGURATIONS WITH FULL ($B = \infty$) AND PARTIAL ($B < \infty$) CSIT. QOS REFERS TO SINR, BER, OR QUEUE STABILITY REQUIREMENTS

	Sum Rate	Fairness	WSR	QoS	Round Robin
MISO, $B = \infty$	[33], [36], [41], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [74], [75], [76], [78], [79], [80], [83], [89], [91], [92], [93], [95], [130], [137], [138], [139]	[41], [63], [66], [72], [75], [78], [88], [96], [137], [138], [152]	[87], [90], [112], [128], [135]	[47], [47], [72], [73], [73], [75], [77], [82], [84], [85], [86], [94], [112], [132], [134]	[63], [66], [79], [91], [96], [137]
MISO, $B < \infty$	[30], [33], [51], [57], [66], [72], [76], [129], [130], [131], [133], [144], [153], [154], [156], [157], [158], [159], [160], [161], [162], [163], [164], [165], [166], [168], [171], [173], [174], [181], [182]	[55], [66], [131], [143], [144], [152], [153], [154], [167], [169], [170], [171]	[81], [143], [155]	[30], [48], [55], [134], [143], [172]	[57], [66], [172]
MIMO, $B = \infty$	[39], [97], [97], [98], [99], [100], [101], [102], [103], [104], [107], [110], [111], [114], [115], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126]	[101], [104], [105], [108], [109], [114], [119], [126], [127]	[122]	[40], [106], [113], [116], [127]	[39], [97], [99], [114]
MIMO, $B < \infty$	[31], [104], [114], [117], [141], [175], [177], [178], [183], [184]	[104], [114], [141], [142], [179], [180]	[147]	[116], [141], [176]	[114], [178]

α-fair function, e.g. if $\omega_k = 1$, then (6) converges to (5b) with $\alpha = 0$. The dynamic adaptation of ω_k can provide different levels of fairness (temporal, utilitarian, rate-proportional, fair-queuing, etc.) over the resource allocation at the *user level* [274].

Another approach to provide fairness is defining $U\left(\{R_k\}_{k=1}^K\right)$ as a function of the rates' statistics. For instance, one might use the cumulative distribution function (CDF) of R_k $\forall k$, which can be empirically estimated at the transmitter [167], [275]. Fairness at the *system level* is another optimization criterion, measured over the average rates, transmit powers, time-slots, or any other allocated resource. Different metrics to quantify fairness can be found in the literature of resource allocation [276]. For a single resource, one can use the Gini index [277], Jain's index [278], or other fairness measures in [279] and reference therein. Generalized fairness metrics over multiple resources allocated simultaneously were defined in [276], [280] and references therein. The weights in (6) might be determined by upper layer requirements, such as buffer sizes, traffic/service priority, packet loads, or any other QoS-related metric. The individual weights affect the scheduling decisions, and can be used to improve precoding design [281].

A special case of (6) is given by setting $\omega_k = 1$ $\forall k$, which represents the sum rate. This criterion defines the maximum amount of error-free information successfully received by a set of co-scheduled users, regardless of fairness [23]. The resource allocation strategies are usually assessed in terms of sum rate, since it quantifies the effectiveness of an algorithm and simplifies scheduling rules. However, maximizing the sum rate in scenarios where the users have heterogeneous SNRs might be very inefficient, since users with poor channel conditions would experience starvation [63], [112]. A classical solution to such a problem is assigning the weights ω_k $\forall k$, according to a fairness criterion. The works listed under the *fairness* category in Table III, optimize the long-term proportional fairness, extending the original idea from [29], [282] to MU-MIMO scenarios.

Round robin is a channel-unaware scheduling method, that allocates the same amount of time to all competing users. It is the simplest form of fair resource-access, but neglects MUDiv, the amount of occupied resources, and QoS requirements [2], [16]. Equal opportunity for resource sharing may not mean equal distribution of resources, which results in wastage, redundant allocation, or resource starvation [276]. Nonetheless, round robin might be useful in ultra dense heterogeneous networks, where transitions from non-LoS to strong LoS channels alter the performance of conventional resource allocation strategies [283]. The works listed under the round robin category in Table III, use the method as a benchmark to assess MUDiv gains in MU-MIMO systems. It is worth mentioning that in the reviewed literature, the system performance is assessed over a large number of channel realizations. For example, using the average sum rate per channel use, also called spectral efficiency, measured in bit/s/Hz. Other common performance metrics are [284]: the *throughput* or average rates achieved for packet transmission with practical modulations and realistic coding schemes; and the *goodput*, which is the total average rate successfully delivered to the scheduled users, including layer 2 overheads and packet retransmission due to physical errors [16].

B. WSR with QoS Optimization

The WSR optimization problem can incorporate different types of QoS requirements. The QoS defines a network-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

or user-based objective [cf. Definition 1], and the weights ω_k $\forall k$, establish priorities according to the service provided to each user. Taking into account QoS requirements reduces the flexibility of the resource allocation strategies. For instance, in the case of opportunistic transmission, achieving a target QoS reduces the value of CSIT, i.e., the channel conditions have to be weighted by requirements coming from the upper layers [18]. System optimization considering QoS has a broader scope and once information from upper layers is considered, a more complex cross-layer optimization problem arises.

The QoS is defined according to the system model, which might include individual traffic characteristics or SINR requirements. Satisfying QoS requirements can be realized at different time granularities: each user must achieve a prescribed performance metric, e.g. instantaneous SINR or data rate [25]; or a network performance metric must reach a stable behavior over time, e.g. attaining finite buffer sizes [81]. Below, we provide a classification of the optimization criteria subject to QoS constraints in MU-MIMO scenarios.

1) Target SINR and Error Rates: The QoS can be guaranteed as long as individual SINR requirements are met, which also satisfies target error and peak rates. The QoS can be defined by a monotonic and bijective function of the SINR: one example is the bit error rate (BER), given a particular modulation and coding scheme (MCS) [73], [85]; another example is the Shannon capacity [25]. Different tradeoffs between global and individual performance are achieved under SINR constraints: maximizing the WSR [227], max-min weighted SINR [1], [24], [248], [277], [285], sum power minimization [24], [25], [83], [224], and other hybrid formulations, see [224] and references therein.

It is worth noticing that in systems with SINR requirements, it is usually assumed that for each channel realization, a set of compatible users $\mathcal{K}$ has been previously selected. The system optimization is realized by power allocation and precoder design [227]. This means that $\mathcal{K}$ must be a feasible set of users [cf. Definition 8], and user scheduling is required only if the constraints associated with $\mathcal{K}$ are infeasible [1], [224]. In that case, it is necessary to relax the initial conditions taking one of the following actions:

i) Reducing the number of selected users applying *admission control* [3], [132], [277] or *user removal* schemes [47], [72], [286], [287]. These procedures are implemented in systems where the achievable SINRs take any positive real value, e.g. [24], [83], [219], [248], [285]. The removal process can be thought as a form of scheduling, used to generate a feasible set of users [cf. Definition 8]. By identifying users with unfeasible constraints, one can drop them temporarily (according to priorities), or re-schedule them over orthogonal resources (e.g. other carriers or time slots) [10].

ii) By relaxing the individual QoS requirements $\forall k \in \mathcal{K}$, one can achieve resource allocation feasibility [cf. Definition 7]. For instance, consider homogeneous target rates, then all scheduled users can be balanced to a common SINR [47]. For heterogeneous QoS, one can iterate over different SINR levels per user (e.g. provided a set of MCSs), so that link adaptation can take place, see [82], [227], [277], [287].

Cross-layer algorithms subject to rate constraints can be designed so that $\mathcal{K}$ and its associated resources are jointly optimized. Analytical results in [48] show that under mild conditions, there exists, with probability one, a feasible set of users $\mathcal{K}$ that can be found though low-complexity algorithms [cf. Section VII-A].

The transmission errors (bit, frame, symbol, packet) are usually due to noisy channels and inaccurate CSIT. For the former factor, the errors occur due to the effect of non-ideal channel coding and finite blocklength channel coding. The error rates can be reduced by implementing stronger channel codes and longer blocklengths.[15] The error rates also occur due to CSI quantization (limited feedback), estimation errors, delays (channel outage), and distortion during the feedback [44], [57], [143]. Therefore, IUI cannot be fully suppressed by precoding techniques [cf. Definition 3], which degrades the achievable SINRs and BER figures. Nonetheless, multi-antenna receivers can implement efficient processing techniques to increase their SINRs and combat quantization errors, e.g., quantization-based combining (QBC) [178]. Under partial CSIT conditions, any practical rate adaptation scheme achieves performance between the worst-case outage rate and the ideal-case, where the achievable rate equals the mutual information [143].

2) Queue Stability in MU-MIMO Scenarios: A network-based optimization requires cross-layer algorithms taking as inputs the queue state information (QSI), CSIT, and their interdependence [84]. There exist a rich literature on optimization subject to queue or packet delay constraints, specially for systems with orthogonal resource allocation (e.g. one user per time-slot or sub-carrier), see [4], [10], [16] for an in-depth overview. According to the QSI relevance, MU-MIMO systems fall in two categories [81], [187]:

i) *Systems with infinite backlogs*: In this scenario, the transmitter is assumed to always have data to send and infinite buffer capacity. In other words, the transmitter fully knows the intended data for each user. The objective of the resource allocation strategies is, in general, to maximize the WSR [cf. Section VI-A]. The weight ω_k of the k-th user is used to reach a desired throughput-fairness tradeoff, instead of expressing the urgency of data flows. Thus, it can be assumed that the queues are balanced and users cannot be discriminated based on them [84]. The service provided is delay-insensitive, and the traffic is managed so that the scheduled users attain non-zero rates [106]. MU-MIMO systems without QSI information fall into this category.

ii) *Systems with bursty traffic and limited buffer size*: In these systems, packet data models with stochastic traffic arrivals are considered. The performance assessment is analyzed from the delaying and queuing perspectives. The WSR maximization attempts to balance between

[15] In practice, for reasonable block length (e.g. 8 kbits) and strong coding (e.g. LDPC), the Shannon capacity can be approached to within 0.05 dB for a target FER of 10^3 [22], [48], [78].

opportunistic channel access and urgency of data flows. The resource allocation algorithm must guarantee finite average buffer occupancy, i.e., *stability of the queues lengths* for all users [176], [219]. It is worth mentioning that by Little's theorem [10], [176], [187], achieving stability in the average queues is equivalent to minimize the average packet delay in the steady state.[16] The system model might consider different sources of delay, e.g., buffer congestion or destination unavailability (outage delay), which define the type of scheduling policy to be implemented [106]. The choice of the system model and problem formulation depend on the desired tradeoff between tractability and accuracy in each particular scenario [187].

In MU-MIMO systems where queue stability is optimized, the scheduling algorithm has to guarantee that the average queue lengths of all users are bounded. Simultaneously, the available CSIT should be opportunistically exploited for throughput maximization [219]. This goal can be achieved by establishing queue-based WSR as the optimization criterion. The weights ω_k in (6) can be defined according to the system requirements: considering the packet queue length or packet departure rates at each scheduling interval [55], [82], [84], [143], [176]; considering fairness with heterogeneous traffic rates [105], [248]; or considering service-oriented requirements, e.g., BER, delay tolerance, and packet dropping ratio for real- or non-real-time services [73], [85], [106]. The queue lengths not only define the priority or urgency of the traffic, but they can also define the encoding order, $\pi(\cdot)$ in (1), see [82], [84], [219]. In general, queue lengths are not symmetric or balanced, which means that the WSR optimization is dominated by the QSI, rather than by the CSIT [55], [84].

The characteristics of the MU-MIMO system model and the set of constraints define the resource allocation strategy that achieves stability. Several factors may affect the performance under QoS constraints: CSIT accuracy [55], [176], CSIT statistics [143], and the number of resources used for channel estimation [82]; the spatial compatibility between co-scheduled user and the SNR regime [84]; the user priority based on the type of service [73], [85]; or the number of simultaneous spatial streams [106]. The overall optimization can also consider delays due to packet losses and retransmissions. This is supported by ACK/NAK (handshake) signaling exchange in the upper layers. Such a protocol is used for automatic repeat request (ARQ), to convey an error-free logical channel to the application layers [43]. Works in [86], [143] and references therein, discuss algorithms to solve the WSR with queue constraints taking into account ARQ protocols. We refer the reader to [187], [288] for a comprehensive texts on cross-layer design under queue and delay constraints.

C. Discussion and Future Directions

The sum-rate maximization is the main criterion to assess conventional cellular system, specially in scenarios with scarce and expensive radio resources, e.g., crowded sub-6 GHz bands. Conventional cellular planning usually considers few high-power transmitters that provide high spectrum efficiency, at expense of other performance metrics, such as energy efficiency (EE) [289]. 5G mobile communications will include dense deployments, operating at higher frequencies, i.e., mmWave, and with very heterogeneous radio resources per base station [14], [290], [291]. The criteria presented in Sections VI-A and VI-B are used to define single objective functions subject to a set of constraints. However, 5G networks will require the simultaneous optimization of multiple criteria: peak data rates, traffic and user load across the network, fairness, quality of service and experience, EE, etc. These multiple objectives are usually coupled in a conflicting manner, such that optimization of one objective degrades the other objectives.

One approach to find a tradeoff between objectives is by deriving an explicit expression that can be used as a single objective function. For instance, EE has been jointly optimized with peak rates, WSR, or load balance in heterogeneous networks, see [291], [292], [293], [294] and references therein. Fundamental tradeoffs among EE and delay, sum rate, bandwidth and deployment cost have been derived in [295]. Balancing fairness and spectral efficiency is a well known problem strived for wireless communications [276], and several works have formulated its joint optimization as a single objective function, e.g., [296], [297], [298]. To illustrate several conflicting objectives, consider that allocating resources to users with strong channels can satisfy QoS requirements and improve EE. However, it might also incur in unfair resource distribution across users and unbalanced load among BSs. Now assume that all users have equally good channels, and transmitting low traffic loads to all users increases the coverage area, but this might be very energy inefficient. Unconstrained EE maximization may result in operating points with low spectral efficiency per user [299].

Another approach for the joint optimization of multiple objectives, is to sample the solution space and chose the operative point that satisfies a predefined tradeoff. A mathematical framework to address multi-objective optimization problems for wireless communications has been proposed in [54], [300]. The solution space of such kind of mathematical problems, generally, does not have a unique point that can optimally satisfy all objectives. The main challenges are to characterize and understand efficient operating points within the solution space, so that the objectives are balanced. Authors in [300] provide an example of this approach by jointly optimizing individual peak rates, average area rates, and EE for a massive MIMO setting. Numerical results illustrate the conflicting nature of the objectives: the average area rates increases with the number of served users; the individual peak rates can be increased when the power is split among few users; and high EE is attained if the rate per user is small.

Techniques such as multi-objective optimization [300] and metaheuristic optimization [cf. Section VII-B], sample the solution space to find close-to-optimal solutions. However, the former is constructed from a mathematical framework, which characterizes the solution space, in particular, the desirable operative points. We foresee that multi-objective optimiza-

[16] For systems with queue or delay requirements, a channel/QoS-aware scheduling algorithm must be supported by admission control mechanisms in order to guarantee feasibility [16].

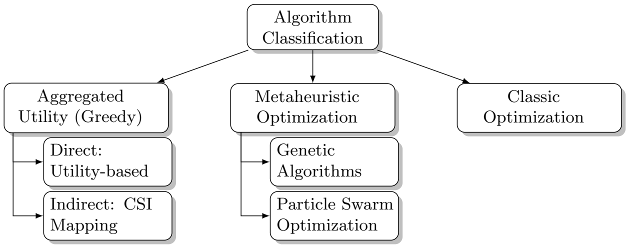


Fig. 8. Classification of the User Scheduling Algorithms for MU-MIMO

tion techniques will play a primal role towards 5G, as the overall network optimization become more complex. Those techniques can provide guidelines for network design, simplify parametrization, and assess compounded optimization criteria for heterogeneous networks. The network designer must establish resource allocation strategies to conciliate conflicted objectives, keeping in mind that the ultimate goals are user experience, satisfaction, and operators' interest [301].

VII. MU-MIMO Scheduling

Resource allocation can be performed over orthogonal radio resources [4], [16], where the allocation decisions are made using individual estimates of BER, SINR, or rate. Scheduling the best user per resource is performed by low-complexity algorithms that do not depend on full CSIT. For instance, a sorting-based algorithm at the transmitter can schedule users based on the CQI. In contrast, the scheduling algorithms for MU-MIMO systems attempt to allocate resources in a non-orthogonal fashion. The users cannot compute their achievable BER or rates since those metrics depend on the CSIT, powers, and precoders assigned to other co-scheduled users.

To illustrate the scheduling in problem (1), consider a MU-MIMO system with a single transmitter equipped with M antennas, and K competing users, each equipped with N antennas. Assuming that linear precoding is implemented [cf. Section IV-B], the maximum number of co-scheduled users per resource is given by $K_{\max} = \lfloor M/N \rfloor$. Define $\hat{\mathcal{K}}^{(ss)} = \{\mathcal{K} \subseteq \{1, 2, \ldots, K\} : 1 \leq |\mathcal{K}| \leq K_{\max}\}$ as the *search space* set containing all possible user groups with limited cardinality.[17] The optimal set, $\mathcal{K}^\star$, that solves (1) lies in a search space of size $|\hat{\mathcal{K}}^{(ss)}| = \sum_{m=1}^{K_{\max}} \binom{K}{m}$. The set $\mathcal{K}^\star$ can be found by brute-force exhaustive search over $\hat{\mathcal{K}}^{(ss)}$, which is computationally prohibitive if $K \gg K_{\max}$. The search space, and in turn the complexity of the problem, varies according to the system model and constraints: the values of M and N, the number of streams per user, individual QoS or power constraints, the set of MCSs, etc.

By plugging the utility function $U(\cdot)$, [cf. Section VI], into (1), we obtain a mathematical formulation of the MU-MIMO resource allocation problem. The optimal scheduling rule that solves (1) in a feasible and efficient way is still an open problem. A number of algorithms have been proposed in the literature to circumvent the high complexity of finding $\mathcal{K}^\star$, and its associated resource allocation. A sub-discipline of optimization theory, known as heuristic search [38], provides solutions (approximation algorithms) to a category of discrete problems. Such problems cannot be solved any other way or whose solutions take a very long time to be computed (e.g. NP-hard problems). User scheduling falls in such a problem category, and the common approach in the literature is to design suboptimal heuristic algorithms that balance complexity and performance. The optimal solution, $\mathcal{K}^\star$, is usually computed for benchmarking in simple MU-MIMO scenarios.

In this section, we provide a classification of the most common scheduling algorithms in the literature of MU-MIMO. Table IV and Fig. 8 show the scheduling algorithms classification, the scenarios in which they are applied, and the methodology they follow. Most algorithms operate at the single resource level, i.e., considering optimization over one sub-carrier or time-slot. Extensions to scenarios with multiple carriers or codes are straightforward, e.g., [39], [41], [73], [85], [92], [101], [109], [111], [114], [115], [116], [126], [147], [155], [171].

A. Aggregated Utility Based Selection

Let us reformulate (1) as a multiuser WSR maximization problem as follows

$$\underbrace{\max_{\mathcal{K} \subseteq \hat{\mathcal{K}}: |\mathcal{K}| \leq c_p} \underbrace{\max_{\mathbf{W}, \mathbf{P}, \pi} \sum_{k \in \mathcal{K}} \omega_k \log_2 (1 + \mathrm{SINR}_k)}_{\text{Inner problem: } R(\mathcal{K})}}_{\text{Outer problem}} \tag{7}$$

where $\mathbf{W}$ and $\mathbf{P}$ summarize the precoding weights and the powers assigned to the users in $\mathcal{K}$, and π defines an encoding order. The formulation in (7) illustrates the fact that the resource allocation problem can be decomposed into different subproblems, which are related to different network layers. Furthermore, under specific system settings and constraints the subproblems can be decoupled.

The weights ω_k, $\forall k \in \mathcal{K}$, can be defined according to the service or application delivered to the users, [cf. Section VI-B]. The inner problem in (7) attempts to maximize the WSR for a given set $\mathcal{K}$. Finding a solution requires joint admission control, precoding design, and power control, e.g. [77], [132], [212], [224]. A solution to the inner problem exists when the multiuser channel associated with $\mathcal{K}$, has specific spatial dimension defined by the precoding scheme. The channel dimensions must be guaranteed by the solution of the outer problem. If linear precoding schemes are implemented, [cf. Section IV-B], the original problem can be relaxed: the cardinality of $\mathcal{K}$ is upper bounded by $K_{\max}$; the encoding order π can be omitted since it does not affect the precoding performance; and power allocation can be computed via convex optimization or by heuristic algorithms [cf. Section IX]. Finding a solution to the combinatorial outer problem implies that the scheduler accomplishes several goals:

[17] $\hat{\mathcal{K}}^{(ss)}$ contains the sets of cardinality one, i.e., SU-MIMO configuration. Efficient scheduling rules must dynamically switch between SU- and MU-MIMO configurations [118], [119], [142], [147], [190], [214].

TABLE IV
USER SCHEDULING METHODS FOR MISO ($N = 1$) AND MIMO ($N > 1$) CONFIGURATIONS, AND FULL ($B = \infty$) AND PARTIAL ($B < \infty$) CSIT

Method	Utility-based	CSI-Mapping	Metaheuristic (stochastic)	Classic Optimization	Exhaustive Search
MISO, $B = \infty$	[41], [47], [59], [60], [62], [67], [72], [78], [81], [95], [128], [138], [152]	[33], [41], [61], [62], [63], [64], [65], [66], [68], [69], [71], [72], [73], [74], [75], [80], [83], [84], [86], [89], [91], [96], [130], [131], [134], [139]	[70], [78], [79], [80], [83], [92], [132]	[72], [77], [93], [94], [132]	[78], [82], [84]
MISO, $B < \infty$	[30], [57], [76], [133], [135], [143], [155], [157], [159], [162]	[33], [48], [55], [66], [76], [129], [130], [131], [134], [135], [144], [153], [154], [155], [156], [158], [159], [160], [161], [162], [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], [174], [181], [182]	[57], [129]	-	[173]
MIMO, $B = \infty$	[39], [98], [100], [102], [105], [107], [108], [109], [110], [111], [113], [114], [116], [118], [122], [125], [126]	[97], [98], [99], [100], [102], [103], [104], [106], [112], [113], [114], [115], [116], [117], [119], [120], [121], [122], [125], [126], [127]	[101], [123], [124]	[39], [40]	[18], [39]
MIMO, $B < \infty$	[141], [142], [175], [178], [184]	[31], [104], [147], [175], [176], [177], [179], [180], [183]	-	[142]	-

it exploits MUDiv to maximize multiplexing gains; and it finds the multiuser channel with the best set of spatially compatible users [cf. Definition 5]. Problem (7) might include constraints over several discrete parameter sets, such as $\mathcal{K}$, the encoding orders, MCSs, spatial streams per user, etc. The complexity of the combinatorial problem grows exponentially with each discrete set size [47].

Due to the hardness of problem (7), one common approach is to find suboptimal solutions through heuristic *greedy opportunistic algorithms*. Given a fixed precoding structure and CSIT, the set $\mathcal{K}$ is constructed by a sequence of decisions. Every newly selected user finds a locally maximum for an objective function. The optimized utility $U(\cdot)$ can be defined by the WSR (direct approach) or by a metric of spatial compatibility (indirect approach). The greedy scheduling rule requires low computational complexity and is easy to implement. However, it does not guarantee neither performance nor convergence to the optimal solution [38].

1) Direct Approach: The seminal works [59], [98] proposed a simple construction of $\mathcal{K}$, by iteratively selecting users as long as the aggregated objective function improves its value. In [59], a greedy user selection (GUS) based on sum-rate maximization was proposed. First, the user with the highest rate is selected, and the following selected user will maximize the total sum-rate. This method was improved in [62], using sequential water-filling to eliminate users with zero transmit power after selection and power allocation. The algorithms in [59], [62] might be computational demanding because calculation of the precoders and power allocation is required each iteration. Furthermore, in the case of imperfect CSI, the above algorithms fail to determine the optimal number of selected users. A generic structure of this kind of user selection is presented in Algorithm 1. In this approach, the transmitter must have channel information so that every iteration the precoders $\mathbf{W}_k$, $\forall k \in \mathcal{K}$, are recalculated to solve the inner problem in (7). If linear precoder schemes are implemented, the maximum number of iterations is at most $K_{\max}$. Since the transmitter knows the precoders and the transmit powers, the achievable rates, SINR, or BERs are known each iteration, allowing the scheduler to identify when it is worth adding more users to $\mathcal{K}$.

This approach is particularly effective if there are no constraints on the cardinality of $\mathcal{K}$, or if the network is sparse, i.e., $|\mathcal{K}| \approx M$ [18]. For instance, if ZF-based precoding is used, the multiplexing gain is not maximized at the low SNR regime, i.e., the total number of scheduled users must be strictly less than $K_{\max}$. In the high SNR regime, multiplexing gains can be maximized if full CSIT is available [52], otherwise the system may become interference limited [cf. Definition 3]. Several works have modified the structure of Algorithm 1 to reduce complexity and improve performance, see [62], [64], [78], [87], [102], [118], [119], [120], [122], [125], [131]. The works [59], [60], [62], [87] showed that the solution space can be shrunk at each iteration, reducing the computational complexity. Based on the characteristics of the powers that solve the inner problem in (7), the scheduler defines a reduced set of candidate users, $\mathcal{K}^{temp}$, for the next iteration.

Nonetheless, such methods also exhibit the flaw of not being able to identify the optimal cardinality, $|\mathcal{K}|$. As such, there are redundant users in the selected user subset, i.e., users

Algorithm 1 Utility-Based Scheduling (Direct)

```
1: Set $\mathcal{K}^{temp} = \bar{\mathcal{K}}$, select the user $k$ with the strongest channel, update $\mathcal{K} = \{k\}$, calculate precoders and powers, compute the achievable rate $R(\mathcal{K})$ in (7).
2: Find the user $i \in \mathcal{K}^{temp} \setminus \mathcal{K}$ that maximizes $R(\mathcal{K} + \{i\})$.
3: If $|\mathcal{K} + \{i\}| \leq K_{\max}$ and $R(\mathcal{K} + \{i\}) > R(\mathcal{K})$, then update $\mathcal{K} = \mathcal{K} + \{i\}$, modify $\mathcal{K}^{temp}$, and go back to step 2.
   Otherwise go to step 4.
4: Compute final precoders, powers, and WSR for $\mathcal{K}$.
```

that can be deleted from the selected user subset to yield a performance increase. This is an inherent flaw of any greedy incremental algorithm, due to the non-iterative cumulative user selection procedure. In [60], [67], it is proposed to add the *delete* and *swap* operations as means to tackle the redundant user issue. However, this approach increases the complexity, even as compared to GUS, as it involves matrix inversions, projections, and an iterative procedure. Moreover, its performance is sensitive to CSI inaccuracies due to the increased error in the projection operation using imperfect CSI, and it does not necessarily find the best user subset in the imperfect CSI case.

2) Indirect Approach: In this approach, the inner and outer problems in (7) are decoupled. To illustrate the structure of the resource allocation strategy, let us reformulate (7) as two problems that must be solved sequentially:

$$\begin{cases} \mathcal{K}^{map} = \underset{\mathcal{K}\subseteq\bar{\mathcal{K}}:|\mathcal{K}|\leq K_{\max}}{\arg\max} \; f(\mathbf{H}(\mathcal{K})) & \text{(8a)} \\ \underset{\mathbf{W},\mathbf{P},\pi}{\max} \; \sum_{k\in\mathcal{K}^{map}} \omega_k \log_2\left(1+\text{SINR}_k\right) & \text{(8b)} \end{cases}$$

Equation (8a) describes a combinatorial user grouping problem, where $f(\mathbf{H}(\mathcal{K}))$ is a metric of spatial compatibility, [cf. Section V]. The user grouping is an NP-C problem, whose solution is found via ExS [72]. Observe that finding a solution to problem (8a) does not require the computation of the precoders and powers. Instead, it depends on CSIT, algebraic operations defined by a given mapping function $f(\mathbf{H}(\mathcal{K}))$, and an iterative procedure. It is worth noting that (8a) may include the weights ω_k to cope with QoS or fairness requirements, see [63]. In contrast, provided the set $\mathcal{K}^{map}$, problem (8b) is tractable and can be solved efficiently via convex optimization [24], [52], [78], [217].

Early works on MU-MIMO systems (e.g. [52], [61], [84]) pointed out that under certain conditions of the weights, ω_k, and SNR regime, the objective function of the inner problem in (7), is dominated by the geometry of the multiuser channel $\mathbf{H}(\mathcal{K})$. The rational behind the reformulation in (8a)-(8b), is that scheduling a set of spatially compatible users [cf. Definition 5], is crucial to suppress IUI. Refining the power allocation and precoding weights in (8b) requires less computational effort than solving the combinatorial problem (8a). A generic structure of this scheduling algorithms is presented in Algorithm 2.

The complexity of finding a solution to problem (8a) can be simplified under certain conditions. Assuming that ZF-based precoding schemes are used, results in [52], [187] establish that the optimal admitted set of users achieves full multiplexing gains, i.e., $|\mathcal{K}| = K_{\max}$, if the system operates in the high SNR regime. Therefore, the cardinality constraint in (8a) is given by $|\mathcal{K}| = K_{\max}$, which shrinks the search space size. A common approach in the literature is to assume that $K_{\max}$ users can be co-scheduled when solving problem (8a), and refining the set $\mathcal{K}$ when solving problem (8b).

The majority of the works that have proposed algorithms to solve problem (8a), e.g. [61], [63], [64], [65], [69], [72], [97], [115], [129], use ZF precoding for the second problem (8b). Consequently, the NSP is the metric of spatial compatibility that maximizes the sum-rate for the set of selected users $\mathcal{K}$, [cf. Section V-A]. A considerable amount of research has been focused on reducing the complexity of the NSP computation, e.g., by reusing previous NSP calculations at each iteration, implementing efficient multiuser channel decompositions, or computing approximations of the NSP. The methods highly depend on M, N, the CSIT accuracy, and the system model.

Different works have proposed a search space reduction per iteration, recalculating and reducing the number of competing users so that spatial compatibility is preserved. A common approach is to preselect a group of candidate users, $\mathcal{K}^{temp}$ in Algorithm 2, based on spatial clustering [cf. Section V-B]. The candidate users at the next iteration will exhibit channel directions fulfilling an ϵ-orthogonality criterion. This approach, coined as semi-orthogonal user selection (SUS), was originally proposed by Yoo and Goldsmith [63], [74]. Several variations of this technique can be found in the literature, see [41], [48], [62], [68], [69], [121], [122], [139], [166], [181]. Once that $\mathcal{K}$ has been constructed, it may be necessary to modify it, so that problem (8b) yields the maximum WSR. A common approach to refine $\mathcal{K}$, is to apply user removal techniques [cf. Section VI-B]. This might be necessary if: the selected channels lacks spatial compatibility; the system operates in extreme SNR regimes; or if the constraints related to $\mathcal{K}$ turn problem (8b) infeasible [72], [125]. If linear precoding and water-filling are used to solve (8b), the set $\mathcal{K}$ can be refined by dropping users that do not achieve a target rate or whose allocated power is zero.

Authors in [114] pointed out that regardless of the number of deployed antennas, it is more efficient to schedule users so that $|\mathcal{K}| < K_{\max}$. This means that the scheduler must seek a tradeoff between maximizing spatial multiplexing gains, and optimizing the WSR for a small set of spatially compatible users. To solve this cardinality problem, some algorithms, e.g. [72], [100], [114], [126], maximize the WSR by comparing the solution of (8b) for multiple sets that solve (8a). This is performed by sequentially constructing groups of spatially compatible users with different cardinalities, and then selecting the best group based on the achievable WSR.

For the general heterogeneous MU-MIMO scenario, $M \geq N_k > 1$, $\forall k$, the function $f(\mathbf{H}(\mathcal{K}))$ in problem (8a) can be defined by a metric of subspace compatibility [cf. Section V-C]. The user grouping procedure using such a metric can operate in a greedy fashion, as in Algorithm 2, but every scheduler design depends on the system model. It is worth pointing out that in heterogeneous systems, maximizing spatial

Algorithm 2 CSI-Mapping-Based Scheduling (Indirect)

```
1: Set $\mathcal{K}^{temp} = \bar{\mathcal{K}}$, select the user $k$ with the strongest channel, update $\mathcal{K} = \{k\}$.
2: Find the user $i \in \mathcal{K}^{temp} \setminus \mathcal{K}$ that maximizes $f(\mathbf{H}(\mathcal{K} + \{i\}))$.
3: If $|\mathcal{K} + \{i\}| \leq K_{\max}$, then update $\mathcal{K} = \mathcal{K} + \{i\}$, preselect users in $\mathcal{K}^{temp}$, and go back to step 2.
   Otherwise go to step 4.
4: Perform operations to refine $\mathcal{K}$.
5: Compute precoders and powers to solve (8b) for $\mathcal{K}$.
```

compatibility is not enough to guarantee WSR maximization. Multidimensional metrics measure the potentially achievable spatial separability but not the effective channel gain per spatial dimension, [cf. Definition 4]. Efficient user selection algorithms in scenarios with heterogeneous users should not only consider spatial compatibility, but also an estimation of the achievable WSR, SINRs, or effective channel gains [120]. Moreover, dynamic allocation of the number of data streams per user must be included in the scheduler, to take into account CSIT accuracy and fulfill QoS constraints. Users with small signal spaces experience less IUI, and results in [302] show that under certain conditions, transforming a MIMO channel into an equivalent MISO channel can improve spatial separability at the expense of multiplexing gain.

The reliability of scheduling based on metrics of spatial compatibility is highly sensitive to CSIT accuracy. Most of the works in the literature (e.g. [97], [98], [99], [100], [102]), perform scheduling by evaluating a form of spatial correlation between MIMO channels of different user. However, the correlation between antennas at each receiver is usually not taken into account. Such a correlation may affect the achievable SINR, depending on the signal processing at the receiver, e.g., receive ZF processing [99], or receive combining [178], [302].

The inner correlation of the channel $\mathbf{H}_k$ of the k-th user can be measured by its rank or by the magnitude (dispersion) of its eigenvalues [115]. Transmitting multiple spatial streams per user cannot be reliable, specially for high inner correlation. A common approach to simplify the scheduling designs is by limiting the number of streams per user, e.g., using only the strongest eigenvector per selected user [97], [99], [302]. For each channel $\mathbf{H}_k$, there exists a dominant spatial direction that can be selected for transmission without creating severe IUI, or performance degradation after precoding. In this way, problem (8a) can be solved based on antenna or eigendirection selection, e.g., [97], [99], [100], [113], [118], [121], [125].

B. Metaheuristic Algorithms

Several works in the literature address problem (7) using stochastic optimization methods, which find close to optimal solutions to complex and non-convex mixed problems. These non-traditional or metaheuristic methods are an alternative to classical programming techniques [38], [303]. The mathematical proof of convergence to the optimal solution cannot be demonstrated, and analytical results or performance analysis cannot be derived. However, these methods may find a close to optimal solution with high probability, without the need of computing derivatives or satisfying convexity requirements (as in standard optimization techniques). The principle behind metaheuristic methods is the following: a set of feasible solutions is iteratively combined and modified until some convergence criterion is met. The term stochastic comes from the fact that initial conditions are generally chosen randomly. For each iteration, the internal parameters change according to dynamic probability functions or even randomly. The algorithms designed under this approach are based on certain characteristics and behavior of natural phenomena, e.g., swarm of insects, genetics, biological or molecular systems [303], [304]. In the context of MU-MIMO systems, two techniques have been used to solve the user scheduling problem: genetic algorithms (GA) and particle swarm optimization (PSO) algorithms.

GA are bio-inspired or evolutionary algorithms, that can solve multiple objective optimization problems with many mixed continuous-discrete variables, and poorly behaved non-convex solution spaces. Unlikely standard optimization methods that rely on a single starting point (e.g. algorithms sensitive to initial conditions [87]), GA start with a set of points (population). These points increase the argument domain of the optimized function and avoid local optimal problems due to the evaluation of the solution space [303]. GA encode potential solutions to the optimization problem in data structures called chromosomes. Due to the binary nature of the variables ξ_k in problem (1), the chromosomes can be represented as binary strings containing valid configurations of the variables ξ_k, [cf. (1c)]. A set of chromosomes is referred to as a population, and at each iteration, the chromosomes combine, exchange critical information, mutate, and eventually evolve toward the optimal solution [78]. Practical application of GA to solve the WSR maximization problem (7) have been proposed for different precoding schemes: ZFBF [78], [187], ZFDP [123], and DPC [101]. Recall that given a set $\mathcal{K}$, the performance of DPC is sensitive to a given user order π selected out of $|\mathcal{K}|!$ valid orders. The works [61], [115] have proposed heuristic algorithms to define π. Since the chromosome structure contains different optimization variables, it can include not only the set of selected users but can also the order in which the users are encoded for ZFDP [123] and DPC [101].

PSO are behavior-inspired algorithms that mimic the distributed behavior of social organisms (particles). The particles use their own intelligence (based on an individual utility function) and the swarm intelligence (global utility function), to discover good directions that can be followed and verified by the swarm. PSO initially defines the swarm as a set of particles randomly sampled from the search space. Each particle has two associated parameters that define how fast it can move toward the optimal solution: position and velocity. In this approach, each particle wanders around in the search space to collect information about the achievable utility function at different locations. This means that the particle looks for the configuration that steers toward the global maximum or the average direction of the swarm. The particles exchange information regarding the best directions, and adjust their positions and velocities accordingly until the swarm converge

Algorithm 3 Metaheuristic Optimization

```
1: For the $i = 1$ iteration, extract $K_{smp}$ elements (chromosomes or particles) from the search space $\hat{\mathcal{K}}^{(smp)}(i) = \{\mathcal{K}_{i,1}, \ldots, \mathcal{K}_{i,K_{smp}}\} \subset \hat{\mathcal{K}}^{(ss)}$.
2: For the $i$-th iteration, evaluate the objective function for each element in $\hat{\mathcal{K}}^{(smp)}(i)$, select the $K_{best}$ elements with the largest objective function, and discard the other bad elements.
3: Perform operations over the remaining elements to generate an improved population/swarm, $\hat{\mathcal{K}}^{(smp)}(i+1)$, for iteration $i+1$.
4: If convergence criterion is met, then compute precoders and powers for the solution set.
   Otherwise go to step 2.
```

to the same solution [303]. Authors in [124] designed PSO algorithms to solve problem (7) using BD precoding. Such an optimization uses the utility function, $U(\cdot)$, of the direct and indirect approaches described in Section VII-A. The PSO approach can define the swarm direction using the function $R(\mathcal{K})$ in the inner problem (7), or the mapping function $f(\mathbf{H}(\mathcal{K}))$ in (8a). This illustrates the fact that the computational complexity of metaheuristic algorithms depends on the optimized objective function.

Algorithm 3 shows a sketch of the steps used to solve problem (7) following the GA or PSO approaches. The constants K_{smp} and K_{best} are arbitrarily defined to limit the overall computational complexity and to speed up convergence. The convergence criterion can be defined by the maximum number of iterations or a performance threshold. The details of the operations required to generate new chromosomes or particles, [Step 3 in Algorithm 3], can be found in the reference aforementioned. The robustness of GA and PSO lies in the fact that the best solutions are systematically combined, and the algorithms have reasonable immunity from getting stuck in local minimums.

Metaheuristic algorithms are efficient techniques for problems where the desired solution does not need to be computed in short time scales. Although bio-inspired algorithms have been used to optimize resource allocation in wireless networks, e.g. [2], [37], [304], [305], they seem to have limited application in practical MU-MIMO systems. This is because the channel conditions, rate demands, and even the number of competing users may change very rapidly over time. Metaheuristic algorithms require centralized processing and the convergence time remains a critical issue [2]. They might suffer from a high computational cost because the objective function is evaluated for each member of the population (swarm). Nevertheless, these methods can provide performance benchmarks to assess other suboptimal, yet practical, resource allocation algorithms.

C. Classical Optimization

Classical optimization techniques [247], [303], have been used for several resource allocation problems in MU-MIMO systems, e.g., WSR maximization, sum power minimization, max-min SINR, and other objective functions, see [1], [54], [224], [273], [306], [307] and references therein. Numerous works have included user scheduling for the optimization of the WSR [39], sum power minimization [40], [77], [190], or some metrics of spatial compatibility [72], [308]. A large number of optimization problems in MU-MIMO systems are non-convex or non-polynomial time solvable, depending on the system models and constraints [224]. Due to the high complexity of the resource allocation problem, the conventional optimization solutions must relax the original problem, approximate non-convex constraints in an iterative fashion, or change the domain of the optimization, sacrificing optimality for the sake of tractability. Nevertheless, by employing a fixed structure for the precoders, the original problem can be simplified, and efficient resource allocation can be performed.

Authors in [39] presented different approaches to solve (1) in a OFDMA MU-MIMO scenario. The mathematical formulation renders the combinatorial problem into a convex problem. By analyzing the entire search space, the proposed algorithms perform channel assignment and signal space selection, i.e., the algorithm defines the scheduled users per carrier and the number of spatial streams per user. Such a reformulation of (1) yields algorithms that decouple the channel assignment in the OFDMA system, i.e., the global optimization is attained by solving problem (7) per sub-carrier. Furthermore, the objective function of the inner problem in (7) is defined in terms of rates and powers, which optimizes the total capacity and power consumption. For classical programming techniques, the set of constraints might turn the problem infeasible. The resource allocation algorithms proposed in [39] identify the assignment sets for which time sharing is the best resource allocation strategy. As discussed in Section VI-B, admission control or user removal techniques can be implemented as a form of scheduling to guarantee feasibility.

Authors in [40] have reformulated the user scheduling problem in multi-carrier scenarios as an extension of problems (8a) and (8b). A close to optimal solution can be found by a sequence of convex and linear programming optimizations. To reduce complexity of the combinatorial problem (8a), the mapping function $f(\mathbf{H}(\mathcal{K}))$ is defined as the magnitude of the MIMO channels, and a close to optimal scheduling sequence is designed based on that metric. According to [40], [190], the generalization of (8b) for OFDMA systems yields a non-convex problem, but quasi-optimal solutions can be attained by applying the classical dual decomposition method. The works [40], [72] have modeled the user scheduling problem in OFDMA systems as a channel assignment problem, which can be solved by efficient algorithms that run in strong polynomial time [309].

A novel reformulation of the WSR maximization problem, (7), as a semi-definite program in [93], sheds some light on non-explicit convex properties of the joint precoding design, power allocation, and users selection problem. The reformulation requires semi-define relaxation, and the solution is found by combining convex optimization and sub-gradient projection methods. Another approach in [94] extends the max-min fair rate allocation problem in [24], [25], [248], [285] so that joint optimization of precoding, power allocation, and user selection was attained. The problem was formulated as the difference of convex functions and solved by the branch and bound (BB) technique, which is a method used to solve mixed-integer programming problems [303]. The approach in [94] yields an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

interference limited systems [cf. Definition 3], i.e., it is allowed that $|\mathcal{K}| \geq M$, which highly increases the search space and computational complexity. Notice that even if the search space is constrained so that $|\mathcal{K}| \leq M$, there are $2^K - 1$ possible user schedules in a MISO configuration that will be enumerated by the BB technique.

As discussed in Section VI, admission control is a form of user selection that may be required to guarantee feasibility. In [77], the joint sum power minimization and maximization of $|\mathcal{K}|$ was formulated as a second-order cone program (SOCP). The set $\mathcal{K}$ is refined iteratively by dropping the user with the largest gap to its target SINR,[18] (i.e. the most infeasible user), and optimal precoders and powers are found by the SOCP optimization. Although the approaches in [77], [93], [94] provide suboptimal solutions due to the relaxation of non-convex problems, they show that joint optimization of the three variable sets $\mathcal{K}$, $\mathbf{P}$, and $\mathbf{W}$ in (7), is feasible through convex optimization.

Other classical optimization methods have been used in the single-carrier MISO scenario with ZF precoding. Quadratic optimization was used in [72] to solve problem (8a), by optimizing a heuristic function $f(\mathbf{H}(\mathcal{K}))$ that linearly combines channel magnitudes and spatial correlation of the MIMO channels. In [308], a heuristic objective function that approximates the NSP has been proposed, and the set $\mathcal{K}$ was found via integer programming. The authors in [176] jointly addressed scheduling and WSR maximization subject to QoS constraints using geometric programming. However, similar to the approach in [39], the optimization in [176] iterates over all solutions of the search space, which limits its application for practical scenarios.

Other works model the resource allocation problem (8a), as mathematical problems that have been extensively studied [309], [310]: the minimum set cover problem from graph theory (e.g., [41], [74]); and the sum assignment problem (e.g., [127], [137], [311], [312]). The modeling used in these approaches requires the calculation of weights or costs values associated with every possible subset in the solution space, i.e., $\{\forall \mathcal{K} \subseteq \bar{\mathcal{K}} : |\mathcal{K}| \leq c_p\}$, where c_p depends on the system parameters. The solutions of such general problems can be found by heuristic algorithms. Due to the exponentially increasing complexity on M, N, and K, the algorithms based on classical optimization are suitable for offline implementation, benchmarking, and to assess other suboptimal heuristic resource allocation strategies for relatively small values of K [224], [273].

D. Scheduling in Multi-Cell Scenarios

In Section III-B, we have described the types of transmission schemes in multi-cellular scenarios. Different levels of coordination are required to improve and maintain the performance of users geographically spread within the coverage area. For users located at the cell edge, the simultaneous transmission from different clustered BSs can boost their throughputs (e.g. CoMP [210]), whereas users close to their BS do not require interaction with adjacent BSs. In full signal-level coordinated systems (e.g., Network MIMO or JT), the global user data and CSI knowledge enables a CU to perform CS using the approaches described in Section VII-A and Section VIII. Assuming partial coordination, the CS/CBF can offer flexibility, scalability, and viability, since centralized or distributed algorithms can be implemented with limited signaling overhead and using local CSIT. The minimum level of cooperation between transmitters can be attained through orthogonal resource allocation. For instance, in LTE systems a technique known as *dynamic point blanking*, prevents transmission at a certain time-frequency resource to reduce interference over such a resource used at a neighboring transmission point [209], [283].

The optimization of CS and CBF can be done based on the methodologies described in Section VII-A, i.e., the two problems can be addressed either independently or jointly. If CS and CBF are decoupled, CS can be tackled by processing shared information related to statistical CSI or performance metrics extracted from local CSI [cf. Section V]. The CBF optimization can be solved by optimal or heuristic techniques for precoding design and power allocation, see [54] for a comprehensive review. A joint optimization of CS and CBF requires iterative calculation of the precoders and selected users, since both optimization spaces are coupled [17], [93]. Efficient scheduling rules must determine the set of users that maximizes the performance, and generates less interference to neighboring cells. The following classification presents some existing methods to implement CS and CBF in multi-cell networks [17], [313].

1) Joint CS-CBF: In full coordinated multi-cellular systems, a cluster of BSs defines a super-cell or virtual distributed antenna system with per-BS power constraints. Full coordination allows different transmission strategies: *i*) all clustered BSs send data to all the users in $\mathcal{K}$; *ii*) some BSs jointly serve a subset of $\mathcal{K}$; *iii*) each BS serves a set of users associated to it; or *iv*) each BS performs single transmission (ST) and serves only one user per scheduling interval, which is known as the interference channel model (IFC), see [54], [222], [223], [268]. The fist and second strategies are exclusive of fully coordinated systems since payload data is shared among BSs. The other strategies are also implemented for distributed CS/CBF. Recent works have extended these transmission schemes to heterogeneous networks, e.g., [92], [93], [111].

Accounting for global CSIT provides flexibility to the CS/CBF design, and several ICI cancellation techniques can be applied. Using classical optimization, the authors in [93] solved the joint CS/CBF optimization, formulated as an extension of problem (7) for heterogeneous networks. The authors in [90], [107] addressed the CS by iteratively evaluating the precoders and powers, following the methodology described in Section VII-A1. The approach in [110] tackled the CBF problem using interference alignment [261], whereas the CS was sequentially solved in a greedy fashion. A dynamic transmission switching between ST and JT was presented in [92], where a centralized CS exploited individual rate statistics (CDF scheduling). In [210], problem (8a) was solved using metrics of spatial compatibility (NSP-based user selection),

[18] See [286] for optimal and suboptimal user removal algorithms for QoS and power constrained systems.

and a centralized resource allocation solved problem (8b), [cf. Section VII-A2]. Practical CS schemes were proposed in [111] for heterogeneous cellular systems, where macro and small cells coordinate scheduling decisions using an the approach described in Section VII-A1. The CBF was optimized using linear precoding and dynamic SU/MU-MIMO switching [147].

2) CS with cyclic CBF: Given a pool of precoding weights, each BS picks a different subset for transmission every scheduling interval, and switches them periodically, in a temporal beam-reuse fashion [179], [180]. This method requires minimum coordination and uses information regarding the allocated beams per BS and the interference environment. This allows practical scheduling and mitigates the flashlight effect (changes of the active beams at neighboring transmitters). Another approach in [137], performs user scheduling by solving a series of assignment problems heuristically, and the precoding weights are optimized every iteration.

3) Sequential CS-CBF: The clustered BSs jointly select users and assign resources in a sequential fashion. The first BS selects its users and broadcast its decision, then the second BS selects its user based on the decision made by the first BS and so on, see [108], [137], [152], [312], [314]. The user selection per BS can be done using the approaches presented in Section VII-A, and the scheduling order of the cells can be assigned according to interference levels or using the round robin approach. The CS can be performed based on interference constraints as in [88]. The BSs schedule their users and perform resource allocation sequentially, so that the ICI generated to previously selected users is below a threshold. If the interference cannot be mitigated as desired, the interfering BSs remain silent during the scheduling interval, which is a low-complexity dynamic clustering strategy.

4) Decoupled CS-CBF: A distributed optimization policy interconnects clustered BSs through a CU or master BS, limiting the message exchange to local CSI and scheduling control signaling [54]. Problem (7) has three optimization variable sets, namely, $\mathcal{K}$, $\mathbf{W}$, and $\mathbf{P}$, and their joint optimization is difficult to solve optimally and distributively. Still, authors in [27] proposed an iterative approach that decouples and optimizes some optimization variables, while the others remain fixed. Yu *et.al* in [109] have followed such an approach and developed a strategy to tackle (7) in a semi-distributed manner, demanding limited communication between BSs and the CU: fix the first two set of variables and optimize the third set, then fix the second and third sets and optimize the first one, and so on until convergence. The BSs jointly update parameters associated with all variable sets in an semi-distributed fashion. By fixing the set of users and precoder weights, the power allocation can be performed based on the information shared between BSs, see [1], [27], [132]. By fixing the precoders and powers, the user scheduling can be performed using the approaches described in Sections VII-A. By fixing the set of users and power allocation, the precoder weights can be computed using distributed algorithms, see [54]. A semi-distributed CS was proposed in [89], where the users are selected using an approximation of the NSP [cf. Section V-A], and the CBF is computed with local CSI. A similar approach was proposed in [91], where the BSs dedicate spatial DoF to cancel ICI for some selected users at neighboring cells. This approach performs semi-distributed spatial clustering for CS [cf. Section V-B], and distributed linear precoding for CBF.

E. Scheduling for Massive MIMO

Massive or large-scale MIMO has been widely envisaged as a key transmission technology for next generation of wireless communication, 5G [13], [14]. In massive MIMO systems, the BS is equipped with a large number of antennas (few hundreds) and serve multiple users (normally few tens). The excess amount of antennas enables focusing the transmission and reception of signal energy into smaller regions of space. Joint optimization over the spatial and multi-user domains can provide significant gains in throughput and EE [13], [49], [315], [316].

In conventional MU-MIMO systems, the number of co-scheduled users K is usually larger than the number of BS antennas M. In contrast, in massive MIMO systems the transmit antennas outnumber the active users ($M \gg K$), which reduces the signal processing complexity and achieves large peak rates. However, the promising performance improvements come at the expense of hardware complexity. Due to cost and power consumption, practical transceiver architectures have a different number of RF chains (M_{RF}) compared to the number of antennas M, [cf. Section IV-G]. If every antenna has its own RF chain, i.e. $M_{RF} = M$, digital beamforming, [cf. Section IV-B], can allocate the whole spectrum to each active user [317]. In practice, the total number of simultaneous users is constrained by the number of RF chains at the base stations [95], [194], [196]. When the number of antennas is larger than the number of RF chains and users, i.e., $M \gg K \geq M_{RF}$, the system performance can be boosted from the diverse path losses and shadow fading conditions of different users.

The channel hardening effect in massive MIMO removes frequency selectivity, and avoids complex scheduling and power control designs [317]. As $M \to \infty$, the MIMO channels become spatially uncorrelated, user separability (in the sense of Definition 5) plays a minor role in the scheduling decisions and the performance optimization relies on the channel magnitudes, see [83], [96], [148], [318]. If sum-rate is the ultimate metric to be optimized, a highly efficient scheduler only needs to assign resources to the users with the largest channel magnitudes.

Most of the scheduling designs for massive MIMO are based on the greedy approaches presented in Section VII-A, e.g. [90], [95], [129], [130], [138], [139], [188]. In these works is common to assume clusters of users sharing similar slow fading characteristics. Approaches such as the *Joint Spatial Division and Multiplexing* (JSDM) [129], [130] first partition cell users into groups with distinguishable linear subspace spanned by the dominant eigenvectors of the groups channel covariance matrix. The transmit beamforming design is performed in two stages: a pre-beamforming that separates groups by filtering the dominant eigenvectors of each groups channel covariance matrix, followed by precoding for separating the users within a group based on the effective channel.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

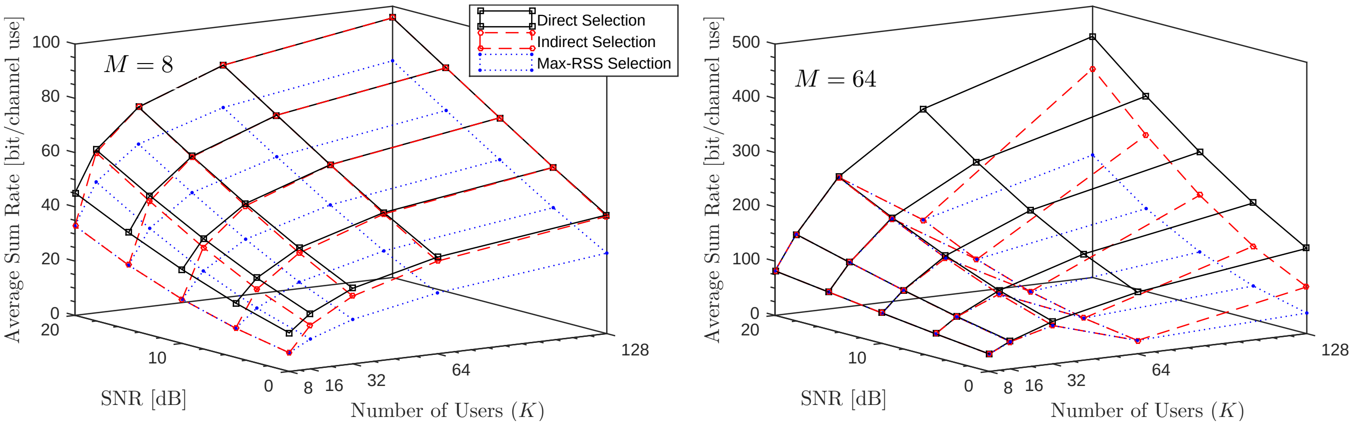


Fig. 9. Average sum rate vs SNR vs the number of users (K), for $M = 8$ (left plot) and $M = 64$ (right plot).

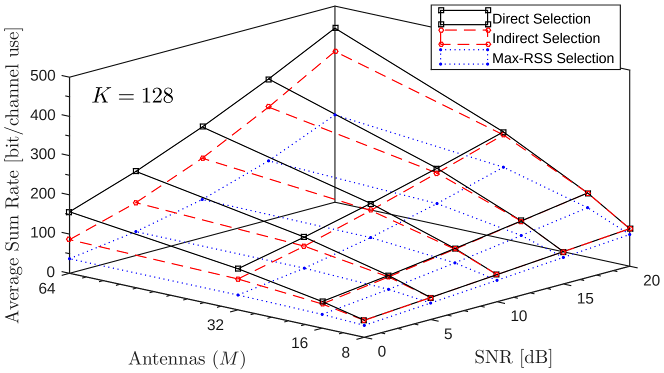


Fig. 10. Average sum rate vs number of antennas (M) vs SNR, and fixed number of active users, $K = 128$.

The user clustering in [129], [130] can be implemented using spatial subspace compatibility metrics [cf. Section V-C]. It is worth noticing that such a grouping process might be necessary to eliminate pilot contamination, by allocating identical pilot sequences to groups of users with similar channel characteristics [50]. The overlap between user clusters and their sizes are features dictated by two factors, the number of RF chains, M_{RF}, and the number of geographically co-located uses [196]. The principle of the two-stage beamforming technique used by JSDM can be applied to multi-cell systems, where each cluster of users is associated to one BS, and the outer precoders are used to cancel ICI [128]. Note that widely used scheduling algorithms in small-scale MU-MIMO, e.g. SUS [63], result in extremely high complexity (of the order of $O(M^3K)$) in the large antenna regime [139], [318]. In contrast, low-complexity scheduling algorithms may attain acceptable performance with full digital beamforming in certain scenarios [96], [318], [319]. However, further research must be done to design efficient scheduling rules for heterogeneous networks using hybrid precoding [cf. Section IV-G].

A process closely related to scheduling is user association, whose mathematical formulation resembles problem (1). Matching a user to a particular transmitter, so that the association optimizes an objective function [cf. Section VI], is a complex combinatorial assignment problem [294]. Ongoing research on this topic have covered cellular (e.g. [307]) and WLAN (e.g. [306]) systems, but literature on heterogeneous networks is still limited [294]. Heuristic and simple association rules employed in current cellular networks, e.g. the max-RSS (received signal strength) or the biased-received-power based criteria, cannot include all system constraints neither fully exploit massive MIMO. In heterogeneous networks, one must consider per-base-station load and power constraints, per-user QoS constraints, and different number of antennas per transmitter. Construction of efficient rules for user association for future heterogeneous 5G networks is an emerging research topic, which must consider channel conditions, load balancing, and EE [294]. Recent results in [292] show that user association to a single BS is optimal most of the times, which can simplify the association rules and algorithms. However, in dense scenarios where the coverage range per transmitter is limited (e.g., due to power constraints), multiple user-base station associations can reduce handover and waste of resources in the backhaul network [194], [196].

F. Discussion and Future Directions

To analyze the performance of the scheduling schemes described in Section VII-A, we consider uniformly distributed users deployed within a single cell with radius 250 m. The users have heterogeneous channel conditions, the path loss exponent is 3.5, and the log-normal shadow-fading has 8 dB in standard deviation. For the direct selection [cf. Algorithm 1], the utility function is the sum rate, in the indirect selection [cf. Algorithm 2], the mapping function is the NSP, and the Max-RSS selects the users with larger channel gains. It is assumed full CSIT at the BS, the precoding scheme is full digital ZFBF (i.e. $M_{RF} = M$), and water-filling is used for power allocation.

The average achievable sum rate is shown in Fig. 9, for a MISO configuration with $N = 1$, $M \in \{8, 64\}$, a cell edge SNR in the range $[0, 20]$ dB, and several number of users, $K \in [8, 128]$. For $M = 8$ (left-plot), we have a conventional MU-MIMO scenario with $K \geq M$, $|\mathcal{K}| \leq M$, and the performance gap between the direct and indirect selection vanishes as K increases. The Max-RSS selection partially exploits the MUDiv by only considering the channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

magnitudes, which results in a considerable performance gap as K grows. For $M = 64$ (right plot), we can divide the users (K) axis in two regions. For $K \leq 32$, the system operates in the massive MIMO region, where there are enough spatial resources to allocate one stream per active user. Notice that as K approaches M, e.g. $K = 32$, the direct selection overcomes the other methods specially in the low SNR regime. For $K = 64$, the indirect and Max-RSS methods attempt to allocate resources to M users, whereas the direct selection optimizes its objective function regardless the attainable multiplexing gain. This result shows that even if M is large, the performance metric is optimized when $|\mathcal{K}| < M$, see [317]. It is worth noticing that for $M = 64$ and $K = 128$, the performance gap between the direct and indirect selection decreases, but the latter approach might require less computational processing. This illustrates the fact that each selection method provides different performance-complexity tradeoffs, depending on the parameters M, N, K, and the SNR regime. Fig. 10 compares the performance of the selection methods with fixed $K = 128$, and different values of M. The performance gap between the direct and indirect selection is small for $M \leq 32$, where the MUDiv is rich. The Max-RSS selection might be an efficient alternative, if M is large and the number of co-scheduled users is small, i.e., $M \gg |\mathcal{K}|$.

Scheduling in massive MIMO systems has not attracted the same attention as in conventional MU-MIMO systems because the excess number of antennas, additional DoF in the scheduling procedure, and the channel hardening effect result in marginal scheduling gains. However, in real-world massive MIMO systems with imperfect CSI, errors in the channel estimation and calibration, there will be some remaining interference among users, especially using linear precoding. The system will inevitably become interference-limited and the sum-rate may saturate (ceiling effect) at high SNR if the CSI and estimation accuracies do not improve accordingly. In this case, it is crucial to select the right number of users to serve, independently of spatial compatibility and channel correlation metrics. Conventional user selection algorithms [cf. Section VII-A], may fail to serve the optimal number of users, even if the selected users have good spatial characteristics. In other words, in massive MIMO systems, it is more important to identify the optimal number of users to serve rather than a set of users with certain channel characteristics [317]. Even random or max-RSS user selection will perform well if they select the right amount of users.

The overhead due to channel estimation is proportional to the number of transmit antennas M. Thus, massive MIMO is more adequate for TDD operation, and its implementation in FDD mode is still an open problem [138], [317]. Although uplink and downlink scheduling are independent tasks with different traffic loads and access techniques [320], the CSI necessary to optimize the downlink transmission is estimated via uplink pilots (channel reciprocity).[19] Thus, further research is necessary to understand the relationship between uplink and downlink scheduling and their associated resource allocation for TDD-based massive MIMO.

VIII. MU-MIMO Scheduling with partial CSIT

In MU-MIMO systems with partial CSIT, resource allocation algorithms can multiplex up to M users accounting for the feedback of one scalar (CQI) and one index (CDI) [31], [cf. Section II-D]. The feedback information load is proportional to the number of deployed users and antennas, but scheduling usually requires rough quantization resolution in the CQI to differentiate between high and low rate users [321]. The CDI plays a more relevant role to achieve spatial multiplexing, thus, it requires a higher quantization granularity. User scheduling requires two main steps: *i*) the transmitter sends pilots for CSI acquisition, the users quantize their channels and feed back the CQI and CDI; *ii*) subsets of users and beams are selected for data transmission based on a particular performance metric [153]. As discussed in Section IV-C, there are two approaches to acquire channel information using codebooks, i.e., quantizing either the channel, or the precoder that better fits the channel. In the following, we classified different scheduling approaches, based on the type of quantization.

A. Scheduling using quantized channels

For the sake of exposition, assume a MISO scenario ($N = 1$), linear precoding, and equal power allocation ($\rho = P_k = P/M, \forall k$). The CQI_k of the k-th user can be given by its SINR defined as [51]:

$$\text{SINR}_k = \frac{\rho \|\mathbf{H}_k\|^2 |\hat{\mathbf{c}}_k \mathbf{W}_k|^2}{1 + \rho \|\mathbf{H}_k\|^2 \sum_{j \neq k} |\hat{\mathbf{c}}_k \mathbf{W}_j|^2}, \quad (9)$$

where $\{\mathbf{W}_i\}_{i=1}^M$, are the precoding vectors extracted from the CDIs (e.g. using the ZFBF). $\hat{\mathbf{c}}_k$ is the actual quantized unit-norm channel or CDI_k given by

$$\hat{\mathbf{c}}_k = \underset{\mathbf{c}_b \in \mathcal{C}}{\arg\max} \quad \cos(\measuredangle(\mathbf{H}_k, \mathbf{c}_b)), \quad (10)$$

where $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_b, \ldots, \mathbf{c}_{2^B}\}$ is a predefined codebook, [cf. Section IV-C]. The k-th user determines its CDI using (10), according to a minimum distance criterion, see [237] for MISO or [175], [236] for MIMO setting. The solution of (10) has the following geometrical interpretation: the user k selects the most co-linear or correlated codeword to its channel, which is equivalent to find the cone containing $\mathbf{H}_k$, as illustrated in Fig. 6. Notice that the SINR_k takes into account channel magnitudes, quantization errors of the CDI, and the spatial compatibility of channel $\mathbf{H}_k$ regarding $\mathcal{C}$. However, to compute the precoders $\{\mathbf{W}_i\}_{i=1}^M$, it is necessary to know a priory $\mathcal{K}$, and the associated CDIs $\{\hat{\mathbf{c}}_i\}_{i=1}^{|\mathcal{K}|}$, which results in a chicken-and-egg problem. Therefore, the exact value of (9) is unknown at the transmitter or receiver sides. Several works have proposed different ways to approximate (9), mainly by its upper/lower bounds or expected value, cf. [155], [156], [158], [159], [162], [170], [177], [182], [322]. If the transmitter has statistical CSI knowledge, the formulation of (9) must take into account the covariance matrices, $\mathbf{\Sigma}_k$, $\forall k$, to have a more accurate estimate of the interference powers [136].

[19] Experimental work for distributed MIMO [216], presented a channel reciprocity protocol that can achieve the performance of explicit channel feedback for mobile users, which is critical for massive MIMO settings.

The CQI can also be given by the channel magnitude, i.e., $\|\mathbf{H}_k\|^2$, but this measure ignores the valuable information contained in the codebook (spatial compatibility), and neglects quantization errors. The number of competing users, the feedback load, and scheduling complexity can be reduced by setting thresholds to the CQIs. This means that only users with strong channels will be considered for scheduling [155], [181]. Multiplexing and MUDiv gains are realized by carefully tuning the threshold according to statistics (e.g. [170]) or numerical characterization (e.g., [158], [166], [181]) of the CQI, for a set of fixed parameters B, M, K, and ϵ, [cf. Section V-B]. In some scenarios where the maximum number of spatial streams per user is at most one, each receiver uses either a weight vector [177] or a combining technique [147], [178], [302] to transform its MIMO channel matrix into an *effective* MISO channel vector.

Different approaches to reduce feedback load and compute the CQI and CDI can be implemented (e.g. [322]). In general, the CQI requires less bits than the CDI, but an optimized bit allocation must take into account the SNR regime, the number of competing users K [158], [170], and practical quantization levels (e.g. MCSs) [157], [cf. Section VIII-C]. Assume that the CQI and CDI are fed back by all competing users, so that the quantized channels can be reconstructed at the transmitter. Resource allocation can be performed using the methods described in Section VII or by low-complexity approaches based on antenna selection, e.g. [89], [171]. ZF precoding is the most common technique used in scenarios with quantized channels, hence, scheduling based on metrics of spatial compatibility [cf. Section V], can be directly applied, e.g., [158], [159], [164], [175]. The user selection for the general MIMO scenario can be performed by computing effective MISO channels or by treating each receive antenna as a virtual user [170]. It is worth mentioning that ZF becomes interference-limited in the high SNR regime with fixed B. However, if K or B scale with the SNR, the achievable rates can be improved and the quantization errors can be mitigated [51], [158].

A common approach to combat IUI (generated by quantization errors), is by reducing the multiplexing gain. Finding the optimal number of data streams is an optimization problem that depends on the system parameters, e.g. SNR regime, mobility, K, B, and M, see [57], [161], [164], [323]. If $K \approx M$, it is likely to have a codebook with limited granularity (fixed B) yielding irreducible IUI. In such a case, the subset of selected users $\mathcal{K}$, must have cardinality strictly less than M, and their CDIs must have high resolution. In that way, the precoding performance can be improved and additional spatial DoF can be used to cancel interference out [162]. The results in [57], [161], [164] show that, the number of selected users highly depends on B, whose optimal value is a function of the SNR regime. For extreme values of the SNR (very low or high), the optimal transmission schemes are TDMA or SU-MIMO, which completely avoid IUI caused by inaccurate CSIT. Experimental results in [8] show that selecting about $\frac{3}{4}$ of the total available beams maximizes performance in different MU-MIMO WLAN configurations.

It is worth noting that the SNR regime plays an important role in the scheduling rule design. Authors in [156] suggested the following guidelines to improve performance:

i) In the high SNR regime with fixed codebook size 2^B, the system becomes interference limited [cf. Definition 3]. The scheduling rules should prioritize users whose fed back channels reduce the quantization error, i.e., the CDI and the available spatial DoF defined the attainable performance. Results in [158] show that the error quantization can be mitigated either in the large user regime ($K \to \infty$), or the high resolution regime ($B \to \infty$).

ii) If the variance of the noise and the IUI are comparable, both the CQI and CDI should be taken into account for user selection.

iii) In the low SNR regime, the system is noise limited and the scheduling rules should prioritize the CQI, since the CDI or error quantization play a negligible role.

These guidelines for scheduling design can be applied to MU-MIMO scenarios with full CSIT, and they have been extended to multi-cellular cooperative systems in [89].

B. Scheduling using RBF

Channel information acquisition based on RBF uses a codebook $\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_b, \ldots, \mathbf{f}_{2^B}\}$, to define the precoders and provide high flexibility for scheduling, [cf. Section IV-C]. In contrast to the methods described in Section VII, where the precoders are unknown before user selection, the approach using RBF can efficiently compute a performance metric that includes the effective channel gain and interference due to quantization errors. However, the accuracy of the quantized channel information is limited by the basis (RBF) or bases (PU2RC) that comprise the codebook [31]. In general, user scheduling is performed by joint user selection and precoding allocation, and the overall complexity depends on the parameters M, B, and K. The set $\mathcal{K}$ that solves problem (7), can be found as illustrated in Phase 1 of Fig. 11, which is based on the algorithm proposed in [31] assuming that $K > M$ and $M \geq N$.

- *Training Phase*: The precoding vectors in $\mathcal{F}$ can be generated according to a known distribution and chosen randomly. The transmitter sends pilots on all the spatial beams $\mathbf{f}_m, \forall m \in \{1, \ldots, 2^B\}$, so that all users can estimate their channels. The codebook design can be simplified by defining codewords as the orthonormal basis of $\mathbb{R}^{M\times 1}$, as is [183]. For this codebook design, antenna selection is implicitly performed at the receiver side. This means that each receive antenna is treated as an individual user, which can reduce scheduling complexity and feedback load. Another approach to construct $\mathcal{F}$, is by eigen-codebook design [133], [178]. Each user computes eigen-decomposition of its covariance matrix $\mathbf{\Sigma}_k$, $\forall k$, and feeds back the eigenvector associated to its maximum eigenvalue. This approach requires extra feedback load, but can provide flexibility to the scheduler. The quantization granularity of $\mathcal{F}$ can be enhanced if the codebook comprises multiple bases, as in PU2RC [162].
- *CSI Feedback*: The users compute a CQI metric per beam, and report their largest CQI to the transmitter. The *effective*

SINR of the k-th user in the m-th beam can be defined as [31]:

$$\mathrm{SINR}_{k,m} = \frac{\rho|\mathbf{H}_k\mathbf{f}_m|^2}{1+\rho\sum_{n\neq m}|\mathbf{H}_k\mathbf{f}_n|^2}, \tag{11}$$

and the index of the best precoder is given by

$$i_k = \arg\max_{m\in\mathcal{B}} \ \mathrm{SINR}_{k,m} \tag{12}$$

where $\mathcal{B} \subseteq \{1,\dots,2^B\}$, is an index subset of active codewords defined by the transmitter so that $|\mathcal{B}| \leq M$. The user designates its $\mathrm{CDI}_k = i_k$ and $\mathrm{CQI}_k = Q(\mathrm{SINR}_{k,i_k})$ with $i_k \in \mathcal{B}$, and $Q(\cdot)$ is a quantization function, see [157], [183] and discussion in Section VIII-C. For $N > 1$, evaluating (11) and (12) can be performed in different ways [147]: using receiver combining techniques, e.g. MRC, if one beam is assigned per user; or equalization techniques, e.g. MMSE [22], when multiple beams are assigned per user. The CQI computation is not limited to the achievable SINR or peak rate, e.g., it can be computed from sufficient statistics of the channel conditions and past CQIs [153]. For instance, authors in [172] defined the CQI as an estimation of the minimum power required to achieve a target SINR. In [167], the CQI is defined as a function of the current and previous channel conditions, which is an approach similar to the CDF-based scheduling [324]. This probabilistic method uses the channel or rate distributions[20] to schedule the users that are more likely to achieve high performance [167].

One can preselect the competing users based on a performance threshold, which reduces feedback and scheduling complexity. Accounting for i.i.d. channels allows to simplify the threshold design, since the channel directions follow a uniform distributions [174]. The k-th user feeds back information only if its associated CQI is above a predefined threshold, γ_{th}, a scheme called *selective multiuser diversity*, see [31], [167], [170], [174], [183], [321]. An extension of such a method is the multi-threshold selection, where the scheduling rule also takes into account the MCSs supported over each link, e.g. [30], [157]. If the system optimization involves queue stability constraints, other thresholds based on QSI can be imposed on top of γ_{th}, provided that the users have knowledge of their respective queue lengths [82], [176].

Preselection can be performed in the spatial domain [cf. Section V-B], where the users meet a constrained in the quantization error. The set of competing users can be defined as $\{k \in \bar{\mathcal{K}} : \cos(\measuredangle(\mathbf{H}_k, \mathbf{f}_{i_k})) < \epsilon\}$, for a predefined threshold ϵ, see [167], [174]. In the large user regime ($K \to \infty$) with fixed M and N, the MUDiv[21] scales as $\log(\log(KN))$ [32], which suggests that preselection based on channel statistics, long-term throughput, or even randomly (e.g., [90]), can reduce the feedback load and achieve MUDiv gains.

- *User Selection*: The transmitter must find the set of users that maximizes the performance metric, e.g. WSR. There are several scheduling approaches, e.g., treating each receive antenna as a single user, assigning at most one beam per user, or allocating multiple beams per user. The simplest selection is assigning the m-th beam to user k_m, which is defined as

$$k_m = \arg\max_{k\in\bar{\mathcal{K}}\ :\ \mathrm{CDI}_k=m} \ \mathrm{CQI}_k \tag{13}$$

If the quantization granularity of the function $Q(\cdot)$ is low, it may happen that (13) accepts more than one solution, in which case k_m can be chosen randomly among the best user for beam m [183]. If the elements of $\mathbf{H}_k$ are i.i.d., each selected user cannot achieve maximum SINR for more than two beams on one antenna provided that $K > M$. This is a valuable design guideline that allows to reduce scheduling complexity [31], [176]. Analytical results in [46] show that M must be scaled proportional to $\log(K)$, so that user starvation is avoided. Once that the set $\mathcal{K} = \{k_m\}_{m=1}^M$ has been found, power allocation and link adaptation can be performed.

Owing to the fact that the quantization granularity is fixed and bounded ($B < \infty$), the precoders $\mathbf{f}_m \in \mathcal{F}$ cannot fully separate users in the spatial domain. Hence, IUI is unavoidable and particularly harmful in high SNR. Moreover, in the sparse user regime, $K \approx M$, RBF cannot benefit from MUDiv and performs poorly. Several user scheduling algorithms have been proposed in [325] to handle sparsity and limited MUDiv over mmWave channels. The optimum number of beams that maximizes the WSR is a function of the number of competing users and the SNR regime [153], [238], [326]. Analytical results establish that for RBF with fixed K, and extreme values of the SNR regime, the optimal number of active beams is one[22] [153], [326]. Depending on the channel conditions and the system constraints, user scheduling might yield a dynamic switching between TDMA (time-sharing) and MU-MIMO transmission. This can take place if IUI is very high, or if the individual rate or power constraints are violated [160], [168]. Therefore, efficient operation in an arbitrary SNR requires a reduced number of scheduled users and active beams, i.e., $|\mathcal{B}| < M$. Analytical results in [153] show that $|\mathcal{K}| = M$ can be attained if $K \to \infty$ for a fixed B. In general, the optimal set of selected users is such that $|\mathcal{K}| < M$, [157]. Finding the best subset of users and beams is a combinatorial problem whose optimal solution is found by ExS [153]. One has to enumerate all user combinations, compute their associated performance metrics $U(\cdot)$, and select the subset with maximum $U(\cdot)$, as illustrated in Phase 2 of Fig. 11. One alternative is to formulate an integer programming optimization, as in [238], or to perform greedy selection, see [157] and Section VII-A.

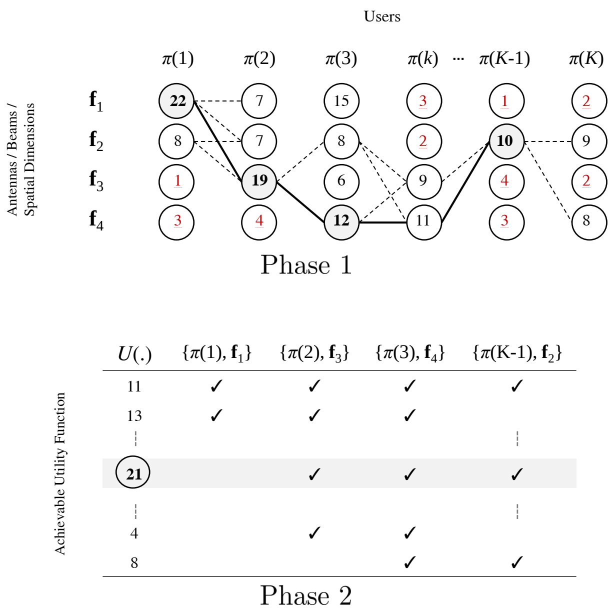
Several extension to the approach presented in Fig. 11 have been developed: improving codebook resolution, e.g. [165], [174]; dynamically updating multi-bases codebooks [169]; computing the CQI by different functions, e.g. [167]; combining RBF with deterministic precoding schemes, e.g. [154]; dynamically adjusting the number of active beams, e.g. [153], [160], etc. To enhanced the quantization granularity, authors in [165] generate several sets of 2^B beams, and the

[20] Observe that the nature of the objective function modifies the statistics of the CQI. For instance, if we want to maximize the sum rate, then users that experience better channel conditions are more likely to be selected [148].

[21] Numerical analysis of capacity versus the number of users K, for both full and partial CSIT, e.g. [63], [66], [81], [98], [100], [102], [104], [110], [120], [156], [165], show diminishing returns of MUDiv over i.i.d. Rayleigh fading channels, i.e., the capacity gain flattens as $K \to \infty$, see [29], [46].

[22] These results are equivalent for the case where channel are quantized in Section VIII-A, see e.g., [57], [161], [164].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials



$U(.)$	$\{\pi(1), \mathbf{f}_1\}$	$\{\pi(2), \mathbf{f}_3\}$	$\{\pi(3), \mathbf{f}_4\}$	$\{\pi(K\text{-}1), \mathbf{f}_2\}$
11	✓	✓	✓	✓
13	✓	✓	✓	
⋮				⋮
21		✓	✓	✓
⋮				⋮
4		✓	✓	
8			✓	✓

Fig. 11. Given the user ordering $\pi(\cdot)$, the user selection is done by finding the strongest user per beam/antenna/spatial dimension associated with $\mathcal{F}$. Phase 1 matches a user with a beam based on the CQI. Phase 2 maximizes the utility function by selecting a subset of users and their associated beams according to some enumeration or search technique.

users report the best CQI and CDI per set. The user selection in [154] combines RBF and SZF for the MISO scenario: given an initial random precoding $\mathbf{f}_1$, its associated user k_1 solves $\max_{k\in\bar{\mathcal{K}}} \cos(\measuredangle(\mathbf{H}_k, \mathbf{f}_1))$, and feeds back a finer CDI that is used to generate the precoder $\mathbf{f}_2$ for the second user, k_2, and so on. In this way, the k-th selected user will only receive interference from the previous $k-1$ selected users. This algorithm fits particularly well in heterogeneous systems, where users can be clustered according to their locations or channel statistics, e.g., [40], [129], [190].

In systems with QoS constraints [cf. Section VI-B], the opportunistic nature of the algorithm in Fig. 11 changes and plays a secondary role, which has a direct impact in the MUDiv gains. This is because the competing users are preselected based on QSI, since scheduling exclusively based on CQIs incurs in delay penalties [46]. In [176], a subset of user is preselected using QSI, and only those users feed back their CSI. The final selection is done as shown in Fig. 11, where the CQIs are computed as the effective SINRs weighted by their respective QSIs. Notice that the parameters of the deployment may affect the latency experienced by the users. For instance, large values of B might result in more scheduled users since IUI can be suppressed more efficiently.

C. Two-stage Feedback Scheduling

Several works in the literature (cf. [18], [153], [160], [163], [167], [170], [171], [173], [182]), have studied the user selection and precoding design problems in terms of the fed back information required to implement them. The total amount of feedback bits per user, B_t, is partitioned such that $B_t = B_1 + B_2$, where each part is used to solve a problem, as illustrated in Fig. 12. Notice that the beam allocation problem in Fig. 11 and the bit allocation problem Fig. 12, bear some resemblance to the methodology described in Section VII-A2. The user selection problem (8a), is solved using some properties of the CSIT of all competing users, and the precoding design and power allocation problem (8b) is solved only for the set of users previously selected or grouped.

The communication phases between the transmitter and the receivers in Fig. 12, are part of a training or pilot mode [44], [153], which is applied to MU-MIMO system with partial CSIT. Taking into account the type of implemented quantization, some remarks are in order [171], [182]:

i) for high-rate feedback systems, a large value of B_t for channel quantization might be more efficient. The accurate CDI can be used for precoding design and to enhance of the multiplexing gain at high SNR [46].
ii) for low-rate feedback systems with small values of B_t, the RBF or PU2RC schemes, [cf. Section VIII-B], are more suitable since they can meet rate constraints and simplify the scheduling.

In scenarios with limited feedback rates, $B_t < \infty$, it is desirable to: *i*) achieve significant MUDiv using B_1 bits for the CQI; *ii*) design codebooks using B_2 bits of resolution for the CDI, to efficiently mitigate quantization errors, IUI, and achieve multiplexing gains; and *iii*) reduce the scheduling complexity and limit the number of users that feedback B_2 bits. The optimal partition of B_t depends on system parameters (N, M, and K), SNR regime, and type of precoding scheme [163], [173]. Moreover, practical systems not only limit the amount of bits per user,[23] B_t, but also the total feedback rate, KB_t, e.g., if the users share the feedback link, see [170], [171], [174], [182]. Observe that the multiuser scheduling in Fig. 12 requires CSI feedback and signaling of the scheduling decisions to the selected users. These feedback and control loop introduce a non-negligible overhead, KB_1, and latency in the system that must be carefully considered in the scheduling rule design [18].

In the first feedback phase in Fig. 12, the competing users send a coarse quantized version of their achievable performance metric. The CQI$_k \in \{q_1, \ldots, q_{2^{B_1}}\}$ is quantized using B_1 bits, where q_i is the i-th quantization level. The set of users in $\mathcal{K}$, can be chosen in a ranking-based per-beam selection, as in Phase 1 of Fig. 11, or per-antenna selection as in [171], [183]. A number of works have shown that depending on the system parameters, choosing $B_1 \in \{1, 2, 3, 4\}$ can be enough to attain the MUDiv gain of the unquantized CQI, see [162], [170], [183]. Results in [171] and references therein, show that efficient precoding design is attained by assigning roughly $1/M$ out of B_t bits to the CQI, and allocating the remaining bits to the CDI. If the system operates in the large user regime ($K \to \infty$), small values of B_1 are enough to achieve asymptotically optimal system performance, see [170] and references therein. MUDiv is defined by users with SINR above a threshold, and constructing $\mathcal{K}$ depends on

[23] In LTE-Advanced, one physical uplink control channel (PUCCH) is assigned per user [209]. The control information such as CQI, CDI, channel rank, scheduling request, and hybrid-ARQ ACK/NAK are conveyed though the PUCCH, which has few bits dedicated to CSI reporting (format 2).

CDIs, since spatially compatible users will be grouped with high probability [31], [176]. In the high resolution regime ($B_t \to \infty$), an efficient partition meets $B_1 > B_2$, and the fraction of bits dedicated for CDI scales proportionally to $\log(B_t)/B_t$, see [171].

If $K \approx M$, larger values of B_1 are required to circumvent the lack of MUDiv and properly identify strong users. However, in the high SNR regime is more beneficial to have $B_1 < B_2$, since quantization errors in the CDI are the main performance-limiting factor [57], [162]. MUDiv clearly affects the optimal partition of B. Numerical results in [44] show the effects of K over the feedback rates, and the interdependence between MUDiv and B_2. Accounting for link adaptation, B_1 must be chosen so that the CQI fully exploits the granularity of the available MCSs, either for channel quantization [177] or RBF [157]. The value of B_1 depends on the type of metric that defines the CQI. For example, the SINR and the achievable rate are real numbers, whereas other metrics might already be defined as integer numbers, see [167].

In the second feedback phase in Fig. 12, each user in $\mathcal{K}$ uses B_2 bits to report its CDI. This mean knowledge of refined channel information at the transmitter, e.g., more accurate spatial directions or effective channel gains. If the system is constrained in the total feedback bandwidth, $|\mathcal{K}|B_2$, heterogeneous and dynamically allocated bits can be used to quantize the CDI. By assigning different number of bits per user, i.e., $B_{2(k)}, \forall k \in \mathcal{K}$, efficiency at the CDI feedback phase can be achieved [182]. The dynamic assignment of $B_{2(k)}$ is more effective in scenarios where the users have heterogeneous average long-term channel gains. The bit allocation rules might consider the following guidelines [171], [182], [327]:

i) for high SNR users, large values of $B_{2(k)}$ reduce quantization errors and mitigate IUI. This is particularly important in scenarios where the CDI is used to compute the precoder weights (e.g., MMSE [168] or ZFBF [170], [171]), [cf. Section VIII-A].
ii) for low SNR users, the performance is noise limited, i.e., the noise variance is larger than the interference, and the value of $B_{2(k)}$ is relatively less important to optimize performance.
iii) dynamic assignment of $B_{2(k)}$ can be extended to sequential transmissions with error correction mechanisms (e.g. ARQ [327]). Such an assignment can be used to further refine $\mathcal{K}$, by temporarily dropping users retransmitting the same packet multiple times.
iv) in multiple-transmitter scenarios, the level of coordination and transmission scheme [cf. Section VII-D], define the methodology to assign the value of $B_{2(k)}$ at each transmitter [171].

Results in [182] suggest that grouping users based on their locations or long-term channel gains (e.g., [40], [129], [190]), yields a more efficient assignment of the feedback bandwidth and simplify the user grouping. Another parameter to be considered when defining B_1 and B_2, is the time used for hand-shaking between feedback phases. The total time used for training and data transmission must be kept below the coherence time of the channel, so that the impact of delayed/outdated CSI is minimized [160], [182]. Otherwise, increasing B_2 would not be enough to achieve MIMO gains [57]. Analytical results in [162] show that for multi-basis codebooks (e.g., PU2RC), large values of B_2 usually do not provide considerable capacity gains, and the overall performance is highly sensitive to MUDiv. A design rule to bear in mind is that the accuracy of the CSIT is more valuable than MUDiv in practical MU-MIMO scenarios [162], [170].

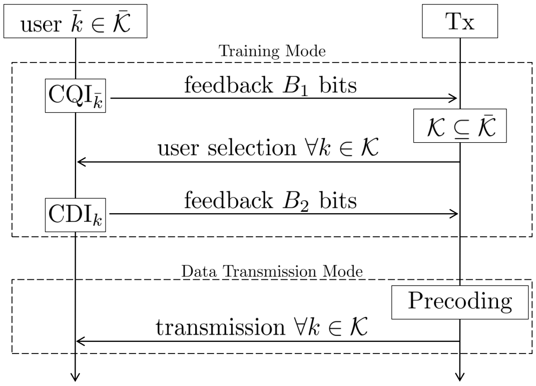

user $\bar{k} \in \bar{\mathcal{K}}$
Tx
Training Mode
$\text{CQI}_{\bar{k}}$
feedback B_1 bits
$\mathcal{K} \subseteq \bar{\mathcal{K}}$
user selection $\forall k \in \mathcal{K}$
CDI_k
feedback B_2 bits
Data Transmission Mode
Precoding
transmission $\forall k \in \mathcal{K}$

Fig. 12. Two-stage feedback MU-MIMO scheduling

D. Implementation in Cellular Scenarios

The scheduling algorithms for MU-MIMO in LTE-Advanced rely on quantized CSI: *i*) the RI/PMI metrics contain information regarding the SU-MIMO channel (due to backward compatibility[24]) projected onto the subspace defined by the codebook; *ii*) the CQI metric indicates the energy of the projection, and it might include ICI and noise [214]. Since the aggregated multiuser channel can be constructed by concatenating the users PMI metrics, and the CQIs are known at the transmitter, the scheduling algorithms described in the subsections above can be directly used in cellular scenarios.

In general, the scheduling algorithms in LTE-Advanced are proprietary and implementation-specific and there is no standardized procedure to define them [214]. It is worth noting that the amount of available CSIT defines the transmission mode (SU- or MU-MIMO), and the scheduling decisions. Practical scheduling rules with switching mode can be defined by evaluating the achievable performance of SU-MIMO and MU-MIMO (with multi-rank transmission) modes, and simply choosing the one with better performance [30], [142], [184]. However, robust switching between these modes still requires research to guarantee adaptability to changes in channel and traffic conditions, as well as viable computational complexity and scalability [17].

Authors in [184], highlight the fact that for known precoding matrices at the transmitter, it is possible to generate look-up-tables (LUTs) that contain information regarding spatial

[24] The limitations due to PMI feedback in SU-MIMO mode can be overcome for certain scenarios by means of efficient estimations of the CDI/CQI. One can take into account spatial compatibility between the MIMO channels and the codebooks, i.e., estimating the effective channel gains [322].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

compatibility and potential interference. Resource allocation based on LUTs could be used to significantly reduce the complexity of the scheduling algorithms. There are several factors that must be considered for CS/CBF design in cellular systems [49], [54], [90], [107], [207], [210]: dynamically determining whether or not coordination is required, switching between JT and CS/CBF, optimizing cluster formation and sizing, and scalability issues. The resource allocation strategies must also include constraints on backhaul bandwidth, CSI acquisition, latency, QoS, mobility, and synchronization.

E. Implementation in WLAN Scenarios

In the IEEE 802.11ac standard, MU-MIMO transmission consists of two tasks [328]: *i*) to identify the users that belong to the transmission set, which is implementation-specific and might depend on priority weights or traffic related parameters [cf. Section VI]; *ii*) the assignment of a Group ID for each set of co-scheduled users, and the downlink transmission signaling. For a given time instance, the AP constructs a Group ID assignment table. The number of rows is defined by the available number of ID groups, and the columns are given by the number of associated users. The condition to have a proper Group ID, is to have different users assigned to each one of the available stream positions. Selecting the proper Group ID and its associated scheduled users is equivalent to a coloring problem over a two dimensional array, whose optimal solution can be found by ExS. A heuristic algorithm was proposed in [328], where the positions of the users in a particular Group ID are determined according to the probability of occurrence. The Group ID selection and success allocation probability can be improved by modeling the discrete searching process as a linear sum assignment problem, for which there exist reliable polynomial-time algorithms [309].

For the general MU-MIMO scenario, a selected user can receive more than one data stream, and all of them must use the same MCS [252]. The authors in [126] proposed a user selection algorithm inspired by [114], originally designed for cellular scenarios. The scheme applied BD with geometric-mean decomposition to guarantee equal MCS allocation over all spatial streams assigned to a single user. That work highlights the need of careful stream allocation per selected user, where the best number of streams is not necessarily equal to rank of the channel. Another approach to guarantee maximum sum rate with joint link adaptation (MCS selection), BD, and user/stream selection was proposed in [141]. The authors extended the scheduling algorithm in [98], originally designed for cellular scenarios, to WLANs with partial CSIT. The link adaptation is performed by a machine learning classifier that provides robustness to CSI inaccuracy. Numerical results suggest that estimation of the IUI is required to allocate MCS more efficiently. The next generation of 802.11 standard will define the conditions and methods for user grouping/association [215]. A critical open problem is to build scheduling algorithms to balance overhead, fairness and individual priorities.

F. Scheduling Complexity

The scheduling complexity can be classified in two types [2]: implementation complexity and computational complexity. The former, refers to the amount of signaling overhead and information exchanged among different network entities. The latter refers to the processing time required to execute a certain algorithm at the transmitter or CU. The implementation complexity is assumed to be reduced whenever the channel information is quantized [cf. Section IV-C]. This type of complexity can be estimated as a function of the system parameters K, B, N, M, and the number of clustered transmitters. Nevertheless, to the best of our knowledge, there is no published work or reference framework that analyze the complexity in a fundamental and general way, at least for the MU-MIMO systems considered here. Most research in the field of limited CSI feedback has focused on the following issues related to complexity [18], [210]: *i*) reducing the number of parameters to be quantized, e.g., bits to quantize the CQI and CDI; *ii*) reducing the required quantized feedback resolution; and *iii*) constraining the signaling and message exchange within a cluster, and minimizing the overall coordination.

In the MU-MIMO literature, most works focus on the computational complexity required to select users, to extract the precoders, and to perform power allocation. There are several metrics to estimate the computational complexity: *i*) number of real or complex operations. *ii*) number of flops, where a flop is defined as a floating point operation [98], e.g., a real addition, multiplication, or division. *iii*) using the big-Oh notation, $O(\cdot)$, which is proportional to the term that dominates the number of elementary operations needed to execute an algorithm [40]. *iv*) other metrics, such as the number of vector/matrix operations, the number of iterations needed for convergence, or results based on simulations and execution time. Therefore, there is no unified approach to characterize the computational complexity for scheduling algorithms, but the aforementioned metrics, summarized in Table V, provide a coarse assessment of the computational load order.

We hasten to say that another sort of complexity analysis can be performed for MU-MIMO systems, in which precoders and powers are jointly optimized for a fixed set of users $\mathcal{K}$, [cf. Section VI-B]. The resource allocation problem only involves $\mathbf{W}$, $\mathbf{P}$, and is formulated as the optimization of a certain system-level utility function, $U(\cdot)$, [cf. Section VI-A], subject to resource budget constraints, see [224] and references therein. In these cases, the set of scheduled users is not optimized, and the complexity is normally defined as a function of M, N, and the characteristics of $U(\cdot)$, e.g., convexity or non-convexity.

G. Discussion

The two approaches in Section VIII-A and VIII-B are used in scenarios where the transmitters have different capabilities and constraints. The former approach is applied when the transmitter can compute linear precoding using the quantized channels [cf. Section IV-B]. The latter approach is used when the precoders are predefined and cannot be modified based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

TABLE V
SUMMARY OF COMPUTATIONAL COMPLEXITY METRICS FOR MU-MIMO SCHEDULING ALGORITHMS

Metric	References
Flop-based	[65], [69], [98], [100], [102], [104], [118], [122], [123], [124]
Real/complex Operations	[59], [60], [62], [67], [69], [71], [72], [99], [113], [155]
$O(\cdot)$	[39], [40], [76], [96], [98], [125], [127], [128], [137], [138], [152], [153], [182]
Other Metrics	[40], [41], [63], [68], [73], [74], [76], [77], [78], [85], [89], [97], [99], [101], [110], [114], [119], [120], [126], [128], [132], [157]

on the instantaneous CSI. The bit allocation for the CQI and CDI, described in Section VIII-C, can be used to analyze both approaches, channel quantization and RBF. Practical resource allocation algorithms work with partial CSI, whose computational complexity depends on the codebook resolution, the number of deployed antennas and active users. On the one hand, high codebook resolution improves peak rates and reduces IUI, but it is a bottleneck for the uplink in FDD mode. On the other hand, dynamic channel-based codebook designs speed up the resource allocation, reduce user-pairing complexity, and mitigate interference [17]. The CQI and CDI reports can be optimized to enhance centralize or distribute resource allocation, and they are also functions of the operative wide band and feedback periodicity [329]. The main issue related to CSI feedback is to find a good trade-off between signaling overhead and accurate channel quality estimation. Enhanced scheduling algorithms for multiuser systems will depend on the type of CSI available at the BSs (e.g. channel quantization or RBF), the deployment configuration, and more complex decision-making rules that can maximize the overall throughput [16].

In practical massive MIMO scenarios, hybrid precoding will be used for MU-MIMO communication, whose performance will depend on the number of RF chains at the transmitter and the accuracy of the fed back CSI. Moreover, the type of transceiver architecture and precoding scheme must be defined according to the objective function (e.g. sum-rate or EE) and the user sparsity [257]. Further research is needed to understand the joint optimization of bit and stream allocation using hybrid precoding, taking into account hardware resolution and channel estimation accuracy [197].

IX. POWER ALLOCATION

The optimal power allocation $\mathbf{P}$ in (7) depends on the CSIT availability, the type of precoding scheme implemented at the transmitter, the individual user priorities, and the objective function. Accounting for full CSIT in single-transmitter scenarios, the optimal power allocation is usually known in analytical closed-form for several performance metrics, e.g., BER, SINR, WSR, or fairness [42], [277]. Assuming that $\mathcal{K}$ is fixed, and linear precoding is implemented at the transmitter, a network-centric power control algorithm finds the optimal $\mathbf{P}$ for WSR maximization through convex optimization, i.e., using the water-filling principle, see [247], [268], [330]. The linear precoding decouples the signals into orthogonal spatial directions mitigating the IUI, and water-filling allocates powers according to the effective channel gains [cf. Definition 4].

In scenarios with a fixed feasible set of users $\mathcal{K}$, [cf. Definition 8], and assuming that $\mathbf{W}$ is coupled with $\mathbf{P}$, the power allocation may involve, for example, a WSR maximization with rate control, sum power minimization with individual SINR constraints, or a max-min SINR problem, see [24], [47], [54], [248]. For this kind of problems, power control algorithms are designed based on optimization theory (see [54], [224]), and the Perron-Frobenius theory[25] for non-negative matrices (see [1], [27]). In contrast to WSR maximization, where water-filling assigns more power to stronger channels, other objective functions require to allocate powers in a different way. For instance, balancing the power over all users so that weak users are assigned more power to improve their error rates [126], assigning powers according to target SINRs [73], [77], [85], [93], [132], [172], or considering queue stability constraints [55]. Some works model $\mathbf{W}$ so that the power is absorbed into it, and optimizing the directions and magnitudes of the precoder weights implicitly performs power control, e.g., [83], [137]. The interested reader is referred to [27] for a comprehensive survey on theory and algorithms for joint optimization of $\mathbf{W}$ and $\mathbf{P}$.

Equal power allocation (EPA) is a suboptimal strategy used to simplify the evaluation of the utility function $U(\cdot)$, especially when the transmitter does not have full CSI knowledge [51]. In certain scenarios, EPA allows a more tractable system performance analysis or derivation of statistics based on K, M, N, or B, which generally yields closed-form expressions. Assuming full CSIT, if the WSR maximization problem [cf. Section VI-A] is optimized using ZF-based or BD-based precoding schemes, the EPA (directly proportional to the individual weights ω_k, $\forall k \in \mathcal{K}$), asymptotically achieves the performance of the optimal water-filling power allocation at high SNR [331]. The authors in [94] discussed the secondary role of power allocation when optimizing problem (7). The paper suggested that more efforts should to be taken to find $\mathcal{K}$ and the corresponding $\mathbf{W}$, which may justify the adoption of EPA under specific SNR conditions.

In systems with partial CSIT, power allocation requires numerical methods that depend on the optimized performance metric [42]. Consider a practical scenario where B is finite and fixed, and $\mathcal{K}$ has been found by one of the methods described in Section VIII. In such a scenario, the system is limited by interference [cf. Definition 3], at the moderate and high SNR regimes, since the interference components in the denominator of (9) and (11) are non-zero. Therefore, one of the main objectives of power control is to perform interference management. However, to compute the optimal $\mathbf{P}$ for WSR maximization, the transmitter must have knowledge of all interference components for all users. This results in

[25] The Perron-Frobenius theory is a fundamental tool to solve congestion control problems and optimize interference limited systems [cf. Definition 3], such as wireless sensor or multi-hop networks [1].

non-convex optimization problems, for which efficient algorithms (depending on the operating SNR regime) can provide suboptimal solutions [160]. Moreover, global knowledge of the interference components at the transmitter may not be attainable in limited feedback systems (e.g. only CQIs are known), and adopting EPA is a common practice in the reviewed MU-MIMO literature.

In multi-transmitter scenarios, the level of coordination, the availability of user data and CSIT, and the precoding scheme determine the best power allocation policy. In full coordinated scenarios, all transmitters belong to the same infrastructure, and a CU determines the power allocation across the cluster. Assuming global CSIT, a fixed set $\mathcal{K}$, and linear precoding schemes, the power allocation that optimizes a global performance metric subject to per-transmitter power constraints can be realized numerically or through water-filling, see [90], [107]. For scenarios with partial coordination between transmitters, assuming that $\mathcal{K}$ is globally known, the power allocation depends on the precoding schemes computed from local CSI and individual priorities ω_k, $\forall k \in \mathcal{K}$, known by each transmitter, see [223], [242], [281].

Table VI summarizes the reviewed scheduling methods and their associated power allocation strategies: EPA, water-filling, and adaptive methods. The references listed under the *adaptive* category depend on the particular objective function and system constraints, see Tables II and III. Those references jointly optimize $\mathbf{P}$ and $\mathbf{W}$ via conventional optimization methods [247] and iterative algorithms.

A. Summary and Future Directions

Power allocation is a dynamic process that compensates instantaneous channel variations to maintain or adjust the peak rates [209]. In MU-MIMO scenarios, power control is required to optimize an objective function, mitigate interference, and satisfy system constraints. Several problem formulations deal with convex objectives and linear constraints, for which water-filling can compute the optimal power allocation. Alternatively, EPA simplifies the optimization by transforming $\mathbf{P}$ into a constant vector, which becomes close-to-optimal for ZF at high SNR. In multi-cell scenarios, power allocation requires exchange of control messages to mitigate ICI. The amount of signaling overhead depends on the type cooperation and coordination between transmitters, and the coordinated signaling (centralized or distributed). Results in [216] show that hardware calibration and distributed power allocation are critical to achieve joint transmission in MU-MIMO settings.

Recently, energy consumption has become a central concern in academia and industry [284], [289]. One of the objectives of 5G is to enhance EE by orders of magnitude [14], tanking into account power consumption at the access (transmit and circuitry power) and core (backhaul) networks. EE depends on the particular network planning and objective function [cf. Section VI]. For instance, in dense heterogeneous networks the distance between transmitters and receivers determine EE [284]. Dense deployments can provide high EE, but the number of transmitter per serving area should be carefully planned to avoid ICI and waste of resources due to handovers [196]. The EE is expressed as a benefit-cost ratio and measured in bits per Joule [284], [299]. Complementary metrics such as Gflops per Watt (to calculate the typical computational efficiency [300]) can be used to assess the power consumption per processing block at the transceivers, or to evaluate the EE of a particular algorithm.

X. Final Remarks

There are many challenges to be solved for resource management in MU-MIMO systems. Identify and standardize the best practices to group and schedule users is fundamental to profit from MUDiv. This is tightly related to affordable-complexity algorithm design to efficiently manage the allocation of multiple users, carriers, time slots, antennas, and transmitters. One of the main requirements to obtain MUDiv and multiplexing gains is knowledge of the CSI. In practical systems, this is challenging because the feedback load rapidly increases with K and M. Multiple transmit antennas require additional pilot overhead proportional to M, if the users need to learn the complete MIMO channel [42]. Further research is required to completely understand performance bounds advance transmission schemes, e.g., dynamic SU-/MU-MIMO switching and multi-/single-stream allocation. The viability of user scheduling and simultaneous transmission in multi-cell deployments must be further investigated. Several topics must be included in future studies: synchronization issues in MU-MIMO, metrics for user grouping in HetNets, scheduling signaling load, coordinated power control, latency and delay due to scheduling, mobility, the impact of realistic traffic models, backhaul constraints, and semi-distributed control algorithms. The scheduling policies for the next generation of wireless technologies must balance between the high cost of CSI acquisition and the benefits of cooperative transmission.

The current performance of MU-MIMO processing is still limited in 4G cellular communication systems. This is because user terminals do not perform interference estimation, and in all cases only CSI feedback for the SU-MIMO mode (SU-RI, PMI, and CQI) is employed [17]. Most of the reviewed works assume continuous power control, but LTE only supports discrete power control in the downlink, through a user-specific data-to-pilot-power offset parameter [209]. These facts must be considered to properly assess current and emerging NOMA schemes in multiuser scenarios, as well as their limitations, see [5], [6], [260]. Resource management rule have a fundamental role to play in future generations of wireless networks, where new and evolving technologies such as mmWave and massive MIMO have already been considered [13], [197], [332], [333]. Such rule will require novel precoding designs and an assertive usage of the multiuser dimension to provide substantial spectral efficiency, energy efficiency and user satisfaction.

A. Heterogeneous Networks

Current cellular networks are facing immense challenges to cope with the ever increasing demand for throughput and coverage owing to the growing amount of mobile traffic. Network densification through deployment of heterogeneous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

TABLE VI
Scheduling Approaches and their associated Power Allocation Methods

Method	EPA	Water-filling	Adaptive
Utility-based	[30], [57], [76], [80], [88], [95], [105], [108], [110], [111], [113], [118], [133], [135], [138], [141], [142], [143], [152], [157], [159], [162], [175], [178], [184]	[36], [39], [59], [60], [62], [67], [90], [95], [98], [100], [102], [118], [122], [125]	[47], [81], [85], [87], [107], [109], [116], [126], [128], [137]
CSI-Mapping	[33], [64], [75], [76], [89], [91], [96], [97], [97], [99], [102], [104], [113], [114], [117], [125], [126], [127], [129], [130], [131], [134], [144], [147], [153], [154], [156], [158], [159], [160], [163], [164], [165], [166], [167], [169], [170], [171], [173], [174], [175], [176], [177], [179], [180], [181], [182], [183]	[41], [48], [61], [63], [65], [66], [69], [71], [72], [74], [84], [86], [98], [99], [103], [106], [114], [115], [117], [119], [120], [121], [139], [160], [168]	[55], [68], [73], [75], [83], [112], [116], [160], [172]
Metaheuristic (stochastic)	[57], [80], [92]	[78], [101], [123], [124]	[70], [79]
Classic Optimization	[72]	[39], [40]	[77], [93], [94], [132]
Exhaustive Search	[82], [173]	[39], [84]	-

infrastructure, e.g., pico BSs and small cells, is envisioned as a promising solution to improve area spectral efficiency and network coverage. Nevertheless, heterogeneous networks and small cell deployments will bring significant challenges and new problems in resource allocation and scheduling for MU-MIMO system. First, network densification reduces MUDiv, since few users are associated to the closest BSs [283]. This can reduce the net throughput and may challenge precoding techniques with imperfect CSI that rely on MUDiv to compensate for the imperfections. Furthermore, the selection metrics may need to change in HetNets as scheduling may be used not for boosting the received signal exploiting MUDiv but for serving users with low ICI. Moreover, putting a large number of antennas in small cells is not envisioned mainly due to excess cost and processing capabilities, thus the promising gains of massive MIMO may not be realized in HetNets. The same applies for cooperative and network MU-MIMO schemes (e.g. CoMP) among small cells, which may require additional signaling and communication among small cells. Access network architectures enhanced with distributed antennas [334] can complement small cells and increase system throughput, but only if large backhaul capacity is available [335]. The existence of signaling interfaces (e.g. X2 in LTE) seem to be sufficient for current deployments, but changes may be required for more advanced coordinated MIMO techniques and joint resource allocation. Network densification will result in major challenges in indoor coverage and services, for which the channel models used in existing MU-MIMO literature may not be relevant.

B. Massive MIMO and Millimeter Waves

Massive MIMO and mmWave technologies have a fundamental role to play in 5G [14], [15], but there are many challenges and open problems. More research is needed to enable FDD mode with acceptable overhead and mitigation of pilot contamination, which will speed up the standardization and commercial adoption of massive MIMO technology. Operating in wideband channels at mmWaves will require efficient hybrid precoding and fast beam adaptation. Low-resolution and cost effective transceiver architectures must be designed to cope with these goals, which will bring new techniques to calibrate the antenna array, define the most efficient array geometry, estimate CSI, and mitigate hardware impairments [216], [336]. Although massive MIMO and mmWave will be implemented in cellular networks [196], currently, they are jointly applied for indoor short-range [197] and outdoor point-to-point applications [194], [337]. Signal processing and medium access techniques for massive MIMO may provide a cost-effective alternative to dense HetNets [283], and full understanding of how these technologies complement each other is matter of future research [299].

Initially, synchronous massive MIMO systems have been implemented with capabilities for joint signal processing [338]. However, coordinated per-user allocation and time-synchronous transmission may be hard to achieve in cellular systems [339]. Therefore, further research is needed to design, prototype, and assess distributed and asynchronous massive MIMO systems, specially in challenging high mobility scenarios. Additionally, the data buses and interfaces at the transmitters must be scaled orders of magnitude to support the traffic generated by many concurrent users. Research on intermittent user activity (bursty traffic) has been presented in [339], but more work is needed to optimize resource allocation for users with heterogeneous data rate requirements and services. These challenges will demand enhanced physical and media access control layer interactions, in order to support fast user/channel tracking and dynamic resource allocation.

C. Summary of Asymptotic Analysis and Scaling Laws

Several authors have focused their work on asymptotic analysis and scaling laws of performance metrics or utility functions $U(\cdot)$. The analytical results depend on the *system parameters*, K, N, M, B, P, and q_k $\forall k \in \mathcal{K}$. The information-theoretic results derived in several works provide fundamental limits of achievable values of $U(\cdot)$ from the user and system perspectives. They also shed light on the relevance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2016.2618870, IEEE Communications Surveys & Tutorials

TABLE VII
ASYMPTOTIC REGIMES IN MULTIUSER SYSTEMS FOR MISO AND MIMO CONFIGURATIONS, WITH FULL ($B = \infty$) AND PARTIAL ($B < \infty$) CSIT

Parameter	**MISO**, $B = \infty$	**MISO**, $B < \infty$	**MIMO**, $B = \infty$	**MIMO**, $B < \infty$
Capacity	[33], [80], [166], [331]	[33], [51], [57], [156], [157], [166], [167], [174], [181], [267], [340]	[32], [46], [331]	[32], [340]
Queues	[73], [82], [187]	[55], [143]	[46]	[176]
High SNR	[52], [331]	[30], [51], [57], [76], [153], [155], [156], [157], [158], [167]	[46], [99], [102], [125], [331], [118], [121], [122]	[42]
Low SNR	[52]	[57], [153], [156], [157], [167]	[99], [102] [121]	[42]
Large K	[29], [62], [63], [64], [66], [70], [74], [90], [91], [139], [166], [112]	[30], [48], [76], [129], [153], [156], [158], [159], [161], [170], [174]	[32], [46], [97], [102], [121], [122]	[177], [340], [31]
Large M	[67], [70], [90], [335]	[48], [129], [130], [131], [161]	[102], [113]	[31]
Bits B_1, B_2	[173]	[51], [55], [57], [76], [153], [157], [158], [161], [162], [163], [166], [167], [171], [181], [267]	-	[177], [178], [183]

TABLE VIII
ABBREVIATIONS

3GPP	3rd Generation Partnership Project	**FDD**	Frequency-division-duplex	**PMI**	Precoding matrix index
ADC	Analog-to-digital converter	**GA**	Genetic algorithms	**QoS**	Quality of service
AP	Access point	**GUS**	Greedy user selection	**QCA**	Quantization cell approximation
ACK/NACK	Acknowledgement handshake	**HetNet**	Heterogeneous network	**QSI**	Queue state information
AWGN	Additive white Gaussian noise	**ICI**	Inter-cell interference	**RF**	Radio frequency
ARQ	Automatic repeat request	**IFC**	Interference channel	**RBF**	Random beamforming
BS	Base station	**IUI**	Inter-user interference	**RI**	Rank indicator
BER	Bit error rate	**JT**	Joint signal transmission/processing	**RSS**	Received signal strength
BD	Block diagonalization	**LoS**	Line-of-sight	**SUS**	Semi-orthogonal user selection
BB	Branch and bound	**LTE**	Long term evolution	**SLNR**	Signal to leakage plus noise ratio
BC	Broadcast channel	**MRT**	Maximum ratio transmission	**SNR**	Signal-to-noise ratio
CU	Central processing unit	**MSE**	Mean-square-error	**ST**	Single transmission
CDI	Channel direction information	**MMSE**	Minimum MSE filter	**SU-MIMO**	Single user MIMO
CQI	Channel quality information	**MCS**	Modulation and coding scheme	**SISO**	Single-input single-output
CSI	Channel state information	**MAC**	Multiple access channel	**SDV**	Singular value decomposition
CDMA	Code division multiple access	**MIMO**	Multiple-input multiple-output	**SDMA**	Space-division multiple access
CIZF	Constructive interference ZF	**MISO**	Multiple-input single-output	**SZF**	Successive ZF
CBF	Coordinated beamforming	**MUDiv**	Multiuser diversity	**TDMA**	Time division multiple access
CoMP	Coordinated multi-point	**MU-MIMO**	Multiuser MIMO	**TDD**	Time-division-duplex
CS	Coordinated scheduling	**NOMA**	Non-orthogonal multiple access	**THP**	Tomlinson-Harashima precoding
CSIR	CSI at the receiver	**NP**	Non-deterministic polynomial time problem	**Wi-Fi**	Trademark of IEEE 802.11
CSIT	CSI at the transmitter	**NP-C**	NP complete	**VP**	Vector perturbation
CDF	Cumulative distribution function	**NSP**	Null space projection	**VQ**	Vector quantization
DoF	Degrees-of-freedom	**OFDMA**	Orthogonal frequency-division multiple access	**WSR**	Weighted sum rate
DPC	Dirty paper coding	**OFDM**	Orthogonal frequency-division multiplexing	**WLAN**	Wireless local area network
EE	Energy efficiency	**PSO**	Particle swarm optimization	**ZF**	Zero forcing
EPA	Equal power allocation	**PU2RC**	Per unitary basis stream user and rate control	**ZFBF**	ZF beamforming
ExS	Exhaustive search	**PHY**	Physical layer	**ZFDP**	ZF dirty paper

of each parameter, the conditions where the parameters are interchangeable, the potential and limitations of MU-MIMO systems, and judicious guidelines for the overall system design. For each utility function $U(\cdot)$, and user priorities ω_k $\forall k$, there exist optimal and suboptimal operation points that can be characterized according to the system parameters and their respective regimes. The capacity of MU-MIMO systems has been assessed in various asymptotic regimes: high SNR ($P \rightarrow \infty$), low SNR ($P \rightarrow 0$), large number of users ($K \rightarrow \infty$), large number of transmit antennas ($M \rightarrow \infty$), and large codebook resolution ($B \rightarrow \infty$).

Table VII summarizes the system parameters and the antenna configurations, i.e., MISO ($N = 1$) or MIMO ($N > 1$), and $M \geq N$. Every single reference in the table has its own system model, assumptions, and constraints, studying one parameter and the corresponding effects on the performance $U(\cdot)$, and other fixed parameters. The table is by no means exhaustive; a comprehensive taxonomy of the asymptotic analytical results is out of the scope of this paper. Our aim is to provide a list of organized results from the reviewed paper, so that the interested reader may use each reference for further studies. Notice that the first row in Table VII points to references that analyze the capacity as a function of different parameters, while the second row refers to works that provide analytical results of MU-MIMO systems with queue constraints.

REFERENCES

[1] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of Resource Allocation in Wireless Networks: Theory and Algorithms*. Berlin, Germany: Springer, 2009.

[2] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent Advances in Radio Resource Management for Heterogeneous LTE/LTE-A Networks,"

IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 2142–2180, Fourth quarter 2014.

[3] M. Ahmed, "Call admission control in wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 7, no. 1, pp. 49–68, First 2005.

[4] A. Asadi and V. Mancuso, "A Survey on Opportunistic Scheduling in Wireless Communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1671–1688, 2013.

[5] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "5G Network Capacity: Key Elements and Technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, March 2014.

[6] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, September 2015.

[7] Q. Li, G. Li, W. Lee, M. i. Lee, D. Mazzarese, B. Clerckx, and Z. Li, "MIMO techniques in WiMAX and LTE: a feature overview," *IEEE Communications Magazine*, vol. 48, no. 5, pp. 86–92, May 2010.

[8] V. Jones and H. Sampath, "Emerging technologies for WLAN," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 141–149, March 2015.

[9] J. Kim and I. Lee, "802.11 WLAN: history and new enabling MIMO techniques for next generation standards," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 134–140, March 2015.

[10] X. Wang, G. Giannakis, and A. Marques, "A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410–2431, Dec 2007.

[11] W. Ajib and D. Haccoun, "An overview of scheduling algorithms in MIMO-based fourth-generation wireless systems," *IEEE Network*, vol. 19, no. 5, pp. 43–48, Sept 2005.

[12] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE J. of S. Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[13] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang, and L. Hanzo, "Survey of Large-Scale MIMO Systems," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1738–1760, 3Q 2015.

[14] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[15] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[16] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, pp. 678–700, Second 2013.

[17] G. Li, J. Niu, D. Lee, J. Fan, and Y. Fu, "Multi-Cell Coordinated Scheduling and MIMO in LTE," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 761–775, Second 2014.

[18] D. Gesbert, M. Kountouris, R. Heath, C.-B. Chae, and T. Salzer, "From Single User to Multiuser Communications: Shifting the MIMO Paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36–46, Sept 2007.

[19] G. Foschini and M. Gans, "On Limits of Wireless Communications in a Fading Environment when using Multiple Antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[20] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[21] J. Mietzner, R. Schober, L. Lampe, W. Gerstacker, and P. Hoeher, "Multiple-Antenna Techniques for Wireless Communications - A Comprehensive Literature Survey," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 2, pp. 87–105, Second 2009.

[22] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[23] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[24] M. Schubert and H. Boche, "Solution of the Multiuser Downlink Beamforming Problem with Individual SINR Constraints," *IEEE Trans. on Vehicular Technology*, vol. 53, no. 1, pp. 18 – 28, jan 2004.

[25] ——, *QoS-based Resource Allocation and Transceiver Optimization*. Hanover, USA: Foundation and Trends(r) in Communications and Information Theory, 2006.

[26] W. Yu and W. Rhee, "Degrees of Freedom in Wireless Multiuser Spatial Multiplex Systems with Multiple Antennas," *IEEE Transactions on Communications*, vol. 54, no. 10, pp. 1747–1753, Oct 2006.

[27] M. Chiang, P. Hande, T. Lan, and C. W. Tan, *Power Control in Wireless Cellular Networks*. Foundations and Trends(r) in Networking, 2008.

[28] L. Zheng and D. Tse, "Diversity and Multiplexing: a fundamental trade-off in Multiple-Antenna Channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[29] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, Jun 2002.

[30] L. Tang, P. Zhu, Y. Wang, and X. You, "Adaptive Modulation in PU2RC Systems with finite rate feedback," *IEEE Transactions on Wireless Communications*, vol. 9, no. 6, pp. 1998–2011, June 2010.

[31] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, Feb 2005.

[32] ——, "A Comparison of Time-Sharing, DPC, and Beamforming for MIMO Broadcast Channels With Many Users," *IEEE Trans. on Commun.*, vol. 55, no. 1, pp. 11–15, Jan 2007.

[33] H.-C. Yang and M.-S. Alouini, *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*. Cambridge University Press, 2011.

[34] R. Heath and A. Paulraj, "Switching between diversity and multiplexing in MIMO systems," *IEEE Transactions on Communications*, vol. 53, no. 6, pp. 962–968, June 2005.

[35] C.-J. Chen and L.-C. Wang, "Enhancing coverage and capacity for multiuser MIMO systems by utilizing scheduling," *IEEE Transactions on Wireless Communications*, vol. 5, no. 5, pp. 1148–1157, May 2006.

[36] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," in *International Zurich Seminar on Communications*, 2006, pp. 10–13.

[37] T. W. Rondeau and C. W. Bostian, *Artificial Intelligence in Wireless Communications*. Artech House, 2009.

[38] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[39] P. Chan and R. Cheng, "Capacity Maximization for Zero-Forcing MIMO-OFDMA Downlink Systems with Multiuser Diversity," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 5, pp. 1880–1889, 2007.

[40] M. Moretti and A. Perez-Neira, "Efficient Margin Adaptive Scheduling for MIMO-OFDMA Systems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 278–287, January 2013.

[41] E. Driouch and W. Ajib, "Efficient Scheduling Algorithms for Multi-antenna CDMA Systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 2, pp. 521–532, Feb 2012.

[42] M. Vu and A. Paulraj, "MIMO Wireless Linear Precoding," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 86–105, Sept 2007.

[43] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, and M. Andrews, "An Overview of Limited Feedback in Wireless Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, October 2008.

[44] M. Kobayshi, N. Jindal, and G. Caire, "Training and Feedback Optimization for Multiuser MIMO Downlink," *IEEE Transactions on Communications*, vol. 59, no. 8, pp. 2228–2240, August 2011.

[45] S. Jafar and A. Goldsmith, "Isotropic Fading Vector Broadcast Channels:The Scalar Upper Bound and Loss in Degrees of Freedom," *IEEE Trans. on Information Theory*, vol. 51, no. 3, pp. 848–857, March 2005.

[46] B. Hassibi and M. Sharif, "Fundamental Limits in MIMO Broadcast Channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1333–1344, September 2007.

[47] I. Koutsopoulos and L. Tassiulas, "The impact of space division multiplexing on resource allocation: a unified treatment of TDMA, OFDMA and CDMA," *IEEE Transactions on Communications*, vol. 56, no. 2, pp. 260–269, February 2008.

[48] V. Lau, "Asymptotic analysis of SDMA systems with near-orthogonal user scheduling (NEOUS) under imperfect CSIT," *IEEE Transactions on Communications*, vol. 57, no. 3, pp. 747–753, March 2009.

[49] T. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[50] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A Comprehensive Survey of Pilot Contamination in Massive MIMO - 5G System," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 905–923, Secondquarter 2016.

[51] N. Jindal, "MIMO Broadcast Channels With Finite-Rate Feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, Nov 2006.

[52] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[53] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept 2006.

[54] E. Bjornson and E. Jorswieck, *Optimal Resource Allocation in Coordinated Multi-Cell Systems*, ser. Foundations and Trends(r) in Communications and Information. Now Publishers Incorporated, 2013.

[55] K. Huang and V. Lau, "Stability and Delay of Zero-Forcing SDMA With Limited Feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6499–6514, Oct 2012.

[56] M. Kobayashi, G. Caire, and D. Gesbert, "Transmit Diversity Versus Opportunistic Beamforming in Data Packet Mobile Downlink Transmission," *IEEE Trans. on Commun.*, vol. 55, no. 1, pp. 151–157, 2007.

[57] J. Zhang, M. Kountouris, J. Andrews, and R. Heath, "Multi-Mode Transmission for the MIMO Broadcast Channel with Imperfect Channel State Information," *IEEE Transactions on Communications*, vol. 59, no. 3, pp. 803–814, March 2011.

[58] X. Gao, B. Jiang, X. Li, A. Gershman, and M. McKay, "Statistical Eigenmode Transmission Over Jointly Correlated MIMO Channels," *IEEE Trans. on Inf. Theory*, vol. 55, no. 8, pp. 3735–3750, 2009.

[59] G. Dimic and N. Sidiropoulos, "On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm," *IEEE Transactions Signal Processing*, vol. 53, no. 10, pp. 3857 – 3868, oct. 2005.

[60] S. Huang, H. Yin, J. Wu, and V. Leung, "User Selection for Multiuser MIMO Downlink With Zero-Forcing Beamforming," *IEEE Trans. on Vehicular Technology*, vol. 62, no. 7, pp. 3084–3097, Sept 2013.

[61] Z. Tu and R. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Communications Letters*, vol. 7, no. 8, pp. 370–372, 2003.

[62] J. Wang, D. Love, and M. Zoltowski, "User Selection With Zero-Forcing Beamforming Achieves the Asymptotically Optimal Sum Rate," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3713–3726, Aug 2008.

[63] T. Yoo and A. Goldsmith, "On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming," *IEEE Journal on Selected Areas in Commun.*, vol. 24, no. 3, pp. 528 – 541, march 2006.

[64] J. Jiang, R. M. Buehrer, and W. Tranter, "Greedy Scheduling Performance for a Zero-Forcing Dirty-Paper Coded System," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 789–793, May 2006.

[65] J. Dai, C. Chang, Z. Ye, and Y. S. Hung, "An efficient greedy scheduler for zero-forcing dirty-paper coding," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1939–1943, July 2009.

[66] G. Lee and Y. Sung, "A New Approach to User Scheduling in Massive Multi-User MIMO Broadcast Channels," *IEEE Transactions on Information Theory*, March 2014, Submitted for publication.

[67] S. Huang, H. Yin, H. Li, and V. Leung, "Decremental User Selection for Large-Scale Multi-User MIMO Downlink with Zero-Forcing Beamforming," *IEEE Wireless Communications Letters*, vol. 1, no. 5, pp. 480–483, October 2012.

[68] A. Razi, D. Ryan, I. Collings, and J. Yuan, "Sum rates, rate allocation, and user scheduling for multi-user MIMO vector perturbation precoding," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 356–365, January 2010.

[69] J. Mao, J. Gao, Y. Liu, and G. Xie, "Simplified Semi-Orthogonal User Selection for MU-MIMO Systems with ZFBF," *IEEE Wireless Communications Letters*, vol. 1, no. 1, pp. 42–45, February 2012.

[70] E. Bjornson, E. G. Larsson, and M. Debbah, "Optimizing Multi-Cell Massive MIMO for Spectral Efficiency: How Many Users Should Be Scheduled?" *ArXiv e-prints*, Oct. 2014.

[71] Y. Shi, Q. Yu, W. Meng, and Z. Zhang, "Maximum Product of Effective Channel Gains: An Innovative User Selection Algorithm for Downlink Multi-User Multiple Input and Multiple Output," *Wireless Communications and Mobile Computing*, 2012.

[72] T. Maciel and A. Klein, "On the Performance, Complexity, and Fairness of Suboptimal Resource Allocation for Multiuser MIMO-OFDMA Systems," *IEEE Trans. Veh. Tech.*, vol. 59, no. 1, pp. 406–419, 2010.

[73] W.-C. Chung, L.-C. Wang, and C.-J. Chang, "A low-complexity beamforming-based scheduling to downlink OFDMA/SDMA systems with multimedia traffic," *Wireless Networks*, vol. 17, no. 3, pp. 611–620, 2011.

[74] T. Yoo and A. Goldsmith, "Sum-rate Optimal Multi-Antenna Downlink Beamforming Strategy based on Clique Search," in *IEEE Global Telecommun. Conf.*, vol. 3, 2005.

[75] D. Christopoulos, S. Chatzinotas, I. Krikidis, and B. Ottersten, "Constructive Interference in Linear Precoding Systems: Power Allocation and User Selection," *arXiv:1303.7454*, 2013. [Online]. Available: http://arxiv.org/abs/1303.7454

[76] X. Xia, G. Wu, J. Liu, and S. Li, "Leakage-based user scheduling in MU-MIMO Broadcast Channel," *Science in China Series F: Information Sciences*, vol. 52, no. 12, pp. 2259–2268, 2009.

[77] E. Matskani, N. Sidiropoulos, Z.-Q. Luo, and L. Tassiulas, "Convex Approximation Techniques for Joint Multiuser Downlink Beamforming and Admission Control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2682–2693, July 2008.

[78] V. Lau, "Optimal Downlink Space-Time Scheduling Design With Convex Utility Functions - Multiple-Antenna Systems With Orthogonal Spatial Multiplexing," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 4, pp. 1322–1333, July 2005.

[79] L. Cottatellucci, M. Debbah, G. Gallinaro, R. Mueller, M. Neri, and R. Rinaldo, "Interference Mitigation Techniques for Broadband Satellite Systems," in *AIAA International Communications Satellite Systems Conference*, June 2006.

[80] M. Wang, T. Samarasinghe, and J. Evans, "Optimizing User Selection Schemes in Vector Broadcast Channels," *IEEE Transactions on Communications*, vol. 63, no. 2, pp. 565–577, Feb 2015.

[81] M. Kobayashi and G. Caire, "Joint Beamforming and Scheduling for a Multi-Antenna Downlink with Imperfect Transmitter Channel Knowledge," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1468–1477, September 2007.

[82] A. Destounis, M. Assaad, M. Debbah, and B. Sayadi, "Traffic-Aware Training and Scheduling for MISO Wireless Downlink Systems," *IEEE Trans. on Inf. Theory*, vol. 61, no. 5, pp. 2574–2599, May 2015.

[83] X. Zhang, E. Jorswieck, B. Ottersten, and A. Paulraj, "User Selection Schemes in Multiple Antenna Broadcast Channels with Guaranteed Performance," in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, June 2007, pp. 1–5.

[84] C. Swannack, E. Uysal-Biyikoglu, and G. Wornell, "Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel," in *Allerton Conf. on Commun., Control, and Computing*, 2004.

[85] C.-F. Tsai, C.-J. Chang, F.-C. Ren, and C.-M. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1734–1743, May 2008.

[86] H. Shirani-Mehr, H. Papadopoulos, S. a. Ramprashad, and G. Caire, "Joint scheduling and ARQ for MU-MIMO downlink in the presence of inter-cell interference," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 578–589, 2011.

[87] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Beamforming and User Selection in SDMA Systems Utilizing Channel Statistics and Instantaneous SNR Feedback," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, April 2007, pp. 113–116.

[88] N. Seifi, M. Matthaiou, and M. Viberg, "Coordinated user scheduling in the multi-cell MIMO downlink," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 2840–2843.

[89] E. Castañeda, A. Silva, R. Samano-Robles, and A. Gameiro, "Distributed Linear Precoding and User Selection in Coordinated Multicell Systems," *IEEE Transactions on Vehicular Technology*, 2015.

[90] H. Huh, A. Tulino, and G. Caire, "Network MIMO With Linear Zero-Forcing Beamforming: Large System Analysis, Impact of Channel Estimation, and Reduced-Complexity Scheduling," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2911–2934, May 2012.

[91] U. Jang, H. Son, J. Park, and S. Lee, "CoMP-CSB for ICI Nulling with User Selection," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 2982–2993, September 2011.

[92] S. Y. Park, J. Choi, and D. Love, "Multicell Cooperative Scheduling for Two-Tier Cellular Networks," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 536–551, February 2014.

[93] M.-L. Ku, L.-C. Wang, and Y.-L. Liu, "Joint Antenna Beamforming, Multiuser Scheduling, and Power Allocation for Hierarchical Cellular Systems," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 5, pp. 896–909, May 2015.

[94] B. Song, Y.-H. Lin, and R. Cruz, "Weighted max-min fair beamforming, power control, and scheduling for a MISO downlink," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 2, pp. 464–469, February 2008.

[95] T. E. Bogale, L. B. Le, and A. Haghighat, "User scheduling for massive MIMO OFDMA systems with hybrid analog-digital beamforming," in *IEEE Int. Conf. on Commun.*, June 2015, pp. 1757–1762.

[96] H. Liu, H. Gao, S. Yang, and T. Lv, "Low-Complexity Downlink User Selection for Massive MIMO Systems," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2015.

[97] A. Bayesteh and A. Khandani, "On the User Selection for MIMO Broadcast Channels," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1086–1107, 2008.

[98] Z. Shen, R. Chen, J. Andrews, R. Heath, and B. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658 –3663, sept. 2006.

[99] L.-C. Wang and C.-J. Yeh, "Scheduling for Multiuser MIMO Broadcast Systems: Transmit or Receive Beamforming?" *IEEE Trans. on Wireless Communications*, vol. 9, no. 9, pp. 2779–2791, 2010.

[100] K. Ko and J. Lee, "Multiuser MIMO User Selection Based on Chordal Distance," *IEEE Transactions on Communications*, vol. 60, no. 3, pp. 649–654, March 2012.

[101] R. Elliott and W. Krzymien, "Downlink Scheduling via Genetic Algorithms for Multiuser Single-Carrier and Multicarrier MIMO Systems With Dirty Paper Coding," *IEEE Transactions Vehicular Technology*, vol. 58, no. 7, pp. 3247 –3262, sept. 2009.

[102] L.-N. Tran and E.-K. Hong, "Multiuser Diversity for Successive Zero-Forcing Dirty Paper Coding: Greedy Scheduling Algorithms and Asymptotic Performance Analysis," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3411–3416, June 2010.

[103] B. H. Wang, H. T. Hui, and Y. T. Yu, "Maximum volume criterion for user selection of multiuser MIMO downlink with multiantenna users and block diagonalization beamforming," in *IEEE Antennas and Propagation Society International Symposium*, July 2010, pp. 1–4.

[104] S. Nam, J. Kim, and Y. Han, "A User Selection Algorithm Using Angle between Subspaces for Downlink MU-MIMO Systems," *IEEE Trans. on Commun.*, vol. 62, no. 2, pp. 616–624, February 2014.

[105] M. Torabzadeh and W. Ajib, "Packet Scheduling and Fairness for Multiuser MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, pp. 1330–1340, March 2010.

[106] O. Souihli and T. Ohtsuki, "Joint feedback and scheduling scheme for service-differentiated multiuser MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 2, pp. 528–533, February 2010.

[107] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1910–1921, April 2009.

[108] Q. Cui, S. Yang, Y. Xu, X. Tao, and B. Liu, "An Effective Inter-Cell Interference Coordination Scheme for Downlink CoMP in LTE-A Systems," in *IEEE Vehicular Tech. Conf.*, Sept 2011, pp. 1–5.

[109] W. Yu, T. Kwon, and C. Shin, "Multicell Coordination via Joint Scheduling, Beamforming, and Power Spectrum Adaptation," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 7, pp. 1–14, July 2013.

[110] H. Park, S.-H. Park, H. Sung, and I. Lee, "Scheduling Methods with MIMO Interference Alignment for Mutually Interfering Broadcast Channels," in *IEEE Global Telecommun. Conf.*, Dec 2010, pp. 1–5.

[111] M. Lossow, S. Jaeckel, V. Jungnickel, and V. Braun, "Efficient MAC protocol for JT CoMP in small cells," in *IEEE International Conference on Communications*, June 2013, pp. 1166–1171.

[112] K. Jagannathan, S. Borst, P. Whiting, and E. Modiano, "Scheduling of Multi-Antenna Broadcast Systems with Heterogeneous Users," *IEEE J. on S. Areas in Commun.*, vol. 25, no. 7, pp. 1424–1434, Sep. 2007.

[113] R. Chen, R. Heath, and J. Andrews, "Transmit Selection Diversity for Unitary Precoded Multiuser Spatial Multiplexing Systems With Linear Receivers," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1159–1171, March 2007.

[114] M. Fuchs, G. Del Galdo, and M. Haardt, "Low-Complexity Space - Time - Frequency Scheduling for MIMO Systems With SDMA," *IEEE Trans. on Vehicular Tech.*, vol. 56, no. 5, pp. 2775–2784, Sept 2007.

[115] P. Tejera, W. Utschick, G. Bauch, and J. Nossek, "Subchannel Allocation in Multiuser Multiple-Input Multiple-Output Systems," *IEEE Trans. on Inf. Theory*, vol. 52, no. 10, pp. 4721–4733, Oct 2006.

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

[117] C. Wang and R. Murch, "Adaptive downlink multi-user MIMO wireless systems for correlated channels with imperfect CSI," *IEEE Trans. on Wireless Commun.*, vol. 5, no. 9, pp. 2435–2446, Sep. 2006.

[118] R. Chen, Z. Shen, J. Andrews, and R. Heath, "Multimode Transmission for Multiuser MIMO Systems With Block Diagonalization," *IEEE Trans. on Signal Processing*, vol. 56, no. 7, pp. 3294–3302, July 2008.

[119] S. Sigdel and W. Krzymien, "Simplified Fair Scheduling and Antenna Selection Algorithms for Multiuser MIMO Orthogonal Space-Division Multiplexing Downlink," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1329–1344, March 2009.

[120] X. Yi and E. Au, "User Scheduling for Heterogeneous Multiuser MIMO Systems: A Subspace Viewpoint," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 4004–4013, Oct 2011.

[121] L. Sun and M. McKay, "Eigen-Based Transceivers for the MIMO Broadcast Channel With Semi-Orthogonal User Selection," *IEEE Trans. on Signal Processing*, vol. 58, no. 10, pp. 5246–5261, Oct 2010.

[122] L.-N. Tran, M. Bengtsson, and B. Ottersten, "Iterative Precoder Design and User Scheduling for Block-Diagonalized Systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3726–3739, July 2012.

[123] R. C. Elliott, S. Sigdel, and W. A. Krzymien, "Low complexity greedy, genetic and hybrid user scheduling algorithms for multiuser MIMO systems with successive zero-forcing," *Transactions on Emerging Telecommunications Technologies*, vol. 23, no. 7, pp. 604–617, 2012.

[124] Y. Hei, X. Li, K. Yi, and H. Yang, "Novel scheduling strategy for downlink multiuser MIMO system: Particle Swarm Optimization," *Science in China Series F: Information Sciences*, vol. 52, no. 12, pp. 2279–2289, Dec. 2009.

[125] B. Lim, W. Krzymien, and C. Schlegel, "Efficient Sum Rate Maximization and Resource Allocation in Block-Diagonalized Space-Division Multiplexing," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 478–484, Jan 2009.

[126] Y. Cheng, S. Li, J. Zhang, F. Roemer, B. Song, M. Haardt, Y. Zhou, and M. Dong, "An Efficient Transmission Strategy for the Multicarrier Multiuser MIMO Downlink," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 628–642, Feb 2014.

[127] G. Aniba and S. Aissa, "Adaptive scheduling for MIMO wireless networks: cross-layer approach and application to HSDPA," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 1, pp. 259–268, Jan 2007.

[128] A. Liu and V. Lau, "Hierarchical Interference Mitigation for Massive MIMO Cellular Networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4786–4797, Sept 2014.

[129] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing: Opportunistic Beamforming, User Grouping and Simplified Downlink Scheduling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 876–890, Oct 2014.

[130] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing: The Large-Scale Array Regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, Oct 2013.

[131] A. Adhikary and G. Caire, "Joint Spatial Division and Multiplexing: Opportunistic Beamforming and User Grouping," *IEEE Transactions on Information Theory*, May 2013, Submitted for publication.

[132] R. Stridh, M. Bengtsson, and B. Ottersten, "System Evaluation of Optimal Downlink Beamforming with Congestion Control in Wireless Communication," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 743–751, April 2006.

[133] M. Kountouris, R. de Francisco, D. Gesbert, D. Slock, and T. Salzer, "Low Complexity Scheduling and Beamforming for Multiuser MIMO Systems," in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, July 2006, pp. 1–5.

[134] M. Kountouris, D. Gesbert, and L. Pittman, "Transmit Correlation-aided Opportunistic Beamforming and Scheduling," in *European Signal Processing Conference*, Sept 2006, pp. 1–5.

[135] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Utilizing the Spatial Information Provided by Channel Norm Feedback in SDMA Systems," *IEEE Trans. on Signal Processing*, vol. 56, no. 7, pp. 3278–3293, July 2008.

[136] ——, "Acquiring Partial CSI for Spatially Selective Transmission by Instantaneous Channel Norm Feedback," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1188–1204, March 2008.

[137] G. Dartmann, X. Gong, and G. Ascheid, "Application of Graph Theory to the Multicell Beam Scheduling Problem," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 4, pp. 1435–1449, May 2013.

[138] Y. Xu, G. Yue, and S. Mao, "User Grouping for Massive MIMO in FDD Systems: New Design Methods and Analysis," *IEEE Access*, vol. 2, pp. 947–959, 2014.

[139] G. Lee and Y. Sung, "Asymptotically optimal simple user scheduling for massive MIMO downlink with two-stage beamforming," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications*, June 2014, pp. 60–64.

[140] V. Raghavan, R. Heath, and A. Sayeed, "Systematic Codebook Designs for Quantized Beamforming in Correlated MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1298–1310, September 2007.

[141] A. Rico-Alvarino and R. Heath, "Learning-Based Adaptive Transmission for Limited Feedback Multiuser MIMO-OFDM," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 7, pp. 3806–3820, July 2014.

[142] J. Fan, G. Li, and X. Zhu, "Multiuser MIMO Scheduling for LTE-A Downlink Cellular Networks," in *IEEE Vehicular Technology Conference*, May 2014, pp. 1–5.

[143] H. Shirani-Mehr, G. Caire, and M. Neely, "MIMO Downlink Scheduling with Non-Perfect Channel State Knowledge," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2055–2066, July 2010.

[144] W. Xu, C. Zhao, and Z. Ding, "Limited Feedback Multiuser Scheduling of Spatially Correlated Broadcast Channels," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4406–4418, Oct 2009.

[145] H. O. W. Weichselberger, M. Herdin and E. Bonek, "Statistical Eigenmode Transmission Over Jointly Correlated MIMO Channels," *IEEE Trans. on Wireless Commun.*, vol. 5, no. 1, p. 90100, Jan 2006.

[146] V. Raghavan, J. Choi, and D. Love, "Design Guidelines for Limited Feedback in the Spatially Correlated Broadcast Channel," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2524–2540, July 2015.

[147] M. Schellmann, L. Thiele, T. Haustein, and V. Jungnickel, "Spatial Transmission Mode Switching in Multiuser MIMO-OFDM Systems With User Fairness," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 1, pp. 235–247, Jan 2010.

[148] E. Bjornson, D. Hammarwall, and B. Ottersten, "Exploiting Quantized Channel Norm Feedback Through Conditional Statistics in Arbitrarily Correlated MIMO Systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 4027–4041, Oct 2009.

[149] E. Bjornson, E. Jorswieck, and B. Ottersten, "Impact of Spatial Correlation and Precoding Design in OSTBC MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3578–3589, November 2010.

[150] B. Godana and T. Ekman, "Parametrization Based Limited Feedback Design for Correlated MIMO Channels Using New Statistical Models," *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 5172–5184, October 2013.

[151] X. Rao, L. Ruan, and V. Lau, "Limited Feedback Design for Interference Alignment on MIMO Interference Networks With Heterogeneous Path Loss and Spatial Correlations," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2598–2607, May 2013.

[152] S.-H. Moon, C. Lee, S.-R. Lee, and I. Lee, "Joint User Scheduling and Adaptive Intercell Interference Cancelation for MISO Downlink Cellular Systems," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 172–181, Jan 2013.

[153] J. Wagner, Y.-C. Liang, and R. Zhang, "On the balance of multiuser diversity and spatial multiplexing gain in random beamforming," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 7, pp. 2512–2525, 2008.

[154] N. Zorba, M. Realp, and A. I. Perez-Neira, "An improved partial CSIT random beamforming for multibeam satellite systems," in *Int. Workshop on Signal Processing for Space Commun.*, Oct 2008.

[155] E. Conte, S. Tomasin, and N. Benvenuto, "A Comparison of Scheduling Strategies for MIMO Broadcast Channel with Limited Feedback on OFDM Systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, no. 1, p. 968703, 2010.

[156] K. Huang, J. Andrews, and R. Heath, "Performance of Orthogonal Beamforming for SDMA With Limited Feedback," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 152–164, Jan 2009.

[157] J. Vicario, R. Bosisio, C. Antón-Haro, and U. Spagnolini, "Beam Selection Strategies for Orthogonal Random Beamforming in Sparse Networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 9, pp. 3385–3396, September 2008.

[158] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Downlink Channels with Limited Feedback and User Selection," *IEEE J. on Selected Areas in Commun.*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.

[159] M. Kountouris, R. de Francisco, D. Gesbert, D. Slock, and T. Salzer, "Efficient Metrics for Scheduling in MIMO Broadcast Channels with Limited Feedback," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, April 2007, pp. III–109–III–112.

[160] M. Kountouris, D. Gesbert, and T. Salzer, "Enhanced multiuser random beamforming: dealing with the not so large number of users case," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1536–1545, October 2008.

[161] W. Dai, Y. Liu, B. Rider, and W. Gao, "How many users should be turned on in a multi-antenna broadcast channel?" *IEEE J. on Selected Areas in Commun.*, vol. 26, no. 8, pp. 1526–1535, Oct. 2008.

[162] N. Ravindran and N. Jindal, "Multi-User Diversity vs. Accurate Channel State Information in MIMO Downlink Channels," *IEEE Trans. on Wireless Commun.*, vol. 11, no. 9, pp. 3037–3046, Sep. 2012.

[163] M. Kountouris, R. de Francisco, D. Gesbert, D. Slock, and T. Salzer, "Multiuser Diversity - Multiplexing Tradeoff in MIMO Broadcast Channels with Limited Feedback," in *Asilomar Conference on Signals, Systems, and Computers*, Oct 2006, pp. 364–368.

[164] C. Wang and R. Murch, "MU-MISO Transmission with Limited Feedback," *IEEE Transactions on Wireless Communications*, vol. 6, no. 11, pp. 3907–3913, November 2007.

[165] W. Choi, A. Forenza, J. Andrews, and R. Heath, "Opportunistic space-division multiple access with beam selection," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2371–2380, Dec 2007.

[166] I. Sohn, J. Andrews, and K.-B. Lee, "MIMO Broadcast Channels with Spatial Heterogeneity," *IEEE Transactions on Wireless Communications*, vol. 9, no. 8, pp. 2449–2454, August 2010.

[167] M. Kountouris, T. Salzer, and D. Gesbert, "Scheduling for Multiuser MIMO Downlink Channels with Ranking-Based Feedback," *EURASIP Journal on Advances in Signal Processing*, no. 1, 2008.

[168] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, June 2005, pp. 975–979.

[169] ——, "Memory-based opportunistic multi-user beamforming," in *Proceedings. International Symposium on Information Theory*, Sept 2005, pp. 1426–1430.

[170] W. Xu and C. Zhao, "Two-Phase Multiuser Scheduling for Multi-antenna Downlinks Exploiting Reduced Finite-Rate Feedback," *IEEE Trans. on Veh. Tech.*, vol. 59, no. 3, pp. 1367–1380, March 2010.

[171] B. Khoshnevis, W. Yu, and Y. Lostanlen, "Two-Stage Channel Quantization for Scheduling and Beamforming in Network MIMO Systems: Feedback Design and Scaling Laws," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 10, pp. 2028–2042, October 2013.

[172] C. Simon and G. Leus, "Round-Robin Scheduling for Orthogonal Beamforming with Limited Feedback," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2486–2496, August 2011.

[173] R. Zakhour and D. Gesbert, "A Two-Stage Approach to Feedback Design in Multi-User MIMO Channels with Limited Channel State Information," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2007, pp. 1–5.

[174] K. Huang, R. Heath, and J. Andrews, "Space Division Multiple Access With a Sum Feedback Rate Constraint," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3879–3891, July 2007.

[175] C.-B. Chae, D. Mazzarese, N. Jindal, and R. Heath, "Coordinated beamforming with limited feedback in the MIMO broadcast channel," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1505–1515, October 2008.

[176] J. Chen and V. Lau, "Large Deviation Delay Analysis of Queue-Aware Multi-User MIMO Systems With Two-Timescale Mobile-Driven Feedback," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4067–4076, Aug 2013.

[177] M. Trivellato, F. Boccardi, and H. Huang, "On transceiver design and channel quantization for downlink multiuser MIMO systems with limited feedback," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1494–1504, October 2008.

[178] N. Jindal, "Antenna combining for the MIMO downlink channel," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 10, pp. 3834–3844, Oct. 2008.

[179] C. van Rensburg and P. Hosein, "Interference Coordination through Network-Synchronized Cyclic Beamforming," in *IEEE Vehicular Technology Conference*, Sept 2009, pp. 1–5.

[180] P. Hosein and C. van Rensburg, "On the Performance of Downlink Beamforming with Synchronized Beam Cycles," in *IEEE Vehicular Technology Conference*, April 2009, pp. 1–5.

[181] M. Min, D. Kim, H.-M. Kim, and G.-H. Im, "Opportunistic Two-Stage Feedback and Scheduling for MIMO Downlink Systems," *IEEE Trans. on Commun.*, vol. 61, no. 1, pp. 312–324, January 2013.

[182] I. Sohn, C. S. Park, and K. B. Lee, "Downlink Multiuser MIMO Systems With Adaptive Feedback Rate," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 3, pp. 1445–1451, March 2012.

[183] W. Zhang and K. Letaief, "MIMO Broadcast Scheduling with Limited Feedback," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1457–1467, September 2007.

[184] M. Wang, F. Li, J. Evans, and S. Dey, "Dynamic Multi-User MIMO scheduling with limited feedback in LTE-Advanced," in *IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, Sept 2012, pp. 1627–1632.

[185] J. Andrews, "Interference Cancellation for Cellular Systems: A Contemporary Overview," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 19–29, April 2005.

[186] A. Lozano, R. Heath, and J. Andrews, "Fundamental Limits of Cooperation," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5213–5226, Sept 2013.

[187] V. Lau and Y. K. R. Kwok, *Channel-adaptive Technologies and Cross-layer Designs for Wireless Systems with Multiple Antennas: Theory and Applications*. John Wiley and Sons, 2006.

[188] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the Number of RF Chains and Phase Shifters, and Scheduling Design With

Hybrid Analog-Digital Beamforming," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3311–3326, May 2016.

[189] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct 1998.

[190] W. Ho and Y.-C. Liang, "Optimal Resource Allocation for Multiuser MIMO-OFDM Systems With User Rate Constraints," *IEEE Trans. on Vehicular Techn.*, vol. 58, no. 3, pp. 1190–1203, March 2009.

[191] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[192] T. Rappaport, R. Heath, R. Daniels, and J. Murdock, *Millimeter Wave Wireless Communications*, ser. Prentice Hall Communications Engineering and Emerging Technologies Series from Ted Rappaport. Pearson Education, 2014.

[193] R. W. Heath, N. Gonzlez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, April 2016.

[194] T. E. Bogale and L. B. Le, "Massive MIMO and mmWave for 5G Wireless HetNet: Potential Benefits and Challenges," *IEEE Vehicular Technology Magazine*, vol. 11, no. 1, pp. 64–75, March 2016.

[195] T. Bai and R. W. Heath, "Coverage and Rate Analysis for Millimeter-Wave Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, Feb 2015.

[196] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter Wave Cellular Networks: A MAC Layer Perspective," *IEEE Trans. on Commun.*, vol. 63, no. 10, pp. 3437–3458, Oct 2015.

[197] S. Kutty and D. Sen, "Beamforming for Millimeter Wave Communications: An Inclusive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 949–973, Second quarter 2016.

[198] E. Jorswieck and H. Boche, *Majorization and matrix-monotone functions in wireless communications*. Foundations and Trends(r) in Communications and Information Theory, 2006, vol. 6.

[199] L. Hanlen and A. Grant, "Capacity Analysis of Correlated MIMO Channels," *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6773–6787, Nov 2012.

[200] D.-S. Shiu, G. Foschini, M. Gans, and J. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. on Commun.*, vol. 48, no. 3, pp. 502–513, Mar 2000.

[201] V. Raghavan, J. H. Kotecha, and A. Sayeed, "Why Does the Kronecker Model Result in Misleading Capacity Estimates?" *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 4843–4864, Oct 2010.

[202] N. Czink, B. Bandemer, G. Vazquez-Vilar, L. Jalloul, C. Oestges, and A. Paulraj, "Spatial separation of multi-user MIMO channels," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2009, pp. 1059–1063.

[203] S. Loyka and A. Kouki, "On MIMO channel capacity, correlations, and keyholes: analysis of degenerate channels," *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1886–1888, Dec 2002.

[204] P. Almers, F. Tufvesson, and A. Molisch, "Keyhole effect in MIMO wireless channels: Measurements and theory," *IEEE Transactions on Wireless Communications,*, vol. 5, no. 12, pp. 3596–3604, Dec 2006.

[205] P. Arapoglou, K. Liolis, M. Bertinelli, A. Panagopoulos, P. Cottis, and R. De Gaudenzi, "MIMO over Satellite: A Review," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 1, pp. 27–51, First 2011.

[206] A. Osseiran, J. Monserrat, and W. Mohr, *Mobile and Wireless Communications for IMT-Advanced and Beyond*. Wiley, 2011.

[207] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-Cell MIMO Cooperative Networks: A New Look at Interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, December 2010.

[208] D. Nguyen and T. Le-Ngoc, *Wireless Coordinated Multicell Systems: Architectures and Precoding Designs*, ser. SpringerBriefs in computer science. Springer, 2014.

[209] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 2nd ed. Elsevier Science, 2014.

[210] P. Marsch and G. Fettweis, *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, 2011.

[211] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," *IEEE Commun. Magazine*, vol. 50, no. 2, pp. 148–155, February 2012.

[212] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated Beamforming for MISO Interference Channel: Complexity Analysis and Efficient Algorithms," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1142–1157, March 2011.

[213] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim, "Recent Trend of Multiuser MIMO in LTE-advanced," *IEEE Communications Magazine*, no. March, pp. 127–135, 2013.

[214] L. Liu, R. Chen, S. Geirhofer, K. Sayana, Z. Shi, and Y. Zhou, "Downlink MIMO in LTE-advanced: SU-MIMO vs. MU-MIMO," *IEEE Commun. Magazine*, vol. 50, no. 2, pp. 140–147, February 2012.

[215] R. Liao, B. Bellalta, M. Oliver, and Z. Niu, "MU-MIMO MAC Protocols for Wireless Local Area Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 162–183, 2016.

[216] E. Hamed, H. Rahul, M. A. Abdelghany, and D. Katabi, "Real-time Distributed MIMO Systems," in *Proceedings of ACM SIGCOMM Conference*. ACM, 2016, pp. 412–425.

[217] M. Bengtsson and B. Ottersten, "Handbook of Antennas in Wireless Communications," L. C. Godara, Ed. CRC Press, 2002, ch. Optimum and Suboptimum Transmit Beamforming.

[218] M. H. M. Costa, "Writing on dirty paper (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[219] G. Caire, "MIMO Downlink Joint Processing and Scheduling: A Survey of Classical and Recent Results," in *Proc. Workshop on Information Theory and Its Applications*, 2006.

[220] B. Hochwald, C. Peel, and A. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication - Part II: Perturbation," *IEEE Transactions on Communications*, vol. 53, no. 3, pp. 537–544, March 2005.

[221] A. Kurve, "Multi-user MIMO Systems: The Future in the Making," *IEEE Potentials*, vol. 28, no. 6, pp. 37–42, November 2009.

[222] E. Larsson and E. Jorswieck, "Competition Versus Cooperation on the MISO Interference Channel," *IEEE Journal on Selected Areas in Commun.*, vol. 26, no. 7, pp. 1059–1069, September 2008.

[223] E. Jorswieck, E. Larsson, and D. Danev, "Complete Characterization of the Pareto Boundary for the MISO Interference Channel," *IEEE Trans. on Signal Processing*, vol. 56, no. 10, pp. 5292–5296, Oct 2008.

[224] M. Hong and Z.-Q. Luo, "Academic Press Library in Signal Processing: Communications and Radar Signal Processing," S. Theodoridis and R. Chellappa, Eds. Elsevier Science, 2013, vol. 2, ch. 8 - Signal Processing and Optimal Resource Allocation for the Interference Channel.

[225] E. Bjornson, M. Bengtsson, and B. Ottersten, "Optimal Multiuser Transmit Beamforming: A Difficult Problem with a Simple Solution Structure," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, July 2014.

[226] M. Sadek, A. Tarighat, and A. Sayed, "A Leakage-Based Precoding Scheme for Downlink Multi-User MIMO Channels," *IEEE Trans. on Wireless Communications*, vol. 6, no. 5, pp. 1711–1721, 2007.

[227] P. C. Weeraddana, M. Codreanu, M. Latva-aho, A. Ephremides, and F. Carlo, *Weighted Sum-Rate Maximization in Wireless Networks: A Review*. Foundations and Trends (r) in Networking, 2012, vol. 6.

[228] S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted Sum-Rate Maximization using Weighted MMSE for MIMO-BC Beamforming Design," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 4792–4799, December 2008.

[229] C. Peel, B. Hochwald, and a.L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication - Part I: Channel Inversion and Regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[230] J. Gentle, *Matrix algebra: Theory, Computations, and Applications in Statistics*. Springer, 2007.

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

[232] V. Stankovic and M. Haardt, "Generalized Design of Multi-User MIMO Precoding Matrices," *IEEE Transactions on Wireless Communications*, vol. 7, no. 3, pp. 953–961, March 2008.

[233] A. Dabbagh and D. Love, "Precoding for Multiple Antenna Gaussian Broadcast Channels With Successive Zero-Forcing," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3837–3850, July 2007.

[234] M. Alodeh, S. Chatzinotas, and B. Ottersten, "Constructive Multiuser Interference in Symbol Level Precoding for the MISO Downlink Channel," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2239–2252, May 2015.

[235] C. Masouros and E. Alsusa, "Dynamic linear precoding for the exploitation of known interference in MIMO broadcast systems," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 3, pp. 1396–1404, 2009.

[236] D. Love, R. Heath, and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2735–2747, Oct 2003.

[237] K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On Beamforming with Finite Rate Feedback in Multiple-Antenna Systems," *IEEE Trans. on Inf. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct 2003.

[238] I.-M. Kim, S.-C. Hong, S. Ghassemzadeh, and V. Tarokh, "Opportunistic Beamforming Based on Multiple Weighting Vectors," *IEEE Trans. on Wireless Commun.*, vol. 4, no. 6, pp. 2683–2687, Nov 2005.

[239] D. Castanheira, A. Silva, and A. Gameiro, "Minimum Codebook Size to Achieve Maximal Diversity Order for RVQ-Based MIMO Systems," *IEEE Commun. Letters*, vol. 18, no. 8, pp. 1463–1466, Aug 2014.

[240] A. Silva, R. Holakouei, D. Castanheira, A. Gameiro, and R. Dinis, "A novel distributed power allocation scheme for coordinated multicell systems," *EURASIP J. Wireless Commun. and Net.*, vol. 30, 2013.

[241] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, "Linear Precoding and Equalization for Network MIMO With Partial Cooperation," *IEEE Trans. on Veh. Tech.*, vol. 61, no. 5, pp. 2083–2096, Jun 2012.

[242] E. Bjornson, R. Zakhour, D. Gesbert, and B. Ottersten, "Cooperative Multicell Precoding: Rate Region Characterization and Distributed Strategies With Instantaneous and Statistical CSI," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4298–4310, 2010.

[243] B. E. Godana and D. Gesbert, "Egoistic vs. altruistic beamforming in multi-cell systems with feedback and back-haul delays," *EURASIP Journal on Wireless Commun. and Net.*, vol. 2013, no. 1, p. 253, 2013.

[244] C. Suh, M. Ho, and D. Tse, "Downlink Interference Alignment," *IEEE Trans. on Communications*, vol. 59, no. 9, pp. 2616–2626, Sep. 2011.

[245] P. Ferrand and J.-M. Gorce, "Downlink Cellular Interference Alignment," INRIA, Research Report RR-8543, May 2014. [Online]. Available: https://hal.inria.fr/hal-00996728

[246] J. Chen and V. Lau, "Two-Tier Precoding for FDD Multi-Cell Massive MIMO Time-Varying Interference Networks," *IEEE J. on Selected Areas in Commun.*, vol. 32, no. 6, pp. 1230–1238, June 2014.

[247] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[248] C. W. Tan, M. Chiang, and R. Srikant, "Fast Algorithms and Performance Bounds for Sum Rate Maximization in Wireless Networks," *IEEE/ACM Trans. on Networking*, vol. 21, no. 3, pp. 706–719, 2012.

[249] R. Zakhour and S. Hanly, "Min-Max Power Allocation in Cellular Networks With Coordinated Beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 287–302, February 2013.

[250] Y. Huang, C. W. Tan, and B. Rao, "Joint Beamforming and Power Control in Coordinated Multicell: Max-Min Duality, Effective Network and Large System Transition," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2730–2742, June 2013.

[251] O. Bejarano, E. Knightly, and M. Park, "IEEE 802.11ac: from channelization to multi-user MIMO," *IEEE Communications Magazine*, vol. 51, no. 10, pp. 84–90, October 2013.

[252] A. Networks, "802.11ac In-Depth," 2014, white Paper. [Online]. Available: http://goo.gl/wT5NOX

[253] E. Bjornson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO Systems With Non-Ideal Hardware: Energy Efficiency, Estimation, and Capacity Limits," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 7112–7139, Nov 2014.

[254] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[255] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[256] F. Sohrabi and W. Yu, "Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, April 2016.

[257] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-Complexity Technologies for 5G Millimeter-Wave MIMO Systems with Large Antenna Arrays," *CoRR*, jul 2016. [Online]. Available: https://arxiv.org/abs/1607.04559v1

[258] X. Yu, J. C. Shen, J. Zhang, and K. B. Letaief, "Alternating Minimization Algorithms for Hybrid Precoding in Millimeter Wave MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 485–500, April 2016.

[259] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, February 2016.

[260] S. Liu, C. Zhang, and G. Lyu, "User selection and power schedule for downlink non-orthogonal multiple access (NOMA) system," in *IEEE Int. Conf. on Commun. Workshop*, June 2015, pp. 2561–2565.

[261] S. A. Jafar, *Interference Alignment - A New Look at Signal Dimensions in a Communication Network*. Foundations and Trends(r) in Communications and Information Theory, 2011.

[262] E. Castañeda, A. Silva, and A. Gameiro, "Contemporary Issues in Wireless Communications," M. Khatib, Ed. InTech, 2014, ch. User Selection and Precoding Techniques for Rate Maximization in Broadcast MISO Systems.

[263] G. Golub and C. Van Loan, *Matrix Computations*, ser. Matrix Computations. Johns Hopkins University Press, 1996.

[264] H. Yanai, K. Takeuchi, and Y. Takane, *Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition*. Springer, 2011.

[265] A. Manolakos, Y. Noam, and A. Goldsmith, "Null Space Learning in Cooperative MIMO Cellular Networks Using Interference Feedback," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3961–3977, July 2015.

[266] P. Lu and H.-C. Yang, "Sum-rate Analysis of Multiuser MIMO System with Zero-Forcing Transmit Beamforming," *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2585–2589, 2009.

[267] C. K. Au-Yeung and D. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 2, pp. 458–462, 2007.

[268] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2006.

[269] A. Barg and D. Nogin, "Bounds on packings of spheres in the Grassmann manifold," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2450–2454, Sep 2002.

[270] X. Li, S. Jin, X. Gao, and R. Heath, "3D Beamforming for Large-scale FD-MIMO Systems Exploiting Statistical Channel State Information," *IEEE Transactions on Vehicular Technology*, 2016.

[271] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[272] J. Mo and J. Walrand, "Fair End-to-End Window-based Congestion Control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, Oct 2000.

[273] W. Utschick and J. Brehmer, "Monotonic Optimization Framework for Coordinated Beamforming in Multicell Networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1899–1909, April 2012.

[274] V. Lau, "Proportional Fair Space - Time Scheduling for Wireless Communications," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1353–1360, Aug 2005.

[275] D. Park, H. Seo, H. Kwon, and B. G. Lee, "Wireless Packet Scheduling Based on the Cumulative Distribution Function of User Transmission Rates," *IEEE Transactions on Communications*, vol. 53, no. 11, pp. 1919–1929, Nov 2005.

[276] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in Wireless Networks: Issues, Measures and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, pp. 5–24, First 2014.

[277] D. Bartolome and A. Perez-Neira, "Spatial Scheduling in Multiuser Wireless Systems: From Power Allocation to Admission Control," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2082–2091, Aug 2006.

[278] R. Jain, W. Hawe, and D. Chiu, "A Quantitative measure of fairness and discrimination for resource allocation in Shared Computer Systems," DEC-TR-301, Tech. Rep., 1984.

[279] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An Axiomatic Theory of Fairness in Network Resource Allocation," in *Proceedings IEEE INFOCOM*, March 2010, pp. 1–9.

[280] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, "Multiresource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework," *IEEE/ACM Trans. on Net.*, vol. 21, no. 6, pp. 1785–1798, Dec 2013.

[281] S.-H. Park, H. Park, H. Kong, and I. Lee, "New Beamforming Techniques Based on Virtual SINR Maximization for Coordinated Multi-Cell Transmission," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1034–1044, March 2012.

[282] E. Chaponniere, P. Black, J. Holtzman, and D. Tse, "Transmitter directed code division multiple access system using path diversity to equitably maximize throughput," May 1999, US Patent 6,449,490.

[283] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in Cellular Systems: Understanding Ultra-Dense Small Cell Deployments," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2078–2101, Fourthquarter 2015.

[284] D. Feng, C. Jiang, G. Lim, L. J. Cimini, G. Feng, and G. Y. Li, "A survey of energy-efficient wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 167–178, First 2013.

[285] D. Cai, T. Quek, and C.-W. Tan, "A Unified Analysis of Max-Min Weighted SINR for MIMO Downlink System," *IEEE Trans. on Signal Processing*, vol. 59, no. 8, pp. 3850–3862, Aug 2011.

[286] H. Mahdavi-Doost, M. Ebrahimi, and A. Khandani, "Characterization of SINR Region for Interfering Links With Constrained Power," *IEEE Trans. on Inf. Theory*, vol. 56, no. 6, pp. 2816 –2828, june 2010.

[287] E. Castañeda, R. Samano-Robles, and A. Gameiro, "Sum Rate Maximization via Joint Scheduling and Link Adaptation for Interference-Coupled Wireless Systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, p. 268, 2013.

[288] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-layer Control in Wireless Networks*, ser. Foundations and Trends(r) in Networking. Now Publishers Incorporated, 2006.

[289] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 18–24, October 2011.

[290] R. Baldemair, T. Irnich, K. Balachandran, E. Dahlman, G. Mildh, Y. Seln, S. Parkvall, M. Meyer, and A. Osseiran, "Ultra-dense networks in millimeter-wave frequencies," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 202–208, January 2015.

[291] T. M. Nguyen, V. N. Ha, and L. B. Le, "Resource Allocation Optimization in Multi-User Multi-Cell Massive MIMO Networks Considering Pilot Contamination," *IEEE Access*, vol. 3, pp. 1272–1287, 2015.

[292] J. Li, E. Bjornson, T. Svensson, T. Eriksson, and M. Debbah, "Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5810–5822, Oct 2015.

[293] T. Yang, F. Heliot, and C. H. Foh, "A survey of green scheduling schemes for homogeneous and heterogeneous cellular networks," *IEEE Commun. Magazine*, vol. 53, no. 11, pp. 175–181, November 2015.

[294] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User Association in 5G Networks: A Survey and an Outlook," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1018–1044, Secondquarter 2016.

[295] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 30–37, June 2011.

[296] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 2, pp. 584–593, Feb. 2008.

[297] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3496–3509, July 2013.

[298] A. B. Sediq, R. Schoenen, H. Yanikomeroglu, and G. Senarath, "Optimized Distributed Inter-Cell Interference Coordination (ICIC) Scheme Using Projected Subgradient and Network Flow Optimization," *IEEE Trans. on Commun.*, vol. 63, no. 1, pp. 107–124, Jan 2015.

[299] E. Bjornson, L. Sanguinetti, and M. Kountouris, "Deploying Dense Networks for Maximal Energy Efficiency: Small Cells Meet Massive MIMO," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 832–847, April 2016.

[300] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective Signal Processing Optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 14–23, Nov 2014.

[301] A. Sharifian, R. Schoenen, and H. Yanikomeroglu, "Joint Realtime and Nonrealtime Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2589–2607, April 2016.

[302] E. Bjornson, M. Kountouris, M. Bengtsson, and B. Ottersten, "Receive Combining vs. Multi-Stream Multiplexing in Downlink Systems With Multi-Antenna Users," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3431–3446, July 2013.

[303] S. Rao, *Engineering Optimization: Theory and Practice*. John Wiley & Sons, Inc., 2009.

[304] Z. Zhang, K. Long, J. Wang, and F. Dressler, "On Swarm Intelligence Inspired Self-Organized Networking: Its Bionic Mechanisms, Designing Principles and Optimization Approaches," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 513–537, First 2014.

[305] L. H. D. Sampaio, M. H. Adaniya, M. de Paula Marques, P. J. E. Jeszensky, and T. Abrão, "Ant Colony Optimization for Resource Allocation and Anomaly Detection in Communication Networks," T. Abrão, Ed. InTech Publisher, 2013.

[306] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 3, pp. 836–850, June 2015.

[307] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal User-Cell Association for Massive MIMO Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1835–1850, March 2016.

[308] E. Castañeda, A. Silva, R. Samano-Robles, and A. Gameiro, "Low-complexity User Selection for Rate Maximization in MIMO Broadcast Channels with Downlink Beamforming," *The Scientific World Journal*, pp. 1–13, 2013.

[309] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems*. Siam, 2009.

[310] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, ser. Algorithms and Combinatorics. Springer, 2006.

[311] E. Castañeda, R. Samano-Robles, and A. Gameiro, "Low Complexity Scheduling Algorithm for the Downlink of Distributed Antenna Systems," in *IEEE Vehicular Technology Conference*, June 2013, pp. 1–5.

[312] Y. Wang, W. Feng, Y. Li, S. Zhou, and J. Wang, "Coordinated User Scheduling for Multi-Cell Distributed Antenna Systems," in *IEEE Global Telecommunications Conference*, Dec 2011, pp. 1–5.

[313] E. Pateromichelakis, M. Shariat, A. u. Quddus, and R. Tafazolli, "On the Evolution of Multi-Cell Scheduling in 3GPP LTE / LTE-A," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, pp. 701–717, 2013.

[314] E. Hossain, D. Kim, and V. K. Bhargava, *Cooperative Cellular Wireless Networks*. Cambridge University Press, 2011.

[315] H. Yang and T. Marzetta, "Performance of Conjugate and Zero-Forcing Beamforming in Large-Scale Antenna Systems," *IEEE J. on Selected Areas in Commun.*, vol. 31, no. 2, pp. 172–179, Feb. 2013.

[316] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan 2013.

[317] E. Bjornson, E. Larsson, and T. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, February 2016.

[318] S. L. H. Nguyen, T. Le-Ngoc, and A. Ghrayeb, "Robust Channel Estimation and Scheduling for Heterogeneous Multiuser Massive MIMO Systems," in *Proceedings of European Wireless Conf.*, May 2014.

[319] O. Bai, H. Gao, T. Lv, and C. Yuen, "Low-complexity user scheduling in the downlink massive MU-MIMO system with linear precoding," in *IEEE Int. Conf.on Commun. in China*, Oct 2014, pp. 380–384.

[320] N. Abu-Ali, A. E. M. Taha, M. Salah, and H. Hassanein, "Uplink Scheduling in LTE and LTE-Advanced: Tutorial, Survey and Evaluation Framework," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1239–1265, Third 2014.

[321] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?" in *IEEE International Conference on Communications*, vol. 1, June 2004, pp. 234–238.

[322] W. Wang, A. Harada, and H. Kayama, "Enhanced Limited Feedback Schemes for DL MU-MIMO ZF Precoding," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1554–1561, April 2013.

[323] M. Kountouris and J. Andrews, "Downlink SDMA with Limited Feedback in Interference-Limited Wireless Networks," *IEEE Trans. on Wireless Commun.*, vol. 11, no. 8, pp. 2730–2741, August 2012.

[324] A. Nguyen and B. Rao, "CDF Scheduling Methods for Finite Rate Multiuser Systems With Limited Feedback," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3086–3096, June 2015.

[325] G. Lee, Y. Sung, and M. Kountouris, "On the Performance of Random Beamforming in Sparse Millimeter Wave Channels," *IEEE J. of Selected Topics in Signal Proc.*, vol. 10, no. 3, pp. 560–575, 2016.

[326] M. Kountouris, "Multiuser Multi-Antenna Systems with Limited Feedback," Ph.D. dissertation, Telecom Paris (ENST), Jan. 2008.

[327] W. Xu, C. Zhao, and Z. Ding, "Limited feedback design for MIMO broadcast channels with ARQ mechanism," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 2132–2141, April 2009.

[328] O. Aboul-Magd, U. Kwon, Y. Kim, and C. Zhu, "Managing downlink multi-user MIMO transmission using group membership," in *IEEE Consumer Commun. and Net. Conf.*, Jan 2013, pp. 370–375.

[329] G. Ku and J. M. Walsh, "Resource Allocation and Link Adaptation in LTE and LTE Advanced: A Tutorial," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1605–1633, thirdquarter 2015.

[330] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686–695, Feb 2005.

[331] J. Lee and N. Jindal, "High SNR Analysis for MIMO Broadcast Channels: Dirty Paper Coding Versus Linear Precoding," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4787–4792, Dec 2007.

[332] Y. Kim, H. Ji, J. Lee, Y.-H. Nam, B. L. Ng, I. Tzanidis, Y. Li, and J. Zhang, "Full dimension MIMO (FD-MIMO): the next evolution of MIMO in LTE systems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 26–33, April 2014.

[333] Y.-H. Nam, M. S. Rahman, Y. Li, G. Xu, E. Onggosanusi, J. Zhang, and J.-Y. Seol, "Full Dimension MIMO for LTE-Advanced and 5G," in *Information Theory and Applications Workshop*, UCSD, USA, 2015.

[334] M. Tolstrup, *Indoor Radio Planning: A Practical Guide for 2G, 3G and 4G*. Wiley, 2015.

[335] H. Q. Ngo, A. E. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO versus Small Cells," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1602.08232

[336] A. Puglielli, A. Townley, G. LaCaille, V. Milovanovi, P. Lu, K. Trotskovsky, A. Whitcombe, N. Narevsky, G. Wright, T. Courtade, E. Alon, B. Nikoli, and A. M. Niknejad, "Design of Energy- and Cost-Efficient Massive MIMO Arrays," *Proceedings of the IEEE*, vol. 104, no. 3, pp. 586–606, March 2016.

[337] R. Taori and A. Sridharan, "Point-to-multipoint in-band mmwave backhaul for 5G networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 195–201, January 2015.

[338] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. wall, O. Edfors, and F. Tufvesson, "A flexible 100-antenna testbed for Massive MIMO," in *IEEE Globecom Workshop*, Dec 2014, pp. 287–293.

[339] E. Björnson and E. G. Larsson, "Three Practical Aspects of Massive MIMO: Intermittent User Activity, Pilot Synchronism, and Asymmetric Deployment," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1509.06517

[340] T. Al-Naffouri, M. Sharif, and B. Hassibi, "How much does transmit correlation affect the sum-rate scaling of MIMO gaussian broadcast channels?" *IEEE Transactions on Communications*, vol. 57, no. 2, pp. 562–572, February 2009.