Mode and User Selection for Multi-User MIMO WLANs without CSI

Narendra Anand
Rice University
Houston, TX, USA
nanand@rice.edu

Jeongkeun Lee
Hewlett-Packard Laboratories
Palo Alto, CA, USA
jklee@hp.com

Sung-Ju Lee
KAIST
Daejeon, Korea
sjlee@cs.kaist.ac.kr

Edward W. Knightly
Rice University
Houston, TX, USA
knightly@rice.edu

***Abstract*—A Multi-User MIMO (MU-MIMO) Access Point (AP) can obtain a capacity gain by simultaneously transmitting to multiple clients. This technique requires Channel State Information (CSI) at the transmitting AP to set antenna gains and phases to enable simultaneous reception through beamforming. The AP must also select both the mode (number of transmit and collective receive antennas) and the user set prior to transmission. While the ideal mode and user selection is a function of CSI, CSI must be estimated with an overhead intensive channel sounding process. We design, implement, and evaluate Pre-sounding User and Mode selection Algorithm (PUMA), a method for mode and user selection *prior* to channel sounding. We show that even without CSI, PUMA *(i)* exploits theoretical properties of MU-MIMO system scaling with respect to mode, *(ii)* characterizes the relative cost of each potential mode, and *(iii)* estimates per-stream transmission rate and aggregate throughput in each mode for a potential user set, all without CSI. Once PUMA has selected the appropriate mode and user group, the chosen protocol's channel sounding method is used on the intended user subset to carry out the transmission. We show that, on average, PUMA selects the mode and group that achieves an aggregate rate within 3% of the saturation throughput of what would have been achieved by sounding all users (which would require significant additional overhead). Moreover, we show that PUMA obtains 30% higher aggregate throughput compared to the best fixed-mode policy that uses the maximum number of available transmit and receive antennas.**

I. Introduction

MU-MIMO (Multi-User Multiple Input Multiple Output) achieves substantial capacity gains by using precoding to support multiple, concurrent data streams to a group of clients. Precoding comprises of computing the transmitter's antenna gains and phases from the channel state information (CSI), i.e., the channel matrix in which each element represents the magnitude and phase offset for each transmitter-receiver antenna path. In this way, each receiver can simultaneously decode its streams [25]. Moreover, the recent IEEE 802.11ac amendment promises multi-Gb/s rates via down-link MU-MIMO with up to 8 transmit antennas at the AP [5], [12].

To realize these capacity gains, in addition to precoding, the AP must also select the *(i) mode:* the number of transmit antennas and collective number of receiving antennas, and *(ii) users:* the set of receiving antenna(s), i.e., clients. For each transmission, the ideal mode and user set is channel-dependent and therefore their selection would require CSI for *all* receivers. However, due to large overhead, it is practically infeasible to collect CSI for all potential receivers prior to each transmission. For example, 802.11ac requires approximately 1.6 kb to 329 kb per client depending on the number of transmitting and receiving antennas.[1]

Previous solutions ([11], [13], [18], [20]) focus on mode and user selection *after* channel sounding is complete. These methods pick the optimal user groups given full CSI of a set of potential receivers or by relying on intermittent probing or stale CSI to estimate the full CSI. The additional overhead these methods require could substantially mitigate the benefits of MU-MIMO transmissions. The key to efficient MU-MIMO transmissions is the amortization of sounding overhead over the transmitted data. Works such as [23] provide methods of compressing the sounding phase thus increasing the amortization of the sounding overhead. However, the benefits of these algorithms are only realized after the transmission mode and group are selected.

To achieve these goals while ensuring the efficient amortization of sounding overhead, we present **P**re-sounding **U**ser and **M**ode selection **A**lgorithm (PUMA), a method of mode and user selection *prior* to channel sounding. The key techniques of PUMA are threefold: *(i)* estimation of expected per-user datarate based on theoretical MU-MIMO system scaling, *(ii)* characterization of relative cost from overhead for each potential mode, and *(iii)* calculation of expected aggregate throughput for a potential group of users given a particular mode. Through this pre-sounding estimation process, PUMA selects the best mode and group of users by maximizing throughput with respect to overhead.

Estimation of Per-User Datarate. By exploiting the properties of theoretical MU-MIMO system scaling, PUMA estimates the potential datarate of a particular receiver prior to sounding. Theoretical MU-MIMO system scaling is based on the *Degrees of Freedom (DoF)* of the transmission mode. DoF refers to how many more transmit antennas (M) than collective receive antennas (K) exist in a particular transmission mode.

The available DoFs manifests as the following tradeoffs with respect to received SINR: increasing the number of

This research is sponsored in part by NSF Grants CNS-1444056 and CNS-1126478.

[1]The 802.11ac standard allows for a transmitter with up to 8 antennas, a receiver with up to 4 antennas, and up to a maximum of 4 concurrent receivers. The respective feedback overhead for 2 Tx antennas, 2 single-antenna Rx, 16 bits per angle, 20 MHz BW is 1664 bits, and for 8 Tx antennas, 2 four-antenna Rx, 16 bits per angle, 160 MHz BW is 329,472 bits [12].

transmit antennas M increases the per-user SINR. Increasing the number of receiving antennas K decreases the per-user SINR but increases the number of parallel data streams (thus resulting in additional, lower datarate, parallel transmissions). Leveraging this tradeoff, PUMA computes the expected per-stream datarate for a potential mode using only M and K without CSI.

While estimating the expected throughput using the available DoFs is not perfect, it is sufficient for mode and user group selection because the indoor Wireless LAN (WLAN) environment usually results in well conditioned channel matrices due to the prevalence of multi-path effects [4]. Thus, the available DoFs has a greater effect on a user's served MU-MIMO SINR than the actual relationship between concurrent users' channel vectors [4].

Characterization of Potential Mode Cost. After computing the expected datarate of a MU-MIMO transmission, PUMA estimates the relative cost of realizing a particular transmission mode. This estimation is based on the aforementioned expected receiver datarate, the net transmission overhead required for a particular MU-MIMO mode, and the amount of available data (backlog) for each receiving node.

The ratio of transmitted data to overhead for a potential mode manifests as the following tradeoff: increasing M and K increases transmission overhead but also allows for increased parallel data streams. Increasing the number of DoFs $(M - K + 1)$ results in increased per-stream datarate (due to the aforementioned per-user SINR increase). Increasing the amount of transmitted data aids in amortizing the MU-MIMO transmission overhead (thus if less data is available to transmit, the overhead cost for an MU-MIMO transmission can outweigh the benefit of the amount of data transmitted).

Calculation of Group and Mode Specific Aggregate Throughput. Given the maximum number of available transmit antennas and all potential user subsets, PUMA computes aggregate throughputs for each potential mode and user group. Thus, every potential MU-MIMO transmission is assigned a quantifiable metric for comparison and maximization. This maximization seeks the best mode and user group based on each user's expected datarate, each user's backlog, and overall cost of serving each potential user group, only leveraging information available *prior* to channel sounding.

We validate PUMA's expected datarate calculation through over-the-air (OTA) experimentation and evaluate the overall performance of PUMA through OTA-trace based emulation. We show that PUMA achieves an aggregate transmission rate that is within 3% of a system that had full CSI for all potential users. Additionally, we show that PUMA achieves at least a 30% higher aggregate throughput than any fixed-mode policy.

The remainder of the paper is organized as follows: Sec. II provides an overview of the PUMA protocol. Sec. III discusses PUMA's functionality specifically with 802.11ac. Sec. IV evaluates the performance of PUMA with OTA experimentation and channel-trace driven emulation. Sec. V describes related work and Sec. VI concludes the paper.

II. PUMA Overview

The PUMA algorithm is executed before the start of any MU-MIMO transmission with only *a priori* information; i.e., without any information garnered from previous multi-stream communication or channel sounding. The necessity of using only immediately available information for mode selection stems from the highly volatile nature of an indoor WLAN environment.

Although this multi-path rich, fading environment results in well conditioned channels, indoor WLAN environments can have an unpredictable coherence time [2], [4], [14]. The measured CSI can easily become stale between packets, making previous transmissions unhelpful for predicting future environments. Additionally, because of the high overhead incurred from measuring CSI, it would be costly for the transmission mode selection to rely on channel sounding. To alleviate the effects of variable and costly CSI measurement, PUMA selects the best mode and user group without CSI.

The following sections detail how, using the information available before channel sounding (Sec. II-A), PUMA predicts the per-user MU-MIMO datarate (Sec. II-B), computes a potential group's throughput (Sec. II-C), and finally selects the appropriate user/group combination (Sec. II-D).

A. Available Pre-sounding Information

Before initiating an MU-MIMO transmission, the AP has the following information: *system state*, *queue state*, and *link state*. By leveraging this *a priori* information, PUMA enables an AP to select the best mode. The system state and queue state are used directly in protocol overhead calculation detailed in Sec. II-C. Link state is leveraged in per-user data rate prediction detailed in Sec. II-B. Combining these components, PUMA estimates the throughput of any possible MU-MIMO transmission an AP can execute.

System State. Before any transmission, the AP knows the hardware configurations of itself and its clients. This includes the available (maximum) number of transmit antennas M_{max} and the available number of associated users' receive antennas K_{max}. PUMA leverages this system state for overhead computation. While a greater number of overall antennas results in increased data transmissions, it also significantly increases sounding overhead.

Queue State. The AP is also aware of each receiver's backlog or queue size. The amount of available data directly affects how much sounding overhead is amortized. If the amount of available data for a particular user is relatively small compared to sounding overhead, the potential gains of a MU-MIMO transmission to that user are severely diminished. We express the available data in terms of available packets b.

Link State. The AP is aware of each user's link state or omnidirectional SNR. The AP automatically gathers this information from periodic beacon messages and updates this information after each received packet. Unlike CSI, this metric need not be instantaneous since received signal strength stays coherent longer than multiple packet transmissions (approxi-

mately 90 ms at 0.9 kph [8]). PUMA leverages each client's link state to estimate the achievable data rate.

B. Predicting User-specific MU-MIMO Datarate

The first key technique of PUMA is the estimation of per-user datarate. PUMA accomplishes this by computing the expected SINR of an MU-MIMO transmission using only pre-sounding information based on theoretical MU-MIMO system scaling. PUMA then estimates the achievable rate using a protocol specific minimum SINR table (such as Table I).

1) Post-Sounding Rate Estimation: Many works provide expressions for the expected received SINR or aggregate capacity of a MU-MIMO transmission (e.g. [25], [17]) such as the following (where C is in b/s/hz):

$$C = \max_{\mathbf{w}_k, P_k} \sum_{k=1}^{K} \log_2 \left(\frac{1 + \sum_{j=1}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2}{1 + \sum_{j=1, j \neq k}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2} \right). \quad (1)$$

However, such methods are not suitable for our purposes because they require post-sounding information (the measured channel matrix) to obtain the channel matrix H. Instead, we seek to estimate the expected performance of a MU-MIMO transmission *before* the channel is sounded.

Given the significant overhead of channel sounding (which we discuss specifically for 802.11ac in Sec. III), a transmitter must serve whatever user it sounds to maximize performance (to be described in Eq. (4)). Additionally, the channel state is highly variable for the frequencies used in WLANs [4] and thus channel sounding must occur before every packet transmission (i.e., previously measured channel matrices cannot be used reliably for future transmissions).

2) Pre-Sounding Rate Estimation: PUMA's pre-sounding rate estimation method is based on theoretical MU-MIMO system scaling. PUMA exploits this scaling by estimating the received SINR for a particular client and converting it into an expected achievable datarate using a standard specific minimum SNR table such as Table I.

The basis of Eq. (1) is the computation of SIR from the multiplication of the h and w vectors. While the H matrix represents the measured CSI, the W matrix is what the AP computes from H to actually construct parallel streams.

A commonly employed MU-MIMO precoding technique, Zero-Forcing [25], requires that the W matrix be computed as the inverse of the H matrix. Beamforming itself is the application of the W steering matrix through the channel H or $H \cdot W$. While a matrix times its inverse should result in the identity matrix, the actual value of H and W may not precisely meet this criterion due to per-user power allocation or an ill-conditioned H.

Eq. (1) is based on this matrix multiplication $H \cdot W$. The additional computations are simply to convert SINR into Shannon Capacity.[2] Thus, the result of the matrix multiplication is the diagonal matrix L and is a representation of the received SINR. Each diagonal element l_i corresponds to each of the K receiving antennas and its magnitude encompasses the beamforming gain (or loss) with respect to the received omnidirectional signal strength P/N_o [16].

Therefore, assuming equal per-user power allocation (and normalized per-antenna power allocation), the actual SINR for a beamforming transmission based on this measured H is:

$$\text{SINR} = 10 * \log_{10} \left(\frac{P/N_o}{M} |l_i|^2 \right). \quad (2)$$

This expression still leaves us in the same position as with Eq. (1). However, instead of attempting to calculate $|l_i|^2$, we consider its distribution.

The distribution of the SINR determining $|l_i|^2$ factor can be shown to be Erlang for Rayleigh matrices [10]. The distribution is dependent upon the dimensions of H (M and K) and has mean $(M-K+1)/K$. Therefore, combining with Eq. (2), we estimate per-user SINR as

$$\mathcal{E}\{\text{SINR}_{\text{BF}}\} = 10 \cdot \log_{10} \left(\frac{M - K + 1}{K} \frac{(P/N_o)}{M} \right). \quad (3)$$

While the resulting expected value computation of per-user SINR shown in Eq. (3) is inherently less precise than Eq. (1) because it does not use CSI, Eq. (3) exploits general system scaling properties of MU-MIMO transmissions to produce a sufficiently accurate result (verified in Sec. IV-B).

3) Model Rationale: This scaling is proportional to the available DoFs of a particular transmission mode. A mode's DoFs refer to how many more transmit antennas there are than receive antennas or $M - K + 1$. The larger this value, the easier it is to construct interference free parallel streams. An [M, K] transmission requires an $M \times K$ channel matrix, which is easier to accurately invert or otherwise decompose when $M > K$ since it will be better conditioned [24].

However, absolute DoFs do not reveal the full solution for theoretical MU-MIMO received SINR scaling; instead we consider normalized DoFs. For example, although both [M_{10}, K_9] and [M_3, K_2] transmissions have equivalent absolute DoFs (2), the per-user SINR increase would be far more noticeable in the latter system because it has relatively more DoFs with respect to K. Thus, MU-MIMO SINR should scale relatively with $(M - K + 1)/K$.

4) Inferring Rate from SINR: PUMA's SINR estimation method, Eq. (3), only requires the M and K of a potential mode and a particular user's omnidirectional signal strength (P/N_o) periodically updated from beacon packets and previous transmissions. Like CSI, the omnidirectional SNR can become stale after a period of time. However, omnidirectional SNR is far more robust to environmental variation than CSI. Channel matrices used for MU-MIMO transmissions are dependent on precise magnitude and phase offsets between each antenna path. Given the wavelengths of the frequencies used for WLANs and their physical interactions with obstacles, slight variations in the transmission environment can render a previously measured magnitude or phase useless. Instead, SNR is a coarse grained measurement that is an aggregate of all amplitudes and thus varies more slowly.

[2]Recall that the Shannon Capacity is computed as $C = \log_2(1 + \text{SINR})$.

PUMA estimates the received SINR for each user in an $[M,\ K]$ system, which allows an AP to not only compare an $[M,\ K]$ system to an $[M',\ K']$ system, but also estimate an approximate MCS rate for each user using the SNR-MCS tables provided by the standard (for 802.11ac see Table I).

C. Computing Expected Throughput

The second component of PUMA is the analysis of the selected MU-MIMO protocol specifically with respect to its aggregate throughput (R) using renewal arguments.

Other than the inferred expected rate (detailed in Sec. II-B), the main components of expected throughput calculation are node backlog (the number of available packets to transmit per receiver) and net transmission overhead. By combining node backlog (client dependent), net overhead (mode dependent), and the expected rate (client and mode dependent), PUMA computes the expected goodput of any possible MU-MIMO transmission. Thus, PUMA provides a quantifiable metric for any possible transmission, allowing an AP to accurately compare potential mode and group selections.

The throughput R for any wireless transmission is generally represented as the amount of data to transmit divided by the total transmit time (including overhead):

$$R = L_D/(T_D + T_{OH}). \tag{4}$$

The total amount of transmitted data across all streams L_D given the maximum packet length L_p is:

$$L_D = \sum\nolimits_{i \in K} b_i \cdot L_p. \tag{5}$$

The overhead time T_{OH} is:

$$T_{OH} = T_S + T_{CF} + T_{ACK} \tag{6}$$

where T_S is the channel sounding time, T_{CF} is the channel feedback time, and T_{ACK} is the receiver acknowledgment time. Thus, T_D, the total data transmission time given the per-user rate r_i is:

$$T_D = \max_{i \in K} (b_i \cdot L_p)/r_i. \tag{7}$$

We express T_D as a maximum value in case the protocol (e.g., 802.11ac) supports different per-user packet aggregation rates b_i or different per-user modulation rates r_i.

Through this formulation of aggregate goodput, we see that the value of T_{OH} limits the performance of a MU-MIMO transmission. A larger user set K served with more antennas M results in a larger amount of transmitted data L_D but it also results in a larger amount of overhead T_{OH} (which, like L_D, also scales with M and K). The appropriate mode is one that maximizes R by efficiently balancing L_D and T_{OH}.

While the basis of PUMA's overhead analysis is applicable to any standard, we focus on 802.11ac in Sec. III.

TABLE I
REQUIRED SNR (FOR 90% PACKET RECEPTION RATE).

MCS	Rate	N_{DBPS}*	SNR(dB)	MCS	Rate	N_{DBPS}*	SNR(dB)
0	BPSK 1/2	117	1.1	5	64-QAM 2/3	936	17.2
1	QPSK 1/2	234	4.1	6	64-QAM 3/4	1053	18.4
2	QPSK 3/4	351	6.7	7	64-QAM 5/6	1170	19.7
3	16-QAM 1/2	468	9.6	8	256-QAM 3/4	1404	23.9
4	16-QAM 3/4	702	12.8	9	256-QAM 5/6	1560	25.5

* N_{DBPS}(number of data bits/symbol) for each MCS with 80 MHz channel bandwidth.

D. PUMA Algorithm

PUMA seeks to jointly minimize the effects of T_{OH} and maximize the value of R from Eq. (4). Essentially, for a set of per-user rates r_i and per-user backlog b_i, PUMA computes:

$$\max_{M \in M_{max}, K \in K_R} R(M, K, b, r) \tag{8}$$

where M_{max} is the maximum possible number of transmit antennas and K_R is the subset of all potential associated receivers with packets in their queues.

One method of maximizing this expression is an exhaustive search. The value for b is set per-user and the value of r is dependent upon how many other concurrent users exist.

Thus, the overall search space is:

$$\sum_{M}^{M_{max}} \sum_{K=1}^{M} \binom{K_R}{K}. \tag{9}$$

This number of potential combinations can be exhaustively searched as long as K_R is not too large. Limiting K_R can be done in any number of ways such as truncating users with small b_i, fairness, or other QoS constraints.

Thus, the exhaustive search method is feasible. When an AP is ready to transmit, the algorithm is executed as follows:

1) Generate $K_R \in K_{max}$ (associated receivers with packets in their queues).
2) $\forall M \in M_{max}$ and sets of $K \in K_R$ (where $K \leq M$), compute the expected per-user SINR for each $[M,\ K]$ combination as shown in Eq. (3).
3) Using the standard's receiver sensitivity table for 90% packet reception, estimate the MCS for each K in every potential group to generate a list of all possible M, K, MCS combinations.
4) Using Eq. (4) with protocol specific values, calculate the expected aggregate throughput for each M, K, user group dependent MCS combination and choose the largest.

While multiple potential modes ($[M,\ K]$ combinations) could have equivalent expected throughputs if they have identical frame aggregation values or similar link qualities, the probability of this occurring is relatively low and the final selection can be chosen randomly.

III. PUMA WITH 802.11AC

While PUMA's basic mechanism is applicable to any random access MU-MIMO protocol, we demonstrate its functionality with 802.11ac. The two protocol-dependent components of PUMA are datarate inference from expected SINR and potential mode aggregate throughput calculation.

A. Datarate Inference from SINR

After computing the expected SINR from Eq. (3), PUMA employs protocol-specific minimum SNR tables to infer the expected per-user datarate. With this per-user datarate, PUMA computes the expected aggregate throughput of any possible mode and user group.

Table I is 802.11ac's minimum SNR table for ensuring a 90% packet reception rate. For each user and potential mode, PUMA selects the Modulation and Coding Scheme (MCS) index whose corresponding SNR is less than or equal to the expected value calculated using Eq. (3). The corresponding number of data bits per symbol (N_{DBPS}) is the per-user datarate.

Note that the difference between MCS index's SNRs is as large as 4.4 db (MCS 4 vs. 5). Although PUMA's SINR estimation method is inherently less accurate than a post-sounding method since it does not utilize CSI, each MCS's large SNR ranges immensely reduce the resulting effect of this error (see Sec. IV-B).

B. Aggregate Throughput Calculation

Aggregate throughput calculation is dependent on protocol overhead in addition to per-user expected datarate and backlog. Analysis of the 802.11ac specification allows for PUMA to precisely compute the expected aggregate throughput for a potential mode and user group. We express each segment of an 802.11ac transmission as generally described in Eq. (4)–(7). An example of an 802.11ac MU-MIMO transmission is depicted as a timeline shown in Fig. 1. After the expected backoff duration EBO=139.5μs (or 15.5 slots at 9μs/slot) and DIFS=34μs, the AP begins transmitting.

Channel Sounding ($\mathbf{T_S}$). The transmitter first announces to all users in transmission range which specific subset is to expect the upcoming transmission with the Null Data Packet Announcement (NDPA=7.4μs) followed by the Null Data Packet (NDP). The NDP contains the sounding pilots used by the receivers to estimate the channel state between itself and each transmitting antenna (thus it scales with M).

Channel Feedback ($\mathbf{T_{CF}}$). Each receiver must sequentially reply with the Compressed Beamforming Report (CBFR) returning a compressed, per subcarrier version of the sounding pilots to the transmitter in the form of angle pairs (12 or 16 bits each for MU-MIMO, 10 or 12 bits for SU-MIMO). The number of angle pairs scales with the number of transmitting antennas M and the number of receiving antennas for that particular node (the 3×1 vector in the example shown in Fig. 1 requires two angle pairs per subcarrier).

The 80 MHz bandwidth has 234 usable subcarriers. One compressed angle set can be used to indicate groups of 1, 2, or 4 subcarriers. This results in a very high overhead due to channel feedback. The example timeline is scaled to 16 bits of feedback with a subcarrier grouping factor of 2. Each receiver has 3,744 bits to transmit back to the AP at the base rate (MCS 0, see Table I) to ensure that the report is not lost.

While the M=3 antenna transmitter sent the NDP in 7.4μs, each receiver must spend 144μs responding with its CBFR.

Between each CBFR, the AP sends a short polling packet requesting the next user to send its report.

Data Transmission ($\mathbf{T_D}, \mathbf{L_D}$). Finally, the AP forms concurrent data streams at varying frame aggregation and MCS rates. If one stream finishes early, the remaining time until the longest stream completes is wasted. PUMA's throughput formulation shown in Eq. (4) accounts for the potential of unequal stream lengths. Instead of arbitrarily trying to avoid this scenario, PUMA searches for the option with the largest throughput. Because this wasted airtime does have a negative effect on throughput, the probability of such a transmission occurring is low. Nevertheless, PUMA can handle this scenario without explicitly considering it.

Block Acknowledgment ($\mathbf{T_{ACK}}$). Once the data transmission has completed, each receiver must sequentially reply with block acknowledgments (BA=6.4μs). Thus, this component of the overhead scales with K.

C. Example PUMA Transmission

Fig. 1 depicts the timeline of two separate transmissions from a 3-antenna transmitter to either three (Fig. 1(a)) or two single antenna receivers (Fig. 1(b)). Each stream consists of 10 aggregated full size (1,500 byte) packets. As previously discussed, the overhead (channel sounding, channel feedback, and acknowledgment) for both transmissions is sent at the base rate although the resulting data rates are different.

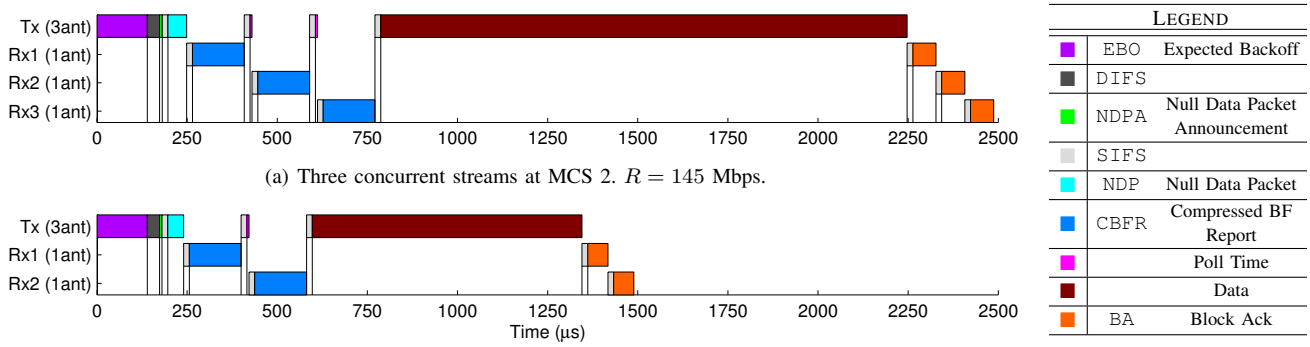
PUMA first calculates the expected datarate of each user. For this example, let each of the three potential users have 18 dB omnidirectional SNRs. Using Eq. (3), PUMA computes expected SINRs of 8.4 and 13.3 dB for users in the [M_3, K_3] and [M_3, K_2] modes respectively. By referencing Table I, PUMA selects MCS 2 for [M_3, K_3] and MCS 4 for [M_3, K_2].

Once the per-user datarates are computed, PUMA computes the aggregate throughput of the potential modes given the expected datarate and node backlog as shown by the to-scale timelines in Fig. 1. Through this computation, PUMA identifies that the aggregate throughput of the [M_3, K_3] transmission is 145 Mbps while the aggregate throughput for the [M_3, K_2] transmission is 161 Mbps. Thus, PUMA selects the [M_3, K_2] transmission mode.

This example highlights a common yet counterintuitive result that PUMA identifies. *A MU-MIMO transmission does not always benefit from using the most antennas.* Not only does the protocol overhead increase with additional antennas, but also the per-user MU-MIMO SINR decreases resulting in lower per-user datarates.

D. Numerical Analysis of Mode Selection

To observe the expected performance and gain an intuition into the effects of M and K, we present a numerical example of PUMA specific to 802.11ac. We consider four separate [M_4, $K_{1:4}$] systems. To evaluate the best case performance for each [M, K] system, all receivers have equivalent omnidirectional SNRs and transmit at the maximum frame aggregation rate (b=64) for maximum overhead amortization.



(a) Three concurrent streams at MCS 2. $R = 145$ Mbps.

(b) Two concurrent streams at MCS 4. $R = 161$ Mbps.

Fig. 1. Example 802.11ac transmission timeline with 3 antenna transmitter sending multiple, 10 aggregated packet streams at 80 MHz (to scale).

For a range of omnidirectional SNRs, we infer the expected per-user datarate given M and K using Eq. (3) and Table I. We then calculate the expected aggregate throughput using Eq. (4) considering 802.11ac-specific overhead and show the performance for each of the four $[M_4, K_{1:4}]$ systems in Fig. 2. The mode that results in the highest aggregate throughput is marked for each omnidirectional SNR.

The key intuition gained from this numerical example is the effect of channel sounding overhead with respect to M and K. While increasing M slightly increases the size of an individual CBFR, additional K increases the number of CBFRs. Thus, to efficiently amortize the overhead induced from increased M and K, the channel state must support high per-user datarates.

The theoretical properties of MU-MIMO system scaling highlighted in Eq. (3) show the effect of higher order modes on per-user SINR. For example, an $[M_4, K_4]$ system results in a per-user SINR approximately 12 dB less than the omnidirectional SNR. Thus, a per-user omnidirectional SNR much higher than 12 dB is required to perform a $[M_4, K_4]$ at a high datarate. The aggregate throughput for a $[M_4, K_4]$ system in Fig. 2 is non-zero at 14 dB, begins to contend with the other modes at 19 dB, and consistently outperforms all other modes starting at 30 dB.

The results for higher order modes in this numerical example further highlight the counterintuitive result shown in Fig. 1: *increasing the number of parallel streams is not always the most efficient transmission mode*. Additionally, because MCS defined datarates are discrete, the aggregate throughput curves for each mode are jagged and result in many intersection points. Thus, each mode does not have a clear SNR range where it exhibits the maximum throughput, necessitating PUMA's dynamic selection method.

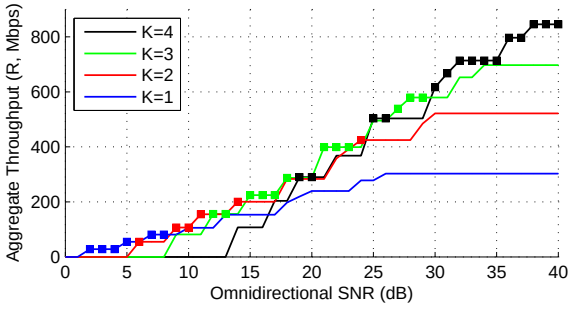


Fig. 2. Theoretical expected throughput (Eq. (4)) of a M_4 antenna transmitter to $K_{1:4}$ parallel single antenna receivers with b=64. For each omnidirectional SNR, maximum expected aggregate throughput represented by square marker.

PUMA's complete analysis of MU-MIMO system scaling and protocol specific overhead allows for the appropriate mode selection decision given the current state of the system. In fact, Fig. 2 graphically represents a PUMA enabled AP's decision engine and the markers represent the decisions themselves: the dynamically calculated, maximum throughput mode.

IV. Experimental Evaluation

A. Experimental Methodology

We first characterize the performance of PUMA through OTA transmissions to verify PUMA's datarate inference method and generate realistic channel traces. We then utilize this realistic OTA data for channel-trace driven emulation.

1) OTA Experimentation: We conduct OTA experiments using the WARP software defined radio [1] utilizing a Zero-Forcing Beamforming framework developed in [4] and expanded in [19]. To perform our experiments, we modify WARPLab, a system that allows for baseband signals to be proccessed in MATLAB, downloaded to the board, and transmitted over-the-air. Because the 802.11ac standard allows up to 8 spatial streams, we connect two 4 antenna WARP boards together to make an appropriate transmitter.

We place 8 receiving antennas in 8 different non-line of sight locations to emulate a typical indoor wireless LAN environment as shown in Fig. 3. We then serve every combination of $[M_{1:8}, K_{1:8}]$ to get a variety of different channel environments for each topology and measure the SINR.

The resulting variability of omnidirectional SNR measurements resulted in an overall mean of 18.3 dB and standard deviation of 5 dB. This allows us to verify the model using a wide range of P/N_o values.

2) Channel-Trace Driven Emulation: We construct a discrete time event emulator in MATLAB to evaluate the efficacy of our mode selection algorithm. For simplicity, we consider a topology wherein an M_4 antenna transmitter serves a subset ($K_{1:4}$) of 8 possible single antenna receivers. We consider $K_{1:4}$ because 802.11ac supports only up to four concurrent receivers (but up to 8 collective receive antennas).

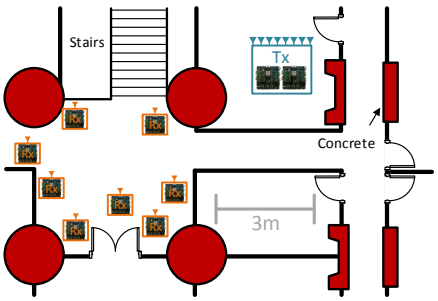


Fig. 3. Experimental topology.

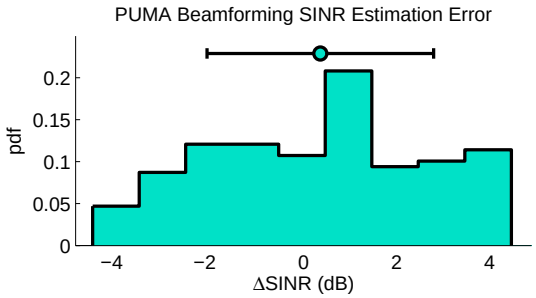


Fig. 4. Measured estimation error of Eq. (3). μ=0.36, σ=2.43 dB.

Because we seek to isolate the effects of mode and user selection, we design an emulation engine that ignores collisions and retransmissions. Transmissions are executed as packets become available on a first-come first-serve basis. At the beginning of every transmission event, each mode and user group is selected based on how many available packets are in each receiver's queue (up to 802.11ac's 64 packet frame aggregation maximum). Packets are modeled to arrive as a Poisson process and the input traffic is defined as aggregate offered load (cumulative generated traffic) to all receivers.

We use our measured, OTA omnidirectional SNR values at each node as a channel trace to determine the expected per-user MCS. The variation in the measured values allows us to consider heterogeneous channels with different omnidirectional SNRs on the AP to client links

Since the 802.11ac standard supports unequal length parallel data streams and unequal per-user datarates, we allow the emulator to transmit parallel payloads with different MCS rates and frame aggregation sizes. Each emulation was run for 100 emulated seconds. The emulation was conducted assuming an 80 MHz bandwidth and 4 μs symbol times. The CBFRs were quantized at 16 bits per subcarrier and a subcarrier grouping of 2. The AP is allowed up to 4 transmit antennas and it serves a group of 8 single antenna receivers (M_{max}=4 and K_{max}=8).

B. Expected SINR Calculation Accuracy

We validate the accuracy of PUMA's SINR estimation method used for datarate inference. Our experiment consists of performing 8 OTA transmissions for all $[M_{1:8},\ K_{1:8}]$ topologies as discussed in Sec. IV-A1.

Using the measured omnidirectional SNR for each receiver, we use Eq. (3) to predict the per-user SINR for each receiver. Perfect SINR results are not expected as Eq. (3) is based on general MU-MIMO system scaling as opposed to CSI. However, the resulting error is almost zero mean and with a standard deviation of 2.43 dB as shown in Fig. 4. This error is tolerable because the standard deviation is approximately equivalent to SINR range for each modulation rate shown in Table I. The use of the MCS table diminishes this error by effectively truncating it. We explore the effect of this error mitigation in Sec. IV-C.

Thus, the expected value equation for per-user SINR is accurate *even without considering measured channel matrices*. This holds true for indoor Wireless LANs because the channel vectors are relatively orthogonal, meaning that the channel matrices used for MU-MIMO transmissions are well conditioned (as experimentally verified in [3]).

C. PUMA Expected Datarate Calculation Accuracy

Using our trace-driven emulation methodology, we compare PUMA against a post-sounding, exhaustive search baseline. This baseline method forgoes the use of the SINR estimation algorithm and employs the actual measured MU-MIMO SINRs. This effectively represents the best case result of using Eq. (1) post-sounding after exhaustively measuring each potential receiver's CSI. Fig. 5 shows the comparative results of our emulation.

A perfect transmitter would send the incoming packets at a rate equivalent to their arrival at the AP. However, because of the overhead time T_{OH} in Eq. (4) required for each packet transmission, this is not possible. Instead, we show that PUMA transmits at the highest feasible portion of that rate by selecting transmission modes that maximize the SINR and thus the MCS while minimizing the transmission overhead. When the aggregate throughput saturates, the maximum possible throughput is achieved.

Observe in Fig. 5 that full knowledge of channel state from an exhaustive sounding process only results in a 3% increase in saturation throughput. Additionally, over all aggregate offered loads, knowledge of full CSI only results in a maximum 7% increase in throughput. Note that the exhaustive search method's performance does not consider the overhead incurred from sounding all potential receivers (like PUMA, it only considers sounding overhead from the users actually

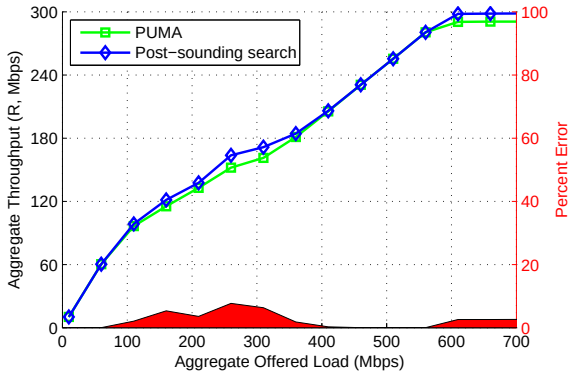


Fig. 5. Comparison between PUMA and post-sounding, exhaustive search.

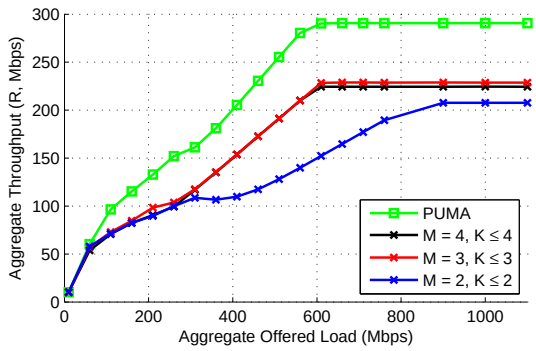


Fig. 6. Comparison between PUMA and fixed modes ($[M_{2:4}, K_{1:4}]$).

served for each transmission). Had the full sounding overhead been considered, the performance of the exhaustive search method would perform significantly worse and its saturation throughput would be far lower.

Fig. 5 also highlights PUMA's datarate inference method's robustness to error. Fig. 4 shows that the variation between the measured and estimated SINR values has a standard deviation of 2.5 dB, approximately equal to the SNR range of each MCS. However, even given this error, we observe that the performance difference between estimating the MU-MIMO SINR and measuring it is minimal. This is a direct result of the effective truncation of the estimated SINR metric, using the minimum SINR table. This truncation essentially smooths the estimated SINR metric and results in similar performance to full knowledge of the MU-MIMO SINR without a post-sounding, exhaustive search.

D. Mode and User Selection Performance

We now compare PUMA to fixed mode selection indicative of the default method used in 802.11ac. Specifically, we compare the performance of PUMA to static $[M_{2:4}, K_{1:4}]$ topologies in Fig. 6 for a range of offered loads. For each possible M, we show the best fixed K value for sake of presentation. These static modes are all potential choices for PUMA's exhaustive search and thus observing how these parts contribute to our collective algorithm illustrates how well PUMA handles varying M, K, b_i, and measured omnidirectional SNR values.

So as not to unfairly disadvantage the fixed modes, we permit the fixed $[M, K]$ topologies to serve any number of users less than or equal to K. For example, given our omnidirectional channel measurements, when M=3, K_{max} is 3 but individual transmissions can be K_2 or K_1, depending on how many users have packets available in their queues.

Although PUMA is inherently a combination of all fixed modes, observe that for any given offered load, no fixed mode performs comparably to PUMA (except at 10 Mbps where all methods are equal). This result suggests that it is not only enough to know the "best" $[M, K]$ combination given some offered load but also it is necessary to dynamically select between potential modes before every transmission, depending on user backlog and omnidirectional SNR.

The relationship between user backlog and omnidirectional SNR is the key interaction the fixed mode topologies fail to consider. Due to the sounding overhead incurred from MU-MIMO transmissions, larger frame aggregation rates are required to properly amortize the cost of employing parallel streams. Thus, users with lower omnidirectional SNRs (resulting in lower achievable datarates) must have more backlogged packets to be efficiently grouped in an MU-MIMO transmission. PUMA considers this interaction and thus does not transmit to low omnidirectional SNR users in higher order modes until the user's backlog is large enough. Thus, the MU-MIMO sounding overhead efficiently and dynamically amortizes the sounding overhead for each transmission.

A concern for PUMA is its potential to unfairly starve users that have consistently less backlogged traffic or poor omnidirectional SNRs. However, unfair scheduling is not a detrimental effect of PUMA, rather, it is a complementary issue. In the most basic sense, PUMA enumerates a list of selections (potential modes and user groups), assigns a metric to each, and selects the best. Our current metric is the aggregate throughput of the system determined by user backlog and expected datarate. To ensure a fair transmission system, an AP can employ any existing proportional fair scheduling algorithm (e.g., [15]) and adjust PUMA's metric accordingly. This would manifest as simply adding a scaling factor generated by a fair scheduling algorithm to step 4 of the PUMA protocol described in Sec. II-D.

Given the measured channels from our experiments, PUMA provides an aggregate saturation throughput increase of approximately 65 Mbps or 30% over the best fixed mode ($[M_3, K_{1:3}]$). The improvement in the saturation region with higher offered loads is of key importance for high congestion scenarios. While the other scenarios all saturate to similar aggregate throughputs (albeit at different rates), PUMA's 30% saturation throughput increase illustrates how efficient overhead amortization through adaptive mode selection based on user link state and backlog allows the same AP to improve the performance of a given topology.

Finally, PUMA's user selection mechanism is a direct result of step 4 of the algorithm discussed in Sec. II-D. Once all potential mode and user groups are enumerated and assigned an aggregate throughput metric, the best mode and user combination is selected. However, the best mode and user combination is not guaranteed to be unique. PUMA's aggregate throughput metric is dependent on per-user omnidirectional SNR and backlog. Thus, multiple mode and user combinations may be assigned the same aggregate throughput metric (e.g., the simplified example shown in Fig. 1). In such cases, a random selection or additional fairness metric can be used.

Nevertheless, real systems rarely have homogenous traffic arrival rates or per-user omnidirectional SNRs. Thus, although PUMA cannot guarantee a unique mode and user combination, it provides a unique selection with high probability.

V. Related Work

Frame Aggregation. Numerous works consider the effects of frame aggregation in MU-MIMO systems. For example, [7] develops a frame aggregation technique and [6] examines the

effects of frame aggregation specifically with 802.11ac. While PUMA selects the most efficient mode based on the number of packets in each receiver's queue, frame aggregation techniques are complementary to PUMA since these techniques can weight or rule out the initial set of potential users.

Implementing a frame aggregation technique with PUMA simply requires the modification of step 1 outlined in Sec. II-D. The selection of K_R from K_{max} can be based on a rule more sophisticated than user packet availability. Additionally, once the list is generated, the values of b_i can be weighted accordingly to implement a frame aggregation protocol.

User Grouping and Selection. Several works focus on user grouping and selection based on channel state and/or unequal transmission length [11], [13], [18], [20]. These works use theoretical expressions similar to Eq. (1) to estimate the aggregate capacities of potential user sets.

The information required to employ these modeling techniques is the channel state. However, obtaining the channel state itself is the overhead that limits the performance of 802.11ac MU-MIMO. Real WLAN environments are too variable to accurately employ stale channel information and the overhead required for channel feedback limits the number of users that can be sounded at a time.

While these algorithms are potentially more accurate than PUMA due to the additional information they require, procuring this information given protocol overhead render them impractical to deploy. Also, due to the volatile nature of indoor WLANs, even once that information is collected, it has a high probability of being outdated further reducing its accuracy. PUMA balances the tradeoff between obtaining the necessary information and the time taken to obtain that information, and thus is more plausible for a true 802.11ac WLAN deployment.

The authors of [22] propose a user selection method that requires substantial modification of 802.11ac's CSI collection exchange to implement a distributed user group probing method. While this method is an improvement over the aforementioned works with respect to sounding overhead, it is not 802.11ac compliant. PUMA, however, is 802.11ac compliant since the additions do not conflict with any standard mandates, allowing for full interoperability with any 802.11ac device.

Mode Comparisons. Survey works exist that compare different MIMO modes such as MU-MIMO and frame aggregation [9] or MU-MIMO and multiple SU-MIMO [21]. Both works highlight the tradeoffs between these schemes but do not provide algorithms for exploiting those tradeoffs such as PUMA. Additionally, neither work verifies these differences with measured over-the-air transmissions.

VI. Conclusion

We present PUMA, a mode and user selection algorithm that allows an MU-MIMO system to efficiently transmit multiple streams by using only pre-sounding information. PUMA estimates the aggregate throughput of all potential mode and user group combinations without knowledge of the channel state. First, PUMA infers the per-user datarate of each user in a potential mode by exploiting theoretical MU-MIMO system scaling properties. PUMA then computes the aggregate throughput of each potential mode and user group combination by considering protocol specific overhead. A PUMA enabled AP selects the best mode and user group based only on information available before channel sounding begins. Given the large amount of overhead required by an 802.11ac transmission, appropriate mode and user selection is the key enabler to reaching gigabit wireless speeds.

References

[1] Rice University WARP project. Available at: http://warp.rice.edu.

[2] F. Adib, S. Kumar, O. Aryan, S. Gollakota, and D. Katabi. Interference alignment by motion. In *Proc. ACM MobiCom*, Miami, FL, Sept. 2013.

[3] N. Anand, S.-J. Lee, and E. Knightly. STROBE: Actively Securing Wireless Communications using Zero-Forcing Beamforming. In *Proc. IEEE INFOCOM*, Orlando, FL, Mar. 2012.

[4] E. Aryafar, N. Anand, T. Salonidis, and E. Knightly. Design and experimental evaluation of multi-user beamforming in Wireless LANs. In *Proc. ACM MobiCom*, Chicago, Illinois, Sept. 2010.

[5] O. Bejarano, M. Park, and E. Knightly. IEEE 802.11ac: From Channelization to Multi-User MIMO. *IEEE Communications Magazine*, 2013.

[6] B. Bellalta, J. Barcelo, D. Staehle, A. Vinel, and M. Oliver. On the Performance of Packet Aggregation in IEEE 802.11ac MU-MIMO WLANs. *IEEE Communications Letters*, 2012.

[7] B. Bellalta and M. Oliver. A space-time batch-service queueing model for MU-MIMO communication systems. In *Proc. ACM MSWiM*, Oct. 2009.

[8] J. Camp and E. Knightly. Modulation rate adaptation in urban and vehicular environments: Cross-layer implementation and experimental evaluation. In *Proc. ACM MobiCom*, San Francisco, CA, Sept. 2008.

[9] J. Cha, H. Jin, B. C. Jung, and D. K. Sung. Performance comparison of downlink user multiplexing schemes in IEEE 802.11ac: MU-MIMO vs. frame aggregation. In *Proc. IEEE WCNC*, Apr. 2012.

[10] A. Edelmen. *Eigenvalues and Condition Numbers of Random Matrices*. PhD thesis, M.I.T., May 1989.

[11] M. Esslaoui, F. Riera-Palou, and G. Femenias. A fair MU-MIMO scheme for IEEE 802.11ac. In *Proc. IEEE ISWCS*, Aug. 2012.

[12] IEEE P802.11ac/D5.0. Specification framework for TGac, Dec. 2013.

[13] Y. Jang and H. K. Y. Lee. Adaptive mode selection for multiuser MIMO downlink systems. In *Proc. IEEE VTC*, Montreal, Canada, Sept. 2006.

[14] F. Kaltenberger, M. Kountouris, D. Gesbert, and R. Knopp. On the trade-off between feedback and capacity in measured MU-MIMO channels. *IEEE Transactions on Communications*, 2009.

[15] H. Kim, K. Kim, Y. Han, and S. Yun. Proportional fair scheduling for multicarrier transmission systems. In *Proc. IEEE VTC*, Sept. 2004.

[16] M. Lopez. *Multiplexing, Scheduling, and Multicasting Strategies for Antenna Arrays in Wireless Networks*. PhD thesis, M.I.T., Aug. 2002.

[17] M. Sharif and B. Hassibi. A Comparison of Time-Sharing, DPC, and Beamforming for MIMO Broadcast Channels With Many Users. *IEEE Transactions on Communications*, 2007.

[18] Z. Shen, R. Chen, J. Andrews, R. Heath, and B. Evans. Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization. In *Proc. Asilomar*, Pacific Grove, CA, Oct. 2005.

[19] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong. Argos: Practical many-antenna base stations. In *Proc. ACM MobiCom*, Istanbul, Turkey, Aug. 2012.

[20] T. Tandai, H. Mori, and M. Takagi. Cross-Layer-Optimized User Grouping Strategy in Downlink MU-MIMO Systems. In *Proc. IEEE VTC*, Apr. 2009.

[21] A. Thapa and S. Shin. A MAC protocol to select optimal transmission mode in very high throughput WLAN: MU-MIMO vs. multiple SU-MIMO. In *Proc. IEEE AH-ICI*, Nov. 2012.

[22] X. Xie and X. Zhang. Scalable user selection for MU-MIMO networks. In *Proc. IEEE INFOCOM*, Toronto, Canada, Apr. 2014.

[23] X. Xie, X. Zhang, and K. Sundaresan. Adaptive feedback compression for MIMO networks. In *Proc. ACM MobiCom*, Miami, FL, Sept. 2013.

[24] H. Yang and T. Marzetta. Performance of Conjugate and Zero-Forcing Beamforming in Large-Scale Antenna Systems. *IEEE JSAC*, 2013.

[25] T. Yoo and A. Goldsmith. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE JSAC*, 2006.