MUTE: Sounding Inhibition for MU-MIMO WLANs

Oscar Bejarano*, Eugenio Magistretti*, Omer Gurewitz†, and Edward W. Knightly*
*ECE Department, Rice University, Houston, TX
{obejarano, emagistretti, knightly}@rice.edu
†CSE Department, Ben Gurion University, Beer Sheva, Israel
{gurewitz}@cse.bgu.ac.il

***Abstract*—In this paper, we present the design, implementation, and evaluation of the novel downlink Multi-User MIMO sounding protocol called MUTE. Our protocol decouples the sounding set selection used to collect Channel State Information (CSI), from the transmission set selection in order to minimize or even eliminate the overhead associated with sounding, while maximizing user selection performance. To this end, MUTE exploits channel statistics to all the different users to predict whether a particular user's channel will remain sufficiently stable, thereby allowing the access point to preclude channel sounding before a MU-MIMO transmission. We show that in indoor WLANs, MUTE can reduce sounding overhead by close to 73% under certain conditions while minimizing rate performance losses due to inaccurate channel estimation.**

I. Introduction

Zero-Forcing Multi-User-MIMO beamforming systems (ZFBF MU-MIMO) rely on *channel sounding* to provide the *beamformer* or Access Point (AP) with Channel State Information (CSI) about each *beamformee* or user. This is necessary to generate the steering beam weights required to perform the zero-forcing precoding prior to a beamformed transmission [13]. Additionally, it is advantageous to acquire CSI from all associated users in order to maximize *user diversity*,[1] thereby improving the user selection process by increasing the likelihood of finding beamformees with orthogonal or semi-orthogonal channel vectors. This can lead to complete suppression of interference between the different data streams serving the different beamformees and therefore to rate maximization at every transmission.

To this end, the beamformer can acquire channel estimates from all potential beamformees before every packet transmission. This provides the AP with accurate, up-to-date CSI about all users to be served, hence improving the performance of the precoding scheme. That is, having the most updated CSI for all users allows the AP to find the optimal user grouping strategy at every transmission. Unfortunately, the overhead required for CSI acquisition is directly proportional to the number of users to be sounded as well as the frequency with which this process takes place. Therefore, in a practical system, the beamformer should find a balance between sounding frequency and CSI accuracy, in the interest of minimizing sounding overhead.

In this paper we propose a multi-user zero-forcing beamforming sounding protocol that addresses the issue of overhead associated with channel sounding, with the goal of eliminating it temporarily based on channel stability. We name our protocol MUTE which stands for Multi-User Transmission Enhancer. In the best case, in the presence of users with stable channels, MUTE will invoke a MU-MIMO transmission without any immediately preceding channel sounding, thereby vastly reducing overhead and correspondingly increasing transmission air time and throughput. Nonetheless, MU-MIMO is very sensitive to the accuracy of CSI, specially as the number of concurrent streams increases. Therefore, MUTE strives to find a balance between CSI degradation and sounding suppression.

We argue that the decoupling of the *sounding* selection procedure from the *transmission* user selection procedure provides the flexibility to choose whether to sound a particular user or not, independently from the set of them to be served in the next ZFBF transmission. This in turn decreases overhead associated with sounding by exploiting the presence of users with stable channels, while independently providing sufficiently accurate information to the AP about channel statistics of associated users. Then, based on this information, the AP can select the combination of users that maximizes an objective function such as achievable rate or a fairness criteria, for example. This is in contrast to existing MU-MIMO implementations where the set of sounded users is the same as the set of users to be served next [4], [7], [12]. Furthermore, two of the major strengths of MUTE are interoperability with IEEE 802.11ac [3] devices as well as the fact that it can operate independently of the scheduler implemented.

MUTE employs a methodology comprised of the following set of mechanisms: *(i) in-situ training* which allows the AP to accumulate information about how rapidly the channels to all associated users are varying in order to generate predictions of current channel conditions based on the time elapsed since the last measurement. These predictions provide the knowledge to decide whether to sound a specific user or not based on channel statistics; and *(ii) idle sounding* which exploits idle channel intervals to opportunistically sound users to constantly update channel measurements to every user.

In particular, our main contributions are threefold:

First, we present a thorough analysis of the sounding overhead incurred in today's MU-MIMO systems, specially in IEEE 802.11ac systems. We demonstrate that even when considering large frame aggregation in order to amortize overhead, the overhead required in a four-user ZFBF transmission can reach 30% of the total transmission time in a 20 MHz channel and 60% in an 80 MHz one. Moreover, our analysis reveals that sounding more than four users before a ZFBF

[1]We define user diversity as the accommodation of a finite set of users with distinct channel characteristics.

transmission in such systems becomes prohibitive and should be completely avoided. This limits the amount of information the AP possesses before every transmission thus leading to a substantial decreases in user diversity, which is necessary in order for the user selection procedure to find the set of users that maximizes performance.

Second, we present the design, implementation, and evaluation of MUTE. MUTE consists of an MU-MIMO sounding protocol that *(i)* identifies the set of users for which sounding is unnecessary based on their channel stability, and *(ii)* relies on channel statistics about all such users to compute the weights needed to perform a ZFBF transmission. Therefore, our protocol minimizes sounding overhead while maintaining high user diversity. We introduce the mechanisms that MUTE employs to minimize sounding by exploiting the presence of users undergoing periods characterized by low channel instability. Additionally, we implement a ZFBF MU-MIMO transmission scheme in the software defined radio platform WARP [1], and rely on a combination of over-the-air transmissions and measurement-driven emulation to evaluate our protocol. Furthermore, we compare MUTE to a benchmark that relies on periodic sounding and up-to-date CSI before every single ZFBF transmission. Although this benchmark scheme does not incur in the rate penalty that our protocol does because of inaccurate channel estimates at the AP, we demonstrate that MUTE can still outperform the benchmark by achieving approximately 70% throughput gains in static environments.

Third, in order to assess the applicability of MUTE under realistic channel conditions we perform a comprehensive over-the-air measurement-based study of channel stability in typical WLAN environments. We explore indoor scenarios comprised of fixed and mobile users in both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS), having static and dynamic channels. We are the first to experimentally characterize the relation between the rate penalty due to residual interference in ZFBF systems and channel information age. Our study reveals that in relatively static environments MUTE can reduce sounding overhead by about 73% without incurring in significant rate losses due to lack of channel estimate accuracy. In addition, in the case of dynamic channels caused by the movement of surrounding objects in a busy university campus environment, we observe close to 55% overhead reduction, and a decrease of only 7% in rate due to outdated channel estimates.

II. MU-MIMO Sounding

ZFBF relies on sounding to acquire channel information about the different users to be served. Channel information is required for two main reasons: Firstly, the zero-forcing precoding technique requires channel estimates to compute the beam steering weights that multiply each of the data streams for the different users [4]; Secondly, the user selection process needs the channel information to each user in order to decide which of them should be served concurrently to minimize inter-user interference [13].

Sounding Overhead in Generic MU-MIMO. The overhead associated with sounding is directly proportional to the number of transmit antennas, users to be sounded, and to the frequency with which this process takes place. Different techniques have been proposed for sounding and retrieving channel estimates from the users. In general, methods can be classified as either explicit or implicit. Explicit sounding [3], [4] requires the AP to broadcast a pilot from each of its antennas so that the user can estimate its channel vector to the AP. Then, the channel information is fed back to the AP in order to use it to generate the beam weights. Let K be the number of single-antenna users, and M be the number of antennas at the AP (total of $M \cdot K$ channels). Then, assuming sounding over a single channel for K users, explicit sounding requires $O(M)$ time to send the pilots, and $O(M \cdot K)$ to feed back the estimated channel information. On the other hand, implicit sounding relies on uplink pilots originated from each user. This reduces the overhead to $O(K)$ [12].

Explicit vs. Implicit Sounding. Although implicit sounding requires less time to obtain channel estimates compared to explicit sounding, it has several drawbacks. First, it requires additional computation to calibrate the transmit and receive chains in each channel to maintain full channel reciprocity. This means that channel information matrices need to undergo a correction process to remove the mismatch between uplink and downlink channels. This lack of reciprocity is caused by imperfect electronic components and other effects such as random phase and amplitude differences in RF hardware [12]. While 802.11n allowed implicit feedback, in 802.11ac it was discarded. Apart from interoperability among chipsets from different vendors, another of the reasons for eliminating it from the standard was the fact that imperfect calibration at the transmitter is less tolerable in multi-user than in single-user beamforming because it leads to harmful interference leakage difficult to remove by the precoder. Likewise, depending on the precoding scheme implemented, calibration may also be required on clients in order to avoid introducing interference leakage. Another reason is that feedback cannot be collected from a beamformee having fewer transmit than receive antennas. More specifically, if the beamformee will be receiving on different antennas and only uses a few of them to transmit, it cannot perform implicit sounding since the AP requires the estimates to all the antennas. Moreover, the pilot transmission from the users needs to be coordinated by the AP; thus, at least one broadcast transmission is required to synchronize and trigger the pilots. In contrast, explicit feedback provides more reliable channel information matrices and does not require such calibration.

Sounding Overhead in 802.11ac [6]. The 802.11ac amendment strives to maintain high accuracy and reliability by proposing a unique explicit feedback method for obtaining channel information to enable MU-MIMO transmissions [3]. Nevertheless, we demonstrate that the cost for the proposed scheme is extremely high and becomes prohibitive as the number of users and antennas at the AP grows. The amendment proposes the following sounding and feedback mechanism (process depicted in Figure 1).

First, a unicast *Null Data Packet Announcement* or NDPA is transmitted by the AP indicating the subset of users required to prepare a compressed beamforming report. The word "compressed" describes the method used by the beamformee to represent the channel information in the form of phase/magnitude in a compressed feedback matrix $\mathbf{V}$.

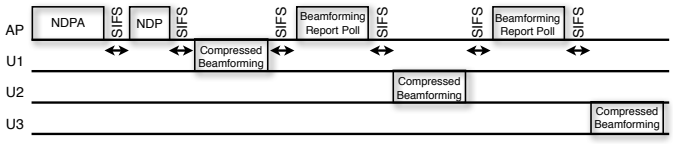

Fig. 1: 802.11ac Sounding and Feedback Timeline

Next, after SIFS, the AP sounds the channel using a *Null Data Packet* (NDP) having the format of the Physical-Layer Convergence Procedure (PLCP) protocol data unit (PPDU), but excluding the data field. The length of the NDP depends on the number of data streams. For instance, if the AP is serving four users, the total amount of time it takes to transmit this frame is greater than 50 μs.

Upon reception of the NDP frame, one of the chosen beamformees waits for SIFS and replies with the compressed beamforming report, which includes the information about the channel between AP and beamformee. Once the AP has received the report from the first beamformee, it polls the rest of the chosen users in order to acquire their beamforming reports as well. These reports represent most of the overhead in the sounding procedure of 802.11ac. Notice that although the amendment specifies the sounding procedure, it does not state which users should be sounded before a MU-MIMO transmission or how often. If we consider an AP with four antennas, the compressed feedback can vary from 180 to 1800 bytes [2]. The size of each feedback report can be estimated by multiplying the number of subcarriers by the number of angles used to represent each subcarrier, and by the number of bits required to represent each of these angles. We use the following 802.11ac parameters in our overhead analysis: SIFS time of 16 μs, Channel Width of 20 MHz, 4 antennas at the AP, 4 users, subcarrier grouping of 4, and quantization $\psi = 5$, $\phi = 7$ bits. These values were obtained from the 802.11ac amendment draft [3]. Based on those parameters, our computations reveal that these reports can take roughly 60% of the total *sounding duration*.

Finally, based on the retrieved channel information, the AP computes the steering weights and performs the MU-MIMO transmission. Based on the parameters above, if sounding is performed at base rate for robustness, we observe that the total amount of time required for sounding is about

$$T_{tot} = t_{NDPA} + 8 \cdot t_{SIFS} + t_{NDP} + 4 \cdot t_{report} + 3 \cdot t_{poll} \approx 745 \mu s$$

In Figure 2 we examine the fraction of airtime consumed by sounding overhead, i.e., the amount of time spent performing sounding out of the total time spent on a transmission. We do this for different transmission rates (from QPSK with $\frac{3}{4}$ FEC to 256-QAM with $\frac{5}{6}$ FEC), number of users (2 to 4), and channel widths (20 and 80 MHz). In order to amortize the sounding overhead over longer transmission durations, we consider frame aggregations varying from 12 frames (18 KB for a maximum packet length of 1500 B) [14] to the maximum A-MPDU aggregation with single MSDU of 64 frames (96 KB) [3]. Although the amendment defines even larger frame aggregations, such large packets can hardly be used under realistic conditions, e.g., see [14]. Notice that in the case of smaller packet sizes, the fraction of airtime spent in sounding would be considerably larger due to the shorter data transmission durations.

The figures shows that sounding overhead is dominant. Notice that the impact of sounding increases as data rates, channel widths and number of users increase, and as packet size decreases. Observe that even in 20 MHz channel, sounding 4 users consumes in all cases more than 10% of the total transmission time when considering 18 KB packets. In fact, as mentioned above, it takes the AP about 745 μs to sound the users; in the same amount of time, a device could transmit more than 1 kB of data at QPSK, and more than 10 kB of data at 256-QAM. In the rest of the paper we propose a scheme to significantly alleviate this overhead.

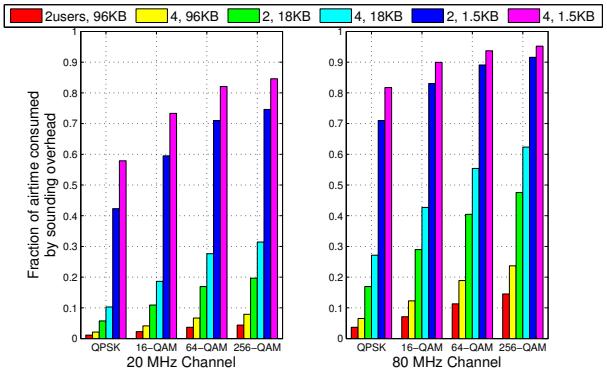

Fig. 2: 802.11ac Sounding Overhead

III. MUTE

MUTE highly reduces the sounding overhead of MU-MIMO ZFBF, while still providing the AP with sufficient channel information about the different associated users. More specifically, MUTE provides the AP with channel statistics about a large set of users, which in turn allows the AP to select the subset of those users that are expected to maximize a certain objective (e.g., rate or fairness), while incurring in only a small fraction of the sounding overhead needed in existing MU-MIMO implementations. Our protocol addresses the issues of which users to sound as well as the frequency with which sounding needs to occur, independently of the scheduling scheme used.

A. Decoupling Sounding & User Selection Procedures

To ensure that CSI is up-to-date before an MU-MIMO transmission, existing solutions consider both sounding and transmission user selection as a single process, i.e., in these systems, before every transmission the AP sounds all users to be served [4], [7], [12]. However, performing the costly sounding operation described in Section II frequently, leads to a low data-to-overhead ratio. Therefore, if channels do not change frequently for some users, the AP unnecessarily spends time sounding users for which the channels have not changed since the last time they were sounded. Moreover, in such implementations, every sounded user will be served in the next transmission. However, if the sounded set contains users with correlated channels then, ZFBF will perform poorly as it will not be able to suppress inter-user interference.

To address these issues, MUTE decouples the sounding user set selection from the transmission user set selection. That is, the former procedure selects the users to be sounded and the frequency at which they should be sounded, whereas the latter process picks the subset of users to be served simultaneously based on the available CSI at the AP. This allows the AP to exploit knowledge about users with static channels, thereby reducing the frequency with which these users need to be sounded. This in turn reduces the overhead required to do a ZFBF transmission with the goal of increasing the overall system efficiency in terms of data to overhead ratio. Notice that *decoupling* does not necessarily mean that both procedures disregard each other. More specifically, the user set selection utilizes information provided by the sounding set selection in order to operate, and vice versa.

With a decoupled system, if channel statistics are available and channels are deemed stable, the AP can sound as few as zero users and rely on their previous estimates to construct the beamformee subset, thus minimizing sounding overhead while serving up to M users. In contrast, a coupled system having channel statistics about the different users can only infer which of them might be good candidates to be sounded next but it would forcibly have to sound all those that the AP wants to serve. Additionally, if no channel statistics are available, a decoupled system could sound all M users before every transmission and determine the subset of those users that maximizes rate, for example. On the other hand, a coupled system would serve all M sounded users without discarding any, even if their combination leads to poor ZFBF performance (e.g., high inter-user channel vector correlation).

B. Sounding Set Selection Procedure

If users are static and the system starts with no information, then the sounding process in MUTE gradually reduces the number of users to be sounded before a transmission. Moreover, it does this while still providing accurate weight calculations for the set of potential users from which the transmission set selection procedure can choose as beamformees for the next transmission. MUTE operates independently of the scheduler and the objective function employed to decide which users should be served next. The sounding set selection procedure in MUTE is based in these mechanisms: *In-Situ training* - mainly to assess the dynamicity of each user's channel, and idle/bootstrapping sounding - to opportunistically collect channel statistics.

1) In-Situ Training: MUTE exploits the presence of users characterized by epochs of quasi-static channels to minimize the amount of users to sound. To achieve this, the AP requires channel statistics for each associated user in order to quantify the variation of each of their channels. We propose a novel *In-Situ Training* mechanism that uses collected channel measurements to determine the expected variation of the current channel to a user. This mechanism provides with a mapping from the time elapsed between every two sounding events for the same user, to the expected degradation in accuracy of the last channel measurement acquired, i.e., a proxy for per-user coherence time based on channel statistics.

Data Collection. The AP obtains channel measurements at every sounding procedure, i.e., magnitude *r* and phase θ of the complex entries in the channel vector fed back from each user. Then, it calculates the absolute magnitude and phase difference between the new sample and all the previously collected ones, i.e., $\delta_{r_{i,j,k}} = |r_i - r_j|$ and $\delta_{\theta_{i,j,k}} = |\theta_i - \theta_j|$ respectively, where i is the index of the most recent sample acquired by the AP for user k on each transmit/receive antenna pair. Here, j represents the index iterating over all previously collected samples ($1 \leq j < i$). This process allows the AP to estimate how much each old measurement has degraded compared to the current channel, i.e., how inaccurately the older measurement represents current channel conditions. Notice we consider single-antenna users, and eliminate the transmit antenna index to simplify notation.

After sounding any user, the AP records the following data: $(t_i, age_{i,j,k}, \delta_{r_{i,j,k}}, \delta_{\theta_{i,j,k}})$, where $age_{i,j,k}$ consists on the time elapsed between the newly obtained sample i and all samples j previously collected for user k. Since we expect the channel to each user to change completely after a certain time, we allow the system to reset and clear all accumulated channel statistics. As explained later in the section, this occurs after three consecutive packet losses or until a Time-To-Live (TTL) limit on the order of several minutes has been reached.

Determination of the Sounding Set. The decision of which set of users to sound in the next transmission is based on two main observations: *(i)* for a wide range of channel dynamics (excluding high mobility), the variability in the most recent samples (e.g., within the last few tens of milliseconds) can provide insights on how volatile the channel will be during the next few milliseconds; and *(ii)* channels in static or slowly varying environments can exhibit clear trends with different ages, i.e., during a given time period the difference between two consecutive channel samples can remain relatively constant (see Section IV);

To determine the sounding set for the next ZFBF transmission, the AP selects a set S of users in the service queue, and decides which subset $\hat{S}$ of them to sound. Notice that MUTE is implemented independently of the scheduling scheme employed, therefore, determining S is out of the scope of this paper. Then, MUTE obtains the set of *Relevant Samples* satisfying any of the following constraints: *(i)* samples that were recorded in the last τ_{recent} milliseconds. In this case, MUTE accounts for the most recent samples; and *(ii)* samples for which their sample age is within $\pm\tau_{age}$ milliseconds of the current sample age. This means that we are choosing a set of samples for which their recorded age is relatively close to the age between the current time t_{now} and the time of the last recorded sample t_{now-1}. Considering both datasets allows the AP to have a more conservative estimation of the expected channel variation, thereby reducing performance degradation due to stale information. The union of these two datasets constitute our *Relevant Samples*, which for the case of the magnitude change it is denoted by Δ_r

MUTE then computes a weighted mean μ^* and variance σ^2 based on the *Relevant Samples* dataset. Only two weight values are assigned to the weight w_i depending on the type of data (i.e., β for each of the *most recent* samples, and $(1-\beta)$ for each *age*-based sample, where we expect β to be higher than 0.5 due to the importance of newer samples). The computed variance indicates the amount of channel variability that the last channel

estimate obtained is expected to undergo by the time that particular user needs to be served. Finally, the weighted variance σ^2 of each transmit/receive path is compared to thresholds $\sigma^2_{r_{Thresh}}$ and $\sigma^2_{\theta_{Thresh}}$ for magnitude and phase, respectively. This threshold indicates the maximum variation in channel magnitude and phase allowed by the system in order to avoid significant losses in rate. That is, if a variation larger than these thresholds occur, sounding is triggered. This threshold reflects a balance of expected losses in rate due to channel estimate inaccuracies and overhead reduction. The values we set for these thresholds are determined experimentally in Section IV. Notice that the penalty in rate due to lack of accuracy of older estimates can be controlled via these thresholds; however, overhead reduction will adjust according to the threshold and current channel dynamics for each user. In other words, based on collected channel statistics, the AP infers a confidence level with respect to how precisely the most recent estimate is able to represent the current channel for a specific user and decides whether the accuracy of the previous estimate is sufficient to avoid sounding that user or not. To stay compatible with 802.11ac, if the variation on the entry for just one of the paths indicates that sounding is needed, then we sound that particular user, i.e., no partial sounding for each individual path.

MUTE's sounding procedure is shown in Algorithm 1. Due to space constraints we only refer to the magnitude change, however, the same procedure applies to the phase change.

Algorithm 1 MUTE's Sounding Procedure

while (1) **do**
  Initialization; Get S
  if $|S| > 0$ **then** ▷ %At least one user to be served
    for $\forall$ Tx/Rx antenna pair and $\forall$ users $s \in S$ **do**
      for $\forall$ i and j **do** ▷ %Select relevant samples
        if $t_{now} - t_i \leq \tau_{recent}$ **then**
          $\Delta_r \leftarrow \beta \times \delta r_{i,1,k}$
        end if
        if $|(t_{now} - t_{now-1}) - age_{i,j,k}| \leq \tau_{age}$ **then**
          $\Delta_r \leftarrow (1-\beta) \times \delta r_{i,j,k}$
        end if
      end for
      ▷ %Compute weighted mean and variance
      $\mu^* = \frac{1}{n}\sum_{l=1}^{n} w_l \Delta_{r_l}$
      $\sigma^2 = \frac{\Gamma_1}{\Gamma_1^2 - \Gamma_2}\sum_{l=1}^{n} w_l (x_l - \mu^*)^2$
      if $\sigma^2 \geq \sigma^2_{r_{Thresh}}$ **then**
        $\hat{S} \leftarrow$ User $s \in S$
      end if
    end for
    Sound users in $\hat{S}$
    Run user selection procedure, serve users $s \subseteq S$
  end if
end while

In certain cases the protocol might be dealing with small sets of samples, therefore we rely on an unbiased estimator of a weighted population variance. Thus, we let $\Gamma_1 = \sum_{i=1}^{n} w_i$, and $\Gamma_2 = \sum_{i=1}^{n} w_i^2$.

2) Idle Sounding: Traffic in WLANs has been shown to be highly bursty [10]. This intrinsic characteristic leads to periods of time when there is no data to be transmitted by the AP to the users. MUTE exploits these downlink idle periods by allowing the AP to opportunistically sound as many users as possible without delaying downlink data traffic for more than the length of a single beamforming report. That is, in the context of 802.11ac, the AP begins sounding by broadcasting NDP packets. As soon as it finishes, it will receive the beamforming report from the first user. Other users will be polled for their beamforming reports. Thus, if outbound data packets arrive at the AP for transmission, it is able to interrupt the polling and return to serve the users instead.

Likewise, to avoid congesting the network and affecting uplink traffic, we design the opportunistic sounding mechanisms to be conservative by doing the following. As soon as the service queue empties, the AP begins contending for sounding following the rules of the Distributed Coordination Function (DCF) in 802.11. To do so, the AP chooses a random number between CW_{last} and CW_{sound}, where the former takes the same value picked for the contention window in the previous beamforming transmission, whereas the latter is fixed to the maximum value proposed in 802.11 for the 6^{th} retransmission, i.e., $CW_{max} = 1023$. These choices allow users to precede the AP in terms of priority when no data needs to be transmitted on the downlink. Once the backoff counter reaches zero, the AP is able to sound users until it finishes or until a data packet arrives at its queue for transmission.

C. Opportunistic Transmission Set Selection Procedure

User selection consists of utilizing all the available information about the channels to the different users to construct beamformee subsets. Classic techniques for user selection rely primarily on the *separation* among the channel vectors of the receivers [13], matrix collinearity, and condition number [9]. Regardless of the metric used by the AP to group users for simultaneous transmission, MUTE provides with the flexibility to choose the subset of them independently of which users were sounded most recently. That is, the protocol provides with sufficient information based on current and past channel statistics so as to allow the AP to make a smart decision about how many and which users to serve next.

D. Packet Loss & Demotion to SU-MISO for Fast Channels

To ensure that our algorithm avoids performance losses caused by users characterized by highly varying channels (e.g, high user mobility), MUTE relies on the following: At the AP, we consider two different criteria to determine if a user is currently being affected by a highly dynamic channel. First, the AP keeps track of the number of consecutive packet losses to each user k. Second, the AP needs to determine if before each packet loss for user k, sounding was required by the *in-situ* training mechanism. A combination of three consecutive packet losses and the second criteria described for each of those packets, indicates a highly dynamic channel to user k. However, in the case of even a single packet loss for user k, we require the AP to sound that user before the next transmission to it. Each packet loss also leads to a one level decrease in the modulation and coding rate, e.g. from 16-QAM $\frac{1}{2}$ to QPSK $\frac{3}{4}$. In terms of retransmissions we follow DCF rules in 802.11 and allow up to seven retransmission for a single packet.

Users that have been flagged as having highly dynamic channels, will not be served via MU-MIMO transmissions. Whenever the first packet in the service queue is destined for a flagged user, no sounding occurs and the AP serves this user via a MISO transmission. Nonetheless, it is expected that some of these user's channels will eventually become more stable, therefore, when that happens it would be beneficial to continue serving them in MU-MIMO transmissions. To this end, the user exploits the intrinsic nature of wireless transmissions to overhear sounding transmissions that were destined to other users in order to calculate the stability of its own channel. Once user k determines that the expected channel variation is below both σ^2_{Thresh}, it informs the AP via a sounding request in the form of a standard uplink transmission packet. Upon reception of the "unexpected" reply from user k, the AP will no longer consider it as flagged and will consider it for sounding and ZFBF transmissions. Notice that we only require one bit of information in a standard 802.11ac packet to enable this mechanism.

IV. Implementing and Evaluating MUTE

MUTE's performance and gains, mainly depend on the key tradeoff between sounding frequency and channel estimation accuracy. In fact, interference nulling via ZFBF requires accurate channel knowledge, which depends on how frequently the channel is sounded and how fast the channel varies. As discussed above, MUTE chooses the sounding frequency based on channel dynamicity, ultimately striking a highly profitable balance between sounding overhead and interference nulling, i.e., the user achievable rates. In this section, first we investigate the relation between channel information age and rate penalty due to residual non-nulled interference, via a comprehensive set of measurements obtained in a testbed including static and mobile terminals forming LOS and NLOS links, in static and dynamic environments Our key finding is that under common channel conditions, channel age of few hundred miliseconds minimally impact the achievable rate. Next, we compare the performance of MUTE with a benchmark sounding scheme based on the standard 802.11ac via an emulation seeded with our real channel measurements, and demonstrate that MUTE's throughput gains can reach 70%. Finally, MUTE's monitoring of user channel dynamicity also benefits the transmission set selection; we conclude the section showing that user selection schemes can highly benefit from larger sets of transmission candidates.

A. Experimental Setup and Evaluation Methodology

We have implemented a ZFBF MU-MIMO transmission scheme in WARP [1]. WARP consists of an FPGA-based platform that includes custom designed radios based on the MAX2829 transceiver chip that operates over 20MHz channels. We perform experiments using WARPLab, a framework that enables the implementation of physical layer algorithms in MATLAB and over-the-air data transmissions by means of a central controller (host PC). WARPLab provides access to analog sample send/receive buffers and RSSI measurements for each experiment.

We consider a WLAN comprised of a single access point (AP) and up to four simultaneous users and we vary their location in order to obtain an extensive and representative set of samples for every scenario. The AP is equipped with four transmit antennas whereas the clients have only a single antenna. For our validation study and evaluation we first collect a comprehensive set of channel measurements in a university campus environment during busy days when the channel is highly dynamic due to environmental mobility (*dynamic environment*). Then, we obtain measurements in the same locations during night hours, in order to capture relatively static channels in the absence of environmental mobility (*static environment*). Additionally, we collect measurements from mobile terminals moving at pedestrian speed.

Emulation Methodology. To evaluate our protocol we collect over-the-air channel measurements and perform a trace-based emulation. Emulation allows us to compare different MU-MIMO schemes by replaying channels (i.e., side-by-side comparison over the exact same channels); notice that this repeatability cannot be achieved in a real-time setup. The emulator takes channel samples measured at each user, computes the ZFBF weights, and uses both the estimates and weights as input to Equation (1) in order to obtain the system's sum rate for a given channel realization. More specifically, consider a system with M antennas at the AP, and a total of K users. Given the $1 \times M$ channel vectors $\mathbf{h}_k$ collected at each receiver k and sent back to the transmitter, we compute the beam steering weight vectors $\mathbf{w}_k$. Then, we compute the sum rate according to Equation (1) [16].

$$R = \sum_{k=1}^{K} log\left(\frac{\sigma^2 + \sum_{j=1}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2}{\sigma^2 + \sum_{j=1, j\neq k}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2}\right) \quad (1)$$

where σ^2 is the noise variance. For more details on how to compute the beam steering weights, refer to [13]. Notice that when computing this rate, channel estimation errors or hardware drift are not taken into account. Nonetheless, we validate our emulator by comparing against our MU-MIMO testbed which considers such estimation errors and drift. To this end, we run over-the-air NLOS experiments for 20 user locations and all possible user combinations in an office environment. First, the AP sounds all users in order to obtain a set of channel measurements $\mathbf{h}_k$. These estimates are fed back to the AP, then the AP computes the weight vectors and performs an over-the-air ZFBF MU-MIMO transmission. Finally, based on the SNR measured at each user, we compute the aggregate rate (similarly to the process reported in [4], [12]). The total power used for transmission is constant, thus with increasing number of antennas, the power allocated to each antenna is reduced to $1/M$. In Figure 3 we present the aggregate rate that we achieve via emulation (theoretical) as well as the rate obtained using the platform (experimental). Observe that in average, the experimental rate reaches 97% of the rate achieved via emulation, which means that our emulator is able to achieve very similar performance to the real testbed.

B. Channel Estimation Accuracy

MUTE chooses to trade accuracy in channel information for a dramatic decrease in sounding overhead, relying on the observation that in MU-MIMO ZFBF small channel inaccuracies may lead to modest rate penalties. Accordingly, when

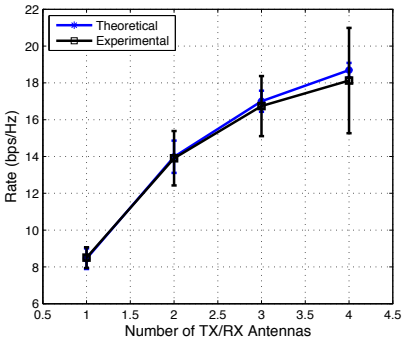

Fig. 3: Aggregate rate for a NLOS MxM system.

user channel varies infrequently, MUTE can systematically avoid sounding for long intervals and rely on information collected hundreds of miliseconds, if not seconds, before. In the following, we present experimental evidences of channel stability and investigate the tradeoff between age of channel information and rate penalty.

Channel Stability. In order to assess the feasibility of MUTE, we experimentally characterize the stability that channels show in terms of magnitude and phase under different environmental conditions and user behaviors. In particular, we investigate how the duration of the interval between successive measurements, i.e., the interval between sounding procedures, affects the accuracy of historical information that the AP possesses. Our experiment consists of collecting over-the-air channel samples between the four AP antennas and all receiving users (up to 8) for indoor LOS and NLOS scenarios for about 160 seconds, with consecutive samples spaced by 400 milliseconds. Then, for each transmit-antenna receive-user pair, we compare phase and magnitude difference of pairs of samples spaced by different time intervals multiple of 400 milliseconds, i.e., by different *ages*.

In Figure 4 we present average and standard deviation of the variation in magnitude and phase between every channel samples spaced by increasing intervals from 0.4 to 6.4 seconds, in LOS and NLOS static indoor environments. The figure is the result of 400 measurements per user per scenario. NLOS channel show larger variability than LOS; note that even in NLOS, the correlation between few hundred milliseconds spaced samples is still high. Specifically, with an age of channel information of 0.4 seconds, we may expect an avg. magnitude variation of less than 0.001 (resp. 0.0022) dB in LOS (resp. NLOS) conditions, and a phase variation of 0.026 (resp. 0.054) radians. In ZFBF, considering a target user, the impact of very small magnitude variations is negligible, while phase variations affect the amount of nulled interference, hence can result in much reduced SNR [13]. In the following, we show the extent to which these variations affect the user achievable rates.

Tradeoff Between Channel Aging and MUTE's Achievable Rate. Using inaccurate channel statistics to perform ZFBF transmissions can lead to degraded performance of the precoding scheme, therefore yielding rate losses due to non-nulled interference. To understand the extent of such effect we investigate the difference in rate achieved with a transmission using the most updated information (ideal case), compared

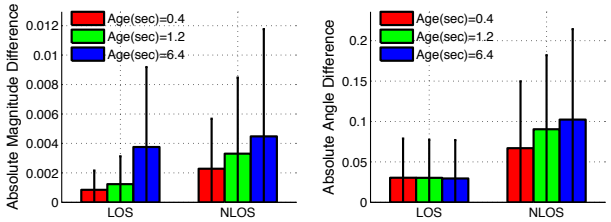

Fig. 4: Average absolute magnitude and phase variation in static LOS and NLOS indoor environments. Error bars show 95% confidence intervals.

to the case where we rely on older information; we term such difference *rate error*. Our experimental procedure is the following. First, we collect the channel matrices $\mathbf{H}_t$ at each sampling time t; then, for the *ideal case*, we obtain the ZFBF precoding weights $\mathbf{W}_t$ and compute the rate according to Equation (1) using the same $\mathbf{H}_t$, i.e., representing a transmission occurring over the same channel used to compute the weight matrix. For MUTE instead, we derive the ZFBF weight $\mathbf{W}$ using $\mathbf{H}_{t-age}$ and calculate the rate using channel $\mathbf{H}_t$ and weight $\mathbf{W}_{t-age}$, i.e., representing a transmission occurring over a channel potentially different from the one used to compute the weight matrix. The *rate error* derives from the fact that, if $\mathbf{H}_{t-age}$ differs from $\mathbf{H}_t$, our precoding will not completely null the interference due to transmissions toward different users, and thus increase the noise toward the intended receiver. As seen in the previous experiment, the larger the *age*, the more likely is the channel to vary between $t - age$ and t.

To investigate the *rate error* we use the same set of measurements we collected above, including LOS/NLOS links, static/dynamic environments. In Figure 5, we plot the CDF of the relative rate error (i.e., for each sample we calculate the rate error and divide it by the actual rate achieved using updated channel information) for a 4-antenna AP, for intervals between two consecutive samples (*age*) in the range between 0.4 and 6.4 seconds for NLOS links. First, we observe that even in the scenarios most adverse to MUTE (dynamic environment), using channel information 400 ms. old, the rate error is below 20% in 65% of the cases. This is a very promising result for MUTE, and shows that the penalty for infrequent sounding can be rather small. More importantly, as the number of streams decreases relative to the number of antennas at the AP, this penalty becomes less significant.

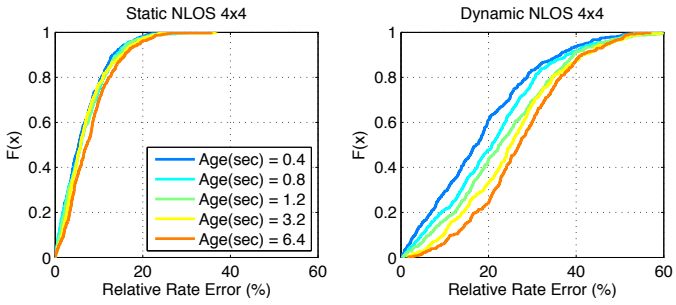

Fig. 5: Relative error rate for different sample *age*.

C. Throughput Gains in MUTE

In this section we assess the gain that MUTE achieves with respect to the benchmark scheme that sounds the channel before each transmission (thus having the most updated/accurate channel estimates possible). We consider three metrics, namely, rate penalty due to infrequently sounding (favorable to the benchmark), sounding overhead reduction (favorable to MUTE), and overall throughput gain of MUTE with respect to the benchmark.

We evaluate two versions of MUTE, each with a different rate loss tolerance. More specifically, in one version we set the thresholds $\sigma^2_{r_{Thresh}}$ and $\sigma^2_{\theta_{Thresh}}$ so as to allow MUTE to incur in a penalty of only about 2 bps/Hz, whereas on the second version we set them to allow only a 1 bps/Hz penalty (a rate loss tolerance). Thus, for the former case we expect a higher overhead reduction at the cost of a greater loss compared to the latter case. Thresholds are assigned based on the analysis of the variance in the measurements shown in Figures 4 and 5. In this experiment, we consider the same set of channel measurements collected above, i.e., all combinations of a 30 users set, in static and dynamic environments; finally, we define a new scenario (named "combined" in the figures), including also a set of users randomly moving in a 3m x 3m area at a speed of 0.5 m/s (about 5% of channels are mobile).

In Figure 6 (*left*) we present average and standard deviation of the per-user rate achieved by MUTE and the benchmark. This plot does not take overhead into account, therefore it represents the rate loss due to inaccurate historical channel information used by MUTE. Observe that at most, MUTE decreases the user rate by 10% and 22% (or 0.9 and 1.9 bps/Hz, respectively, for the two MUTE versions) with respect to the benchmark. However, Figure 6 (*right*) shows that this penalty permits a large reduction of the sounding overhead, ranging from about 55% to 95%. If we only allow a loss of 1 bps/Hz, this translates to an average sounding frequency decreasing from 400 ms of the benchmark, to about 2s/1s/1s of MUTE for static/dynamic/combined cases, respectively. In conclusion, MUTE can be tuned according to a configurable rate loss tolerance to achieve a large sounding overhead decrease (55-95%) at the price of a rate penalty (as low as 7% for a 1 bps/Hz tolerance).

Finally, we investigate the throughput gain that MUTE can attain compared to the benchmark; these results take both channel information inaccuracy and overhead reduction into consideration. When serving 4 users our system is constrained to transmit every two consecutive packets 400 ms apart; accordingly, for each packet transmission, we measure the airtime consumption based on rate achieved, packet size (from 1.5 kB to 18 kB), and sounding overhead (as detailed in Section 2). In this case, the throughput gain is the ratio between airtime consumptions of MUTE (with 1 bps/Hz tolerance) and that of the benchmark. Even though we neglect the time between two consecutively transmitted packets because of our system limitations, this procedure provides an estimate of back-to-back transmissions seeing channels within the statistical distribution of the interval extremes. In Figure 7 we plot the percent throughput gain achieved by MUTE for different frame sizes. The gain decreases as the packet length -and duration- increases; this is because the portion of time spent in sounding decreases. However, observe that in static conditions, our scheme reaches up to 70% gains for 1.5 kB frames and up to 28% with very large 18 kB frames, due to the significant reduction in sounding overhead. In the worst case, i.e., dynamic scenarios with 18 kB frames, MUTE can still attain 17% gains due to ∼55% overhead reduction, while incurring in small rate inaccuracies. In conclusion, in a variety of WLAN scenarios, MUTE largely outperforms periodic-sound based schemes.

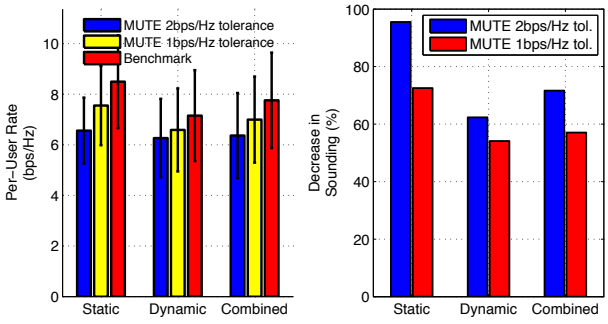

Fig. 6: Performance of MUTE under different scenarios. *Left* plot does not consider impact of overhead in rate performance

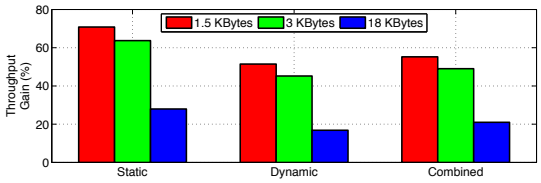

Fig. 7: MUTE's Throughput Gain - 1 bps/Hz tolerance

MUTE Leverages User Diversity. Rate selection in MU-MIMO should try to avoid grouping together users with correlated channels. Thus, knowledge of the channels of a large set of users (*user diversity*), i.e., ideally much larger than the number of users expected to be served, leads to the possibility of selecting a higher-rate user set. At each transmission the benchmark knows only the channels of the users it sounded immediately before transmitting. In contrast, MUTE simultaneously monitors the channels of multiple users, thus allowing the user selection procedure to choose among a larger set of users. The effect of *user diversity* in MU-MIMO systems has been previously studied from an information theoretical perspective [13]. In this paper, we isolate and explore this effect experimentally in order to quantify the gains that MUTE can attain by leveraging *user diversity*. We consider a network comprised of a single AP with 4 antennas, and 30 single-antenna users, and we repeat the experiment for 400 different channel instances. For each experiment, i.e., for each channel instance, the AP chooses to serve the combination of m users that maximizes the aggregate rate from a set of n users uniformly selected from the 30-user population. Additionally, we compare against an exhaustive search approach that selects the combination that maximizes the rate among all users by choosing the best combination of m users.

In Figure 8 we present the aggregate rate for both $m = 1$ (i.e., a 4x1 system) and $m = 4$ (i.e., 4x4), for n increasing from 1 to 10, in order to evidence how MUTE benefits

the user selection process as the number of user channels monitored increases. First, we observe that coupled schemes (e.g., benchmark), represented by the value of $m = n = 1$ in the left figure, and $m = n = 4$ in the right, are highly suboptimal, renouncing to 47% of the capacity in the $m = 1$ case, and to 48% in the $m = 4$ case. Interestingly, if the channel conditions (as we explored in the previous section) allow MUTE to systematically add even only 5 users out of 30 (i.e., a mere 1 user every 6), the capacity gap would decrease to 12% and to 21%, respectively, for a gain of 68% and 45% with respect to the benchmark; this translates to 5.1 and 10.8 bps/Hz increases. Such large gains for even a small number of users monitored are made possible by the diminishing returns of sounding an increasing number of users. In conclusion, in contrast to a conventional coupled system where only up to four users can be sounded and served due to prohibitive sounding overhead, MUTE permits the selection scheme to leverage the knowledge of the channels of multiple users and achieve larger gains via user diversity.

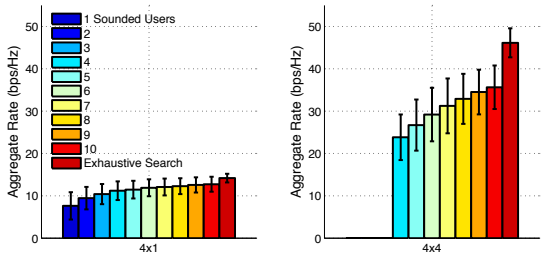

Fig. 8: User diversity in SU-MISO and MU-MIMO. No data displayed for the 4x4 system below 4 users.

V. Related Work

Prior work is comprised of theoretical work on MU-MIMO WLANs and cellular networks including channel feedback analysis, as well as MU-MIMO implementations proposing sounding, grouping, and rate adaptation techniques. Theoretical works most relevant to MUTE address the issue of user selection, e.g., [7], [13]. All these schemes are orthogonal to MUTE and, as shown in the last section, can take advantage of the larger set of user channels reported by MUTE.

Several MU-MIMO ZFBF experimental works have recently appeared in literature. In particular, the work in [4] implements and evaluates the performance of ZFBF MU-MIMO schemes; however, the scheme proposed is still based on explicit channel sounding at each transmission similarly to 802.11ac. [12] proposes the Argos design, extending the implementation of ZFBF MU-MIMO to a 64-antenna APs. While [12] employs a calibration scheme for implicit sounding to reduce the sounding overhead, such a scheme still requires *all* users to be sounded before *each* transmission, introduces additional rate errors, and could equally benefit from the infrequent sounding in MUTE. [5] compares alternative MU-MIMO schemes; their results confirm that schemes based on CSI largely outperform alternative schemes. Finally, [11] and [15] address the uplink case where the main challenge is the coordination of user transmissions, and [8] explores diversity and multiplexing gains in MU-MIMO. [17] addresses feedback overhead by means of CSI compression in frequency and time. In contrast to all MU-MIMO implementations, we propose a sounding protocol for downlink beamforming transmissions that reduces the overhead by exploiting channel statistics and channel stability.

VI. Conclusion

In this paper we have analyzed the overhead associated with sounding in indoor MU-MIMO WLANs and proposed MUTE which exploits the presence of users with slowly-varying channels in order to minimize this overhead. To this end, the protocol relies on historical CSI obtained via previous soundings to predict the variation in channel magnitude and phase given the amount of time that has passed since the last measurement for a specific user was collected. Using testbed experiments and measurement-driven emulation, we show that MUTE can significantly reduce sounding overhead without incurring in meaningful rate penalties.

VII. Acknowledgements

This research was supported by NSF grants CNS-1314822, CNS-1126478, CNS-1012831, CNS-1012921 and by a grant from Cisco Systems Inc.

References

[1] Rice University WARP project. Available at: http://warp.rice.edu.
[2] Cisco inc. *802.11ac: The Fifth Generation of Wi-Fi, Technical White Paper*, August 2012.
[3] IEEE 802.11ac/D3.0, Enhancements for Very High Throughput for Operation in Bands Below 6 GHz. 2012.
[4] 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*, pages 197–208, 2010.
[5] H. V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire. Achieving high data rates in a distributed MIMO system. In *Proc. of ACM MobiCom*, 2012.
[6] O. Bejarano, E.W. Knightly, and M. Park. IEEE 802.11ac: from channelization to multi-user MIMO. *IEEE Comm. Mag.*, 51(10):84–90, 2013.
[7] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran. Multiuser MIMO achievable rates with downlink training and channel state. *IEEE Trans. Inf. Theor.*, 56(6):2845–2866, June 2010.
[8] B. Chen, K. C. Lin, and H. Wei. Harnessing receive diversity in distributed multi-user MIMO networks. In *Proc. of ACM SIGCOMM*, 2013.
[9] N. Czink, B. Bandemer, G. Vazquez-Vilar, L. Jalloul, C. Oestges, and A. Paulraj. Spatial separation of multi-user MIMO channels. In *Symposium PIMRC*, 2009.
[10] I. W. C. Lee and A. O. Fapojuwo. Analysis and modeling of a campus wireless network tcp/ip traffic. *Comput. Netw.*, 53(15):2674–2687, October 2009.
[11] W. Shen, Y. Tung, K. Lee, K. Lin, S. Gollakota, D. Katabi, and M. Chen. Rate adaptation for 802.11 multiuser MIMO networks. In *Proc. of ACM MobiCom*, 2012.
[12] 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.
[13] Y. Taesang and A. Goldsmith. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE JSAC*, 24(3):528–541, 2006.
[14] K. Tan, J. Fang, Y. Zhang, S. Chen, L. Shi, J. Zhang, and Y. Zhang. Fine-grained channel access in wireless LAN. In *Proc. of ACM SIGCOMM*, 2010.
[15] K. Tan, H. Liu, J. Fang, W. Wang, J. Zhang, M. Chen, and G. M. Voelker. SAM: Enabling practical spatial multiple access in wireless LAN. In *Proc. of ACM MobiCom*, 2009.
[16] H. Viswanathan, S. Venkatesan, and H. Huang. Downlink capacity evaluation of cellular networks with known-interference cancellation. *IEEE SAIC*, 21(5):802 – 811, June 2003.
[17] X. Xie, X. Zhang, and K. Sundaresan. Adaptive feedback compression for MIMO networks. In *Proc. of ACM MobiCom*, 2013.