Abstract—Multi-User (MU) MIMO schemes introduced to IEEE 802.11ac amendments can achieve high throughput by transmitting different data streams to multiple users simultaneously. However, it requires the user Channel State Information (CSI), which is not trivial to obtain in practice. Alternatively, the access point can balance the expected rate with the CSI processing overhead by choosing an appropriate number of transmitting and receiving antennas, as well as the number of users according to client channel and backlog. We propose the application of a user selection algorithm based on MU-MIMO theoretical properties and user omnidirectional SNR named PUMA (Pre sounding User and Mode selection Algorithm) and its application on a diversity scenario. We compare the data rate of the proposed scheme with traditional fixed-mode and exhaustive search user selection, and the equivalent effects on energy efficiency. Results show that the PUMA advantages are enhanced in a diversity scenario when client grouping causes dramatic changes in the sum rate. Besides, by using different diversity schemes on the same network configuration, we obtained different rates and consumption.

I. INTRODUCTION

The introduction of MU-MIMO (Multi-User Multiple Input Multiple Output) schemes to IEEE 802.11ac amendment [1] is the answer for the demand for new technologies that improve the capacity and spectral efficiency on the Industrial, Scientific, and Medical (ISM) band for internet access. An Access Point (AP) equipped with such technique can transmit different data streams to multiple users simultaneously, obtaining an extensive capacity gain.

The standardization of MU and Single-User (SU) beamforming techniques have started in the 802.11n release. In that release the method of acquiring the user Channel State Information (CSI), which is required for MU-MIMO, was left open for industry development, causing non-interoperable implementation throughout different suppliers. In the 802.11ac amendment, the feedback format and frames are standardized, leaving only the method of applying the beamforming weights open to developers [1]. Additionally, the 802.11ac amendment allows user selection algorithms to be integrated with the sounding process to maximize the MU-MIMO gains, by balancing the overhead of client channel matrices calculation for a selected receiving user set.

The spectral efficiency achieved by MU-MIMO comes with trade-offs that must be taken into account. In [2], the relation between feedback bits, the number of users, and the Signal-to-Noise Ratio (SNR) is explored. It was shown that spatial separations of users and the number of quantization bits can increase the throughput. As the number of selected users increases, the feedback size expands proportionally, which may amortize the multiplexing gains with the sounding overhead.

The optimal feedback length from client to AP is a relevant research topic for its impacts on the network sum-rate. In an optimal case, all the users are sounded, and a complex matrix combination algorithm selects the best grouping according to the instantaneous user CSI. This operation presents high computation complexity and overhead, resulting in much more resource-intensive and costly APs, which are undesirable for the consumer market.

For this reason, reduced complexity techniques to increase network throughput are crucial. One of these techniques is user selection, which relies on previous transmission statistics to select users that can achieve the best rate, without recurring to instantaneous CSI. One of such techniques is the Pre sounding User and Mode selection Algorithm (PUMA) [3], which requires the client backlog (number of packets to transmit) and omnidirectional SNR to estimate the best rate, which is the balance of payload and overhead size, and the time needed to send it. PUMA algorithm is based on a search of the optimal user combination that gives the highest rate in the defined search space. After finding the optimal user set and backlog, the selected users are sounded via the 802.11 process and the precoding coefficients are estimated according to the channel matrix.

Besides PUMA, other techniques have been developed. In [4], a technique to lessen the overhead effects is proposed. The CUIC (Concurrent Uplink Control Messages) process uses equalization techniques such as Zero-Forcing (ZF) or Minimum-Mean Square Error (MMSE) at the AP, to receive client CSI in parallel streams and decode them correctly. However, this technique includes an extra step that brings incompatibility among 802.11ac standard devices. In [5], the authors present IMMULA (Intelligent MU-MIMO User selection with Link Adaptation), a dynamic user grouping selection that presents promising results compared to other techniques discussed in [6].

The previously mentioned techniques use multiple antennas to transmit several parallel streams, but the receivers are equipped with a single antenna. Nevertheless, receive diversity employing multiple antennas can also be effective in increasing energy efficiency and system reliability, through SNR improvement. As so, the AP creates collective streams...
at the transmitter that are received through several paths at the users. At this point, the receiver can use techniques such as Maximal Ratio Combining (MRC) or Selection Combining (SC) [7] to improve the SNR.

In this work, we investigate PUMA algorithm maximum throughput performance and energy efficiency on MU-MIMO 802.11ac, but considering that users have multiple antennas. PUMA was chosen because it is the simplest method among the previously mentioned, only needing readily available information (omnidirectional SNR) and a simple exhaustive search with a small search space. The obtained results are compared with a fixed mode scheme, where no user selection algorithm is applied, with the exhaustive search, which incurs in sounding all users and with the original PUMA where each receiver has only a single antenna [3]. Numerical results show that receive diversity can further extend the throughput and energy efficiency gains in MU-MIMO technology. Moreover, PUMA selection algorithm can experience up to 30% of throughput improvement while the fixed mode can only achieve 26% when users are equipped with two antennas, which is a proper improvement while the fixed mode can only achieve 30% of throughput improvement while the fixed mode can only achieve 26% when users are equipped with two antennas, which is a proper balance between hardware complexity and diversity gains.

The subsequent sections are organized as follows. In Sec. II, the system model, the original PUMA, and the 802.11ac consumption model are presented. The numerical analysis is outlined in Sec. III and Sec. IV concludes the paper.

II. BACKGROUND

A. Notation

Bold uppercase and lowercase letters are used for matrices and vectors, respectively, $H^\dagger$ is the Moore-Penrose pseudo-inverse. $E \{ \cdot \}$ represents the mathematical expectation and $h^\ast(H^\dagger)$ is the conjugate transpose of a vector (matrix). Vectors are represented as columns throughout the paper.

B. System Model

![Fig. 1. System topology with 802.11 AP and $K_{\text{max}}$ users.](image)

We consider an indoor Wireless LAN (WLAN) MU-MIMO system as depicted in Figure 1, where the AP has $M_{\text{max}}$ antennas, and $K_{\text{max}}$ associated stations, each with $N$ antennas. The AP transmits with $M \leq M_{\text{max}}$ antennas and can create $K \leq M$ parallel streams to transmit to a subset of users $S$, with $|S| = K$. The clients can be single or multiple antenna devices, and when having more than one antenna, MRC or SC is implemented [7]. Then, the received signal at the user $U_k$ is given by [8]

$$y_k = \sqrt{P_k} h_k x + n_k,$$

where $P_k$ represents the path-loss and shadowing effect, $x \in \mathbb{C}^{M \times 1}$ and $y_k \in \mathbb{C}^{N \times 1}$ are the transmitted and received signals. The user channel $h \in \mathbb{C}^{1 \times M}$ is composed of the channel fading coefficients and the system matrix $H \in \mathbb{C}^{K \times M}$ is composed by the $h_k$ channel vectors, where $k \in S$. Each fading element is modeled according to a Rayleigh random variable $CN(0, 1)$ [7] that remains constant between the sounding process and frame transmission time, however varies between transmissions. The additive white Gaussian noise is represented by $n_k \in \mathbb{C}^{N \times 1}$, with variance $N_0/2$ per dimension, where $N_0$ is the noise unilateral power spectral density [7]. The signal vector is restricted to a transmit power $P \leq E \{ x^\ast x \}$.

Under these circumstances, the AP must coordinate the user and antenna selection to achieve the best throughput and energy efficiency. This can be done by transmitting to the selected clients simultaneously using a beamforming technique. This method comprises the use of precoding coefficients for transmission, which can eliminate mutual interference among the users. A multiple signal vector that can support $K$ streams can be represented as a linear combination of the user data streams and beamforming weights as [8]

$$x = \sum_{k \in S} w_k \sqrt{p_k} s_k,$$

where $w_k \in \mathbb{C}^{M \times 1}$ is the user beamforming vector, $p_k$ is a transmit power scaling factor, and $s_k$ represents the information symbol of user $k$.

A common beamforming technique is ZF [9], which inverts the channel matrix to create orthogonal channels among the clients. Thus, the beamforming matrix is obtained as $W = H^\dagger$, where $W = [w_1, w_2, \ldots, w_k]$. Rewriting (1) with the precoding weights we have

$$y_k = (h_k w_k \sqrt{p_k}) s_k + \sum_{i \in S, i \neq k} (h_i w_i \sqrt{p_i}) s_i + n_k.$$  (3)

ZF beamforming removes completely the interference among clients, therefore for $i \neq k$, $h_i w_k$ is negligible even considering quantization errors [8]. Thus, user signal multiplication factor $h_i w_k \sqrt{p_k}$ in (3) is the effective channel gain.

C. PUMA

To compute the precoding coefficients $w_k$, CSI between all transmitting/receiving channel paths must be obtained. This process imposes a large overhead to each transmission round and is described in the 802.11 standard as the sounding process. The sounding process, as illustrated by Figure 2 [3], starts with the AP sending a Null Data Packet Announcement (NDPA) containing a training waveform and the users selected for transmission. The Null Data Packet (NDP) phase contains the sounding pilots for each user to estimate its channel. Then, the previously selected users send back the Compressed Beamforming Report (CBFR), which is the $h_k$ user channel vector quantized and grouped over multiple subcarriers [6]. CBFR size is proportional to the number of antennas involved in the transmission, i.e., $[M, K]$. Therefore having more parallel streams increases the CBFR size, but more data can be sent in...
parallel to the users. After the sounding process, the AP sends the multiple user data stream and waits for each independent Block Ack (BA) in the uplink.

\[
\begin{array}{|c|c|c|}
\hline
\text{AP} & \text{NDPA} & \text{NDP} \\
\hline
\text{U}_1 & \text{CBFR} & \text{BA} \\
\hline
\text{U}_2 & \text{CBFR} & \text{BA} \\
\hline
\end{array}
\]

Fig. 2. Multi-User sounding procedure for a 2 user case.

Since there is an overhead that depends on the number of antennas, in some cases transmitting with all available antennas may not be the optimal solution. In addition, considering equal power allocation for each user, we can represent the SNR for user \( k \) as

\[
\text{SNR} = 10 \log_{10} \left( \frac{(P \cdot P_k)/(N_0 B)}{M} (h_k w_k) \right),
\]

where \( P \) is the AP transmit power and \( B \) is the channel bandwidth. Therefore, as the AP creates more parallel streams, the number of transmit antennas \( (M) \) increases, and user SNR decreases, as the transmit power is divided among the transmit antennas according to (4).

The 802.11ac standard presents the relationship between user SNR and the appropriate Modulation and Coding Scheme (MCS) for 10% Packet Error Rate (PER). Table I shows part of this relation, while the complete information can be found in [1]. The MCS scheme is applied only in the data transmission phase, while packet overhead is sent at MCS0 to improve reception reliability. It can be seen from Table I that higher SNRs support high throughput modes, increasing energy efficiency and network rate. Then, PUMA’s goal is to choose a user/antenna set that yields the maximum throughput, taking the sounding overhead and the estimated SNR of individual users into account. On the other hand, computing the SNR using (4) requires all potential users’ CSI, as the \( w_k \) vectors are calculated from the channel matrix, imposing even higher overhead.

**TABLE I**

<table>
<thead>
<tr>
<th>MCS</th>
<th>Rate</th>
<th>SNR (dB)</th>
<th>MCS</th>
<th>Rate</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BPSK1/2</td>
<td>1.1</td>
<td>5</td>
<td>64-QAM2/3</td>
<td>17.2</td>
</tr>
<tr>
<td>1</td>
<td>QPSK1/2</td>
<td>4.1</td>
<td>6</td>
<td>64-QAM3/4</td>
<td>18.4</td>
</tr>
<tr>
<td>2</td>
<td>QPSK1/4</td>
<td>6.7</td>
<td>7</td>
<td>64-QAM5/6</td>
<td>19.7</td>
</tr>
<tr>
<td>3</td>
<td>16-QAM1/2</td>
<td>9.6</td>
<td>8</td>
<td>256-QAM3/4</td>
<td>23.9</td>
</tr>
<tr>
<td>4</td>
<td>16-QAM1/4</td>
<td>12.8</td>
<td>9</td>
<td>256-QAM5/6</td>
<td>25.5</td>
</tr>
</tbody>
</table>

From [3] it is known that the terms \( (h_k w_k) \) follow an Erlang distribution for Rayleigh fading, as well as they depend on the size of channel matrix \( \mathbf{H} \) having a mean of \( (M - K + 1)/K \). Thus, combining it with (4) we obtain the PUMA SNR estimation rule

\[
E \{ \text{SNR} \} = 10 \log_{10} \left( \frac{(M - K + 1) (P \cdot P_k)/(N_0 B)}{K M} \right),
\]

from which PUMA seeks the best throughput

\[
R = L_D / (T_{OH} + T_D).
\]

To estimate \( R \) in (6), PUMA only needs accessible information on the AP, that is overhead time \( (T_{OH}) \) for a given mode \( [M, K] \), transmitted data size \( (L_D) \) and omnidirectional SNR. Then, PUMA evaluates \( R \) for all possible combinations among \( K \) users, backlog \( b_k \) and mode \( [M, K] \), where the data size for a given group is

\[
L_D = \sum_{k \in S} b_k \cdot L_p,
\]

with \( L_p \) as the maximum packet size and \( b_k \) is defined as the minimum between the user packet queue and frame aggregation rate, as users may have more packets to transmit than a frame can aggregate.

Moreover, the overhead time \( (T_{OH}) \) in (6) is defined as

\[
T_{OH} = T_S + T_{CF} + T_{ACK},
\]

where \( T_i \in \{S, CF, ACK\} \) are the sounding, beamforming feedback and receiver ACK transmission times, respectively. Finally, the data transmission time is defined as

\[
T_D = \max(h_k \cdot L_p)/r_k,
\]

where \( r_k \) is the \( k \)th user’s MCS data rate, as shown in Table I, which is obtained with the estimated SNR in (5). In this sense, PUMA’s most contribution is to find an estimate for (6) before the sounding process, which is mainly a modification from (4) to (5).

It can be seen from (6)-(9) that user combination may cause great changes in the network rate. Selecting users with small backlog \( b_k \) may increase \( T_{OH} \) without sufficiently increasing \( L_D \). In addition, (9) depends only on the user that is slower in transmitting its backlog queue. In other words, a single user with low instantaneous channel gain can affect the whole transmission round throughput. For this reason, receive diversity techniques can positively impact the overall data rate of the MU-MIMO transmissions when users with multiple antennas are combined with single-antenna users with better channel gain at transmission time [7].

In this paper, two diversity techniques are implemented: i.) Selection Combining (SC), in which the user selects only the antenna experiencing the highest instantaneous SNR for each frame, and ii.) Maximal Ratio Combining (MRC), where the resulting signal vector is a weighted sum of the signals received from all antennas. The advantage of SC lies in requiring just one RF receive chain that is switched into the active antenna branch, while the advantage of MRC is that it can achieve an equivalent SNR that is the sum of the SNR seen at each receiving antenna. However, despite MRC advantages in terms of diversity, increasing the number of antennas also increases the number of RF chains, consequently increasing the energy consumption [10], [11]. For this reason, the most energy-efficient transmission mode is not always the one that presents the best SNR or throughput.

**D. Energy Consumption Model**

As shown in [11], an important amount of energy is spent by the RF circuits during frame transmission. Thus, the total power consumption does not only depends on the transmit
power $P$, but also on the other electronic blocks to transmit and receive. In this work we employ the model from [11], so that the power consumption during an AP or a user transmission are, respectively,

$$\mathcal{P}_A (M, K, S) = (1 + \delta) P + M P_{TX} + \sum_{k \in S} N k \cdot P_{RX}, \quad (10)$$

$$\mathcal{P}_U = (1 + \delta) P + P_{TX} + M_{max} P_{RX}, \quad (11)$$

where $\delta$ represents the additional energy spent at the power amplifier due to nonlinearities, $P_{TX}$ and $P_{RX}$ represent the RF chain power consumption at the transmitter and receiver, respectively. As the AP chooses higher transmission modes $[M, K]$, the number of RF chains increases proportionally, implying more energy consumption. In this model, clients send data back to AP with only one antenna, and the AP receives data in the uplink with $M_{max}$ antennas.

An 802.11ac frame is composed of several phases, each one the user or the AP transmits its information. Then, we define the frame energy consumption in two main parts, corresponding to upstream and downstream energies, and each of these is a sum of the phases each device is transmitting:

$$E_{frame} = \sum_a \mathcal{P}_A \cdot t_a + \sum_b \mathcal{P}_U \cdot t_b, \quad (12)$$

where $a \in \{\text{NDPA, NDP, Data}\}$ and $b \in \{\text{CBFR, BA}\}$, are the AP and user transmission phases, respectively, as can be seen in Figure 2. The $t_a$ and $t_b$ phase times are estimated with the information in [11], considering the overhead phases are sent at MCS0, and Data phase is sent with the estimated MCS.

Finally, energy efficiency can be defined as

$$EE = b_s/E_{frame}, \quad (13)$$

where $b_s$ represents the number of bytes successfully received.

### III. NUMERICAL ANALYSIS

For the simulation results, we consider a WLAN as depicted in Figure 1 with $K_{max}=8$ users as in [3] and an AP with $M_{max}=4$ antennas. For each mode/scheme, the number of users with two antennas changes from 0 to $K$, implying $K+1$ simulations. Two main modes are considered, denoted by Fixed Mode and PUMA Mode, where Fixed Mode selects the transmission group randomly at the beginning of each frame transmission (only nodes with packets to transmit are selected), while the user selection strategy depicted in II-C is used for the PUMA Mode. For exhaustive search evaluation, only the selected transmission mode sounding overhead is considered. When the full-sounding process is considered, the exhaustive search performance is much lower.

The simulations are based on the offered load, aggregate/saturation throughput and energy efficiency metrics. The offered load is defined as the packet generation rate in bits per second, which is considered Poisson distributed in each node. The aggregate throughput represents the data delivery rate in bits per second from the AP to all users, whose value is obtained in (6) for each transmission. The saturation throughput is the data rate where the aggregate throughput no longer increases when the offered load is increased, i.e., is the maximum possible throughput achieved by the network. Finally, the energy-efficiency is given by (13), where all the values consider the mean of 300 simulation iterations.

The transmit power is $P=100 \text{ mW}$ [1], the carrier frequency is 5.1 GHz and the path loss exponent is 2.5 [7]. Each user is positioned at a distance $d_k=10 \text{ m}$, while the RF chain consumption values for transmission and reception are used according to [11]. Moreover, shadow fading is simulated with a standard deviation $\sigma_{dB}=5 \text{ dB}$. The channel fading coefficients are assumed to change at the beginning of each frame, and shadowing changes every 100 ms according to [12]. Transmissions may aggregate up to 64 packets, each one with a length $L_p=1500$ bytes. For the PUMA simulation, user selection runs at each 100 ms with SNR data gathered from the last 5000 packets. Packet loss is modeled using the BER curves for 802.11ac obtained in [13] for the different MCS.

Figure 3 shows the aggregate throughput as a function of the offered load for PUMA and Fixed Mode. We can observe that PUMA achieves a higher saturation throughput when compared to the Fixed Mode with the same number of parallel streams. Also, when all users are equipped with multiple antennas using SC, the saturation throughput of the PUMA Mode is much more positively influenced by the receive diversity than the Fixed Mode. In this particular example, 30% of improvement is obtained for the PUMA Mode, while 26% of improvement is obtained for the Fixed Mode. Exhaustive search mode achieves a higher saturation throughput, as it can predict user channel deep fades.

This result can be explained by the fact that the Fixed Mode has a higher probability of selecting a low data rate user for a transmission round, which increases $T_D$ in (9) and reduces the network throughput. For this reason, an in-depth investigation of the impact of the use of multiple antenna users for the different modes and diversity is presented in Figure 4. The saturation throughput is evaluated for both SC and MRC, and Fixed Mode is simulated for each possible maximum number of transmission streams ranging from 2 to $M_{max}$.

![Fig. 3. Exhaustive, PUMA and Fixed Modes with and without diversity.](image)
inherently better SNR because of the receive diversity. In addition, employing MRC instead of SC incurs in increased saturated throughput, which may increase energy consumption. This is because MRC uses all RF receive chains during the whole transmission. Finally, it can be seen that the saturation throughput does not significantly increase for more than 4 multiple antenna users, which also motivates the investigation on the energy efficiency, since less multiple antenna users may result in lower overall energy consumption.

The energy efficiency of the PUMA Mode using both SC and MRC is presented in Figure 5. As we can observe, MRC achieves the best results, even using more receiving chains. This behavior can be explained by the fact that MRC receivers have a higher average SNR, which allows them to use higher-order modulations, consuming less energy due to the reduced time on air. Furthermore, another observation is that increasing the number of users with multiple antennas does not increase the energy efficiency in the saturation regime. On the other hand, using the mode with more diversity paths can be beneficial in the lower traffic regime. For example, for an offered load of 300 Mbps, a network with 8 multiple antenna users can be 36% more energy efficient than a network with only 7 multiple antenna users.

![Fig. 4. PUMA and Fixed Mode saturation throughput when increasing users with diversity.](image)

![Fig. 5. PUMA Mode Energy Efficiency comparing the different number of users with SC and MRC diversity.](image)

IV. CONCLUSIONS

In this paper, we evaluated the PUMA mode selection algorithm in an 802.11ac WLAN system with receive diversity. Our analysis shows that a simple mode selection algorithm such as PUMA can exploit receive diversity fully. Moreover, we show that the Fixed Mode performs poorly than PUMA in a diversity scenario. The energy efficiency analysis shows an interesting view of the diversity schemes under MU-MIMO: although MRC power consumption is greater than that of SC, MRC receivers have a higher average SNR, which allows them to use higher-order modulations, being more energy efficient in general. Furthermore, concerning the number of multiple antenna users, our analysis shows that the best option depends on the average offered load of the system. Lightly loaded systems will benefit from multiple diversity enabled receivers, while heavily saturated systems will achieve the same energy efficiency with less diversity.

REFERENCES


