RICE UNIVERSITY

Protocol Design and Experimental Evaluation for Efficient Multi-User MIMO Wireless Networks

by

Oscar Bejarano

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Edward W. Knightly, Chair
Professor of Electrical and Computer Engineering

Behnaam Aazhang
J.S. Abercrombie Professor of Electrical and Computer Engineering

Ashutosh Sabharwal
Professor of Electrical and Computer Engineering

T. S. Eugene Ng
Associate Professor of Computer Science and Electrical and Computer Engineering

Houston, Texas
April, 2015
ABSTRACT

Protocol Design and Experimental Evaluation for Efficient Multi-User MIMO Wireless Networks

by

Oscar Bejarano

Information theoretic results on Multi-User MIMO (MU-MIMO) have demonstrated a many-fold increase in capacity compared to Single-Input Single-Output. By leveraging multiple antennas at the Access Point (AP) and beamforming techniques, MU-MIMO enables simultaneous transmissions of multiple independent streams on the downlink. Ideally, with sufficient antennas at the AP, MU-MIMO can attain capacity gains proportional to the number of streams. However, the cost required to enable efficient and robust multi-stream transmissions is much higher than that for the single-stream case and worsens with increasing number of streams. More specifically, two key factors hinder the potential gains that can be attained via MU-MIMO: (i) To serve multiple users simultaneously, the AP needs to collect Channel State Information (CSI) from all users to be served (i.e., sounding). Sounding overhead reduces the effective data airtime utilization of the overall system. (ii) Multi-stream transmissions are highly susceptible to inter-stream interference originated due to inaccurate or outdated CSI, thereby reducing packet reception performance. We demonstrate that in practice, the costs of MU-MIMO not only decrease the gains demonstrated by theory but can completely outweigh the benefits. We identify those adverse situations and propose several techniques that alleviate the negative impact caused by sounding overhead and CSI inaccuracies.
First, we design CUiC and MUTE, two protocols that address MU-MIMO sounding overhead by performing overhead compression along spatial and temporal domains, respectively. CUiC exploits the available Degrees-of-Freedom (DoF) at the AP to allow multiple users to reply with their control messages (e.g., channel estimates and acknowledgements) simultaneously, therefore reducing the time required for users to reply, to a constant. MUTE exploits epochs characterized by slowly moving channels to reduce the frequency of channel sounding. Second, we design CHRoME, a protocol that addresses interference-leakage caused by outdated and inaccurate CSI as well as out-of-cell interference. CHRoME proposes a bit rate selection strategy that re-tunes the selection according to current channel and interference conditions. Additionally, if necessary, CHRoME realizes a fast soundless retransmission that exploits liberated DoF at the AP to minimize retransmission overhead. We implement and evaluate all three schemes using a combination of WARP FPGA-based transceivers, and custom emulation platforms.
Acknowledgments

I would like to begin by expressing my deepest gratitude to Professor Edward Knightly for being not only an excellent academic advisor but also an exceptional career mentor. Thank you for your encouragement and support throughout these many years, and for allowing me to grow both as an individual and as a researcher. I would also like to thank the members of my thesis committee, Dr. Ashutosh Sabharwal, Dr. Behnaam Aazhang, and Dr. Eugene Ng. This work would not have been possible without their invaluable guidance and support. Ashu and Behnaam, thank you for your patience and encouragement. Both of you have played an important role in my personal and professional growth.

I am very fortunate to have been surrounded by brilliant friends/fellow Ph.D. students. Thank you Pedro and Achal for your extraordinary help and patience. I would have never finished (or probably even actually started) without your help. Ryan, Naren, and Clay, I have looked up to you since I got to Rice. You better hire me before any of your companies gets acquired by Google. David, Adriana, and Pablo, thank you for all your help, advice, and friendship. I also want to thank Samantha for believing in me. She might not know this, but a few words from her convinced me to continue after my Master’s degree. I learned so much from my friends and collaborators Eugenio Magistretti, Omer Gurewitz, and Sadia Quadri. Eugenio and Omer, thank you for being exceptional mentors for many years, and for teaching me how to be rigorous and how to do things the right way.

In addition to my friends at Rice, I am deeply indebted to so many other people. Alondra, Ceci, and Luisa, thank you for your unconditional support and for helping me get through difficult times. Andrea, if anyone deserves thanks, it’s you. Thank you for believing in me and for being so patient, supportive, and encouraging. Lastly, I wish to thank my entire family for their unwavering support and faith in me. All that I am, I owe it to them.
To my parents, my sisters, and my brother.
# Contents

Abstract ii  
Acknowledgments iv  
List of Illustrations ix  
List of Tables xi

1 Introduction 1

2 Literature Review 5  
  2.1 MU-MIMO Control Messages Overhead Reduction 5  
  2.2 MU-MIMO Resilience to Mobility, Interference, and Imperfect Channel State Information 8

3 Background on Multi-User MIMO and the IEEE 802.11ac Amendment 10  
  3.1 Multi-User MIMO Primer 10  
    3.1.1 Multi-User Beamforming 11  
    3.1.2 Zero-Forcing Beamforming 12  
    3.1.3 Experimental Platform - WARP and WARPLab [1] 13  
    3.1.4 Performance of ZFBF 14  
  3.2 Sounding and Acknowledgement Processes in MU-MIMO 16

4 Concurrent Uplink Control Messages 26  
  4.1 Introduction 26  
  4.2 Background on Receiver-Side Beamforming 29
4.3 Concurrent Uplink Control Messages ........................................ 31
  4.3.1 Enabling Spatially Multiplexed Feedback ............................... 32
  4.3.2 Variable Length Reports, Missing Uplink Messages, and
       Retransmissions ............................................................... 37
  4.3.3 User Selection and Retransmissions ................................... 38
  4.3.4 Overhead Analysis ....................................................... 39
4.4 Experimental Evaluation of CUiC ............................................. 41
  4.4.1 Decoding Reliability of CUiC .......................................... 42
  4.4.2 Improving Sounding Resilience in Dynamic Scenarios .............. 49
  4.4.3 Achieving Constant Sounding Overhead with CUiC ................... 50
  4.4.4 Efficient Acknowledgement Process With CUiC ...................... 51
  4.4.5 Throughput of CUiC Policies .......................................... 53
4.5 Chapter Concluding Remarks .................................................. 54

5 Feasibility of Sounding Suppression ......................................... 55
  5.1 Introduction ........................................................................... 55
  5.2 MUTE .................................................................................. 58
    5.2.1 Decoupling Sounding & User Selection Procedures ................. 58
    5.2.2 Sounding Set Selection Procedure .................................... 60
    5.2.3 Opportunistic Transmission Set Selection Procedure ............... 66
    5.2.4 Packet Loss & Demotion to SU-MISO for Fast Channels .......... 66
  5.3 Implementing and Evaluating MUTE ......................................... 67
    5.3.1 Experimental Setup and Evaluation Methodology .................... 68
    5.3.2 Channel Estimation Accuracy .......................................... 73
    5.3.3 Throughput Gains in MUTE ............................................. 76
    5.3.4 Discussion: A First Look at MUTE in an OFDM System .......... 80
  5.4 Chapter Concluding Remarks .................................................. 82
6 Interference Resilient Multi-User Beamforming WLANs 83

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>83</td>
</tr>
<tr>
<td>6.2</td>
<td>M$^3$CS: Beamformed Probing for “Just-in-Time” MCS Selection</td>
<td>86</td>
</tr>
<tr>
<td>6.2.1</td>
<td>CSIT-Based MCS Selection</td>
<td>87</td>
</tr>
<tr>
<td>6.2.2</td>
<td>MCS Selection via MUBF Probing</td>
<td>89</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Incurred Overhead</td>
<td>93</td>
</tr>
<tr>
<td>6.3</td>
<td>SPARE-DoF: Inter-Stream Interference Mitigation and Recovery</td>
<td>93</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Receiver Strategy in CHRoME</td>
<td>94</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Inference of Per-Stream Multiuser Interference</td>
<td>95</td>
</tr>
<tr>
<td>6.4</td>
<td>Multi-User Interference-Aware Fast Recovery</td>
<td>99</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Overview</td>
<td>100</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Retransmission Overhead</td>
<td>101</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Receiver-Dimensionality Reduction and Inter-Stream Interference</td>
<td>103</td>
</tr>
<tr>
<td>6.5</td>
<td>System Implementation, Measurements, and Evaluation</td>
<td>107</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Implementation and Experimental Methodology</td>
<td>107</td>
</tr>
<tr>
<td>6.5.2</td>
<td>M$^3$CS: Probing-Based MCS Selection</td>
<td>109</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Interference Mitigation and Recovery</td>
<td>118</td>
</tr>
<tr>
<td>6.5.4</td>
<td>CHRoME’s Fast Recovery</td>
<td>122</td>
</tr>
<tr>
<td>6.6</td>
<td>Chapter Concluding Remarks</td>
<td>125</td>
</tr>
</tbody>
</table>

7 Conclusion and Future Directions 126

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Summary of Contributions and Significance</td>
<td>126</td>
</tr>
<tr>
<td>7.2</td>
<td>Future Directions</td>
<td>127</td>
</tr>
</tbody>
</table>

Bibliography 129
## Illustrations

3.1 Zero forcing beamforming experimental rate: LOS and NLOS  
3.2 802.11ac Sounding and Feedback Timeline  
3.3 802.11ac Sounding Overhead  
3.4 802.11ac acknowledgement process timeline  
3.5 Complete 802.11ac Timeline  

4.1 Geometric representation of ZF and MMSE  
4.2 CUiC Staggered Feedback Format  
4.3 Control Frame Error Rate of CUiC and 802.11ac  
4.4 Receiver diversity gain in CUiC  
4.5 Matrix condition number and signal variance  
4.6 Channel autocorrelation and overhead impact  
4.7 CUiC’s overhead reduction and throughput gains  
4.8 Throughput of CUiC’s policies  

5.1 MUTE’s platform validation: aggregate rate for an NLOS MxM system  
5.2 Inferred SINR vs. Condition Number  
5.3 Measured magnitude and phase variation in static environments  
5.4 Relative error rate for different sample age  
5.5 MUTE’s per-user rate and overhead reduction  
5.6 MUTE’s throughput gain  
5.7 User diversity in SU-MISO and MU-MIMO
5.8 Preliminary tests of MUTE: 16-QAM, 2x2 constellation  
6.1 Illustration of single-stream vs. multi-stream systems  
6.2 MUBF probing and CSS feedback  
6.3 MUBF ISIT preamble  
6.4 Illustration of channel estimates for interleaved subcarriers  
6.5 Retransmission strategies  
6.6 Configuration minimizing retransmission time  
6.7 SINR gain due to reduction in receiver-dimensionality  
6.8 RBIR vs. SNR and PER vs. SNR  
6.9 MCS selection schemes  
6.10 Throughput of MCS selection schemes  
6.11 MCS selection timeline  
6.12 Throughput and MCS timeline as a function of time gap between sounding and data transmission  
6.13 MCS selection accuracy as a function of time gap between sounding and data transmission  
6.14 Feedback quantization  
6.15 Measured inter-stream interference  
6.16 EVM performance of CHRoME’s IC strategy  
6.17 System’s throughput of CHRoME’s retransmission scheme
# Tables

3.1 802.11ac analysis parameters ........................................... 20
3.2 802.11ac timeline parameters ........................................... 24

4.1 Sounding Overhead Reduction in CUIC ............................... 51

6.1 Percent of fully successful retransmissions ......................... 124
Chapter 1

Introduction

Multi-user MIMO or Multi-User Beamforming (MU-MIMO or MUBF) has emerged as a technology capable of increasing the spectral efficiency of a wireless system by allowing the Access Point (AP) to precode multiple independent streams such that these can be transmitted simultaneously to multiple users. In Single-User MIMO, the capacity gain over Single-Input Single-Output (SISO) scales approximately with $\min\{M, N\}$, where $M$ and $N$ are the number of antennas at the transmitter and receiver, respectively [63]. Consequently, for typical scenarios where the form factor of user terminals, e.g., mobile devices, constrains the amount of radios and antennas they can have, these gains can be very limited. Conversely, in MU-MIMO the AP can serve as many single-antenna users as it has transmit antennas thus eliminating such tight physical constraint. That is, MU-MIMO can attain maximal multiplexing gains with minimal number of antennas on the different user terminals.

Signal precoding leverages the spatial separation between different users to partially or completely cancel the interference among their corresponding streams (inter-stream interference) by multiplying each independent stream by beam-steering weights [5, 52, 63]. Notice that while beamforming is a suboptimal strategy, with optimally chosen precoding vectors and as the number of available users goes to infinity, the sum rate of beamforming has been shown to approach that of the optimal strategy, dirty paper coding (DPC) [42, 63].

An extensive body of literature, both theoretical and experimental, has demonstrated that MU-MIMO can achieve substantial physical-layer (PHY) gains over single-stream sys-
tems, e.g., SISO [5, 7, 45, 52, 63]. Unfortunately, at the system level the cost required to enable robust multi-stream transmissions is much higher than that for the single-stream case. In this thesis we identify and address two major MU-MIMO performance hinderers:

**Inefficient Sounding and Acknowledgement Processes.** First, to enable precoding the AP needs to collect channel state information (CSI or CSIT – CSI at the Transmitter) from all users to be served via a process known as *sounding*. That is, for $S$ users, sounding requires $S$ transmission exchanges (i.e., $O(S)$) between the AP and users in order to obtain their channel estimates. We demonstrate that this leads to a substantial amount of overhead, specially with increasing number of concurrent users. Likewise, the MU-MIMO acknowledgement process also requires the same amount of exchanges, thus reducing the performance gains attained via the simultaneous downlink transmissions. The high overhead incurred by both processes reduces the system’s data air-time utilization (i.e., air-time that could be devoted to the transmission of data packets).

**High Vulnerability to Channel Dynamics and Inaccurate CSIT.** Second, in contrast to single-stream systems, MU-MIMO is significantly more sensitivity to channel dynamics or inaccurate CSIT. In particular, fading and inaccurate or outdated CSIT translates into a decrease in signal strength in single-stream systems. However, in MU-MIMO this translates not only into a reduction in signal strength but also into an increase in inter-stream interference, which aggravates with increasing number of concurrent streams.

In this thesis we propose three different protocols to address the aforementioned MU-MIMO gains hinderers. These protocols introduce a suite of strategies for diminishing the negative impact of sounding overhead and inter-stream interference on next generation MU-MIMO WLAN implementations. In particular this thesis presents the following contributions:

- First, we design, implement, and experimentally evaluate Concurrent Uplink Control
Messages (CUiC) to scale the MU-MIMO control information exchange process and vastly improve the efficiency of 802.11ac-based MU-MIMO networks. The key technique is the design of new channel sounding and acknowledgment mechanisms that enable multiple users to transmit their reverse-direction control messages (i.e., beamforming reports and acknowledgments) concurrently to the AP, in $O(1)$ transmission slots. We implement CUiC and perform an extensive set of experiments and demonstrate throughput gains of more than 100% compared to 802.11ac.

Second, 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 CSIT, 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.

Finally, we introduce the design of CHRoME, a downlink multi-user beamforming protocol that addresses the inherent sensitivity of multi-stream systems to user and environmental mobility, and imperfect channel state information. CHRoME presents the following key strategies: (i) a technique for accurately selecting the downlink bit rate in the presence of inter-stream interference via a MUBF probe prior to the data transmission, (ii) a mechanism for identifying and removing the most harmful interfering stream when the available degrees of freedom are not sufficient to cancel all interferers, and (iii) a fast retransmission scheme that exploits liberated antenna resources to increase the expected per-user SINR and retransmit without having to re-sound the channel. We implement each mechanism
and evaluate them via a combination of indoor over-the-air experiments and trace-driven
 emulation. We demonstrate that CHRoME increases the resilience of MUBF systems to
 inter-stream interference and achieves substantial throughput gains compared to 802.11ac.
Chapter 2

Literature Review

This chapter presents an overview of prior work in the multiple areas covered in this thesis.

2.1 MU-MIMO Control Messages Overhead Reduction

Overhead Reduction. Prior work is comprised of both theoretical and practical MU-MIMO implementations for WLANs and cellular networks. Information theoretical works that are most relevant to CUIC and MUTE address the issue of sounding feedback compression, user selection and grouping, as well as alternative coding and sounding feedback strategies, e.g., [13, 31–33]. The work in [13] presents an information theoretical based characterization of the ergodic achievable rates of a zero-forcing beamforming under imperfect and quantized channel state feedback. A comprehensive tutorial on limited or finite-rate feedback MIMO wireless systems is presented in [31] (see references therein). Mostly, the focus of such work is on quantization mechanisms employed to compress feedback. Similarly, in [32] the authors provide a codebook-based technique to reduce feedback to only an index that indicates the best code to be used for signal precoding. In [33] the authors propose a multiplexing scheme that obviates the need for accurate or updated CSI at the transmitter (CSIT) by triggering a multi-stage transmission that allows each user to reconstruct its intended symbol based on both a copy of that particular symbol as well as a linear combination of its intended and interfering symbol.

Recently, some experimental works on MU-MIMO overhead reduction have appeared
in the literature [4, 37, 45, 61]. A differential feedback strategy for reducing the overhead of the explicit sounding process in 802.11ac is proposed in [37]. This differential feedback technique exploits any high correlation between the current and previous channel realizations in order to provide the AP with sufficient information to precode the signals. That is, feedback conveys only the change in the channel, relative to the previous realization. Alternatively, Argos [45] presents an implicit feedback scheme to avoid the substantial overhead incurred by explicit feedback systems. Moreover, to ensure full channel reciprocity Argos proposes a calibration technique. In [4] the authors present a user and mode selection algorithm that does not require CSIT collection. More specifically, this scheme implements a two-step process that avoids sounding all users associated to the AP and instead pre-selects a subset of them to be sounded. While this approach significantly improves the user selection process, it still requires all pre-selected users to be sounded before the downlink data transmission. Similarly, in [61] the authors present a practical feedback compression approach based on 802.11ac that reduces feedback over time, frequency, and quantization domains.

Our work contrasts from the rest in the following way. First, via a testbed implementation of MUTE we demonstrate that in a wide range of scenarios sounding is not needed before every single MU-MIMO data transmission. That is, the overhead incurred by the sounding process can be alleviated by learning about the statistics of the channel and determining whether the previously collected CSIT is useful for future transmissions. Second, our work implements an explicit feedback scheme that already incorporates channel compression techniques by quantization as mandated in 802.11ac, and introduces a fourth dimension in compression, i.e., along the spatial domain, to further alleviate the problem of sounding overhead. Moreover, by relying on explicit feedback, our protocols guarantee better accuracy compared to implicit systems and avoid complex and expensive trans-
mit/receive chains calibration [24, 61].

**Multiplexed Stream Detection and Transmit Synchronization.** Our protocol CUiC implements a space division multiple access (SDMA) technique to parallelize CSI feedback transmissions as well as other control messages such as acknowledgements. The seminal work introducing the concept of spatial demultiplexing for single-user MIMO systems is presented in [59] via the design of the V-BLAST architecture. In [23] a joint interference alignment and cancellation technique is proposed to decode concurrent streams in MIMO LANs. An extensive body of literature on receiver architectures for decoding multiple streams in MIMO scenarios can be found in [47, 52], and references therein. The focus of such prior work is on single-user MIMO. Due to the lack of synchronization between different users, such MIMO techniques can perform poorly in MU-MIMO scenarios. In [50] the authors target a distributed, uncontrolled environment where each user contends on random access fashion. [50] proposes an uplink chain decoding scheme to perform interference cancellation on the composite received signals in order to decode independent streams. Additionally, in [43] a rate adaptation scheme the complements the work in [50] is presented. Furthermore, [15] provides a technique to exploit not only multiplexing gains in MU-MIMO but also diversity gains in the case of multiple antennas at the receiver. Likewise, [62] presents a MU-MIMO scalable architecture targeting the problems of real-time signal processing and the increased computational complexity in such systems, specially as the number of antennas and streams increases.

In contrast to all this work, we focus on the recent 802.11ac amendment and leverage the structure of the proposed sounding and acknowledgement processes to enable concurrent uplink control messages, thus avoiding the transmission alignment complexity that arises in random access transmissions. That is, CUiC operates on control frames instead of data, consequently targeting overhead reduction and significantly simplifying protocol
requirements. Moreover, we perform a thorough evaluation of the demultiplexing and decoding capabilities of the AP in different indoor scenarios and demonstrate the feasibility of such system in next generation 802.11-based networks. In contrast, our protocol provides with the synchronization required to perform the simpler decoding based on single-user MIMO techniques on a multi-user MIMO environment.

In LTE-A, uplink MU-MIMO is enabled via centralized signaling provided by the base station (eNodeB) in order to simplify the decoding process. That is, at each scheduling assignment the eNodeB provides a 3-bit reference signal indicator that assigns the orthogonal code and phase rotation for each terminal (user) [20]. On the other hand, in 802.11-based uplink MU-MIMO, the lack of such signaling leads to additional requirements and challenges for synchronization and decoding processes.

2.2 MU-MIMO Resilience to Mobility, Interference, and Imperfect Channel State Information

Achieving interference resilient downlink MUBF requires the implementation of a variety of mechanisms throughout different phases of the MUBF transmission. Based on the mechanisms we have implemented in CHRoME, prior literature can be divided into pre- and post- MUBF transmission techniques.

Pre-MUBF Transmission. Both theoretical and practical works have focused on developing user, mode, and MCS selection strategies that attempt to minimize the effects of inter-stream interference. User and mode selection (or spatial scheduling) involves grouping users based on their spatial correlation with the purpose of maximizing the SINR at each user, consequently enabling the use of higher MCS. Existing work relies on the CSIT measured during the sounding phase in order to allow the AP to determine the post-
processing SINRs of a particular set of users, and select their independent MCS accordingly [4, 16, 17, 27, 40, 44]. In contrast, our work focuses on enabling resilience regardless of the pre-chosen user set and their corresponding MCS. Therefore, CHRoME complements these protocols. In addition, notice that other approaches aimed at eliminating inter-stream interference in MUBF focus on the implementation of accurate CSIT estimation and quantization techniques [3,52], as well as a wide variety of precoding strategies [17,51,58]. CHRoME works in combination with any type of legacy CSIT estimation and quantization, as well as precoding strategy thereby also complementing all of these downlink systems.

**Post-MUBF Transmission** Upon a failed MUBF transmission, previous work on downlink MUBF WLAN systems either follow the retransmission mechanism proposed in 802.11 (via binary exponential backoff) or completely ignore the retransmission process. On the other hand, cellular systems like LTE Advanced employ mechanisms such as Automatic Repeat-reQuest (ARQ) and Hybrid-ARQ [20,46] to fast and efficiently recover from such losses.

In contrast to MUBF WLAN systems, CHRoME avoids the costly overhead incurred by the combination of sounding and binary exponential backoff via a one-time immediate retransmission thus reusing the same CSIT but exploiting additional degrees of freedom. Moreover, CHRoME benefits from the added robustness that can be added via ARQ schemes such as Hybrid ARQ with incremental redundancy.
Chapter 3

Background on Multi-User MIMO and the IEEE 802.11ac Amendment

3.1 Multi-User MIMO Primer

Single-User Multiple-Input Multiple-Output (SU-MIMO) communication systems with $M$ antennas at the transmitter and $N$ antennas at the receiver have been shown to provide a capacity gain of $\min\{M, N\}$ times that of Single-Input Single-Output (SISO) systems [22, 63]. Nevertheless, while the AP can usually accommodate several antennas, the form factor (size) of mobile devices restrict the number of antennas that these users can have. A similar argument can be made about the cost involved.

Multi-User MIMO (MU-MIMO) eliminates such restriction by enabling the multi-antenna AP to serve multiple users with one or more antennas, simultaneously [21]. Moreover, results have shown that MU-MIMO achieves similar capacity scaling to that of SU-MIMO when an $M$ antenna AP communicates with as many as $K = N$ single-antenna users [57]. That is, in theory, MU-MIMO can potentially achieve maximum multiplexing gain even when users are limited to a single antenna. Many different baseband techniques have been proposed to enable multiple concurrent downlink transmissions via the implementation of linear and non-linear precoding schemes at the transmitter. With the assumption of full channel state information at the transmitter (CSIT), an interference pre-subtraction technique known as Dirty Paper Coding (DPC) (in combination with optimal power allocation) [18] has been shown to be the capacity-achieving strategy. Nonetheless,
DPC is difficult to implement due to its prohibitive computational complexity, especially with increasing number of users [63].

3.1.1 Multi-User Beamforming

System Model and Notation

Let $M$ denote the number of antennas at the AP and $N_k$ represent the number of antennas at user $k = 1, \ldots, K$, where $K$ is the total number of users associated with a particular AP. We follow the narrowband channel model in [63] where the channel gain between a user and an antenna at the AP is described as a zero-mean circularly symmetric complex Gaussian random variable. User $k$ receives signal

$$y_k = H_k x + z_k, \quad k = 1, \ldots, K \tag{3.1}$$

where $H_k \in \mathbb{C}^{N_k \times M}$ represents the channel gain matrix between the AP and user $k$, $x \in \mathbb{C}^{M \times 1}$ represents the symbol transmitted by the base station, and $z_k \in \mathbb{C}^{N_k \times 1}$ is the additive white Gaussian noise (AWGN) at the $k^{th}$ user. The elements of $H_k$ are assumed to be independent, and both the channel and the noise are normalized such that noise and the entries of $H_k$ have unit variance. In addition, we let the transmitter have an average power constraint $\mathbb{E}\{xx^*\} \leq P$ [63]. We use uppercase (lowercase) boldface to denote matrices (vectors). $H^*$ ($h^*$) denotes the complex conjugate transpose of matrix $H$ (vector $h$), $H^T$ ($h^T$) the transpose, and $\mathbb{E}$ denotes the expectation operator.

Beamforming

Multi-User Beamforming (MUBF) is a suboptimal strategy that allows the AP to simultaneously transmit independent data streams to multiple users by precoding these streams. Each of the data streams is multiplied by a beam steering weight vector for transmission.
across all the available transmit antennas. The weight vectors are used to reduce or eliminate interference among the different streams [63]. In a multi-user scheme the transmitted signal $\mathbf{x}$ is given by $\sum_{k=1}^{K} \sqrt{P_k} \mathbf{w}_k \hat{x}_k$, where $\hat{x}_k$, $\mathbf{w}_k$, and $P_k$ denote the data symbol, beam steering weight vector, and transmit power scaling factor for user $k$, respectively [63]. The received signal for user $k$ is given by:

$$y_k = (\sqrt{P_k} \mathbf{h}_k \mathbf{w}_k) \hat{x}_k + \sum_{j \neq k} \sqrt{P_j} \mathbf{h}_j \mathbf{w}_j \hat{x}_j + z_k$$ (3.2)

where the first element corresponds to the intended signal, the second one represents the interference, and the third element $z_k$ is the additive noise with unit variance, at the $k$th user.

The sum rate achieved by beamforming is [54]

$$R = \max_{\mathbf{w}_k, P_k} \sum_{k=1}^{K} \log \left( \frac{1 + \sum_{j=1}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2}{1 + \sum_{j=1, j \neq k}^{K} P_j |\mathbf{h}_k \mathbf{w}_j|^2} \right)$$ (3.3)

subject to

$$\sum_{k=1}^{K} ||\mathbf{w}_k||^2 P_k \leq P$$

### 3.1.2 Zero-Forcing Beamforming

In this work we focus mostly on Zero-Forcing Beamforming (ZFBF) which consists of a suboptimal beamforming strategy commonly used in practice due to its performance which in the limit of large $K$ it is comparable to the optimal approach [63], and due to its ease of implementation. ZFBF requires CSIT and therefore relies on channel sounding to obtain each users’ channel estimates in order to generate the beam steering weights that enable MUBF [54, 58]. The ZFBF weights are chosen such that a zero-interference condition at the intended user is satisfied, i.e., $\mathbf{h}_k \mathbf{w}_j = 0$ for $j \neq k$. Therefore, the second term in equation 3.2 becomes zero.
For a set of users $S$ to be concurrently served by the AP ($S \subset \{1, \ldots, K\}$, $|S| = S \leq M$), let $W(S)$ and $M(S)$ be the sub matrices of $W = [w_1, \ldots, w_K]$ and $H = [h_1^T, \ldots, h_K^T]^T$, respectively. The choice of weights $W(S)$ that yields zero interference is the pseudo-inverse of $H(S)$ [58],

$$W(S) = H(S)^\dagger = H(S)^\ast(H(S)H(S)^\ast)^{-1}$$  \hspace{1cm} (3.4)

Notice that with ZFBF, the maximum number of independent data streams that can be transmitted is equal to the number of antennas at the transmitter.

**Note:** Throughout this thesis we will use the terms MUBF and MU-MIMO interchangeably. Moreover, unless otherwise stated, we consider an equal power allocation where the AP allocates equal power to all concurrently served users/streams. Results presented in [5] have shown that in high SNR regimes (experiments performed in typical indoor office environments), the performance of equal power allocation is very close to that of the optimal power allocation scheme, i.e., waterfilling approach.

### 3.1.3 Experimental Platform - WARP and WARPLab [1]

The Wireless Research Open Access (WARP) platform consists of an FPGA-based software defined radio platform, interfaced with custom designed radios based on the MAX2829 chipset. The MAX2829 transceiver chip operates at both the 2.4 and 5 GHz bands over 20MHz channels. This platform allows for the implementation of clean-slate PHY and MAC protocols. We perform experiments using WARPLab, a programming environment that integrates MATLAB tools with WARP to control the platform via a host computer for running experiments and collecting data. This allows us to run all the baseband signal processing at a central controller (hostPC), and then transmit data over-the-air. WARPLab provides access to analog sample send/receive buffers and RSSI measurements for each experiment.
3.1.4 Performance of ZFBF

We implement a generic narrowband multi-user ZFBF system in WARP and investigate its performance under different indoor scenarios.

**Serving Maximum Number of Users Possible**

We demonstrate the performance of MU-MIMO without considering the effects of sounding overhead on the overall performance of the system. We do this to demonstrate the potential gains of MU-MIMO in indoor WLANs. First, we consider the case where the number of users served equals the Degrees-of-Freedom (DoF) of the system.

To this end, we conduct over-the-air experiments where we vary the number of antennas at the AP as well as the number of users from 1 to 4. In order to compare the 4x4 case with the 1x1 case we set a constraint on the transmission power. That is, the total power used for transmission is constant, thus with increasing number of antennas, the power allocated to each antenna is reduced to \( \frac{1}{M} \), where \( M \) is the number of antennas at the AP. We explore Line-Of-Sight (LOS) and Non-LOS (NLOS) scenarios in an indoor environment to understand the behavior under different conditions. We run experiments for more than 50 user locations and for all possible user combinations out of the total set of 4.

In Figure 3.1 we present both the average empirical and theoretical per-user rate. Theoretical rates are obtained by using collected channel traces as input to Equation (3.3). Observe that in both figures, the per user rate drops as the number of data streams increases. This is mainly caused by the equal power constraint we impose. More importantly, notice that the difference between our experimental results and theory are very small. These small discrepancies appeared whenever the ZFBF technique was not able to completely suppress inter-user interference as expected in theory. This is more evident as the number of users increases thus increasing the potential for higher interference levels.
Figure 3.1: Per-user rate with same number of AP antennas and single-antenna users in LOS (top) and NLOS (bottom). For aggregate rate each result needs to be multiplied times the number of users served.
These results reveal that the aggregate rate achieved by a 4x4 system is much higher than that achieved by a 1x1 system, i.e., MU-MIMO vs. SU-MISO. Namely, our evaluation shows an average rate increase of approximately 220% and 86% for both LOS and NLOS, respectively.

### 3.2 Sounding and Acknowledgement Processes in MU-MIMO

Multi-user beamforming strategies (e.g., ZFBF) rely on channel sounding to acquire channel information about the different users to be served. Channel information is required for two main reasons: Firstly, the precoding technique requires channel estimates to compute the beam-steering weights that multiply each of the data streams for the different users [5]; Secondly, in order to serve multiple users simultaneously, the AP triggers a user selection procedure to find such group. To this end, the AP needs the channel information to each user in order to decide which of them should be served concurrently to minimize inter-user interference (or inter-stream interference) [63]. Ideally, the selected group consists of users with orthogonal or semi-orthogonal channel vectors between them.

**Sounding Overhead in Generic MU-MIMO.** The overhead associated with sounding is directly proportional to the number of transmit antennas, to the number of 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, 5] requires the AP to broadcast a pilot from each of its antennas so that the users can estimate their channel vectors from the AP. Then, the channel information is fed back to the AP in order to generate the beam-steering weights. Let $S$ be the number of concurrently served single-antenna users, and $M$ be the number of antennas at the AP (total of $M \cdot S$ channels). Then, assuming sounding over a single channel for $S$ users, explicit sounding
requires $O(M)$ time to send the pilots, and $O(M \cdot S)$ 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 $O(S)$ [45].

**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 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 [45]. 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, at the expense of higher overhead, explicit feedback provides more reliable channel information matrices and does not require

---

*802.11ac only supports downlink MUBF. Throughout this work we refer also to users as *beamformees* and to APs as beamformer.
such calibration.

**Sounding Overhead in 802.11ac.** 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, 9]. 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 3.2).

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 (phase/magnitude) in a parameterized compressed feedback matrix $V$.

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 single-stream users, the total amount of time it takes to transmit this frame is about $52 \mu s$.

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

<table>
<thead>
<tr>
<th>AP</th>
<th>NDPA</th>
<th>SIFS</th>
<th>NDP</th>
<th>SIFS</th>
<th>Beamforming Report Poll</th>
<th>SIFS</th>
<th>Beamforming Report Poll</th>
<th>SIFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Compressed Beamforming</td>
<td></td>
<td>Compressed Beamforming</td>
<td></td>
</tr>
<tr>
<td>U2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Compressed Beamforming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Compressed Beamforming</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3.2 : 802.11ac Sounding and Feedback Timeline (not to scale).*

The process continues in a cyclic manner, allowing for efficient and accurate channel information exchange among the AP and its associated users.
channels 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 generate most of the overhead incurred in the sounding pro-
cedure 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]. In Table 3.1 we present the 802.11ac parameters employed in the
following overhead analysis. 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 beam-steering
weights and beamforms to the same subset of users. The total overhead for a 4x4, 20 MHz
system with lowest quantization (5 and 7 bits, see Table 3.1), and 40 µs PLCP header per
frame, is

\[
T_{total} = t_{NDPA} + 8 \cdot t_{SIFS} + t_{NDP} + 4 \cdot t_{report} + 3 \cdot t_{poll} \\
\approx 1.2 \text{ms}
\]

In Figure 3.3 (top) 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 trans-
mission. We do this for different transmission rates (from QPSK with \( \frac{3}{4} \) FEC to 256-QAM
with \( \frac{5}{6} \) FEC). The rest of the parameters used are shown in Table 3.1. 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 Bytes) [49] to the
maximum A-MPDU aggregation with single MSDU of 64 frames (96 KB) [3].

\[†\] Although

\[†\] A-MPDU stands for Aggregate MAC Protocol Data Unit, and MSDU stands for MAC Service Data Unit.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFS</td>
<td>$16 \mu s$</td>
</tr>
<tr>
<td>Chan. Width</td>
<td>20 and 80 MHz</td>
</tr>
<tr>
<td>AP Antennas</td>
<td>4</td>
</tr>
<tr>
<td>Number of Users</td>
<td>2 and 4</td>
</tr>
<tr>
<td>Subcarrier Grouping</td>
<td>4</td>
</tr>
<tr>
<td>Quantization $\psi$, $\phi$</td>
<td>5, 7 bits</td>
</tr>
</tbody>
</table>

**Table 3.1 : 802.11ac Parameters Used in Analysis**

The amendment defines even larger frame aggregations, such large packets can hardly be used under realistic conditions, e.g., see [49]. 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. In addition we assume 400 ns guard interval (GI).

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 35% of the total transmission time when considering 18 KB packets. In fact, as mentioned above, it takes the AP 1.2 milliseconds to sound all users; in the same amount of time, a device could transmit more than 3 kB of data at QPSK, and more than 13 kB of data at 256-QAM. Also, observe that sounding even only 2 users consumes 20% of the total transmission time in a 256-QAM transmission with 96 KB packets and almost 60% with 18 KB packets (at 80 MHz). It can be shown that if a device can transmit at 433 Mbps, the amount of time used for sounding could have been used to transmit 13,000 more bytes [2]. Moreover, our analysis reveals that sounding more than four users before a ZFBF
transmission in such systems becomes prohibitive. 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.

To better illustrate the impact of sounding overhead on system performance we compare the throughput achieved by MU-MIMO and Single-Input Single-Output (SISO) systems whenever this overhead is considered. Figure 3.3 (bottom) depicts the throughput of both systems for a 4-antenna AP serving 2 and 4 users, as a function of the number of aggregate frames (number of MPDUs within one A-MPDU). We consider a modulation and coding scheme of 256-QAM, 1500 byte packets, lowest feedback quantization, 80 MHz channel width, and maximum subcarrier grouping of 4. Notice that the SISO system implements a TDMA (Time Division Multiple Access) scheme that avoids contending for the medium (consecutive transmissions). We only consider the time it takes for transmission therefore we assume perfect reception (perfect loss-less channels). Observe that even when considering perfect MU-MIMO, it is outperformed by SISO at low-aggregation regimes. Therefore, for such regimes it is absolutely unnecessary to enable multi-user transmissions since a SISO transmission would suffice to achieve equal or better performance. In subsequent chapters we present the design and implementation of two strategies for minimizing the impact of sounding on the performance of MU-MIMO systems.

**Structure of the Acknowledgement Process in 802.11ac.** The MU-MIMO acknowledgement process (shown in Figure 3.4) has a similar structure to the sounding process. In general, it consists of an exchange of Block ACK Requests (BAR) and Block ACKs (BA) between the AP and the different users. Notice that in 802.11ac all MPDUs are A-MPDUs. Therefore, all replies take the form of block ACKs. After an MU-MIMO downlink data transmission, one of the users replies immediately with its corresponding BA, whereas the
Figure 3.3: 802.11ac Sounding Overhead: (Top) Fraction of airtime consumed by sounding overhead; (Bottom) throughput comparison between MU-MIMO and SISO, considering sounding overhead.
rest of them are polled for their BAs via a BAR sent by the AP. To determine which user has to reply first, the ACK policy for the data MPDU of one specific user is set to implicit BA.

![Diagram](attachment:image.png)

**Figure 3.4 : 802.11ac acknowledgement process timeline (not to scale).**

**Complete 802.11 Timeline.** Finally, in Figure 3.5 we present a timeline considering both the sounding and acknowledgement processes at their corresponding time-scales. Data transmits a maximum of 65535 octets (A-MPDU size is 65535, enclosed within one Very-High Throughput PLCP Protocol Data Unit or *VHT-PPDU*). The timeline was generated using the values shown in Table 3.2. The legend shows the following components (from left to right): Exponential Backoff (EBO), Short Interframe Space (SIFS), Null Data Packet Announcement (A-NDP or NDPA), Null Data Packet (NDP), Compressed Beamforming Report (cBFr), Report Poll (pTime), Data (A-MPDU), Block Acknowledgement (BA), and Block Ack Request (BAR).
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFS</td>
<td>16 $\mu$s</td>
</tr>
<tr>
<td>DIFS</td>
<td>34 $\mu$s</td>
</tr>
<tr>
<td>PLCP header</td>
<td>40 $\mu$s</td>
</tr>
<tr>
<td>MAC header</td>
<td>30 Bytes</td>
</tr>
<tr>
<td>BAR</td>
<td>26 Bytes</td>
</tr>
<tr>
<td>BA</td>
<td>32 Bytes</td>
</tr>
<tr>
<td>$N_{ss}$ per user</td>
<td>1</td>
</tr>
<tr>
<td>Channel Width</td>
<td>40 MHz</td>
</tr>
<tr>
<td>AP Antennas</td>
<td>4</td>
</tr>
<tr>
<td>Number of Users</td>
<td>4</td>
</tr>
<tr>
<td>Subcarrier Grouping</td>
<td>1</td>
</tr>
<tr>
<td>Quantization $\psi, \phi$</td>
<td>7, 9 bits</td>
</tr>
<tr>
<td>MCS</td>
<td>64-QAM $\frac{5}{6}$</td>
</tr>
</tbody>
</table>

Table 3.2: 802.11ac parameters used to construct the timeline
Figure 3.5: 802.11ac Timeline
Chapter 4

Concurrent Uplink Control Messages

This work has been previously published by IEEE in [11]

4.1 Introduction

Downlink Multi-User MIMO (MU-MIMO) enables capacity gains via simultaneous transmission from the access point (AP) to multiple users [21,52]. Consequently, downlink MU-MIMO has been incorporated into the Wi-Fi standard via the IEEE 802.11ac amendment, which promises multi-Gb/sec data rates [3,9]. Unfortunately, despite such high physical layer rates, a significant amount of airtime is devoted to providing the access point (transmitter) with the required Channel State Information (CSI or CSIT).* Indeed, in 802.11ac, the number of required CSI feedback messages increases linearly with the number of users simultaneously served. Similarly, after every downlink MU-MIMO transmission, the number of acknowledgements (ACKs) and ACK-request exchanges between the AP and the served users also increases linearly with the number of users. We will show that the time-resources devoted to these exchanges severely reduce throughput.

In this chapter, we present the design, implementation, and experimentally validation of a protocol for concurrent uplink transmission of control messages from multiple users.

*In this chapter we use CSI and CSIT interchangeably. These refer to the channel state information used by the AP to precode the downlink data streams. In contrast, CSIR denotes the channel state information measured by the AP during uplink transmissions (i.e., user to AP).
Namely, we present CUiC (Concurrent Uplink Control) to replace the aforementioned sequential CSI feedback and ACK messaging system with a design that achieves parallel transmission. In particular, we present the following contributions.

First, we demonstrate how to scale uplink feedback by employing only a single transmission slot to simultaneously send the control messages from all users (independently of the number of users). To realize concurrent single-slot feedback, we demonstrate that preamble staggering of user-to-AP packets enables the AP to obtain “clean” measurements for performing automatic gain control (AGC) and estimating the channel (channel state information at the receiver or CSIR), carrier frequency offsets, and other timing offsets due to digital and RF mismatches between users and AP. Unlike random access uplink MU-MIMO MAC protocols which require (i) additional user-AP frame exchanges to train the AP [29, 43], or (ii) special-purpose space-division multiple access (SDMA) techniques to counteract the effects of multi-stream signal overlap (on both preamble and data portions of the packet) [50], we design CUiC such that it emulates a multi-stream SU-MIMO beamforming system. Namely, in CUiC the AP pre-determines the timings required to align the transmissions from all concurrent users to within $800 \text{ ns}$, and acquires the training needed for stream demultiplexing without additional signaling or frame exchanges; Consequently, this enables the use of legacy signal detection and correction techniques by relying solely on packet preambles, and SU-MIMO spatial demultiplexing techniques [52]. Note that our mechanism only applies to uplink control messages in which the users are pre-selected (the set of users selected by the AP for downlink data is a subset of the users sending CSI to the AP) and timing can be tightly controlled by the AP (see Chapter 2 for further discussion of related work). Furthermore, we demonstrate that as a desirable side effect, CUiC also increases robustness of the sounding process by reducing the time between CSIT estimation and downlink data transmission, which is critical in fast fading channels or in the presence
of highly mobile users.

Second, we implement CUiC in an 802.11 OFDM MU-MIMO platform prototype based on WARP and WARPLab [1], and demonstrate the feasibility of decoding multiple simultaneous wide-band 802.11ac control messages in uplink MU-MIMO random access systems via preamble staggering and traditional spatial demultiplexing techniques. Moreover, we propose a frequency offset correction method comprised of a combination of pre-compensation and iterative correction. This allows the AP to apply the offset of a single user to the composite stream while treating the rest of the signal components as noise at each iteration. We demonstrate that in a system comprised of four simultaneously transmitting single-antenna users and a 4-antenna AP, CUiC can achieve a frame error rate of only 6% even at low SINR regimes, i.e., -5 dB.

Third, we propose and evaluate a suite of sounding policies that enable different user selection algorithms, e.g., [4, 60]. More specifically, we design CUiC so that it can be tailored to achieve one of the following targets: user diversity maximization, sounding air-time minimization, and matching the amount of data collected to the number of degrees of freedom (DoF) available at the transmitter (802.11ac-like).

Finally, we evaluate the performance of CUiC via a combination of over-the-air, and channel emulator based experiments and compare against an 802.11ac benchmark. Our evaluation reveals that all CUiC policies outperform the benchmark for a wide range of data packet lengths. In particular, we demonstrate the importance of adjusting the system by selecting a different policy for different frame aggregation regimes. More importantly, we show that even when large frame aggregation is utilized to amortize overhead (e.g., 96 kB), our scheme can attain close to 40% throughput gains compared to the benchmark. Similarly, for short packet sizes (e.g., 1.5 kB), CUiC can achieve more than 1.3x throughput gains.
4.2 Background on Receiver-Side Beamforming

CUiC relies heavily on decoding spatially multiplexed streams (also known as receiver-side beamforming) in order to reduce the amount of time spent on sounding feedback and data acknowledgement acquisition. Therefore, we dedicate this section to provide a brief overview of spatial multiplexing detection and decoding techniques.

Chapter Notation. As presented in Chapter 3.1.1, $M$ denotes the number of antennas at the AP and $N_k$ represent the number of antennas at user $k = 1, \cdots, K$, where $K$ is the total number of users associated with a particular AP. In this Chapter we use $S$ to denote the number of users transmitting concurrent messages to the AP. Although users can have more than one antenna, for simplicity we consider one data stream per user at any time. $H \in \mathbb{C}^{S \times M}$ represents the channel gain matrix between the AP and all users served simultaneously. We use uppercase (lowercase) boldface to denote matrices (vectors). Thus, $h_s \in \mathbb{C}^{1 \times M}$ represents the channel vector to user $s$.

Spatial Multiplexing Detection and Decoding. In an SDMA transmission, the AP receives a linear combination of $S$ signals on each antenna. Thus, if the AP receives $M$ linear combinations, for $M = S$ unknown transmissions, each of these multiplexed streams can be estimated. To accomplish this, several linear schemes such as Maximum Likelihood (ML), Minimum-Mean Square Error (MMSE), and Zero-Forcing (ZF) have been proposed [12, 47, 52].

Although ML is the optimal, capacity-achieving scheme, it requires the receiver to test all possible input values or transmitted space-time symbols. Therefore, its complexity is proportional to $\Psi^S_c$ where $\Psi_c$ is the number of symbols in the modulation constellation [47], which makes this scheme highly impractical and even prohibitive with increasing number of transmit antennas and constellation size [52]. The complexity of MMSE and ZF are much lower since they only involve a linear transformation requiring $M \times S$ complex
multiplications. They also require an inverse calculation which in slowly fading channels can be done only once per frame and is not required per space-time symbol [47].

With full CSIR at the AP, ZF can be used to completely suppress inter-stream interference. Figure 4.1(a) illustrates the operation of ZF [12]. Notice that $h_1 x_1$ is projected onto the space orthogonal to the interferer (null space). This projection is represented by $h_1^{* \text{proj}} h_1 x_1$. The main drawback of ZF is that it leads to significant noise enhancement if the channel matrix $H$ is ill-conditioned. That is, if the angle between the two vectors is small, the projection becomes very small as well, which means that the resulting SINR is significantly decreased. In contrast to ZF, Figure 4.1(b) illustrates that MMSE does not suppress the interference completely but provides the optimal compromise between interference suppression and maximizing the signal strength from the intended user [52].

![ZF and MMSE](image)

**Figure 4.1:** Geometric representation of ZF and MMSE. In this 2x2 system, stream 1 with channel $h_1$ is the intended stream (stream to decode), whereas stream 2 with channel $h_2$ is the interferer.

MMSE minimizes the expected value of the mean square error between the transmitted signal vector $x$ and the received vector $W^* y$. $y$ is the linear combination of all received
signals and $W^*$ is given by

$$W^* = (H^*H + \sigma^2 I)^{-1}H^*$$

where $\sigma^2$ represents the noise variance. Notice that at high SNR, the term $\sigma^2 I$ becomes negligible, therefore MMSE achieves the asymptotic performance of ZF [12].

Additionally, a combination of MMSE and a non-linear decoding technique termed Successive Interference Cancellation (SIC) achieves the capacity of the MIMO channel when fading is i.i.d. Rayleigh [52]. SIC consists of an iterative receiver, that is, it decodes one stream at a time. More specifically, when decoding a stream, it considers the remaining signal as interference. Then, after decoding a particular stream, its contribution to the overall signal is reconstructed and removed from it. This procedure continues until all streams are decoded.

### 4.3 Concurrent Uplink Control Messages

Today, uplink control messages in MU-MIMO WLANs are transmitted in the same way as other control frames, i.e., clients transmit feedback sequentially, one at a time. Consequently, both the channel state acquisition process (sounding) and the acknowledgment process yield a time-resource-intensive frame exchange between the AP and each individual user before and after a MU-MIMO transmission. This translates into a corresponding reduction in effective airtime utilization, proportional to the number of users comprising the transmission. In contrast, CUiC realizes concurrent uplink transmission of beamforming reports and acknowledgements from a set of users, thereby achieving the same control message acquisition goal in only a fraction of the time.
4.3.1 Enabling Spatially Multiplexed Feedback

**Feedback Timing Structure.** CUiC introduces an explicit channel estimation feedback technique that allows the multiple users sounded to reply with their channel information simultaneously. In addition, it also introduces an acknowledgement process where users reply simultaneously with their corresponding block acknowledgements (BAs) after a downlink MU-MIMO transmission. These procedures lead to a complex spatial multiplexing scenario in which the AP has the task of detecting and decoding several signals with no inherent orthogonality among them in time, or code. To achieve this, we design a scheme that implements a multiplexed signal decoder, and modifies the structure of the feedback and ACK processes implemented in the 802.11 standard (i.e., 802.11ac amendment), to easily coordinate transmissions from all users.

Spatial multiplexing decoding techniques have been widely employed in single-user MIMO. Nonetheless, the application of these techniques in uplink random access SDMA systems poses several challenges that have prevented their implementation in existing network deployments. Most of these challenges reduce to timing constraints required for user synchronization as well as estimation and correction of transceiver RF chain impairments. CUiC provides concurrent uplink control messaging and establishes the timing structure necessary to synchronize users and provides the mechanisms to estimate and correct for RF chain impairments, thus enabling spatial demultiplexing via MMSE-SIC decoding.

To perform MMSE-SIC decoding, the AP requires channel information and timing offset estimates, e.g. frequency offsets, to all concurrently transmitting users. CUiC exploits the fact that all packet transmissions require a preamble for signal detection, channel estimation, etc., and modifies the preambles of the CSI feedback reports (i.e., compressed beamforming reports) and block ACKs to provide the AP with the information necessary to decode the composite signal. However, to obtain an accurate representation of the chan-
nel to each user and of the different RF mismatches, the preambles need to be clean, that is, without interference from other streams. For that reason, CUiC introduces a preamble time-based staggering technique that prevents all these preambles from overlapping with one another.

![Figure 4.2: CUiC Staggered Feedback Format (not to scale).](image)

**Scalable Sounding and Acknowledgement Structure in CUiC.** CUiC leverages the basic sounding and acknowledgement structures in 802.11ac to align all transmissions such that the preambles from all users are time-staggered.\(^1\) That is, for the case of the sounding process, upon reception of the NDP all users that were addressed in the NDPA packet compute their channel vectors to the AP. Then, based on the order specified in the NDPA, they align their preambles and beamforming reports. Figure 4.2 shows the proposed alignment for the case of three different users. By zero-padding between preambles and the payload carrying the beamforming reports, we can align the reports from all users. This allows the AP to receive each preamble corresponding to a different stream in a clean manner which in turn allows it to estimate the carrier frequency offset, symbol timing, and channel coefficients necessary to decode the different frames. Likewise, for concurrent ACKs, the

---

\(^1\)Refer to Chapter 3 for an overview of the sounding and acknowledgement structures in 802.11ac.
end of the downlink MU-MIMO A-MPDUs serves as the trigger for the multiple users to reply simultaneously. This leads to the suppression of all Block ACK Requests (BARs) and inter-frame spacings required to coordinate channel access between successive replies.

**Synchronization Challenges in Spatial Demultiplexing.** In contrast to single-user MIMO architectures in which all transmitted streams originate from the same device with RF interfaces sharing a common clock and a fully synchronized transmission trigger, in uplink MU-MIMO (SDMA), transmissions from different users do not have these characteristics. Three major synchronization challenges arise when attempting to decode a composite signal [53]: First, a combination of Doppler shifts and discrepancies between the oscillators used to generate the carrier frequency at the AP and users can lead to several different carrier frequency offsets (CFOs) thus causing irreducible inter-carrier interference (ICI). Second, symbol synchronization is necessary to align the signal reception from the multiple users with the observation window at the AP. That is, if the symbols transmitted from all the individual users do not fall within the timing window of the AP, inter-symbol interference (ISI) and ICI can be introduced. Third, the sampling clock frequency is obtained from a local oscillator, consequently, any mismatch between the oscillators of the different devices can lead to a misalignment of the digital sampling clock frequency of all users with respect to that of the AP. Therefore to prevent these issues, the AP or users need to compensate for these offsets.

Although CFO estimation is greatly simplified by using clean preambles, compensation and correction becomes more difficult since the received signal consists of a linear combination of multiple independent streams, where each can have a different offset. CUIC relies on a combination of pre-compensation [53] and residual CFO correction at the AP. That is, in CUIC, each user utilizes both the NDPA and NDP to obtain an accurate CFO estimate and then applies a pre-compensation factor on the next uplink transmission. Nonetheless,
all residual CFO components that remain after pre-compensation need to be removed by
the AP. CUiC proposes a method for correcting such remaining offset from each stream.
In general, the technique treats the signal components of all users except the one currently
being decoded, as noise, and iteratively removes each component. More specifically, after
CFO estimation for all streams, the composite signal is passed through the MMSE-SIC re-
ceiver. Once the AP knows which stream to decode first, it applies CFO correction for that
specific stream to the entire compound signal. Next, the MMSE-SIC algorithm decodes
this first stream, removes the applied CFO correction, and removes the decoded signal
component from the original compound signal. The process is repeated until all streams
are decoded. This processes can be represented as follows: Let \( \hat{x}_i(n) \) be the time domain
symbol \( n \) transmitted by the \( i \)th user, and \( h_i(n) \) be the impulse response of the \( i \)th channel
(for each AP antenna). Thus, after passing the signal of the \( i \)th user through the channel
we obtain \( x_i(n) = \hat{x}_i(n) * h_i(n) \). Ignoring noise terms, the composite received baseband
signal is given by

\[
r(n) = \sum_{i=1}^{S} x_i(n) e^{j2\pi f_i n}
\]

where \( S \) is the number of users transmitting simultaneously to the AP, and \( \Delta f_i \) denotes the
\( i \)th user’s CFO normalized by symbol period. Assuming that the AP attempts to decode the
signal from user 1 first, it will correct the CFO in time domain by multiplying \( r(n) \) with the term \( e^{-j2\pi \Delta f_1 n} \). Thus, in the first MMSE-SIC iteration the compound signal obtained
after correcting the CFO of the first stream is given by

\[
y(n) = r(n) e^{-j2\pi \Delta f_1 n}
\]

\[
= x_1(n)(1) + x_2(n)e^{j2\pi n(\Delta f_2-\Delta f_1)} + \cdots \\
+ x_S(n)e^{j2\pi n(\Delta f_S-\Delta f_1)}
\]
Notice that the MMSE-SIC receiver can now start decoding the component of the first stream. After the first stream has been decoded, the compound signal is multiplied by $e^{j2\pi f_1 n}$ to remove the CFO component from this first stream. This yields a compound signal comprised of all original streams except the first one, where each of them can be decoded in the same way. In our system this method allowed us to remove an offset of up to several kHz.

Pre-synchronization by each user has also been shown to be effective for symbol synchronization and sampling clock frequency misalignment correction [53]. Moreover, in terms of symbol synchronization the use of the OFDM cyclic prefix allows the AP to successfully decode frames for offsets that are less than 0.8 $\mu$s (length of the cyclic prefix). Nevertheless, the implementation of longer prefixes has been shown to enhance robustness [65]. Finally, in addition to pre-synchronization, the correction of the residual sampling clock frequency misalignment can be achieved by performing oversampling, i.e., via a fractional spaced equalizer (FSE) [50].

**Avoiding Signal Saturation via Power Control.** Although SIC is known to perform better when the multiple streams have different magnitudes, if the difference between these concurrent signals is not within the dynamic range of the Analog-to-Digital Converter (ADC), the AGC can erroneously tune the system therefore leading to frame losses, e.g., due to ADC saturation. CUiC allows the AP to do AGC at each individual preamble, and set the RF and baseband gains accordingly. That is, instead of relying on the gains set for the first arriving stream, the AP can tune the gains to the value that maximizes the likelihood of successful reception of all streams. While the development of an adaptive AGC and power control methods is out of the scope of this thesis, we implement a simple AGC scheme in which we tune the system at each preamble and re-tune again during the arrival of the concurrent signals. Nonetheless, we expect that implementing an adaptive
AGC would further increase the performance of the system. In addition, power control techniques in prior work [38] can be incorporated into our scheme.

**Reducing Iterative Error Propagation via Ordered SIC.** Due to the iterative nature of the MMSE-SIC decoder, this scheme is susceptible to error propagation. More specifically, any errors that originate from decoding the first stream (first layer) will propagate to the decoding process for the rest of the layers. We employ ordered-SIC [47] to establish a decoding order among all simultaneous streams. More specifically, based on the *post-processing* SINR of all users, the AP arranges them in descending order and begins to decode each stream one by one. Notice that the post-processing SINR is obtained after the MMSE weights have been applied to the composite signal. Thus, at every iteration the AP selects the strongest stream, which leads to lower decoding errors and therefore to lower error propagation. The fact that CUiC requires all users to transmit their control frames at the same rate means that the stream with highest post-processing SINR has a higher probability of being successfully decoded.

**4.3.2 Variable Length Reports, Missing Uplink Messages, and Retransmissions**

**Variable Length Reports.** Although acknowledgements from the different users have the same length, this might not be the case for the beamforming reports. If simultaneous reports have different lengths, the stream overlap for some symbols will be reduced. Thus, CUiC benefits from an increase in diversity for a few symbols. Notice that the decoder is agnostic to packet length.

**Missing Reports and Retransmissions.** For best reliability we send acknowledgements and beamforming reports at base rate, however, if some of them are missing, CUiC imitates the polling/reply mechanisms in 802.11ac. That is, if an uplink control message is missing, the AP sends either a report poll or a block ACK request (according to the type of
missing frame), until the AP successfully decodes the frame.

**Differentiating Missing Reports and No Feedback.** In order for the AP to know if a user decides not to transmit a feedback report, it relies on the existence of a preamble from that specific user. More specifically, if a user does not need to reply with its feedback, it does not send a preamble. Otherwise if a preamble for a given user is sensed by the AP, but the latter one does not decode the rest of the stream, then it is able to determine if a decoding failure occurred.

### 4.3.3 User Selection and Retransmissions

User selection in downlink MU-MIMO is tightly coupled to the sounding process. Selecting the best subset of users to serve simultaneously can occur before [4] or during [60] sounding. In the context of CUIC, we propose a set of policies for combining sounding and user selection, each with a different objective: (i) *Basic operation (802.11ac-like)* - The AP collects CSIT from as many users as it has transmit antennas (maximum DoF). Regardless of the user selection algorithm employed prior to sounding, the AP collects \( M \) beamforming reports and then serves the subset of those \( M \) users that maximize the aggregate rate (two-round user selection). (ii) *Maximize user diversity (mDiv)* - Increased user diversity leads to increased rate performance of downlink MU-MIMO [10, 63]. With more information about the channels to different users, the AP can select combinations of them that satisfy a certain rate or fairness criteria. CUIC increases the number of CSIT reports provided at each feedback slot by \( M \)-fold, thus giving the AP \( M \) times more information compared to the single-user per feedback-slot case. If we assume that the AP can spend the same amount of time performing sounding as 802.11ac, then it can potentially acquire CSIT from \( M^2 \) users (vs. \( M \) in 802.11ac). In case there are fewer than \( M^2 \) associated users, this scheme could lead to maximum achievable rate. (iii) *Minimize sounding air-time uti-
In this policy, only one feedback slot is allowed regardless of the number of successfully decoded (concurrent) streams. Similarly to the basic operation, the AP can implement a two-round user selection. Notice that retransmissions are allowed in the first two policies in case the AP fails to decode at least one stream. A retransmission would allow other users to “piggyback” and join the uplink transmission on the next feedback slot (thus increasing user diversity). While in our evaluation we let the AP chooses a random set of users to join the retransmission, the user selection algorithm presented in [60] could improve the process by enabling the AP to report vectors orthogonal to the already sounded users, and allowing best candidates to reply over that next slot. Nonetheless, notice that this would lead to a short contention period thus leading to an increase in sounding air-time.

4.3.4 Overhead Analysis

Feedback Overhead. The 802.11ac sounding procedure requires one beamforming report and one report poll per each user (except for the first user which replies after receiving the NDP, without poll). In contrast, CUIC can use only a single feedback slot for the different sounded users to reply with their beamforming reports. In the case of 802.11ac, if a 4-antenna AP sounds only two single-antenna users in a 20 MHz channel, the sounding procedure takes approximately between 631 µs and 727 µs, depending on the number of bits used for matrix quantization (we assume the beamforming reports are sent at QPSK with $\frac{1}{2}$ coding). Evidently, this number is much higher when considering more than a single user. Considering that an 802.11ac device can transmit at 433 Mbps, during that same amount of time the AP could transmit an extra 13 kB, thus making this process very expensive [2]. The following expression represents the amount of time needed for sounding,
given that $S$ users will be served in the next MU-MIMO transmission.

$$T_{802.11ac} = t_{NDPA} + (2 \cdot S + 1) \cdot t_{SIFS} + t_{NDP}$$
$$+ S \cdot t_{report} + (S - 1) \cdot t_{poll}$$

CUiC removes the need for multiple SIFS, reports, and polls. Moreover, it relies on the 802.11ac Long Training Field (VHT-LTF) from each of the staggered preambles to do channel estimation, where each VHT-LTF is 4 $\mu$s long. Thus, to minimize additional overhead in CUiC, instead of staggering the entire preambles of all received reports, we could stagger only a smaller portion of them, e.g., the VHT-LTF of each user, which allows a 4 $\mu$s gap between every two non-overlapping symbols, at the expense of triggering AGC only once. Therefore, the overhead increase incurred in CUiC could be as low as $4S$ $\mu$s, where $S$ is the number of users to be served concurrently. Consequently, the overall sounding overhead required by CUiC reduces to

$$T_{CUiC} = t_{NDPA} + 2 \cdot t_{SIFS} + t_{NDP} + t_{report} + S \cdot LTF$$

**Acknowledgements Overhead.** With respect to the acknowledgment process in MU-MIMO, in the best case CUiC removes the need for any block ACK request and any subsequent SIFS. That is, it reduces the transmission time required for the acknowledgment process by about 120 $\mu$s and 60 $\mu$s for each BAR (at 6 and 24 Mbps, respectively) and by 16 $\mu$s for each SIFS. Additionally, it can reduce the number of transmission slots used to send compressed block ACKs (BA) from one per user to a total of only one, i.e., reduces the total time spent transmitting BA’s from 512 $\mu$s to 128 $\mu$s at 6 Mbps, and from 248 $\mu$s to 62 $\mu$s at 24 Mbps (parameters used: 40 $\mu$s PLCP, 34 bytes for MAC header and FCS, 26 bytes for BAR, and 32 bytes for BA). By saving that amount of time, in a 4-user system (at 20 MHz, 64-QAM $\frac{8}{6}$), the AP could otherwise transmit up to an additional 13 kB. No-
tice that CUiC incurs in a minor overhead of at least 4 $\mu$s for each staggered preamble (or VHT-LTF) transmitted, i.e., one for each user.

In summary, for both the feedback and acknowledgment processes, instead of requiring $S$ transmission slots for all $S$ feedback reports/ACKs and $S - 1$ polling frames, plus $2S + 1$ SIFS ($16\mu s$ each), our scheme only requires one transmission slot and an almost negligible overhead of $4S \mu s$.

### 4.4 Experimental Evaluation of CUiC

In this section we present an experimental validation and evaluation of CUiC under a wide variety of indoor WLAN scenarios. Our investigation focuses on (i) reliability of our CUiC demultiplexer implementation, (ii) overhead reduction of CUiC compared to the sounding and ACK procedures in 802.11ac, and (iii) throughput performance of CUiC.

**Experimental Methodology.** We implement CUiC in the FPGA-based WARP platform [1] and perform over-the-air (OTA) experiments, channel emulator based experiments, and trace-driven emulation, to validate and evaluate our system implementation. First, to validate our schemes we perform controlled experiments using a channel emulator and trace-driven emulation thus allowing us to independently tune different variables at a time. Next, we perform OTA transmissions to investigate performance gains in real channel conditions. We implement an OFDM 802.11-based physical layer and use the 2.484 GHz band, i.e., channel 14, which is not in use in our experimental region, therefore limiting the amount of out-of-network interference affecting our measurements.

**WARP, WARPLab, and Channel Emulator.** The WARP platform consists of an FPGA-based software defined radio, interfaced with custom designed radios based on the MAX2829 chipset. This platform allows for the implementation of clean-slate PHY and MAC protocols. WARPLab is a programming environment that integrates MATLAB tools
with WARP to control the platform via a host computer for running experiments and collecting data. To perform experiments under controlled and repeatable channel conditions we utilize the Azimuth ACE 400WB channel emulator, which supports 4x4 channel configurations [26].

4.4.1 Decoding Reliability of CUiC

Relative Signal Strength. The performance of spatial demultiplexers is highly dependent on the difference between the power of the signal to be decoded next, and the interference plus noise components of the composite signal. Notice that at each stage in the CUiC decoder, only one stream is of interest, whereas the superposition of the rest of the streams is considered interference. Therefore, to determine the SINR at which the intended stream can be reliably decoded with our 4x4 CUiC decoder, we investigate the system’s performance at different SINR values.

To this end, we perform a controlled experiment using the Azimuth channel emulator. WARP boards connect to each 4-input/output RF port, and to the host PC that manages the emulator and experimental settings, as well as data collection. We follow the channel model employed in [5], which consists of a 9-tap Rayleigh fading channel with a delay per path profile going from 0 to 80 \( \text{ns} \) and a path loss per path having a range from 0 to 22 \( \text{dB} \). We vary the attenuation of the intended user’s signal and the three other interferers. We transmit one thousand 802.11 frames and determine if each of them was successfully received after our decoding process.

Figure 4.3 (left) depicts the control frame error rate (CF-ER), defined as the error rate for sounding feedback control frames transmitted on the uplink, for different SINR levels. We compare the performance of our MMSE-SIC demultiplexer to a ZF-based demultiplexer. First, the control frame error rate of our system decreases rapidly with increasing
SINR. Notice that as the SINR reaches 0 dB, performance can reach near zero CF-ER. Additionally, MMSE-SIC outperforms ZF specially for low SNR regimes as expected. Therefore, we conclude that our MMSE-SIC based CUiC decoder, can reliably decode at least one stream in a 4x4 system even for SINR below 0 dB.

![Graph showing control frame error rate vs SINR for MMSE-SIC and ZF decoder](image)

**Figure 4.3**: (Left) CUiC Decoder: Control Frame Error Rate (CF-ER) vs SINR for MMSE-SIC CUiC and ZF decoder. (Right) Control frame error rate of CUiC and 802.11ac in real WLANs.

**Decoding Performance in Real Indoor WLAN Scenarios.** Next, we examine the ability of CUiC to decode concurrent streams under different scenarios in a realistic indoor WLAN environment. Additionally, we investigate the benefit of leveraging additional antennas at the AP for increased receive diversity, and robustness. We deploy four single-antenna users and one AP with four antennas in a medium size conference room. Then we evaluate the control frame error rate performance of CUiC in static and mobile conditions. We perform five transmission runs for a different network configuration or mobility pattern.
For the static case, the position of each node at each run is different, but remains fixed for the duration of each run. For the mobile case, at each run we arbitrarily move the position of the AP following a different pattern, thus varying the speed, acceleration, and distance to the users. Nonetheless, the maximum speed reached is walking pedestrian speed. In both cases, i.e., static and mobile, environmental mobility due to dynamic scatterers is present.

Figure 4.3 (right) depicts the average control frame error rate and standard deviation in concurrent uplink transmissions, over each different configuration (topology), for both static and mobile conditions. 802.11ac feedback exhibits the most reliable behavior due to the fact that it transmits each packet individually and sequentially via base-rate SIMO (Single-Input Multiple-Output). Namely, to ensure a fair comparison, we enhance our 802.11ac feedback implementation with a linear combining technique to exploit diversity gains by leveraging all AP antennas. However, for static conditions, the control frame error rate for CUIC is below 10%. In a following section we demonstrate that even considering the higher reliability of sequential SIMO feedback as in 802.11ac, CUIC is superior when overhead is considered. More importantly, observe that there is a fast improvement in performance as the number of user transmitting simultaneously decreases compared to the number of receiving antennas at the AP. The reason for this is a combination of increased receiver diversity and a reduction in sources of errors due to timing misalignments and synchronization mismatches.

**Antenna Diversity as Robustness Enhancer.** As shown in Figure 4.3, with additional number of antennas at the AP compared to the number of concurrent users ($M \geq S + 1$), the system provides with a substantial decrease in CF-ER. One of the key factors driving this result is the increase in diversity gain provided by having more receiving antennas than aggregate transmitting antennas. We demonstrate how much this factor influences the performance of our system by quantifying the CF-ER for a fixed number of users and
The experiment consists of four single-antenna users (fixed number of users) simultaneously transmitting to an AP having between four and eight antennas. We study 10 different indoor topologies with arbitrary node location and transmit more than one thousand packets at each topology. Figure 4.4 depicts the percent CF-ER as a function of the number of antennas at the AP. Notice the clear decrease in error rate as the number of antennas changes from four to eight, i.e., from 22.8% error rate to 7.6%, respectively. More interestingly, observe the sharp decrease from about 22.8% to 12.6% error rate caused by adding only one additional antenna. Thus, the figure shows that with only one additional antenna at the AP we can reach a large portion of the gains attained by adding as many as four additional antennas. Moreover, the 44% reduction in error rate attained by adding one antenna is very close to the average improvement of about 42% observed in the previous plot in Figure 4.3 (when going from 4x4 to 4x3), meaning that out of all the probable causes leading to such improvement when the number of users in the system decreases, diversity gains provide the most substantial advantage.

**Channel Correlation and Relative Node Positioning.** Channel vector correlation and the spread in signal strengths among concurrently transmitting users are considered factors that influence successful decoding of a composite stream. More specifically, low correlation between the channel vectors of the different concurrently transmitting users allows the AP to better separate the streams after computing and applying either the MMSE or ZF weights. On the other hand, the relative signal strength of all users (influenced by the diverse node positions) affects the ordered SIC process by allowing the AP to first decode the stronger and more robust user therefore reducing the amount of errors propagated from one stage to the other. At the beginning of this section, via a controlled experiment we observed that the difference in power between the signal to be decoded next and the inter-
fering streams is of high importance for reliable frame decoding. However, the system’s performance in a real deployment where the AP continuously computes the post processing SINR of each user and determines which one of them should be decoded next, remains unknown.

To study both the impact of channel correlation and signal power variation among users on the ability of the AP to decode the different signals, we deploy a 4x4 system and investigate the channel orthogonality among the four users and the group’s signal power variance. In particular, to determine channel orthogonality of the users, we compute the infinity norm condition number of the channel matrix $H$, i.e., $C_\infty(H)$ [60]. For every SDMA transmission we record the number of failed frames and the infinity norm condition number averaged over all 48 data subcarriers, as well as the variance in signal power among all users. We evaluate the system for a total of 24 indoor topologies with arbitrary node positioning. Figure 4.5 (left) shows the cumulative distribution function of the matrix condition number

Figure 4.4: Performance improvement due to receiver diversity gain.
for all five different cases of frame failures. Surprisingly, there is no clear distinction between the different cases except when there are no frame decoding failures. Nonetheless, as expected, the case of no failures shows lower condition number with a higher probability compared to the rest of the cases. Similarly, for the variance in signal power, there is no evident distinction between all the cases. This means that in a real indoor environment, node positioning and channel correlation do not have a significant impact on the AP’s ability to decode the different frames. These results have important implications on the design of user selection and power control protocols and algorithms.

Figure 4.5: (Left) CDF of channel matrix infinity condition number. (Right) CDF of variance in signal power among concurrent users.
Figure 4.6: (Left) Channel autocorrelation and TX trigger for CUIC and 802.11ac-based SIMO feedback. Figure’s legend in text below. (Right) Over-the-air experiment showing overhead impact on packet transmissions for 64-QAM 5/6 coding, 20 MHz channels with 400 nS guard intervals (basic CUIC policy).
4.4.2 Improving Sounding Resilience in Dynamic Scenarios

By shortening the amount of time between sending the sounding frames and triggering a downlink MU-MIMO transmission, CUiC can also increase the robustness of the system in the presence of fast moving channels. In this section we demonstrate how much protection our scheme can provide compared to the sequential SIMO feedback of 802.11ac-based sounding. To this end, we plot the channel correlation as a function of time for different channel profiles and identify the difference in channel variation between the sounding time required in CUiC vs. that required in 802.11ac. Figure 4.6 (left) depicts the channel correlation coefficient as computed in [55] for static (black) and mobile (red) nodes. Solid lines represent the averages over thousands of realizations. Additionally, for illustration purposes we show individual realizations with dotted lines. Assuming that sounding occurs at time $t = 0$, the first (green) vertical line shows when a downlink transmission is triggered with CUiC. On the other hand, the second vertical line (blue) shows when the downlink transmission is triggered in 802.11ac. That is, the time between $t = 0$ and the green and blue lines depict the amount of time spent doing sounding in CUiC and in 802.11ac, respectively. These sounding times were computed for 20 MHz channels, maximum subcarrier grouping of four, minimum quantization bits, and base rate. Notice however that for other parameters, the 802.11ac sounding time can reach close to 1 millisecond [2, 9, 10].

Observe that for several dynamic profiles the difference in channel correlation is significant. For some users, the coefficient drops from about 0.7 to a value below 0.4. Therefore, for many channel profiles, the amount of time saved via CUiC can make the difference between receiving a subset of the frames transmitted vs. not being able to receive anything.
4.4.3 Achieving Constant Sounding Overhead with CUiC

CUiC’s objective is to maintain constant overhead even as the number of users to be sounded increases (in contrast to current implementations, e.g., 802.11ac). We investigate the amount of overhead that CUiC can suppress compared to 802.11ac for two variations of our scheme, i.e., with and without power control.

**Power Control.** Without an adaptive AGC algorithm, large differences in the power of the multiple signals can lead to reception issues such as ADC saturation. Although we reserve the design of such adaptive system for future work, in addition to our original scheme CUiC, we also present an idealized protocol where receive gains are tuned manually so as to guarantee that all signals fall within the ADC’s dynamic range. We denote such scheme as CUiC-PC (Power Control), where retransmissions due to saturation are mostly non-existent. The idealistic scheme provides with a notion of the impact of power control on the overhead reduction that CUiC can attain.

We perform an extensive set of over-the-air (OTA) experiments in a rich scattering indoor office environment at daytime, where all sounded users transmit simultaneously to the AP with their channel estimation feedback. Based on the number of users sounded, and the total number of retransmissions we calculate the amount of overhead involved in the sounding procedure using the 802.11ac timings for each sounding packet (i.e, NDPA, NDP, beamforming report, report poll, and packet duration) [3]. Using these timings we compute the fraction of airtime consumed by sounding overhead, out of the total transmission time, and compare to the SIMO based feedback in 802.11ac. We deploy thirty different topologies and for each topology we transmit over one thousand 802.11 frames.

Table 4.1 shows the percent reduction in overhead for each of the evaluated schemes. For the 4x4 case, we achieve nearly 70% reduction, however, as we consider systems with fewer users these gains decrease due to the lower amount of overhead incurred in 802.11ac-
based systems. Furthermore, we explore how this overhead reduction translates into performance gains when considering data packet transmissions. Figure 4.6 (right) depicts the fraction of airtime consumed by sounding overhead for the cases where two, three, or four users reply simultaneously with their beamforming reports. We plot the averages over the thirty different topologies for 802.11ac, and both CUiC and CUiC-PC. Observe that while the overhead in 802.11ac increases rapidly going from 31% to 47% (two users vs. four users), for the concurrent uplink schemes this overhead stays almost constant by going only from 21 to 22%. In conclusion, even without user transmission power control, CUiC can maintain an almost constant overhead as the number of users increases from two to four.

<table>
<thead>
<tr>
<th>Table 4.1 : Sounding Overhead Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4x2</td>
</tr>
<tr>
<td>CUiC</td>
</tr>
<tr>
<td>CUiC-PC</td>
</tr>
</tbody>
</table>

4.4.4 Efficient Acknowledgement Process With CUiC

To investigate the potential gains attained via concurrent block ACK transmissions, we implement a complete acknowledgement process and evaluate its performance in different topologies. We use similar procedure and methodology to that employed in Section 4.4.3. However, in this case we evaluate the performance using two different base rates for both BAs and BARs, i.e., 24 and 6 Mbps. Figure 4.7 (left) depicts the percent reduction in acknowledgement process time compared to 802.11n/ac systems utilizing these two base rates. Similarly to the sounding procedure, we allow BA retransmissions whenever there is
Figure 4.7: (Left) CUiC’s achieved time reduction in the MU-MIMO acknowledgement process. (Right) Throughput gain of CUiC policies compared to 802.11ac (20 MHz channel, subcarrier grouping of 2, and feedback quant. of $\phi = 9$ and $\psi = 7$ bits).

Figure 4.8: Aggregate throughput of CUiC policies compared to 802.11ac (20 MHz channel, subcarrier grouping of 2, and feedback quant. of $\phi = 9$ and $\psi = 7$ bits).
a failed transmission. Observe that even with retransmissions, 4x4 CUíc is able to achieve
close to 44% gains and about 25% gains for 6 and 24 Mbps, respectively. Therefore,
in a similar way to the sounding process, these gains contribute to an overall system-level
efficiency improvement of a complete MU-MIMO system compared to legacy 802.11 SISO
ones.

4.4.5 Throughput of CUíc Policies

Finally, we evaluate the throughput performance of each of the proposed policies and report
their throughput gains over 802.11ac. We consider a system comprised of a 4-antenna AP
and 8 single-antenna users (deployed across more than 30 topologies to emulate a larger
number of users). Due to hardware constraints, our system cannot meet the fast 802.11ac
timing constraints. Consequently, to ensure fair comparison of our policies and the bench-
mark we rely on a combination of uplink OTA transmissions and downlink MU-MIMO
emulation to perform this evaluation. For the uplink we implement all four CUíc policies
on WARP and record all channel estimates for all transmissions, as well as the number
of successful/unsuccessful user transmissions, and retransmissions at each instance. Then,
using the collected channel estimates we compute the achievable rate [10] at each trans-
mission, for each policy, which we use to compute packet and sounding durations. Thus,
our throughput evaluation considers accurate 802.11ac timings, real uplink performance of
CUíc, and theoretical downlink ZFBF MU-MIMO rate based on collected (OTA) channels.
Moreover, channels are assumed to remain static between sounding and data transmission,
for all schemes. This emulation allows us to replay the same channels thus enabling exper-
iment repeatability.

In Figure 4.7 (right) and Figure 4.8 we show that CUíc significantly and consistently
outperforms 802.11ac for a wide range of packet sizes, i.e., from single packet to frame
aggregation of 64 frames. For shorter packet lengths, the minimum sounding (mSo) policy performs best because it only requires one transmission slot to sound all users and it can serve users much faster. Therefore, as long as the control frame error rate remains low, the best strategy is to truncate the sounding process to the minimum number of slots to maintain constant overhead. Nonetheless, notice that as the number of aggregate frames increases, the maximum diversity (mDiv) strategy outperforms the rest. That is, with higher overhead amortization it is best to utilize the same number of sounding feedback slots as 802.11ac but collect channels from as many users as possible, thus improving user diversity and consequently, increasing throughput.

4.5 Chapter Concluding Remarks

In this chapter we present the design, implementation, and evaluation of a novel MU-MIMO scheme that enables the use of concurrent uplink transmissions by multiple users in order to simultaneously send control and management frames to the AP. More specifically, we present CUiC, a scheme that addresses the high sounding overhead of 802.11ac systems and the inefficiencies of the acknowledgement process in MU-MIMO networks. CUiC implements a concurrent uplink feedback approach to reduce the amount of time spent feeding back channel estimates from the users to the AP as well as block ACKs to acknowledge frame reception. We demonstrate that CUiC can achieve near 67% sounding overhead reduction as well as close to 45% reduction in the amount required for 4 users to reply with their ACKs.
Chapter 5

Feasibility of Sounding Suppression

This work has been previously published by IEEE in [10]

5.1 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) to 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 [63]. Additionally, it is advantageous to acquire CSI from all associated users in order to maximize user diversity, thereby improving the user selection process by increasing the likelihood of finding beamformees with 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

---

*In this chapter we use CSI to refer to CSI at the transmitter, i.e., CSIT.
†We define user diversity as the accommodation of a finite set of users with distinct channel characteristics.
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 chapter 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.

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 [5, 6, 13, 45]. 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 chan-
nels 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; (ii) idle sounding which exploits idle channel intervals to opportunistically sound users to constantly update channel measurements to every user; (iii) bootstrap sounding which takes advantage of the required association procedure between the user and AP to provide the latter one with initial channel measurements towards each new user.

In particular, our main contributions in this chapter are the following:

First, 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.

Second, 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.

5.2 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.

5.2.1 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 [5, 6, 13, 45]. However, performing the costly sounding operation described in Chapter 3 with high frequency
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 highly 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 CSI statistics to 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.

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 \( S = M \) 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 it wants to serve. Additionally, if no channel statistics are available, a decoupled system could sound all \( S \) 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 \( S \) sounded users without discarding any, even if
their combination leads to poor ZFBF performance (e.g., high inter-user channel vector correlation).

5.2.2 Sounding Set Selection Procedure

The sounding process in MUTE gradually reduces the number of users to be sounded before every transmission 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.

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 two consecutive soundings 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 historical measurements.

Data Collection. The AP obtains channel measurements from every sounding procedure, i.e., magnitude $r$ and phase $\theta$ of the complex entries in the channel vector fed back from each user. Then, the it calculates the absolute magnitude and phase difference be-
tween 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 ($1 \leq j \leq i$). This process allows the AP to estimate how much each older measurement has degraded compared to the current channels. That is, how inaccurate is the older measurement representing the current conditions of the channel. Notice however that phase wraps around and this needs to be taken into account. Moreover, part of our ongoing work is focused on exploring other methods of dealing with this issue such as using the relative phase considering the different antennas at the AP, instead of the absolute phase. This would lead to much higher accuracy. For this reason, the performance of MUTE is limited by how much the channel changes, and performs better when there is little to no movement in the environment.

After sounding any user, the AP records the following information:

$$(t_i, \text{age}_{i,j,k}, \delta_{r_{i,j,k}}, \delta_{\theta_{i,j,k}}),$$

where $\text{age}_{i,j,k}$ 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. (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 4.4);

To determine the sounding set, before the next ZFBF transmission the AP selects a set $\mathcal{S}$ of users in the service queue (where $|\mathcal{S}| \leq M$), and decides which subset $\mathcal{S}'$ of them to sound. Notice that MUTE is implemented independently of the scheduling scheme employed, therefore, determining $\mathcal{S}$ is out of the scope of this thesis. Then, MUTE obtains the set of Relevant Samples satisfying any of the following constraints: (i) samples that were recorded in the last $\tau_{\text{recent}}$ milliseconds. In this case, MUTE accounts for the most recent samples; and (ii) samples for which their sample age is within $\pm \tau_{\text{age}}$ 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 $t_{\text{now}}$ and the time of the last recorded sample $t_{\text{now}-1}$. 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 and for the case of the magnitude change it is denoted by $\Delta_r$.

MUTE then computes a weighted mean $\mu^*$ and variance $\sigma^2$ 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., $\beta$ for each of the most recent samples, and $(1 - \beta)$ for each age-based sample, where we expect $\beta$ to be higher than 0.5 due to the relevance 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 $\sigma^2$ of each transmit/receive path is compared to thresholds $\sigma^2_{r\text{Thresh}}$ and $\sigma^2_{\theta\text{Thresh}}$ 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
statistics inaccuracies and overhead reduction. The values we set for these thresholds are
determined experimentally in Section 4.4. Notice that the penalty in rate due to lack of
accuracy of older estimates can be controlled via these thresholds; however, overhead re-
duction 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.

In the algorithm we compute the weighted mean $\mu^*$ and variance $\sigma^2$ based on $\mu^*$. Since
in certain cases the protocol might be dealing with small sets of samples, we rely on an
unbiased estimator of a weighted population variance. Thus, we let $\Lambda_1 = \sum_{i=1}^{n} w_i$, and
$\Lambda_2 = \sum_{i=1}^{n} w_i^2$.

Idle Sounding

Traffic in WLANs has been shown to be highly bursty [30]. 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
Algorithm 1 MUTE’s Sounding Procedure

1: while (1) do
2:     Initialization; $\Phi_{\Delta r} = \emptyset$; Get $S$ (where $|S| \leq M$)
3:     if $|S| > 0$ then $\triangleright$ At least one user to be served
4:         for $\forall$ Tx/Rx antenna pair and $\forall$ users $s \in S$ do
5:             for $\forall$ $i$ and $j$ do $\triangleright$ Select relevant samples
6:                 if $t_{now} - t_i \leq \tau_{recent}$ then
7:                     $\Delta_r \leftarrow \beta \times \delta r_{i,1,k}$
8:                 end if
9:                 if $|(t_{now} - t_{now-1}) - age_{i,j,k}| \leq \tau_{age}$ then
10:                    $\Delta_r \leftarrow (1 - \beta) \times \delta r_{i,j,k}$
11:                end if
12:            end for $\triangleright$ Compute weighted mean and variance
13:            $\mu^* = \frac{1}{n} \sum_{l=1}^{n} w_l \Delta_{r_l}$
14:            $\sigma^2 = \frac{1}{N_{\text{c}} - N_1} \sum_{l=1}^{n} w_l (x_l - \mu^*)^2$
15:            if $\sigma^2 \geq \sigma^2_{\text{thresh}}$ then
16:                $\hat{S} \leftarrow \text{User } s \in S$
17:            end if
18:        end for $\triangleright$ Sound users in $\hat{S}$
19:    end if
20:    Run user selection procedure, serve users $s \subseteq S$
21: end while
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, it 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.

**Bootstrap Sounding**

In any WLAN, users undergo a process of association and authentication when connecting to an AP. Therefore an initial sounding to every single user could be performed at the same time as the association without requiring significant additional overhead to do so. That is, in the context of 802.11ac sounding mechanisms, since the AP only needs to sound one user at a time, then the NDPA or polling are not strictly necessary. This means that the AP only needs to transmit the NDP packet for channel sounding and wait for a single compressed beamforming report. Depending on the number of bits used for quantization of the angles required for generating the feedback matrix $V$, this process could take approximately 150 $\mu$s longer than association times in current WLANs.
5.2.3 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 [63], matrix collinearity, and condition number [19]. 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.

5.2.4 Packet Loss & Demotion to SU-MISO for Fast Channels

MUTE operates in a wide range of regimes going from largely stable channels to channels changing as often as every consecutive sounding procedure. For instance, in channels where the coherence time is shorter than the time between two consecutive sounding procedures, the AP cannot rely on collected channel statistics, but it will force a sounding to those users before the data transmission. Unfortunately, if the channel to a user changes between sounding and the ZFBF transmission, the beam weights will become outdated and will not correspond to the actual channel observed at the time of the beamforming transmission.

To ensure that our algorithm avoids performance losses caused by those users characterized by such highly varying channels, 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 a single packet loss for user $k$, we require the AP to sound that user before the next transmission to it. Furthermore, a 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 by means of MU-MIMO transmissions. That is, 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 become more stable at some point, 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 $\sigma^2_{Thresh}$, 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 a bit of information in a standard 802.11ac packet to enable this mechanism.

5.3 Implementing and Evaluating MUTE

MUTE’s performance and gain, 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 frequently the channel varies. As discussed above, MUTE chooses the sounding frequency based on channel dynamicity, ultimately striking a highly profitable balance be-
tween 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 - in total we deployed more than 30 different setups. We are the first to experimentally study this issue in MU-MIMO. Our key finding is that under common channel conditions, channel age of few hundred milliseconds to few seconds 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 transmit set selection; we conclude the section showing that user selection schemes can highly benefit from larger sets of transmission candidates.

5.3.1 Experimental Setup and Evaluation Methodology

We have implemented a ZFBF MU-MIMO transmission scheme in WARP [1] and performed all our experiments using the WARPLab framework. 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 use over-the-air channel measurements collected 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 (5.1) 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 (5.1) [54].

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

(5.1)

For more details on how to compute the beam steering weights, refer to Chapter 3.

One big challenge in using this expression in our emulator is to match the measured (experimental) channel gains to those used in the analytical expression above. To achieve this, we run a calibration procedure just before performing all the channel data collection. This calibration procedure consists of first computing the SNR between each transmitter and receiver antenna (each individual path) as well as the corresponding channel gain. Notice that our computation of SNR is based on the measured Receive Signal Strength Indicator (RSSI) for both the intended signal and the noise, and normalizing by the receiver gain setting. Then, we compute a scaling factor \( \alpha = \frac{\text{SNR}_{k,m}}{|h_{k,m,\text{cal}}|^2} \), where \( \text{SNR}_{k,m} \) represents the signal-to-noise ratio between AP antenna \( m \) and single-antenna user \( k \) (in linear scale). Similarly, \( h_{k,m,\text{cal}} \) is the complex scalar representing the channel gain between each antenna element. This is the channel gain measured during the calibration phase. Finally, the
channel entries used as input to Equation (5.1) are given by

\[ h_{k,m} = \sqrt{\alpha} \cdot \hat{h}_{k,m} \]  

(5.2)

where \( \hat{h}_{k,m} \) represents each of the measured channel gains during our channel data collection (subsequent to the calibration process).

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 presented in \([5,45]\)). 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 5.1 we present the per-user 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.

**Inferring Per-User MUBF SINR.** Predicting or inferring the SINR at each user during the downlink beamforming transmission is important to determine the bit rate (or modulation and coding scheme, in the context of 802.11) required for transmission. In single-stream systems, knowing the individual user’s SNR or channel vector is enough to select a given bit rate (e.g., SNR-based rate selection \([14]\)); however, in MU-MIMO the SINR at each user depends on the channel vectors between the rest of the concurrent users and the AP.
Using empirical measurements from our testbed we compare the accuracy of employing individual SNRs for each user against the joint SINR (considering the entire channel matrix) obtained from Equation (5.1). To this end, we gather measurements in an indoor environment for more than 20 client locations in both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS). Moreover, we vary the number of antennas at the AP as well as the number of concurrent users, i.e., $M$ and $S$, respectively. Notice that for this experiment we let $M = S$.

In Figure 5.2 we plot the condition number of matrix $H$ against the absolute SINR difference between the measured SINR at a particular user during beamforming, and both $(i)$ the inferred SINR based on the joint channels to all users (prior to beamforming) and $(ii)$ the individual SNR measured at a particular user prior to the beamformed transmission.
The condition number of matrix $\mathbf{H}$ is computed as follows:

$$\kappa(\mathbf{H}) = \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \geq 1 \quad (5.3)$$

where $\sigma_{\text{max}}$ and $\sigma_{\text{min}}$ are the largest and smallest singular values of $\mathbf{H}$, respectively. When considering ZFBF, the channel matrix conditioning is relevant to determine how sensitive $\mathbf{H}$ is to matrix invertibility. Therefore, if $\mathbf{H}$ is ill-conditioned (large condition number), then the zero-forcing scheme will lead to a high amplification of the noise, thus leading to a lower per-user SINR.

Notice in Figure 5.2 that while the difference in SINR is much higher when considering only the individual SNR compared to the inferred SINR, the performance obtained by relying only on such inference can be rather poor. That is, none of these two techniques are satisfactory to accurately determine the beamforming SINR, specially with increasing matrix condition number. In Chapter 6 we propose a mechanism to address this issue.

![Figure 5.2 : Inferred SINR vs. Condition Number](image-url)
5.3.2 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 small rate penalties. Accordingly, when user channel varies infrequently, MUTE can systematically avoid sounding for long intervals and rely on information collected hundreds of milliseconds, 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 environment 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 transmitting 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 5.3 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 that LOS; note that even for the 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 small magnitude variations is negligible, while phase variations affect the amount of nulled interference, hence can result in much reduced SNR. In the following, we show the extent to which these variations affect the user achievable rates.

![Figure 5.3](image.png)

Figure 5.3: Average absolute magnitude and phase variation in static LOS and NLOS indoor environments. Error bars show 95% confidence intervals.

**Tradeoffs 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 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 $H_t$ at each sampling time $t$; then, for the *ideal case*, we obtain the ZFBF precoding weights $W_t$ and compute the rate according to Equation (5.1) using the same $H_t$, i.e., representing a transmission occurring over the same channel used to compute the weight matrix. For MUTE instead, we derive the ZFBF weight $W$ using $H_{t-age}$ and calculate the rate using channel $H_t$ and weight $W_{t-age}$,
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 $H_{t-age}$ differs from $H_t$, 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 5.4, 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.

![Figure 5.4: Relative error rate for different sample age.](image-url)
5.3.3 Throughput Gains in MUTE

MUTE performance benefits derive from reaching a profitable balance between sounding overhead and rate penalty. 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 $\sigma^2_{\text{rThresh}}$ and $\sigma^2_{\text{qThresh}}$ so as to allow MUTE to lose only about 2 bps/Hz, whereas on the second version we set them to allow only a 1 bps/Hz loss (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 Figures 5.3 and 5.4. 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).

Figure 5.5 (left) present average and standard deviation of the per-user rate achieved by MUTE and our 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 5.5 (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 trans-
lates 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 achieves a large sounding overhead decrease (55-95%) at the price of a rate penalty (as low as 7% for a 1 bps/Hz loss).

Finally, we investigate the throughput gain that MUTE can attain compared to the benchmark; this 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 802.11ac 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 5.6 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.

**MUTE Leverages User Diversity.** To maximize rate, in MU-MIMO user selection it is important to avoid grouping together users with correlated channels. Thus, the knowl-
Figure 5.5: Performance of MUTE under different scenarios. *Left* plot does not consider impact of overhead in rate performance.

Figure 5.6: MUTE’s throughput gain - 1 bps/Hz tolerance
edge of the channels of a large set of users (*user diversity*), i.e., ideally much larger than the number of users that are 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 large set of users, i.e., MUTE potentially leads to higher ZFBF rates at each given transmission. In this section, we show how a larger set of transmission candidates benefits the rate achieved by user set selection.

The effect of *user diversity* in MU-MIMO systems has been previously studied from an information theoretical perspective [63]. In this section, we isolate and explore this effect experimentally in order to quantify the gains that MUTE can allow by taking advantage of *user diversity*. To this end, 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 achievable, from a set of $n$ users uniformly selected among the 30 users population. Additionally, we compare against the exhaustive search approach that selects the combination that maximizes the rate among the whole of 30 users by choosing the best combination of $m$ users.

In Figure 5.7 we present the aggregate rate for both a $m = 1$ (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 selection 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 increasing number of users. In conclusions, 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.

![Figure 5.7 : User diversity in SU-MISO and MU-MIMO. The legend shown corresponds to both plots, however, notice that there is no data displayed for the 4x4 system below 4 users.](image)

5.3.4 Discussion: A First Look at MUTE in an OFDM System

Some time after introducing MUTE we were able to implement a 20 MHz OFDM beamforming system in WARPLab and performed some preliminary tests in an attempt to validate MUTE. Although our platform is in its early stages and does not operate in real-time,
we have tested it in an indoor office environment with low environmental mobility and static nodes. We have observed that we can successfully decode 16-QAM 2x2 MUBF transmissions even with time gaps between sounding and data transmission of 320 milliseconds. Figure 5.8 depicts the 16-QAM constellation received at the two different users. The figure illustrates more than 10 consecutive transmissions (overlapping symbols) with a 320 milliseconds gap between sounding and data transmission.

![Figure 5.8: 2x2 Constellation.](image-url)
5.4 Chapter Concluding Remarks

In this chapter 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. Our scheme 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.
Chapter 6

Interference Resilient Multi-User Beamforming WLANs

6.1 Introduction

Downlink Multi-User Beamforming (MUBF) is a key technique to scale throughput in dense WLANs as it enables an Access Point (AP) to simultaneously transmit multiple independent data streams to different users in the same frequency resource block.* Such multi-user transmission has been demonstrated in WLAN systems (e.g., [5, 7]), massive MIMO systems (e.g., [45, 62]), and is now standardized in IEEE 802.11ac [3, 9] and commercialized.

To achieve concurrent transmission, the AP precodes the independent streams by multiplying them by a beam-steering weight matrix in a way that reduces or removes inter-user or inter-stream interference. Such precoding requires knowledge of the channels between the antenna array at the AP and each concurrently served user. In protocols such as IEEE 802.11ac, this Channel State Information (CSI) is obtained via a sounding process in which predefined pilots are transmitted by the AP so that channel state is estimated by the receiver and fed back to the transmitter.

Unfortunately, client mobility, environmental mobility, and any source of precoding error (e.g., due to CSI feedback compression/quantization) can vastly degrade performance. In particular, imperfect beam steering does not merely result in a poorer quality signal at the receiver due to energy being directed away from the receiver: in a multi-user system,

---

*In this work we use MU-MIMO and MUBF interchangeably.
imperfect beam steering also increases inter-stream and inter-user interference, i.e., influencing both the signal $S$ and interference $I$ components of the Signal to Interference plus Noise Ratio (SINR) (the argument can be made rigorous via capacity analysis [66]).

In this chapter we present the design, implementation, and experimental evaluation of CHannel Resilient Multi-user bEamforming (CHRoME) and make the following contributions:

First, we propose M$^3$CS (Multi-user Multi-stream MCS), a technique for “just-in-time” multi-user bit-rate selection. In contrast to single-stream systems, multi-stream modulation and coding scheme (MCS) selection introduces the challenge of selecting multiple and potentially unequal MCS instead of just a single one. Schemes where the AP selects the MCS based on collected CSIT at the transmitter (CSIT) have been demonstrated to have strong performance in SU-MIMO systems [25]. However, we will show that in MUBF systems, if the AP selects the MCS for each stream based solely on collected CSIT [8, 43], performance will rapidly degrade with increasing mobility and estimation error. In principle, as mobility and other uncontrollable factors degrade SINR, the AP can maintain successful frame reception at the users by sufficiently reducing the MCS.

Our key technique is to make the selection as late as possible (immediately prior to data transmission) and by using a beamformed probe so that clients can assess the actual SINR of the beamformed transmission vs. the predicted SINR due to measurements of the channel training sequences. In this way, the AP can re-tune its selections accordingly, just-in-time for the downlink data transmission. We will show that with mobile environments, mobile users, or imperfect CSIT, the additional overhead introduced by the MUBF probing and feedback is far outweighed by avoiding rate under-selection (unnecessarily low MCS that wastes airtime) or over-selection (excessively high MCS that yields frame loss).

Second, we design and implement spare-DoF, a mechanism for mobile clients to in-
crease resilience to inter-stream inter-user interference by exploiting all “spare” degrees of freedom (DoF) to enhance signal reception. Unfortunately, in most scenarios, clients lack a sufficient number of antennas to perfectly cancel all interfering streams using classical techniques [41, 52]. However, we observe that a typical WLAN client may have a second spare antenna (or even more, but we use two antenna receivers for ease of exposition). Consequently, we design a scheme to enhance interference resilience, despite having only a single extra DoF by modifying the training preamble in the data frame so that each user can obtain an estimate of the single other user that causes it the greatest SINR degradation. Armed with sufficient information about the worst “culprit” in signal degradation, we emulate a classical two-user interference cancellation system. We show that even by ignoring all interfering streams other than the strongest one, CHRoME receivers vastly improve frame reception rate.

Third, despite the aforementioned resilience mechanisms, frames will occasionally be non-decodable due to excessive co-stream interference or mobility. Unfortunately, current retransmission strategies, inherited from the original CSMA design [28], require re-contention after a doubled backoff window. Consequently, physical layer parameters such as beam-steering weights are likely stale by the time retransmission is feasible, and therefore the time and resource penalty of channel sounding must be incurred again. In contrast, we design a soundless fast retransmission strategy in which the AP triggers a one-time immediate retransmission using the same CSIT as in the original transmission. Yet, because the original transmission failed, it is clear that the re-transmission strategy must be changed. Thus, because only a subset of the users’ transmissions will have failed, the retransmission will exploit the “liberated” degrees of freedom (e.g., an 8 antenna transmission to 8 users with 2 failed frames will have an additional 6 DoFs for the retransmission to 2 users). To avoid resounding the channel, we design a scheme in which the user’s block ACKs that fol-
low the failed transmission, piggyback a measurement of inter-stream interference obtained during the data transmission. With this hint, the AP can characterize the expected retransmission SINR, and reset beam-steering weights and bit rates such that they are sufficiently robust to enable reception despite the use of increasingly outdated CSIT.

Finally, we implement all three components of CHRoME on the WARP platform [1], and perform an extensive set of over-the-air experiments combined with trace-driven emulation. Our evaluation reveals that the MCS selection mechanism in CHRoME can achieve between 7% and 280% throughput gains under mobility and dynamic channel scenarios, compared to CSIT-based MCS selection schemes. Likewise, under non-ideal quantization, CHRoME can reach between 9% and 600% throughput gains. Similarly, the fast retransmission scheme in CHRoME outperforms 802.11ac by at least 53% in terms of throughput.

6.2 M³CS: Beamformed Probing for “Just-in-Time” MCS Selection

Multi-user Multi-Stream MCS, M³CS, assesses the channel and inter-stream interference affecting each user, just prior to the data MUBF transmission, and adapts each stream’s MCS accordingly. Modulation and coding scheme selection in MUBF is fundamentally different to the case of single-input single-output (SISO) systems. SISO transmitters typically rely on the SNR of previous packets as well as packet loss history to determine the best MCS to be used in the next transmission, i.e., SNR-based and packet loss-based algorithms [14]. Nevertheless, a MUBF AP cannot rely on individual SNR knowledge unless the channels to all users are completely orthogonal. Otherwise, any dependence among channel vectors to the multiple users would introduce an interference signal component. Similarly, using packet loss as a MUBF MCS indicator would require the set of concurrently served users to be the same for successive transmissions. Otherwise, the angle between the channel vectors of different user groups would lead to different MCS require-
ments.

Figure 6.1 illustrates the key difference between how single-stream and multi-stream systems are affected by errors in channel estimation or by environmental/user mobility. In particular, the figure shows that beamforming errors in the single-stream case merely result in a decrease in signal strength whereas in the multi-stream case they can also lead to an increase in inter-stream interference.

6.2.1 CSIT-Based MCS Selection

In order to accurately determine the most appropriate MCS for each individual user within a concurrent group, the AP needs to estimate the expected SINR with which the transmitted signals will arrive at each user during the data MUBF transmission. In the hypothetical case of the AP acquiring perfect CSIT (no quantization errors or channel variations), the SINR
for every user can be directly calculated as follows. Consider a narrowband channel model and a network comprised of a single AP with \( M \) transmit antennas and \( K \) users. Let \( \Upsilon \) be the set of transmit antennas at the AP (\(|\Upsilon| = M\)),\(^{†}\) and \( S \) be the set of users selected by the AP to be served in the next MUBF transmission, i.e., \( S \subseteq \{1, \ldots, K\} \), \(|S| = S \leq M\). Also, let \( \Gamma \) be the SINR matrix at the users and each entry \( \gamma_{ji} \in \Gamma \) be the power from stream \( j \) measured at user \( i \). \( W = [w_1 \ldots w_S] \) represents the precoding matrix applied by the AP to generate a single stream to each of the users in \( S \). Thus, given the collected CSIT, the AP computes the SINR of the intended stream at user \( i \) as follows:

\[
\text{SINR}_i = \frac{\gamma_{ji}}{N_0 + \sum_{j:j \neq i} \gamma_{ji}} = \frac{|\sum_{m=1}^{\Upsilon} h_{im} w_{mi}|^2}{N_0 + \sum_{j:j \neq i} |\sum_{m=1}^{\Upsilon} h_{im} w_{mj}|^2}
\]

(6.1)

(6.2)

where \( h_{im} \) denotes the complex channel gain between AP antenna \( m \) and user \( i \). Similarly, \( w_{mi} \) represents the complex weight applied to AP antenna \( m \) and \( N_0 \) is the noise variance at user \( i \). More specifically, via the sounding process, the AP learns the complex channel vector representing the path between all transmitting antennas and the users, i.e., channel matrix \( H \), and uses this information to compute the corresponding precoding weight matrix \( W \). As shown in Equation (6.2), this information is sufficient to determine the MUBF SINR given the current CSIT and current channel conditions.

**Limitations of CSIT-based MCS selection.** Selecting MCS based on CSIT alone has several drawbacks. First, Equation (6.2) assumes that channels between sounding and MUBF data transmission remain static, whereas channel variation will decrease SINR. Because a decrease as low as 2 or 3 dB in SINR requires a reduction in MCS, throughput can be severely degraded with such an SINR decrease. Feedback compression or feedback reduction schemes in which sounding does not take place before every data transmission

\(^{†}\)Symbol \(|\cdot|\) indicates set cardinality.
[10,61] are therefore particularly vulnerable, since channel variation over multiple packet transmissions is more significant.

Second, inaccurate CSIT estimation due to quantization or inter-stream interference will have a similar effect. Quantization primarily affects explicit sounding systems where users estimate the channel based on a training sequence transmitted by the AP, and then feed back a quantized version of these estimates (i.e., CSIT) utilizing only a small number of bits to limit overhead. Similarly, in implicit feedback systems where users transmit the training sequences and the AP estimates the channel based on this training, any other sources of interference (or noise) at the users will lead to inaccurate CSIT estimation because the channel measurements take place at the AP and not at the users. Consequently, this interference information is not considered in such estimates which in turn can lead to poor MCS selection.

6.2.2 MCS Selection via MUBF Probing

The combination of quantization errors, channel dynamics, and inaccurate information with respect to the noise and interference observed at each user, affect the performance of MCS selection schemes in MUBF systems. Therefore, the amount of inter-stream interference affecting each user, as well as the negative effects due to current channel conditions can only be known during the actual downlink MUBF transmission. We design a multi-user inter-stream interference probing mechanism that proactively evaluates the MCS selection resulting from predicting the per-user SINR based on the acquired CSIT. In particular, CHRoME proposes a multi-stage MCS selection scheme that probes the multi-user channels to evaluate the accuracy of zero forcing and adapts each stream’s MCS, just-in-time for data transmission. While the first two stages are dedicated to acquiring CSIT and to probing the channel to adapt the MCS for all users, a third stage consists of reporting back
this information to the AP. Figure 6.2 depicts the entire sounding, probing, and feedback process, i.e., all three stages.

![Diagram](image)

Figure 6.2 : MUBF probing and CSS feedback.

**Multi-User Inter-Stream Interference Probing**

Since CHRoME probes themselves are transmitted at a particular MCS, we use (necessarily sub-optimal) CSIT-based selection to set this initial MCS for each stream of the probe. Thus, using the most recent CSIT for each user $s \in \mathcal{S}$, the AP computes the beam-steering weight matrix and applies it to the independent data streams for the probe. The AP then triggers a multi-user probe by transmitting a minimum-length downlink multi-user frame at the rates determined using CSIT-based MCS selection. The multi-user probing frame enables each user to infer channel variations since sounding occurred, as well as the inter-stream interference affecting the transmission. Thus, upon reception of this probing frame,
each user measures its effective SINR ($\text{SINR}_{\text{eff}}$) and maps it to the corresponding preferred MCS. Notice that in the ideal case that channels are completely static and CSIT is perfectly estimated, the measured MUBF SINR corresponds to the MCS previously estimated by the AP via the CSIT-based MCS selection scheme. In contrast, in non-ideal cases, the AP can now adapt to the true conditions.

CHRoME is agnostic to the feedback mechanism implemented and can operate with implicit or explicit systems. Similarly, as shown in Figure 6.2 we do not make any assumption with respect to how frequently sounding occurs. That is, regardless of when sounding took place, the MUBF probing frame is triggered prior to the MUBF data transmission and utilizes the most recent CSIT collected for each user.

**Correlatable MCS Feedback**

In order for the AP to readjust the MCS according to the current channel conditions, each user needs to report back the computed MCS to the AP. Moreover, this feedback process needs to take place within the shortest time frame possible to minimize the overhead incurred by our system. The fact that MCS are identified with an index (0 to 9 in 802.11ac), means that we can represent each MCS selection with only a few bits. CHRoME maps each MCS index to a predefined pseudo noise binary codeword (i.e., a correlatable symbol sequence or CSS [34]). The transmission length and processing required to identify these sequences is significantly lower than what is required for decoding a packet, thus making them ideal for this application. More specifically, upon MCS selection at the users, they reply with a corresponding CSS. Figure 6.2 depicts both (a) the timeline (not to scale) showing where the MCS probing and feedback take place within a given MUBF transmission, as well as (b) a simplified representation of CSS usage.

**Signaling MCS with a CSS.** Broadly, correlatable symbol sequences are BPSK se-
quences that are filtered, up-sampled, and transmitted via wideband techniques. While CSS preserve the statistical properties of sampled white noise, cross-correlation of any CSS with a matching copy will produce a spike indicating a positive match. The advantages of CSS over decodable packets include higher detection reliability, higher robustness to radio parameter imperfections, and substantial transmission time reduction. More specifically, as demonstrated in [34], 127-symbol Gold sequences can be reliably detected at low SINR (-6 dB) with 5.7% false negatives and no false positives. Consequently, these can be detected at 10 dB lower compared to 6 Mbps OFDM frames. Moreover, CSS do not require a preamble or data processing thus reducing the amount of time needed for their transmission to only 6.35 µs [34].

CHRoME’s dictionary. 802.11ac features 10 different modulation and coding schemes indexed from 0 to 9. Given that 127-symbol Gold sequences allow 127 different sequences while retaining a low theoretical cross-correlation among them, these can easily support a mapping to 10 different MCS indexes. In order to support all 10 codewords the AP’s hardware could either use 10 simultaneous correlators (higher design complexity and cost), or buffer the received sequences and evaluate them sequentially one at a time (longer processing time).

Feedback Processing. CSS do not require data decoding. Out of the 16 µs that 802.11 allocates for SIFS, 14 µs are used for such processing and the rest for switching the radio modality (between TX and RX). Given that during CSS reception the AP does not need to switch from TX chain to RX chain, and vice versa, the transmission of consecutive CSS requires only up to 1 µs in between to account for signal propagation delay. Notice that the order in which users reply follows the order established in the sounding frames (e.g., Null Data Packet Announcement NDPA, in 802.11ac).
6.2.3 Incurred Overhead

The additional time required to trigger the probing and feedback mechanism in CHRoME can be broken down into the following components.

\[ T_{\text{overhead}} = \text{SIFS} + \text{Probe} + \text{SIFS} + S \cdot \text{CSS} + \ldots \]

\[ (S - 1) \cdot (1\,\mu s) \]

where SIFS and RIFS take 16 \( \mu s \) and 2 \( \mu s \), respectively. As previously mentioned, each CSS requires 6.35 \( \mu s \). Finally, the length of the probing frame depends on the minimum MCS used to transmit one A-MPDU to each user. That is, if four users are probed and three of them are served at MCS-3 (16-QAM, \( \frac{1}{2} \)) but the remaining one is probed at MCS-0 (BPSK, \( \frac{1}{2} \)), then the maximum time length is computed based on the MCS-0 transmission. We implement A-MPDUs as short as RTS packets. Therefore, the probing frame would take approximately 52 \( \mu s \) including preambles and assuming 6 Mbps transmission rate. For a 4-user system employing the lowest transmission rate (worst case), the total overhead \( T_{\text{overhead}} \) reaches a maximum of 112.4 \( \mu s \). Notice however that increasing the transmission rate of the probing frame would significantly decrease the total overhead.

6.3 SPARE-DoF: Inter-Stream Interference Mitigation and Recovery

\textit{Spare-DoF} obviates the full-antenna requirement necessary to cancel all interfering streams by analyzing each independent interference component and attempting to remove the \( N_s - 1 \) most harmful interferers.‡ In multi-stream systems, users encounter interference-limited channels where inter-stream interference becomes the major performance hinderer. More

‡We assume each user only receives one stream at a time, therefore \( N_s - 1 \) denotes the number of spare antennas at user \( s \).
importantly, the susceptibility of MUBF systems to inter-stream interference aggravates with increasing number of concurrent independent streams. As theory has shown, degrees of freedom available at the receivers can be leveraged to either increase the system’s diversity or to cancel interfering streams [52]. While in single-stream beamforming systems (characterized by noise-limited channels), a multi-antenna receiver highly benefits from the additional DoF to maximize SNR via receiver diversity, e.g., maximal ratio combining (MRC), in multi-stream systems interference cancellation becomes the desired strategy [41].

6.3.1 Receiver Strategy in CHRoME

While multi-antenna users can attempt to simultaneously decode their corresponding data streams as well as cancel the interference from other streams, their interference cancellation capabilities are limited by the number of degrees of freedom they possess. That is, a user $s$ with $N_s$ antennas can attempt to decode one stream while canceling up to $N_s - 1$ interfering ones. Consequently, in typical scenarios where the AP concurrently serves more users than the number of antennas at each of them, the composite signal – comprised of the intended stream and multiple strong interferers – cannot be separated, thus decreasing the likelihood of successful decoding.

The *Spare-DoF* mechanism identifies each individual interference component and cancels $N_s - 1$ of these streams in order, starting from the strongest component. Therefore, users face the challenge of not only training their receivers to decode their intended signals, but also of learning the channel corresponding to the rest of the streams while also identifying the most harmful ones. After canceling all possible $N_s - 1$ interfering streams, the remaining ones are treated as noise. We implement an ordered successive interference cancellation (O-SIC) scheme in which we first separate the multiple $N_s$ streams via
a Minimum Mean Square Error (MMSE) equalizer and then iteratively cancel the interfering symbols starting with the stream with strongest post-processing SINR first in order to minimize errors due to propagation [29, 52, 64].

### 6.3.2 Inference of Per-Stream Multiuser Interference

The MMSE-SIC receiver requires knowledge about the channel corresponding to each stream. In this subsection we present two alternatives to achieve this. First, we present the approach implemented in the current IEEE802.11 standard; and second, we present our own approach. While the former method allows each user to solve a system of equations in order to obtain the channel estimates of each stream, our method allows users to directly measure each of these channel components.

**Receiver Training Structure in 802.11n/ac**

The High Throughput (HT) and Very High Throughput (VHT) preambles mandated by the 802.11n and 802.11ac amendments, respectively, implement multiple pre-defined sequences that are prepended to the packet’s data field in order to train the receivers. In particular, the standard mandates the use of long training fields (LTF) to achieve this. These LTFs (HT-LTF and VHT-LTF) are necessary for each user to perform channel estimation (MIMO channel estimation in the case of multi-antenna receivers). Notice that in order for a user to obtain an estimate for each stream, the AP needs to prepend as many LTFs as the number of concurrent streams (except for 3, 5, and 7 streams where it requires 4, 6, and 8 LTFs respectively). An orthogonal mapping matrix is applied to the generated VHT-LTFs in order for users to detect its intended training sequence and remove the effects of the overlapping LTFs.

**Illustrative example.** Consider a 2x2 system where a 2-antenna AP transmits to two
2-antenna users (see the network in Figure 6.3). Let $h_{i,m}$ be the combination of the cyclic shift applied to the training symbols at the transmitter and channel tap for the $i^{th}$ receiving antenna in user $k$ and the $m^{th}$ spatial stream. Also, let $\xi$ be the frequency-domain VHT-LTF generated by the AP, and $y_{i,\tau}$ be the frequency-domain representation of the received signal (for each subcarrier) at the $i^{th}$ receiving antenna in user $k$ at time $\tau$. Notice that for simplicity we have removed the subcarrier index from all the variables. After processing the LTFs by removing the cyclic prefix and applying an FFT to extract the training subcarriers, the received signals can be represented as follows [36].

\[
\begin{align*}
y_{1,t_1} &= h_{11} \cdot \xi + h_{12} \cdot \xi + z_{1,t_1} \\
y_{2,t_1} &= h_{21} \cdot \xi + h_{22} \cdot \xi + z_{2,t_1} \\
y_{1,t_2} &= (-1)h_{11} \cdot \xi + h_{12} \cdot \xi + z_{1,t_2} \\
y_{2,t_2} &= (-1)h_{21} \cdot \xi + h_{22} \cdot \xi + z_{2,t_2}
\end{align*}
\]

where the $-1$ factor is introduced by the orthogonal mapping matrix specified in the standard, and $z_{i,\tau}$ represents the AWGN. Time $t_1$ and $t_2$ indicate the first and second symbol (VHT-LTF) transmission, respectively. By combining the subcarriers of the two consecutive symbols (for each antenna independently), we obtain the following:

\[
\begin{align*}
h_{11} &= \frac{y_{1,t_1} - y_{1,t_2}}{2 \cdot \xi} \\
h_{12} &= \frac{y_{1,t_1} + y_{1,t_2}}{2 \cdot \xi}
\end{align*}
\]

The same procedure is used to obtain the estimates at other receiver antennas.

**Receiver Training Structure in CHRoME**

We design a multi-user inter-stream interference training (ISIT) preamble that orthogonalizes the training symbols of each stream such that users can obtain a clean (interference-free) channel estimate for each signal component of the MUBF transmission. To guarantee
that there is no overlap among the symbols corresponding to different streams, we imple-
ment a subcarrier interleaving mechanism where the total number of subcarriers used for
training is divided among the number of streams. That is, each stream occupies a different
subcarrier in each LTF. Compared to multi-stream 802.11 training, our ISIT preamble al-

Illustrative example. Consider the simple 4x1x3 topology \((M, N_{ss{s}}, S)\) and frame
structure in Figure 6.3;\(^\S\) During the first LTF symbol \(\xi\), the AP uses subcarriers \([1:3:64]\) (64 subcarriers in 20 MHz channels) to beamform the training sequence corresponding to
spatial stream 1 (SS1), to user 1. Likewise, during the first LTF the AP uses subcarriers
\([2:3:64]\) and \([3:3:64]\) to beamform the training sequence corresponding to spatial streams
2 (SS2), and stream 3 (SS3), respectively. After the first LTF symbol has been trans-
mitted, user 1 has acquired not only the estimates corresponding to its intended stream (from
subcarriers \([1:3:64]\)), but also estimates of any interference component generated by SS2
and SS3, i.e., \(I_{2\rightarrow1}\) and \(I_{3\rightarrow1}\). That is, since the LTF is beamformed, those subcarriers in
which only SS2 and SS3 allocate any information should ideally be only sensed by users
2 and 3 but not by user 1. Any energy sensed by user 1 in those subcarriers is considered
inter-stream interference (regardless of how low the interference strength might be). How-
ever, notice that up until this point, user 1 has no knowledge about the channel of SS1 for
subcarriers \([2:3:64]\) and \([3:3:64]\). Therefore, on the second and third LTF symbols the sub-
carrier order is rearranged such that user 1 measures SS1 in these remaining subcarriers.
This process applies to all users/streams.

\(^\S\)\(N_{ss{s}}\) is the number of spatial streams intended for user \(s = 1, \cdots, S\).
Figure 6.3: MUBF ISIT preamble: example topology, subcarrier allocation and training timeline.

The received signal at the two antennas $A$ and $B$ in user 1 are given by:

\[
y_{1,A,t_1} = h_{1,A}w_1\zeta_1 + h_{1,A}w_2\zeta_2 + h_{1,A}w_3\zeta_3
\]
\[
y_{1,B,t_1} = h_{1,B}w_1\zeta_1 + h_{1,B}w_2\zeta_2 + h_{1,B}w_3\zeta_3
\]
\[
y_{1,A,t_2} = h_{1,A}w_1\zeta_2 + h_{1,A}w_2\zeta_3 + h_{1,A}w_3\zeta_1
\]
\[
y_{1,B,t_2} = h_{1,B}w_1\zeta_2 + h_{1,B}w_2\zeta_3 + h_{1,B}w_3\zeta_1
\]

where $y_{i,n,t_1}$ is the received signal at antenna $n$ of user $i$ during the first LTF symbol (i.e., during $t_1$); $h_{i,n}$ is the channel vector between all AP transmit antennas and antenna $n$ in user $i$. $w_j$ represents the beam-steering weight vector corresponding to data stream $j$ and $\zeta_{sc}$ denotes the set of subcarriers $sc$ of an LTF symbol, e.g., $\zeta_1 =$LTF(1:3:64). In this

---

*Third LTF omitted due to space constraints.*
example, all second terms represent the interference of stream 2 onto user 1 (i.e., $I_{2\to1}$) and all third terms represent the interference of stream 3 onto user 1 (i.e., $I_{3\to1}$).

Figure 6.4 illustrates the channel magnitude for two consecutive LTFs of our ISIT preamble. Due to space constraints and for ease of illustration, we only show the subcarrier allocation in a 2x1x2 system. Notice that while in the first LTF the AP allocates the odd subcarriers to SS1, in the second LTF the subcarrier order is reversed. Also, notice that in these plots both the interference and intended signal have comparable magnitudes. This example is easily extended to more than 2 streams by proportionally increasing the number of LTFs with respect to the number of streams (1-to-1 relationship), and by allocating the corresponding subcarriers according to the example provided above.

![Image](image.png)

**Figure 6.4** : Illustration of channel estimates for interleaved subcarriers (2x1x2 system).

### 6.4 Multi-User Interference-Aware Fast Recovery

In this section we describe the design and implementation factors for our multi-user fast recovery scheme.
6.4.1 Overview

Compared to single-user systems, failed transmissions in multi-user schemes incur a longer recovery time thus reducing the system’s efficiency, i.e., data airtime utilization. In particular, multi-user retransmissions typically involve not only a contention phase as in the case of 802.11 legacy retransmissions but also a re-sounding phase. Namely, in 802.11-based MUBF systems, upon a failed transmission, the AP triggers a binary exponential backoff process and begins contending for the medium. Once the AP gains access to the medium, it re-sounds the channel to generate the beam-steering weights needed for MUBF.

In contrast, we propose a multi-user retransmissions scheme that precludes the need to re-sound the channel by triggering a one-time immediate retransmission. That is, our scheme targets to realize a throughput gain by reducing the overhead incurred from repeated channel estimation. Nonetheless, by doing this, the AP faces the challenge of precisely determining the MCS which yields a successful retransmission even when the CSIT it possess for each user is increasingly outdated and inaccurate. Merely selecting MCS based on previously collected CSIT would likely lead to a failed retransmission, especially given the fact that the original transmission using this information has already failed. Similarly, arbitrarily decreasing the MCS to account for uncertainty in the current channel conditions might become overly conservative. Our joint retransmission and MCS selection scheme considers two key concepts, decreased receiver-dimensionality due to liberated antenna resources after successfully serving at least one user in the original user set, and per-user inter-stream interference awareness at the AP obtained via feedback during the acknowledgment process.
6.4.2 Retransmission Overhead

The retransmission process in multi-user systems is less efficient than that in single-stream systems due to the need to re-sound the channel. Consider the 4x1x4 example in Figure 6.5; the top figure illustrates the MUBF retransmission process in 802.11-based networks. First, the AP beamforms four different streams to four different users but only two of them are completely decoded. Consequently, the AP initiates a contention phase after DIFS time and then triggers a sounding phase to acquire the channel estimates of the two remaining users as well as two other users. In contrast, CHRoME eliminates the need to re-sound the two remaining users by immediately attempting a retransmission (Figure 6.5 - bottom). Notice that the potential gains that our scheme can provide in terms of overhead reduction are both due to eliminating re-sounding as well as to avoiding any increase in the contention window (CW) that is readjusted (incremented) after a frame loss. While the legacy system can maximize the number of streams served in a particular MUBF transmission, we demonstrate that the availability of additional degrees of freedom provides us with the opportunity to use the same CSIT collected in the previous sounding phase and still attain gains compared to legacy retransmission.

In CHRoME, before the retransmission is triggered the AP evaluates whether to serve all users in a multi-user MIMO fashion or via a TDMA MISO (Time Division Multiple Access Multiple-Input Single-Output) transmission. In particular, the AP assesses the time required to complete the transmissions in the two different modes and selects the configuration that minimizes the retransmission time (considering the respective MCS to each user). Figure 6.6 presents a hypothetical scenario where the AP compares the two retransmission modes. That is, it evaluates the time it takes to serve all users one at a time (sequentially) vs. serving them concurrently. Notice that in the TDMA MISO case, the MCS for each user is expected to be higher due to a higher expected SINR enabled by a power and diver-
Figure 6.5: Illustration of the retransmission strategy in legacy 802.11 (top) and in CHRoME (bottom)

sity gain. Consequently, the increase in MCS could lead to a faster overall transmission. As shown in the figure, in this case the AP would select the TDMA transmission since it would be completed in less time.

Figure 6.6: Hypothetical illustration comparing different retransmission configurations
6.4.3 Receiver-Dimensionality Reduction and Inter-Stream Interference Awareness

Since CHRoME avoids re-sounding the channel, we allow the AP to reuse the CSIT employed in the original transmission in order to generate the beam-steering weights needed in the beamformed retransmission.

**SINR Enhancement Due to Liberated Antennas.** Our rationale for allowing the reuse of possibly outdated channel information is based on the counteracting effect provided by the sudden availability of additional (liberated) antennas at the transmitter. That is, if any users were successfully served in the original transmission, every additional degree of freedom that becomes available at the transmitter (with respect to the number of concurrent users) yields an increase in per-user SINR due to both antenna diversity gain and per-stream transmit power increase. CHRoME exploits these gains in order to counteract the SINR reduction that is due to the use of inaccurate or outdated channel estimates to generate the beam-steering weights. Without re-sounding the channel, the AP reshapes the channel matrix to account only for the remaining users and computes the beamforming weights for those users. Reducing the number of users to be served relative to the number of transmit antennas simplifies the construction of non-interfering streams, thus leading to a lower SINR penalty due to imperfect beamforming weights.

To show the potential SINR gains that can be attained by MUBF systems when the number of transmit antennas increases relative to the number of simultaneous users we simulate a scenario with one multi-antenna AP and 100 single-antenna users in a MIMO Gaussian channel. The AP employs a zero forcing precoding strategy and we assume perfect CSIT estimation (no quantization). Moreover, the channel input is subject to an average power constraint $E[||x||_2^2] \leq \text{SNR}$, where we let $\text{SNR} = 10 \text{ dB}$. The user group selected at each transmission is based on the individual channel norm for each user, i.e., $||h_s||$. Therefore, at every transmission the AP serves the $M$ users with highest channel
norm. Each data point consists of an average obtained over 10000 channel realizations. Figure 6.7 (left) shows the post-processing per-user SINR as a function of the number of transmit and receive antennas.

While the increase in SINR due to an additional transmit antenna varies depending on the overall configuration, the minimum increase we observed is roughly 2 dB (8x6 to 8x5 configuration). Moreover, these results demonstrate that the steepest increases occur at both extremes, that is, when the system approaches the maximum diversity gain, i.e., Mx1, as well as in the case where there is only one single additional antenna. More importantly, notice that the SINR increase observed in our simulations, closely match the expected value that is roughly approximated by Equation (6.3) [4] and plotted in Figure 6.7 (right). Therefore, the SINR of a signal transmitted with power $P$ scales proportionally to $\frac{M-S+1}{S}$.

$$E\{\text{SINR}_{BF}\} = 10 \cdot \log_{10} \left( \frac{M - S + 1}{S} \frac{P}{N_0} \right)$$ (6.3)

Although there is a clear scaling difference between both plots in Figure 6.7, the difference in SINR from increasing or decreasing the number of users relative to the number of transmit antennas remains the exact same. Based on these results, we can observe that in the case of the scenario presented in Figure 6.5, the per-user SINR in the retransmission can increase by close to 7 dB. Considering the required 802.11 receiver sensitivity this could mean an increase of more than two MCS indexes under the assumption of static channel conditions. Therefore, we expect that the decrease in SINR due to outdated CSIT can be significantly mitigated in CHRoME via receiver-dimensionality reduction.

**Inter-Stream Interference-Aware Retransmission.** While a failed transmission indicates that the channel cannot support the current MCS given the current transmission resources and conditions (i.e., transmit antennas and concurrent users), the AP has no other
information to update its MCS selection according to current channel conditions. The default approach would be to let the AP select the MCS based on the CSIT for each user in the retransmission set. Nonetheless, as previously discussed, this would not consider the effects of inter-stream interference on each individual user. Similarly, the AP could conservatively select an MCS by simply decreasing the CSIT-based selection by one or two, e.g., MCS-4 to MCS-3 or to MCS-2. Notice however, that in this approach the AP merely relies on speculation that would possibly lead to an inaccurate selection (either by under- or over-selecting).

In CHRoME we enable users to *piggyback* information with respect to the SINR measured at each of these users, in the block acknowledgements (BA). More specifically, in CHRoME users report to the AP the information provided by the ISIT preamble we pre-
sented in Section 6.3. That is, upon a failed transmission, users append the individual SINR components to their block acknowledgement. Each individual SINR corresponds to the individual components induced by each independent data stream. For instance, if the original transmission to user 2 failed, this user reports three individual SINR values based on the measured $I_{1 \rightarrow 2}$, $I_{3 \rightarrow 2}$, and $I_{4 \rightarrow 2}$ components. Assuming that the retransmission user set contains both users 2 and 4 (reduced from four to two users), the AP considers only the SINR induced by user 4 onto user 2 in order to select the highest possible MCS according to the 802.11 receiver sensitivity specifications [3, 36]. Notice that this requires extending the BA frame by $|S| - 1$ fields, each consisting of only one octet. Recall $S$ denotes the set of users selected by the AP to be served in the previous MUBF transmission. If the original transmission to a particular user was successful, a regular 802.11 BA is used.

Also notice that recalculating the beam-steering matrix with fewer users would yield a higher per-user SINR compared to the SINR measured during the beamformed transmission. However, we argue that by relying on such SINR we increase the robustness of the system to outdated CSIT.

**Discussion on TXOP and Channel Release Mechanism.** In the context of 802.11, a modification to the retransmission strategy would need to consider its effect on the transmit opportunity TXOP mechanisms. While we have not explicitly addressed this issue in our implementation, we argue that the TXOP can be adjusted so as to allow one fast retransmission at a minimum, for delay sensitive traffic such as voice and video. Similarly, in the case that no retransmission is required and the AP has no more data to transmit at a particular TXOP, channel release mechanisms such as the one proposed in [34] can be easily implemented.
6.5 System Implementation, Measurements, and Evaluation

We validate CHRoME via an implementation and an extensive set of testbed and system emulation experiments. First, we describe our implementation and experimental methodology. Then, we investigate the performance of each individual technique in CHRoME using a combination of over-the-air (OTA) transmissions as well as trace-driven emulation to accurately model 802.11 timings while transmitting over collected channel traces.

6.5.1 Implementation and Experimental Methodology

Implementation and testbed. We implemented CHRoME in the WARP and WARPLab framework [1]. The WARPLab environment allows us to perform all the signal processing including encoding in a PC, and then transmit these signals over the air for decoding on the receiver side. Nonetheless, in WARPLab, reading (writing) from (to) the board’s buffers do not allow us to evaluate our protocol in real-time; therefore, to accurately represent the time-scales at which 802.11 operates, we also rely on trace-driven emulation where we first collect continuous channel samples and then use these to evaluate our scheme. More importantly, by doing this we ensure we can replay the same channels for the schemes to be compared, therefore achieving repeatability and a rigorous evaluation.

We implement a 20 MHz OFDM zero-forcing MUBF system and evaluate the performance of our inter-stream interference mitigation and resolution scheme. That is, we transmit thousands of OFDM frames using our ISIT preamble and measure the effectiveness of our interference detection and cancellation mechanism. We measure performance based on bit error rate as well as error vector magnitude (EVM) in the 5.8 GHz band.

Trace-driven emulation. To accurately model the 802.11 time-scales we implement an 802.11ac-based MUBF trace-driven emulator featuring an entire OFDM transmit and receive RF chain. We collect a comprehensive set of channel traces with our testbed plat-
form and use those as input to our emulator. Precoding consists of a zero-forcing scheme with equal power allocation. Transmit side EVM is determined according to the highest MCS within a user group. We implement least squares channel estimation based on our ISIT preamble and the resulting per subcarrier SINR (post processing SINR at the MIMO detector output) is used to compute the effective SINR ($\text{SINR}_{\text{eff}}$) which we then map to an MCS. Notice that the channel traces also include the per-stream interference measurement provided by our ISIT preamble.

**From single-carrier to OFDM: effective SINR to MCS mapping.** In contrast to single carrier systems where SINR can be directly mapped to an MCS, in multi-carrier systems an intermediate step is necessary to map the per-subcarrier SINR to an effective SINR scalar metric, and in turn to a given MCS. The SINR in MIMO OFDM systems operating over frequency selective fading channels presents a highly dynamic range among the subcarriers. The performance of OFDM coded systems over these multi-carrier channels depends on the joint statistics of the SINR considering all data subcarriers, therefore, average SINR is not a useful metric to accurately estimate the system’s performance. The 802.11ax task group (TGax) is considering the use of a mutual information based MCS mapping method to achieve this PHY abstraction [48,56]. This method uses the per-subcarrier SINR to compute a received bit information rate (RBIR) metric which is then mapped to an $\text{SINR}_{\text{eff}}$. Further, this $\text{SINR}_{\text{eff}}$ is used to compute the packet error rate (PER) for different MCS.

In both plots in Figure 6.8 we present the curves we generated in order to map the per-subcarrier SINR to an MCS. That is, the figure on the left shows the relation between the RBIR metric and $\text{SINR}_{\text{eff}}$, whereas the curve on the right shows the relation between $\text{SINR}_{\text{eff}}$ and PER.
Figure 6.8: (Left) RBIR as a function of SINR for least square channel estimation. (Right) PER vs. SNR with an MMSE receiver. From left to right: MCS0 to MCS9.

6.5.2 M³CS: Probing-Based MCS Selection

MCS selection accuracy in explicit and implicit MUBF sounding systems. While MCS over-estimation can lead to significant frame losses, under-estimation leads to an opportunity loss in which the current channel conditions could have supported higher rates thereby leading to a throughput increase. We investigate the MCS selection accuracy of our probing scheme compared to the baseline MCS selection as well as a more conservative approach in which we decrease the baseline MCS by one. To this end, we evaluate the extent to which these schemes under- or over- select the MCS. Notice that the baseline MCS selection represents a scheme where each stream’s MCS is chosen according to the collected CSIT (i.e., a purely CSIT-based technique). We collect channel traces for over 28 different user locations and run 15,000 frame transmissions. For each transmission and MCS selection
scheme we measure the number of frames in which the MCS was over-, under-, and accurately selected. Moreover, we consider two different feedback algorithms: Explicit: The AP sounds the channel before every packet transmission and follows the feedback process mandated in 802.11ac. Implicit: All users transmit a training pilot sequentially to allow the AP to estimate the channel. We modify our emulator to achieve perfect channel reciprocity (including transmit and receive RF chains) to eliminate calibration effects from our study.

Figure 6.9 depicts the MCS selection accuracy of each scheme. The top plots correspond to experiments where we generated out-of-cell interference from neighboring APs and their users, i.e., interference that is not inter-stream interference. This out-of-cell yielded additional interference to the in-cell multi-user clients ranging from -70 to -90 dBm. Likewise, the left plots correspond to systems that obtain CSIT with explicit feedback measured by the clients and the right plots obtain CSIT with implicit measurements at the AP.

The results indicate that CHRoME is highly resilient to out-of-cell interference in both explicit and implicit systems. This is because CHRoME re-adjusts the MCS selection according to the interference learned and observed by each user during the probe. In contrast, the baseline schemes perform poorly in implicit systems (right plots) because sounding does not take out-of-cell or inter-stream interference into consideration, therefore leading to substantial over-selection. On the other hand, the relatively fair performance of the two baseline schemes for explicit feedback systems with interference is due to the fact that this interference forces a dramatic drop in MCS to the lowest indexes thus avoiding significant over-selection. Consequently, CHRoME yields greater gains when the channel supports a wide range of MCS. Notice however that our scheme is also affected by having this wider MCS operation range. Moreover, the inaccuracies in CHRoME arise from changes in the channel between sounding and MUBF data transmission. The results we report in
Figure 6.9 correspond to a large combination of indoor scenarios (with and without environmental mobility); however, by looking at those scenarios where environmental mobility was induced, the percent of packets with erroneous MCS selection drops to 15%. On the other hand, for scenarios with induced environmental mobility, this percent of packets with mistaken MCS rises to 33%. Thus, over- and under-selection instances are more evident in fast fading scenarios.

For the same experiments described above, we investigate the aggregate throughput of the multiple schemes. That is, we consider the MAC/PHY overhead involved in the transmission process, including the additional overhead incurred by CHRoME. Notice that CHRoME outperforms the baseline schemes in all instances, achieving gains ranging from 16% to 280% in the case of implicit systems and between 16% and 42% in the case of explicit systems. Therefore, substantial gains outweigh the limited overhead necessary to enable MUBF probing and feedback in CHRoME.

**Adaptation response time.** To illustrate how well MCS selection in CHRoME follows the best possible MCS (i.e., the highest MCS that can be supported during the data MUBF transmission) compared to baseline schemes, we plot a timeline showing about 40 samples in time and the MCS selected at each instance. Figure 6.11 shows the MCS index as a function of time for the cases where there is interference at the users, and no interference, top and bottom respectively. Observe that the green (CHRoME) curve closely matches the best MCS whereas the basic baseline scheme frequently over-selects. This demonstrates the capability of CHRoME to rapidly track the ideal selection even with drastic changes in channel conditions where the desired MCS jumps as high as eight MCS indexes.

**Robustness to suppression of channel sounding.** As shown in prior work [10, 61], the overhead incurred by 802.11ac explicit feedback can be a significant fraction of the total air-time. These same works (including the work in this thesis) have proposed suppression
Figure 6.9: Selection accuracy for all three MCS selection schemes. Top (Bottom) with (without) out-of-cell interference at users; BL=Baseline CSIT-Based/BLc=Baseline Conservative/CH=CHRoME.

Figure 6.10: Throughput of each feedback system and MCS scheme. One-to-one correspondence with plots in Figure 6.9.
Figure 6.11: MCS selection timeline for scenarios with out-of-cell interference (top) and without it (bottom). The CSIT-Based selection corresponds to the baseline scheme we have previously introduced.
of channel sounding in order to reduce overhead by avoiding sounding before every packet transmission. Nonetheless, these schemes are susceptible to transient channel variations and stale CSIT. Therefore, we explore the ability of CHRoME to protect the system against these changes.

In Figure 6.12 (top) we present the throughput performance of CHRoME compared to the other MCS selection schemes as the time gap between sounding and the beamformed transmission is increased. That is, we vary the amount of time between sounding and the downlink MUBF data transmission to evaluate the system’s robustness against outdated CSIT. Nonetheless, in our scheme the probing mechanism is triggered before every single one of these transmissions. Notice that we take into account the additional overhead required to trigger CHRoME. The reported results correspond to an evaluation performed over a subset of the total number of scenarios used in the experiments of Figure 6.10.

First, observe that even with small time gaps between sounding and data transmission, the baseline is significantly outperformed by the more conservative approach and by our scheme. In particular, for certain scenarios (those more severely affected by doppler), even a 1 ms. gap leads to MCS over-selection (shown in Figure 6.12 (bottom)). Therefore, by choosing a lower rate as in the case of the conservative scheme, over-selection is avoided.

Second, the slope of the baseline schemes is very steep, specially before reaching 50 ms. In contrast, the slope of CHRoME’s curve decreases at a much slower rate. More specifically, with increasing time gap, the system becomes more vulnerable to over-selection thus closing the gap between the two baseline schemes. Unlike the baselines, CHRoME keeps updating the MCS according to the degradation of CSIT. Notice that there is a key difference between the reason behind the throughput decay of CHRoME and that of the baseline scheme. More specifically, as shown in Figure 6.12 (bottom) and Figure 6.13, the baseline incurs in a substantial amount of MCS over-selection, which translates into a significant
packet loss. On the other hand, throughput losses in CHRoME are mostly due to the drop in MCS which is required to ensure that the channel can support the selected transmission rate. Figure 6.12 (bottom) shows the average MCS index selection for the baseline, CHRoME, and the ideal MCS selection scheme (i.e., ground truth). This ground truth is obtained by measuring the SINR at each user during the downlink MUBF data transmission and mapping it to an MCS index.

In Figure 6.13 we present the fraction of packets in which MCS was accurately selected, over-selected, or under-selected. As expected, the amount of over-selection incurred by the baselines increases with increasing time gap. Even for the conservative scheme, such increase leads to a switch from under- to over-selection. That is, the conservative scheme reaches a point in which decreasing the MCS by only one index is not sufficient. Surprisingly, while we expected CHRoME’s accuracy to remain the same with increasing time gap, there is a noticeable increase in accuracy. The results revealed that with highly degraded CSIT the minimum MCS becomes the best option. Therefore, the system forces the scheme to keep selecting the lower MCSs thus preventing it from choosing an erroneous higher rate.

**Resilience to feedback quantization error.** Similarly to changes in the environment and user mobility, errors due to poor CSIT quantization will also hinder performance by inducing errors on the beam-steering weights calculation thereby increasing the inter-stream interference. We evaluate performance as a function of the number of bits used to quantize the channel estimates such that each user quantizes its feedback using $B$ bits. We perform scalar quantization [35, 39] where the total number of bits are evenly allocated to magnitude and phase components. Following the explanation in [39], the elements of the channel vector $\mathbf{h}_k = [h_1, \ldots, h_M]^T$ – where $M$ is the number of transmit antennas (and in this case, the number of concurrently served users) – are divided by the element $h_1$ to yield $M - 1$
Figure 6.12: Throughput and MCS timeline as a function of time gap between sounding and data transmission. MCS indexes are averaged over the concurrently served users, as well as over all packet transmissions.
Figure 6.13: MCS selection accuracy as a function of time gap between sounding and data transmission

complex elements. Then, the $M - 1$ phases (relative) are quantized individually using uniform quantization in $[-\pi, \pi]$. On the other hand, the inverse tangents $\tan^{-1}\left(\frac{|h_m|}{|h_1|}\right)$ of the relative magnitudes for $m = 2, \cdots, M$, are quantized uniformly in the interval $[0, \frac{\pi}{2}]$

Figure 6.14 depicts the MCS selection accuracy of all three selection schemes (i.e., (a) to (c)) as well as their corresponding per-user throughput performance (i.e., (d)). To guarantee that our results only reflect the effect of quantization, we have performed this experiment under controlled scenarios where the channel obtained during sounding is the exact same as the one during the MUBF transmission. Therefore, besides the varying AWGN noise observed at the users and the circular symmetric complex Gaussian random variable used to map per subcarrier SINR to effective SINR (form the RBIR mapping process), only the quantization level varies. Observe that the selection accuracy of CHRoME is much higher than both baseline schemes in all cases. The extremely low quantization level leads to the selection of the lowest MCS which in turn leads to almost perfectly accurate
MCS selection (no “room” for error). While there is a slight decrease in selection accuracy for CHRoME with improved quantization, this remains close to 90% as we approach perfect quantization. This decrease is due to the fact that the channel supports a wider range of MCS thus the likelihood of making a mistake increases as well.

In contrast, for the baseline schemes, lower quantization translates into a high over-selection rate. Notice that with improved quantization the selection inaccuracy diminishes. The selection accuracy is clearly reflected in the throughput gain attained by CHRoME in Figure 6.14 (d). Since channels remain the same, the gap between schemes narrows as the number of bits needed to represent the actual channel vector increases. As shown in the figure, gains of CHRoME compared to the best of the baseline schemes range from 600% ($B = 20$) to 9% ($B = 100$). Interestingly, the baseline curves cross after approximately $B = 50$. The reason for this is that the basic baseline scheme begins to converge towards the best selection (like CHRoME) whereas the conservative scheme keeps under selecting.

### 6.5.3 Interference Mitigation and Recovery

CHRoME leverages multiple antennas at the users to cancel the most harmful interfering stream, i.e., the strongest source of inter-stream interference. To understand the direct adverse effect that such interference has on users we first quantify the error rate that it introduces into the system.

**Inter-stream interference performance impact.** We perform an over-the-air experiment where a 4-antenna AP (with antenna separation of $\frac{\lambda}{2}$) beamforms data to two concurrent 2-antenna users (with separation of $4\lambda$ between the users).\(^8\) For this experiment, users only utilize one antenna for data reception. We transmit over 1000 frames where half of them correspond to line-of-sight (LOS) scenarios and the other half to non-LOS (NLOS).\(^8\)

\(^8\) $\lambda$ represents the wavelength of the signal.
Figure 6.14: Feedback quantization: throughput and MCS selection accuracy; (a) baseline, (b) CHRoME, and (c) conservative baseline.
scenarios. At each transmission, users rely on our proposed ISIT preamble to measure the inter-stream interference, and similarly, immediately after reception users measure the noise.

Figure 6.15 (left) depicts the bit-error-rate (BER) as a function of the SINR at each user. The figure shows results for the measured inter-stream interference (ISI) as well as the co-channel interference (CCI) which we define as the amount of out-of-cell interference. That is, CCI does not take into consideration any inter-stream interference, and if no other AP or user in neighboring cells is active then the measurement is simply the noise. Observe that while the CCI measurement remains relatively constant at around 30 dB SNR, the SINR due to ISI has a significant impact on error rate. Therefore, dealing with such interference is critical to reduce the BER and consequently improve frame reception in MUBF systems.

**Removal of only the strongest interferer.** Similarly, the performance of the interference cancellation mechanism in CHRoME depends on its ability to identify the individual interference components from the multiple streams. With limited DoF at the receiver, the difference between the multiple interference components provides insights as to how well a user can cancel the former one while treating the rest as noise. Therefore, we investigate the difference in strengths between the interference components generated by each stream. To this end, we set up a 4x1x4 system (4-antenna AP and four single-antenna users) where the AP transmits over 4000 MUBF frames across four different topologies. For each transmission, the intended user measures its signal strength as well as the inter-stream interference, and computes the individual SINR considering each individual interference component.

Figure 6.15 (right) shows the SINR difference between the two highest measurements. Notice that in almost 90% of the transmissions, the difference reaches 4 dB. This demonstrates that under perfect and ideal interference cancellation, this difference would be enough to decode a stream that is at least one or two MCS higher than what current SINR
could support. Thus, in cases where the strongest interfering stream is at least 2 or 3 dB higher than the second strongest, cancellation would enable users to potentially decode frames that would not be possible to decode otherwise.

Figure 6.15: (Left) SINR co-channel and inter-stream interference; (Right) SINR difference between strongest and next strongest interfering stream.

**Interference cancellation in CHRoME.** Next, we investigate how the identification and removal of the strongest interferers – in the absence of enough DoF to cancel all interfering streams – improves the system’s performance. We deploy four different topologies where a 4-antenna AP beamforms three concurrent streams at all times (one for each user). Moreover, one user features two antennas whereas the other two only have one antenna each. We compare CHRoME’s interference cancellation scheme against a simple antenna diversity scheme where only the antenna receiving the strongest signal is considered, as well as a receiver diversity scheme that applies a stronger weight (between 0.5 and 0.8) to the antenna that replied with channel feedback.**

**Notice that the beam-steering weights are generated according to the feedback provided only by the
Figure 6.16 depicts the received EVM as a function of the SINR. Observe that throughout the entire range of SINR, the interference cancellation scheme outperforms the rest. While the performance of the antenna selection and receiver diversity strategies improves rapidly with increasing SINR, their EVM at low SINR regimes is significantly lower than that of the cancellation technique. That is, CHRoME’s IC scheme is only mildly affected by low SINR compared to the other schemes. This demonstrates the critical need for leveraging both antennas to remove the strongest interferer using CHRoME vs. using the antennas for legacy multi-antenna reception mechanisms.

![Figure 6.16: EVM performance of CHRoME’s IC strategy](image)

**6.5.4 CHRoME’s Fast Recovery**

Throughput gain/loss of our retransmission system compared to an 802.11 MUBF approach depends on two main factors: incurred overhead and success rate of retransmission frames.

---

antenna measuring the strongest SNR.
Avoiding a re-sounding phase should decrease overhead yet could also decrease the likelihood of a successful retransmission if the same channel information is used. In contrast, while 802.11 incurs higher overhead, it also uses more updated CSIT estimates to perform a beamforming transmission. In principle, CHRoME attempts to increase the likelihood of a successful retransmissions while exploiting this decrease in overhead.

We implement CHRoME’s retransmission scheme and compare against 802.11 with re-sounding and against two other baselines. The first baseline consists of a MUBF retransmission scheme where users are always served simultaneously, whereas the second one consists of a MISO TDMA scheme that serves all users sequentially (one at a time). We provide the 802.11 scheme an advantage by assuming that all the retransmissions use all available DoF to maximize the multiplexing gain. That is, in a system with 4 antennas at the AP, all retransmissions consider 4 concurrent users. Moreover, only for the 802.11 case, our evaluation assumes that all retransmissions are always successful (best case scenario for 802.11).

We transmit a total of 12,000 packets and plot the system’s throughput in Figure 6.17. First, observe that the additional overhead incurred by the combination of doubling of the backoff window and channel re-sounding in the 802.11 scheme leads to a significant throughput penalty. Second, the difference in MCS due to higher number of concurrently served users causes a large throughput difference. Although the TDMA scheme serves each user at a higher rate compared to the MU-MIMO case, serving users sequentially leads to a similar drop in throughput. Therefore, we observe that with outdated CSIT and a small number of failed users, an MU-MIMO retransmission scheme has similar performance to a MISO TDMA scheme. Finally, observe that CHRoME performs at least as well as the best performing scheme (i.e., either MU-MIMO or MISO TDMA). That is, by selecting the configuration that minimizes the retransmission time we can outperform the rest of the
strategies.

In Table 6.1 we present the percent of fully successful retransmissions. In particular, for each number of failed users, we compute the amount of times that a retransmission was 100% successful (i.e., no failed users during the retransmission). As expected, for the MU-MIMO scheme, there is an increase in percentage as the number of failed users increases; nonetheless, this decrease only goes from 96.6% to 80.5%. Similarly we observe that for the TDMA scheme it ranges from 98% to 86% due to having a more aggressive MCS selection mechanism therefore incurring in over-selection. Furthermore, in our evaluation we observed an overall overhead reduction of 64.6% compared to the 802.11 baseline.

In summary, this evaluation demonstrates that the significant reduction in overhead due to avoiding sounding, as well as the resilience provided by the liberated degrees of freedom which enable a high successful retransmission rate, shift the accuracy/overhead tradeoff in favor of CHRoME’s fast retransmission scheme leading to high throughput gains. Therefore, even with outdated CSIT, the advantage provided by the additional antenna resources at the transmitter is sufficient to ensure that almost all retransmissions are successfully decoded.

<table>
<thead>
<tr>
<th></th>
<th>R=1 user</th>
<th>R=2 users</th>
<th>R=3 users</th>
</tr>
</thead>
<tbody>
<tr>
<td>4xR MU-MIMO (%)</td>
<td>96.6</td>
<td>89.8</td>
<td>80.5</td>
</tr>
<tr>
<td>4x1 TDMA (%)</td>
<td>98</td>
<td>86</td>
<td>92</td>
</tr>
</tbody>
</table>

*Table 6.1: Percent of fully successful retransmissions*
Figure 6.17: System’s throughput of CHRoME’s retransmission scheme compared to 802.11.

6.6 Chapter Concluding Remarks

In this chapter we present the design and implementation of CHRoME, a novel MUBF scheme that increases the system’s resilience against downlink inter-stream interference due to channel variations, user mobility, and poor feedback quantization. CHRoME features an inter-stream-interference-robust MCS selection technique, a mechanism for detecting and mitigating interference-leakage, and a fast retransmission scheme that obviates the need to re-sound the channel therefore minimizing overhead while guaranteeing robust retransmissions. We demonstrate that by obtaining and incorporating knowledge with respect to inter-stream interference into design decisions, our protocol can attain significant throughput gains compared to legacy systems.
Chapter 7

Conclusion and Future Directions

In this chapter we present a summary of our contributions and potential future directions for this work.

7.1 Summary of Contributions and Significance

Recent advances in integrated circuit (IC) design and software defined radio (SDR) have paved the way for the implementation and commercialization of Multi-user MIMO systems. Similarly, recent academic and industry work have demonstrated that these implementations are capable of reaching the astounding PHY rates previously claimed by theoretical work. Nevertheless, as we have shown in this thesis, these substantial gains are severely diminished at the system level due to a combination of factors. In particular, this thesis identifies two key factors that lead to a significant reduction of these gains at the higher layers, i.e., MAC and above; namely, control messages overhead (e.g., sounding overheard) and the lack of resilience to inter-stream interference. More specifically, we present the following contributions:

First, we demonstrate the impact of overhead due to control messages such as sounding and multi-user acknowledgements, on the overall data air-time utilization of the system. Likewise, we show that inter-stream interference due to inaccurate and outdated channel state information leads to a significant decrease in SINR at the users and consequently to a severe throughput degradation.
Second, we propose two protocols (CUiC and MUTE) for alleviating the amount of overhead necessary to enable MU-MIMO. These protocols rely on temporal and spatial compression in order to achieve this. On one hand, CUiC exploits the available degrees of freedom at the AP to allow users to concurrently send their control messages. On the other hand, MUTE relies on channel statistics to determine whether sounding can be suppressed for users with slowly fading channels. Both protocols are shown to substantially decrease the air-time devoted to transmitting control messages instead of actual data.

Finally, we propose CHRoME, a protocol for addressing the losses that originate due to inter-stream interference. In fast fading environments where even sounding prior to every single data transmission does not guarantee that the collected CSIT will not become immediately outdated, throughput performance decays dramatically. CHRoME provides a suite of mechanisms for preventing erroneous MCS selection, and for enabling fast and immediate recovery. We demonstrate that our scheme outperforms existing solutions by accurately selecting the MCS as well as reducing the overhead required to recover from losses during MU-MIMO data transmissions.

7.2 Future Directions

Based on the results presented in this thesis we identify several new directions for future work on this area of MU-MIMO:

- Temporal sounding suppression requires the AP to evaluate a certain metric and in turn decide whether the CSIT it currently possesses is accurate enough for calculating the beam-steering weights to be used in the next data transmission. In Chapter 5 we demonstrated the feasibility of sounding suppression in time and relied on the change in magnitude and phase of the distinct channel components to make such decision.
Nevertheless, a unified and more accurate metric can significantly reduce the system’s implementation complexity as well as the memory requirements imposed by our scheme. Moreover, this metric should easily extend to multi-carrier systems such as OFDM. Recall that our evaluation of MUTE considered a narrowband system. In our opinion, this metric has yet to be found.

- From an overhead minimization perspective, implicit feedback is the best sounding strategy. Nonetheless, explicit feedback systems have prevailed due to their higher accuracy and ease of implementation, i.e., no RF calibration required. Another factor we believe is critical for achieving highly accurate implicit feedback is to find a mechanism that informs the AP about the interference or noise observed at the users. That is, in implicit feedback the AP precodes the signal based on its view of the channel; unfortunately, in real deployments the physical channel is not reciprocal due to other sources of interference. Solutions to these two problems would potentially reduce the need for explicit feedback therefore making MU-MIMO systems more efficient.

- Much of the future research focus needs to be set on scaling MU-MIMO networks in order to exploit their maximum potential. With increasing number of users the complexity of currently employed precoding schemes can become prohibitive. Similarly, the level of inter-stream interference is significantly more dominant with more concurrent streams. While the former issue can be alleviated with more efficient hardware platforms and parallelization algorithms, loss-prevention and recovery mechanisms for large-scale MU-MIMO (e.g., massive MIMO) will represent one of the major challenges in the near future.
Bibliography


of ACM SIGCOMM, 2013.


