# User Association and Scheduling Based on Auction in Multi-Cell MU-MIMO Systems

Mengjie Xie, *Student Member, IEEE*, Tat-Ming Lok, *Senior Member, IEEE*, and Qing Yang, *Member, IEEE*

***Abstract*— We study the user association and scheduling problem in multi-cell multi-user multiple-input multiple-output systems with dynamic traffic. Two successive linear beamforming schemes, namely, zero forcing-successive interference cancellation and adaptive orthogonal beamforming, are adopted in the uplink channel and the downlink channel, respectively. The confidence level of signal-to-interference-plus-noise ratio of a random user is analyzed, in order for base stations (BSs) to decide on the maximum number of active users while guaranteeing the worst-case quality of service. We also derive the conditional expected rate of a random user knowing its local channel state information, in order for users to evaluate the performance and select the BS. Noticing that the performance can be divided into different classes, we model the user association and scheduling problem as an auction game. The auction is equipped with a strategy-proof pricing scheme and can be conducted in a distributed fashion with a limited information exchange. Shown by the simulation, the proposed mechanism has the dominant performance in respect of social welfare, throughput, and load balancing.**

***Index Terms*— MU-MIMO, user association, user scheduling, zero forcing-successive interference cancellation, adaptive orthogonal beamforming, auction, pricing.**

## I. Introduction

MULTI-USER MIMO has been supported by 4G and wireless local area network (WLAN) standards since LTE release 8, 802.16m and 802.11ac [1], [2]. MU-MIMO is also known as MIMO broadcast channel (BC) in the downlink case and MIMO multiple access channel (MAC) in the uplink case. It is known that dirty paper coding (DPC) achieves the capacity region of MIMO BC [3]. Successive interference cancellation (SIC), first proposed in [4], can be viewed as a reciprocal approach of DPC in MIMO MAC. However, such capacity-achieving schemes have high computational complexity in implementation. Linear precoding and decoding schemes, such as zero-forcing beamforming (ZFBF) [5]–[7] and orthogonal beamforming (OBF) [8], are more practical techniques. In [5], Caire et al. combine ZF with DPC to cancel the non-causally known interference. This scheme is referred to as ZF-DP in [5] and ZF-DPC in [9]. ZF-DPC requires coding with known interference, which is still complicated and does not suit dynamic traffic. A dual scheme named ZF-SIC is adopted in [9] and [10] for MIMO MAC. It applies the ZF decoder stream by stream and removes decoded streams successively as in SIC. In this paper, we adopt ZF-SIC in MIMO MAC, considering its promising performance and adaptability in a dynamic scenario. In MIMO BC, successive design of beamformers is also applicable to OBF [11]–[14]. A user's beamformer can be matched with its channel projected to the null space of all beamformers of the precedent users. As a result, precedent users do not experience interference from later users. This scheme is referred to as adaptive OBF (AOBF) in [11] and [14] and this paper, multilayer OBF (MOBF) in [12] and successive OBF (SOBF) in [13]. AOBF is a practical scheme in the downlink, because successive pre-subtraction of interference is achieved by beamforming rather than coding in ZF-DPC. In addition, AOBF adapts better to a small and varying number of users [13].

ZF-SIC and AOBF share the feature of successive design of decoders or beamformers and yield classified service to users. However, the classified service is not exploited in the corresponding works [9]–[14]. In this paper, we propose a user scheduling mechanism to exploit this feature of ZF-SIC and AOBF from the economic perspective, in the light of the Vickrey-Clarke-Groves (VCG) auction [15]. Auction theory has been widely applied for resource allocation in networks. Early in 1993, [16] has suggested a smart market mechanism to price congestion. Later, [17] proposes a smart pay admission control (SPAC) mechanism to price different classes of QoS. Both mechanisms follow the principle of VCG auction.

In addition to user scheduling, user-BS association plays a crucial role in enhancing the performance as the network becomes dense and heterogeneous in 5G systems [18]–[22]. In [21], the authors studied the energy efficiency problem of joint user association and power allocation in a two-tier heterogeneous network with small cells, and used a continuous and convex relaxation method to maximize the energy efficiency. Ultra-dense network and energy harvesting are considered in [22] and an iterative gradient user association and power allocation algorithm is proposed and shown to converge rapidly to an optimal point. In our work, we employ game theory, instead of optimization tool, to attack the user

Manuscript received July 26, 2017; revised January 8, 2018; accepted March 14, 2018. Date of publication April 5, 2018; date of current version June 8, 2018. This work was supported by the General Research Fund from the Research Grants Council, Hong Kong, under Project CUHK 14203616. This paper was presented at the IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, May 2016. The associate editor coordinating the review of this paper and approving it for publication was S. K. Jayaweera. *(Corresponding author: Qing Yang.)*

M. Xie and T.-M. Lok are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: xm012@ie.cuhk.edu.hk; tmlok@ie.cuhk.edu.hk).

Q. Yang is with College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: yang.qing@hotmail.com).

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.2018.2820711

association and pricing problem. We also show that this auction method can be conducted in a distributed manner with limited information exchange. In this paper, we consider a multi-cell system and a bidirectional user-BS association problem coupled with the user scheduling problem in each cell. Another problem in 5G systems is more frequent handovers of mobile users due to smaller cell size [23]–[25]. In terms of user-BS association, a handover event can be viewed as a departure event followed by an arrival event. In this paper, we assume that users arrive and depart dynamically, which is also an inherent feature in communication networks. Dynamic traffic renders the user association and scheduling problem more challenging.

The contribution of this paper is stressed below. First, we derive the exact numerical solution to the SINR confidence level of a random user under AOBF. This is a remaining problem in previous works [11]–[14]. Second, we derive the conditional distribution of SINR and conditional expected rate of a random user under ZF-SIC and AOBF, given the user's local CSI. This is uniquely concerned in our work for users to perform BS selection in multi-cell systems. Finally, we propose an auction-based user association and scheduling mechanism equipped with a strategy-proof pricing scheme. The mechanism is different than the traditional VCG auction in the sense that there is no centralized auctioneer and that all users are accommodated with limited resources. In our previous work [26], the mechanism is demonstrated with ZF-SIC in the uplink. It follows the VCG auction, which may exclude users from the system during heavy traffic due to limited resources. In this paper, we perform a comprehensive analysis of the rate and SINR under both ZF-SIC and AOBF. The mechanism adapts to both schemes with unified expressions of the utility functions and QoS criteria and is improved to accommodate all users.

The rest of the paper is organized as follows. In Section II we introduce the MU-MIMO system model performing ZF-SIC in the uplink and AOBF in the downlink. Section III analyzes a user's SINR confidence level, both unconditionally and conditioning on local CSI of the user, as well as its conditional expected rate. In Section IV, we give the utility function and QoS criteria, propose the auction-based mechanism for user association and scheduling and prove that it is strategy-proof, individually rational and socially near-optimal. Section V demonstrates the simulation results in a dynamic scenario. Section VI concludes the paper.

## II. System Model

We consider a general multi-cell MU-MIMO system. There are $K$ BSs, indexed by the set $\mathcal{K} = \{1, 2, \dots, K\}$, in the network. The $k$th ($k \in \mathcal{K}$) BS has $M^{[k]}$ antennas. Users have a single antenna and arrive and depart dynamically. We assume block Rayleigh fading that the CSI stays unchanged during the packet transmission. Each user is associated with one BS and each BS can serve multiple users. We refer to the users associated with the same BS as a cluster. Different BS-cluster pairs are assumed to transmit on different sets OFDM subcarriers.

Next we illustrate a static status of the concerned multi-cell MU-MIMO OFDM system with dynamic traffic. There are $N$ users, indexed by the set $\mathcal{N} = \{1, 2, \dots, N\}$, in the network. The index set for users in the $k$th cluster is denoted by $\mathcal{N}^{[k]}$ such that $\mathcal{N}^{[k]}$ are disjoint for different $k$ and $\mathcal{N} = \cup_{k\in\mathcal{K}}\mathcal{N}^{[k]}$. Partitioning $\mathcal{N}$ into $\mathcal{N}^{[k]}$ captures the inter-cell user association problem. For concise reference to the users within the same cluster, we further index the $n$th ($n \in \mathcal{N}^{[k]}$) user by $l = \phi_n^{[k]}$ where $\phi_n^{[k]}$ is an integer-valued function mapping $n \in \mathcal{N}^{[k]}$ to $l \in \mathcal{L}^{[k]} = \{1, 2, \dots, L^{[k]}\}$. $L^{[k]} = \min\{|\mathcal{N}^{[k]}|, \bar{L}^{[k]}\}$ represents the number of simultaneously active users in the $k$th cluster. $\bar{L}^{[k]}$ is the maximum number of supportable users satisfying a worst-case QoS requirement. If $|\mathcal{N}^{[k]}| \le \bar{L}^{[k]}$, $\phi_n^{[k]}$ is a bijective mapping between $\mathcal{N}^{[k]}$ and $\mathcal{L}^{[k]}$. In the case of $|\mathcal{N}^{[k]}| > \bar{L}^{[k]}$, the remaining $(|\mathcal{N}^{[k]}| - L^{[k]})$ users wait in the queue with $l = 1$ as well. Along the timeline, users with $l = 1$ are scheduled in a round-robin manner. $|\mathcal{N}^{[k]}|$ and $\bar{L}^{[k]}$ are referred to as the cluster size and cluster capacity, respectively. As is to be specified later, the index $l = \phi_n^{[k]}$ indicates the order of designing the decoders or beamformers. Assignment of $\phi_n^{[k]}$ captures the intra-cell user scheduling problem.

In the rest of this section and next section, we investigate a static status of the MU-MIMO channel of the $k$th BS-cluster pair and omit the superscript $[k]$ in all notations for briefness. For instance, we say that BS has $M$ ($M^{[k]}$) antennas and serves $L$ ($L^{[k]}$) single-antenna users simultaneously. The uplink channel adopts ZF-SIC and the downlink channel adopts AOBF, which is to be discussed separately. We follow the definitions of ZF-SIC in [9] and [10] and AOBF in [11] and [14] and derive the equations using the notations defined in this section.

### A. ZF-SIC in the Uplink Channel

In the following analysis, we take the $k$th BS-cluster as an example, noted that the following analysis is applicable for other BS-cluster as well. In the uplink, the $l$th ($l \in \mathcal{L}^{[k]}$) user transmits a data stream $x_l(t)$ with $\mathbb{E}[x_l(t)x_l^*(t)] = 1$ using power $p_l$. The received signal at the BS is

$$\mathbf{y}(t) = \sum_{l\in\mathcal{L}^{[k]}} \sqrt{p_l}\mathbf{h}_l\, x_l(t) + \mathbf{z}(t), \tag{1}$$

where $\mathbf{h}_l$ is the CSI vector from user $l$ to all $M$ antennas of the BS. We assume a Rayleigh flat fading channel such that the entries of all $\mathbf{h}_l$ are i.i.d. zero-mean unit-variance circularly symmetric complex Gaussian random variables (r.v.s). $\mathbf{z}(t) \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_M)$ is the additive white Gaussian noise (AWGN) vector.

To estimate each data stream $x_l(t)$, the BS adopts ZF-SIC by applying a unitary matrix[1] $\mathbf{U}_l$ that zero-forces multi-user interference. In addition, the successfully decoded data streams are removed from $\mathbf{y}(t)$. Assume the BS decodes the data streams in ascending order of $l$. The interference space for the $l$th data stream can be represented by the $M \times (L - l)$ matrix

$$\mathbf{F}_l = \left[\, \mathbf{h}_{l+1}\ \mathbf{h}_{l+2}\ \cdots\ \mathbf{h}_L \,\right]. \tag{2}$$

[1] A matrix having orthonormal columns is referred to as "unitary matrix" for consistency of notations in this work.

Then the decoding matrix $\mathbf{U}_l$ of dimension $M \times d_{\mathbf{U}_l}$ satisfies the following conditions:

$$\mathbf{U}_l^H \mathbf{F}_l = \mathbf{0}, \tag{3}$$

$$\mathbf{U}_l^H \mathbf{U}_l = \mathbf{I}_{d_{\mathbf{U}_l}}. \tag{4}$$

Because $\mathbf{F}_l$ is an i.i.d. Gaussian random matrix, it is almost surely full rank [27], thus has rank $(L-l)$ with probability 1. Therefore, $\text{rank}(\mathbf{U}_l) = d_{\mathbf{U}_l} = M - L + l$. Now the estimated $l$th data stream is given by

$$\hat{x}_l(t) = \mathbf{U}_l^H \mathbf{y}(t) = \sqrt{p_l}\mathbf{U}_l^H \mathbf{h}_l \, x_l(t) + \mathbf{U}_l^H \mathbf{z}(t). \tag{5}$$

With optimal combining, the SINR and rate of the $l$th user are given by

$$\Gamma_l = \frac{p_l}{\sigma^2} ||\mathbf{U}_l^H \mathbf{h}_l||^2, \tag{6}$$

$$R_l = \log(1 + \Gamma_l). \tag{7}$$

### B. AOBF in the Downlink Channel

In the downlink, the BS multiplexes and transmits $x_l$ ($l \in \mathcal{L}^{[k]}$) with the $M \times 1$ beamforming vector $\mathbf{v}_l$ and an equal power $p = \frac{P}{L}$. The received signal at the $l$th user is

$$\hat{x}_l(t) = \sum_{m \in \mathcal{L}^{[k]}} \sqrt{p}\, \overleftarrow{\mathbf{h}}_l^H \mathbf{v}_m \, x_l(t) + \overleftarrow{z}_l(t), \tag{8}$$

where $\overleftarrow{\mathbf{h}}_l$ is the CSI vector between the BS and the $l$th user in the downlink channel with the same distribution as $\mathbf{h}_l$, and $\overleftarrow{z}_l(t)$ is the AWGN with noise power $\sigma^2$.

With AOBF, precedent users will receive better service than later users. We let the beamformers be designed in descending order of $l$. The first designed beamformer $\mathbf{v}_L$ is given by

$$\mathbf{v}_L = \frac{\overleftarrow{\mathbf{h}}_L}{||\overleftarrow{\mathbf{h}}_L||}. \tag{9}$$

The remaining $\mathbf{v}_l$ for $l < L$ are given recursively by the following procedure.

$$\mathbf{W}_l = \left[\, \mathbf{v}_{l+1} \; \mathbf{v}_{l+2} \; \cdots \; \mathbf{v}_L \,\right], \tag{10}$$

$$\tilde{\mathbf{v}}_l = \left[\mathbf{I}_M - \mathbf{W}_l \mathbf{W}_l^H\right] \overleftarrow{\mathbf{h}}_l, \tag{11}$$

$$\mathbf{v}_l = \frac{\tilde{\mathbf{v}}_l}{||\tilde{\mathbf{v}}_l||}. \tag{12}$$

By observing $\mathbf{W}_l^H \tilde{\mathbf{v}}_l = (\mathbf{I}_{d_{\mathbf{W}_l}} - \mathbf{W}_l^H \mathbf{W}_l)\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l = 0$, it can be verified recursively that

$$\mathbf{W}_l^H \mathbf{W}_l = \mathbf{I}_{d_{\mathbf{W}_l}}, \tag{13}$$

where $d_{\mathbf{W}_l} = L - l$. We can also verify that

$$\tilde{\mathbf{v}}_m^H \overleftarrow{\mathbf{h}}_l = \tilde{\mathbf{v}}_m^H \tilde{\mathbf{v}}_l + \tilde{\mathbf{v}}_m^H \mathbf{W}_l \mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l = 0, \quad \forall m < l. \tag{14}$$

Therefore, the $l$th user will not receive interference from the $m$th user where $m < l$. Equation (8) can be rewritten as

$$\hat{x}_l(t) = \sqrt{p}\, \overleftarrow{\mathbf{h}}_l^H \mathbf{v}_l \, x_l(t) + \sqrt{p}\, \overleftarrow{\mathbf{h}}_l^H \mathbf{W}_l \bar{\mathbf{x}}_l(t) + \overleftarrow{z}_l(t), \tag{15}$$

where $\bar{\mathbf{x}}_l(t) = [x_{l+1}(t), \ldots, x_L(t)]^T$. Now the SINR of the $l$th user in the downlink channel with AOBF is given by

$$\overleftarrow{\Gamma}_l = \frac{||\mathbf{v}_l^H \overleftarrow{\mathbf{h}}_l||^2}{||\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l||^2 + \sigma^2/p} = \frac{||\overleftarrow{\mathbf{h}}_l||^2 - ||\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l||^2}{||\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l||^2 + \sigma^2/p}. \tag{16}$$

The corresponding rate is expressed as

$$\overleftarrow{R}_l = \log(1 + \overleftarrow{\Gamma}_l). \tag{17}$$

It is noticed that the SINR and rate are dependent on $||\mathbf{U}_l^H \mathbf{h}_l||^2$ with ZF-SIC in the uplink case and $||\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l||^2$ with AOBF in the downlink case, where $\mathbf{U}_l$ and $\mathbf{W}_l$ are both unitary matrices. As will be shown in later sections, the performance analysis of the two schemes can be partially unified and the user association and scheduling mechanism suits both schemes.

## III. Performance Analysis

This paper involves the user association problem in a multi-cell system with dynamic traffic. Users have the incentive to evaluate its performance served by different BSs. Under ZF-SIC and AOBF, a user's performance is dependent on its local CSI as well as the CSI of all precedent users. However, due to the dynamic traffic, the information of precedent users needs to be updated frequently, incurring high computational complexity. Therefore, in the interests of users, we evaluate the performance conditioning on local CSI between users and different BSs. In the interests of BSs, the unconditional performance is also analyzed.

We first show the common properties regarding the ZF-SIC beamformers $\mathbf{U}_l$ and the AOBF beamformers $\mathbf{W}_l$. The distributions of $\mathbf{U}_l^H \mathbf{h}_l$ and $\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l$ are given by $\mathbf{U}_l^H \mathbf{h}_l \sim \mathcal{CN}(0, \mathbf{I}_{d_{\mathbf{U}_l}})$ and $\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l \sim \mathcal{CN}(0, \mathbf{I}_{d_{\mathbf{W}_l}})$, because $\mathbf{h}_l$ has i.i.d. complex Gaussian entries, entries of $\mathbf{U}_l^H \mathbf{h}_l$ are also complex Gaussian r.v.s. For any deterministic $\mathbf{U}_l$ satisfying $\mathbf{U}_l^H \mathbf{U}_l = \mathbf{I}_{d_{\mathbf{U}_l}}$, the mean and covariance matrix of $\mathbf{g}_l$ are given by $\mathbb{E}[\mathbf{g}_l] = \mathbf{U}_l^H \mathbb{E}[\mathbf{h}_l] = \mathbf{0}$ and $\mathbb{E}\left[\mathbf{g}_l \mathbf{g}_l^H\right] = \mathbf{U}_l^H \mathbb{E}\left[\mathbf{h}_l \mathbf{h}_l^H\right] \mathbf{U}_l = \mathbf{I}_{d_{\mathbf{U}_l}}$. The result holds in general because $\mathbf{U}_l$ is determined by $\mathbf{F}_l$ and independent of $\mathbf{h}_l$. Therefore, $\mathbf{g}_l \sim \mathcal{CN}(0, \mathbf{I}_{d_{\mathbf{U}_l}})$. The distribution of $\mathbf{W}_l^H \overleftarrow{\mathbf{h}}_l$ is proved exactly the same way.

*Proposition 1: $\mathbf{U}_l$ and $\mathbf{W}_l$ are isotropically random unitary matrices.*[2]

*Proof:* Let $\mathbf{F}_l' = \mathbf{\Omega}^H \mathbf{F}_l$ for some independent square unitary matrix $\mathbf{\Omega}$. The columns of $\mathbf{F}_l'$ are independent and the distribution of each column can be inferred by complex Gaussian distribution as shown in the previous paragraph. Obviously $\mathbf{F}_l'$ and $\mathbf{F}_l$ have the same i.i.d. complex Gaussian distribution and have the same probability density $f(\mathbf{F}_l) = f(\mathbf{F}_l')$. For any unitary $\mathbf{U}_l$ being a solution to $\mathbf{U}_l^H \mathbf{F}_l = \mathbf{0}$, $\mathbf{U}_l' = \mathbf{\Omega}\mathbf{U}_l$ is a corresponding solution to $\mathbf{U}_l'^H \mathbf{F}_l' = \mathbf{0}$. Because $f(\mathbf{F}_l) = f(\mathbf{F}_l')$, it follows that $f(\mathbf{U}_l) = f(\mathbf{\Omega}\mathbf{U}_l)$. Therefore, $\mathbf{U}_l$ is an isotropically random unitary matrix.

The proof of $\mathbf{W}_l$ being an isotropically random unitary matrix follows a similar procedure. Let $\overleftarrow{\mathbf{h}}_l' = \mathbf{\Omega}\overleftarrow{\mathbf{h}}_l$ for all $l \in \mathcal{L}$ for some independent square unitary matrix $\mathbf{\Omega}$. Let $\mathbf{v}_l$ be the unique solution given by (9) to (12) with the channel realization $\overleftarrow{\mathbf{h}}_l$, and $\mathbf{v}_l'$ be the solution with the

[2] An isotropically random unitary matrix is a matrix whose columns are orthonormal and whose probability density is invariant to left multiplication by an independent square unitary matrix.

channel realization $\overleftarrow{\mathbf{h}}'_l$. It can be proved recursively that $\mathbf{v}'_l = \mathbf{\Omega}\mathbf{v}_l$ for all $l \in \mathcal{L}$. Accordingly, $\mathbf{W}'_l = \mathbf{\Omega}\mathbf{W}_l$. Since $f(\overleftarrow{\mathbf{h}}_{l+1}, \ldots, \overleftarrow{\mathbf{h}}_L) = f(\overleftarrow{\mathbf{h}}'_{l+1}, \ldots, \overleftarrow{\mathbf{h}}'_L)$, we have $f(\mathbf{W}_l) = f(\mathbf{\Omega}\mathbf{W}_l)$ and finish the proof. □

With the goal of evaluating a user's SINR and rate conditioning on local CSI, we want to derive the conditional distribution of $\mathbf{U}^H\mathbf{h}$ for a deterministic vector $\mathbf{h}$ and an isotropically random unitary matrix $\mathbf{U}$. We start with Theorem 1 as follows.

*Theorem 1: $\boldsymbol{\alpha} = \mathbf{U}^H\mathbf{n}$, where $\mathbf{U} \in \mathcal{C}^{M\times d}$ is an isotropically random unitary matrix and $\mathbf{n} \in \mathcal{C}^M$ is a deterministic unit vector, is a random vector with the following probability density function*

$$f_{\boldsymbol{\alpha}}(\mathbf{x}) = \frac{1}{\pi^d}\frac{\Gamma(M)}{\Gamma(M-d)}(1-||\mathbf{x}||^2)^{M-d-1}, \tag{18}$$

*where $\Gamma(\cdot)$ is the Gamma function.*

*Proof:* See Appendix A. □

Based on (18), it is not difficult to derive the distribution of $\beta = ||\boldsymbol{\alpha}||^2$ where $\boldsymbol{\alpha} = \mathbf{U}^H\mathbf{n}$, given by the following corollary.

*Corollary 1: $\beta = ||\boldsymbol{\alpha}||^2$ is Beta$(d, M-d)$ distributed with the probability density function*

$$f_\beta(x) = \frac{1}{B(d, M-d)}x^{d-1}(1-x)^{M-d-1}. \tag{19}$$

Theorem 1 suggests that $\mathbf{U}_l^H\mathbf{h}_l$ conditioning on $\mathbf{h}_l$ and $\mathbf{W}_l^H\overleftarrow{\mathbf{h}}_l$ conditioning on $\overleftarrow{\mathbf{h}}_l$ are regardless of the direction but only the norm of $\mathbf{h}_l$ and $\overleftarrow{\mathbf{h}}_l$. Therefore, only the norm instead of the full CSI vector needs to be fed back in the uplink for users to evaluate the performance.

In the following, we derive the unconditional SINR confidence level for BSs to evaluate the worst-case QoS and determine the cluster capacity $\bar{L}$. More importantly, we derive the conditional SINR confidence level and the conditional expected rate for users to evaluate the performance and select the BS.

### *A. Performance of ZF-SIC*

With ZF-SIC, the SINR and rate of user $l$ are given by (6) and (7). Since $\mathbf{U}_l^H\mathbf{h}_l \sim \mathcal{CN}(0, \mathbf{I}_{d_{\mathbf{U}_l}})$ according to Proposition 1, it is well known that $||\mathbf{U}_l^H\mathbf{h}_l||^2$ is $\chi^2$-distributed with DoF being 2 $d_{\mathbf{U}_l} = 2(M-L+l)$. Therefore, the confidence level of $\Gamma_l \geq \gamma_l$ for some SINR target $\gamma_l$ can be calculated as follows.

$$\begin{aligned}\Pr(\Gamma_l \geq \gamma_l) &= 1 - F_{\chi^2}\left(\frac{\sigma^2\gamma_l}{p_l}; 2(M-L+l)\right)\\ &= \frac{\Gamma(M-L+l, \frac{\sigma^2\gamma_l}{2p_l})}{\Gamma(M-L+l)},\end{aligned} \tag{20}$$

where $F_{\chi^2}(x;k)$ is the cdf of $\chi^2(k)$ distribution, $\Gamma(\cdot)$ is the Gamma function, and $\Gamma(\cdot,\cdot)$ is the upper incomplete Gamma function.

Table I lists a few examples with the SINR target $\gamma_l = 10$ dB and SNR $= p_l/\sigma^2$ varying from 0 to 20 dB. Under ZF-SIC, the first user with index $l = 1$ receives the worst service. The BS can adjust the cluster capacity $\bar{L}$ as shown to ensure that the worst user has a confidence level >90% to achieve SINR >10 dB.

TABLE I
CONFIDENCE LEVEL OF $\Gamma_l \geq \gamma_l$

| $\gamma_l$ | $p_l/\sigma^2$ | $M-L+l$ | $\bar{L}$ | Confidence level |
|---|---|---|---|---|
| 10 | 100 | 1 | $M$ | $> 95\%$ |
| 10 | 10 | 2 | $M-1$ | $> 90\%$ |
| 10 | 1 | 5 | $M-4$ | $> 97\%$ |

Following Theorem 1 and Corollary 1, the conditional SINR $\Gamma_l$ given $\mathbf{h}_l$ can be expressed in terms of the r.v. $\beta_l \sim Beta(M-L+l, L-l)$.

$$\Gamma_l|\mathbf{h}_l = \frac{p_l}{\sigma^2}||\mathbf{h}_l||^2\beta_l. \tag{21}$$

The conditional confidence level for the SINR threshold $\gamma_l$ is

$$\begin{aligned}\Pr(\Gamma_l \geq \gamma_l|\mathbf{h}_l) &= 1 - F_{\beta_l}\left(\frac{\sigma^2\gamma_l}{p_l||\mathbf{h}_l||^2}\right)\\ &= 1 - I_{\frac{\sigma^2\gamma_l}{p_l||\mathbf{h}_l||^2}}\left(M-L+l, L-l\right),\end{aligned} \tag{22}$$

where $F_{\beta_l}(x)$ is the cdf of $\beta_l$ and can be represented by the regularized incomplete beta function $I_x(\cdot,\cdot)$.

Accordingly, the conditional expected rate is

$$\mathbb{E}\left[R_l|\mathbf{h}_l\right] = \int_0^1 \log\left(1 + \frac{p_l}{\sigma^2}||\mathbf{h}_l||^2\, x\right) f_{\beta_l}(x)\, dx, \tag{23}$$

where $f_{\beta_l}(x)$ is the pdf of $\beta_l$. (23) can be calculated numerically. However, it is difficult to derive a closed-form expression. Instead, we can approximate $R_l$ by $\log(\Gamma_l)$. This approximation becomes accurate at high SNR. Based on the geometric mean of beta distribution, we have

$$\begin{aligned}&\mathbb{E}\left[R_l|\mathbf{h}_l\right]\\ &\quad\approx \log\left(\frac{p_l}{\sigma^2}\right) + \log\left(||\mathbf{h}_l||^2\right) + \psi(M-L+l) - \psi(M),\end{aligned} \tag{24}$$

where $\psi(\cdot)$ is the digamma function. It is clear that the conditional expected rate is determined by the transmit SNR of the user, the strength of the channel, the order of decoding and the antenna number of the BS.

### *B. Performance of AOBF*

With AOBF, the SINR and rate of user $l$ are given by (16) and (17). Shown by the literature [12]–[14], it is hard to derive exact and closed-form expression of the CDF of the SINR for a random user. Reference [12] and [13] apply the concepts of orthogonal efficiency and orthogonal deficiency to approximate the distribution of SINR. Reference [14] applies order statistics and analyzes the performance of the first three users. It turns out that a unified and computationally friendly numerical solution to the confidence level, or cdf, of the SINR for a random user can be derived with the help of Theorem 1 and Corollary 1.

According to Theorem 1 and Corollary 1, $||\mathbf{W}_l^H\overleftarrow{\mathbf{h}}_l||^2$ can be written as $\overleftarrow{\beta}_l||\overleftarrow{\mathbf{h}}_l||^2$, where $\overleftarrow{\beta}_l \sim Beta(L-l, M-L+l)$ and $||\overleftarrow{\mathbf{h}}_l||^2 \sim \chi^2(2M)$ are two independent r.v.s. Therefore, $\overleftarrow{\Gamma}_l \geq \overleftarrow{\gamma}_l$ for some SINR target $\overleftarrow{\gamma}_l$ can be equivalently expressed in terms of an auxiliary r.v. $X_l$ as following

$$X_l = (1+\overleftarrow{\gamma}_l)\overleftarrow{\beta}_l + \frac{\sigma^2\overleftarrow{\gamma}_l}{p}\frac{1}{||\overleftarrow{\mathbf{h}}_l||^2} \leq 1, \tag{25}$$

where $X_l$ is the weighted sum of two independent r.v.s $\overleftarrow{\beta}_l$ and $\frac{1}{||\overleftarrow{\mathbf{h}}_l||^2} \sim$ inv-$\chi^2(2M)$. The pdf of $X_l$ is the convolution of the pdfs of the scaled beta r.v. and the scaled inv-$\chi^2$ r.v., written as

$$f_{X_l}(x) = \left[\frac{1}{1+\overleftarrow{\gamma}_l} f_{\overleftarrow{\beta}_l}\left(\frac{x}{1+\overleftarrow{\gamma}_l}\right)\right] * \left[\frac{p}{\sigma^2 \overleftarrow{\gamma}_l} f_{\frac{1}{||\overleftarrow{\mathbf{h}}_l||^2}}\left(\frac{p}{\sigma^2\overleftarrow{\gamma}_l} x\right)\right]. \quad (26)$$

The characteristic functions of beta distribution and inv-$\chi^2$ distribution are well known. Then $X_l$ can be described by its characteristic function given by

$$\begin{aligned}\phi_{X_l}(t) &= \phi_{\overleftarrow{\beta}_l}\Big((1+\overleftarrow{\gamma}_l)t\Big)\phi_{\frac{1}{||\overleftarrow{\mathbf{h}}_l||^2}}\left(\frac{\sigma^2\overleftarrow{\gamma}_l}{p}t\right)\\ &= 1F_1\Big(L-l; M; (1+\overleftarrow{\gamma}_l)jt\Big)\\ &\quad \times \frac{2}{\Gamma(M)}\left(-\frac{\sigma^2\overleftarrow{\gamma}_l}{2p}jt\right)^{\frac{M}{2}} K_M\left(\sqrt{-\frac{2\sigma^2\overleftarrow{\gamma}_l}{p}jt}\right),\end{aligned} \quad (27)$$

where ${}_1F_1(\cdot;\cdot;\cdot)$ is the confluent hypergeometric function of the first kind and $K_n(\cdot)$ is the modified Bessel function of the second kind. Now the SINR confidence level for the threshold $\overleftarrow{\gamma}_l$ can be computed using the Gil-Pelaez Theorem [28]

$$\begin{aligned}\Pr(\overleftarrow{\Gamma}_l \geq \overleftarrow{\gamma}_l) &= \Pr(X_l \leq 1)\\ &= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\Im\left[e^{-jt}\phi_{X_l}(t)\right]}{t}dt.\end{aligned} \quad (28)$$

Equation (28) can be applied to any user $l$ and is more accurate and computationally friendly,[3] compared with the existing methods which are either approximations or entail multiple integrals. Moreover, $\Im\left[e^{-jt}\phi_{X_l}(t)\right]/t = 0$ as $t \to \infty$ [29, Lemma 1], so the function can be considered to have finite support in computation.

In the case of large $M$, the law of large numbers can be used to approximate $||\overleftarrow{\mathbf{h}}_l||^2$, the sum of $2M$ i.i.d. $\chi^2(1)$ r.v.s, by the value $2M$. This result is also implied by the pdf of inv-$\chi^2(2M)$ for a large $M$, which appears to be a much narrower impulse located around $\frac{1}{2M}$, compared with the pdf of $Beta(L-l, M-L+l)$. Therefore, in (25), the inv-$\chi^2$ r.v. can be taken as a constant $\frac{1}{2M}$ during the convolution. Moreover, by comparing the scaling factors of the beta r.v. and the inv-$\chi^2$ r.v., we see that the narrow-impulse effect is more evident at high SNR. Now the SINR in a large-scale (LS) MIMO system can be approximated by $\overleftarrow{\Gamma}_l^{\text{LS}} = \frac{1-\overleftarrow{\beta}_l}{\overleftarrow{\beta}_l + \sigma^2/(p \cdot 2M)}$ and the corresponding confidence level is

$$\begin{aligned}\Pr(\overleftarrow{\Gamma}_l^{\text{LS}} \geq \overleftarrow{\gamma_l}) &= \Pr\left(\overleftarrow{\beta}_l \leq \frac{1-\frac{\sigma^2\overleftarrow{\gamma}_l}{p}\frac{1}{2M}}{1+\overleftarrow{\gamma_l}}\right)\\ &= I_{\frac{1-\frac{\sigma^2\overleftarrow{\gamma}_l}{p}\frac{1}{2M}}{1+\overleftarrow{\gamma_l}}}\left(L-l, M-L+l\right).\end{aligned} \quad (29)$$

TABLE II

CONFIDENCE LEVEL OF $\overleftarrow{\Gamma}_l \geq \overleftarrow{\gamma}_l$

| $\overleftarrow{\gamma}_l$ | $p/\sigma^2$ | $M$ | $\bar{L}$ | Eq.(28) | Eq.(29) |
|---|---|---|---|---|---|
| | | 1 | 1 | 60.65% | 1 |
| 1 | 1 | 3 | 2 | 60.65% | 65.97% |
| | | 12 | 6 | 67.19% | 67.67% |
| 1 | 1 | 1 | 1 | > 60% | 1 |
| | 10 | | | > 95% | 1 |
| 1 | 1 | $3\bar{L}$ | $2 \leq \bar{L} \leq 6$ | > 94% | |
| | 10 | | | > 96% | |
| 10 | 10 | $16\bar{L}$ | $2 \leq \bar{L} \leq 6$ | > 93% | |
| | 100 | | | | |

In addition, at asymptotically high SNR, we can approximate the SINR by $\overleftarrow{\Gamma}_l^{\text{high}} = \frac{1-\overleftarrow{\beta}_l}{\overleftarrow{\beta}_l}$ and further simplify (29) to

$$\Pr(\overleftarrow{\Gamma}_l^{\text{high}} \geq \overleftarrow{\gamma}_l) = I_{\frac{1}{1+\overleftarrow{\gamma_l}}}\left(L-l, M-L+l\right). \quad (30)$$

Table II lists the confidence level of the worst user, i.e., $l = 1$, calculated by (28) and (29) with different parameters. The first three rows aim to show the accuracy of the approximation (29). With SNR $= 1$, the error is acceptable at moderate $M$. As explained before, to achieve the same accuracy, $M$ can be further decreased as SNR increases. The next two rows show that the performance of the best user is noise-limited and can be improved by raising SNR. However, starting from the second best user, the performance is interference-limited, which is dominated by the number of antennas $M$ rather than SNR.

Next, we give the confidence level of the conditional SINR given $\overleftarrow{\mathbf{h}}_l$, as well as the conditional expected rate as below

$$\begin{aligned}&\Pr(\overleftarrow{\Gamma}_l \geq \overleftarrow{\gamma}_l | \overleftarrow{\mathbf{h}}_l)\\ &\qquad = I_{\frac{1-\frac{\sigma^2\overleftarrow{\gamma}_l}{p}\frac{1}{||\overleftarrow{\mathbf{h}}_l||^2}}{1+\overleftarrow{\gamma_l}}}\left(L-l, M-L+l\right),\end{aligned} \quad (31)$$

$$\begin{aligned}&\mathbb{E}\left[\overleftarrow{R}_l | \overleftarrow{\mathbf{h}}_l\right]\\ &\qquad = \log\left(1+\frac{\sigma^2}{p||\overleftarrow{\mathbf{h}}_l||^2}\right)\\ &\qquad\quad - \int_0^1 \log\left(x+\frac{\sigma^2}{p||\overleftarrow{\mathbf{h}}_l||^2}\right) f_{\overleftarrow{\beta}_l}(x)\,dx.\end{aligned} \quad (32)$$

As discussed in the previous subsection, the conditional expected rate requires numerical computation. For further analysis, we adopt its high SNR approximation given by

$$\mathbb{E}\left[\overleftarrow{R}_l | \overleftarrow{\mathbf{h}}_l\right] \approx \mathbb{E}\left[-\log(\overleftarrow{\beta}_l)\right] = \psi(M) - \psi(L-l). \quad (33)$$

(33) implies that at sufficiently high SNR, the expected rate under AOBF is regardless of the transmit SNR nor the channel strength because it is limited by interference. For the best user who experiences no interference, the approximation will be infinity. In fact, the best user can obtain its exact rate easily by knowing its own CSI vector norm and transmit SNR. Because approximating a sufficiently high rate by infinity will not affect

[3] It can be computed using MATLAB functions integral(), hypergeom() and besselk().

what follows in this paper, (33) is formally adopted as a unified characterization of the conditional expected rate for all users.

From the above results, we see that all other things being equal, the conditional expected rate and the SINR confidence level under both ZF-SIC and AOBF increases with $l$. Under ZF-SIC, larger $l$ indicates larger diversity gain in the effective interference-free channel. Under AOBF, larger $l$ indicates less interference. Therefore, $l$ can be used to represent the class of service within the cluster. Users urge for higher classes of service. This observation inspires us to design a user association and scheduling mechanism based on the concept of auction in next section.

## IV. Auction-Based User Association and Scheduling

In this section, we resume the BS index $[k]$ on the superscript and replace the class of service $l$ by the mapping function $\phi_n^{[k]}$. We first specify the utility function and QoS criteria according to the performance analysis in the previous section, then propose an auction-based user association and scheduling mechanism adapting to both ZF-SIC and AOBF, and finally discuss the economic properties of the mechanism.

### A. Utility Function and QoS Criteria

It is reasonable to assume that the utility of a user depends on its rate in communication systems. Due to dynamic traffic concerned in this paper, the conditional expected rate is a suitable metric. We use a unified notation $r_n^{[k]}$ to represent the conditional expected rate of user $n$ admitted into cluster $k$, according to (24) with ZF-SIC in the uplink and (33) with AOBF in the downlink. When user $n$ is active, the conditional expected rate is given by

$$\tilde{r}_n^{[k]} = \eta_n^{[k]} + \rho_{\phi_n^{[k]}}, \tag{34}$$

where $\eta_n^{[k]}$ captures the terms regardless of $\phi_n^{[k]}$:

$$\eta_n^{[k]} = \begin{cases} \log\left(\frac{p_n}{\sigma^2}\right) + \log\left(\|\mathbf{h}_n^{[k]}\|^2\right) - \psi(M^{[k]}), & \text{ZF-SIC} \\ \psi(M^{[k]}), & \text{AOBF}, \end{cases} \tag{35}$$

and $\rho_{\phi_n^{[k]}}$ is the part that increases with $\phi_n^{[k]}$:

$$\rho_{\phi_n^{[k]}} = \begin{cases} \psi(M^{[k]} - L^{[k]} + \phi_n^{[k]}), & \text{ZF-SIC} \\ -\psi(L^{[k]} - \phi_n^{[k]}), & \text{AOBF} \end{cases} \tag{36}$$

Recall that users with $\phi_n^{[k]} = 1$ share the channel via TDMA. The resulted rate is averaged over time. Therefore, $r_n^{[k]}$ can be expressed as

$$r_n^{[k]} = \begin{cases} \frac{1}{|\mathcal{N}^{[k]}| - L^{[k]} + 1} \tilde{r}_n^{[k]}, & \phi_n^{[k]} = 1 \\ \tilde{r}_n^{[k]}, & 2 \le \phi_n^{[k]} \le L^{[k]}. \end{cases} \tag{37}$$

It is clear that the conditional expected rate $r_n^{[k]}$ increases with the class of service $\phi_n^{[k]}$.

Now we can assume that the utility of a user $n$ is a function of its conditional expected rate $r_n^{[k]}$, its valuation of the rate $\theta_n$ and the price it pays $\pi_{\phi_n^{[k]}}$, which takes the following form

$$u_n^{[k]} = \theta_n \, r_n^{[k]} - \pi_{\phi_n^{[k]}}. \tag{38}$$

In support of the expected rate, a suitable measure of QoS is the outage probability for a given threshold rate, which can be easily converted to the SINR confidence level. We assume user $n$ has a QoS requirement given by

$$\Pr(\tilde{\Gamma}_{\phi_n^{[k]}}^{[k]} \ge \gamma_n | \tilde{\mathbf{h}}_n^{[k]}) \ge \zeta_n, \tag{39}$$

where $\tilde{\Gamma}_{\phi_n^{[k]}}^{[k]}$ and $\tilde{\mathbf{h}}_n^{[k]}$ are unified representations of the SINR and CSI vector of user $n$ accepted into cluster $k$ in either uplink or downlink. The left side of (39) can be calculated by (22) under ZF-SIC or (31) under AOBF. For given $M^{[k]}$, $L^{[k]}$ and transmit SNR, (39) can be further converted to a minimum class requirement as follows

$$\phi_n^{[k]} \ge \bar{l}_n^{[k]}, \quad \forall k \in \mathcal{K}. \tag{40}$$

For BS $k$ seeing a dynamic traffic, the unconditional SINR confidence level can be adopted to guarantee the worst-case QoS. To be specific, the SINR confidence level of the user in service class 1 should be larger than some threshold:

$$\Pr(\tilde{\Gamma}_1^{[k]} \ge \gamma^{[k]}) \ge \zeta^{[k]}. \tag{41}$$

The left side of (41) can be calculated by (20) under ZF-SIC and (28) under AOBF. For given $M^{[k]}$ and transmit SNR, (41) can be converted to a cluster capacity requirement as follows

$$L^{[k]} \le \bar{L}^{[k]}. \tag{42}$$

For example, with ZF-SIC in the uplink, letting $\bar{L}^{[k]} = M^{[k]}$ guarantees >95% confidence for the worst-case SINR to exceed 10 dB at transmit SNR $=$ 20 dB. With AOBF in the downlink, letting $\bar{L}^{[k]} = 8$ guarantees >80% confidence for the worst-case SINR to exceed 10 dB given $M^{[k]} = 100$ and total transmit SNR $=$ 20 dB. These parameters are also adopted in simulation in Section V.

### B. Auction-Based Mechanism and Pricing Scheme

The goal of user $n$ is to select the BS that maximizes its utility. In mathematical terms, it solves the following problem:

$$\text{(P1)} \qquad \max_{k \in \mathcal{K}} u_n^{[k]},$$

where $u_n^{[k]}$ is given by (38). (P1) is formulated without additional QoS requirement. We consider (P1) because it can be shown that the utility is always positive (the proof is in Proposition 3 of Section IV.C) so that the user always has incentive to select one BS. To relieve users' concern about QoS, we assume that BSs set the cluster capacity $\bar{L}^{[k]}$ to guarantee an acceptable worst-case QoS.

Another rational assumption is that BSs' joint goal is to maximize the social welfare $\sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}^{[k]}} \theta_n \, r_n^{[k]}$. A serious concern in achieving this goal is accommodation of all users with limited resource. The classic VCG auction is a well-known approach for socially optimal resource allocation. However, in the VCG auction, the resource is limited so that users with lowest bids may get nothing, which leads to the admission and outage control problem in our scenario. In the SPAC mechanism in [17], the resource in the lowest QoS level is unlimited so that all remaining users can be accommodated without affecting the QoS, which may not be practical in genuine communication systems. A second concern is the computational complexity. The VCG auction and the SPAC mechanism in [17] require all $K$ BSs to make joint decisions. In addition, arrival or departure of a user may cause re-association and re-scheduling of all underlying users in lower classes.

To address the first concern, we divide the users into two categories:

1) paying users with the $(L^{[k]}-1)$ highest bids, assigned with class $\phi_n^{[k]} > 1$ and price $\pi_{\phi_n^{[k]}} > 0$.
2) free users with the $(|\mathcal{N}^{[k]}| - L^{[k]} + 1)$ lowest bids, assigned with class $\phi_n^{[k]} = 1$ and $\pi_1 = 0$;

The social welfare maximization problem is relaxed to maximization of all paying users' welfare, formulated as follows

$$\text{(P2)}\quad \max_{\phi_n^{[k]}>1,\quad \forall n,k} \sum_{k\in\mathcal{K}} \sum_{n:\phi_n^{[k]}>1} \theta_n\, r_n^{[k]}$$
$$\text{s.t.} L^{[k]} \le \bar{L}^{[k]},\ \forall k \in \mathcal{K},$$

To address the second concern, we assume that users comply with a no-regret association policy considering their inertia to stick with the same BS. Then the post-admission sub-problem of (P2) at each BS $k$ is

$$\text{(P3)}\quad \max_{\phi_n^{[k]}>1,\quad \forall n\in\mathcal{N}^{[k]}} \sum_{n:\phi_n^{[k]}>1} b_n\, r_n^{[k]}$$
$$\text{s.t.} L^{[k]} \le \bar{L}^{[k]},$$

where $b_n$ denotes the declared valuation of user $n$. The solution to (P3) is obvious. Since $r_n^{[k]}$ increases with $\phi_n^{[k]}$, BS $k$ assigns higher class $\phi_n^{[k]}$ to user $n$ with higher bid $b_n$.

Notice that (P3) uses the declared valuation $b_n$ instead of the true valuation $\theta_n$. The remaining problem is to design a strategy-proof pricing scheme, given by

$$\pi_{\phi_n^{[k]}} = \begin{cases} 0, & \phi_n^{[k]} = 1 \\ (r_n^{[k]}(2) - r_n^{[k]}(1))\bar{b}_1, & \phi_n^{[k]} = 2 \\ \pi_{\phi_n^{[k]}-1} + (\rho_{\phi_n^{[k]}} - \rho_{\phi_n^{[k]}-1})\bar{b}_{\phi_n^{[k]}-1}, & 2 < \phi_n^{[k]} \le L^{[k]}, \end{cases} \tag{43}$$

where $\bar{b}_l$ is a brief representation of the highest bid among user(s) in class $l$, i.e., $\bar{b}_l = \max_{n:\phi_n^{[k]}=l}\{b_n\}$. $\bar{b}_{l-1}$ belongs to the user in class $l-1$ if user $n$ is admitted into a class $\phi_n^{[k]} \ge l$, but class $l$ if user $n$ is not admitted. $r_n^{[k]}(l)$ is a brief representation of $r_n^{[k]}$ when $\phi_n^{[k]} = l$.

The interpretation of (43) is to charge a user for the implicit cost of scheduling it as a paying user instead of a free user. Admitting user $n$ as a free user has no additional cost, so $\pi_1 = 0$. To see the effect of admitting user $n$ as a paying user, we denote the bidding profile of users in cluster $k$ by $\mathbf{b}$ and the bidding profile without $b_n$ by $\mathbf{b}_{-n}$. The optimal class assignment depends on the bidding profile and can be written as $\phi_n^{[k]*}(\mathbf{b})$. We first examine the explicit effect on the declared welfare of all current users in cluster $k$ by admitting user $n$ as a paying user:

$$\begin{aligned} \Delta W^{[k]} &= \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(\mathbf{b}_{-n})) - \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(\mathbf{b})) \\ &= \sum_{q=3}^{\phi_n^{[k]*}(\mathbf{b})} (\rho_q - \rho_{q-1})\bar{b}_{q-1} + (r_x^{[k]}(2) - r_x^{[k]}(1))\bar{b}_1 + \Delta w_1^{[k]}, \end{aligned} \tag{44}$$

where $\Delta w_1^{[k]}$ denotes the inevitable effect on the free users' welfare due to admission of one more user, which is not counted in the cost of scheduling user $n$ as a paying user. $x$ is the index of the user in class 2 if user $n$ is not admitted. $(r_x^{[k]}(2) - r_x^{[k]}(1))\bar{b}_1$ is the explicit effect on user $x$.

However, user $n$ has no reason to remedy this explicit effect, which depends on the local transmit power and channel strength of user $x$ in the uplink case. Instead, this part of implicit resource occupied by user $n$ should be $(r_n^{[k]}(2) - r_n^{[k]}(1))\bar{b}_1$. The implicit effect on all users in cluster $k$ is denoted by

$$\begin{aligned} \Delta\tilde{W}^{[k]} &= \sum_{q=3}^{\phi_n^{[k]*}(\mathbf{b})} (\rho_q - \rho_{q-1})\bar{b}_{q-1} + (r_n^{[k]}(2) - r_n^{[k]}(1))\bar{b}_1 + \Delta w_1^{[k]} \\ &= \pi_{\phi_n^{[k]*}(\mathbf{b})} + \Delta w_1^{[k]}. \end{aligned} \tag{45}$$

Therefore, $\pi_{\phi_n^{[k]*}(\mathbf{b})} = \Delta\tilde{W}^{[k]} - \Delta w_1^{[k]}$ can be interpreted as the implicit cost of scheduling user $n$ as a paying user. In fact, in the downlink case where $(r_x^{[k]}(2) - r_x^{[k]}(1))$ does not depend on local parameters of user $x$, the explicit effect is equal to the implicit effect, i.e., $\Delta W^{[k]} = \Delta\tilde{W}^{[k]}$.

A direct consequence of the pricing scheme is that after admitting a new paying user, the declared utilities of paying users remain unchanged, because the change in price offsets the change in welfare

$$\pi_l - \pi_{l-1} = (r_n^{[k]}(l) - r_n^{[k]}(l-1))\bar{b}_{l-1}, \quad 2 \le l \le \phi_n^{[k]}, \tag{46}$$

where $\bar{b}_{l-1}$ is defined after (43). User $n$ receives rate $r_n^{[k]}(l)$ and rate $r_n^{[k]}(l-1)$ before and after the BS admitting the new paying user, respectively.

Now we can list the steps of the auction-based user association and scheduling mechanism as follows.

It can be seen that the proposed mechanism requires limited information exchange in steps 2 and 6. The full CSI vector for beamformer design is only needed at the designated BS after the user-BS association is established. The computational complexity of the proposed algorithm comes from the BS side and the user side. For the BS side, the complexity is mainly from computing the temporary price by equation (43) whose complexity is linear with the number of users. So the complexity of BS side is $O(KN)$, where $K$ is the number of BS and $N$ is the number of users. For the user side, the complexity is mainly from solving the problem (P1) whose

**User Association and Scheduling Mechanism**

```
1  if user $n$ arrives then
2      User $n$ reveals its bid $b_n$ to all $K$ BSs;
3      for $k = 1$ to $K$ do
4          BS $k$ assigns the temporary class $\phi_n^{[k]}$ in ascending
           order of all bids $b_m, \forall m \in \mathcal{N}^{[k]}$;
5          BS $k$ sets the temporary price $\pi_{\phi_n^{[k]}}$ by (43);
6          BS $k$ feeds back $||\mathbf{h}_n^{[k]}||^2$ (only in the uplink case),
           $M^{[k]}$, $L^{[k]}$, the class $\phi_n^{[k]}$ and the price $\pi_{\phi_n^{[k]}}$ to the
           user.
7      end
8      User $n$ selects a BS $k_n^*$ by solving (P1);
9      BS $k_n^*$ admits user $n$, sets classes and prices as in
       steps 4 and 5 for all user $m \in \mathcal{N}^{[k_n^*]}$ with $b_m \leq b_n$;
10 end
11 if user $n$ departs then
12     BS $k_n^*$ assigns classes in ascending order of $b_m$ and
       sets prices by (43), for all user $m \in \mathcal{N}^{[k_n^*]}$ with
       $b_m < b_n$;
13 end
```

complexity is linear with the number of BS, thus the overall user side complexity is also $O(KN)$. Therefore, the proposed algorithm has low complexity thus is feasible for practical system.

### *C. Discussion on Economic Properties*

The proposed mechanism inherits the VCG auction. The main differences are clarified as follows: 1) our mechanism accommodates all users with limited resources; 2) the pricing scheme is based on the implicit cost of admitting a paying user instead of the explicit cost; 3) our mechanism is conducted in a multi-cell system and in a distributed manner subject to the no-regret association policy. Therefore, it is worth further discussion on some crucial economic properties, e.g., truthfulness, individual rationality and social optimality, declared as follows.

*Proposition 2: User $n$ will reveal its bid $b_n$ truthfully to maximize its utility.*

*Proof:* Assume without loss of generality that user $n$ is admitted into cluster $k$. We first consider the case that user $n$ is scheduled as a paying user. We express the utility as a function of the bidding profile and suppose to the contrary that there exists a $b_n \neq \theta_n$ such that

$$u_n^{[k]}(b_n, \mathbf{b}_{-n}) > u_n^{[k]}(\theta_n, \mathbf{b}_{-n}). \tag{47}$$

According to (44) and (45),

$$\begin{aligned}
&\pi_{\phi_n^{[k]*}(b_n,\mathbf{b}_{-n})} \\
&\quad = \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(\mathbf{b}_{-n})) \\
&\qquad - \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(b_n,\mathbf{b}_{-n})) + (r_n^{[k]}(2) - r_n^{[k]}(1))\bar{b}_1 \\
&\qquad - (r_x^{[k]}(2) - r_x^{[k]}(1))\bar{b}_1 - \Delta w_1^{[k]}.
\end{aligned} \tag{48}$$

Except for the second term, all other terms are regardless of the bid $b_n$. Substitute (48) into the utility given by (38) and cancel the same terms, (47) is equivalent to

$$\begin{aligned}
&\theta_n r_n^{[k]}(\phi_n^{[k]*}(b_n,\mathbf{b}_{-n})) \\
&\quad + \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(b_n,\mathbf{b}_{-n})) \\
&\quad > \theta_n r_n^{[k]}(\phi_n^{[k]*}(\theta_n,\mathbf{b}_{-n})) + \sum_{m\neq n} b_m r_m^{[k]}(\phi_m^{[k]*}(\theta_n,\mathbf{b}_{-n})),
\end{aligned} \tag{49}$$

where $\phi_n^{[k]*}(b_n,\mathbf{b}_{-n})$ is based on optimal solution to (P3) for the bidding profile $(b_n,\mathbf{b}_{-n})$. For the bidding profile $(\theta_n,\mathbf{b}_{-n})$, it is an arbitrary solution. (49) suggests that the arbitrary solution yields a larger value in (P3) than the optimal solution $\phi_n^{[k]*}(\theta_n,\mathbf{b}_{-n})$, which is a contradiction.

Next we show that a paying user has no incentive to understate its bid for being a free user. Assume that when user $n$ reveals its bid truthfully, it gets class $l \geq 2$ and $\bar{b}_{l-1}$ is the highest bid among user(s) in class $l-1$. It follows that $\theta_n > \bar{b}_{q-1}$ for $2 < q \leq l$. According to (46), the price paid by user $n$ is

$$\pi_l = \sum_{q=2}^{l}(r_n^{[k]}(q) - r_n^{[k]}(q-1))\bar{b}_{q-1} < (r_n^{[k]}(l) - r_n^{[k]}(1))\theta_n. \tag{50}$$

Therefore, $u_n^{[k]}(\theta_n,\mathbf{b}_{-n}) = \theta_n r_n^{[k]}(l) - \pi_l$ is greater than $\theta_n r_n^{[k]}(1) = u_n^{[k]}(b_n,\mathbf{b}_{-n})$.

Similarly, a free user $n$ has no incentive to overstate its bid. In this case, assume that when user $n$ is overstating, it gets class $l \geq 2$ and $\bar{b}_{l-1}$ is the highest bid among user(s) in class $l-1$. It follows that $\theta_n < \bar{b}_{q-1}$ for $2 < q \leq l$. When user $n$ is overstating, the price is $\pi_l = \sum_{q=2}^{l}(r_n^{[k]}(q) - r_n^{[k]}(q-1))\bar{b}_{q-1} > (r_n^{[k]}(l) - r_n^{[k]}(1))\theta_n$. Therefore, $u_n^{[k]}(\theta_n,\mathbf{b}_{-n}) = \theta_n r_n^{[k]}(1) > \theta_n r_n^{[k]}(l) - \pi_l = u_n^{[k]}(b_n,\mathbf{b}_{-n})$. □

*Proposition 3: The auction is individual rational at sufficiently high SNR.*

*Proof:* Individual rationality means that each user gets a non-negative utility and therefore has the incentive to participate in the auction. The condition of sufficiently high SNR is to ensure that the approximation of the rate under ZF-SIC is non-negative, which is trivially satisfied in reality.

First, a free user $n$ pays zero price and gets a rate $r_n^{[k]}(1) > 0$. Therefore, its utility $u_n^{[k]} = \theta_n r_n^{[k]}(1) > 0$. For a paying user $n$ in class $l \geq 2$, the utility $u_n^{[k]} = \theta_n r_n^{[k]}(l) - \pi_l$. Substitute (50) into $\pi_l$ we obtain

$$u_n^{[k]} = \theta_n r_n^{[k]}(l) - \sum_{q=2}^{l}(r_n^{[k]}(q) - r_n^{[k]}(q-1))\bar{b}_{q-1}. \tag{51}$$

According to the inequality in (50), we obtain the r.h.s. of (50) is $\theta_n r_n^{[k]}(l) - \sum_{q=2}^{l}(r_n^{[k]}(q) - r_n^{[k]}(q-1))\bar{b}_{q-1} > \theta_n r_n^{[k]}(1)$. Because $\theta_n r_n^{[k]}(1)$ is always positive as defined in (38), so $u_n^{[k]} > 0$. □

*Proposition 4: The mechanism is socially near-optimal.*

*Proof:* The mechanism is implemented in a distributed manner at each BS and each user. At BS side, (P3) maximizes the true social welfare of paying users within each cluster, because users reveal the bids truthfully. At user side, solving (P1) is equivalent to maximizing the implicit gain in social welfare of all paying users. According to (45), $\pi_{\phi_n^{[k]}} = \Delta\tilde{W}^{[k]} - \Delta w_1^{[k]}$ is the implicit effect on all other paying users due to admission of a paying user $n$. Therefore, the utility $u_n^{[k]} = \theta_n r_n^{[k]} - \pi_{\phi_n^{[k]}}$ in (P1) is the implicit gain in social welfare of all paying users in all clusters. If the explicit gain $\theta_n r_n^{[k]} - (\Delta W^{[k]} - \Delta w_1^{[k]})$ is maximized in (P1), the solution given by (P1) and (P3) coincides with the optimal solution to (P2) subject to the no-regret policy. The error term between the explicit gain and implicit gain is $\epsilon = \Delta\tilde{W}^{[k]} - \Delta W^{[k]}$. In the downlink case, the error term $\epsilon = 0$. Therefore, the proposed mechanism is socially near-optimal in the sense that it maximizes the social welfare of paying users subject to the no-regret association policy with an error term $\epsilon = \Delta\tilde{W}^{[k]} - \Delta W^{[k]}$. □

Notice that the welfare is a scaled measure of the rate. The price is a penalty term that represents the harm on the welfare of other users. The mechanism can be viewed as a decentralized algorithm besides the economic interpretation. BSs feed back the penalty as if the user was admitted. The user chooses the BS that maximizes its welfare adjusted by the penalty, which also maximizes the net gain in welfare in the whole system.

## V. Simulation Results

We model the dynamic traffic as an M/M/1 queue. Users arrive as a Poisson process at rate $\lambda$ and stay in the system for an exponentially distributed time with parameter $\mu$. All users' valuation $\theta_n$ is uniformly distributed on $[0, 2]$. We start with an empty system and monitor the system for $Q$ consecutive users, or equivalently saying, along with $2Q$ arrival and departure events in the long term.

Two comparison schemes, RSS-based selection and the greedy selection (GS), are provided as benchmarks. With RSS-based selection, user $n$ selects the BS with the largest CSI vector norm $||\mathbf{h}_n^{[k]}||$. With the greedy selection, user $n$ selects the BS with the largest exact rate; In the case of heavy traffic and no BS can offer a positive rate, user $n$ selects a BS randomly. For a fair comparison, the two schemes also adopt ZF-SIC in the uplink and AOBF in the downlink. However, it is hard to design a sophisticated user scheduling mechanism for the two comparison schemes. Greedy user scheduling together with the greedy BS selection can offer the optimal performance at the cost of huge complexity and is not practical in dynamic traffic. Therefore, the users are scheduled in order of arrival time in the comparison schemes.

Fig. 1 and Fig. 2 plot the social welfare and aggregate throughput for $Q = 20000$ consecutive users in an uplink MU-MIMO OFDM system of $K = 3$ BSs, each with $M^{[k]} = 8$ antennas. The statistics are averaged over time. ZF-SIC is adopted for all presented mechanisms. The cluster capacity $\bar{L}^{[k]}$ is set to be the same as $M^{[k]}$, which yields >95% confidence for the worst-case SINR to exceed 10 dB at SNR = 20 dB.

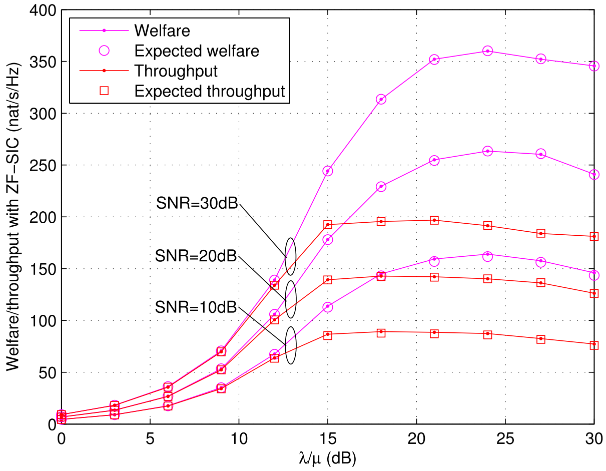


Fig. 1. Expected and actual social welfare and aggregate throughput by the proposed mechanism, with ZF-SIC in the uplink MU-MIMO network where $K = 3$, $M^{[k]} = \bar{L}^{[k]} = 8$ and SNR $= p_n/\sigma^2 = 10$ to 30 dB.

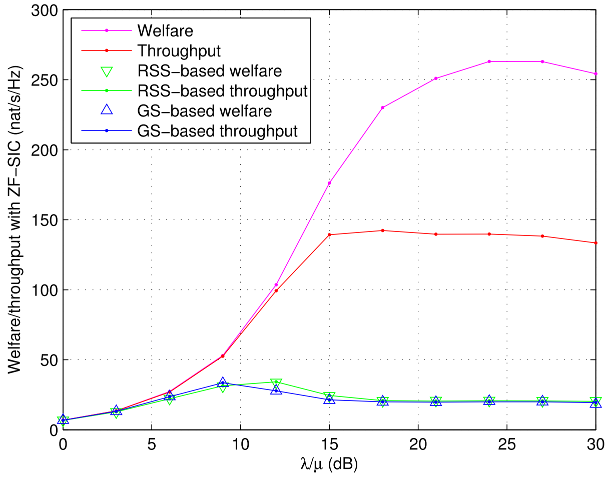


Fig. 2. Social welfare and aggregate throughput by different mechanisms, with ZF-SIC in the uplink MU-MIMO network where $K = 3$, $M^{[k]} = \bar{L}^{[k]} = 8$ and SNR $= p_n/\sigma^2 = 20$ dB.

It can be seen in Fig. 1 that the expected welfare and expected throughput, which are the metrics adopted in the proposed mechanism, highly coincide with the actual welfare and throughput at medium and high SNR. Since users' valuation is uniformly distributed on $[0, 2]$, the social welfare is almost twice the aggregate throughput at high $\lambda/\mu$ ratio in the proposed mechanism. This is because with sufficient users, the proposed mechanism assigns better resources to users with higher valuation.

Fig. 2 provides comparisons with RSS-based selection and greedy selection at SNR = 20 dB. The proposed mechanism outperforms RSS-based selection and greedy selection in both welfare and throughput, and the gain becomes more substantial at a higher $\lambda/\mu$ ratio. At low $\lambda/\mu$ ratio, there may be only one or two users in each cluster constantly. All mechanisms offer the highest classes to the users, so the performance is not differentiated. As mentioned before, our mechanism achieves a social welfare almost twice as high as the aggregate throughput at high $\lambda/\mu$ ratio. On the other hand, the comparison

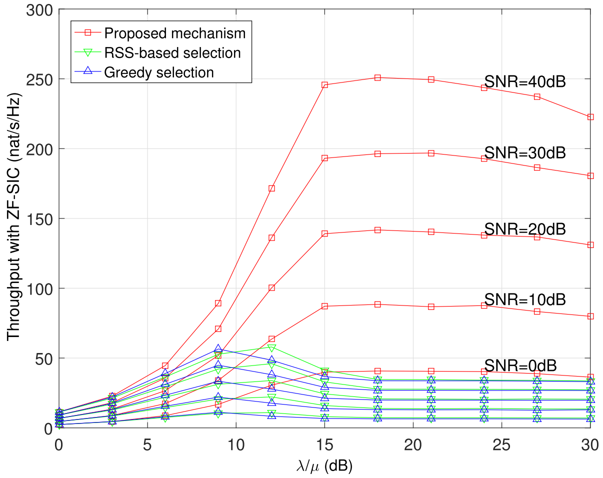


Fig. 3. Aggregate throughput with ZF-SIC in the uplink MU-MIMO network where $K = 3$, $M^{[k]} = \bar{L}^{[k]} = 8$ and $\mathrm{SNR} = p_n/\sigma^2 = 0$ to 40 dB.

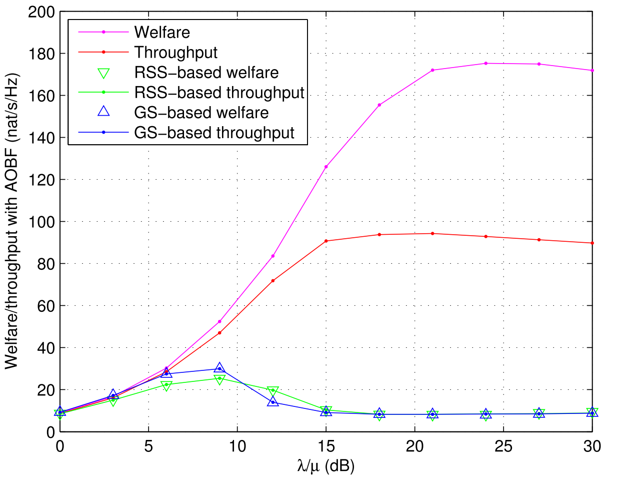


Fig. 5. Social welfare and aggregate throughput by different mechanisms, with AOBF in the downlink MU-MIMO network where $K = 3$, $M^{[k]} = 100$, $\bar{L}^{[k]} = 8$ and $\mathrm{SNR} = P^{[k]}/\sigma^2 = 20$ dB.

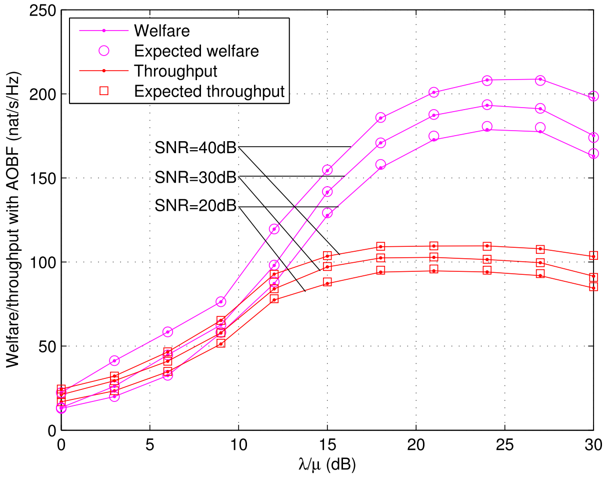


Fig. 4. Expected and actual social welfare and aggregate throughput by the proposed mechanism, with ZF-SIC in the uplink MU-MIMO network where $K = 3$, $M^{[k]} = \bar{L}^{[k]} = 8$ and $\mathrm{SNR} = p_n/\sigma^2 = 10$ to 30 dB.

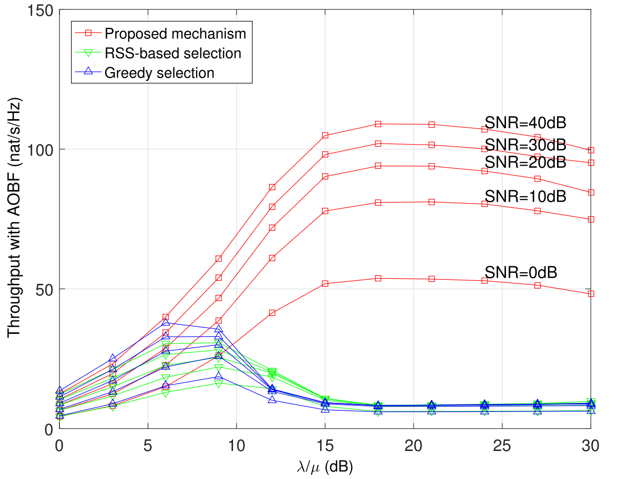


Fig. 6. Aggregate throughput with AOBF in the downlink MU-MIMO network where $K = 3$, $M^{[k]} = 100$, $\bar{L}^{[k]} = 8$ and $\mathrm{SNR} = P^{[k]}/\sigma^2 = 0$ to 40 dB.

mechanisms lack the ability to maximize true social welfare, because they are not equipped with a truth-telling mechanism for acquisition of users' valuation. Our mechanism also has a dominant performance in terms of throughput, which is the social welfare normalized by valuation. The underlying reason is that in the comparison schemes, users are selfish and competitive; while in our mechanism, a user maximizing its own utility is approximately maximizing the gain in social welfare and the throughput.

Fig. 3 shows the aggregate throughput at various SNR levels under the same system setting as in Fig. 1 and Fig. 2. The proposed mechanism is dominant over all SNR regimes. Raising SNR can enhance the performance in all schemes, because the performance is noise-limited under ZF-SIC.

Fig. 4 to Fig. 6 demonstrate the social welfare and aggregate throughput with AOBF in a downlink large-scale MIMO system. The system has $K = 3$ BSs, each with $M^{[k]} = 100$ antennas and a cluster capacity $\bar{L}^{[k]} = 8$. With the total transmit $\mathrm{SNR} = P^{[k]}/\sigma^2 = 20$ dB as in Fig. 5, there is >80% confidence for the worst-case SINR to exceed 10 dB. The performance of the presented mechanisms under AOBF follows the same trend as under ZF-SIC. It is worth mentioning that for the best user in the highest class, we use the exact welfare and throughput to replace the expected welfare and throughput, which will be infinity according to the approximation given by (33). This replacement is practical, because the best user can easily evaluate the exact rate in the downlink.

It is noticed in Fig. 4 that at low and medium SNR, the expected welfare and throughput are slightly larger than the exact values at high $\lambda/\mu$ ratio. This is because the transmit power $P^{[k]}$ is equally allocated to more users so that the high SNR approximation (33) is not accurate enough. For example, when $\mathrm{SNR} = P^{[k]}/\sigma^2 = 20$ dB and the cluster of capacity $\bar{L}^{[k]} = 8$ is full, each user's SNR is only around 11 dB. Despite making the approximations more accurate, high SNR does not offer a substantial gain in throughput, as shown in Fig. 6, because of the interference-limited nature of AOBF.

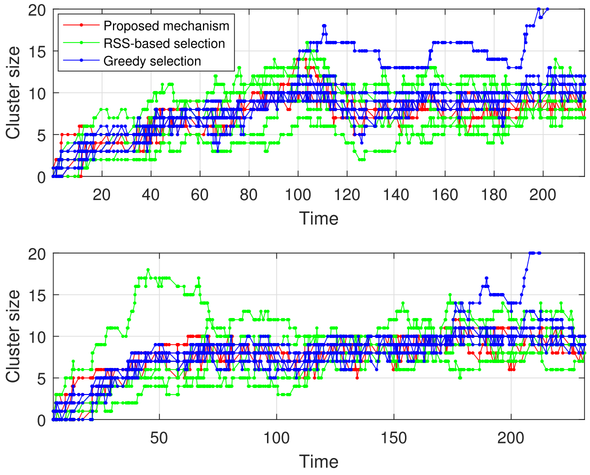


Fig. 7. Real-time cluster size in two independent realizations with ZF-SIC in the uplink MU-MIMO network where $K = 5$, $M^{[k]} = \bar{L}^{[k]} = 10$, SNR $= p_n/\sigma^2 = 20$ dB and $\lambda/\mu = 50$ or 17 dB.

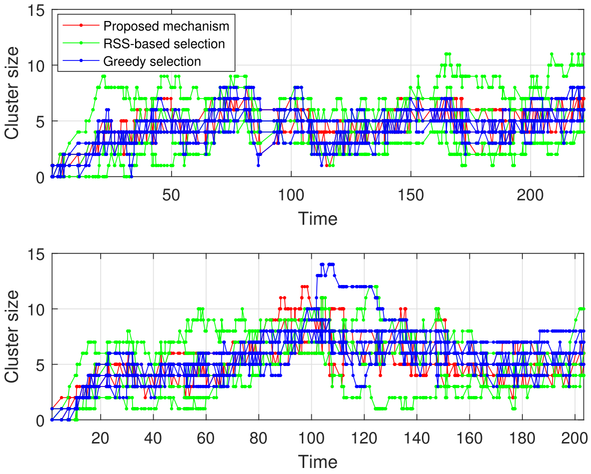


Fig. 8. Real-time cluster size in two independent realizations with AOBF in the downlink MU-MIMO network where $K = 5$, $M^{[k]} = 100$, $\bar{L}^{[k]} = 8$, SNR $= P^{[k]}/\sigma^2 = 20$ dB and $\lambda/\mu = 30$ or 12.7 dB.

Fig. 7 and Fig. 8 monitor the real-time cluster size over the first 400 arrival/departure events with $K = 5$ BSs in the systems adopting ZF-SIC and AOBF, respectively. The $\lambda/\mu$ ratio is selected so that the mean of the cluster size is around $\bar{L}^{[k]}$. The five intertwined red lines display the sizes of five clusters under the proposed mechanism, while the five green lines and five blue lines display the cluster sizes under the comparison schemes.

It is observed in all the realizations that the red and blue lines have a smaller spread than the green lines, which means the cluster size is more balanced under the proposed mechanism and greedy selection. However, greedy selection may have unexpected bad performance in heavy traffic when the clusters are all full and the newly arrived users can only make random selections. When the traffic stays heavy, the effect of unbalanced load might sustain for a fairly long time under RSS-based selection and greedy selection. On the contrary, the proposed mechanism is more robust to heavy traffic. Assume the system parameters are equal at different BSs. When all clusters are full, a newly arrived user scheduled as a free user tends to select the cluster with a smaller size. A newly arrives bidding higher than some paying user will be scheduled as a paying user. The cluster admitting the new user tends to consist of more users with higher bids, which poses a higher barrier for the next user to join. Therefore, the traffic load under the proposed mechanism is more balanced, stable, and robust to heavy traffic.

## VI. Conclusion

We propose a user association and scheduling mechanism in multi-cell MU-MIMO systems adopting ZF-SIC in the uplink and AOBF in the downlink. A comprehensive analysis is conducted for ZF-SIC and AOBF, for BSs and users to evaluate the SINR confidence level and rate, and develop the QoS criteria and utility function. A user association and scheduling mechanism is then proposed to exploit the differentiated classes of services in ZF-SIC and AOBF. From the perspective of game theory, the proposed mechanism is an auction equipped with a strategy-proof pricing scheme. From the perspective of algorithm design, it is a distributed algorithm with limited information exchange. The proposed mechanism is near-optimal in social welfare maximization and adapts to dynamic traffic. It also has superior performance in maximizing the throughput and balancing the load.

## Appendix
## Proof of Theorem 1

Before proving the theorem, we introduce the following notations. For a complex matrix $\mathbf{A}$, write $\mathbf{A}{=}\Re\mathbf{A}{+}j\Im\mathbf{A}$ where $\Re\mathbf{A}$ and $\Im\mathbf{A}$ are the real and imaginary parts of $\mathbf{A}$. Denote $\mathbf{A}^{(1)}{=}\begin{bmatrix}\Re\mathbf{A}\\ \Im\mathbf{A}\end{bmatrix}$, $\mathbf{A}^{(2)}{=}\begin{bmatrix}-\Im\mathbf{A}\\ \Re\mathbf{A}\end{bmatrix}$ and $\mathbf{A}^{\ddagger}{=}\begin{bmatrix}\mathbf{A}^{(1)} & \mathbf{A}^{(2)}\end{bmatrix}$. Let $\mathbf{a}_i$ be the $i$th column of matrix $\mathbf{A}$. Then for an isotropically random unitary matrix $\mathbf{U} \in \mathcal{C}^{M\times d}$, we can prove the following lemmas for $\mathbf{U}^{\ddagger} \in \mathcal{R}^{2M\times 2d}$.

*Lemma 1: Columns of $\mathbf{U}^{\ddagger}$ form an (incomplete) orthonormal and symplectic basis of $\mathcal{R}^{2M}$.*

*Proof:* Lemma 1 can be verified easily according to the way we construct $\mathbf{U}^{\ddagger}$. □

*Lemma 2: For $\mathbf{\Omega} \in \mathcal{U}(M)$, $\mathbf{\Omega}^{\ddagger} \in \mathcal{O}(2M) \cap \mathcal{SP}(2M, \mathcal{R})$ is transitive, which means it can transform any (incomplete) orthonormal and symplectic basis to any other.*

*Proof:* Notice that the orthogonal and symplectic structure of $\mathbf{\Omega}^{\ddagger}$ is impervious to matrix multiplication or inverse. It suffices to prove that there exists such $\mathbf{\Omega}^{\ddagger}$ that can transform the standard basis $(\mathbf{e}_1, \ldots, \mathbf{e}_d; \mathbf{f}_1, \ldots, \mathbf{f}_d)$, where $\mathbf{e}_i$ is the $i$th column of $\mathbf{I}_{2M}$ and $\mathbf{f}_i = -\mathbf{J}\mathbf{e}_i$ for $\mathbf{J} = \begin{bmatrix}\mathbf{0} & \mathbf{I}_M\\ -\mathbf{I}_M & \mathbf{0}\end{bmatrix}$, to any orthogonal and symplectic basis $(\mathbf{x}_1, \ldots, \mathbf{x}_d; \mathbf{y}_1, \ldots, \mathbf{y}_d)$. Such $\mathbf{\Omega}^{\ddagger}$ can be found by a method similar to the Gram-schmidt process: (1) For $i = 1, \ldots, d$, let $\boldsymbol{\omega}_i = \mathbf{x}_i$ and $\boldsymbol{\omega}_{M+i} = \mathbf{y}_i$; (2) For $i = d+1, \ldots, M$, find $\boldsymbol{\omega}_i$ orthogonal to all $\boldsymbol{\omega}_j$ and $\boldsymbol{\omega}_{M+j}$ for $j = 1, \ldots, i-1$ as in the Gram-Schmidt process and let $\boldsymbol{\omega}_{M+i} = -\mathbf{J}\boldsymbol{\omega}_i$. □

*Lemma 3: The distribution of $\mathbf{U}^{\ddagger}$ is invariant to orthogonal and symplectic transformations, i.e., $f(\mathbf{U}^{\ddagger}) = f(\mathbf{\Omega}^{\ddagger}\mathbf{U}^{\ddagger})$ for all $\mathbf{\Omega}^{\ddagger} \in \mathcal{O}(2M) \cap \mathcal{SP}(2M, \mathcal{R})$.*

*Proof:* Lemma 3 results directly from $f(\mathbf{U}) = f(\mathbf{\Omega}\mathbf{U})$ for the isotropically random unitary matrix $\mathbf{U}$ and all $\mathbf{\Omega} \in \mathcal{U}(M)$. □

*Lemma 4: $\mathbf{u}_i^{(p)} (i = 1, \ldots, d; p = 1, 2)$ is uniformly distributed on the unit sphere $\mathcal{S}^{2M-1}$.*

*Proof:* According to the transitivity of $\mathbf{\Omega}^{\ddagger}$ in Lemma 2, $\mathbf{\Omega}^{\ddagger}\mathbf{u}_i^{(p)}$ can be any $2M \times 1$ unit vector for $\mathbf{\Omega}^{\ddagger} \in \mathcal{O}(2M) \cap \mathcal{SP}(2M, \mathcal{R})$. According to Lemma 3, $f(\mathbf{u}_i^{(p)}) = f(\mathbf{\Omega}^{\ddagger}\mathbf{u}_i^{(p)})$. Therefore, $\mathbf{u}_i^{(p)}$ is uniformly distributed on $\mathcal{S}^{2M-1}$. □

Based on the above lemmas, Theorem 1 can be proved as follows. Consider the $i$th entry of $\boldsymbol{\alpha}$. We have $\alpha_i^{\ddagger} = \begin{bmatrix} \Re\alpha_i & -\Im\alpha_i \\ \Im\alpha_i & \Re\alpha_i \end{bmatrix} = \mathbf{u}_i^{\ddagger T}\mathbf{n}^{\ddagger}$, where $\mathbf{u}_i^{\ddagger}$ and $\mathbf{n}^{\ddagger}$ both consist of a pair of orthogonal and symplectic basis vectors. As stated before, the distribution of the orthogonal and symplectic basis is invariant to orthogonal and symplectic transformation. Therefore, we can assume without loss of generality that $\mathbf{n}^{\ddagger} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{f}_1 \end{bmatrix}$ where $\mathbf{e}_1$ and $\mathbf{f}_1$ are a pair of standard basis vectors as in Lemma 2.

We first find the distribution of $\Re\alpha_1 = \mathbf{u}_1^{(1)T}\mathbf{n}^{(1)}$ on $[-1, 1]$. Recall that $\mathbf{u}_1^{(1)}$ is uniformly distributed on $\mathcal{S}^{2M-1}$. The probability that $\Re\alpha_1 \in [x, x + dx]$ is proportional to the "lateral area" of a truncated hypercone built out of an (2M-2)-sphere of radius $\sqrt{1 - x^2}$ and of slant height $\frac{dx}{\sqrt{1-x^2}}$, which entails $f_{\Re\alpha_1}(x)dx \sim \left(1 - x^2\right)^{M - \frac{3}{2}} dx$. Therefore, the pdf of $\Re\alpha_1$ can be expressed as

$$f_{\Re\alpha_1}(x) = C(M)\left(1 - x^2\right)^{M - \frac{3}{2}}, \tag{52}$$

for a normalizing constant $C(M)$. Noticing that $x^2$ is Beta$(\frac{1}{2}, M - \frac{1}{2})$ distributed, we have

$$C(M) = \frac{1}{B(\frac{1}{2}, M - \frac{1}{2})}, \tag{53}$$

in terms of the Beta function $B(\cdot, \cdot)$.

Next we find the conditional distribution of $\Im\alpha_1$ given $\Re\alpha_1$. $\Re\alpha_1$ is also the projection of $\mathbf{u}_1^{(2)}$ onto $\mathbf{n}^{(2)}$. Given $\Re\alpha_1 = x$, $\mathbf{u}_1^{(2)}$ has $(2M - 1)$ DoF. The assumption $\mathbf{n}^{(2)} = \mathbf{f}_1$ makes it more comprehensible by fixing the $(M + 1)$th entry of $\mathbf{u}_1^{(2)}$ to be $x$ and leaving the other $(2M - 1)$ entries uniformly distributed on a $(2M - 2)$-sphere of radius $\sqrt{1 - x^2}$. So the conditional pdf of $\Im\alpha_1$ is

$$\begin{aligned} &f_{\Im\alpha_1|\Re\alpha_1}(y|x) \\ &\quad = C(M - \frac{1}{2})\left(1 - x^2\right)^{-\frac{1}{2}}\left(1 - \frac{y^2}{1 - x^2}\right)^{M-2}. \end{aligned} \tag{54}$$

By multiplying (52) and (54), we obtain the joint distribution of $\Re\alpha_1$ and $\Im\alpha_1$

$$f_{\alpha_1}(a) = \frac{1}{\pi}\frac{\Gamma(M)}{\Gamma(M - 1)}(1 - ||a||^2)^{M-2}. \tag{55}$$

The conditional distribution of the rest $\Re\alpha_i$ and $\Im\alpha_i$ can be obtained inductively. By multiplying the conditional probability density functions according to the chain rule, we obtain (18) and end the proof.

## References

[1] Q. Li *et al.*, "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[2] R. van Nee, "Breaking the gigabit-per-second barrier with 802.11AC," *IEEE Wireless Commun.*, vol. 18, no. 2, p. 4, Apr. 2011.

[3] 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.

[4] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 2–14, Jan. 1972.

[5] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

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

[7] 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.

[8] R. de Francisco, M. Kountouris, D. T. M. Slock, and D. Gesbert, "Orthogonal linear beamforming in MIMO broadcast channels," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2007, pp. 1210–1215.

[9] L. N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Beamformer designs for MISO broadcast channels with zero-forcing dirty paper coding," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1173–1185, Mar. 2013.

[10] S. Wu, W. Mao, and X. Wang, "Performance study on a CSMA/CA-based MAC protocol for multi-user MIMO wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3153–3166, Jun. 2014.

[11] J. Duplicy, D. P. Palomar, and L. Vandendorpe, "Adaptive orthogonal beamforming for the MIMO broadcast channel," in *Proc. 2nd IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, Dec. 2007, pp. 77–80.

[12] J. Xie, J. Zhang, and L. Bai, "Multilayer orthogonal beamforming for priority-guaranteed wireless communications," *Int. J. Distrib. Sensor Netw.*, vol. 8, no. 8, p. 307467, 2012.

[13] Y. Li, L. Bai, C. Chen, Y. Jin, and J. Choi, "Successive orthogonal beamforming for cooperative multi-point downlinks," *IET Commun.*, vol. 7, no. 8, pp. 706–714, May 2013.

[14] S. Ozyurt and M. Torlak, "Unified performance analysis of orthogonal transmit beamforming methods with user selection," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1026–1037, Mar. 2013.

[15] V. Krishna, *Auction Theory*. New York, NY, USA: Academic, 2009.

[16] J. K. MacKie-Mason and H. R. Varian, "Pricing the Internet," in *Proc. Comput. Econ. EconWPA*, Jan. 1994, pp. 1–39. [Online]. Available: https://ideas.repec.org/p/wpa/wuwpco/9401002.html

[17] J. Shu and P. Varaiya, "Pricing network services," in *Proc. 22nd Annu. Joint Conf. IEEE Comput. Commun.*, vol. 2. Mar. 2003, pp. 1221–1230 vol.2.

[18] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

[19] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.

[20] J. B. Ernst, S. Kremer, and J. J. P. C. Rodrigues, "A utility based access point selection method for IEEE 802.11 wireless networks with enhanced quality of experience," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 2363–2368.

[21] G. Ye, H. Zhang, H. Liu, J. Cheng, and V. C. M. Leung, "Energy efficient joint user association and power allocation in a two-tier heterogeneous network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–5.

[22] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[23] A. C. Morales, A. Aijaz, and T. Mahmoodi, "Taming mobility management functions in 5G: Handover functionality as a service (FaaS)," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–4.

[24] H. Zhang, C. Jiang, and J. Cheng, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92–99, Jun. 2015.

[25] X. Duan and X. Wang, "Authentication handover and privacy protection in 5G HetNets using software-defined networking," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 28–35, Apr. 2015.

[26] M. Xie and T. M. Lok, "Access point selection and auction-based scheduling in uplink MU-MIMO WLANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[27] N. R. Goodman, "The distribution of the determinant of a complex Wishart distributed matrix," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 178–180, 1963.

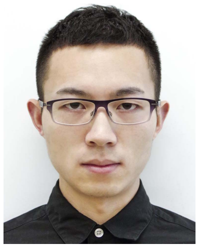
[28] J. G. Wendel, "The non-absolute convergence of Gil–Pelaez' inversion integral," *Ann. Math. Statist.*, vol. 32, no. 1, pp. 338–339, Mar. 1961. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177705164

[29] V. Witkovskỳ, "On the exact computation of the density and of the quantiles of linear combinations of t and F random variables," *J. Statist. Planning Inference*, vol. 94, no. 1, pp. 1–13, Mar. 2001.

**Mengjie Xie** (S'16) received the B.Eng. degree and the Ph.D. degree under the supervision of Prof. T.-M. Lok in information engineering from The Chinese University of Hong Kong, Sha Tin, Hong Kong, in 2012 and 2017, respectively. Her research interests include wireless communications, multiple-input multiple-output (MIMO) systems, multiuser MIMO systems, interference alignment, resource allocation and scheduling, and network economics.

**Tat-Ming Lok** (SM'03) received the B.Sc. degree in electronic engineering from The Chinese University of Hong Kong, Sha Tin, Hong Kong, in 1991, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1992 and 1995, respectively.

He was a Post-Doctoral Research Associate with Purdue University. He then joined The Chinese University of Hong Kong, where he is currently an Associate Professor. His research interests include communication theory, communication networks, signal processing for communications, and wireless systems. He was a Co-Chair of the Wireless Access Track of the IEEE Vehicular Technology Conference in 2004. He has served on Technical Program Committees of different international conferences, including the IEEE International Conference on Communications, the IEEE Vehicular Technology Conference, the IEEE Globecom, the IEEE Wireless Communications and Networking Conference, and the IEEE International Symposium on Information Theory. He also served as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2002 to 2008. Since 2015, he has been serving as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.



**Qing Yang** (M'18) received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2015. From 2015 to 2016, he was a Post-Doctoral Researcher with the Information Engineering Department, The Chinese University of Hong Kong. In 2016, he joined P2 Wireless Technologies Co., Ltd., as the Principal Engineer, where he was involved in the research and development of outdoor wireless mesh networks. In 2018, he joined the College of Information Engineering, Shenzhen University. His research interests include PHY/MAC cross-layer design and optimization, ultra-dense network, physical-layer network coding, visible light communication in IoT, and software-defined radio.