Scalable User Selection for MU-MIMO Networks

Xiufeng Xie and Xinyu Zhang
University of Wisconsin-Madison
Email: {xiufeng, xyzhang}@ece.wisc.edu

***Abstract*—In a multi-user MIMO (MU-MIMO) network, an AP with M antennas can only serve up to M users out of a large user population. The M users' rates are inter-coupled and depend on their channel orthogonality. Substantial theoretical studies focused on selecting users to maximize capacity, but they require feedback of channel state information (CSI) from all users. The resulting overhead can easily overwhelm useful data in large scale networks. In this paper, we propose a scalable user selection mechanism called *orthogonality probing based user selection* (OPUS). OPUS only requires up to M rounds of CSI feedback. In each round, it employs a novel probing mechanism that enables a user to evaluate its orthogonality with existing users, and a distributed contention mechanism that singles out the best user to feedback its CSI. Software-radio based implementation and experimentation shows that OPUS significantly outperforms traditional user selection schemes in both throughput and fairness.**

I. Introduction

Multi-user MIMO (MU-MIMO) downlink transmission has been enabled in state-of-the-art wireless network standards including 802.11ac WLAN, WiMax and LTE cellular networks. MU-MIMO holds the potential to substantially improve spectrum efficiency, by allowing concurrent transmissions from a multi-antenna access point (AP) to multiple users. In theory, downlink capacity grows linearly with the number of transmit or receive antennas, whichever is smaller [1]. In practice, an AP's number of transmit antennas is always limited, *e.g.*, up to 8 in existing standards, yet the number of users can grow to hundred-scale. Thus, the AP needs to select a limited number of users to serve in each MU-MIMO transmission. A user selection strategy must be judiciously devised, because the users are coupled and their achievable rates depend on the orthogonality of their instantaneous channel states.

Substantial theoretical research has focused on the user selection problem for MU-MIMO [2], with an objective of maximizing downlink capacity. This optimization problem can be solved by assuming that the AP knows instantaneous channel state information (CSI) of all users. In practice, CSI has to be obtained from users' feedback and entails formidable overhead, which grows linearly with the number of users and can overwhelm the actual channel time spent in data transmission [3]. Moreover, the feedback may span a longer duration than channel coherence time — some users' CSI may become outdated by the time the AP is ready to make the decision on user selection.

Therefore, a user selection mechanism must be *scalable* — given a growing user population, it should bound the CSI overhead while maximizing the downlink capacity, in order to optimize the overall *throughput*. This objective puts traditional user selection schemes [2] in a dilemma: for scalability, only the "best" user group should be served in each MU-MIMO transmission, but selecting the best group requires *all* users to feedback their CSI, which compromises scalability. CSI compression algorithms [3] alleviate the overhead, but do not stop its growth with user population. Random user selection obviates the need for full CSI feedback, yet it neglects the coupling among users and reduces downlink capacity [4].

In this paper, we propose *Orthogonality Probing based User Selection* (OPUS), a scalable user selection mechanism for MU-MIMO networks. OPUS is compatible with the 802.11ac MU-MIMO standard, which requires an AP to sequentially poll/receive CSI feedback. Standard 802.11ac assumes the set of users to be served is already given (*e.g.*, via random selection). OPUS replaces this assumption by integrating a light-weighted user selection procedure into each round of "probe-and-feedback". In each round, each unselected user estimates its *potential* to boost capacity when grouped with existing selected users. Then, the unselected users initiate a distributed *feedback contention*, where the one with the highest capacity potential wins and immediately sends its CSI back to the AP. The entire round of operations repeats until the number of selected users reaches the upper-bound, *i.e.*, the number of antennas on the AP. In this way, OPUS bounds the overhead even if the user population grows. Meanwhile, it optimizes downlink capacity by selecting users properly.

Since 802.11ac reserves the medium before a downlink transmission, and OPUS only takes effect in the reserved duration, it would not affect the performance of legacy 802.11ac devices nearby.

Practical implementation of OPUS entails unique challenges. First, how can a user estimate its contribution to downlink capacity when grouped with those already selected ones? A straightforward solution may run a "test transmission" that serves the selected users plus an unselected one, and then measures the resulting bit-rate. But the number of test transmissions grows with the number of users, incurring huge overhead. OPUS overcomes this problem using a novel *orthogonality probing* scheme. It reengineers the 802.11ac probing/polling frame, and embeds a training preamble simultaneously steered to all channel directions that are orthogonal to the selected users in the signal space. Upon receiving the probing frame, each user evaluates its *preference metric*, which reflects its channel quality and orthogonality to selected users, and can be used to infer its potential contribution to downlink capacity.

Thereafter, OPUS's AP needs to identify the user with the highest preference metric, but again at low overhead — without requiring all users to report their metrics one by one. OPUS's feedback contention mechanism meets this challenge

978-1-4799-3360-0/14/$31.00

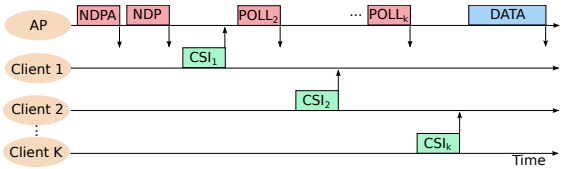

Fig. 1. CSI feedback mechanism in 802.11ac.

through a distributed protocol, whereby the best user is ensured to be singled out among all unselected ones. The protocol incurs a small, fixed overhead that does not grow with the user population. In addition, OPUS optimizes the number of concurrent users, in order to balance a traceoff between MIMO multiplexing gain and diversity gain. Instead of maximizing the number of concurrently served users (default operation in 802.11ac [5]), it can intelligently terminate the user selection when adding more users compromises channel orthogonality.

We have built a prototype of OPUS on the WARP [6] software radio platform. OPUS's components are implemented on top of a MU-MIMO modulation/decoding module that we built following the 802.11ac PHY. Extensive experiments show that the overhead of existing user selection schemes [2], [7] can easily nullify their capacity gain, even in a medium sized network with 8 to 20 users. In contrast, OPUS enables efficient user selection, leading to low overhead and high throughput even if the network size grows to hundred-scale — arguably the maximum size of a WLAN cell. In a typical 802.11ac MU-MIMO WLAN with a 2-antenna AP and 20 users, OPUS achieves a throughput gain of 1.4× to 5.1× compared with three state-of-the-art user selection schemes.

To our knowledge, this work represents the first *experimental study* of state-of-the-art user selection schemes [2], [7] in MU-MIMO networks, and the first user selection scheme that achieves *scalability* in medium to large sized networks.

The rest of this paper is organized as follows. Sec. II presents background for CSI feedback and user selection in MU-MIMO. In Sec. III, we introduce the motivation behind a scalable user selection scheme. We describe the design choices and components of OPUS in Sec. IV and validate its performance in Sec. V. Finally, Sec. VII concludes the paper.

II. Background

In this section, we briefly review the CSI feedback and user selection problems in MU-MIMO. We focus on a multi-antenna AP and single-antenna users — the default configuration for 802.11ac MU-MIMO [5]. The results can be easily extended to multi-antenna users following [8].

A. CSI feedback mechanism

Before a MU-MIMO transmission, the AP needs to know the CSI (*i.e.*, the *channel matrix* from its transmit antennas to intended users) based on users' feedback. Fig. 1 illustrates 802.11ac's MAC operations related with CSI feedback.

First, the AP sends a null data packet announcement (NDPA) frame to notify intended users for beamforming and reserve channel from neighboring WLAN cells. Immediately afterwards, it sends a null data packet (NDP) which contains a training preamble. Upon receiving the NDP, each intended user k estimates its CSI, *i.e.*, channel gain vector between transmit antennas and itself. The first user sends its CSI back to the AP immediately. Other intended users each provides feedback only upon receiving a probing/polling packet from the AP. The users' ordering is conveyed in the NDPA.

As the number of users increases, CSI feedback costs more time, but is always sent at the lowest modulation rate for reliability. Unfortunately, the data packet's duration does not increase — it can even decrease with data rate. Thus the overhead becomes overwhelming in high-rate 802.11ac networks.

To alleviate the feedback overhead, 802.11ac quantizes the CSI numerical values into 4 to 8 bits fixed-point numbers, and allows up to 4 adjacent OFDM frequency bins to share CSI. Yet after such compression, the per-user overhead still ranges from 100 to 800 bytes, and grows linearly with the number of intended users [3]. Alternatively, CSI report can be sent less frequently. But to ensure accuracy, the feedback period must be much shorter than the channel coherence time. In indoor environment with static or walking users, feedback period needs to be shorter than 15ms to maintain CSI accuracy [9].

Besides, frame aggregation can reduce the relative overhead of CSI. In large-scale wireless networks, contention causes long inter-packet service delay, leaving more opportunities for accumulating and aggregating frames from the upper layers. 802.11ac allows up to 5.5ms of aggregated frame duration [5].

B. Incremental User Selection

For an AP with M antennas, MU-MIMO requires selecting a subset of up to M users to serve based on the CSI of all K users in a network. This problem involves MAC-layer fair scheduling as well as PHY-layer capacity maximization. The latter is particularly important and unique to MU-MIMO, because the sum-rate of a downlink transmission is determined by channel orthogonality among served users [1]. But even with full CSI knowledge at the AP, it is still computationally prohibitive to find the optimal user set that maximizes the sum-rate, especially when total user number K is large [8].

Suboptimal algorithms that perform incremental user selection (*e.g.*, SUS [2]) have been shown to well approximate the optimal capacity at low computational complexity [8]. The AP can choose the first user with the highest channel quality. Then, it selects the next user that provides the best potential performance when grouped with those selected ones. The procedure repeats until M users are selected. However, in each iteration, identifying the "next best" user still requires full CSI from all unselected users.

III. Motivation and Challenges

In this section we establish a theoretical model to understand how user selection affects MU-MIMO performance and why existing schemes are insufficient.

A. Impact of user selection in practical MU-MIMO networks

Practical MU-MIMO implementation in 802.11ac need to satisfy two constraints: per-antenna power budget and low computational complexity. The first constraint roots in the hardware

structure of 802.11ac transceivers, which accompanies each antenna with a separate RF front-end. Existing analysis of user selection mostly focused on cellular networks with total-power constraint [2]. Few works considered optimizing MU-MIMO under per-antenna power constraint (*e.g.*, [10]), but the solutions involve sophisticated non-linear optimization. In what follows, we analyze the impact of user selection under the two practical constraints. The analysis will provide guidelines for designing the orthogonal probing mechanism in OPUS.

1) MU-MIMO system model: Owing to its low complexity, Zero-forcing beamforming (ZFBF) is widely adopted for implementing MU-MIMO downlink transmissions. ZFBF allows an AP to *precode* data symbols and steer them towards desired users, while precanceling the mutual interference. Consider an AP with M transmit antennas and a set of selected users $\mathcal{S}$. Let d_k be the data symbol intended for user k; $\boldsymbol{h}_k$ the $1 \times M$ *channel state vector* between transmit antennas and user k. With precoding, each transmit antenna emits a weighted combination of the users' data. For user k, the weights on different antennas form a $M \times 1$ *precoding weight vector* $\boldsymbol{w}_k$. Let n_k denote the noise level of user k, its received signal after AP's precoding and channel distortion becomes:

$$y_k = \boldsymbol{h}_k \boldsymbol{w}_k d_k + \textstyle\sum_{j \in \mathcal{S}, j \neq k} \boldsymbol{h}_k \boldsymbol{w}_j d_j + n_k,\ k \in \mathcal{S} \quad (1)$$

To realize MU-MIMO, ZFBF enforces the following constraint: $\boldsymbol{h}_k \boldsymbol{w}_j = 0, \forall j \in \mathcal{S}, j \neq k$. Consequently, user k only receives its desired symbol d_k, whereas other users' symbols are cancelled owing to the composite effects of precoding and channel distortion.

Suppose $|\mathcal{S}| = K$. Given a *channel state matrix* $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_K]^T$ from AP's transmit antennas to all users, the *beamforming weight matrix* $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M]$ that satisfies the ZFBF constraints is usually computed from the pseudo inverse: $\boldsymbol{W} = \boldsymbol{H}^\dagger = \boldsymbol{H}^*(\boldsymbol{H}\boldsymbol{H}^*)^{-1}$. Note that pseudo inverse implicitly enforces a constraint $\boldsymbol{h}_k \boldsymbol{w}_k = 1$.

2) Impact of user selection under per-antenna power constraint: Since data symbols have unit power, the per-antenna power constraint implies $\sum_{k=1}^{K} |W_{mk}| \leq \sqrt{P}$ for each antenna m with power budget P. To satisfy both this constraint and the ZFBF constraints, we must divide the precoding vector of all users by the same factor $f_s = \max_m \sum_{k=1}^{K} |w_{mk}| = ||\boldsymbol{W}||_\infty$.

Then the received signal at user k becomes:

$$y_k = \boldsymbol{h}_k (\frac{\sqrt{P}\boldsymbol{w}_k}{\|\boldsymbol{W}\|_\infty} d_k) + n_k = \frac{\sqrt{P}}{\|\boldsymbol{W}\|_\infty} d_k + n_k$$

Also note that $\boldsymbol{W} = \boldsymbol{H}^\dagger$, we can obtain the SINR_k at receiver k as follows.

$$\text{SINR}_k = \frac{P}{\sigma^2 \|\boldsymbol{W}\|_\infty^2} = \frac{P}{\sigma^2} (\frac{\|\boldsymbol{H}\|_\infty}{\mathcal{C}_\infty(\boldsymbol{H})})^2 \quad (2)$$

where $\mathcal{C}_\infty(\boldsymbol{H})$ denotes the infinity norm condition number [11] that reflects channel orthogonality among users, and is related with $\boldsymbol{H}$ by: $\mathcal{C}_\infty(\boldsymbol{H}) = \|\boldsymbol{H}\|_\infty \|\boldsymbol{H}^\dagger\|_\infty = \|\boldsymbol{H}\|_\infty \|\boldsymbol{W}\|_\infty$. It is known that $\mathcal{C}_\infty(\boldsymbol{H})$ approaches infinity when any two of the users' channels are linearly correlated and approaches 1 when the users are fully orthogonal [11].

Based on the analysis above, we can conclude that *under per-antenna power constraint, the MU-MIMO SINR is determined by both the condition number $\mathcal{C}_\infty(\boldsymbol{H})$ which reflects selected user's mutual channel orthogonality, and the infinity norm $\|\boldsymbol{H}\|_\infty = \max_{k \in \mathcal{S}} \sum_{m=1}^{M} |h_{km}|$ which reflects channel quality.* It is critical to select a proper subset of users with low correlation (small $\mathcal{C}_\infty(\boldsymbol{H})$) and high channel quality (large $\|\boldsymbol{H}\|_\infty$), in order to maximize the sum rate of all users.

Note that it is the infinity norm condition number $\mathcal{C}_\infty(\boldsymbol{H})$ of the channel matrix $\boldsymbol{H}$ that affects the performance of MU-MIMO. This fundamentally differs from single-link MIMO, whose capacity is known to be determined by the channel matrix's 2-norm condition number $\mathcal{C}_2(\boldsymbol{H})$ [1].

Also note that the work in [12] argues that selecting different user subsets can cause at most 3-4 dB SINR difference, so user selection would not affect the throughput performance much. However, the conclusion is based on a total power constraint. In Sec.V-B, we will show through experiments that under a per-antenna power constraint, improper user grouping can significantly affect the users' SINR.

B. Challenges in User Selection

An optimal user selection scheme should evaluate each subset of users with size M, which entails an exhaustive search over all possible $\binom{K}{M}$ combinations. Incremental user selection can reduce the computational complexity to $M \times K$ [8], but it requires full CSI from all users. As mentioned above, per-user CSI feedback incurs 100 to 800 bytes overhead sent at the lowest modulation rate of 6Mbps. Suppose there are 20 users, then the channel time cost can be up to $800 \times 20/6 = 2666\mu s$, equivalent to multiple data packets' durations. Consider the requirement of $15ms$ feedback period. Due to channel contention latency, only one or two transmissions may be delivered within this period. In other words, even with 20 users, the channel time cost of existing schemes can be comparable to or even exceed that of actual data transmission.

In addition, existing user selection algorithms evaluate the potential capacity of each user based on theoretical channel models [2], [7], [8]. It remains an open problem how this can be realized in practical MU-MIMO protocols. Our OPUS mechanism is designed to overcome such limitations.

IV. OPUS Design

A. Design Overview

OPUS inherits the low-complexity of incremental user selection, but it dramatically reduces the feedback overhead, thus achieving scalability. Instead of requiring the AP to obtain all users' CSI, OPUS runs distributed user selection, whereby each user evaluates its potential contribution to downlink capacity when grouped with those already selected ones. Such potential is characterized using a *preference metric*.

Fig. 2 illustrates a typical flow of operations in OPUS. OPUS preserves the basic operations in 802.11ac, except that it adds a fixed-duration contention period before CSI feedback, and reengineers the polling packet to facilitate its orthogonal probing mechanism. At a high level, OPUS works as follows:

(i) First, the AP announces its intention for MU-MIMO downlink transmission through the NDPA and NDP packets.

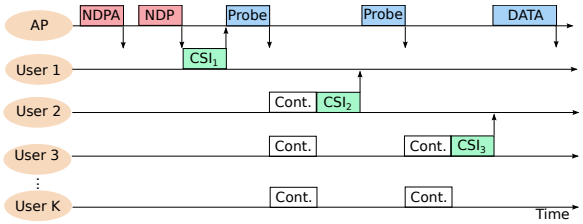

Fig. 2. Orthogonality probing based user selection.

The first user ("core" user) is selected by the AP and announced in the NDPA based on a throughput fairness criteria (Sec. IV-E).

(ii) Each user k estimates its CSI, *i.e.*, *channel state vector* $\boldsymbol{h}_k$ independently based on the NDP. The core user returns its CSI report to the AP, completing the first round of user selection.

(iii) In subsequent rounds, given CSI from already selected users, the AP computes possible directions to probe in the signal space, which are orthogonal to the selected users' channels. It then beamforms a probing frame towards these directions, with different training data sequences for each direction.

(iv) After receiving the probing frame, each unselected user evaluates its maximum SINR along all training directions, which is later used as its preference metric.

(v) All unselected users join a *feedback contention*. The one with maximum preference metric wins and sends its CSI to the AP who adds it to the set of selected users.

(vi) Repeat steps (iii)-(v) (a *user-selection round*) until the number of selected users reaches M. The AP may terminate the procedure early if adding any unselected user hurts the MU-MIMO performance. Afterwards, based on the collected CSI, the AP runs ZFBF and delivers data frames to all selected users. Note that multiple data frames can follow one user-selection procedure, provided that they span much shorter duration than the coherence time, as we mentioned in Sec. II.

We proceed to describe all components of OPUS in detail.

B. Orthogonality Probing Mechanism

OPUS's orthogonality probing mechanism allows individual users to evaluate their potential contribution to downlink capacity based on a probing frame from the AP.

1) Basic operations: We describe how orthogonality probing works using a simple example in Fig. 3(a). In the beginning of user-selection round n, the AP already has CSI from $(n-1)$ users selected in previous rounds. It then constructs a probing frame pointing to $M-(n-1)$ directions that are orthogonal to the selected users' channels. Each unselected user k receives the probing frame, evaluates its overhearing SINR for each probing direction, and subsequently computes its preference metric g_k. Simply put, g_k is associated with the direction that user k is best aligned to, thus reflecting channel quality and orthogonality with selected users. Fig. 3(b) illustrates user channels and probing directions of this user-selection round in vector space.

2) Designing the probing frame: The probing frame must "point" to directions that are orthogonal to selected users in the signal space. OPUS meets this goal by redesigning a *VHT-LTF preamble* in the probing frame. In legacy 802.11ac, the VHT-LTF carries a ZFBF-precoded known data sequence to facilitate

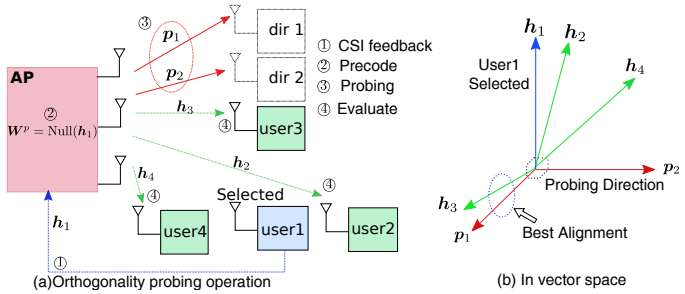

Fig. 3. An example of orthogonality probing.

packet decoding. In OPUS, we define M known training sequences. In user-selection round n, the AP precodes the first $M-(n-1)$ known training sequences and steers them to desired $M-(n-1)$ probing directions. The precoded sequences are embedded into the VHT-LTF and emitted simultaneously through M antennas. So, how to design the precoding weights? We answer this question in the following claim.

Claim 1 *For user-selection round n, suppose $\boldsymbol{H}_{n-1}$ is the channel matrix formed by the $(n-1)$ users selected in previous $(n-1)$ rounds. The set of directions to probe form a basis of the null space of $\boldsymbol{H}_{n-1}$, which can be probed all at once by precoding $M-(n-1)$ known data sequences, and sending the precoded sequences simultaneously through 802.11ac's VHT-LTF preamble. The precoding matrix should be a $M \times (M-(n-1))$ matrix: $\boldsymbol{W}^p = Null(\boldsymbol{H}_{n-1})$.*

Proof: Given $\boldsymbol{H}_{n-1}$, the AP needs to compute all $M-(n-1)$ probing directions that are orthogonal to all the selected users' channel directions. Let $\boldsymbol{p}_i$ be a $1\times M$ unit vector denoting the i-th probing direction. Then the AP beamforms the probing frame to all probing directions as if it is serving a group of "fake users" with the $(M-(n-1))\times M$ *fake channel matrix* $\boldsymbol{P} = [\boldsymbol{p}_1^T, \boldsymbol{p}_2^T, ..., \boldsymbol{p}_{M-(n-1)}^T]^T$.

Let $\mathcal{S}_{n-1}$ be the set of selected users. The orthogonality requirement for $\boldsymbol{p}_i$ entails that: $\boldsymbol{h}_k \cdot \boldsymbol{p}_i = 0,\ k \in \mathcal{S}_{n-1}$; or $\boldsymbol{H}_{n-1}\boldsymbol{p}_i^* = 0$ in matrix form. The set of all $\boldsymbol{p}_i^*$ satisfying this equation forms a basis of the null space of $\boldsymbol{H}_{n-1}$, *i.e.*,

$$\boldsymbol{P}^* = [\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, ..., \boldsymbol{p}_{M-S_{n-1}}^*] = \text{Null}(\boldsymbol{H}_{n-1}) \tag{3}$$

To probe the $M-(n-1)$ directions in the fake channel matrix $\boldsymbol{P}$, similar to ZFBF, the precoding matrix can be designed by pseudo-inverse: $\boldsymbol{W}^p = \boldsymbol{P}^\dagger$. Note that the set of all probing directions are orthonormal, thus the probing direction matrix $\boldsymbol{P}^*$ is a semi-orthogonal matrix, with an intrinsic property: $\boldsymbol{P}^* = \boldsymbol{P}^\dagger$. It then follows that $\boldsymbol{W}^p = \boldsymbol{P}^*$, which can also be written for each beamforming direction as: $\boldsymbol{w}_i^p = \boldsymbol{p}_i^* = \text{Null}(\boldsymbol{H}_{n-1})$, thus completing the proof. □

3) Designing the preference metric: Upon receiving the probing frame, each unselected user needs to compute its *preference metric*, which reflects its potential contribution to downlink capacity when grouped with selected users. More specifically, the preference metric needs to reflect both its channel quality and mutual orthogonality with selected users' channels. OPUS meets this goal as follows.

Claim 2 *For user-selection round n, the preference metric of user k can be defined based on its overhearing SINR$_k^i$ for each probing direction $\boldsymbol{p}_i$ as:* $g_k = \max_{\boldsymbol{p}_i \in \boldsymbol{P}}(SINR_k^i)$.

Proof: Let σ_k^2 denote noise floor of user k, then SINR_k^i can be modeled as follows.

$$\text{SINR}_k^i = \frac{|\boldsymbol{h}_k \cdot \boldsymbol{p}_i|^2}{\sum_{\boldsymbol{p}_j \in \boldsymbol{P}, j \neq i} |\boldsymbol{h}_k \cdot \boldsymbol{p}_j|^2 + \sigma_k^2} \tag{4}$$

The correlation between two channel vectors can be evaluated using the Hermitian angle [2], [13]. Let $0 \leq \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_i) \leq \frac{\pi}{2}$ denote the Hermitian angle between channel vector $\boldsymbol{h}_k$ and probing direction $\boldsymbol{p}_i$, we have

$$|\boldsymbol{h}_k \cdot \boldsymbol{p}_i| = \|\boldsymbol{h}_k\| \|\boldsymbol{p}_i\| \cos \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_i) \tag{5}$$

Note that $\boldsymbol{p}_i$ is unit vector, then Equ. (4) can be written as

$$\text{SINR}_k^i = \frac{(\cos \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_i))^2}{\sum_{\boldsymbol{p}_j \in \boldsymbol{P}, j \neq i} (\cos \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_j))^2 + \frac{\sigma_k^2}{\|\boldsymbol{h}_k\|^2}} \tag{6}$$

From Equ. (6), we can observe that *the SINR$_k^i$ of user k reflects both its channel direction and channel quality.* High SINR_k^i requires $\cos \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_i)$ to be large and all $\cos \Theta_H(\boldsymbol{h}_k, \boldsymbol{p}_j)$ small, which implies that channel of user k has good alignment with one of the probing directions, *i.e.*, good orthogonality with the channels of all selected users. Meanwhile, high SINR_k^i also requires a large channel magnitude $\|\boldsymbol{h}_k\|^2$, which reflects high channel quality.

In selection round n, there are $M-(n-1)$ probing directions that are "equally" orthogonal with all selected users' channel directions. But for an unselected user k, its channel quality may differ along these directions. Thus its preference metric should be defined according to the direction with highest overhearing SINR as: $g_k = \max_{\boldsymbol{p}_i \in \boldsymbol{P}}(\text{SINR}_k^i)$. □

The analysis above is only used to justify our design of preference metric. The actual OPUS implementation adopts a practical way to evaluate SINR_k^i: First we view it as the SINR of an *equivalent channel*, whose input is the *training sequence* (Sec. IV-B2) for probing direction $\boldsymbol{p}_i$ and output is the data symbols user k receives from the probing frame. Then, with both input and output known, user k estimates the channel gain and noise floor for each equivalent channel separately, based on which it computes the SINR_k^i (similarly to [3]). Note that a user does not have to differentiate different directions. It only needs to evaluate the maximum SINR among them.

C. Distributed Feedback Contention

For scalability, OPUS's feedback contention mechanism needs to single out the user with the highest preference metric in a distributed manner.

1) Overview: The feedback contention mechanism works as follows (Fig. 4). After receiving the probing frame, each unselected user k computes and quantizes its preference metric g_k into N bits. These bits are mapped to a N-stage contention procedure. In each stage, users with "1" on the corresponding bit send a short energy burst, whereas those with "0" listen to the channel. If a listening user senses a busy channel in any stage, then it infers that someone else has a larger preference metric, hence it quits the contention immediately and will not join later stages. Meanwhile, since 802.11 radios are half-duplex, users with bit "1" cannot listen to the channel and will survive this stage. Finally, only the one who survives all N stages wins the contention. In this way, the user with the highest preference metric will finally feedback its CSI.

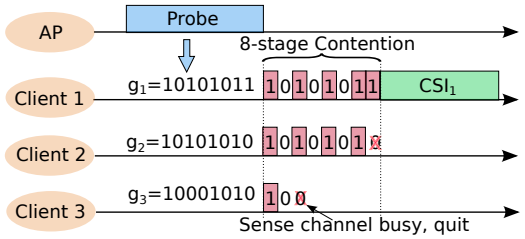

Fig. 4. An example for the CSI feedback contention mechanism.

2) Energy burst design and detection: OPUS's contention mechanism is designed for compatibility with 802.11 hardware. Its energy burst is produced by sending an 802.11 null data packet (NDP) which only contains a preamble. Since the minimum preamble length in 802.11 is $20\mu s$ and the switching time between transmitting and receiving mode for the RF front-end is typically smaller than $5\mu s$ [14], we design the duration of each contention stage as $27\mu s$ (3 time slots in 802.11).

During the contention a user with bit "0" only performs *energy detection* instead of decoding. This design has two advantages. (1) The energy burst can be detected from a long distance, thus all users can participate in the contention. (2) When more than one users send energy bursts in a contention stage, contention still works because the listening users only need to know if there is *at least one* user sending energy bursts.

3) Synchronization: Users' contention stages should be aligned with each other in time. OPUS leverage the probing frame as a reference broadcast to synchronize users at the beginning of the contention. The feasibility and effectiveness of such reference broadcast has been validated in OFDMA based wireless LANs [14]. OPUS's synchronization requirement is even lower than OFDMA because it only relies on detecting energy bursts that span multiple slots. It will not be affected by small jitters in synchronization, *e.g.*, those caused by propagation delay, which is typically below $800ns$ [15] and much shorter than the $9\mu s$ time slot in 802.11.

4) Contention overhead: A notable feature of the feedback contention is that its channel time cost remains the same irrespective of the user population. With N contention stages, the total contention overhead is fixed to $3N$ time slots.

When users' channels are correlated and have similar quality, their preference metric may be similar as well. If the number of contention stages (*i.e.*, quantization bits for the preference metric) N is not large enough, there can be more than one winners sharing the highest preference metric, incurring collision. OPUS has several intrinsic measures to reduce such risks.

First, collision probability of such binary-countdown like mechanisms decreases exponentially with N [16]. In experiments, we find that $N = 8$ ($2^8 = 256$ quantization levels) is sufficient to keep collision to a minimum (Sec. V-C). Second, OPUS enforces a lower limit to the preference metric, corresponding to the minimum SINR required to support the lowest

modulation rate, which can be obtained from a vendor-specific look-up table (we use the one from Cisco [17]). Users with preference metric below the limit will not participate in the feedback contention. Similarly, OPUS enforces an upper limit to the preference metric, corresponding to the maximum SINR that AP can observe in its WLAN. Together with the lower limit, this bounds the range of preference metric to quantize, resulting in a high resolution after quantization. We choose $[7, 30]$ dB as the default range, corresponding to a resolution of $23/256 = 0.09$dB when $N = 8$. Thus, users can seldom share the same quantized preference metric value.

In the rare case when collision occurs, OPUS's AP stops the user selection and directly starts MU-MIMO beamforming to the selected users, so as not to waste the channel time.

D. User Number Selection (UNS): Early Termination

Instead of forcing M concurrent transmissions (*i.e.*, greedily exploiting multiplexing gain), OPUS can intelligently terminate the user selection, if a higher downlink capacity can be achieved by selecting a smaller user set (*i.e.*, exploiting diversity gain).

The termination decision is made in two ways. First, after n-th round ($n < M$) of CSI feedback, the AP estimates the total downlink rate, by computing the per-user SINR as in Equ. (2) and mapping it to bit-rate following a look-up table [17]. If the sum rate is lower than the $(n-1)$-th round, the AP discards the n-th user, and beamforms to the first $(n-1)$ users instead.

Second, in the n-th user-selection round, if a user's preference metric is lower than the minimum SINR required to support its lowest modulation rate, then it gives up the feedback contention. If the AP hears no energy burst during the contention stages, it infers all unselected users have low rate, and starts beamforming to the $(n-1)$ users already selected.

E. Proportional Throughput Fairness

OPUS employs a simple randomized algorithm to arbitrate proportional throughput fairness among users. To initiate user selection, the AP appoints the first ("core") user with certain probability weighted by users' historical throughput record, which the AP can learn from ACKs. For proportional fairness, users with lower throughput should have a higher probability to be selected as the core user. To approach this principle, the AP keeps a moving average of all users' throughput. Let $\overline{R_k}$ denote the average throughput of user k, the probability that user k is selected as the "core" user is $p_k = \frac{1/\overline{R_k}}{\sum_{j\in\mathcal{K}} 1/\overline{R_j}}$, where $\mathcal{K}$ is the set of all users. Alternative fairness measures can be implemented by weighting the users in different ways.

F. Summary of Protocol Properties: Why is OPUS Scalable?

(i) Bounded CSI feedback overhead: Unlike existing schemes (*e.g.*, [2]) that need CSI from all K users as input, OPUS only requires CSI from at most M selected users. K can grow to hundred-scale, whereas M equals the number of transmit antennas on the AP, which is bounded (*e.g.*, to 8 in 802.11ac).

(ii) Fixed feedback contention overhead: The overhead of OPUS's feedback contention mechanism is only determined by the number of stages, which is fixed (to 8 by our default setting) regardless of the total user number K.

(iii) Low feedback collision probability: As we have justified, even with a fixed number of stages, the collision probability during feedback contention can be kept to a minimum.

(iv) Capacity maximization: While fixing the user selection overhead, OPUS leverages its orthogonality probing mechanism to ensure the best set of users are grouped, so that the MU-MIMO capacity gain is fully exploited and can scale with M.

(v) Limitation & Possible Improvements: OPUS's feedback contention protocol balances performance and compatibility with existing 802.11 devices. However, it assumes all users can overhear each other and may cause collision with hidden terminals. To overcome this limitation and further reduce contention overhead at the cost of increasing hardware complexity, frequency-domain contention mechanisms [14] can be integrated. This is left as our future work.

V. Performance Evaluation

In this section, we validate OPUS's performance through testbed experiments, aiming to answer the following questions:

- Is user selection necessary in real MU-MIMO networks?
- How much throughput gain can OPUS achieve in comparison with existing schemes?
- How much overhead does OPUS incur?
- Could OPUS work in mobile channel conditions?
- Is OPUS scalable?

A. Implementation and Experimental Setup

We have prototyped OPUS on top of an 802.11ac-compatible MU-MIMO OFDM library that we built on WARP. The library implements OFDM modulation, packet detection/synchronization, channel estimation and symbol demodulation, ZFBF-based MU-MIMO precoding, along with the MAC-layer probe-and-CSI-feedback (see [3] for details). To realize OPUS, we modified the 802.11ac polling frame into the probing frame that enables orthogonality probing (Sec. IV-B). We further implement the computation and quantization of the preference metric, and the user number selection mechanism. Each user's preference metric is derived from its measured average SINR among all its OFDM subcarriers. Due to the interface latency of WARP, we cannot directly implement a real-time version of the feedback contention. However, since all WARP radios in our testbed are connected to a PC controller, we emulate the contention on the PC, which then commands the winning contender to start the CSI feedback.

For performance comparison, we have also implemented three state-of-the-art user selection schemes: (i) Semi-orthogonal User Selection (SUS) [2]. SUS runs incremental user selection, but requires full CSI and evaluates users in a different way than OPUS. It sets a threshold for orthogonality and then selects qualified users with the highest channel quality. (ii) Random User Selection (RUS), essentially the 802.11ac default, which selects M users randomly with equal probability, and only requires M CSI feedbacks. (iii) Random Beamforming (RBF) [7], which randomly beamforms to multiple directions,

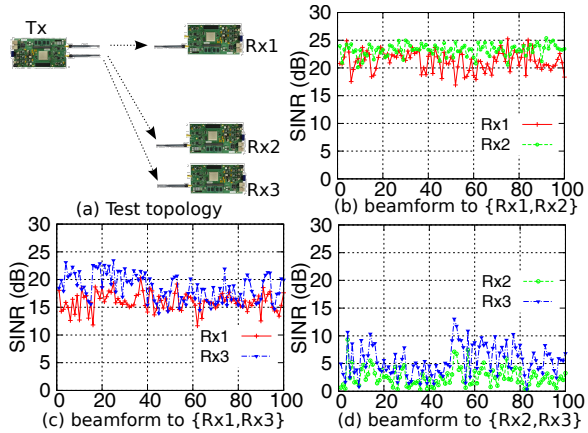

Fig. 5. Receiving SINR for all possible user combinations in a 3-user network (x-axis is the index of frame transmissions over time).

requires users to feedback their alignment with each direction, and then beamform to users with best alignment.

Our experiments are conducted on a testbed comprised of 5 WARP nodes, located in an office environment. All MU-MIMO transmissions run on a 2.4GHz channel unused and non-overlapping with ambient wireless devices. Other PHY parameters follow the 802.11ac default (*e.g.*, 20MHz bandwidth and 64 subcarriers). CSI values are compressed to 4 bits for both real and imaginary components. Packet size is 1.5KB unless noted otherwise.

B. Why Is User Selection Necessary?

We first verify the impact of user selection in a benchmark topology (Fig. 5(a)) with a 2-antenna AP and 3 users. Fig. 5(b), (c) and (d) plot the MU-MIMO performance of all possible user combinations. We can observe that the combination {Rx2,Rx3} results in around 15dB lower SINR compared to others. Obviously, user selection can significantly affect network capacity.

To reveal the deeper reasons behind, we examine the users' channel characteristics in Fig. 6. From Equ. (2), we know the receiving SINR is determined by the users' channel orthogonality (reflected by infinity norm condition number $C_\infty(\boldsymbol{H})$) and channel quality (reflected by $\|\boldsymbol{H}\|_\infty$). Fig. 6(a) shows that when $\boldsymbol{H} = [\boldsymbol{h}_2^T, \boldsymbol{h}_3^T]^T$, $C_\infty(\boldsymbol{H})$ is always larger than all other possible combinations. Meanwhile, Fig. 6(b) shows that both Rx2 and Rx3 have lower channel magnitude than Rx1, which results in low $\|\boldsymbol{H}\|_\infty$. This explains why {Rx2,Rx3} has the worst performance and validates our analysis in Sec. III.

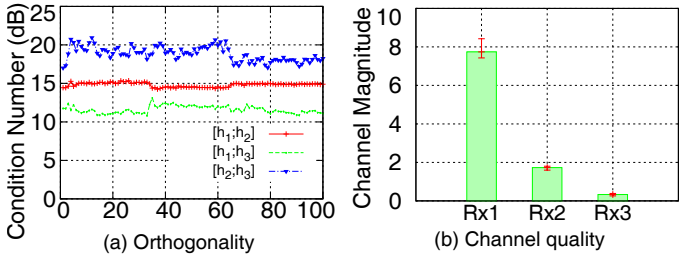

Fig. 6. Channel orthogonality and channel quality in the 3-user network.

C. Micro-benchmark Performance

1) Capacity gain from user selection: To validate that OPUS's orthogonality probing mechanism can identify the user combination with high performance, we compare its PHY capacity (no MAC overhead) with the benchmark schemes in the same topology as Fig. 5(a). From the results in Fig. 7(a), we see that OPUS outperforms all other user selection schemes. It achieves 49.2% capacity gain over RUS, which selects users randomly and may occasionally beamform to the suboptimal user combination {Rx2,Rx3}. SUS also results in lower capacity than OPUS, since it evaluates channel orthogonality/quality in a suboptimal way. RBF can only beamform to randomly generated directions. Since users can be very far away from its randomly pointed beams, RBF leads to poor PHY capacity. As shown in Fig. 7(b), OPUS outperforms SUS and RBF for the PHY capacity of all users, but RUS has the best fairness among them. Note that the fairness control component of OPUS is not activated here, its effectiveness will be discussed later.

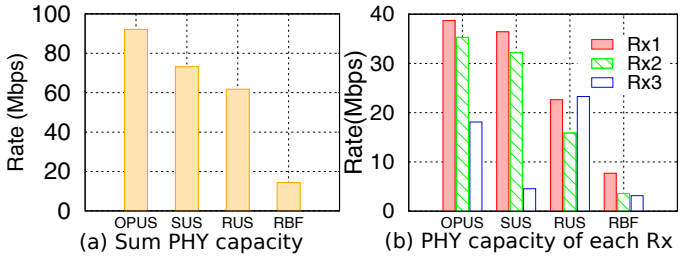

Fig. 7. Effectiveness of orthogonality probing during user selection.

2) MAC-layer overhead of OPUS: In this micro-benchmark, we compare the overhead of different user selection schemes. To isolate PHY effects, the results are obtained from the emulated MAC layer. In the emulation, we assume that all users can overhear each other in the feedback contention, *i.e.*, no hidden terminals. From Fig. 8(a), we can observe that: (i) The extra overhead induced by OPUS (mainly from feedback contention, cf. Fig. 2 and Fig. 1) is only a small increment to the total CSI feedback overhead in 802.11ac (reflected in RUS). Note that the CSI feedback from M selected users is a must for MU-MIMO, no matter whether user selection is used. (ii) OPUS's overhead increases negligibly with user population K, and thus it can be scalable. We have also observed that OPUS's feedback contention causes negligible collision — only around 1.1% with 20 users and 5.3% for up to 100 users. We omit the detailed plots due to space constraint. (iii) RBF does not require CSI feedback from users, thus has the smallest overhead, but this comes at the cost of poor PHY capacity.

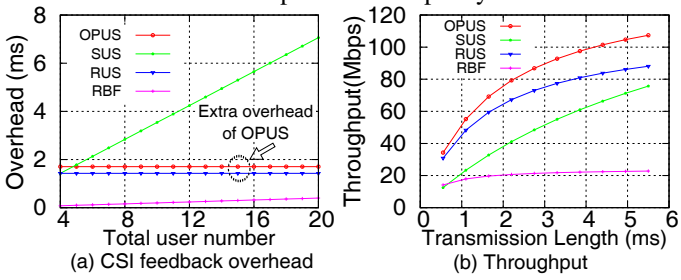

Fig. 8. Overhead analysis and the effect of frame aggregation on throughput.

We now evaluate impact of frame aggregation, which is used in 802.11ac to amortize the CSI overhead. From Fig. 8(b), we can see the net throughput of all user selection schemes increases with frame duration (up to the 5.5ms limit in 802.11ac),

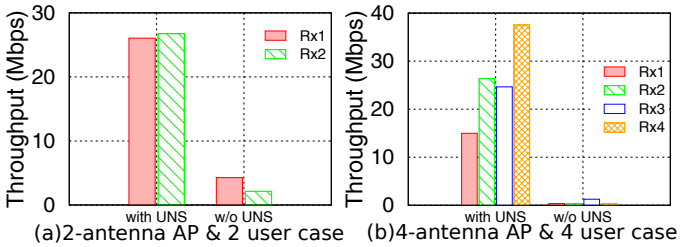

Fig. 9. Throughput with or without user number selection.

and OPUS maintains the highest throughput in all cases.

3) Effect of user number selection (UNS): Equ. (2) indicated that if a user with bad orthogonality to selected users is included to make up a user combination of size M, it will result in a large condition number C_∞ and ruin the performance of all selected users. OPUS's UNS mechanism avoids such pathological cases. Fig. 9(a) shows results from a 2-antenna AP serving two users with strongly correlated channels. MU-MIMO beamforming to both users leads to poor performance. With UNS, the AP judiciously beamforms to a single user, leading to more than $4\times$ throughput gain. Similar observations can be made in the experiments in Fig. 9(b), which contains a 4-antenna AP and 4 users, with Rx1 and Rx2 having strongly correlated channels.

4) Throughput fairness: In this micro-benchmark, we evaluate the fairness control component of OPUS. For each user number, we test it under 10 different topologies and evaluate the Jain's index based on the average throughput. Due to limited hardware in our testbed, trace-driven emulation based on real channel traces from WARP is used when user number is larger than 4. Accuracy of such emulation has been validated in our recent work [3]. The results in Fig. 10 show that OPUS's fairness control mechanism effectively maintains Jain's Index to close to 1. Meanwhile, the mechanism causes almost no throughput loss compared to OPUS without fairness control (*i.e.*, randomly selecting the "core" user). This is because the fairness control does not incur MAC-layer overhead.

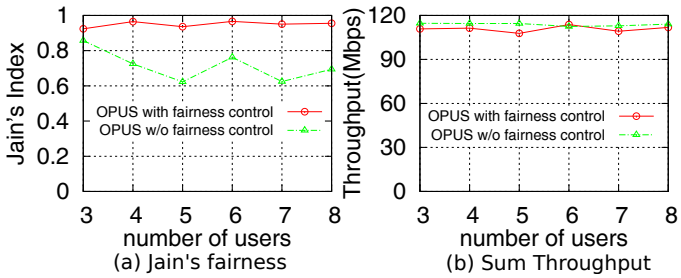

Fig. 10. Effectiveness of fairness control and its impact on throughput.

5) MU-MIMO user selection with mobility: Mobility affects the channel coherence time, which in turn dictates the CSI feedback period and overhead. We test such effects in the 3-node benchmark topology (Fig. 5(a)), but with Rx3 moving at walking speed. The results in Fig. 11(a) show that the PHY capacity of OPUS drops with increasing user selection and CSI feedback interval (in terms of the number of packet transmissions L within the feedback period). This roots in the staleness of CSI over time, and affects all MU-MIMO transmission schemes. The MAC-layer throughput first goes up because the average per-frame overhead decreases with increasing L. However, it starts to drop when $L \geq 8$ because the PHY capacity drops too much with outdated CSI.

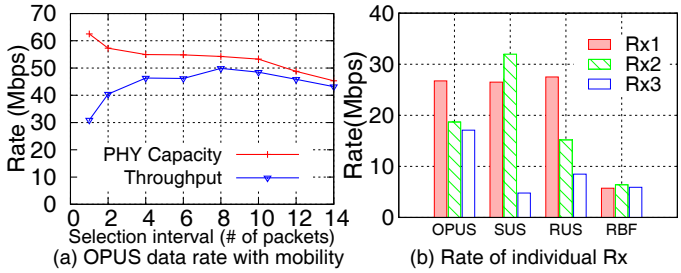

Fig. 11. PHY-layer capacity and network-level throughput under mobility.

Fig. 11(b) depicts the capacity of all schemes under this mobile scenario. OPUS still achieves the best sum-rate and fairness. SUS's capacity is only slightly lower than OPUS. However, this comes at the cost of fairness: SUS tends to prioritize the user group {Rx1,Rx2} and causes low performance for the mobile node Rx3.

D. Network-scale scalability test for practical number of users

We now test OPUS's network-level performance under practical number of users based on trace-driven emulation. The test topology for channel trace collection is shown in Fig. 13(a). We run each emulation under given user population for 30 minutes and record the resulting throughput of each user.

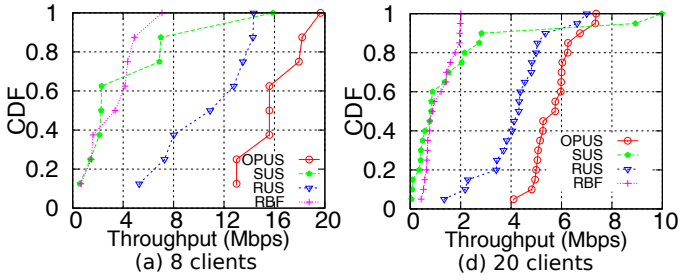

Fig. 12. Throughput comparison under practical user population.

Fig. 12 plots the CDFs of network throughput, where we can observe that: (i) OPUS always outperforms all other user selection schemes in both throughput and fairness (reflected in a steeper CDF). With 20 users, it achieves $1.4\times$ average throughput gain over SUS and $5.1\times$ over RUS/RBF. (ii) Random User Selection (RUS) naturally have comparable fairness, but much lower throughput than OPUS. (iii) SUS causes very low fairness and delivers high throughput only for few users.

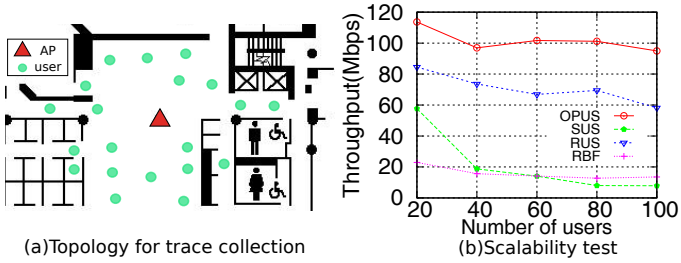

Fig. 13. The test topology and throughput comparison in scalability test.

To demonstrate the scalability of OPUS, we run trace-driven emulation by varying user population K from 20 to 100. The traces are generated by distorting a baseline 20-user trace with random complex multipliers. The resulting throughput

in Fig. 13(b) shows that OPUS significantly outperforms all other schemes under all network sizes. Moreover, its throughput performance is affected negligibly by increasing user number.

VI. Related Work

Existing MU-MIMO based network standards, including 802.11ac, WiMax and LTE, can run MU-MIMO only for a given set of users. They leave user selection as a vendor-specific operation. Substantial theoretical work has modeled the impacts of user selection on MU-MIMO downlink capacity and devised approximation algorithms to optimal user selection [2], [18]. For tractability, these works adopt simplified channel models, and assume CSI of all users is available as input. Low overhead user selection algorithms have been proposed [8] which use the statistical features instead of full CSI. But practical wireless channels are hard to be characterized by a stationary model, especially in indoor multipath environment and for mobile users. The work in [19] reduces the overhead by selecting users incrementally. However, it requires the AP to broadcast the CSI of all selected users before adding one user, which introduces more overhead in the downlink.

The majority of work in reducing CSI feedback overhead focused on designing compression algorithms and modeling their impact on MU-MIMO downlink capacity [3]. However, such mechanisms do not fundamentally solve the scalability problem, and incur non-trivial overhead as the user population grows. Implicit feedback can substantially reduce CSI overhead by leveraging channel reciprocity [20], but it still requires each user to send a packet, based on which the AP can infer the downlink channel state. It can be integrated into OPUS and replace the explicit CSI feedback from each client.

Experimental studies of MU-MIMO emerged only recently. Feasibility of MU-MIMO was validated in [12] through a software radio prototype. The NEMOx system [21] runs a MU-MIMO communications algorithm, and an incremental user selection mechanism that uses time-averaged CSI as input, which only fits relatively static network environment.

The principle of user selection is also reflected in opportunistic scheduling for cellular and ad-hoc networks (see *e.g.*, [22]). Such protocols select one link at a time — the one with highest channel quality, subject to certain fairness constraint. But MU-MIMO transmission involves a group of users whose rates are coupled through their channel orthogonality. Hence its user selection problem calls for brand new design choices.

OPUS's feedback contention mechanism inherits the principles of binary-countdown MAC protocols [16]. These protocols promote random access to ensure roughly equal access opportunity among contenders. In contrast, OPUS leverages binary-countdown to design a decentralized protocol that singles out a user with the highest preference metric.

VII. Conclusion

In this paper, we have introduced OPUS, a scalable user selection scheme for MU-MIMO networks. OPUS adopts a novel orthogonality probing mechanism that enables effective user selection at low overhead, regardless of the number of users in the network. It further incorporates a feedback contention protocol and a user number selection mechanism, which facilitate the orthogonality probing and aim to optimize network throughput with bounded overhead. We have prototyped OPUS along with three other state-of-the-art user selection schemes. In a typical medium-sized 802.11ac WLAN, OPUS can achieve a multi-fold throughput gain over other schemes while maintaining fairness. More importantly, it maintains high performance even when the user population rises to hundred-scale.

References

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
[2] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Downlink Channels with Limited Feedback and User Selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, 2007.
[3] X. Xie, X. Zhang, and K. Sundaresan, "Adaptive Feedback Compression for MIMO Networks," in *Proc. of ACM MobiCom*, 2013.
[4] T. Yoo and A. Goldsmith, "On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, 2006.
[5] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," *IEEE Std. 802.11ac Draft 3.0*, 2012.
[6] Rice University, "Wireless Open-Access Research Platform," http://warp.rice.edu/trac/wiki, 2013.
[7] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channels with Partial Side Information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, 2005.
[8] D. Gesbert, M. Kountouris, R. W. Heath, C.-B. Chae, and T. Salzer, "Shifting the MIMO Paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, 2007.
[9] R. Kudo, K. Ishihara, and Y. Takatori, "Measured Channel Variation and Coherence Time in NTT Lab," *IEEE 802.11-10/0087r0*, 2010.
[10] K. Karakayali, R. Yates, G. Foschini, and R. Valenzuela, "Optimum Zero-Forcing Beamforming with Per-Antenna Power Constraints," in *Proc. of IEEE International Symposium on Information Theory*, 2007.
[11] V. Sundarapandian, *Numerical Linear Algebra*. PHI Learning, 2008.
[12] E. Aryafar, N. Anand, T. Salonidis, and E. W. Knightly, "Design and Experimental Evaluation of Multi-user Beamforming in Wireless LANs," in *Proc. of ACM MobiCom*, 2010.
[13] K. Scharnhorst, "Angles in Complex Vector Spaces," *Acta Applicandae Mathematica*, vol. 69, no. 1, 2001.
[14] K. Tan, J. Fang, Y. Zhang, S. Chen, L. Shi, J. Zhang, and Y. Zhang, "Fine-Grained Channel Access in Wireless LAN," in *SIGCOMM*, 2010.
[15] E. Magistretti, K. K. Chintalapudi, B. Radunovic, and R. Ramjee, "WiFi-Nano: Reclaiming WiFi Efficiency Through 800 ns Slots." in *Proc. of ACM MobiCom*, 2011.
[16] H. Wu, A. Utgikar, and N.-F. Tzeng, "SYN-MAC: A Distributed Medium Access Control Protocol for Synchronized Wireless Networks," *ACM Mobile Networks and Applications Journal*, vol. 10, no. 5, 2005.
[17] Cisco Systems Inc., "Wireless Mesh Access Points, Design and Deployment Guide," *Release 7.3*, 2012.
[18] T. Ji, C. Zhou, S. Zhou, and Y. Yao, "Low Complex User Selection Strategies for Multi-User MIMO Downlink Scenario," in *Proc. of IEEE WCNC*, 2007.
[19] J. Mundarath, P. Ramanathan, and B. Van Veen, "A Distributed Downlink Scheduling Method for Multi-User Communication with Zero-Forcing Beamforming," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 11, 2008.
[20] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical Many-Antenna Base Stations," in *Proc. of ACM MobiCom*, 2012.
[21] X. Zhang, K. Sundaresan, M. A. A. Khojastepour, S. Rangarajan, and K. G. Shin, "NEMOx: Scalable Network MIMO for Wireless Networks," in *Proc. of ACM MobiCom*, 2013.
[22] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic Media Access for Multirate Ad Hoc Networks," in *Proc. of ACM MobiCom*, 2002.