Joint User Scheduling and Beam Selection Optimization for Beam-Based Massive MIMO Downlinks

Zhiyuan Jiang, *Member, IEEE*, Sheng Chen, *Student Member, IEEE*, Sheng Zhou, *Member, IEEE*, and Zhisheng Niu, *Fellow, IEEE*

***Abstract*— In beam-based massive multiple-input multiple-output systems, signals are processed spatially in the radio-frequency (RF) front end and thereby the number of RF chains can be reduced to save hardware cost, power consumptions, and pilot overhead. Most existing work focuses on how to select or design analog beams to achieve performance close to full digital systems. However, since beams are strongly correlated (directed) to certain users, the selection of beams and scheduling of users should be jointly considered. In this paper, we formulate the joint user scheduling and beam selection problem based on the Lyapunov-drift optimization framework and obtain the optimal scheduling policy in a closed form. For reduced overhead and computational cost, the proposed scheduling schemes are based only upon statistical channel state information. Towards this end, asymptotic expressions of the downlink broadcast channel capacity are derived. To address the weighted sum rate maximization problem in the Lyapunov optimization, an algorithm based on block coordinated update is proposed and proved to converge to the optimum of the relaxed problem. To further reduce the complexity, an incremental greedy scheduling algorithm is also proposed, whose performance is proved to be bounded within a constant multiplicative factor. Simulation results based on widely-used spatial channel models are given. It is shown that the proposed schemes are close to optimal and outperform several state-of-the-art schemes.**

***Index Terms*— Massive MIMO, user-scheduling, hybrid beamforming, statistical CSI.**

I. Introduction

IN MASSIVE multiple-input multiple-output (MIMO) based wireless communication systems, the spectral and radiated energy efficiency can be both boosted by the deployment of massive number of antennas [2]. Moreover, the high beamforming gain of a massive antenna array is the main enabler for millimeter-wave systems against high pathloss. Therefore, it is extremely important to design high-performance, efficient and practical transmission strategy in massive MIMO systems for the emerging 5G cellular system.

Under the assumption that full digital signal processing is performed at the base station (BS) side with massive antenna arrays, the system performance has been widely investigated, e.g., in [2]–[4]. However, it is widely accepted that full digital signal processing implementation encounters very severe challenges in practice, on account of the following impediments.

Radio-Frequency (RF) Chain Hardware Cost and Power Consumptions: Full digital signal processing requires that all antennas can be digitally controlled from baseband. Hence, one dedicated RF chain, including e.g., low-noise amplifier, analog-digital-converter (ADC), power amplifier and etc., is needed for each antenna. In massive MIMO systems, not only is this requirement entails a dramatic increase in the deployment cost of the system, but also that the power consumption would be driven up to a prohibitive level. As indicated in [5] and [6], concretely, a BS with 256 RF chains consumes about 10 times the power (only the RF chains) as compared with an entire current long-term-evolution (LTE) BS.

Baseband Signal Processing Complexity: The spatial baseband processing includes multiple kinds of matrix operations, such as inversions and singular-value-decompositions (SVDs) whose complexity scales with M^3 where M is the number of antenna elements for full digital processing. Moreover, these extremely demanding matrix operations are required to be executed very frequently (once every 1 ms for spatial precoding in LTE systems). This is very challenging to the design of baseband processing units, both in terms of chip costs and power consumption.

System Specific Limitations: Aside from the first two challenges, there are some other practical considerations which are system-specific. For example, the fronthaul interface in cloud radio access networks (C-RAN) poses a serious limitation in the number of data streams that can be transmitted between the remote-radio-units (RRUs) and the baseband units (BBUs). Considerable amounts of work has been dedicated to the signal spatial compression in C-RAN [7]. Moreover, the channel state information (CSI) acquisition overhead in frequency-division-duplexing (FDD) system scales with the number of digitally controllable antennas. It constitutes a

Manuscript received April 17, 2017; revised September 12, 2017 and November 18, 2017; accepted December 30, 2017. Date of publication January 12, 2018; date of current version April 8, 2018. This work was supported in part by the Nature Science Foundation of China under Grant 61701275, Grant 91638204, Grant 61571265, and Grant 61621091, in part by the China Postdoctoral Science Foundation, and in part by the Intel Collaborative Research Institute for Mobile Networking and Computing. Part of this paper was presented at the 23rd Asia–Pacific Conference on Communications [1]. The associate editor coordinating the review of this paper and approving it for publication was E. A. Jorswieck. *(Corresponding author: Sheng Zhou).*

The authors are with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhiyuan@tsinghua.edu.cn; chen-s16@mails.tsinghua.edu.cn; sheng.zhou@tsinghua.edu.cn; niuzhs@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2789895

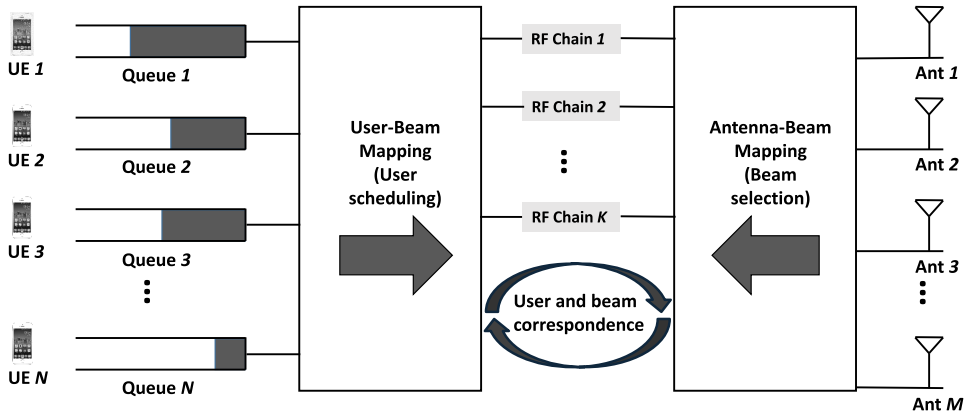


Fig. 1. An illustration of the beam-based massive MIMO systems where user scheduling and beam selection are correlated.

major bottleneck in realizing the massive MIMO gain in FDD systems.

In view of these challenges, architectures with low RF- and processing-complexity have been proposed extensively, e.g., in [8]–[15]. The existing literature can be divided into three categories. The first is *hybrid beamforming*, which adopts an RF front end with an analog beamforming module such that the number of RF chains is significantly reduced [10]. Although the analog beamforming module is usually composed of phase shifters with constant-amplitude beamforming weights to save hardware cost, the high-speed phase shifters, whose quantity is the same with the number of antenna elements, pose a drastic challenge to the cost of RF front ends. In this regard, the recently proposed *beamspace MIMO* architecture [12] adopts a lens antenna array which acts analogously like a lens focusing on light beams from different directions. It transforms the signal into the angular domain and thus reduces the number of RF chains due to signal angular sparsity. Since it does not require any phase shifters, the total cost is reduced, and therefore it is considered to be one of the candidate solutions to the 5G millimeter-wave massive MIMO systems. The other approach is based on digital beamforming which involves *multi-layer signal processing* [5], [13], [16]. Although the number of RF chains is not reduced, the processing complexity and pilot overhead problems are partly solved.

In essence, all the aforementioned solutions aim at providing comparable performance as full digital processing systems with limited number of RF chains and reduced complexity in massive MIMO systems. Based on Fig. 1, the current literature mainly focuses on the right side of the figure, i.e., the antenna to beam mapping and beam selection schemes, which leverages the angular domain power sparsity of the channel to transform the signals from the antenna domain to the beam domain. Towards this end, beam sweeping and steering methods in the hybrid beamforming architecture can be used to capture the signal direction [17]. Other methods adopt joint analog and digital precoding design [11], [18]. In beamspace MIMO systems, the lens antenna array can be regarded as a directional beamforming module with low cost. On the other hand, the left side of the Fig. 1 which represents user-beam mapping, is scantly treated. The user-beam mapping essentially deals with **user scheduling in the beam domain**. Unlike the previous user scheduling related work, e.g., in [19]–[21], the user scheduling problem in the beam domain is tangled with the beam selection. In reality, due to the angular sparsity of the massive MIMO channel [22], the beams, which represent the signal directions, are strongly related to the users, in the sense that each beam usually contains signals of very few (possibly one) users. Therefore, the user scheduling and beam selection have to be jointly considered to avoid possible performance degradation due to user-beam mismatch.

The channel state information (CSI) is of vital importance to the system. The CSI can be categorized as instantaneous CSI and statistical CSI. It is worthwhile to emphasize the specific CSI usage at each stage (time scale) of the beam-based massive MIMO transmissions since most existing work ignores this and assumes instantaneous CSI is always available [23], [24]. We propose that beam-based downlink scheduling should be performed *only* based on the **statistical CSI**. The reason is two-fold. From an implementation perspective, the statistical CSI is much easier to obtain than instantaneous CSI, attributing to the fact that statistical CSI can be obtained with much lower cost because a) it can be estimated without dedicated pilots [25]; b) it varies at a lower speed (in the order of 1 second to 10 seconds) compared with instantaneous CSI (in the order of 1 ms to 100 ms) [26]. Moreover, in C-RAN systems, the beamforming module is integrated with the remote radio heads (RRHs) and hence limited computation capability is expected [27] which prevents us from using complicated channel estimation schemes. On the other hand, it is also theoretically possible to only rely on statistical CSI in the user scheduling and beam selection phase, since beams are essentially long-term statistics. Furthermore, the statistical CSI can be obtained efficiently with a limited number of RF chains based on, e.g., compressive sensing based channel estimation schemes [9], [28].

In this paper, we aim to address the user and beam joint scheduling problem in beam-based massive MIMO downlinks. The contributions include:

- We formulate the problem based on the Lyapunov-drift optimization framework. An optimal scheduling policy is proposed thereby to achieve optimum utilities.

The optimality proof is given which shows the achieved utility is arbitrarily close to the optimum.

- To address the queue weighted sum rate maximization problem arisen in optimizing the Lyapunov-drift, which is a mixed integer programming (MIP) problem, the block-coordinated-update-based (BCU-based) algorithm which deals with the continuous convex relaxation of the MIP problem is proposed. In order to implement the algorithm based on statistical CSI, a deterministic equivalence of the downlink broadcast channel capacity in the large antenna array regime is derived, depending only on statistical CSI. The BCU-based algorithm is proved to converge to the global optimum of the relaxed problem. An iterative water-filling based approach is also proposed to reduce the number of iterations.
- Furthermore, a low-complexity incremental greedy algorithm is proposed. We prove that the greedy algorithm can achieve near-optimal performance, within a multiplicative factor due to the submodular property of the problem.
- By simulations, it is shown that the proposed algorithms outperforms several state-of-the-art beam selection schemes. Moreover, since it is based on statistical CSI, the frequency of executing the algorithm is significantly reduced, making it more preferable in practice.

A. Related Work

The proposed joint user and scheduling schemes are related to the beam selection problem in beamspace MIMO systems [23], [24], [29]–[31], or more generally antenna selection problem in MIMO systems [32]. However, the considered joint scheduling problem in beam domain is unique, in the sense that the beam magnitudes are strongly correlated with users. The beam-user scheduling problem is also considered in a switched-beam based massive MIMO system in [29], where one pre-defined (*fixed*) beam is associated with one user. In [30], a greedy joint scheduling of beams of users is proposed and we will compare our results with it in the simulations. Concerning the literature related to the mathematical treatment adopted in the paper, the Lyapunov-drift optimization framework is attributed to the pioneer work by Neely [33]. The large system deterministic equivalence to derive the downlink channel capacity is related to the celebrated random matrix theory [34]. Furthermore, in the algorithm design, the BCU technique dates back to multi-convex optimization, e.g., in [35], and the approximation factor of the greedy algorithm is related to the submodular set function optimization problem as in [36].

B. Paper Organizations and Notations

The remainder of the paper is organized as follow. In Section II, the system model and channel model are presented, and the problem is formulated. In Section III, the Lyapunov-drift approach is used to design an optimal scheduling policy. In Section IV, a BCU-based scheduling algorithm is described to address the queue weighted sum rate maximization problem. In Section V, a low-complexity greedy algorithm is presented. The simulation results are conveyed in Section VI. Finally, we conclude our work in Section VII.

Throughout the paper, we use boldface uppercase letters, boldface lowercase letters and lowercase letters to designate matrices, column vectors and scalars, respectively. $\boldsymbol{X}^T$ and $\boldsymbol{X}^\dagger$ denotes the transpose and complex conjugate transpose of matrix $\boldsymbol{X}$, respectively. $X_{i,j}$ and x_i denotes the (i, j)-th entry and i-th element of matrix $\boldsymbol{X}$ and vector $\boldsymbol{x}$, respectively. $\mathrm{tr}(\boldsymbol{X})$ denotes the trace of matrix $\boldsymbol{X}$. Denote by $\mathbb{E}(\cdot)$ as the expectation operation. Denote by $\boldsymbol{I}_N$ as the N dimensional identity matrix. The logarithm $\log(x)$ denotes the binary logarithm.

II. System Model and Problem Formulation

A. Signal Model

The single-cell system downlink is considered in this paper, where a BS with M co-located antennas transmits to N_t single-antenna users.[1] The BS has K RF chains ($K \leq M$). Assuming narrow-band and time-invariant channels,[2] the receive signal of user-n is,

$$y_n = \boldsymbol{h}_n^\dagger \boldsymbol{x} + n_n, \tag{1}$$

where $\boldsymbol{h}_n$ is an M-dimensional channel vector, $\boldsymbol{x}$ is the downlink transmit signals, and n_n denotes the i.i.d. Gaussian additive noise with unit variances. The downlink channel matrix is denoted by $\boldsymbol{H} = \left[\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{N_\mathrm{t}}\right]^\dagger$. The transmit signal after beamforming can be written as

$$\boldsymbol{x} = \boldsymbol{B}_\mathrm{a}\boldsymbol{s}, \tag{2}$$

where $\boldsymbol{s}$ denotes the K-dimensional digitally precoded data symbols for the scheduled, i.e., spatial-multiplexed N_s users ($N_\mathrm{s} \leq K$ such that a linear digital precoder such as zero-forcing precoder can eliminate the inter-user-interference,[3] and obviously $N_\mathrm{s} \leq N_\mathrm{t}$). The RF (analog) beamforming, which can be realized by the lens antenna array and beam selection in beamspace MIMO or general analog beamforming in hybrid beamforming architectures, is denoted by $\boldsymbol{B}_\mathrm{a}$ with dimension $M \times K$. On account of the analog beamforming, the effective channel observed from baseband is

$$\bar{\boldsymbol{H}} = \boldsymbol{H}\boldsymbol{B}_\mathrm{a}, \tag{3}$$

where the effective channel vector corresponding to user-n is denoted by $\bar{\boldsymbol{h}}_n$ and $\bar{\boldsymbol{H}} = \left[\bar{\boldsymbol{h}}_1, \ldots, \bar{\boldsymbol{h}}_{N_\mathrm{t}}\right]^\dagger$. The RF beamforming considered in the paper is a widely adopted directional

[1] The proposed schemes can be straightforwardly extended to multiple-antenna-user case by treating multiple antennas of a user as multiple users with an identical channel correlation matrix [25] and setting the backlog pressure $Q_n(t)$ in **P3** to be the same for these antennas.

[2] Wideband channels can be decomposed to a set of parallel narrow-band channels by, e.g., orthogonal-frequency-division-multiplexing (OFDM) modulations. The time-invariant channel assumption essentially deals with data transmission inside the channel coherence time (or block length in the block fading channel model).

[3] In the simulations, we assume $K = N_\mathrm{s}$ and hence $K \leq N_\mathrm{t}$.

beamforming scheme,[4] and hence

$$\boldsymbol{B}_{\mathrm{a}} = \boldsymbol{B}_{\mathrm{DFT}}\boldsymbol{\Sigma}_{\mathrm{b}}, \tag{4}$$

where $\boldsymbol{B}_{\mathrm{DFT}}$ is the equivalent discrete-Fourier-transform (DFT) matrix (or Kronecker product of DFT matrices for uniform planar antenna arrays (UPAs)). The beam selection decision is denoted by the diagonal matrix $\boldsymbol{\Sigma}_{\mathrm{b}}$ whose entries are binary, i.e., $(\boldsymbol{\Sigma}_{\mathrm{b}})_{i,i} \in \{0, 1\}$, $\forall i$.

B. Channel Model

Based on a geometry-based channel model [37], the channel vector of the n-th user can be written as

$$\boldsymbol{h}_n = \sqrt{\frac{M}{L_n}} \sum_{l=1}^{L_n} \beta_l^{(n)} \boldsymbol{\alpha}\left(\theta_l^{(n)}, \psi_l^{(n)}\right), \tag{5}$$

where L_n denotes the total number of multi-path components (MPCs) in the propagation medium including line-of-sight (LoS) and none-line-of-sight (NLoS) MPCs. The amplitude of each MPC is denoted by $\beta_l^{(n)}$, and $\theta_l^{(n)}$ and $\psi_l^{(n)}$ denote the azimuth and elevation angles of the l-th arriving MPC, respectively. Thereby, the steering vector for one MPC is given by (assuming UPA whereas one-dimensional uniform-linear-array (ULA) can be regarded as a special case)

$$\begin{aligned}\boldsymbol{\alpha}^{\mathrm{UPA}}&\left(\theta_l^{(n)}, \psi_l^{(n)}\right) \\ &= \frac{1}{\sqrt{M}}\Bigg[1, \ldots, e^{-j2\pi\left(m\lambda_h \sin\theta_l^{(n)} \cos\psi_l^{(n)} + n\lambda_v \cos\theta_l^{(n)} \sin\psi_l^{(n)}\right)}, \\ &\quad \ldots, e^{-j2\pi\left((H-1)\lambda_h \sin\theta_l^{(n)} \cos\psi_l^{(n)} + (V-1)\lambda_v \cos\theta_l^{(n)} \sin\psi_l^{(n)}\right)}\Bigg]^T,\end{aligned} \tag{6}$$

where H and V denote the number of columns and rows in the UPA, respectively, and λ_h and λ_v are the antenna spacing in the horizontal and vertical domains, respectively. The order of elements in the steering vector is mapped to the indexing order of antennas in the UPA. The physical meaning of the steering vector and channel representation in (5) is that for an MPC with direction-of-arrival (DoA) $\left(\theta_l^{(n)}, \psi_l^{(n)}\right)$, the array response is given by (6). Summing up all the contributing MPCs, we obtain the compound channel representation in (5). Based on the channel model, the RF beamforming in (4) can take advantage of the limited number of MPCs as compared with the number of antennas, and only selects a subset of the beams to attain equivalent performance (signal power) with a smaller number of RF chains.

However, the beam selection cannot be isolated from the user scheduling problem. Apart from the reasons given in Section I, from a throughput perspective, different users have different transmission needs resulting from traffic demand or fairness considerations. Therefore, the user scheduling and beam-domain CSI should also be jointly considered. The joint beam-domain massive MIMO downlink scheduling problem is formulated as follows.

[4]Note that the RF beamforming in this paper can be readily generalized to arbitrary beam pattern, e.g., the eigenvector-based beam pattern in [16], by replacing the DFT-based beamforming matrix B_{DFT} with the desired beamforming matrix. Also note that although directional beamforming is adopted, the proposed schemes can adapt to the channel variation more flexibly compared with traditional directional antenna based systems, by updating the channel statistics estimations and adjusting beam patterns.

C. Problem Formulations

The long-time average rate of user n is denoted by $\bar{R}_n$, and the instantaneous rate of user n at time t is denoted by $R_n(\boldsymbol{H}(t), \pi(t))$, given the channel coefficients $\boldsymbol{H}(t)$ and control policy (user scheduling and beam selection as far as the paper is concerned) $\pi(t)$. Note that this does not mean the scheduling decision relies on the availability of instantaneous CSI, as in Proposition 1 an deterministic equivalence of the rate expression will be derived which is based solely upon statistical CSI. Based on ergodicity, $\bar{R}_n = \mathbb{E}\{R_n(\boldsymbol{H}, \pi)\}$, $\forall n \in \{1, \ldots, N_{\mathrm{t}}\}$, where the expectation is taken over channel coefficients $\boldsymbol{H}(t)$ and possibly $\pi(t)$ when a stochastic control policy is considered. The achievable ergodic rate region can be characterized as

$$\mathcal{R} = \mathrm{coh} \bigcup_{\pi \in \mathcal{X}} \left\{\bar{\boldsymbol{R}} : 0 \le \bar{R}_n \le \mathbb{E}\left[R_n(\boldsymbol{H}, \pi)\right]\right\}, \tag{7}$$

where $\mathcal{R}$ is a N_{t}-dimensional region, $\bar{R}_n$ is its n-th component, and "coh" denotes the closure of a convex hull. The set of all feasible scheduling policies is denoted by $\mathcal{X}$. The utility maximization problem is formulated as

$$\textbf{P1:} \underset{\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_b}{\mathrm{maximize}} U\left(\bar{\boldsymbol{R}}\right) \quad \text{s.t.,}\ \bar{\boldsymbol{R}} \in \mathcal{R}, \tag{8}$$

where $\boldsymbol{\Sigma}_{\mathrm{u}}$ is a diagonal matrix denoting user scheduling decision at time t, i.e., $s_i \triangleq (\boldsymbol{\Sigma}_{\mathrm{u}})_{i,i} \in \{0, 1\}$, and $b_i \triangleq (\boldsymbol{\Sigma}_{\mathrm{b}})_{i,i}$ denotes the beam selection decision. The network utility function $U\left(\bar{\boldsymbol{R}}\right)$ is defined as a function of the long-time average rate for each user, e.g.,

$$U_{\mathrm{sum}}\left(\bar{\boldsymbol{R}}\right) = \sum_{n=1}^{N_{\mathrm{t}}} \bar{R}_n \tag{9}$$

for sum rate maximization,

$$U_{\mathrm{pfs}}\left(\bar{\boldsymbol{R}}\right) = \sum_{n=1}^{N_{\mathrm{t}}} \log(\bar{R}_n + c_n) \tag{10}$$

for proportional-fairness scheduling (PFS) [38], where c_n's are non-negative constants to regularize the logarithm expressions, and typical value is $c_n = 0$, $\forall n \in \{1, \ldots, N_{\mathrm{t}}\}$ for exact PFS or $c_n = 1$, $\forall n \in \{1, \ldots, N_{\mathrm{t}}\}$ to ensure positive objective function value which is a mathematical convenience. Basic properties of the utility function $U\left(\bar{\boldsymbol{R}}\right)$ are required, e.g., concavity and monotonicity [39], over the rate vector $(R_1, \ldots, R_{N_{\mathrm{t}}})$.

III. Optimal Beam-Based Joint Scheduling Policy

To solve **P1**, it is found that two severe challenges exist. First, the ergodic capacity region $\mathcal{R}$ does not yield a closed-form expression. The work in [40] and [41] characterizes the broadcast channel (BC) capacity region and a duality between BC and multiple-access-channel (MAC) in the sense of both capacity region and outage probability is found. Moreover, an iterative water-filling algorithm is proposed to calculate it, given the *instantaneous* channel coefficients. Nevertheless, no closed-form expressions are available except

for capacity bounds [42].[5] Secondly, the scheduling and beam selection decisions should be made *dynamically* to match the channel variations and user traffic in time. To address these issues, we seek to leverage a powerful tool of Lyapunov-drift optimization which is shown to have superior performance compared to static solutions [43] with simple decision structures (max-weight structure [44]); thereby, **P1** is decomposed into **P2** and **P3** described in the following.

Essentially, **P1** is a time-average network utility maximization problem which is hard to solve directly. The Lyapunov-drift approach decomposes the time-average optimization into optimization in each scheduling step; and the resultant subproblems are formulated as **P2** and **P3**. By applying the solutions to **P2** and **P3** at each scheduling step, the time-average network utility can be optimized.

A. Lyapunov-Drift Based Network Utility Maximization

To maximize the network utility function in (8), the transmission need of each user, which is determined by the transmission history and utility function, is represented by a set of virtual queues. The arrival process is designed to reflect the transmission urgency of each user and a max-weight algorithm is applied to stabilize the queues whenever possible. Specifically, let $Q_n(t)$ denote the virtual queue length in bits of user n at the beginning of t-th scheduling step, let $a_n(t)$ denote the number of (virtual) arrival bits which are optimization variables for utility maximization, and let $\mu_n(t)$ denote the allocated number of service bits to queue-n, which equals the allocated number of service bits between scheduling steps. The queuing dynamics are written as

$$Q_n(t+1) = Q_n(t) - \tilde{\mu}_n(t) + a_n(t), \tag{11}$$

where $\mu_n(t) = \sum_{\tau=1}^{T} R_n(\boldsymbol{H}(\tau), \pi(\tau))$, the number of channel uses between scheduling steps is denoted by T, and $\tilde{\mu}_n(t) = \min\{Q_n(t), \mu_n(t)\}$ denotes the number of actual service bits, considering the circumstances that sometimes the queue is emptied given the amount of allocated service bits. Notice that the queues here are created virtually to facilitate the utility maximization and thus they are not real traffic patterns. In Section VI (Fig. 5), we extend to stochastic real traffic scenarios in simulations. The optimal beam-based downlink scheduling policy at a given scheduling time t, i.e., a dynamic policy which achieves the solution to (8), can be described as below.

Admission Control: For virtual queue $\boldsymbol{Q}(t) = [Q_1(t), \ldots, Q_{N_t}(t)]$, let the number of arrival bits, i.e., $\boldsymbol{a}(t)$, be the solution of

$$\begin{aligned} \textbf{P2:} \quad & \underset{\boldsymbol{a}(t)}{\text{maximize}} \; V U(\boldsymbol{a}(t)) - \boldsymbol{a}(t)^T \boldsymbol{Q}(t), \\ & \text{s.t., } 0 \le a_n(t) \le A_{\max}, \quad \forall n \in \{1, \ldots, N_t\}, \end{aligned} \tag{12}$$

where V and $A_{\max}$ are pre-defined constants.[6] □

[5]The broadcast channel capacity formula is adopted as the optimization objective in this paper due to its better generality compared with, e.g., achievable rates based on linear precoding schemes. It is also because that non-linear downlink transmission schemes, e.g., non-orthogonal multiple access schemes, are attracting more and more attention recently.

[6]For Typical values, V and $A_{\max}$ can be approximately 100-fold of the service rate.

Scheduling: Given the arrival process determined above, the service, i.e., the joint scheduling decisions, is based on the solution of the following problem:

$$\textbf{P3:} \quad \underset{\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_b, \boldsymbol{p}}{\text{maximize}} \; \sum_{n=1}^{N_t} \left[Q_n(t) s_n \sum_{\tau=1}^{T} R_n(\boldsymbol{H}(\tau), \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_b, \boldsymbol{p}) \right] \tag{13}$$

$$\text{s.t., } \sum_{n=1}^{N_t} p_n \le P, \quad \sum_{i=1}^{N_t} s_i = N_s, \quad \sum_{i=1}^{M} b_i = K, \tag{14}$$

$$s_i \in \{0, 1\}, \quad b_i \in \{0, 1\}, \tag{15}$$

where $\boldsymbol{p} = [p_1, \ldots, p_{N_t}]$ denotes the transmit power corresponding to N_t user data streams and hence P in (14) is the sum power constraint. The scheduling decisions are denoted by binary variables s_i and b_i. $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_u$ are diagonal matrices consisting of s_i and b_i, respectively. The downlink instantaneous transmission rate $R_n(\boldsymbol{H}(t), \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_b, \boldsymbol{p})$ is a function of the downlink transmit power allocation, channel coefficients, and scheduling decisions. The departure from the n-th virtual queue is $\mu_n(t) = s_n \sum_{\tau=1}^{T} R_n(\boldsymbol{H}(\tau), \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_b, \boldsymbol{p})$. □

It is observed that the admission control problem **P2** is a convex problem and hence is easy to solve. For instance with PFS utility, the optimal admission control is given by

$$a_n^*(t) = \min\left\{ \frac{V}{Q_n(t)}, A_{\max} \right\}, \quad n \in \{1, \ldots, N_t\}. \tag{16}$$

However, the problem **P3** is an MIP problem, which is NP-complete [45]. Before diving into details on solving **P3** in the following sections, we assume the optimal solutions to both problems are obtained for the moment, which is denoted by π^*. The optimality of the algorithm is established in the following theorem.

Theorem 1: Denote

$$\bar{\boldsymbol{R}}^* \triangleq \underset{\bar{\boldsymbol{R}} \in \mathcal{R}}{\arg\max}\, U\left(\bar{\boldsymbol{R}}\right). \tag{17}$$

Suppose the transmission rate is bounded, i.e., $0 \le \bar{R}_n \le R_{\max}$, $\forall n \in \{1, \ldots, N_t\}$, the utility function $U(\cdot)$ is concave and entry-wise non-decreasing, and bounded on $[0, R_{\max}]$. The channel coefficients $\boldsymbol{H}(t)$ are i.i.d. over different scheduling periods, then based on the scheduling algorithm resulting from **P2** and **P3**, the following conditions are met.

$$\liminf_{\tau \to \infty} U\left(\frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}[\boldsymbol{R}(t)] \right) \ge U\left(\bar{\boldsymbol{R}}^*\right) - C/V, \tag{18}$$

$$\lim_{\tau \to \infty} \frac{\mathbb{E}[Q_n(\tau)]}{\tau} = 0, \quad \forall n. \tag{19}$$

□

Proof: See Appendix A. ■

Remark 1: Theorem 1 reveals that the utility function of the time-averaged transmission rate based on the scheduling decisions derived in **P2** and **P3** is within a constant (arbitrary small if V is large) to the optimum and the virtual queues are mean-rate-stable, where C is a constant related to $A_{\max}$ (41). In the following, we will dig into the methods to solve **P3**.

IV. Block Coordinate Update Based Method for P3

This section is dedicated to solving the scheduling problem of **P3** only based on the knowledge of statistical CSI. The previous section establishes the optimality of the proposed beam-based scheduling algorithm given the solutions of **P2** (generally easy to solve) and **P3**. However, due to the NP-hardness of **P3**, explicit solutions are hard to attain. More importantly, it is proposed that the scheduling decisions of **P3** should only rely on statistical CSI, rendering the solution even more intractable. Towards this end, an algorithm based on solving the convex relaxation of the original problem leveraging the BCU technique and random matrix theory is proposed. First, **P3** is transformed for better exposition based on the uplink-downlink duality [40], [46]. The instantaneous achievable rate in **P3** is evaluated by the MIMO broadcast channel capacity. The following Proposition 1 derives an implicit asymptotic expression of the objective function in **P3** such that the optimization is only dependent on statistical CSI which in this case is the channel correlation matrices.

Proposition 1: In the large system regime, i.e., $K \to \infty$ and $K/N_{\rm s} \to \beta$, the queue-weighted downlink channel capacity in **P3** is asymptotically equivalent to

$$\mathcal{D}(\boldsymbol{Q}, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{N_{\rm t}}, \boldsymbol{\Sigma}_{\rm u}, \boldsymbol{\Sigma}_{\rm b}, \boldsymbol{p}) \triangleq \sum_{n=1}^{N_{\rm t}} q_n T \log\left(1 + p_n s_n \mathrm{tr}\left[\boldsymbol{\Sigma}_{\rm b}\boldsymbol{B}_{\rm DFT}^{\dagger}\boldsymbol{R}_n\boldsymbol{B}_{\rm DFT}\boldsymbol{\Sigma}_{\rm b} \cdot \left(\frac{1}{M}\sum_{j=1}^{n-1}\frac{p_j s_j \boldsymbol{\Sigma}_{\rm b}\boldsymbol{B}_{\rm DFT}^{\dagger}\boldsymbol{R}_j\boldsymbol{B}_{\rm DFT}\boldsymbol{\Sigma}_{\rm b}}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right]\right), \tag{20}$$

where q_i's are arranged in non-increasing order, and $e_{n,i}$ is the unique solution of the following equations.

$$e_{n,i} = \mathrm{tr}\left[\boldsymbol{\Sigma}_{\rm b}\boldsymbol{B}_{\rm DFT}^{\dagger}\boldsymbol{R}_i\boldsymbol{B}_{\rm DFT}\boldsymbol{\Sigma}_{\rm b} \cdot \left(\frac{1}{M}\sum_{j=1}^{n-1}\frac{p_j s_j \boldsymbol{\Sigma}_{\rm b}\boldsymbol{B}_{\rm DFT}^{\dagger}\boldsymbol{R}_j\boldsymbol{B}_{\rm DFT}\boldsymbol{\Sigma}_{\rm b}}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right]. \ \square \tag{21}$$

Proof: See Appendix B. ■

A. Convex Relaxation

Although the original MIP **P3** is NP-hard, it can be relaxed to a multi-convex problem by replacing the binary constraints with real-value constraints. The convex relaxation of an MIP is a widely-used technique to achieve near-optimal solutions to the original problem [47], [48]. The relaxed version of **P3** is stated below.

$$\textbf{P4:} \quad \underset{\boldsymbol{\Sigma}_{\rm b}, \boldsymbol{w}}{\text{maximize}} \ \mathcal{D}(\boldsymbol{Q}, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{N_{\rm t}}, \boldsymbol{I}_{N_{\rm t}}, \boldsymbol{\Sigma}_{\rm b}, \boldsymbol{w}) \tag{22}$$

$$\text{s.t.}, \ \sum_{n=1}^{N_{\rm t}} w_n \le P, \quad \sum_{i=1}^{M} b_i = K, \ 0 \le b_i \le 1, \ \forall i, \tag{23}$$

where $w_n = p_n s_n$, and $\mathcal{D}$ is defined in (20). Define the optimum solution of **P4** as b_i^*'s and w_n^*'s, respectively. Then the scheduling decision is to schedule the beams and users corresponding to the largest K b_i^*'s and $N_{\rm s}$ w_n^*'s, respectively. It is observed that **P4** is a multi-convex problem [35] since the objective function is concave in both $\boldsymbol{\Sigma}_{\rm b}$ and $\boldsymbol{w}$. In view of this, the Algorithm 1 which bases upon the BCU technique is proposed.

Algorithm 1 BCU-Based Scheduling

Initialization: $\boldsymbol{\Sigma}_{\rm b}^{(0)} = \boldsymbol{I}_M$;
Iteration: **for** $t = 1 : T$ **do**
  User scheduling update based on iterative water filling: $\forall n \in [1, N_{\rm t}]$, $\omega_n^{(0)} = P/N_{\rm t}$; **for** $t_{\rm w} = 1 : T_{\rm w}$ **do**
    Compute for each n,

$$\beta_n^{(t_{\rm w})} = \mathrm{tr}\left[\boldsymbol{\Sigma}_{\rm b}^{(t-1)}\boldsymbol{B}_{\rm DFT}^{\dagger}\boldsymbol{R}_n\boldsymbol{B}_{\rm DFT}\boldsymbol{\Sigma}_{\rm b}^{(t-1)} \cdot \left(\frac{1}{M}\sum_{j=1}^{n-1}\frac{w_j^{(t_{\rm w}-1)}\boldsymbol{R}_j}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right], \tag{24}$$

    where $e_{n,i}$ is the unique solution of the equations in (21);
    Apply the classical water filling algorithm with water levels defined by $\boldsymbol{\beta}^{(t_{\rm w})}$

$$\boldsymbol{\gamma}^{(t_{\rm w})} = \underset{\sum_{n=1}^{N_{\rm t}}\gamma_n \le P, \boldsymbol{\gamma}\ge\boldsymbol{0}}{\arg\max} \sum_{n=1}^{N_{\rm t}} q_n \log\left(1 + \gamma_n \beta_n^{(t_{\rm w})}\right)$$

    Update $\boldsymbol{\omega}$ as
    $\boldsymbol{\omega}^{(t_{\rm w})} = (1 - 1/M)\,\boldsymbol{\omega}^{(t_{\rm w}-1)} + (1/M)\boldsymbol{\gamma}^{(t_{\rm w})}$
    if $\|\boldsymbol{\omega}_n^{(t_{\rm w})} - \boldsymbol{\omega}_n^{(t_{\rm w}-1)}\| < \epsilon$ **then**
      $\boldsymbol{w}^{(t)} = \boldsymbol{\omega}^{(t_{\rm w})}$, break;
  Beam selection: Solve for $\boldsymbol{\Sigma}_{\rm b}^{(t)}$, which is the solution to the convex optimization problem of **P4** with $\boldsymbol{w} = \boldsymbol{w}^{(t)}$.
  Stopping criterion: **if** $\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\| < \epsilon_1$ *and* $\|\boldsymbol{\Sigma}_{\rm b}^{(t)} - \boldsymbol{\Sigma}_{\rm b}^{(t-1)}\| < \epsilon_2$ **then**
    $\boldsymbol{w}_{\rm opt} = \boldsymbol{\omega}^{(t)}$, $\boldsymbol{\Sigma}_{\rm b,opt} = \boldsymbol{\Sigma}_{\rm b}^{(t)}$, break;
Output: The scheduling user set is the users with the largest $N_{\rm t}$ values in $\boldsymbol{w}_{\rm opt}$. The selected beams are the ones with the largest K values in the diagonal entities of $\boldsymbol{\Sigma}_{\rm b,opt}$.

The basic idea of the proposed BCU-based user and beam joint scheduling is that an iterative method which cyclically optimizes user scheduling and beam selection with the other fixed is guaranteed to converge to the global optimum of **P4**. In order to accelerate the iteration, an iterative water filling approach which is based on [49] and deals with user scheduling is adopted.

Convergence of the Proposed Algorithm: The convergence to the global optimum is due to the convergence results of the BCU algorithm [50]. The details of the proof is given in Appendix C. It is found through simulations that the BCU-based algorithm normally converges after 2-3 iterations. Therefore, the computational complexity and convergence time are acceptable in simulated scenarios.

Remark 2: The Algorithm 1 can solve the joint user scheduling and beam selection problem based on statistical CSI. Therefore, it is applicable before channel estimations. After the system selects users and beams, the instantaneous channel estimations can be implemented and subsequently digital precoding and decoding can follow. This is in line with the multi-layer signal processing concept proposed in [5] and [16], which proposes that the pre-beamforming should be done based on channel statistics to save RF chains, complexity and system overhead.

V. Incremental Selection Based Method for P3

Although the BCU-based algorithm is guaranteed to converge to the optimum of the relaxed convex optimization problem, it still has high complexity due to the iterative algorithm design. Therefore it may take a long time to converge. In this regard, an algorithm which selects users and beams incrementally with low complexity is proposed. The key to the design of the algorithm is to derive the incremental selection criterion. Thanks to the results in Proposition 1, the structure of the asymptotic rates in (20) can be utilized to give such a criterion. Thereby, the incremental selection algorithm is described in Algorithm 2, which assumes $K = N_{\mathrm{s}}$.[7]

Remark 3: Due to the successive interference cancellation (SIC) structure in the broadcast channel queue-weighted capacity expression in Proposition 1, the rates of the users decoded (selected in IGS) first will not be affected by the users decoded (selected in IGS) later. Therefore, the IGS algorithm is viable because the earlier decisions are decoupled from later ones.

Complexity Analysis: The IGS has a complexity of $\mathcal{O}(N_{\mathrm{s}}N_{\mathrm{t}}M)$, because in each step it involves an exhaustive search over $\mathcal{O}(MN_{\mathrm{t}})$ possible user and beam combination, and there are N_{s} iterations. Compared with an exhaustive search over all user and beam subsets with complexity of $\binom{N_{\mathrm{t}}}{N_{\mathrm{s}}}\binom{M}{K}$, and the BCU-based scheduling which is difficult to quantize the complexity due to the iterative design of the algorithm, the IGS has a relatively very low complexity.

Due to the greedy nature of the IGS algorithm, it may fail to find the optimum sets of users and beams to schedule. Hence, it is better to have a *worst-case performance bound* of the IGS, such that potentially arbitrarily bad solutions are excluded. Fortunately, this is the case for the proposed IGS, due to the submodularity property [36] of the problem.

[7]The assumption is justified by arguing that maximum degree-of-freedom is achieved with $K = N_{\mathrm{s}}$.

Algorithm 2 Incremental Greedy Scheduling (IGS)

Initialization: $\mathsf{U} = \mathsf{B} = \emptyset$; $\mathsf{U}_{\mathrm{all}} = [1 : N_{\mathrm{t}}]$, $\mathsf{B}_{\mathrm{all}} = [1 : M]$

Incremental Selection: **for** $t = 1 : N_s$ **do**

Find the user-n_t and the beam-b_t that maximize:

$$(n_t, b_t) = \underset{n\in\mathsf{U}_{\mathrm{all}}\backslash\mathsf{U}, b\in\mathsf{B}_{\mathrm{all}}\backslash\mathsf{B}}{\arg\max} q_n \log\left(1 + \frac{P}{N_{\mathrm{s}}}\mathrm{tr}\left[\boldsymbol{\Sigma}_b\bar{\boldsymbol{R}}_n\boldsymbol{\Sigma}_b \cdot \left(\frac{1}{M}\sum_{j\in\mathsf{U}}\frac{\frac{P}{N_{\mathrm{s}}}\boldsymbol{\Sigma}_b\bar{\boldsymbol{R}}_j\boldsymbol{\Sigma}_b}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right]\right)$$

where $\bar{\boldsymbol{R}}_j = \boldsymbol{B}_{\mathrm{DFT}}^{\dagger}\boldsymbol{R}_j\boldsymbol{B}_{\mathrm{DFT}}$,

$$(\boldsymbol{\Sigma}_b)_{i,i} = \begin{cases} 1, & \text{for } i \in \mathsf{B}\cup\{\mathrm{b}\} \\ 0, & \text{else,} \end{cases}$$

and

$$e_{n,i} = \mathrm{tr}\left[\boldsymbol{\Sigma}_b\bar{\boldsymbol{R}}_i\boldsymbol{\Sigma}_b\left(\frac{1}{M}\sum_{j\in\mathsf{U}}\frac{\frac{P}{N_{\mathrm{s}}}\boldsymbol{\Sigma}_b\bar{\boldsymbol{R}}_j\boldsymbol{\Sigma}_b}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right].$$

Update: $\mathsf{U}\cup\{n_t\}\rightarrow\mathsf{U}$, $\mathsf{B}\cup\{b_t\}\rightarrow\mathsf{B}$

Output: The scheduling user set is U, and the selected beam set is B.

Theorem 2: Denote the queue weighted sum rate achieve by the IGS algorithm as D_{IGS}, and the global optimum as D_{opt}, then it is satisfied that

$$D_{\mathrm{IGS}} \geq (1 - e^{-1})D_{\mathrm{opt}}.\square \tag{25}$$

Proof: The proof is based on the submodularity of the queue-weighted sum rate maximization problem in **P3**. Informally, the submodularity property indicates the problem has diminishing returns, i.e., in this case the sum rate increase by scheduling a user or a beam is larger when scheduled with a smaller user/beam set, i.e.,

$$D(\mathsf{U}_1\cup u) - D(\mathsf{U}_1) \geq D(\mathsf{U}_2\cup u) - D(\mathsf{U}_2), \tag{26}$$

for any $u \in \mathsf{U}_{\mathrm{all}}\backslash(\mathsf{U}_1\cup\mathsf{U}_2)$ and $\mathsf{U}_1 \subseteq \mathsf{U}_2$. This is easily validated since the same user will suffer from more interference with a larger scheduled user set. There are two additional conditions for submodularity, which is that the function should be *nondecreasing* and *nonnegative*. The nondecreasing property, i.e.,

$$D(\mathsf{U}\cup u) \geq D(\mathsf{U}), \quad \forall u \in \mathsf{U}_{\mathrm{all}}\backslash\mathsf{U}, \tag{27}$$

can be proved by arguing that at least, zero power can be allocated to the user or beam to obtain equal performance without the user or beam since the objective function is a maximization over all user and beam selection schemes.

It should be noted that although the nonnegative condition is easily validated, e.g., for the sum rate maximization or max-min rate maximization, it is not met exactly for the PFS sum logarithm rate optimization. However, if we fix $c_n = 1$, $\forall n$ in (10), the objective function is non-negative and thus the submodularity property is upheld.

Based on the submodularity, it can be proved that the IGS achieves a near-optimum performance. The remaining details of the proof is well-known in [36] and [51] and therefore omitted for brevity. ■

Remark 4: Although with Theorem 2, it is only proved that the IGS achieves at least about 60% throughput of the optimum scheme, the performance in reality is much better than that, as which will be shown by simulations. Existing work which also utilizes the greedy algorithm with submodularity property also agrees with this finding [51].

VI. Simulation Results

In this section, simulation results are presented. The channel model is as described in Section II-B, where the number of MPCs for each user is $L_n = 3$, $\forall n$ (including one LoS MPC), unless stated otherwise. The ULA is used in the simulations and the amplitude of the LoS MPC is 10 times the one of the NLoS MPCs. The DoAs of the signals are generated from i.i.d. uniform distributions. The antenna spacing $d = \lambda/2$, where λ denotes the carrier wavelength. In some of the following cases where users' pathlosses are not identical, the distances of users are generated based on an i.i.d. uniform distributions from 30 to 200 meters and the pathloss γ_n is

$$\gamma_n = \left(\frac{d_n}{d_0}\right)^{-\gamma}, \tag{28}$$

where $\gamma = 2$ which is in line with mm-wave channel measurements [52] and d_0 is some reference point distance. The regularized zero-forcing (RZF) precoder is adopted for system evaluation, i.e., define $\boldsymbol{K}_{\text{rzf}} = \left(\bar{\boldsymbol{H}}^{\dagger}\bar{\boldsymbol{H}} + M\alpha \boldsymbol{I}_M\right)^{-1}$. The RZF precoding matrix is expressed as $\boldsymbol{B}_{\text{d}} = \zeta \boldsymbol{K}_{\text{rzf}}\bar{\boldsymbol{H}}^{\dagger}$, where ζ is a normalization scalar to fulfill the power constraint, and α is the regularization factor [53]. Although RZF precoder is not the optimal coding scheme for Gaussian broadcast channel (dirty-paper-coding with minimum-mean-square-error (MMSE) precoder is proved for optimality but limited in reality due to high complexity), it can achieve full degree-of-freedom (DoF) in the high signal-to-noise-ratio (SNR) region and is easy to implement. In the simulations, $\alpha = N_{\text{s}}/\rho$, where ρ is the receive SNR [53]. The user instantaneous rate is calculated by the Shannon formula. The block fading model is adopted, where the channel stays constant for 10 time slots and evolves to another realization based on an i.i.d. distribution. The phase and amplitude of each MPC is generated randomly. The simulation runs for 1000 such blocks and calculate the time-averaged downlink transmission rates. The constants used in the Lyapunov-drift optimization are set to be $V = A_{\max} = 10^2\, r_{\max}$, where $r_{\max}$ is the maximum rate of the users. The ϵ, ϵ_1 and ϵ_2 in the stopping criterion in BCU-based algorithm are set to be $10^{-2}\rho/K$. In comparisons, the state-of-the-art interference aware beam selection scheme (IA-BS) in [23], and the BDMA scheme in [30] are also simulated. These schemes for comparisons are shown to perform better than several other existing schemes, e.g., [24].

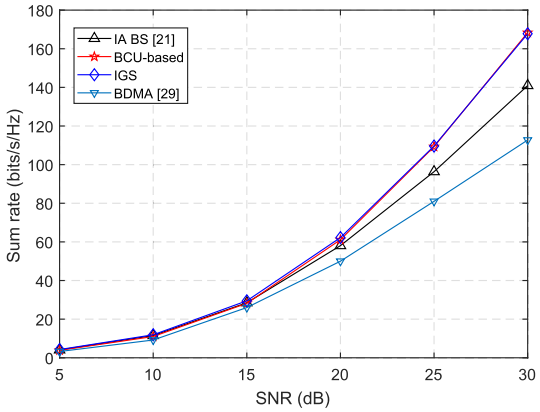


Fig. 2. Sum rate comparisons with identical user pathloss. The number of BS antennas is 256, the number of users is 100, the number of scheduled users and beams are both 40.

In Fig. 2, our proposed BCU-based algorithm and the IGS algorithm are compared with the IA-BS algorithm [23] and the BDMA scheme [30]. Since the sum rates are considered, the utility function in (9) is adopted. The scheduled user set in IA-BS is assumed to have the most channel power. It is observed that by jointly considering user scheduling and beam selection, the sum rate performance can be improved in the high-SNR regime. The reason is that in the high-SNR regime, the interference is dominating the performance, and thus by jointly considering the user scheduling and beam selection by the proposed schemes, the interference is better suppressed. The BDMA scheme simply adopts a sequential approach which selects the beam-user pairs incrementally, and it builds on optimizing a sum-rate upper bound; both factors lead to performance degradation. Thus, the resultant performance is not as good. Nonetheless, it should be noted that the BDMA scheme is designed for multiple-antenna users and hence the interference can be suppressed thereby whereas such effects are not captured in the presented simulations. Therefore, we focus on the IA-BS scheme for comparisons in the following. Furthermore, the figure also shows that the BCU-based and IGS algorithms achieve very similar performance in this scenario.

Considering the utility functions with user-fairness considerations, e.g., the PFS utility function in (10), the performance advantage of the proposed schemes is more obvious as shown in Fig. 3 (left). Based on the IA-BS, the beams with stronger channels are always preferred, corresponding to users with small pathloss, resulting in ignorance of the fairness among users. In the proposed joint scheduling schemes, the admission control in **P2** utilizes a virtual queue to control the fairness among users. Note that since there are always unscheduled users in the IA-BS due to their small pathloss, the sum log rate of the IA-BS scheme in this case is negative infinity. In Fig. 3 (right), the performance with identical pathloss is presented. Due to the fact that users have identical large-scale fading, and hence equal probability to have good channels, the IA-BS can achieve reasonably good fairness performance.

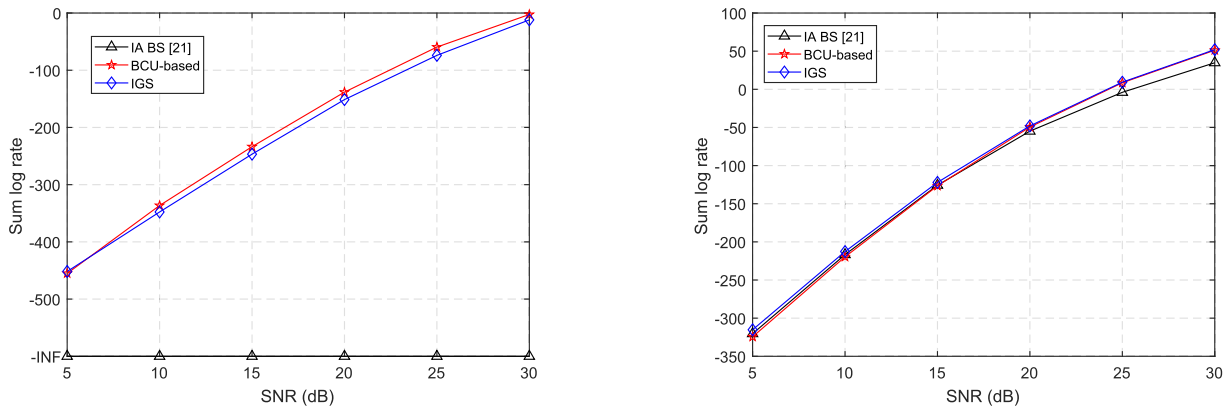


Fig. 3. Sum log rate comparisons with (left) and without (right) non-identical user pathloss. The users with non-identical pathloss are generated with distances i.i.d. uniformly distributed between 20 m to 200 m. The number of BS antennas is 256, the number of users is 100, the number of scheduled users and beams are both 40.

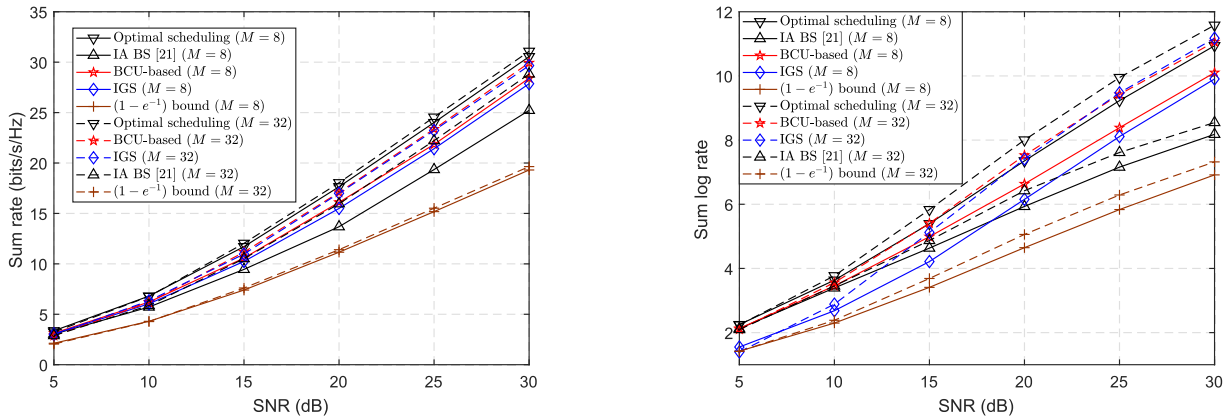


Fig. 4. Comparisons with optimal scheduling. The number of users is 8, the number of scheduled users and beams are both 4. $c_n = 1$, $\forall n$ as in (10).

In order to show the performance loss of our proposed schemes compared with optimal scheduling, an exhaustive search over all the feasible user and beam sets is conducted to solve **P3** and the optimal scheduling performance is obtained accordingly. Due to the prohibitive high complexity of exhaustive search, we consider a small-scale problem where there are 8 users, 8 BS antennas and 4 scheduled users and beams. Nonetheless, it is found by many existing works, e.g., [54], [55], that the impact of imperfect downlink scheduling decreases with the increase of antenna dimension due to the channel hardening effect. Therefore, the relative performance gap with a larger system dimension should be smaller, or at least similar with that in Fig. 4. In Fig. 4, the left and right figures show sum-rate and sum-log-rate optimizations, respectively. It is observed that in general the proposed schemes can achieve near-optimal performance. The BCU-based scheme is shown to have better performance compared with the IGS, but with higher complexity. The performance bound which we prove for the IGS algorithm is also plotted in the figure. The IGS scheme always performs better than the $(1 - 1/e)$ bound. In the low SNR regime, it is observed that the IA-BS scheme outperforms the IGS scheme, due to the reason that the IA-BS scheme always selects the user and its corresponding beam which have the strongest channel, and that in Fig. 4 we set $c_n = 1$ in the log rate to ensure positive utilities and thus less penalty on the unfairness among users is accounted for.

A throughput comparison with a stochastic traffic model is carried out and shown in Fig. 5 and 6. Instead of considering full-buffer greedy transmitting sources, each user's arrival traffic is modeled as a Bernoulli process, i.e., the arrival process for user-n

$$\alpha_n = b_n r_c, \quad \forall n, \tag{29}$$

where b_n is i.i.d. Bernoulli distributed with expected mean values of p_n, r_c is a constant which denotes approximately the service rate of each user, and $r_c = \frac{N_s}{N_t} \log(1 + \eta \frac{\rho}{N_s})$ where η denotes the approximate SNR loss coefficient introduced by interference. Therefore, based on this setting, the mean values of b_n can be regarded as the traffic intensity where $p_n = 0$ denotes zero traffic, and p_n close to 1 denotes heavy traffic. The queuing dynamic is, with slightly abusing the notations,

$$q_n(t+1) = q_n(t) - \tilde{\mu}_n(t) + \alpha_n(t), \tag{30}$$

where $\tilde{\mu}_n(t) = \min\{q_n(t), \mu_n(t)\}$ denotes the actual service rate taken into account of empty queues. The throughput is calculated by averaging the sum actual service rate of each user $\tilde{\mu}_n(t)$.

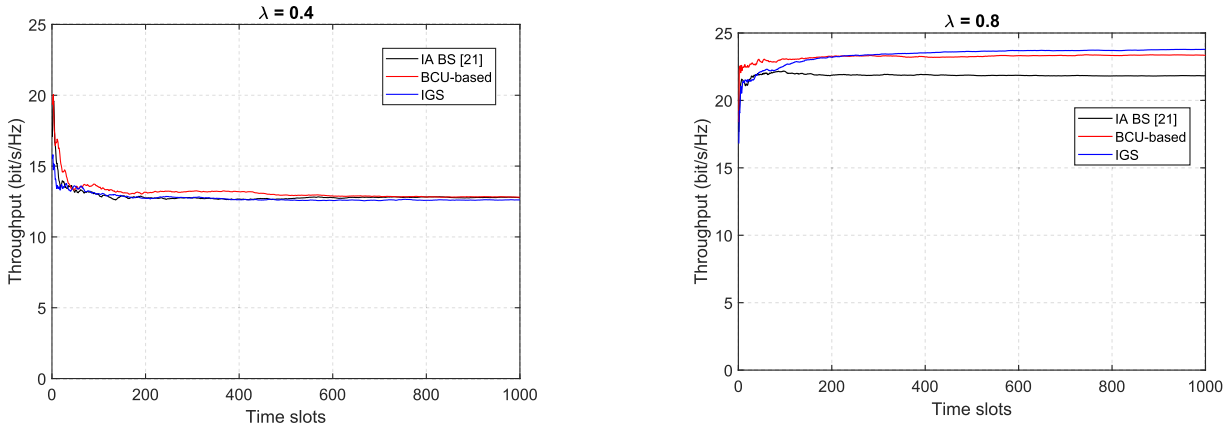


Fig. 5. Sample paths of the system average throughput evaluations. The number of BS antennas is 64, the number of users is 40, the number of scheduled users and beams are both 20. $\eta = 0.4$.

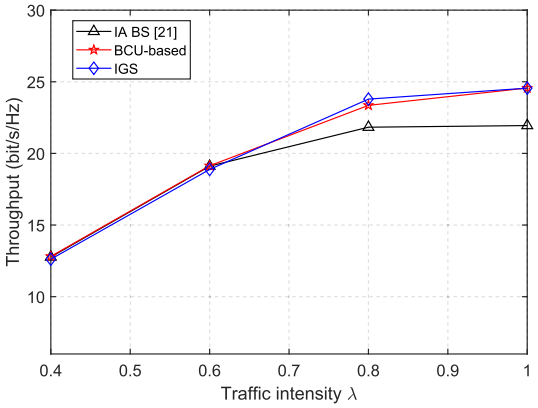


Fig. 6. The number of BS antennas is 64, the number of users is 40, the number of scheduled users and beams are both 20. $\eta = 0.4$.

Our proposed Lyapunov-drift based schemes, i.e., BCU-based and IGS, can be easily adapted to this traffic model, by replacing the virtual queues in (11) with (30) and eliminating the admission control step in **P2**. We assume the IA-BS scheme [23] schedules N_s users with the N_s-largest queue lengths and selects beams accordingly, which is in line with the methodology that upper layers, e.g., medium-access-control (MAC) layer, schedules some users and push the bits to physical layer. In comparisons, the proposed schemes jointly considers user scheduling with traffic demands and beam selection. One sample path of the system is depicted in Fig. 5, where the left and right sub-figures denote relatively low and heavy traffic, respectively. The system average throughput is stationary after about 100 time slots, which equals about 100 ms given the LTE numerologies where one time slot (transmission-time-interval) is one subframe (1 ms). With this convergence time, it is found that the proposed schemes can effectively converge to a reasonably good solution before the statistical CSI varies (usually at a time scale of several seconds). In Fig. 6, the average throughput is compared among different scheduling schemes under different traffic intensities. It is observed that the proposed schemes outperform the IA-BS scheme when the system is with high traffic load. Note that when the system is not fully loaded, i.e., traffic intensity is lower than about 0.6, the average throughput equals the sum arrival rate and thus the performance advantage of the proposed schemes have not emerged. It is shown that joint considerations of user scheduling and beam selection leads to better system throughput.

VII. Conclusions

In this paper, the BCU-based scheduling scheme and the IGS scheme are proposed to address the join user scheduling and beam selection optimization problem in beam-based massive MIMO systems based only on statistical CSI. The problem is formulated under the Lyapunov-drift optimization framework. In order to solve the weighted rate maximization problem therein, the proposed BCU-based scheme leverages the convex relaxation of the problem and adopts the BCU technique with the iterative water-filling approach. It is proved that the BCU-based scheduling scheme iteratively converges to the optimum of the relaxed problem. Due to its iterative algorithm structure, relatively high complexity is required. Towards this end, the IGS algorithm is proposed which is based on a greedy approach. Nevertheless, it is proved that the IGS scheme can achieve performance within a multiplicative factor of $(1 - e^{-1})$ to the optimum. In simulations, it is shown that the proposed schemes can achieve near-optimal performance and outperform the state-of-the-art beam selection schemes, with utilities such as sum rate and proportional fairness. While existing works focuses on the beam selection, which effectively strive to maximize the sum rate performance, they are not optimized when the user scheduling and beam selection are jointly considered especially when user fairness is taken into account. The performance bound we derive for the IGS scheme is also shown to be well observed.

Appendix A
Proof of Theorem 1

We first briefly review the Lyapunov-drift approach, which is the main mathematical tool in our proof. Define the

Lyapunov function as

$$L(t) \triangleq \frac{1}{2}\sum_n Q_n^2(t), \tag{31}$$

and the Lyapunov drift as

$$\Delta(t) \triangleq \mathbb{E}[L(t+1) - L(t)|\,\boldsymbol{Q}(t)]. \tag{32}$$

Lemma 1: If there exist constants B and ϵ, which satisfy

$$\Delta(t) \le B - \epsilon\sum_n Q_n(t), \tag{33}$$

then we have:

1) If $\epsilon \ge 0$, then all queues are mean rate stable.
2) If $\epsilon > 0$, then

$$\limsup_{\tau\to\infty}\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\sum_n Q_n(t)\right] \le \frac{B}{\epsilon}, \tag{34}$$

and hence all queues are strongly stable. □

Proof: The proof of Lemma 1 follows the standard procedure as in [33]. ■

Given the queuing dynamics (11) and based on the definition in (32), we have

$$\begin{aligned}
\Delta(t) &\le \mathbb{E}\left[\frac{1}{2}\sum_n\left(Q_n^2(t+1) - Q_n^2(t)\right)\Big|\,\boldsymbol{Q}(t)\right]\\
&= \mathbb{E}\left[\frac{1}{2}\sum_n\left((Q_n(t) - \tilde{\mu}_n(t) + a_n(t))^2 - Q_n^2(t)\right)\Big|\,\boldsymbol{Q}(t)\right]\\
&= \mathbb{E}\left[\sum_n\left(\frac{1}{2}\tilde{\mu}_n^2(t) - \tilde{\mu}_n(t)Q_n(t) + \frac{1}{2}a_n^2(t)\right.\right.\\
&\quad\left.\left. + Q_n(t)a_n(t) - a_n(t)\tilde{\mu}_n(t)\right)\Big|\,\boldsymbol{Q}(t)\right]\\
&\le \mathbb{E}\left[\sum_n\left(\frac{1}{2}\mu_n^2(t) - \mu_n(t)Q_n(t) + \frac{1}{2}a_n^2(t)\right.\right.\\
&\quad\left.\left. + Q_n(t)a_n(t)\right)\Big|\,\boldsymbol{Q}(t)\right]\\
&= \mathbb{E}\left[\sum_n\frac{\mu_n^2(t) + a_n^2(t)}{2}\Big|\,\boldsymbol{Q}(t)\right] - \sum_n Q_n(t)\mathbb{E}[\mu_n(t)\\
&\quad - a_n(t)|\,\boldsymbol{Q}(t)].
\end{aligned} \tag{35}$$

Observing that

$$\mathbb{E}\left[\sum_n\frac{\mu_n^2(t) + a_n^2(t)}{2}\Big|\,\boldsymbol{Q}(t)\right] \le \frac{T^2}{2}\left[r_{n,\max} + A_{\max}^2\right] \triangleq B, \tag{36}$$

where $r_{n,\max} = \log\left(1 + \|\boldsymbol{h}_n\|^2\,P\right)$ denotes the maximum transmission rate in one channel use since $r_{n,\max}$ is the channel capacity as if the user n was alone, it follows that

$$\Delta(t) \le B - \sum_n Q_n(t)\mathbb{E}\left[\mu_n(t) - a_n(t)|\,\boldsymbol{Q}(t)\right]. \tag{37}$$

Therefore, for any arrival rate inside the ergodic capacity region, since the scheduling problem in **P3** minimize the right-hand side of (37), the condition in Lemma 1 (33) is upheld with $\epsilon \ge 0$, i.e., all queues are mean rate stable. In order to show the throughput-optimality in (18), subtract a term related to the utility function,

$$\begin{aligned}
&\Delta(t) - V\mathbb{E}\left[U(\boldsymbol{a}(t))|\,\boldsymbol{Q}(t)\right]\\
&\le B - \underbrace{\sum_n Q_n(t)\mathbb{E}\left[\mu_n(t)|\,\boldsymbol{Q}(t)\right]}_{\textbf{Scheduling}}\\
&\quad + \underbrace{\mathbb{E}\left[\sum_n Q_n(t)a_n(t) - VU(\boldsymbol{a}(t))\Big|\,\boldsymbol{Q}(t)\right]}_{\textbf{Admission control}},
\end{aligned} \tag{38}$$

It is observed that the admission control and scheduling problems in **P2** and **P3** are equivalent to minimize the related terms in (38) as labeled. Therefore, given the solution to **P2** and **P3**, the left-hand side of (38) is less than the term on the right-hand side with any queue-independent scheduling and admission control. Concretely,

$$\begin{aligned}
\Delta(t) - V\mathbb{E}\left[U(\boldsymbol{a}(t))|\,\boldsymbol{Q}(t)\right] \le B - \sum_n Q_n(t)\bar{R}_n\\
+ \sum_n Q_n(t)z_n - VU(\boldsymbol{z}),
\end{aligned} \tag{39}$$

where $\bar{R}_n$ and z_n denotes any queue-independent service rate and admission rate, respectively. Taking expectations on both sides over $\boldsymbol{Q}(t)$, and taking the telescoping sum yields (assuming $\boldsymbol{Q}(0) = 0$ for better exposition),

$$\begin{aligned}
&\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\left[Q_n(t)\right](\bar{R}_n - z_n)\\
&\qquad \le B + V\left(U\left(\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\boldsymbol{a}(t)\right]\right) - U(\boldsymbol{z})\right)
\end{aligned} \tag{40}$$

Let $\boldsymbol{z} = \bar{\boldsymbol{R}}^*$ which is the rate point in $\mathcal{R}$ that achieves the optimum utility function. Based on the fact that all queues are mean rate stable as shown before, the left-hand side is non-negative, it then follows that

$$\begin{aligned}
U(\bar{\boldsymbol{R}}^*) &\le U\left(\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\boldsymbol{a}(t)\right]\right) + B/V\\
&\le U\left(\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\boldsymbol{R}(t)\right]\right) + B/V
\end{aligned} \tag{41}$$

Let $\tau \to \infty$, take the lim sup and rearrange the terms yields the optimality condition in (18). The inequality in (41) is based on the fact the queues are all mean rate stable the utility function is non-decreasing. This completes the proof.

APPENDIX B
PROOF OF PROPOSITION 1

Let $q_n = Q_n(t)$ denote the virtual queue state at the scheduling time t, define

$$\begin{aligned}
&\mathcal{G}(\boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{\Sigma}_{\mathrm{u}}, \boldsymbol{\Sigma}_{\mathrm{b}}, \boldsymbol{p}) \\
&\triangleq \sum_{n=1}^{N_{\mathrm{t}}} q_n s_n \sum_{t=1}^{T} R_n\left(\boldsymbol{H}(t), \boldsymbol{\Sigma}_{\mathrm{u}}, \boldsymbol{\Sigma}_{\mathrm{b}}, \boldsymbol{p}\right) \\
&= \sum_{n=1}^{N_{\mathrm{t}}} q_{\pi_n} s_{\pi_n} \\
&\quad \cdot \sum_{t=1}^{T} \log \frac{\det\left(\boldsymbol{I} + \sum_{j=1}^{n} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{h}_{\pi_j} \boldsymbol{h}_{\pi_j}^{\dagger} \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} p_{\pi_j} s_{\pi_j}\right)}{\det\left(\boldsymbol{I} + \sum_{j=1}^{n-1} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{h}_{\pi_j} \boldsymbol{h}_{\pi_j}^{\dagger} \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} p_{\pi_j} s_{\pi_j}\right)} \\
&= \sum_{n=1}^{N_{\mathrm{t}}} q_{\pi_n} s_{\pi_n} \\
&\quad \cdot \sum_{t=1}^{T} \log\left(1 + \boldsymbol{h}_{\pi_n}^{\dagger} \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{h}_{\pi_n} p_{\pi_n} s_{\pi_n}\right) \\
&= \sum_{n=1}^{N_{\mathrm{t}}} q_n \sum_{t=1}^{T} \log\left(1 + \boldsymbol{h}_n^{\dagger} \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{h}_n p_n s_n\right) \quad (42)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{I} + \bar{\boldsymbol{H}}_{[n-1]} \bar{\boldsymbol{H}}_{[n-1]}^{\dagger}, \\
\bar{\boldsymbol{H}}_{[n-1]} &= \sqrt{p_{\pi_j} s_{\pi_j}}\, \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \left[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{n-1}\right], \qquad (43)
\end{aligned}$$

and $\pi_i \in [1, \ldots, N_{\mathrm{t}}]$ is a permutation of the user index which satisfies $q_{\pi_1} s_{\pi_1} \geq \ldots \geq q_{\pi_{N_{\mathrm{t}}}} s_{\pi_{N_{\mathrm{t}}}}$ representing the decoding order in the dual uplink multiple-access-channel [41]. The last equality is based on the fact that

$$x \log\det(\boldsymbol{I} + \boldsymbol{A}x) = \log\det(\boldsymbol{I} + \boldsymbol{A}x), \quad \forall x \in \{0, 1\}, \quad (44)$$

and without loss of generality, we assume q_i's are arranged in non-increasing order. We invoke [56, Th. 1] which is stated at the end of the proof as Lemma 2 for reading convenience. Denote the channel correlation matrix for user-n as $\boldsymbol{R}_n$, and it yields

$$\begin{aligned}
&\mathcal{G}(\boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{\Sigma}_{\mathrm{u}}, \boldsymbol{\Sigma}_{\mathrm{b}}, \boldsymbol{p}) \\
&\quad = \sum_{n=1}^{N_{\mathrm{t}}} q_n \sum_{t=1}^{T} \log \\
&\qquad \times \left(1 + p_n s_n \cdot \mathrm{tr}\left[\boldsymbol{x}_n^{\dagger} \boldsymbol{R}_n^{\frac{1}{2}} \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_n^{\frac{1}{2}} \boldsymbol{x}_n\right]\right) \\
&\xrightarrow{K \to \infty} \sum_{n=1}^{N_{\mathrm{t}}} q_n \sum_{t=1}^{T} \log\left(1 + p_n s_n \cdot \mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_n \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{A}^{-1}\right]\right) \qquad (45) \\
&\xrightarrow{K \to \infty} \sum_{n=1}^{N_{\mathrm{t}}} q_n T \log\left(1 + p_n s_n \mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_n \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \right.\right. \\
&\qquad \left.\left. \cdot \left(\frac{1}{M} \sum_{j=1}^{n-1} \frac{p_j s_j \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_j \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}}}{1 + e_{n,j}} + \boldsymbol{I}\right)^{-1}\right]\right), \qquad (46)
\end{aligned}$$

where $e_{n,i}$ is the unique solution of the following equations.

$$\begin{aligned}
e_{n,i} = \mathrm{tr}\Bigg[&\boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_i \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \\
&\cdot \left(\frac{1}{M} \sum_{j=1}^{n-1} \frac{p_j s_j \boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_j \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}}}{1 + e_{n,j}} + \boldsymbol{I}\right)^{-1}\Bigg]. \qquad (47)
\end{aligned}$$

The inequality (45) is based on [57, Lemma 14.2], and (46) is based on the following lemma.

Lemma 2: Let $\boldsymbol{B_N} = \boldsymbol{X}_N^{\dagger} \boldsymbol{X}_N + \boldsymbol{S}_N$ with $\boldsymbol{S}_N \in \mathbb{C}^{N \times N}$ Hermitian nonnegative definite and $\boldsymbol{X}_N \in \mathbb{C}^{n \times N}$ random. The ith column $\boldsymbol{x}_i$ of $\boldsymbol{X}_N^{\dagger}$ is $\boldsymbol{x}_i = \boldsymbol{R}_i^{\frac{1}{2}} \boldsymbol{y}_i$, where the entries of $\boldsymbol{y}_i \in \mathbb{C}^{r_i}$ are i.i.d. of zero mean, variance $1/N$ and have eighth-order moment of order $\mathcal{O}\left(\frac{1}{N^4}\right)$. The matrices $\boldsymbol{R}_i$'s are channel correlation matrices for each user, and $\boldsymbol{Q}_N \in \mathbb{C}^{N \times N}$ is deterministic. Assume $\limsup_{N \to \infty} \sup_{1 \leq i \leq N} \|\boldsymbol{R}_i\| < \infty$ and let $\boldsymbol{Q}_N$ have uniformly bounded spectral norm (with respect to N). Define

$$m_{\boldsymbol{B}_N, \boldsymbol{Q}_N}(z) = \frac{1}{N} \mathrm{tr}\left[\boldsymbol{Q}_N \left(\boldsymbol{B}_N - z \boldsymbol{I}_N\right)^{-1}\right]. \qquad (48)$$

Then, for $z \in \mathbb{C} \backslash \mathbb{R}^{+}$, as n, N grow large with ratios $\beta_{N,i} = N/r_i$ and $\beta = N/n$ such that $0 < \liminf_N \beta_N \leq \limsup_N \beta_N < \infty$ and $0 < \liminf_N \beta_{N,i} \leq \limsup_N \beta_{N,i} < \infty$, we have that

$$m_{\boldsymbol{B}_N, \boldsymbol{Q}_N}(z) - m^{o}_{\boldsymbol{B}_N, \boldsymbol{Q}_N}(z) \to 0 \qquad (49)$$

almost surely, with $m^{o}_{\boldsymbol{B}_N, \boldsymbol{Q}_N}(z)$ given by

$$m^{o}_{\boldsymbol{B}_N, \boldsymbol{Q}_N}(z) = \frac{1}{N} \mathrm{tr} \boldsymbol{Q}_N \left(\frac{1}{N} \sum_{j=1}^{n} \frac{\boldsymbol{R}_j}{1 + e_{N,j}(z)} + \boldsymbol{S}_N - z \boldsymbol{I}_N\right)^{-1} \qquad (50)$$

where the functions $e_{N,j}(z)$ form the unique solution of

$$e_{N,i}(z) = \frac{1}{N} \mathrm{tr} \boldsymbol{R}_i \left(\frac{1}{N} \sum_{j=1}^{n} \frac{\boldsymbol{R}_j}{1 + e_{N,j}(z)} + \boldsymbol{S}_N - z \boldsymbol{I}_N\right)^{-1}. \square \qquad (51)$$

APPENDIX C
PROOF OF THE CONVERGENCE OF THE BCU-BASED ALGORITHM

First, the proof of the iterative water filling approach in the user scheduling part is given. Consider the user scheduling problem with beam selection fixed, i.e.,

$$\begin{aligned}
\textbf{P5:} \quad &\underset{\boldsymbol{w}}{\text{maximize}} \quad \sum_{n=1}^{N_{\mathrm{t}}} q_n \log\left(1 + \mathsf{f}_n(\boldsymbol{w})\right) \qquad (52) \\
&\text{s.t.,} \quad \sum_{n=1}^{N_{\mathrm{t}}} w_n \leq P, \qquad (53)
\end{aligned}$$

Denote

$$\mathsf{f}_n(\boldsymbol{w}) = \mathsf{f}_n(w_1, w_2, \ldots, w_n) = w_n \mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathrm{b}} \boldsymbol{B}_{\mathrm{DFT}}^{\dagger} \boldsymbol{R}_n \boldsymbol{B}_{\mathrm{DFT}} \boldsymbol{\Sigma}_{\mathrm{b}} \cdot \left(\frac{1}{M}\sum_{j=1}^{n-1} \frac{w_j \boldsymbol{R}_j}{1+e_{n,j}} + \boldsymbol{I}\right)^{-1}\right]. \quad (54)$$

The key to the proof is to construct the equivalent optimization problem as stated below.

$$\textbf{P6:} \quad \underset{\boldsymbol{w}(m),\, 0\le m\le N_{\mathrm{t}}-1}{\text{maximize}} \; \frac{1}{N_{\mathrm{t}}} \sum_{m=0}^{N_{\mathrm{t}}-1} \sum_{n=1}^{N_{\mathrm{t}}} q_n \log\big(1 + \mathsf{f}_n(w_1([m+1]_{N_{\mathrm{t}}}), w_2([m+2]_{N_{\mathrm{t}}}), \ldots, w_n([m+n]_{N_{\mathrm{t}}}))\big)$$

$$\text{s.t.,} \; \sum_{n=1}^{N_{\mathrm{t}}} w_n(m) \le P, \; \forall m. \quad (55)$$

The reason that **P5** and **P6** are equivalent is straightforward due to the Shur-concavity of the objective function of **P6** [58]. Therefore, the solution of **P6** is obtained at the point which satisfies

$$\boldsymbol{w}(m) = \boldsymbol{w}, \quad \forall m. \quad (56)$$

Since **P6** is concave in $\boldsymbol{w}(m)$, the BCU technique which cyclically optimizes $\boldsymbol{w}(m)$ with others fixed is guaranteed to converge to the global optimum, which yields the same procedure as in Algorithm 1 with some minor mathematical manipulations [59]. Therefore, we conclude that the iterative water filling approach adopted in Algorithm 1 converges to the optimum in the user scheduling step.

Next, it will be shown that the BCU technique which cyclically update user scheduling and beam selection converges to the optimum. Based on [35], it is sufficient to check if the problem satisfies the following two conditions:

- The objective function, denoted by $\phi(\boldsymbol{x})$, is continuously differentiable in some neighborhood of every stationary point of $\phi(\boldsymbol{x})$.
- For every k, $1 \le k \le n$, $\phi(\boldsymbol{x})$ is a strictly concave function of x_k, the other points x_j, $j \ne k$, being arbitrarily chosen in their respective domains.

The above two conditions are easily met in this problem since $\mathcal{D}(\boldsymbol{Q}, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{N_{\mathrm{t}}}, \boldsymbol{I}_{N_{\mathrm{t}}}, \boldsymbol{\Sigma}_{\mathrm{b}}, \boldsymbol{w})$ is continuously differentiable in the whole domain and concave in $\boldsymbol{\Sigma}_{\mathrm{b}}$ and $\boldsymbol{w}$, respectively. Therefore, the proposed BCU-based scheduling scheme is guaranteed to converge to the global optimum.

REFERENCES

[1] Z. Jiang, S. Zhou, and Z. Niu, "A block coordinated update method for beam-based massive MIMO downlink scheduling based on statistical CSI," in *Proc. Asia–Pacific Conf. Commun. (APCC)*, Dec. 2017.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] F. Rusek, A. Lozano, and N. Jindal, "Mutual information of IID complex Gaussian signals on block Rayleigh-faded channels," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 331–340, Jan. 2012.

[4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[5] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[6] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[7] Z. Jiang, S. Zhou, and Z. Niu, "Antenna-beam spatial transformation in c-RAN with large antenna arrays," in *Proc. IEEE Int. Conf. Commun. (ICC Workshops)*, May 2017, pp. 1215–1220.

[8] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, 2016.

[9] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[10] A. F. Molisch *et al.*, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[11] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for MmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[12] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[13] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2882, May 2015.

[14] Z. Jiang, S. Zhou, and Z. Niu, "On dimensionality loss in FDD massive MIMO systems," in *Proc. IEEE Conf. Wireless Commun. Netw. (WCNC)*, Mar. 2015, pp. 399–404.

[15] Z. Jiang, S. Zhou, R. Deng, Z. Niu, and S. Cao, "Pilot-data superposition for beam-based FDD massive MIMO downlinks," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1357–1360, Jun. 2017.

[16] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[17] A. F. Molisch and X. Zhang, "FFT-based hybrid antenna selection schemes for spatially correlated MIMO channels," *IEEE Commun. Lett.*, vol. 8, no. 1, pp. 36–38, Jan. 2004.

[18] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

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

[20] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 2055–2066, Jul. 2010.

[21] H. Kim, K. Kim, Y. Han, and S. Yun, "A proportional fair scheduling for multicarrier transmission systems," in *Proc. IEEE 60th Veh. Technol. Conf. (VTC-Fall)*, vol. 1. Sep. 2004, pp. 409–413.

[22] C. U. Bas *et al.* (2017). "A real-time millimeter-wave phased array MIMO channel sounder." [Online]. Available: https://arxiv.org/abs/1703.05271

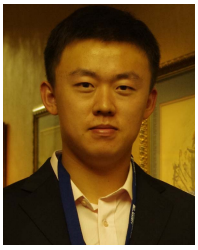
[23] X. Gao, L. Dai, Z. Chen, Z. Wang, and Z. Zhang, "Near-optimal beam selection for beamspace mmwave massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1054–1057, May 2016.

[24] P. V. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2212–2223, Jun. 2015.

[25] Y.-C. Liang and F. P. S. Chin, "Downlink channel covariance matrix (DCCM) estimation and its applications in wireless DS-CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 2, pp. 222–232, Feb. 2001.

[26] A. Adhikary *et al.*, "Joint spatial division and multiplexing for mm-Wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, Jun. 2014.

[27] A. Checko *et al.*, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.
[28] X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6010–6021, Sep. 2017.
[29] J. Wang, H. Zhu, L. Dai, N. J. Gomes, and J. Wang, "Low-complexity beam allocation for switched-beam based multiuser massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8236–8248, Dec. 2016.
[30] C. Sun, X. Q. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam division multiple access transmission for massive MIMO communications," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2170–2184, Jun. 2015.
[31] L. You, X. Gao, G. Y. Li, X.-G. Xia, and N. Ma, "BDMA for millimeter-wave/terahertz massive MIMO transmission with per-beam synchronization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1550–1563, Jul. 2017.
[32] A. F. Molisch and M. Z. Win, "MIMO systems with antenna selection," *IEEE Commun. Mag.*, vol. 5, no. 1, pp. 46–56, Mar. 2004.
[33] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.
[34] A. Edelman and N. R. Rao, "Random matrix theory," *Acta Numer.*, vol. 14, pp. 233–297, May 2005.
[35] J. Warga, "Minimizing certain convex functions," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 3, pp. 588–593, 1963.
[36] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Program.*, vol. 14, no. 1, pp. 265–294, 1978.
[37] A. F. Molisch, "A generic model for MIMO wireless propagation channels in macro- and microcells," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 61–71, Jan. 2004.
[38] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.
[39] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.
[40] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
[41] W. Yu, "Uplink-downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.
[42] Z. Jiang, S. Zhou, and Z. Niu, "Capacity bounds of downlink network MIMO systems with inter-cluster interference," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 4612–4617.
[43] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Downlink power allocation for multi-class CDMA wireless networks," in *Proc. IEEE INFOCOM*, vol. 3. Jun. 2002, pp. 1480–1489.
[44] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*. Breda, The Netherlands: Now Publishers, 2006.
[45] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations*. New York, NY, USA: Springer, 1972, pp. 85–103.
[46] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.
[47] S. E. Nai, W. Ser, Z. L. Yu, and H. Chen, "Beampattern synthesis for linear and planar arrays with antenna selection by convex optimization," *IEEE Trans. Antennas Propag.*, vol. 58, no. 12, pp. 3923–3930, Dec. 2010.
[48] A. Dua, K. Medepalli, and A. J. Paulraj, "Receive antenna selection in MIMO systems using convex optimization," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2353–2357, Sep. 2006.
[49] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.
[50] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
[51] R. Kim, H. Lim, and B. Krishnamachari, "Prefetching-based data dissemination in vehicular cloud systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 292–306, Jan. 2016.
[52] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[53] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.
[54] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sep. 2004.
[55] Z. Jiang, S. Zhou, and Z. Niu, "Dynamic channel acquisition in MU-MIMO," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4336–4348, Dec. 2014.
[56] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.
[57] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[58] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, vol. 143. New York, NY, USA: Springer, 1979.
[59] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1640–1646, Aug. 2006.



Zhiyuan Jiang (S'12–M'16) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2010 and 2015, respectively. From 2013 to 2014, he was a Visitor of the WiDeS Lab, University of Southern California, CA, USA. From 2015 to 2016, he was an Experienced Researcher with Ericsson Research. He currently holds a post-doctoral position with the Department of Electronic Engineering, Tsinghua University. His research interests include massive multiple-input multiple-output systems, stochastic optimization, and information theory.

Sheng Chen (S'17) was born in 1995. He received the bachelor's degree from Tsinghua University, Beijing, in 2016, where he is currently pursuing the Ph.D. degree. His research interests focus on channel estimation in massive multiple-input multiple-output systems.

Sheng Zhou (S'06–M'12) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2005 and 2011, respectively. In 2010, he joined the Wireless System Lab, Department of Electrical Engineering, Stanford University, Stanford, CA, USA, where he was a Visiting Student. From 2014 to 2015, he was a Visiting Researcher with the Central Research Lab, Hitachi Ltd., Japan. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include cross-layer design for multiple antenna systems, mobile edge computing, and green wireless communications.

Zhisheng Niu (M'98–SM'99–F'12) received the bachelor's degree from Beijing Jiaotong University, China, in 1985, and the M.E. and D.E. degrees from the Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992–1994, he was with the Fujitsu Laboratories Ltd., Japan. In 1994, he joined Tsinghua University, Beijing, China. He was a Visiting Researcher with the National Institute of Information and Communication Technologies, Japan, from 1995 to 1996; the Hitachi Central Research Laboratory, Japan, from 1997 to 1998; Saga University, Japan, in 2001; the Polytechnic University of New York, NY, USA, from 2002 to 2002; the University of Hamburg, Germany, in 2014; and the University of Southern California, CA, USA, in 2014. He is currently a Professor with the Department of Electronic Engineering, Tsinghua University. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.

He has served as a Chair for the Emerging Technologies Committee from 2014 to 2015, the Director for Conference Publications from 2010 to 2011 and the Asia–Pacific Board of the IEEE Communications Society from 2008 to 2009, the Councilor for IEICE, Japan, from 2009 to 2011, and a member of the IEEE Teaching Award Committee from 2014 to 2015 and the IEICE Communication Society Fellow Evaluation Committee from 2013 to 2014. He has also served as an Associate Editor-in-Chief for the IEEE/CIC joint publication *China Communications* from 2012 to 2016, and an Editor for the IEEE WIRELESS COMMUNICATION from 2009 to 2013 and the *Wireless Networks* from 2005 to 2009. He is currently serving as an Area Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATION AND NETWORKS and the Director for Online Content of IEEE ComSoc from 2018 to 2019.

Dr. Niu has authored or co-authored over 100 journal and over 200 conference papers in IEEE and IEICE publications. He co-received the best paper awards from the 13th, 15th, and 19th Asia–Pacific Conference on Communication in 2007, 2009, and 2013, respectively, the International Conference on Wireless Communications and Signal Processing (WCSP 2013), and the Best Student Paper Award from the 25th International Teletraffic Congress (ITC25). He received the Outstanding Young Researcher Award from the Natural Science Foundation of China in 2009 and the Best Paper Award from the IEEE Communication Society Asia–Pacific Board in 2013. He was a Distinguished Lecturer of the IEEE Communication Society from 2012 to 2015 and the IEEE Vehicular Technologies Society from 2014 to 2016. He is a fellow of IEICE.