Algorithms and Proofs of Concept for Massive MIMO Systems

Vieira, Joao

2017

Document Version:
Publisher's PDF, also known as Version of record

Algorithms and Proofs of Concept for Massive MIMO Systems

João Vieira

Lund 2017

Department of Electrical and Information Technology
Lund University
Box 118, SE-221 00 LUND
SWEDEN

This thesis is set in Computer Modern 10pt
with the LaTeX Documentation System

Series of licentiate and doctoral theses
No. 108
ISSN 1654-790X
ISBN 978-91-7753-442-6 (print)
ISBN 978-91-7753-443-3 (digital)

Printed in Sweden by *Tryckeriet i E-huset*, Lund.
October 2017.

To the altruists

Popular Science

Imagine that you want to experience the best mobile data connection ever. Feeling sleepy already? Ok, let's switch perspective! Let's then say you want that WiFi Netflix (or SVT) stream from your modem to your 4K Smart-TV to be super reliable, i.e., you should be able to see at least 10 drama scenes of a SKAM episode without any cut. Or say you want that Call of Duty connection to be flawless, so that you don't get killed because of "lag and stuff". How about streaming live videos from your cellphone Facebook app? It would be perfect if every wireless connection would be flawless, right? Turns out every one owning a mobile device thinks exactly like you!

Mobile services are expected to be fast, reliable, and of course, cheap. For scientists and engineers, building systems that meet today's traffic user demands is a challenging task already, not to mention that the number of users (or more precisely, the number of connected devices) are expected to grow exponentially in the next upcoming years. To support the customer demands, scientists and engineers roll up their sleeves in order to build such systems. When it comes to cellular communications (the ones your mobile operator charges you monthly so that you can enjoy your 10 or 100 Gigabytes of incoming data traffic), there are several ideas on how to improve the overall data traffic of the network without compromising on the quality of service to the end user. One such idea is to have base stations with 100's or even 1000's of antennas co-operating together. Think of antennas as you think about Hebrews: it's more likely that one Hebrew crosses the Red Sea alive if the number of Hebrews crossing the Red sea is large, right? Same with antennas and signals: it is more likely that at least one of the 100 transmitted signals (each signal transmitted from a single base station antenna) arrives reliably to your mobile device than if only a single signal was originally transmitted.

"But 100 co-operating antennas sounds complicated, right?" Right! Nevertheless, this novel technology is one of the main candidates to be integrated in fifth generation (5G) mobile systems, which are expected to start operating in 2020. This thesis focuses on solving different problems associated with this

technology. It addresses some of the following questions:

1. *Does this idea work in practice?* We tried to answer this question by building the first non-proprietary system that demonstrates this technology, and meets the real-time requirements that a 4G communication link should ensure (although it is envisioned to integrate 5G systems). We named our prototype the LuMaMi testbed, which stands for Lund University Massive MIMO testbed.

2. *Can it be made cheap?* Each antenna needs to have an associated circuit to process its incoming/outgoing signals. As a result, scaling the number of antennas also scales the number of associated circuits. These circuits need to be cheap in order for the whole system to be cost effective. The main problem with cheap circuits is that they do not behave ideally, for example, they are not "reciprocal". We addressed part of this problem by proposing an algorithm which mitigates (i.e., calibrates) the non-idealities of a cheap circuit from a reciprocity point of view.

3. *Can this system track me well?* One requirement of next generation wireless systems, it that they need to localize and track users reliably, which can be done by taking advantage of the hundreds of antennas available. Using tools from artificial intelligence, we found suitable algorithms for positioning and tracking of mobile devices.

This thesis addresses different problems associated with base stations operating with 100's or even 1000's of co-operating antennas. The scope of the thesis is broad: the problems addressed range from algorithm design to implementation, from communications to localization, etc. It resulted from an interdisciplinary project between the Communications Engineering group and the Integrated Electronic Systems group at Lund University, and several industrial and academic partners. It, hopefully, makes a small contribution when it comes to pushing the envelope of knowledge in the area, and helps maturing this conceptual technology so it can integrate our daily lives in a near bright future.

Abstract

This thesis focuses on algorithms and proofs of concepts in the area of wireless systems operating with a large number of antennas, especially at the base station side.

The first studied topic concerns the design and implementation of massive multiple-input multiple-output (MIMO) testbeds, primarily for communications. This is an entirely new engineering challenge on its own, due to the unprecedented use of a large number of base station antennas together with time division duplex (TDD)-based operation. We consider hardware and system-level aspects of extending current Long Term Evolution (LTE) systems in order to integrate a massive number of antennas at the base station side. We materialize our testbed design into the Lund University massive MIMO (LuMaMi) testbed, and finalize with (measured) proof-of-concept results to validate our design claims.

The second researched topic addresses transceiver calibration to re-establish the reciprocity assumption of a wireless link. This aspect is crucial to be dealt due to the preferred operation mode of massive MIMO, i.e. TDD. To overcome the practical hassles of hardware-based calibration schemes, we propose a convenient over-the-air sounding method between all pairs of base station antennas that allows gathering enough measurements in order to estimate robust calibration coefficients. We provide algorithmic contributions and experimental evidence that corroborate the use of this calibration methodology in practice. This calibration approach is also applied to the case of calibrating the transmitter and receiver chains individually, for classical array beamforming applications.

The topic of detection in block fading (massive) SIMO systems is also addressed. This system setup is very representative to those of many existing systems as of today, e.g., in low power sensor networks. Using an estimation framework learned from our work in transceiver calibration, namely the generalized method of moments (GMM), we study a closed-form estimator that balances complexity and performance nicely.

The last part of the thesis aims to bring together the emerging topic of Deep Learning with fingerprint-based terminal positioning using uplink massive MIMO channels. The key idea is that the intricate structure of *raw* massive MIMO channels can be learned by deep learning networks and therefore used for positioning purposes. We study the applicability of a particular case of deep learning methods, namely, convolutional neural networks, which are state-of-the-art learning machines in the context of image processing.

Preface

This doctoral thesis is an outcome of my personal interest for the subject of statistical signal processing. An interest that grew, in part, due to the existence of influential figures in my education journey.

The impact of influential figures in my education journey has significantly played out in terms of which academic paths I ended up undertaking. For example, my primary school teacher, which I had great affection for, was a native Spanish speaker. Her shortcomings speaking the national language of her current country of residence (i.e., Portugal) was compensated by excelling when it came to teaching scientific disciplines. I believe it back then when my mind started to accept, and enjoy, exact sciences more than, say, *sciences of the other kinds*. During secondary school I was fortunate enough to cross paths with one the most charismatic math teachers I ever had–Father Vieira. He was a ferocious chess player who hated to lose more than enjoyed winning. During Bachelor studies I had an extremely exigent and rigorous teacher, Prof. Amândio, who awakened my personal interest for signal processing. He made the academic journey of students hard, but fun–the true way to craft champions. Finally, during my Master and Ph.D. studies I found a mentor from the Eslöv's metropolis that shared many of the attributes of my previous teachers (even a language defect) and much more: he is also a dear friend.

Each of these individuals had captivating personalities: they were authentic, altruistic, and genuinely passionate towards their teaching subject. They kept it real for themselves and for others, and therefore were extremely influential figures in my academic life. They were landmarks in my entire education that, at key moments, steered my academic path leading to this doctoral thesis in the discipline of applied signal processing for wireless systems. This thesis is my best effort to let them known that all their guidance mattered.

This doctoral thesis summarizes my research as Ph.D. student at the department of Electrical and Information Technology at Lund University from September-2013 to October-2017. It is comprised of two parts. Part I addresses the fundamental concepts in the area of multi-antenna systems that

the general reader needs to be familiar with in other to follow the rest of the thesis. Part I also provides a brief introduction to each of the (four) researched topics. These research efforts are synthesized in Part II of the thesis by means of a compilation of scientific publications. Their published details are listed below.

[1] Joao Vieira, Steffen Malkowsky, Karl Nieman, Zachary Miers, Nikhul Kundargi, Liang Liu, Ian Wong, Viktor Öwall, Ove Edfors and Fredrik Tufvesson, "A flexible 100-antenna testbed for Massive MIMO," in *Proc. IEEE Global Communications Conference (GLOBECOM) Workshop on Massive MIMO from Theory to Practice*, Austin, USA, pp. 287-293, Dec. 2014.

Personal Contributions: This paper results from a project which was a collaboration between the Communications Engineering and the Integrated Electronics Systems groups at Lund University, together with an industrial partner. I was one of the main responsibles to carry out this project. I am one of the main contributors of this paper, and I took the lead in writing the paper. Together with Steffen Malkowsky, I was responsible for assembling the testbed, testing individual testbed components, and programming the testbed FPGAs. I also contributed with suggestions in the architectural design of the testbed. Once the testbed was operational, I designed the measurement campaign, performed the measurements, and did the statistical analysis for proof-of-concept.

[2] Steffen Malkowsky, Joao Vieira, Liang Liu, Paul Harris, Karl Nieman, Nikhil Kundargi, Ian Wong, Fredrik Tufvesson, Viktor Öwall and Ove Edfors, "The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation," in *IEEE Access,* vol. 5, 9073-9088, May 2017.

Personal Contributions: This paper is a follow up to Paper I, where a novel centralized architectural design for a massive MIMO base stations is the focus. The (Lund) Ph.D. students involved in the project took turns when it came to the main authorship of the publication outputs. Steffen Malkowsky is the main contributor of this paper. From my side, I contributed with writing some sections of the paper, and participated in the measurement campaign and respective statistical analysis. I also contributed with FPGA programming to implement the new baseband architecture, and provided suggestions from a broad point-of-view.

[3] Joao Vieira, Fredrik Rusek and Fredrik Tufvesson, "Reciprocity calibration methods for massive MIMO based on antenna coupling," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Austin, USA, pp. 3708-3712, Dec. 2014.

Personal Contributions: I am the main contributor of this paper, and I took the lead writing the paper. I came up with the idea of the proposed mutual coupling based calibration method used for co-located massive MIMO base station arrays. I was responsible for the algorithmic contributions, performing the measurements, and doing the simulations and statistical analysis.

[4] Joao Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu and F. Tufvesson, "Reciprocity Calibration for Massive MIMO: Proposal, Modeling, and Validation," in *IEEE Transactions in Wireless Communications*, vol. 13, no. 4, pp. 3042-3056, May 2017.

Personal Contributions: I am the main contributor of this paper, and I took the lead writing the paper. I was responsible for proposing the novel estimator associated with the paper, and analyzing its statistical performance. I also implemented the entire calibration methodology in our testbed, and designed and executed the measurement campaigns. I also performed the analysis of the experimental results.

[5] Joao Vieira, Fredrik Rusek and Fredrik Tufvesson, "A receive/transmit calibration technique based on mutual coupling for massive MIMO base stations," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, Spain, pp. 1-6, Sept. 2016.

Personal Contributions: I am the main contributor of this paper, and I took the lead writing the paper. I utilized the previous idea of mutual coupling based calibration originally envisioned to re-establish reciprocity, and applied it to calibrate the transmitter and/or receiver radio frequency chains individually. I was responsible for development the algorithmic adjustments, and performing the statistical analysis.

[6] Joao Vieira, Fredrik Rusek and Fredrik Tufvesson, "A Generalized Method of Moments Detector for Block Fading SIMO Channels," in *IEEE Communications Letters,* vol. 20, Issue 7, 1477-1480, Jul. 2016.

Personal Contributions: I am one of the main contributors of this paper. I was responsible for deriving the estimator in closed-form, and performing the numerical analysis. Fredrik Rusek did the analytical asymptotic analysis, and we shared the paper writing equally.

[7] Joao Vieira, Erik Leitinger, Muris Sarajlic, Xuhong Li and Fredrik Tufvesson, "Deep Convolutional Neural Networks for Massive MIMO Fingerprint-Based Positioning," in *IEEE 28th Annual International Symposium on Per-*

sonal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, Oct. 2017.

Personal Contributions: I am the main contributor of this paper, and I took the lead writing the paper. I came up with the main idea of the paper: using deep learning methods, i.e., convolutional neural networks, to process transformed massive MIMO channels for positioning. I re-adapted and implemented the neural network for the current application, and conducted the numerical analysis. The channel realizations used as fingerprints were generated from an already existing implementation of the COST 2100 channel model at our research group.

Acknowledgements

Some people thank God in this first paragraph. I thank Fredrik Rusek. Working alongside with Fredrik from an early stage of my postgraduate studies allowed me to fulfill many of the expectations I originally had, and to have a blast while doing it. Only his twisted sense of humor outshines his ability to mentor and to do good science.

I am also grateful to my supervisor Professor Fredrik Tufvesson, for providing me a life changing opportunity to pursue postgraduate studies and to integrate his research group. All of our interactions made me a smarter person from many points of view. I also want to thank my co-supervisor Professor Ove Edfors for, among many things, being a role model when it comes to lead (by example) the Communications Engineering group at Lund University.

To my older brother in science, and dear friend, Muris Sarajlić. (How come a consonant has an accent?) Thanks for being available for all kinds of discussions without constantly looking at the watch.

A special thanks to the true heroes behind the testbed project, Steffen Malkowsky and Associate Professor Liang Liu. You are the ones that endure when times are tough.

Thanks to Erik Leitinger and to Jose Flordelis, for voluntarily reviewing parts of the thesis, and for sharing a genuine interest for research. Also, thanks to all my other research partners and work colleagues whom I have, in one way or another, interacted during my post graduate studies.

To the ones I think about every day: my mother, my father and my sister.

To my love Fátima, my son Sebastian, and to the rest of my unborn children: I will make you the happiest persons in this damn Earth.

João Vieira
Lund, October 2017

List of Acronyms and Abbreviations

ADC Analog-to-Digital Converter

BB BaseBand

BS Base Station

CNNs Convolutional Neural Networks

DL Deep Learning

EM Expectation Maximization

EVM Error Vector Magnitude

FPGA Field-Programmable Gate Array

GMM Generalized Method of Moments

GPS Global Positioning System

HSPA High Speed Packet Access

IID Independent and Identically Distributed

LLNs Law of Large Numbers

LOS Line-of-Sight

LS Least Squares

LTE Long Term Evolution

LuMaMi Lund University Massive MIMO

MIMO Multiple-Input Multiple-Output

MISO Multiple-Input Single-Output

ML Maximum-Likelihood

MR Maximum-Ratio

MMSE Minimum Mean Squared Error

MU Multi-User

OAM Orbital Angular Momentum

OFDM Orthogonal Frequency Division Multiplexing

PMF Probability Mass Function

QAM Quadrature Amplitude Modulation

RELU REctified Linear Unit

RF Radio Frequency

RIM Rusek Is the Man

RT Real-Time

SER Symbol Error Rate

SIMO Single-Input Multiple-Output

SINR Signal-to-Interference-and-Noise Ratio

SISO Single-Input Single-Output

SNR Signal-to-Noise Ratio

SU Single-User

TDD Time Division Duplex

UE User Equipment

Contents

Part I

Introduction to the Researched Topics

Chapter 1

Background

This chapter provides a brief overview of some of the existing multiple-input multiple-output (MIMO) technologies preceding massive MIMO—the focus of the thesis.

1.1 Point-to-Point and Multi-User MIMO

The essence of wireless MIMO technology is to equip both ends of a wireless link with multiple antennas and respective processing units. Compared to single-input single-output (SISO) systems, i.e., systems that operate with solely one antenna at each end of the link, the benefits brought by multiple antenna systems have been thoroughly studied for the last decades and are well understood [1, 2]. There are a number of approaches that can be used to realize a wireless MIMO system, two of which are described below.

MIMO has two main flavors. The classical flavor is widely known as point-to-point MIMO or single-user (SU)-MIMO [1]. The second flavor, an extension to the multi-user (MU) case, is widely known as MU-MIMO [1]. These two different flavors of a MIMO link are sketched in Fig. 1.1. The setup of a point-to-point MIMO link is simple: at a given time/frequency resource there is a single transmitter and a single receiver. Both entities are equipped with antenna arrays, and signal processing units that jointly process the inputs/outputs of their respective antenna array. The transmitter and receiver communicate through a medium termed the propagation channel, which is the *interface* between the two antenna arrays. If both ends of the link have antenna arrays of the same size (i.e., $M = K$ in Fig. 1.1) the MIMO system is symmetric in the sense that the overall behavior is the same, no matter which link direction is considered.

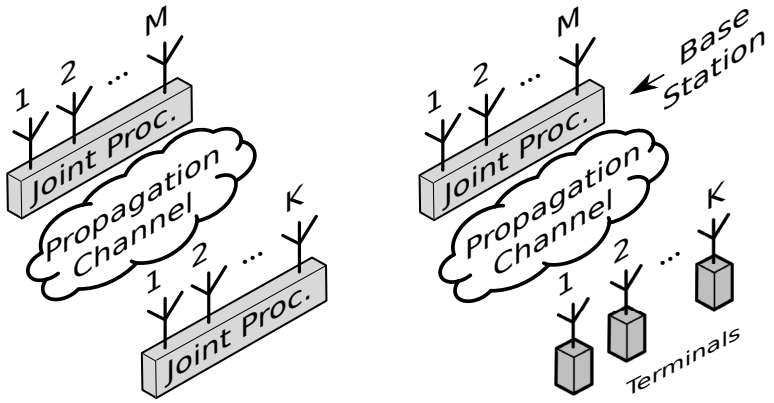

Figure 1.1: Left—A sketch of a point-to-point MIMO system. Each end of the link is equipped with an antenna array and a unit that jointly processes the array's transmit/received signals. Right—A sketch of a MU-MIMO system. The BS sits at the end of the link that is capable of joint processing. The mobile terminals sit on the other end of the link and are only capable of processing their own transmit or received signals.

This is not true in a MU-MIMO context. In MU-MIMO, one side of the link is not capable of performing joint processing. This side of the link represents the users end of the link, where K mobile terminals lie. In this work we consider single-antenna terminals only, thus K is also the total number of antennas at the terminals side. In the other end of the link lies an M-antenna base station (BS) which, in general, is capable of joint processing. As a result, there is the concept of uplink and downlink, i.e., when the BS acts as a receiver and as a transmitter, respectively.

1.2 Applications

One of the most dominant applications of the MIMO concept is in the field of digital communications. There the goal is to replicate some information-bearing binary sequence at the receive end of the link, equal to that existing at the transmit end of the link, with some specified error probability. It was *back in the late 1990's* when most of the fundamental results of point-to-point MIMO were established [3–5]. Similarly, MU-MIMO's research golden years were during the 2000's [6,7]. Since then, MIMO technology has been integrated

in several communication standards, for example IEEE 802.11n, IEEE 802.16, high speed packet access (HSPA) 3GPP Release 7, and Long Term Evolution (LTE) 3GPP Release 8 [8–10].

Other applications of MIMO are in *pure* channel sounding-based systems.[1] The goal of channel sounding is to simply measure properties of a wireless channel (e.g., impulse response or angular spread) [11]. The goal of channel sounding-based systems is to use the measured channel properties for subsequent ends. Pure channel sounding-based systems include, for example, radar-based applications such as detection of a target and its motion, positioning, localization and tracking [12].

This thesis covers two fields where MIMO can be applied: communications and positioning.

1.3 Shortcomings in Cellular Communications

MIMO was a brilliant invention with most early theoretical work pointing towards practical success in cellular communications. Despite its current presence in many wireless standards, some argue that no flavor of MIMO has yet lived up to its initial expectations. The common consensus appear to be that, up to the existing proposals, MIMO cellular technology does not scale very well with the system size, where most of the promised gains can be harvested—see [13] for an extended discussion. This happens for different reasons in the different MIMO flavors. For example, in point-to-point MIMO, most of the promised gains can only be harvested in ideal propagation conditions, where channels between different antenna pairs can be modeled as statistically independent. This is often not the case in practice due to the limited aperture of conventional user terminals, and simply increasing the number of terminal antennas does not improve the situation significantly. With MU-MIMO in its originally envisioned form and many of its extensions, e.g. distributed MU-MIMO, the main problems are the overhead for training which scales linearly in M and K,[2] and the need to feedback the trained downlink channels back to the BS for precoding purposes. Even with relatively small-sized systems, the scalability in cellular communications becomes compromised due to the fast time-varying nature of cellular channels which need to be typically re-learned on a millisecond basis [14]. (Also, from a practical stand-point, another problem is the latency and deployment hassle associated with *backhauling* distributed BSs.)

In this thesis, we study a relatively novel extension of MU-MIMO—massive

[1] We refer to pure channel sounding-based systems as systems where the receiver has always knowledge of the transmitted signals. A communication system therefore does not qualify as a pure channel sounding-based system.

[2] This is also the case for point-to-point MIMO.

MIMO—an approach where the number of antennas at both ends of the link, especially at BS side, scales massively. To the current date, certain operation modes of massive MIMO communications appear to bypass most of the fundamental scaling bottlenecks of the previous MIMO flavors, and therefore appear to scale well with the system size. We introduce this topic in Chapter 2.

1.4 Thesis Structure

This thesis focuses on four research topics in the realm of systems that operate with a massive number of antennas. More specifically: *i)* testbed design and implementation, *ii)* transceiver calibration, *iii)* detection algorithms, and *iv)* radio-based positioning. We first provide an introduction to the field of massive MIMO in Chapter 2, and then introduce each of the four research topics in Chapters 3, 4, 5, and 6, respectively. Finally, Chapter 7 provides conclusions and lessons learned from the work, and suggests future research directions. This finalizes Part I of thesis. Then, Part II of the thesis includes the publications listed in the Preface.

We remark that the notation used in Chapters 3, 4, and 5 follows the same conventions as those defined in Chapter 2. However, since Chapter 6 introduces a different signal model together with a positioning framework, we re-define some previously used variables. Thus, Chapter 6 is stand-alone when it comes to most notation used.

Chapter 2

Massive MIMO

This chapter introduces massive MIMO. It introduces the signal models used in the thesis, and comments on suitable processing schemes. It finalizes by addressing the advantages of using systems with a massive number of antennas for communications and for pure channel sounding-based applications.

2.1 Definition and General Remarks

Massive MIMO is the term used to describe a MU-MIMO system with a large number of antennas at both ends of the link, but especially at the BS side. The BS serves a multitude of mobile terminals in the same time-frequency resource. These appear to be the most specific statements that describe a massive MIMO communication system without conflicting with the different approaches available in literature. Some of these approaches utilize: time division duplex (TDD) or frequency division duplex (FDD) operation [15, 16], fully digital or hybrid architectures [17], codebook-based processing, full-dimensional processing, or processing using measured channels [16, 18], single or multi-antenna terminals [19], co-located or distributed BSs arrays [20, 21], etc. All of these mentioned approaches have something in common - they use the excess number of BS antennas, $M - K$, in an advantageous manner from multiple points of view. While still in active research and development, massive MIMO is a very promising technology, and most likely, will be integrated in next generation wireless systems (e.g., 5G systems) in one form or another.

The research presented in this thesis is based on an approach to massive MIMO very similar to that originally envisioned [22]. Its main characteristics are:

- *TDD operation:* the uplink and downlink channel uses occur at the same frequency but at different time instances. The main motivation for TDD-based operation is as follows. Under coherent detection in time-varying reciprocal channels, the measured uplink channel can be used, not only for equalization of the received uplink signals, but also for downlink precoding. With that, the number of training resources scale linearly only with K, and not with M as in previous MIMO flavors. This is because one single uplink pilot can train an unlimited number of BS antennas. Moreover, there is no feedback of the trained downlink channels back to the BS [13].

- *Fully Digital Architecture:* This architecture yields one transceiver chain per BS-antenna, and assumes that *measured* channels are used for both for uplink equalization and for downlink precoding. In a given channel use (we introduce this concept in Section 2.2), there is always one digital baseband (BB) signal per BS antenna. The BS processes all M digital BB signals jointly. As a result, this is the architecture that yields the most processing flexibility compared to its counterparts. From both theory to measurements, it appears to be the benchmark when it comes to communication performance [16, 23].

- *Co-located BS:* All work presented in this thesis is done in the context of having one single and co-located BS. This means that all BS antennas are physically located at one specific site.

- *Single-antenna terminals:* Each user terminal is equipped with only one antenna, and it only processes the signals associated with its antenna. As a result, most of the signal processing is conducted at the BS side.

2.2 Signal Models

Given the previously described massive MIMO approach, let us now introduce the BB signal models for transmission over a time-invariant narrowband channel.[1] First, we shortly introduce the general case of signaling over a block fading channel, as it covers all channel cases considered in this thesis. Then, we introduce a special case of this model—transmission in a single channel use—as it is the most used case in this thesis.

[1] The time-invariant narrowband channel assumption holds in many cases when using, for example, properly parametrized multicarrier modulation techniques as orthogonal frequency division multiplexing (OFDM) in communications [24], or by using properly parameterized orthogonal sounding sequences as Zadoff-Chu sequences in channel sounding [11].

We start with the uplink. Let K single-antenna users simultaneously transmit signals at time t of a block fading MIMO channel with $T+1$ coherent channel uses. Here we look at t, with $0 \leq t \leq T$, as the index of the coherent dimension of the channel. Let $x_{k,t}$ be the transmitted signal by the kth terminal at time t, and define the vector of transmitted signals from all K terminals at that instance as $\boldsymbol{x}_t = [x_{1,t} \cdots x_{K,t}]^T$. By collecting all $(T+1)K$ transmit signals in the matrix $\boldsymbol{X} = [\boldsymbol{x}_0 \cdots \boldsymbol{x}_T]$, we can write the received signals by an M-antenna BS array as

$$\boldsymbol{Y} = \sqrt{\rho_u}\boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}. \tag{2.1}$$

The matrix $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_K] \in \mathbb{C}^{M \times K}$ models the uplink radio channel with stochastic entries that account for small-scale fading effects—they are modeled as zero-mean unit-variance independent and identically distributed (IID) circularly symmetric complex Gaussian random variables. Other models exist for $\boldsymbol{H}$ that account for other channel effects (e.g. large-scale fading [22]), but we choose this simple model for the sake of our exposition—many insights that we will obtain also hold for other models. The entries of $\boldsymbol{N} = [\boldsymbol{n}_0 \cdots \boldsymbol{n}_T]$ typically model the combined effect of different system sources of noise, and are modeled as zero-mean IID circularly symmetric complex Gaussian random variables with variance σ^2. The variable ρ_u can be thought as the average energy transmitted during one uplink channel use by each user if $\mathrm{E}\left\{|x_{k,t}|^2\right\} = 1, \forall k, t$ holds—we assume this is the case throughout this thesis.

We now introduce the special case of signaling in a single channel use, i.e., when $T = 0$. We focus on this case for the rest of the thesis unless explicitly mentioned otherwise. With $\boldsymbol{Y} = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_T]$, $\boldsymbol{y} \triangleq \boldsymbol{y}_0$, $\boldsymbol{x} \triangleq [x_1 \ldots x_K]^T \triangleq \boldsymbol{x}_0$ and $\boldsymbol{n} \triangleq \boldsymbol{n}_0$, we write this case explicitly as

$$\boldsymbol{y} = \sqrt{\rho_u}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}. \tag{2.2}$$

In the context of communications, we assume the entries of $\boldsymbol{x}$ to be IID random variables with unit variance, and that each entry is sampled from a distribution with probability mass function (PMF), $\mathrm{pmf}(\cdot|\mathcal{A})$, defined as

$$\mathrm{pmf}(z|\mathcal{A}) = \begin{cases} \frac{1}{\#\{\mathcal{A}\}}, & z \in \mathcal{A} \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\mathcal{A}$ is a set whose elements are quadrature amplitude modulation (QAM) symbol alternatives [24], and $\#\{\cdot\}$ denotes the set cardinality.

The received signal vector, $\boldsymbol{y}$, is typically post-processed so that a quantity of interest is inferred. The post-processing function, $\mathrm{post} : \mathbb{C}^{M \times 1} \to \mathbb{C}^{N \times 1}$, depends on the current MIMO application, but in most cases we are interested

in performing some kind of estimation or detection of a parameter vector $\boldsymbol{p} \in \mathbb{C}^{N\times 1}$. For future use, we write this inferred parameter vector explicitly as

$$\hat{\boldsymbol{p}} = \text{post}(\boldsymbol{y}). \tag{2.3}$$

In the context of communications in a single channel use, we typically have $\boldsymbol{p} = \boldsymbol{x}$.

We now proceed by introducing downlink transmission. Let the M transmitted signals in the downlink, one signal per BS antenna, be stacked in the vector $\boldsymbol{z}$ as $\boldsymbol{z} = [z_1 \cdots z_M]^T$. By stacking each received user terminal signal in the vector $\boldsymbol{y}'$, the downlink model is written as

$$\boldsymbol{y}' = \sqrt{\rho_d}\boldsymbol{B}\boldsymbol{z} + \boldsymbol{n}'. \tag{2.4}$$

We also assume $\mathrm{E}\{z_k\} = 0$ and $\mathrm{E}\{\boldsymbol{z}^H\boldsymbol{z}\} = 1$, and thus ρ_d denotes the average energy transmitted during one downlink channel use. The matrix $\boldsymbol{B}$ is the downlink radio channel matrix, and $\boldsymbol{n}'$ is a vector with stochastic entries modeling downlink noise. In Section 4.1 we remark to what extent $\boldsymbol{B}$ resembles $\boldsymbol{H}^T$.

Each transmit signal vector, $\boldsymbol{z}$, is typically a function of a more generic quantity of interest, $\boldsymbol{p}' \in \mathbb{C}^{N'\times 1}$. The mapping function is denoted by pre : $\mathbb{C}^{N'\times 1} \to \mathbb{C}^{M\times 1}$. For sake of notation, define the transmit signal vector $\boldsymbol{z}$ as

$$\boldsymbol{z} = \text{pre}(\boldsymbol{p}'). \tag{2.5}$$

In the context of communications in a single channel use, we often have $\boldsymbol{p}' = \boldsymbol{x}'$, where the kth entry of $\boldsymbol{x}' \in \mathbb{C}^{K\times 1}$ is the data symbol intended for the kth user.

2.3 ML detection and Sum-Rate Capacity

There exist fundamental communication metrics to describe the performance of the links introduced in Section 2.2. Next, we introduce two such error metrics for the uplink case (2.2).

One metric of interest is the error probability of detecting $\boldsymbol{x}$. Given that the BS has perfect knowledge of $\boldsymbol{H}$, and under the modeling assumptions of (2.2), then maximum-likelihood (ML) detection of $\boldsymbol{x}$ is optimal in terms of minimizing the error probability.[2] ML detection is performed by solving the following constrained least squares (LS) problem

$$\hat{\boldsymbol{x}}_{\text{ML}} = \arg \min_{\substack{\boldsymbol{x} \\ \text{s.t. } \boldsymbol{x}\in\mathcal{A}^K}} J_{\text{ML}}(\boldsymbol{x}), \tag{2.6}$$

[2] The equivalence between the minimum error probability detector and the ML detector can be found in [24]. We also emphasize that the error probability is defined in terms of detection of the $K \times 1$ vector $\boldsymbol{x}$, and not in terms of its entries.

where $J_{\mathrm{ML}}(\boldsymbol{x}) = ||\boldsymbol{y} - \sqrt{\rho_u}\boldsymbol{H}\boldsymbol{x}||^2$.

Another example of a fundamental single scalar metric is the instantaneous channel sum-rate capacity C. It is an achievable upper bound on the sum of the signaling rates at which information can be exchanged between both ends of the link with arbitrarily low error probability [25].[3] If the BS has $\boldsymbol{H}$ at hand, the sum-rate capacity, measured in bits-per-channel-use, for that channel realization is given by

$$C = \log_2 \det\left(\boldsymbol{I} + \mathrm{snr}\, \boldsymbol{H}^H \boldsymbol{H}\right) \tag{2.7}$$

$$= \sum_{\ell=1}^{K} \log_2\left(1 + \mathrm{snr}\lambda_\ell\right), \tag{2.8}$$

where $\mathrm{snr} = \rho_u/\sigma^2$ is the, so-called, uplink signal-to-noise ratio (SNR) per receive antenna per user, λ_k is the kth largest eigenvalue of $\boldsymbol{G} \triangleq \boldsymbol{H}^H\boldsymbol{H}$ [1]. We remark that there no is concept of inter-user interference in this context—we define this concept later.

The complexity of the detector that is optimum in the sense of (2.6) scales exponentially with K. The ML decoding principle can also be used to reach capacity in the sense of (2.7). Alternatively, the minimum mean squared error (MMSE) receiver with interference cancellation can also do the same job at cubic complexities in K [1]. These last approaches rely on asymptotic arguments with respect to the number of channels uses to *orthogonalize* the receive information streams, which is far from being practical. Thus, it is of interest to use signal processing strategies that orthogonalize information streams on a symbols basis, and that do so at moderate complexities so that the system can scale well in practice.

2.4 Linear Signal Processing for Communications

One typical setting in massive MIMO communications is to constrain the outputs of post$(\cdot)$ and pre$(\cdot)$, see (2.3) and (2.5), to be linear functions of their inputs. Fortunately, the favorable properties of massive MIMO channels allow near-optimal performance to be obtained with such linear processing strategies. This is addressed in detail in Section 2.5. For now, let us write this linear dependence for the detector and for the precoder explicitly. With that (2.3) and (2.5) take the form of

$$\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{y} \qquad \text{and} \qquad \boldsymbol{z} = \boldsymbol{P}\boldsymbol{x}', \tag{2.9}$$

[3] To achieve this upper bound, the channel inputs $x_{k,t}$ are in general not drawn from finite–sized alphabets as $\mathcal{A}$, and $T \to \infty$.

where $\boldsymbol{W} \triangleq \left[\boldsymbol{w}_1^T \cdots \boldsymbol{w}_K^T\right]^T$ and $\boldsymbol{P}$ denote the general linear equalizer and precoder matrices, respectively. In the uplink, the kth entry of $\hat{\boldsymbol{x}}$ can be written as

$$\hat{x}_k = \underbrace{\sqrt{\rho_u}\,\boldsymbol{w}_k\boldsymbol{h}_k x_k}_{\triangleq \mathrm{sig}_k} + \underbrace{\sqrt{\rho_u}\sum_{\ell=1,\ell\neq k}^{K}\boldsymbol{w}_k\boldsymbol{h}_\ell x_\ell}_{\triangleq \mathrm{int}_k} + \underbrace{\boldsymbol{w}_k\boldsymbol{n}}_{\triangleq \tilde{n}_k}. \tag{2.10}$$

In a linear processing setup, int_k is usually termed the inter-user (or inter-stream) interference experienced by user k when decoding symbol x_k. Also, $\tilde{n}_k$ denotes post-processed noise, and sig_k is the desired signal that allows detection of x_k.

Most of the work in this thesis is based on two different linear processing approaches, namely, maximum-ratio (MR) and LS. We now describe how to obtain closed-forms for their equalizer matrices, $\boldsymbol{W}_{\mathrm{MR}}$ and $\boldsymbol{W}_{\mathrm{LS}}$, respectively. The derivation of their downlink counterparts, i.e. precoders, is similar and thus omitted from our exposition (conveniently left as a homework for the interested reader).

- *Maximum-Ratio Equalizer*: This equalizer maximizes the SNR ratio per user, disregarding inter-user interference. Since the optimization problem is decoupled between users, the MR objective function can be defined as

$$J_{\mathrm{MR}}(\boldsymbol{W}) = \mathrm{E}\left\{\sum_{\ell=1}^{K}\frac{|\mathrm{sig}_\ell|^2}{|\tilde{n}_\ell|^2}\right\} = \mathrm{snr}\sum_{\ell=1}^{K}\frac{|\boldsymbol{w}_\ell\boldsymbol{h}_\ell|^2}{\boldsymbol{w}_\ell\boldsymbol{w}_\ell^H}. \tag{2.11}$$

  Applying the Cauchy-Schwarz inequality [24] to each element of the summation at the right hand side of (2.11), gives $J_{\mathrm{MR}}(\boldsymbol{W}) \leq \mathrm{snr}\sum_{\ell=1}^{K}\boldsymbol{h}_\ell^H\boldsymbol{h}_\ell$. Thus, equality holds for $\boldsymbol{W} = \boldsymbol{D}\,\boldsymbol{H}^H$, regardless of snr, where $\boldsymbol{D}$ is a diagonal matrix with arbitrary non-zero diagonal elements. Defining the maximizer of (2.11) as $\boldsymbol{W}_{\mathrm{MR}}$, the vector estimate resulting of MR equalization is given by

$$\hat{\boldsymbol{x}}_{\mathrm{MR}} = \boldsymbol{W}_{\mathrm{MR}}\boldsymbol{y} \tag{2.12}$$

$$= \sqrt{\rho_u}\boldsymbol{D}\boldsymbol{G}\boldsymbol{x} + \boldsymbol{W}_{\mathrm{MR}}\boldsymbol{n}. \tag{2.13}$$

- *Least-Squares Equalizer*: This equalizer is obtained by solving (2.6) under a looser constraint, namely, $\boldsymbol{x} \in \mathbb{C}^{K\times 1}$. Its solution, $\hat{\boldsymbol{x}}_{\mathrm{LS}}$, is the classical projection of the received signal into the Moore-Penrose pseudoinverse of $\sqrt{\rho_u}\boldsymbol{H}$. Using the resulting solution $\boldsymbol{W}_{\mathrm{LS}} = \frac{1}{\sqrt{\rho_u}}\boldsymbol{G}^{-1}\boldsymbol{H}^H$, the equalized

signal is written as

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{W}_{\mathrm{LS}}\boldsymbol{y} \tag{2.14}$$

$$= \boldsymbol{x} + \frac{1}{\sqrt{\rho_u}}\boldsymbol{G}^{-1}\boldsymbol{H}^H\boldsymbol{n}. \tag{2.15}$$

The matrix $\boldsymbol{W}_{\mathrm{LS}}$ can also be seen as a zero-forcing equalizer, in the sense that forces the inter user interference to zero, i.e., $\mathrm{int}_k = 0, \forall k$.

We now address how to perform detection in a single channel use. It follows from (2.10) that the unbiased estimate of symbol x_k is given by

$$\hat{x}_k^{\mathrm{unb}} = \hat{x}_k / \left(\sqrt{\rho_u}\boldsymbol{w}_k\boldsymbol{h}_k\right) \tag{2.16}$$

$$= x_k + f_k(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_K, \boldsymbol{n}). \tag{2.17}$$

With that, detection of x_k after linear equalization is performed by solving

$$\arg \min_{\substack{x \\ \text{s.t. } x\in\mathcal{A}}} ||\hat{x}_k^{\mathrm{unb}} - x||. \tag{2.18}$$

The drawback of this sub-optimal detection approach is that, in general, it ignores the effect of inter-user interference. The advantage is that, compared to a brute force ML search, the computational complexity is reduced to K searches where each search is done over $\#\{\mathcal{A}\}$ elements. Thus, the overall processing complexity of linear equalization is mainly dictated by the computation of the equalizer $\boldsymbol{W}$ (e.g., the computation of $\boldsymbol{G}$ and its inversion are the most computationally demanding operations when doing least-squares equalization).

Interestingly, the sum-rate capacity of the linearly equalized models is the same as (2.2), i.e., $I(\boldsymbol{y};\boldsymbol{x}) = I(\boldsymbol{W}_{\mathrm{MR}}\boldsymbol{y};\boldsymbol{x}) = I(\boldsymbol{W}_{\mathrm{LS}}\boldsymbol{y};\boldsymbol{x})$, where $I(\cdot;\cdot)$ denotes the mutual information between its two inputs, and we assume that the joint distribution of the entries of $\boldsymbol{x}$ is optimal for capacity [25]. However, a typical assumption in the informational theoretical analysis of linear equalizers is to treat int_k as noise, rather than a decodable signal. One advantage of this approach is similar to that of (2.18)—decoding can be done separately user-by-user (stream-by-stream). Define the signal-to-interference-and-noise ratio (SINR) of the kth linearly equalized stream as

$$\mathrm{sinr}_k = \frac{\mathrm{E}\left\{|\mathrm{sig}_k|^2\right\}}{\mathrm{E}\left\{|\mathrm{int}_k|^2\right\} + \mathrm{E}\left\{|\tilde{n}_k|^2\right\}} \tag{2.19}$$

$$= \frac{\mathrm{snr}|\boldsymbol{w}_k\boldsymbol{h}_k|^2}{\mathrm{snr}\sum_{\ell=1,\ell\neq k}^{K}|\boldsymbol{w}_k\boldsymbol{h}_\ell|^2 + \boldsymbol{w}_k\boldsymbol{w}_k^H} \tag{2.20}$$

The resulting sum-rate is given by

$$C_{\text{lin}} = \sum_{k=1}^{K} \log_2 (1 + \text{sinr}_k). \tag{2.21}$$

One general drawback of linear equalization is clearly seen by comparing (2.21) with (2.7). That is, some channel energy portraying interference, appears in the denominator of sinr_k resulting in a sum-rate loss.[4] Fortunately, the impact of the interference in the sum-rates is reduced as M grows, as it will be shown next.

2.5 Benefits of Operating with Many Antennas

Many of the benefits of using massive MIMO systems, for both communication and pure channel sounding systems, result from *averaging* effects when M grows large. Central in this analysis is the application of the weak law of large numbers (LLNs). For the model in (2.2), an application of the weak LLNs provides

$$\frac{1}{M} \boldsymbol{h}_k^H \boldsymbol{h}_\ell \to 0, \quad \ell \neq k, \quad \text{as} \quad M \to \infty, \tag{2.22}$$

and

$$\frac{1}{M} ||\boldsymbol{h}_k||^2 \to 1, \quad \text{as} \quad M \to \infty, \tag{2.23}$$

where convergence is defined in probability.

In massive MIMO jargon, the term *favorable propagation* is used to refer to the asymptotic result (2.22), and the term *channel hardening* is used to refer to the asymptotic result (2.23) [13]. Note that in the limit, these properties have no physical significance since no practical system can accommodate an infinite number of fixed-size antennas. Nevertheless, it allows for an understanding of, e.g., the fundamental behavioral trends associated with increasing M to a large number. Only mild channel conditions need to hold for the asymptotic effects of (2.22) and (2.23) to be present [26]. We will comment on the impact of such effects to the system performance shortly.

Most of the massive MIMO benefits are reciprocal (i.e., they exist in both uplink and downlink directions in roughly the same form). Thus, for simplicity we only address the uplink case. Below, we list some of the advantages of

[4] This is not the case in LS detection, since no inter-user interference exists. Instead, the drawback is noise amplification, especially when $\boldsymbol{H}$ is an ill-conditioned matrix—see the closed-form of its SINRs in [1].

operating with a large number of BS antennas but before doing so, we note that

$$C = \sum_{\ell=1}^{K} \log_2 (1 + \text{snr}\lambda_\ell) \tag{2.24}$$

$$\overset{a)}{\leq} \sum_{\ell=1}^{K} \log_2 \left(1 + \text{snr}||\boldsymbol{h}_\ell||^2\right) \tag{2.25}$$

$$\overset{b)}{\approx} \sum_{\ell=1}^{K} \log_2 \left(1 + \text{snr}\text{E}\left\{||\boldsymbol{h}_\ell||^2\right\}\right). \tag{2.26}$$

$$= K \log_2 (1 + M\text{snr}). \tag{2.27}$$

The inequality *a)* follows from Hadamard's inequality. We now remark on the benefits of massive MIMO.

- *Channel Hardening*: At first glance, it might appear that the approximation *b)* does not reach equality as $M \to \infty$. This is because $||\boldsymbol{h}_\ell||^2$ does not converge to $\text{E}\left\{||\boldsymbol{h}_\ell||^2\right\}$ as $M \to \infty$. However, (2.25) converges to (2.26) in probability—see proof in Appendix A. The main reason is because of (2.23), which is a very desirable effect in massive MIMO termed channel hardening.

- *Favorable Propagation*: An outcome of (2.22) is that scaled inter-user channel vectors *become* orthogonal as $M \to \infty$. As a result, we have $\lambda_k/||\boldsymbol{h}_k||^2 \to 1$. This tightens inequality *a)* at large M (in a similar fashion to that of the proof given in Appendix A).

- *Array Gain*: In general, array gain represents an increase in the effective SNR due to an increased number of antennas and coherent combination of the respective received signals. (For example, verify the growth of (2.19) when MR equalization is performed.) In the case of sum-rate capacity, it is seen by the factor M in (2.27)—a scaling of the effective SNR, i.e. $M\text{snr}$. In communications, array gain only increases capacity logarithmically, but there are other MIMO applications where array gain has a more prominent role (see end of this section).

- *Multiplexing Gain*: It is originally defined in point-to-point MIMO systems as the ratio of the capacity slope with $\log_2(\text{snr})$ at large values of snr, i.e., $\lim_{\text{snr}\to\infty} C(\text{snr})/\log_2(\text{snr}) = K$. The benefict of massive MIMO is that this multiplexing gain K can be harvested at moderate snr values, due to the large M.

- *Deterministic Uniform Coverage*: With increasing M, the instantaneous rates become deterministic. Moreover, we approach uniform coverage of $\log_2(1 + M\text{snr})$ bits-per-channel-use to all K users.
- *Sum-Rate Capacity*: All of the previous mentioned properties play in favor of one of the most fundamental metrics in communication systems, i.e., the sum-rate capacity. This motivates the use of massive MIMO for communications.

The previous properties motivate the use of large antenna arrays from a fundamental point-of-view, as they were obtained directly from the original model (2.2). However, there are also practical reasons to employ massive MIMO. From a computational point-of-view we have

- *Linear Processing:* For a fixed noise variance σ^2, we have that $\text{sinr}_k, \forall k$ in (2.21) scales linearly with M for both ZF or MR equalizers. Using L'Hôpital's rule, one can show that

$$C/C_{\text{lin}} \to 1, \quad \text{as} \quad M \to \infty, \tag{2.28}$$

  which makes linear processing asymptotically optimal from a sum-rate point-of-view. When it comes to symbol detection in a single channel use, ML detection can be replaced by linear equalization at large M since the SINR of the latter scale linearly with M.

There are also beneficial practical outcomes from the increased array gain characteristic of massive MIMO. They include

- *Coverage:* The large array gain provided by increasing M can be exploited in different practical ways. For example, *i)* the system can extend its coverage for the same transmit power, *ii)* the system can maintain its coverage when moving up to higher carrier frequencies—yielding higher path-losses, or *iii)* the system can scale down the radiated power by M and maintain its coverage. In all of the above, the overall receive SNR can be maintained constant and therefore there is no performance degradation.
- *Hardware Requirements:* One practical outcome of operating with many antennas and being able to reduce the radiated power, is that the hardware requirements of the system, mainly on the analog front-ends, can be relaxed. The system can therefore operate with low-end (and therefore cheaper) hardware components. Under some conditions, some examples are the reduction of number of analog-to-digital converter (ADC) quantization bits [27, 28], and reduction of the linear range of power amplifiers [29].

For pure channel sounding-based applications we have

- *Parameter Inference:* Increasing the number of antennas typically brings many advantages from an inference point-of-view. For example, there are occasions where in a single channel snapshot, M can be seen as the number of independent observations of a random event. Thus, for a parametrized model, a higher number of model parameters can be estimated when M is higher. Also, and perhaps more interestingly, a higher M typically translates into inference robustness. This is because, in statistical signal processing, the variance of an efficient estimator decreases linearly with the number of independent observations.

The benefits mentioned above motivate the use of massive MIMO from many standpoints. However, with such unprecedented MIMO setups also comes a number of interesting challenges that need to be investigated. These challenges range from fundamental ones (e.g., inter-cell pilot contamination [22], the extent of channel reciprocity in TDD radio systems, and training overhead in FDD operation [23]), to practical challenges (e.g., aspects of system design and optimization [30], the impact of hardware non-idealities in the system performance [31], and out of band radiation aspects [32]), to challenges in the applicability of the technology (e.g., for security [33], and for positioning [34]). This thesis studies several of such challenges. We start with the first research topic in the next chapter.

Chapter 3

Realizing massive MIMO communications

In this chapter we put the first research topic of this thesis into context: the design and implementation of a real-time (RT) massive MIMO testbed primarily for communications. We provide a brief survey on existing testbeds, and comment on the challenges associated with designing RT testbeds that operate with a large number of antennas at the BS side. We finalize with a summary of the contributions presented in this chapter.

3.1 The Role of Testbeds and Prototyping

One of the main goals of wireless research is to provide insights on design and resulting performance of wireless systems. Within the extensive list of researched topics in communications, there are many instances of proposed technologies claiming very promising practical outcomes. Consider, e.g., the recent cases for physical layer technologies such as Orbital Angular Momentum (OAM) [35], millimeter-wave [36], massive MIMO, and visible light-based communications [37]. Once the understanding of such topics matures and of the most body of theoretical research point coherently in the same direction, the next typical step is to realize such conceptual technologies and verify their feasibility. This crucial step can be achieved by means of proof-of-concept platforms, or testbeds. Testbeds are key elements in the path *from theory to practice* from many points of view. We remark on some of them below.

1) Testbeds can be used as tools to do scientific studies. For example, they can be used to understand to which extent most theoretical claims can be

harvested in practice [38]. Also, testbeds can be used to validate new methodologies (e.g., an algorithm for calibration [39]) by means of implementation and analysis of measured performance metrics. Testbeds can also be used to diagnose if (each of) the modeling assumptions and/or eventual analytical approximation steps during the design of a method hold in practice.

2) Testbeds aid the system design from a hardware point of view. Explained briefly, the design criteria for testbeds are typically not in line with those that are used to define a (finalized) commercial product. Testbeds are often over-designed for flexibility and performance, and under-designed for power consumption and other practical aspects as physical dimensions. Nevertheless, the experience of building and operating such prototypes often provide insights and guidelines for an optimized system design. For example, a common use for testbeds in a massive MIMO context is to aid inferring minimum hardware requirements in order to harvest most performance under realistic channel conditions, e.g., *how many antennas and how many ADC quantization bits are needed in practice?* [27, 40].

3) Functional communication prototypes are engineering milestones. They serve as key vehicles to enforce an industrial push of the respective technology towards commercialization. This includes, e.g., product development and standardization. Testbeds are also useful assets for technological statements by research institutions. These statements are typically done by reporting experimentally achieved spectral efficiencies or throughputs (e.g., [41]).[1]

We now provide an overview of existing massive MIMO testbeds as of today.

3.2 Existing massive MIMO testbeds

Currently, there exist massive MIMO testbeds of different kinds. Some of their categories include: RT or not, for communications or for pure channel sounding only,[2] standard compliant or not, proprietary or not. By RT operation, we refer to systems equipped with processing capabilities and other functionalities able to meet certain throughput/latency criteria depending on the targeted application. For example, cellular communication testbeds may qualify for RT if they yield aggregated processing latencies lower than the coherence times of time-varying cellular channels (typically hundreds of micro-seconds) [9]. Let us discuss some of the existing massive MIMO testbeds and put them into context. The testbeds we will discuss, and some of their key features, are listed in Table

[1] Unfortunately, scientific aspects such as associated error metrics are often forgotten to be reported in such press-releases.

[2] We remark that, in coherent communications [24] some system resources are used for training, which can also qualify as sounding. Thus, existing testbeds performing coherent communications also have incorporated channel sounding capabilities.

3.1. We start with testbeds owned by academic institutions and finalize with testbeds owned by cellular infrastructure manufacturers.

Table 3.1: Existing Massive MIMO testbeds (ordered by year of publication). The testbed project, written in bold letters, is one of the main contributions of this thesis.

Institution	Year	BW [MHz]	$M \times K$	RT	Spectral eff.
Rice U. [42]	2012	0.625	64×15	No	–
CSIRO ICT. [43]	2012	14	$- \times 14$	No	67 bps/Hz
Microsoft [44]	2013	-	$- \times 12$	Yes	-
LUND U. [21, 45]	2014	20	100×12	Yes	145 bps/Hz
Facebook [46]	2016	-	94×24	Yes	71bps/Hz
EURECOM [47]	2017	5	64×4	Yes	$LTE-like$
ZTE [48]	2017	-	128×16	Yes	2.1 Gbps
Huawei [49]	2017	20	$- \times 16$	Yes	33 bps/Hz

Perhaps the first known testbed since massive MIMO's seminal paper [22] is the Argos 64-BS antenna testbed [42] from Rice University.[3] It qualifies as a channel sounder, which was mainly used to validate the initial massive MIMO sum-rate claims by means of SINR measurements. There is also EURECOM's testbed [47]—an open source TDD-based RT massive MIMO testbed aiming to be compliant with the LTE standard. Also, in collaboration between Universitè de Sherbrooke, Nutaq claims to prototype TDD massive MIMO using Virtex-6 field-programmable gate arrays (FPGAs) for processing in order to enable RT operation [51]. Their system specifications are still not very clear. Other testbed projects include Microsoft's BigStation [44], Facebook's Aries project [46], and the Ngara demonstrator by CSIRO ICT Centre [43]. Here we have not mentioned massive MIMO testbed projects as the ones at the University of Bristol, NTNU, Southeast University (China), and KU LEUVEN [52, 53]. The reason is that they share most architectural principles, including some hardware components, with the testbed studied in this thesis, namely the Lund University massive MIMO (LuMaMi) testbed [21, 45].[4] A picture of the LuMaMi testbed is shown in Fig. 3.1 for completeness, but we postpone the discussion of our work in this matter to Paper I and Paper II, which can be found in Part II of the thesis.

Several cellular infrastructure manufacturers have made press releases about their massive MIMO testbeds, field trials, and achieved spectral efficiencies. For example, Optus and Huawei claimed an infield trial with a spectral efficiency

[3] Prior to the Argos work, other many-antenna measurement systems have long existed, .e.g., the RUSK channel sounder [50]. However, in the context of this thesis they do not qualify as, so-called, *testbeds* since their are not flexible and modular.

[4] The spectral efficiency value shown in Table 3.1 was achieved with the University of Bristol 128×22 testbed setup based on the same framework as the LuMaMi testbed, see [41].

Figure 3.1: The BS of the LuMaMi testbed. In the front sits an antenna array made out of 160 patch antennas.

of about 2 bps/Hz/user over sixteen terminals [49]. In a demonstration at the Mobile World Congress (MWC) 2017, ZTE reported a field trial where sixteen 256-QAM modulated data streams were spatial multiplexed in order to achieve a sum-throughput of 2.1 Gbps [48] (no bandwidth value was reported).

3.3 Missing Pieces in the Literature

As described in Section 3.2, efforts from many research institutions have been put into the design and implementation of massive MIMO testbeds. However, up to until Paper I and Paper II of this thesis, several aspects still deserved further investigation.

First, compared to the entire body of theoretical work, only a relatively small amount of publications addressing the performance of massive MIMO in realistic environments did exist. One of the main reasons for this can be seen in Table 3.1 and respective references: a significant number of the existing systems are proprietary. Most published material by such proprietary projects is, to a large extent, in the form of institutional statements rather than scientific reports that truly aim to advance the overall understanding of the field. Moreover, the list of existing non-proprietary systems reduces even further when they are constrained to yield RT processing requirements. This is of primary interest when the goal is to analyze the impact of specific system blocks (e.g., different precoding algorithms) or even of the entire system under RT conditions. For example, only RT testbeds can provide reliable information to whether massive MIMO *works* in practical dynamic channels (e.g., to perform propagation studies in dynamic environments as in [54]).

Second, up to until Paper I and Paper II of this thesis, only a few publications addressed architectural guidelines, both from system-level and hardware requirements, for the design of RT massive MIMO TDD-based systems.[5] Architectural frameworks for the design of TDD-based massive MIMO systems are important since many novel design challenges arise due unprecedented use of a large number of BS antennas and reciprocity-based operation.

3.4 Design Challenges

Let us shed some light on some design challenges related to RT massive MIMO systems by putting them into perspective with the following use case. In the

[5] As remarked previously, the design challenges lie mostly at the BS side, since one *welcome consequence* of massive MIMO is that the single-antenna terminals only perform simple processing operations. Hence, we focus our discussion on BS aspects.

uplink, say that an 100-antenna BS quantizes the received BB (complex) signals at 120 MS/s with 32 bits per sample—16 bits for the in-phase component and 16 bits for the quadrature component. This constitutes a total amount of incoming sampled BB signals of $R_{\text{ag}} = 384$ Gbps that needs to be processed in RT by the BS. Moreover, hardware synchronization during the acquisition of these samples needs to occur in order to enable coherent processing. To handle the nature of TDD systems, the hardware must also switch from uplink to downlink operation, and vice-versa, in fractions of milli-seconds—the coherence times of cellular channels. Thus, with most of today's off-the-shelf hardware options, realizing a flexible and modular BS architecture capable of synchronously acquiring, shuffling, and processing such orders of magnitude of BB data in RT is far from being a trivial task. Below, we detail several challenges that compose this problem.

- *Synchronization*: a level of synchronization between the BS radio frequency (RF) front-ends must occur. In a TDD massive MIMO context, the BS requires time and frequency synchronization between each of the M RF front-ends for coherent processing of signals. Time synchronization ensures that low deviations from the specified sampling rate (e.g. 120 MS/s) occur at the M ADCs, while frequency synchronization ensures that the carrier offsets are *small*, e.g., much smaller than the subcarrier spacing in an OFDM system.[6] There are also synchronization aspects that need to be considered from a BB point-of-view. These aspects are considered next.

- *Data Shuffling*: the BS needs to route large amounts of BB data between the RF front-ends and the (possibly centralized) processing units. Since hardware components have finite throughputs for their input/output ports, as well as for the buses that interconnect each of the hardware components, proper system design is required in order to avoid routing bottlenecks. In addition to such throughput limitations, each data transfer between system elements needs to be executed in a (close to) deterministic low-latency fashion. As justified previously, such strict data transport requirements are needed to support massive MIMO operation in, e.g., high mobility environments.

- *Data Processing*: In conjunction with deterministic low-latency data

[6] Note that these requirements are still less strict than those of classical array processing systems. Here, the antenna array response and the frequency responses between each transceiver chain must also be calibrated such that the system is able to beamform in (physical) angles [2]. For uplink (massive) MIMO transmission, these two effects are embedded in the uplink channel estimate and thus do not need to be taken explicitly into account when coherently detecting uplink payload data.

transportation, it is crucial to have low-latency processing units able to perform matrix operations. In massive MIMO, one challenge is that the size of these matrices scale up with M and K—thus optimized methods for operations such as matrix inversion or factorization are of interest. They are crucial to ensure, for example, that precoding coefficients are always up-to-date. This requires that the following sequence of events is executed with latencies much smaller than that of the channel coherence time, namely, *i)* per-antenna uplink channel estimation, *ii)* routing channels estimates to a centralized processing unit, *iii)* computation of MIMO precoder and *iv)* routing precoded signals to their respective RF front-ends. On a different note, it is also important to consider architectures that can parallelize the entire processing load as much as possible. For example, it is convenient—from an implementation point-of-view—to process different sub-carriers of an MIMO OFDM signal independently.

- *Scalability and Modularity*: Testbeds suitable for research should yield scalable and modular architectures. The former allows flexibility in the system parameters, e.g., with respect to M and K. This is important to explore the impact of different parameterizations in the overall performance of the system. The latter, offers the possibility to easily and rapidly swap system units. Among other factors, this allows for verification of the performance impact of different implementations of the same system block, e.g., the performance impact of different MIMO precoders.

3.5 Contributions

The contributions associated with this chapter are summarized below.

3.5.1 Paper I: A flexible 100-antenna testbed for Massive MIMO

This paper presents the design of a massive MIMO testbed, where the BS operates with 100 coherent radio-frequency transceiver chains based on software-defined radio technology. It addresses the design challenges discussed in Section 3.4. The design considers RT MIMO precoding and decoding, which is distributed across 50 Xilinx Kintex-7 FPGAs with PCI-Express interconnects. Our design yields unique features as: (i) high throughput processing of 384 Gbps of RT BB data in both transmit and receive directions, (ii) low-latency architecture with channel estimate to precoder turnaround of less than 500 micro seconds, and (iii) a flexible extension up to 128 antennas. In this paper,

we detail the goals of the testbed, discuss the signaling and system architecture, and show initial non-RT measured results for an uplink Massive MIMO over-the-air transmission from four single-antenna user equipments (UEs) to 100 BS antennas. Note that this paper constitutes the first public presentation of a massive MIMO testbed design capable of RT operation with an LTE-like physical layer.

3.5.2 Paper II: The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation

This paper provides a framework for designing massive MIMO testbeds in general. It therefore generalizes the work of Paper I. The framework considers hardware and system-level requirements, such as processing complexity, duplexing mode and frame structure. In contrast to Paper I, it also addresses a generic system and processing partitioning which allows centralized signal processing operations to be distributed onto a multitude of physically separated processing units. The design is validated with RT proof-of-concept measurements, both in the uplink and downlink directions, in a TDD fashion. This 100-antenna testbed BS multiplexes 12 UEs in the same time/frequency resource using LTE-like OFDM signaling and 20 MHz of bandwidth.

Chapter 4

Transceiver Calibration for Reciprocity

In this chapter, we address the second research topic of this thesis: transceiver calibration. The chapter is structured so that it sheds some light on the different areas that this topic spans. These range from wireless propagation aspects, to impairments in the transceiver chains. We conclude the chapter by clarifying some of the terminology used in the publications of the thesis, and by summarizing the scientific contributions.

4.1 Reciprocity of Radio Channels

Many works in the context of wireless systems invoke the reciprocity argument for the propagation channel. This is also the case in many standard textbooks, including those focusing on channel propagation aspects as modeling and characterization from a system perspective [11, 55, 56]. More often than not, reciprocity is solely invoked as argument to justify a specific purpose (e.g., to justify TDD-based approaches), and a thorough discussion on channel reciprocity, including its conditions, is hard to find. (Instead, this topic can be found in classical electromagnetic textbooks, e.g. [57], but the involved mathematics may be hard to follow to the standard signal processing engineer.) Next, we try to shed some light on this concept as it is an essential background assumption of this thesis.

One element of any antenna system is the propagation channel, i.e., the medium between the transmitting and receiving antennas. Recalling a well-known reciprocity theorem in electromagnetics—the Lorentz Reciprocity

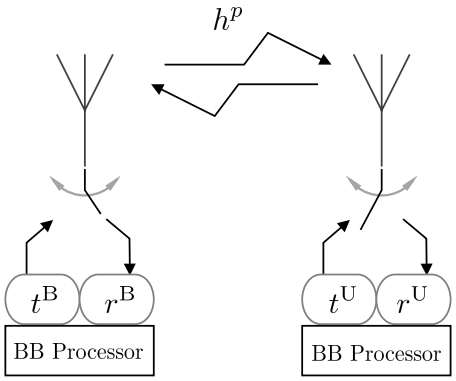

Figure 4.1: Radio link between a BS and a UE. From a BB point-of-view, the radio channel is composed by the cascade of the following components: the transmitter RF-front end, the reciprocal part of the channel h^P, and the receiver RF-front end.

Theorem—a medium is considered to be reciprocal if, quoting [58], *"the coupling of one set of (radiating) sources* $\{\boldsymbol{J}_1, \boldsymbol{M}_1\}$ *with the corresponding (electric and magnetic) fields of another set of sources* $\{\boldsymbol{E}_2, \boldsymbol{H}_2\}$ *must be equal to the coupling of the second set of sources* $\{\boldsymbol{J}_2, \boldsymbol{M}_2\}$ *with the corresponding fields of the first set of sources* $\{\boldsymbol{E}_1, \boldsymbol{H}_1\}$*, and vice versa."*. This implies that

$$\int_{\mathbb{V}} (\boldsymbol{E}_1 \cdot \boldsymbol{J}_2 - \boldsymbol{H}_1 \cdot \boldsymbol{M}_2)\,\mathrm{dV} = \int_{\mathbb{V}} (\boldsymbol{E}_2 \cdot \boldsymbol{J}_1 - \boldsymbol{H}_2 \cdot \boldsymbol{M}_1)\,\mathrm{dV}, \tag{4.1}$$

where $\mathbb{V}$ is a volume enclosed in the medium, and $\cdot$ denotes the scalar product.[1] Conditions for reciprocity also exist: the two involved sets of sources radiate at the same frequency and in the same propagation medium, which is assumed to be linear, isotropic, but not necessarily homogeneous. Noticeably, these conditions for the medium are in line with our understanding about the dominant propagation mechanisms of wireless propagation channel at radio frequencies (e.g., free space propagation, reflection, and diffraction)—see Chapter 4 of [11]. Therefore, in the context of the theorem stated above, the propagation channel should not compromise the reciprocity of an antenna system. In fact, this

[1] We can get a bit of intuition from (4.1) if we consider the special case when the two radiating sources are electric point dipoles. With that, (4.1) simplifies to $\boldsymbol{E}_1 \cdot \boldsymbol{J}_2 = \boldsymbol{E}_2 \cdot \boldsymbol{J}_1$ which resembles the Rayleigh-Carson Reciprocity theorem [59] widely known in circuit analysis.

reciprocity assumption is widely accepted within the wireless community, and can be found in many standard textbooks.

However, most of today's antenna systems are also equipped with non-linear unilateral components at each end of the wireless link. A medium with such components do not satisfy the conditions of the theorem stated above. In order to gain understanding on a generic antenna system setup, consider the sketch of a narrowband TDD-based SISO link illustrated in Fig. 4.1. The BS is equipped with an antenna, a transceiver RF front-end, and a BB processing unit. Its RF front-end has a transmitter and a receiver chain, each with non-zero responses t^B and r^B, respectively. In general, we assume $t^B \neq r^B$-we remark on this at the end of the section. Each chain is multiplexed depending on the link direction, thus the transceiver element is, in general, unilateral. The same description holds for the UE, where the transmitter and receiver non-zero responses are written as t^U and r^U, respectively. In this work we assume all transceiver responses to be linear.[2] The term h^P denotes the part of the link that can be considered linear and reciprocal. It consists of the cascade of the propagation channel and antenna responses (since they are assumed to have reciprocal radiation patterns [58]). With that, the BB responses of the uplink and downlink radio channels can be written as

$$h = r^B h^P t^U \quad \text{and} \quad b = r^U h^P t^B, \tag{4.2}$$

respectively.

This model can be easily extended to the MIMO case. Again, let the subscript m index the transceivers at the BS end of the link—for example, t_m^B and r_m^B are the transmitter and receiver responses of the mth BS transceiver. Similarly, let t_k^U and r_k^U be the transmitter and receiver responses of the kth user transceiver. Define $\boldsymbol{T}^U = \text{diag}\left\{\left[t_1^U \cdots t_K^U\right]\right\}$, $\boldsymbol{T}^B = \text{diag}\left\{\left[t_1^B \cdots t_M^B\right]\right\}$, $\boldsymbol{R}^U = \text{diag}\left\{\left[r_1^U \cdots r_K^U\right]\right\}$, and $\boldsymbol{R}^B = \text{diag}\left\{\left[r_1^B \cdots r_M^B\right]\right\}$, where the operator $\text{diag}\left\{[a_1 \cdots a_M]\right\}$ creates a diagonal matrix with diagonal entries given by $a_1, \cdots, a_M$, respectively. Using the same notation as in Section 2.2, the generalization of (4.2) to a MIMO system can be written as

$$\boldsymbol{H} = \boldsymbol{R}^B \boldsymbol{H}^P \boldsymbol{T}^U \qquad \text{and} \qquad \boldsymbol{B} = \boldsymbol{R}^U \left(\boldsymbol{H}^P\right)^T \boldsymbol{T}^B. \tag{4.3}$$

From a radio channel point-of-view, the condition for reciprocity is that

$$\boldsymbol{H} = \boldsymbol{B}^T \iff \boldsymbol{H}^P = \left(\boldsymbol{T}^B\right)^{-1} \boldsymbol{R}^B \boldsymbol{H}^P \boldsymbol{T}^U \left(\boldsymbol{R}^U\right)^{-1}, \tag{4.4}$$

[2] Certain care needs to be taken when using this assumption. For example, it requires operation in the linear region of the transmitter power amplifier [60].

which, for any uplink propagation channel $\boldsymbol{H}^P \in \mathbb{C}^{M\times K}$, boils down to

$$\frac{t_m^B}{r_m^B} = \frac{t_k^U}{r_k^U}, \quad \forall m, k. \tag{4.5}$$

There are two main things to note from (4.5). First, most practical systems as of today are not designed to satisfy (4.5). One can conjecture that attaining (4.5) solely with proper system design may be a very constraining and challenging task. For example, the transmitter and receiver of the same transceiver chain should share the same oscillator reference signal, otherwise there will be an independent phase ambiguity in both of their responses during their boot—and hence in their responses' ratio. Also, due to process, voltage and temperature (PVT) variations during circuit manufacturing and operation, and due to possibly different designs in general, the transmitter and receiver chains in general have independent frequency responses. For this reason a radio MIMO channel, in the form of (4.3), is not considered reciprocal in general. Second, the condition to achieve reciprocity is that the *ratios* of the transmit and receive responses of all system RF-chains must be equal. In terms of calibration, this is a looser constraint than aligning all transmitter and receiver responses individually, as in many classical array beamforming applications. This also tells us that, if some sort of calibration that attains reciprocity needs to occur—so that, e.g., downlink precoding can be performed using uplink channel estimates—it can focus on aligning the ratios only.

4.2 Calibration Strategies for enabling Reciprocity

Below, we overview some of the existing calibration approaches that enable the reciprocity assumption for a radio MIMO channel.

One standard approach to re-establish reciprocity is based on dedicated hardware circuitry and digital signal processing techniques. This solution is very common, not in the context of massive MIMO communications, but in pure MIMO channel sounding(-based) systems [55]. Certain hardware-based calibration setups may provide means to calibrate $\boldsymbol{T}^U, \boldsymbol{T}^B, \boldsymbol{R}^U$, and $\boldsymbol{R}^B$ individually. Also, they require a special design and integration of the calibration hardware with the entire measurement system. All the intricacies of designing and implementing dedicated hardware for calibration—which may scale up with M—may not make this approach suitable to integrate massive MIMO communication systems. Moreover, it is also not a primary calibration option to be integrated in most testbeds and prototypes, since here the priority during

the system design is typically to focus on the fundamental design challenges (e.g., as described in section 3.4). Instead, a more convenient calibration option relies on over-the-air measurements (and no extra calibration hardware). Since the non-reciprocal channel entities lie at both ends of the radio link, the most straightforward conceptual approach is to perform bi-directional measurements between the BS and the UEs. This allows capturing all non-reciprocal system elements in the measurement set. This measurement set is then used to calibrate $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ and $\left(\boldsymbol{R}^U\right)^{-1}\boldsymbol{T}^U$, e.g., see [61]. The main drawback of this approach is its lack of robustness since the calibration quality depends on the over-the-air link from the BS to the UEs, which may fade or have high path-loss. The remedy for this is to focus only on calibrating the non-reciprocal part of the channel at the BS side only, i.e., $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$. Estimating $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ can be done with over-the-air measurements at the BS side only. This approach is stable in the sense that, with a co-located array, the channels between BS antennas have a dominant deterministic component (Paper IV discusses this in detail). Moreover, less training overhead is in general required for calibration compared to estimating both $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ and $\left(\boldsymbol{R}^U\right)^{-1}\boldsymbol{T}^U$. The price to pay for not estimating $\left(\boldsymbol{R}^U\right)^{-1}\boldsymbol{T}^U$ is an eventual sum-rate loss, which depends on the magnitude variations of the non-calibrated diagonal entries of $\left(\boldsymbol{R}^U\right)^{-1}\boldsymbol{T}^U$- see [62] for a thorough analysis on this matter. Nevertheless, the calibration approaches considered in this thesis for massive MIMO TDD calibration focus only on estimating the diagonal entries of $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$.

4.3 A Remark on the Model

The model (4.3) assumes that $\boldsymbol{R}^B$ and $\boldsymbol{T}^B$ are diagonal matrices. However, there may exist impairments that compromise this diagonal structure (and therefore the diagonal structure of $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$). According to [63,64], impairments that qualify for this matter are, e.g., cross-talk between BS RF-chains, and strong parasitic interaction between closely spaced BS antenna elements (usually termed mutual coupling[3]). The nature of such impairments introduce a non-reciprocal dependence between the entries of $\boldsymbol{H}$ and $\boldsymbol{B}^T$, and therefore cannot be embedded in $\boldsymbol{H}^P$—it needs to be modeled in the non-diagonal entries of $\boldsymbol{R}^B$ and $\boldsymbol{T}^B$.

The question that remains is how reasonable this diagonal assumption is in practical systems. This exact topic was addressed in [64]. The authors estimated $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ by means of experimental data obtained using a 4×1

[3] We note that the term mutual coupling may refer to different phenomena depending on the context. We clarify this in Section 4.4.

multiple-input single-output (MISO) prototype. One focus on the experiment was to estimate $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ with and without the diagonal assumption, and see how the results differ. There are many interesting aspects of this experiment, however one of main interest is the aspect of BS antennas being a quarter of wavelength apart. This close spacing makes the current case study of high interest, as strong mutual coupling is more likely to occur with such small antenna spacings. Nevertheless, one of the main outcomes of the work was that the diagonal elements of $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ dominate over the non-diagonal elements, and therefore the diagonal assumption is "reasonable" [64]. This, of course, depends mainly on the properties of the system at hand, but just like other works in MIMO reciprocity calibration [20, 61], the work of this thesis is done based on the assumption that $\boldsymbol{R}^B$ and $\boldsymbol{T}^B$ are diagonal matrices.[4] As it will be seen in the publications associated with this chapter, this assumption will prove to be of very useful analytical convenience.

4.4 A Remark on Mutual Coupling

The term mutual coupling (or antenna coupling) is often used to describe different phenomena in different areas of study. For example, in antenna array design the term mutual coupling is typically used to refer to the undesirable parasitic interaction between closely spaced antennas [58]. Possible outcomes of these *ping-pong* effects are the altering of the mutual impedances of the antennas, and consequently their radiated contributions to the far-field pattern (compared to the ideal case without coupling). However, from a circuit analysis point-of-view, the term (antenna) coupling is simply the phenomenon of one circuit (with an antenna) inducing a current, or voltage source, onto another circuit (with an antenna). Therefore, it occurs not only in the context of closely spaced antennas but when antennas are in the far-field. Strictly speaking, all antenna systems interact (i.e., communicate) through antenna coupling.

In the context of the publications associated with this thesis, we do not use the term mutual coupling to refer to parasitic effects between closely spaced antennas (and respective outcomes thereof). We use mutual coupling to refer to a channel component that allows reliable signaling between two nearby BS antennas (e.g., two BS antennas in a co-located array). The energy of this component typically depends on factors such as the array configuration (including the antenna elements themselves), and it does not follow the standard

[4] Assuming $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ to be a diagonal matrix as in [64] does not necessary imply that $\boldsymbol{R}^B$ and $\boldsymbol{T}^B$ are diagonal matrices. However, the exceptional cases where this occur are not likely to occur in realistic situations, and thus for all practical purposes we use the diagonal assumption on $\boldsymbol{R}^B$ and $\boldsymbol{T}^B$.

d^{-n} decay power law, with $1.5 \leq n \leq 4$ and d being the distance between antennas [11]. This channel component is assumed to be non-fading, reciprocal, and typically dominant over other channel multipath components. This mutual coupling terminology has also been used in other works [65, 66].

4.5 Contributions

The contributions associated with this chapter are summarized below.

4.5.1 Paper III: Reciprocity calibration methods for massive MIMO based on antenna coupling

This paper considers reciprocity calibration of a massive MIMO system, similar to the system introduced in Section 2.1. It proposes a novel calibration method which is conducted entirely at the BS side by sounding the BS antennas one-by-one while receiving with the other BS antennas. It also deals with modeling of the dominant component of the channels between BS antennas, i.e., the component due to mutual coupling. We study a number of approaches, mostly existing in the literature, suitable to estimate calibration coefficients. Our theoretical study indicates that it should be possible to calibrate the transceivers of a massive MIMO BS array in order to re-establish the reciprocity assumption.

4.5.2 Paper IV: Reciprocity Calibration for Massive MIMO: Proposal, Modeling and Validation

This paper wraps up the work of Paper III as follows. It proposes an estimator for the calibration coefficients that outperforms all estimators found in literature. The estimator is a special case of the Expectation Maximization (EM) algorithm [67]. We implement our proposed antenna coupling-based calibration proposal in the massive MIMO testbed described in Papers I and II, and verify experimentally that our calibration method works. We also study how the calibration error behaves in frequency, i.e., across subcarriers of an OFDM system, and propose a wideband estimator that reduces the error across frequency. Finally, we propose a model for the calibration error, and validate this model with measurement results.

4.5.3 Paper V: A Receive/Transmit Calibration Technique based on Mutual Coupling for Massive MIMO Base Stations

This paper considers the calibration methodology proposed in Paper III. It applies it in order to the calibrate the transmitters $\{t_m^B\}$ and/or receiver $\{r_m^B\}$ radio frequency chains individually, as in classical array processing applications. It verifies that, in this context, more information about the model is needed in order to estimate the transmitters and receivers responses individually, compared to estimating their ratios only. This work opens up opportunities for pure channel sounding-based applications to be integrated with massive MIMO BSs.

Chapter 5

Detection in block-fading SIMO channels

In this chapter, we introduce the third research topic of this thesis. The main contribution of its associated publication is a symbol detection method for block-fading SIMO channels. To put this detection method into context, we briefly review two standard detection approaches, and introduce the framework of our proposed method.

5.1 Signal Model

Let us now re-introduce the model for block-fading communications. From (2.1), we see that block-fading channels have Rayleigh fading IID entries that are drawn once every block, and are invariant through an entire block. Each block is drawn in an IID fashion as well. The coherence time of the block—the block length $(T+1)$—is the number of channel uses in which the channel is considered to be invariant, and the index is t. Here we assume the coherent dimension to be time, but in general it can be any other available dimension, e.g., frequency. In this chapter we assume that an uplink block-fading SIMO transmission is performed.[1] With that, we can write the uplink signal model as

$$\boldsymbol{y}_t = \sqrt{\rho_u}\boldsymbol{h}x_{1,t} + \boldsymbol{n}_t, \quad 0 \leq t \leq T, \tag{5.1}$$

where $\boldsymbol{h} \triangleq \boldsymbol{h}_1$. The vectors $\boldsymbol{h}$ and $\boldsymbol{n}_t$ have the same Gaussian multivariate distributions as in (2.1). Also, we assume that $x_{1,0} = 1$ is a known inserted

[1] We motivate why this is the case in Paper VI.

training symbol, and thus $\boldsymbol{y}_0$ is the LS estimate of $\boldsymbol{h}$. Only in this chapter, we assume that there is no apriori knowledge of the current channel realization $\boldsymbol{h}$ nor of its distribution at the receiver side. Similarly to Section 2.2, we assume the elements of $\{x_{1,\ell}\}_{\ell=1}^{T}$ to be IID with unit energy, and that each element is sampled from a distribution with PMF $\mathrm{pmf}(\cdot|\mathcal{A})$. For later use, define

$$\left[\boldsymbol{y}_0 \; \tilde{\boldsymbol{Y}}\right] \triangleq \boldsymbol{Y}, \tag{5.2}$$

and

$$[1 \; \tilde{\boldsymbol{x}}] \triangleq \boldsymbol{x} = [x_{1,0} \cdots x_{1,T}]. \tag{5.3}$$

Remarkably, the SIMO block fading channel model (5.1) is suitable for many wireless scenarios of today. For example, in low-power sensor networks, where static non-synchronized single antenna nodes wake up sporadically to transmit a few packets of data in a burst.

5.2 Exploiting the Block-fading Structure for Detection

There are several ways to decode $\tilde{\boldsymbol{x}}$ based on $\boldsymbol{Y}$. Below we address the two corner cases when it comes to performance and complexity.

The simplest way is to perform channel estimation and data detection separately. For example, one can assume that the LS channel estimate, $\boldsymbol{y}_0$, is correct and perform MR combining [24], i.e.,

$$\hat{\boldsymbol{x}} = \frac{1}{||\boldsymbol{y}_0||^2} \boldsymbol{y}_0^H \tilde{\boldsymbol{Y}}. \tag{5.4}$$

The computational complexity of this approach is linear in MT. This approach is optimal, in the sense that it minimizes the symbol error rate (SER), if $\boldsymbol{y}_0 = \boldsymbol{h}$. This is true with probability 0.

When it comes to minimizing error rates, the optimal detector performs joint ML channel estimation and data detection [68]. Since each entry of $\tilde{\boldsymbol{x}}$ can take $\#\{\mathcal{A}\}$ distinct values, there are $\#\{\mathcal{A}\}^T$ equally likely hypotheses that need to be tested. With $1 \leq \ell \leq \#\{\mathcal{A}\}^T$, define hypothesis $\mathcal{H}_\ell$ as

$$\mathcal{H}_\ell : \boldsymbol{Y} = \sqrt{\rho_u}\boldsymbol{h}\boldsymbol{x}_\ell + \boldsymbol{N},$$

where $\boldsymbol{x}_\ell$ is one possible outcome of $\boldsymbol{x}$. Under $\mathcal{H}_\ell$, the channel estimate that maximizes the log-likelihood of (5.1), $p(\boldsymbol{Y}|\boldsymbol{h}) \propto -||\boldsymbol{Y} - \sqrt{\rho_u}\boldsymbol{h}\boldsymbol{x}||^2$, is

$$\hat{\boldsymbol{h}}_{\mathrm{ML}} = \boldsymbol{Y}\boldsymbol{x}^\dagger / \sqrt{\rho_u}. \tag{5.5}$$

Define the projection matrix $\boldsymbol{Q}\boldsymbol{Q}^H \triangleq \boldsymbol{I} - \boldsymbol{x}^\dagger\boldsymbol{x}$. It follows that $||\boldsymbol{Y} - \sqrt{\rho_u}\hat{\boldsymbol{h}}_{\mathrm{ML}}\boldsymbol{x}||^2 \propto ||\boldsymbol{Y}\boldsymbol{Q}||^2$, and therefore the solution to the joint ML channel estimation and data detection problem is given by

$$\hat{\boldsymbol{x}} = \arg \min_{\substack{\tilde{\boldsymbol{x}} \\ \text{s.t. } \tilde{\boldsymbol{x}} \in \mathcal{A}^{1\times T}}} ||\boldsymbol{Y}\boldsymbol{Q}||^2. \tag{5.6}$$

Without clever searching algorithms [68], the complexity of optimal detection is exponential in T.

5.3 The Generalized Method of Moments

The estimation approach for $\tilde{\boldsymbol{x}}$ proposed in our work, is an intermediate case between the previous two corner cases both in terms of performance and complexity. We came to know about this estimation approach due to the state-of-the-art estimation proposal for the calibration coefficients $\left(\boldsymbol{R}^B\right)^{-1}\boldsymbol{T}^B$ (prior to our proposed EM algorithm). We verified that the authors of [20] empirically set up a constrained LS cost function to estimate the calibration coefficients. It appears that this estimation approach was pursued mainly due to suitability purposes as it makes use of no probabilistic model assumptions—in fact, it is derived assuming noiseless received signals [20]. As it turns out, this estimation approach is not as uncommon as one might initially think. It is, in fact, one instance of the generalized method-of-moments (GMM) estimator [69]. An estimation method that contributed, in part, for the 2013 Nobel Prize award in Economics to its author Lars Hansen.

The GMM is an estimation approach which is highly used in econometrics, but not so much in communications. The key point to build a GMM estimator is to empirically find a suitable function, $\boldsymbol{f}'(\cdot)$, of the model observations $\boldsymbol{Y}$ and model variables $\boldsymbol{h}$ and $\boldsymbol{x}$, such that

$$\mathrm{E}\left\{\boldsymbol{f}'(\boldsymbol{Y}, \boldsymbol{h}, \boldsymbol{x})\right\} = \boldsymbol{0}. \tag{5.7}$$

Here $\boldsymbol{0}$ is a vector of zeros. Each entry of the column vector $\boldsymbol{f}'(\boldsymbol{Y}, \boldsymbol{h}, \boldsymbol{x})$ is termed a moment condition. To ensure a closed-form solution for the estimator, we constrain the output of $\boldsymbol{f}'(\cdot)$ to be linear in the, here assumed, deterministic but unknown parameter vector $\boldsymbol{\theta} = \left[\boldsymbol{x}\ \boldsymbol{h}^T\right]^T$. With that, we can write (5.7) as

$$\mathrm{E}\left\{\boldsymbol{f}(\boldsymbol{Y})\right\}\boldsymbol{\theta} = \boldsymbol{0}. \tag{5.8}$$

The estimator is obtained by solving the following quadratic form

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} ||\boldsymbol{V}\boldsymbol{\theta}||^2, \tag{5.9}$$

where $\boldsymbol{V}^H\boldsymbol{V} \triangleq \boldsymbol{f}(\boldsymbol{Y})^H\boldsymbol{W}\boldsymbol{f}(\boldsymbol{Y})$. Here $\boldsymbol{W}$ is a positive semi-definite matrix, so called, weighting matrix. In applications where $x_{1,0}$ is not constrained to be 1 (as it is in our case), quadratic constraints can be imposed in (5.9), e.g. $||\boldsymbol{\theta}||^2 = 1$, to avoid the all-zero solution without compromising the closed-form of $\hat{\boldsymbol{\theta}}$ due to the quadratic form of (5.9).

There is an entire theoretical framework built behind this estimation approach, see [70]. The main properties of interest are that, under an optimal setting for $\boldsymbol{W}$ and remaining conditions for optimality, $\hat{\boldsymbol{\theta}}$ is an asymptotically unbiased, asymptotically Gaussian, and asymptotically efficient estimator.

5.4 Contributions

The contributions associated with this chapter are summarized below.

5.4.1 Paper VI: A Generalized Method of Moments Detector for Block Fading SIMO Channels

This paper applies the GMM estimator, an estimator we learned from our work in reciprocity calibration, for detection in block fading SIMO channels. The estimator is obtained in closed-form, and the closed-form has an intuitive formulation. The results of this estimator are contrasted with the results of the benchmark detection approaches introduced in sec. 5.2. The paper verifies how the number of BS antennas M and the coherence interval of the channel $T + 1$ influence performance and complexity. It turns out that the obtained SNR gains are linear in M which makes this estimator very suitable for block fading *massive* SIMO channels.

Chapter 6

Combining Deep Learning and Positioning with Large Antenna Arrays

In this chapter, we address the fourth research topic of this thesis. The chapter begins by introducing the different topics relevant to understand our research contribution, which is summarized at the end of the chapter.

Also, as remarked previously in Chapter 1.4, some parts of the notation used in previous thesis chapters are re-defined here.

6.1 Radio-based Positioning Approaches

Most of the methods used for radio-based positioning fall into three categories: *i)* triangulation, *ii)* proximity, and *iii)* fingerprinting—see [71] for an overview on these approaches. Triangulation methods (based on, e.g., time-of-arrival or signal strength criteria) are widely used for outdoor positioning in propagation scenarios where one multipath component, typically the line-of-sight (LOS) component, dominates over all other components. In such scenarios, the channel(s) can be modeled accurately and the position of the mobile terminal can be triangulated reliably (e.g., the case of satellite navigation in the Global Positioning System (GPS)). However, this is typically not the case in most indoor scenarios, or even in outdoor scenarios (such as in dense urban environments) where strong shadowing effects, obstructed LOS conditions, or rich scattering may occur. Instead, the preferred approach in many areas for positioning in

such propagation conditions is fingerprinting [72].[1] Contrary to many other positioning approaches, fingerprinting can be done with only one reference anchor node [73], and therefore it can easily integrate existing standards without the need to deploy dedicated infrastructure solely for positioning purposes.

Fingerprinting using machine learning methods has traditionally been considered together with signal strength measurements [71, 74]. This was also the case in [34] when the authors attempted to perform positioning based on Gaussian processes with distributed massive MIMO BSs.[2] However, this approach is bounded to be sub-optimal since it discards (differential) phase information, which in general is of tremendous interest for positioning. It is therefore of interest to process *raw* channel fingerprints and understand their inherent structure for positioning purposes. Luckily, the field of Deep Learning has already a long history and is very suitable to understand the structure of intricate input data sets. The next sections will introduce Deep learning and formulate a learning method that is well suited for positioning based on fingerprinting.

6.2 Deep Learning

Many Deep Learning (DL) tools have already been around for several decades (e.g., the Perceptron algorithm [76]), but it was only in the latest years that they have peaked in applicability and popularity [77]. This is mainly due to the computational capabilities of today's processors that can now efficiently process data sets of large sizes, together with the ongoing development of efficient optimization techniques. Nowadays, DL methods are widely accepted as the most powerful class of learning methods, within those in the realm of Machine Learning, when it comes to tackle learning tasks of complex and non-trivial nature [78]. In recent years, DL has had tremendous success in image processing and speech recognition [79, 80], and its applicability to the field of communications is also a matter of rising interest [81, 82].

One can find many definitions for DL (or hierarchical learning), depending on the context of its application. Some definitions focus on the common structure of such learning methods with that of the human brain and therefore on one of its main applications: artificial intelligence [78]. A more pragmatic definition of DL is perhaps as follows: DL is *"a class of machine learning techniques that exploit many layers of non-linear information processing for supervised or*

[1] Proximity methods are conceptually simple and simple to implement, but the order of their error is proportional to the distance between the anchor reference nodes. They are not suitable for highly accurate positioning in general, but still have many applications. For example, one main application of proximity methods in cellular communications is for the hand-over of a user terminal between BSs [11].

[2] One of main motivation of using many antennas for positioning, is that under some channel assumptions, they can trade-off against signaling bandwidth [75].

unsupervised feature extraction and transformation, and for pattern analysis and classification." [83]. These "many layers of non-linear information processing" make DL methods very suitable to process *raw* real-world data sets with arbitrary structures such as photographic images, speech signals, etc. The main purpose of these layers is to perform parametric feature learning from the data set at hand at one stage, and projecting "new" inputs into the learned features in another stage. As a result, a DL network with optimized features (and optimized features weights) can *make use of* the intricate information structure in raw data sets to solve the problem at hand (e.g. face recognition for sexual orientation [84]).

Many DL networks still remain a black box when it comes to their understanding—many advances are attained by trial-and-error, followed by behavioral post-rationalization. Fortunately, there exist several research works that have tried to fundamentally understand these networks [85,86]. Examples of DL networks are feed forward neural networks, deep belief networks, and recurrent neural networks. In this work we focus on one special case of feed forward neural networks, namely, convolutional neural networks.

6.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are the DL networks with the largest success in image processing tasks [79, 87]. A CNN is nothing more than a tunable function $f(\cdot, \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ denoting the network parameter vector. With a given input $\boldsymbol{X}^0 \in \mathbb{X}$, the main goal of a CNN is that its output

$$\boldsymbol{y} = f(\boldsymbol{X}^0, \boldsymbol{\theta}) \tag{6.1}$$

approximates, in some sense, the output

$$\boldsymbol{y}^* = f^*(\boldsymbol{X}^0) \tag{6.2}$$

of the true underlying function we wish to learn $f^*(\cdot)$. Here $\mathbb{X}$ is the domain of $f(\cdot)^*$, which is assumed to be the set of $S_1^0 \times S_2^0$ real matrices that is spawned by $\boldsymbol{X}^0$.

The structure of a generic CNN with L layers and K convolutional kernels per layer is as follows. Layer ℓ, with $1 \leq \ell \leq L$, employs

1. A convolution of its input, $\boldsymbol{X}^{\ell-1} \in \mathbb{R}^{S_1^{\ell-1} \times S_2^{\ell-1} \times S_3^{\ell-1}}$, with K different kernels. The kth kernel is here denoted as $\boldsymbol{W}_k^\ell \in \mathbb{R}^{D_1^\ell \times D_2^\ell \times S_3^{\ell-1}}$, and we typically have $D_1^\ell << S_1^{\ell-1}$ and $D_2^\ell << S_2^{\ell-1}$. Define the output of the

convolution[3] with the kth kernel as $\boldsymbol{C}_k^\ell = \boldsymbol{W}_k^\ell \star \boldsymbol{X}^{\ell-1} + b_k^\ell$, where b_k^ℓ is a bias parameter, and $\star$ denotes the 2 dimensional convolution operator. Here $\boldsymbol{C}_k^\ell \in \mathbb{R}^{S_1^{\ell-1} \times S_2^{\ell-1} \times 1}$. For latter use, we stack all the K elements of $\left\{\boldsymbol{C}_k^\ell\right\}_k$ into the tensor $\boldsymbol{C}^\ell \in \mathbb{R}^{S_1^{\ell-1}, S_2^{\ell-1}, K}$.

2. A non-linear transformation, $\psi(\cdot)$, to each entry of $\boldsymbol{C}^\ell$. We obtain $\boldsymbol{G}^\ell = \psi\left(\boldsymbol{C}^\ell\right)$, with $\boldsymbol{G}^\ell \in \mathbb{R}^{S_1^{\ell-1}, S_2^{\ell-1}, K}$.

3. A pooling operation, $\gamma(\cdot)$, to $\boldsymbol{G}^\ell$ mainly for dimensionality reduction purposes (see [78] for pooling options). With that, we obtain $\boldsymbol{X}^\ell = \gamma\left(\boldsymbol{G}^\ell\right)$, which is the input to layer $\ell + 1$. The output dimensionality satisfy $S_1^\ell \leq S_1^{\ell-1}$, $S_2^\ell \leq S_2^{\ell-1}$, and $S_3^\ell = K$.

Here S_3^0 depends on the application but in our contribution we have $S_3^0 = 2$.[4] The non-linear transformation $\psi(\cdot)$ can take many forms, but today's default choice is the (discontinuous) REctified Linear Unit (RELU), i.e., $\psi(\boldsymbol{X}) = \max(\boldsymbol{X}, 0)$ where the maximum is applied entry-wise. The output of the CNN, $\boldsymbol{y} \in \mathbb{R}^{D_{\text{out}} \times 1}$, is typically obtained after a (full) linear transformation $\boldsymbol{y} = \boldsymbol{W}\text{vec}\left\{\boldsymbol{X}^L\right\} + \boldsymbol{b}$. The CNN parameters

$$\boldsymbol{\theta} \triangleq \left[\left[\text{vec}\left\{\boldsymbol{W}_1^1\right\}^T \ldots \left\{\boldsymbol{W}_K^L\right\}^T\right] \text{vec}\left\{\boldsymbol{W}\right\}^T \left[b_1^1 \ldots b_L^K\right] \boldsymbol{b}^T\right]^T,$$

are optimized depending on the approach at hand, .e.g., regression or classification.

A final important remark follows. Remembering that a convolution is a special case of a linear map, there are two assumptions for the inputs of CNNs, that need to hold in order for them to retain most of the learning capabilities of standard feed forward neural networks [78]. Those assumptions are:

1. The features of $\boldsymbol{X}^0$ are "local", i.e., each feature fits within a kernel size;

2. The features of $\boldsymbol{X}^0$ are invariant to translations in the input space.

CNNs are very computationally efficient learning machines in applications where these assumptions hold.

6.4 Contributions

The contributions associated with this chapter are summarized below.

[3] Note that the convolution operation is only performed in the first two dimensions. Also, the convolution output has the same dimensions than the convolution input in these two dimensions. This can be achieved by using variants of the original convolutional operation [78].

[4] We explain why this is so in Section 6.4.

6.4.1 Paper VII: Deep Convolutional Neural Networks for Massive MIMO Fingerprint-Based Positioning

This paper provides an initial investigation on the application of CNNs for fingerprint-based positioning using measured massive MIMO channels. It claims that CNNs can efficiently learn the intricate structure massive MIMO channels. The only condition is that such channels are represented in a domain yielding a sparse structure, so that the CNN input assumptions mentioned in Section 6.3 hold. The channel fingerprint set is generated by the COST 2100 channel model [88], and is used to experimentally verify our claims. In the context of this paper, $\boldsymbol{X}^0$ is a massive MIMO channel fingerprint. When it comes to its dimensions, S_1^0 is the number of antennas at the BS, S_2^0 is the number of measured frequency points of a wideband channel, and $S_3^0 = 2$ corresponds to two real components necessary to describe a complex fingerprint (e.g., magnitude and phase). Finally, the desired output $\boldsymbol{y}^*$ is the spatial coordinate associated with the fingerprint $\boldsymbol{X}^0$.

Chapter 7

Conclusions and Future Work

Massive MIMO is an emerging physical layer technique with potential to integrate many of the future wireless communication standards, in one form or another. From a positioning and localization points-of-view, the use of large antenna arrays at the BS side also opens up many interesting possibilities. Contrary to past system setups, we envision both communication and positioning technologies to be incorporated in one single cellular base station infrastructure. This would be an elegant solution from many points-of-view, not to mention that both technologies could aid each other.

Having BSs operating with a massive number of antennas opens up research questions from many perspectives. This thesis addressed some of them. We list below some of the main lessons learned and also directions that what we think would yield interesting future work.

- When it comes to transceiver calibration, one of our main contributions was an asymptotically efficient estimator. This basically closes the door when it comes to proposing better estimators that retrieve reciprocity, since calibration should typically occur at high enough signal-to-noise ratios. Not surprisingly (from a signal processing point-of-view), its asymptotic error is non-white Gaussian. Noticeably, most literature that analyzed the impact of the calibration error in the system performance do not make use of such error model, in fact, most of it did not consider the calibration error to be Gaussian at all [89, 90]. Therefore, it would be of interest to verify if most of the insights attained from the studies that considered other error models, also hold for the Gaussian case. Now, from

an experimental point-of-view, our work showed that our novel method for calibration works in practice. This was verified by means of measured downlink error vector magnitudes (EVMs) at the users side. The EVM curves saturate at high enough calibration signal-to-noise ratios. It would be interesting to understand which impairments or modeling misassumptions are responsible for this saturation, and their respective order of importance in this matter. Candidates include, e.g., the assumption that the transceiver is a linear unit, undesired aspects of strong mutual coupling between closely spaced antennas, the extent of reciprocity in practical propagation channels, and of course, measurement noise. Only with such in-depth understanding, we will be able to pinpoint which of these phenomena dominate when it comes to their impact on the error performance. This would provide better insights on an optimized design of reciprocity-based massive antenna systems.

- When it came to detectors for SIMO channels, our work led us to the generalized method of moments estimator. This estimator uses no probabilistic information on the model at hand, and therefore it can be defined for a number of applications. Even with seemingly complicated models, as the model for calibration in Paper IV, the estimator bypassed the latent variables elegantly in order to reach a nice closed-form. Also, based on the results of Paper VI, it appears to provide a fairly good trade-off when it comes to its sub-optimal performance and complexity. For all the reasons mentioned above, and also because not so many applications of this estimation approach are found in wireless literature, we believe a potential application of the GMM would be for parameter estimation in channel sounding-based applications, especially those with complicated models and low complexity (e.g., real-time) requirements.

- Finally, when it comes to our fingerprint-based positioning study, the main lesson learned is that *raw* channel snapshots can be transformed such that they fall into the category of inputs which can be efficiently processed by convolutional neural networks. Such networks can therefore understand the intricate structure of a wireless channel (which is related to its complex geometry) and use it for positioning purposes. In our experiments, it attained positioning errors in the order of fractions of wavelengths. When it comes to future work, there exist two promising research aspects. Firstly, it appears that Deep Learning is equipped with useful tools that can be used to tackle the main problem of fingerprint-based positioning, i.e., channel variations that were not captured during the fingerprinting process. This field in Deep Learning is termed *regularization*, and has been extensively studied and developed within image

processing. Secondly, our work focused on a standard implementation of a real-valued CNN for proof-of-concept purposes and it did not cover the possibility to handcraft the network to the application at hand. We have several ideas on this matter, e.g., due to the periodic complex-valued structure of (sampled) measured channels, a circular complex-valued convolutional network would fist best. Equivalently, it could be implemented by means of a cascade of fast Fourier transforms and multiplications with transformed kernels. With this design option, the CNN behavior may be simpler and more intuitive to understand. Insights obtained with this study could aid in the optimization of the design of positioning systems. Overall, given the current hype on Deep Learning methods and our investigations, I personally believe that this is the thesis topic with more potential and possibilities for future research.

References

[1] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*, 1st ed. New York, NY, USA: Cambridge University Press, 2008.

[2] T. Kaiser, *Smart Antennas: State of the Art*, ser. EURASIP book series on signal processing and communications. Amercain University in Cairo Prees, 2005.

[3] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

[4] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, pp. 585–595, 1999.

[5] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, Mar 1998.

[6] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[7] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, Aug 2003.

[8] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband*, 2nd ed. Academic Press, 2008.

[9] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 1st ed. Academic Press, 2011.

[10] P. Nuaymi, *WiMAX: Technology for Broadband Wireless Access*. John Wiley & Sons, 2007. [Online]. Available: https://books.google.se/books?id=Kvf5bdM9QlYC

[11] A. Molisch, *Wireless Communications*, ser. Wiley - IEEE. Wiley, 2010.

[12] Y. Bar-Shalom and X. Li, *Multitarget-multisensor Tracking: Principles and Techniques*. Yaakov Bar-Shalom, 1995. [Online]. Available: https://books.google.se/books?id=GfOoMQEACAAJ

[13] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of Massive MIMO*, ser. Fundamentals of Massive MIMO. Cambridge University Press, 2016.

[14] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, Feb 2002.

[15] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable Rates of FDD Massive MIMO Systems With Spatial Channel Correlation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2868–2882, May 2015.

[16] J. Flordelis, F. Rusek, F. Tufvesson, E. G. Larsson, and O. Edfors, "Massive MIMO Performance - TDD Versus FDD: What Do Measurements Say?" *CoRR*, vol. abs/1704.00623, 2017. [Online]. Available: http://arxiv.org/abs/1704.00623

[17] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid Beamforming for Massive MIMO: A Survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[18] Y. H. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 172–179, June 2013.

[19] E. L. Bengtsson, P. C. Karlsson, F. Tufvesson, J. Vieira, S. Malkowsky, L. Liu, F. Rusek, and O. Edfors, "Transmission Schemes for Multiple Antenna Terminals in Real Massive MIMO Systems," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, USA, Dec 2016, pp. 1–6.

[20] R. Rogalin, O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable Synchronization and Reciprocity Calibration for Distributed Multiuser MIMO," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1815–1831, April 2014.

[21] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Owall, O. Edfors, and F. Tufvesson, "A flexible 100-antenna testbed for Massive MIMO," in *2014 IEEE Globecom Workshops (GC Wkshps)*, Austin, USA, Dec 2014, pp. 287–293.

[22] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[23] E. Bjornson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, February 2016.

[24] M. Salehi and J. Proakis, *Digital Communications*. McGraw-Hill Education, 2007.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[26] H. Q. Ngo and E. G. Larsson, "No Downlink Pilots Are Needed in TDD Massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2921–2935, May 2017.

[27] M. Sarajlic, L. Liu, and O. Edfors, "When Are Low Resolution ADCs Energy Efficient in Massive MIMO?" *IEEE Access*, vol. 5, pp. 14 837–14 853, 2017.

[28] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink Performance of Wideband Massive MIMO With One-Bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 87–100, Jan 2017.

[29] S. K. Mohammed and E. G. Larsson, "Per-Antenna Constant Envelope Precoding for Large Multi-User MIMO Systems," *IEEE Transactions on Communications*, vol. 61, no. 3, pp. 1059–1071, March 2013.

[30] S. Malkowsky, J. Vieira, K. Nieman, N. Kundargi, I. Wong, V. wall, O. Edfors, F. Tufvesson, and L. Liu, "Implementation of Low-Latency Signal

Processing and Data Shuffling for TDD Massive MIMO Systems," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, Dallas, TX, USA, Oct 2016, pp. 260–265.

[31] U. Gustavsson, C. Sanchz-Perez, T. Eriksson, F. Athley, G. Durisi, P. Landin, K. Hausmair, C. Fager, and L. Svensson, "On the impact of hardware impairments on massive MIMO," in *2014 IEEE Globecom Workshops (GC Wkshps)*, Austin, USA, Dec 2014, pp. 294–300.

[32] C. Mollen, U. Gustavsson, T. Eriksson, and E. G. Larsson, "Out-of-band radiation measure for MIMO arrays with beamformed transmission," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[33] D. Kapetanovic, G. Zheng, and F. Rusek, "Physical layer security for massive MIMO: An overview on passive eavesdropping and active attacks," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 21–27, June 2015.

[34] V. Savic and E. G. Larsson, "Fingerprinting-Based Positioning in Distributed Massive MIMO Systems," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, Boston, MA, USA, Sept 2015, pp. 1–5.

[35] B. Thidé, H. Then, J. Sjöholm, K. Palmer, J. Bergman, T. D. Carozzi, Y. N. Istomin, N. H. Ibragimov, and R. Khamitova, "Utilization of Photon Orbital Angular Momentum in the Low-Frequency Radio Domain," *Phys. Rev. Lett.*, vol. 99, p. 087701, Aug 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.99.087701

[36] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[37] M. Z. Afgani, H. E. H. Haas, and D. A. K. D. Knipp, "Visible light communication using OFDM," in *Proc. 2nd Int. Conf. Testbeds Res. Infrastruct. Develop. Netw. Communities*, 2006, pp. 1–6.

[38] X. Gao, M. Zhu, F. Rusek, F. Tufvesson, and O. Edfors, "Large antenna array and propagation environment interaction," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov 2014, pp. 666–670.

[39] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson, "Reciprocity Calibration for Massive MIMO: Proposal, Modeling, and Validation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3042–3056, May 2017.

[40] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in Real Propagation Environments: Do All Antennas Contribute Equally?" *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 3917–3928, Nov 2015.

[41] 5G Researchers Set New World Record For Spectrum Efficiency. Accessed: 2017-10-03. [Online]. Available: https://spectrum.ieee.org/tech-talk/telecom/wireless/5g-researchers-achieve-new-spectrum-efficiency-record

[42] C. Shepard et al., "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 53–64.

[43] H. Suzuki, R. Kendall, K. Anderson, A. Grancea, D. Humphrey, J. Pathikulangara, K. Bengston, J. Matthews, and C. Russell, "Highly spectrally efficient Ngara Rural Wireless Broadband Access Demonstrator," in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, Gold Coast, QLD, Australia, Oct 2012, pp. 914–919.

[44] Q. Yang, X. Li, H. Yao, J. Fang, K. Tan, W. Hu, J. Zhang, and Y. Zhang, "BigStation: Enabling Scalable Real-time Signal Processing in Large MU-MIMO Systems," in *ACM SIGCOMM*. ACM, August 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/bigstation-enabling-scalable-real-time-signal-processing-in-large-mu-mimo-systems/

[45] S. Malkowsky, J. Vieira, L. Liu, P. Harris, K. Nieman, N. Kundargi, I. C. Wong, F. Tufvesson, V. Owall, and O. Edfors, "The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation," *IEEE Access*, vol. 5, pp. 9073–9088, 2017.

[46] Introducing Facebook's new terrestrial connectivity systems Terragraph and Project ARIES. Accessed: 2017-10-03. [Online]. Available: https://code.facebook.com/posts/1072680049445290/introducing-facebook-s-new-terrestrial-connectivity-systems-terragraph-and-project-aries

[47] X. Jiang and F. Kaltenberger, "Demo: an LTE compatible massive MIMO testbed based on OpenAirInterface," in *WSA 2017; 21th International ITG Workshop on Smart Antennas*, March 2017, pp. 1–2.

[48] ZTE Pre5G TDD Massive MIMO 2.0 sets new record for single-site peak rate at MWC 2017. Accessed: 2017-10-03. [Online]. Available: http://www.zte.com.cn/global/about/press-center/news/201702Ma/0228ma4

[49] OPTUS AND HUAWEI TAKE ANOTHER STEP TOWARDS 5G. Accessed: 2017-10-03. [Online]. Available: https://media.optus.com.au/media-releases/2017/optus-and-huawei-take-another-step-towards-5g

[50] R. S. Thoma, D. Hampicke, A. Richter, G. Sommerkorn, A. Schneider, U. Trautwein, and W. Wirnitzer, "Identification of time-variant directional mobile radio channels," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 2, pp. 357–364, Apr 2000.

[51] TitanMIMO: A 100x100 Massive MIMO Testbed Based on xTCA Standards. Accessed: 2017-10-03. [Online]. Available: https://www.nutaq.com/blog/titanmimo-100x100-massive-mimo-testbed-based-xtca-standards

[52] P. Harris, S. Zang, A. Nix, M. Beach, S. Armour, and A. Doufexi, "A Distributed Massive MIMO Testbed to Assess Real-World Performance and Feasibility," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, Glasgow, UK, May 2015, pp. 1–2.

[53] X. Yang, W. Lu, N. Wang, K. Nieman, S. J. an Hongbo Zhu, X. Mu, I. C. Wong, Y. Huang, and X. You, "Design and Implementation of a TDD-Based 128-Antenna Massive MIMO Prototyping System," *CoRR*, vol. abs/1608.07362, 2016. [Online]. Available: http://arxiv.org/abs/1608.07362

[54] P. Harris, S. Malkowsky, J. Vieira, E. Bengtsson, F. Tufvesson, W. B. Hasan, L. Liu, M. Beach, S. Armour, and O. Edfors, "Performance Characterization of a Real-Time Massive MIMO System With LOS Mobile Channels," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1244–1253, June 2017.

[55] S. Salous, *Radio Propagation Measurement and Channel Modelling*. Wiley, 2013. [Online]. Available: https://books.google.se/books?id=3Al4uqB66KUC

[56] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[57] J. Van Bladel, *Electromagnetic Fields*, ser. IEEE Press Series on Electromagnetic Wave Theory. John Wiley & Sons, 2007. [Online]. Available: https://books.google.se/books?id=1tPotI3DJuEC

[58] C. A. Balanis, *Antenna Theory: Analysis and Design*. Wiley-Interscience, 2005.

[59] J. R. Carson, "A generalization of the reciprocal theorem," *The Bell System Technical Journal*, vol. 3, no. 3, pp. 393–399, jul 1924.

[60] T. Schenk, *RF Imperfections in High-rate Wireless Systems: Impact and Digital Compensation*. Springer, 2008.

[61] F. Kaltenberger, H. Jiang, M. Guillaud, and R. Knopp, "Relative channel reciprocity calibration in MIMO/TDD systems," in *ICT Mobile Summit 2010, 19th Future Network & Mobile Summit, June 16-18, 2010, Florence, Italy*, Florence, ITALY, 06 2010. [Online]. Available: http://www.eurecom.fr/publication/3082

[62] W. Zhang et al., "Large-Scale Antenna Systems With UL/DL Hardware Mismatch: Achievable Rates Analysis and Calibration," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1216–1229, April 2015.

[63] H. Wei, D. Wang, and X. You, "Reciprocity of mutual coupling for TDD massive MIMO systems," in *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, Nanjing, China, Oct 2015.

[64] J. Xiwen et al., "MIMO-TDD reciprocity under hardware imbalances: Experimental results," in *2015 IEEE International Conference on Communications ICC, 8-12 June 2015, London, United Kingdom*, London, U.K., 2015.

[65] R. Jedlicka, M. Poe, and K. Carver, "Measured mutual coupling between microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 29, no. 1, pp. 147–149, Jan 1981.

[66] H. Wei, D. Wang, H. Zhu, J. Wang, S. Sun, and X. You, "Mutual Coupling Calibration for Multiuser Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 606–619, Jan 2016.

[67] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[68] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, "GLRT-Optimal Noncoherent Lattice Decoding," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3773–3786, July 2007.

[69] L. P. Hansen, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, vol. 50, no. 4, pp. 1029–54, July 1982. [Online]. Available: http://ideas.repec.org/a/ecm/emetrp/v50y1982i4p1029-54.html

[70] A. Hall, *Generalized Method of Moments*, ser. Advanced Texts in Econometrics. OUP Oxford, 2004.

[71] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, Nov 2007.

[72] S. He and S. H. G. Chan, "Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 466–490, 2016.

[73] C. Steiner, *Location Fingerprinting for Ultra-wideband Systems: The Key to Efficient and Robust Localization*, ser. Series in Wireless Communications. Logos Verlag Berlin, 2010.

[74] A. Farshad, J. Li, M. K. Marina, and F. J. Garcia, "A microscopic look at WiFi fingerprinting for indoor mobile phone localization in diverse environments," in *International Conference on Indoor Positioning and Indoor Navigation*, Montbeliard-Belfort, France, Oct 2013, pp. 1–10.

[75] K. Witrisal, E. Leitinger, S. Hinteregger, and P. Meissner, "Bandwidth Scaling and Diversity Gain for Ranging and Positioning in Dense Multipath Channels," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 396–399, Aug 2016.

[76] F. Rosenblatt, *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*, ser. Report (Cornell Aeronautical Laboratory). Spartan Books, 1962.

[77] Yann Lecun and Yoshua Bengio and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015.

[78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges,

L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[80] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[81] S. Dorner, S. Cammerer, J. Hoydis, and S. ten Brink, "Deep learning-based communication over the air," *ArXiv preprint arXiv1707.03384*, 2017.

[82] A. G. Nariman Farsad, "Detection Algorithms for Communication Systems Using Deep Learning," *ArXiv preprint arXiv1705.08044*, 2017.

[83] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1561/2000000039

[84] Y. W. Michal Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology (in press)*, Sept 2017.

[85] N. T. R. Shwartz-Ziv, "Opening the Black Box of Deep Neural Networks via Information," *ArXiv preprint arXiv1703.00810*, 2017.

[86] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.

[87] J. J. Tompson, A. Jain, Y. Lecun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1799–1807.

[88] L. Liu et al., "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, December 2012.

[89] X. Luo, "Multi-User Massive MIMO Performance with Calibration Errors," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4521–4534, Jul 2016.

[90] D. Liu, W. Ma, S. Shao, Y. Shen, and Y. Tang, "Performance Analysis of TDD Reciprocity Calibration for Massive MU-MIMO Systems With ZF Beamforming," *IEEE Communications Letters*, vol. 20, no. 1, pp. 113–116, Jan 2016.

[91] A. Gut, *An Intermediate Course in Probability*, 2nd ed. Springer Publishing Company, Incorporated, 2009.

Appendix A

Define

$$D_k = \log_2\left(1 + \text{snr}||\boldsymbol{h}_k||^2\right) - \log_2\left(1 + \text{snrE}\left\{||\boldsymbol{h}_k||^2\right\}\right). \tag{7.1}$$

We want to show that

$$\sum_{k=1}^{K} D_k \to 0 \quad \text{as} \quad M \to \infty. \tag{7.2}$$

That is, (2.25) and (2.26) converge in probability when $M \to \infty$. Before doing so, define

$$\begin{aligned}\epsilon \triangleq &||\boldsymbol{h}_k||^2 - \text{E}\left\{||\boldsymbol{h}_k||^2\right\} \\ = &||\boldsymbol{h}_k||^2 - M.\end{aligned}$$

Note that the LLNs in (2.23) implies that

$$\epsilon/M \to 0 \quad \text{as} \quad M \to \infty. \tag{7.3}$$

Now, write (7.1) as

$$\begin{aligned}D_k =& \log_2\left(1 + \text{snr}\left(M + \epsilon\right)\right) - \log_2\left(1 + \text{snr}M\right) && (7.4)\\ =& \log_2\left(\frac{1 + \text{snr}\left(M + \epsilon\right)}{1 + \text{snr}M}\right) && (7.5)\\ =& \log_2\left(1 + \frac{\text{snr}\ \epsilon}{1 + \text{snr}M}\right). && (7.6)\end{aligned}$$

For a fixed snr, it follows from (7.3) that $(\text{snr}\ \epsilon)/(1+\text{snr}M) \to 0$ as $M \to \infty$. Since $\log_2\left(1 + x\right)$ is a continuous function at $x = 0$, a consequence of the Continuous mapping theorem–theorem 6.7. in [91]–is that

$$D_k \to 0 \quad \text{as} \quad M \to \infty. \tag{7.7}$$

Due to $D_k \to 0, \forall k$, it follows directly that $\sum_{k=1}^{K} D_k \to 0$ as $M \to \infty$ which concludes the proof.

Part II

Included Papers

Paper I

A flexible 100-antenna testbed for Massive MIMO

Massive multiple-input multiple-output (MIMO) is one of the main candidates to be included in the fifth generation (5G) cellular systems. For further system development it is desirable to have real-time testbeds showing possibilities and limitations of the technology. In this paper we describe the Lund University Massive MIMO testbed LuMaMi. It is a flexible testbed where the base station operates with up to 100 coherent radio-frequency transceiver chains based on software radio technology. Orthogonal Frequency Division Multiplex (OFDM) based signaling is used for each of the 10 simultaneous users served in the 20 MHz bandwidth. Real time MIMO precoding and decoding is distributed across 50 Xilinx Kintex-7 FPGAs with PCI-Express interconnects. The unique features of this system are: (i) high throughput processing of 384 Gbps of real time baseband data in both the transmit and receive directions, (ii) low-latency architecture with channel estimate to precoder turnaround of less than 500 micro seconds, and (iii) a flexible extension up to 128 antennas. We detail the design goals of the testbed, discuss the signaling and system architecture, and show initial measured results for a uplink Massive MIMO over-the-air transmission from four single-antenna UEs to 100 BS antennas.

Joao Vieira, Steffen Malkowsky, Karl Nieman, Zachary Miers, Nikhil Kundargi, Liang Liu, Ian Wong, Viktor Owall, Ove Edfors, and Fredrik Tufvesson,
"A flexible 100-antenna testbed for Massive MIMO,"
in *Proc. IEEE GLOBECOM Workshop Massive MIMO, Theory Pract., Dec. 2014, pp. 287-293*

1 Introduction

Massive MIMO is a promising technology and a strong candidate for future-generation wireless systems. Compared to conventional MIMO, potential benefits brought by the extra degrees-of-freedom due the excess number of BS antennas include [1] [2]: (i) both system capacity and radiated energy efficiency can be improved by several orders of magnitude; (ii) hardware requirements on the base station (BS) radio frequency (RF) chains can be greatly relaxed; (iii) simplification of the multiple-access layer; all of this with (iv) reduced complexity at the user equipment (UE). To take the next steps in the development and verification of the potential, it is necessary to have proof-of-concept platforms, i.e. testbeds, where Massive MIMO can operate under real-life conditions (e.g., with analog front-end impairments and real wave propagation conditions) to assist further algorithm development and circuit design. Testbeds can improve the overall understanding of, so far conceivable, issues and help maturing the technology for standardization.

Table 1: Existing Massive MIMO testbeds.

Institution	**Band (GHz)**	**Hardware**	**# of BS antennas**	**# of users**	**(< 1 ms) turnaround?**
Lund [3]	2.6	RUSK channel sounder [4]	128 (cylinder)	6	No
Rice [5]	2.4	WARP, powerPC	64 (planar)	15	No
Samsung [6]	1-28	Propr-ietary	64 (planar)	?	?

Table 1 lists existing many-antenna testbeds as of today. The first system is a channel sounding system used at Lund University to measure the wireless channel with a large number of antennas to validate theoretical gains [3]. 50 MHz channel measurements were taken over slow continuous user movements and then processed offline. The results confirm favorable propagation for measured channels with low eigenvalue spread. Second, Rice University [5] constructed a testbed and evaluated pratical performance gains of Massive MIMO in indoor environments. Channel measurements were collected over a 0.625 MHz bandwidth for both LOS and NLOS conditions, and promising capacity results based on SINR computations were presented. Third, researchers at Samsung [6] recently made their work in many-antenna MIMO systems public. This testbed is targeted at millimeter wave bands but can be applied to cellular band applications. The press release is not very detailed, though it is mentioned that a throughput of 1 Gbps is achieved at 2 km range.

Despite prior work in large scale MIMO systems, many shortcomings are evident. Existing testbeds are either proprietary, non-real-time, or both. These limitations hinder researchers from developing algorithms tied to real wireless channels. To address this, we have developed an extensible platform, the LuMaMi testbed, to realize up to 20 MHz bandwidth 100-antenna MIMO. It is built up of commercial off-the-shelf hardware, making it accessible and modifiable. The main objectives for this

testbed are:

- implementing BS architectures to meet high-throughput/low-latency processing requirements;
- evaluating practical performance of different baseband processing algorithms;
- implementing time and frequency synchronization solutions between BS RF chains;
- identifying scenarios where favorable propagation conditions for Masssive MIMO exist (or do not exist);
- demonstrating a Massive MIMO proof-of-concept by concurrent high-speed data streaming to and from multiple users, via high-density spatial multiplexing within the same time-frequency resource. The link quality can be accessed either by: (i) evaluating performance metrics by streaming pseudo-noise (PN) sequences, such as bit-error-rate (BER), error vector magnitude (EVM), etc; (ii) visualizing streamed high-definition (HD) videos;

The remainder of this paper is structured as follows: Sec. 2 details the system architecture and hardware components implementing the BS; Sec. 3 addresses different aspects of the communication protocol; Sec. 4 presents the initial testbed results in terms of RF-chain synchronization and illustrations of received signal constellations under maximum-ratio combining (MRC) and zero-forcing (ZF) uplink spatial multiplex; and Sec. 5 presents conclusions drawn from the work.

2 Testbed design

2.1 Problem formulation

In a massive MIMO context, a potential BS architecture designed to yield low processing latency, transport latency and high transport reliability would

- use an all-mighty central controller (CC) aggregating and processing data from/to all (100) antennas;
- be architected in a star-like fasion yielding hundreds of input/output ports;
- shuffle large amounts of baseband data between the CC and RF front ends through high bandwidth/low latency interconnects;
- operate with hundreds of perfectly synchronized RF chains with low RF impairments;

While the second point imposes a tight hardware constraint, potentially preventing flexibility and scalability of the system, the first is the toughest to meet with today's off-the-shelf solutions since 100 antennas of baseband data far exceeds the input/output (IO) capabilities of most practical hardware. Flexible implementations of massive MIMO BSs with real-time processing requirements are thus non-trivial.

2.2 Hierarchical overview

Fig. 3 shows the hierarchical overview of our system, whose main blocks are detailed as follows:

Central controller (CC)

A master chassis embeds a x64 controller (NI PXIe-8135) which runs LabVIEW on a Windows 7 64-bit OS and serves three primary functions: (*i*) it provides a user interface for radio configuration, deployment of FPGA bitfiles, system control, and visualization of the system, (*ii*) it acts as source and sink for the user data—e.g. HD video streams—sent across the links, and (*iii*) the CC measures link quality with metrics such as BER, EVM, and packet-error rate (PER). It connects to three switches through cabled Gen 2 x8 PCI Express (MXIe) in a star fashion.

Switches

The switches consist of three (NI 1085 PXIe) 18-slot chassis. The first slot is reserved for the modules (NI PXIe-8381) that connect to the master chassis, and the remaining slots hold MXIe interface cards (NI PXIe-8374) to link with the SDRs. The MXIe interface between the Gen 2 x8 PCIe backplane and the SDRs is Gen 1 x4. Switches yield no processing but allow data to be transferred between SDRs using peer-to-peer direct memory access (DMA) streaming and between SDRs and the CC using target-to-host and host-to-target DMA transfers.

Software defined radios

The SDRs (NI 2943R/USRP-RIO) each contain a reconfigurable (Xilinx Kintex-7) FPGA and two full-duplex 40 MHz RF bandwidth transceivers that can be configured for center frequencies 1.2-6 GHz, and can transmit with up to 15 dBm. Baseband processing is partitioned and distributed across the fifty FGPAs, as detailed in Sec. 2.4, and the RF transceivers connect to the antenna array.

Please check [7] for further hardware specifications.

2.3 Streaming IO rates

For proper baseband processing partition, the limitations of the hardware components implementing the system in Fig. 3 are:

- Each Gen 2 x8 PCI Express interface linking the three chassis handles up to 3.2 GBps bidirectional traffic.
- Two Gen 2 x8 switches link the interface cards through the backplane of the chassis. Their streaming rate is bounded to 3.2 GBps of bidirectional traffic in each slot with an aggregate total of 32 GBps inter-switch traffic.
- Each SDR has 13 available DMA channels (three are used for the radio configuration) that share the total IO rate for Gen 1 x4 PCIe of 800 MBps bidirectional.

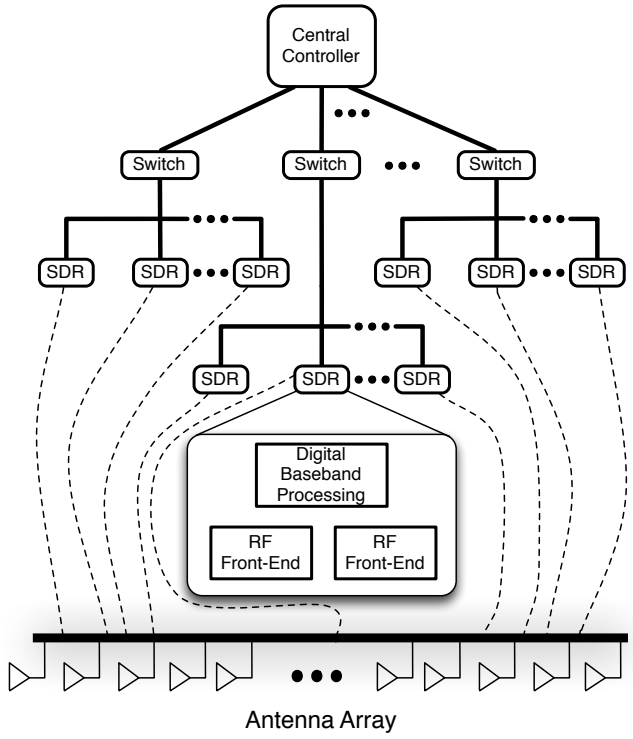

Figure 1: Hierarchical overview of the base station.

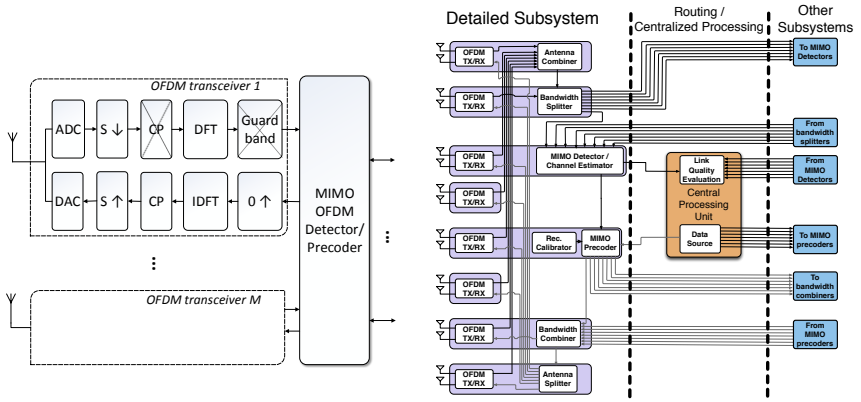

Figure 2: Left: Main blocks of a typical MIMO OFDM transceiver. Right: BS subsystem for partitioned baseband processing.

2.4 Sub-system partitioning

Below we detail how the baseband processing is partitioned across FPGAs. The functional representation of an OFDM Massive MIMO system is shown at the left part of Fig. 2. To map this to hardware, the occupied bandwidth is divided into eight OFDM sub-bands which are processed independently to relax the IO requirements of a single FPGA. One subsystem of eight FPGAs, shown in the right part of Fig. 2, operate per sub-band. Additional folding of MIMO detectors and precoders (since we do not have eight subsystems) is performed and end nodes are inserted to achieve a full 100-antenna platform.

For each RX chain, the received RF signals are digitized, followed by analog front-end calibration and time/frequency synchronization. From the synchronized data, the cyclic prefix (CP) is removed, followed by FFT OFDM demodulation and guard-band removal. Note that the OFDM symbols contain the superposition of the transmitted signals by all users. In each sub-system, consisting of 16 receive antennas, the yet unequalized OFDM symbols are streamed into an FPGA with an "Antenna Combiner" function. This combines all the uplink streams from the 16 antennas and passes the result to another FPGA in the sub-system with a "Bandwidth splitter" function, which splits the signals into eight bandwidth chunks. In each sub-system, we have one FPGA with a "MIMO Detector" function collecting data of a given bandwidth chunk from the other seven sub-systems. Using the channel matrix estimated from uplink pilots, the "MIMO Detector" cancels interference and detects the frequency-domain symbols from each user equipment. The detected symbols are then sent to the CC for further processing, such as link quality evaluation.

At the downlink, the channel estimates and reciprocity calibration estimated weights are passed to the "MIMO Precoder", and reciprocal processing is performed,

e.g., modulation instead of demodulation.

It can be noted that each subcarrier data sample is quantized with 12 bits for each in-phase and quadrature component. This allows meeting the SDRs IO rate limitations listed in 2.3.

2.5 Latency analysis

To support fast precoder turnaround time, the system has been architected to provide low latency in the signal path from channel estimation to MIMO precoding, shown in Figure 2. The turnaround time must meet the frame structure shown in Figure 6. This structure leaves 214 μs for total latency

$$\Delta = \Delta_f^{rx} + \Delta_o^{rx} + \Delta_e + \Delta_p + \Delta_o^{tx} + \Delta_f^{tx} + N_h \Delta_h + \phi \quad (1)$$

in the critical path, including RX front-end delay Δ_f^{rx}, OFDM RX (CP removal, FFT, guard subcarrier removal) Δ_o^{rx}, channel estimate calculation Δ_e, precoder calculation Δ_p, OFDM TX (guard subcarrier interleave, IFFT, CP addition) Δ_o^{tx}, and TX front-end delay Δ_f^{tx}. Additional sources of latency include overheads in data routing, packing, and unpacking ϕ as well as latency for each hop across the PCIe backplane $N_h \Delta_h$. The worst-case latency of each hop is $\Delta_h = 5$ μs for the seven-hop path ($N_h = 7$), resulting in a worst-case total PCIe latency of $N_h \Delta_h = 35$ μs in the critical signal path. $\Delta_f^{rx} + \Delta_f^{tx}$ was measured to be $\approx 2.25\mu s$, $\phi \approx 0.1\mu s$, $\Delta_o^{rx} \approx \Delta_o^{tx} \approx 27.5\mu s$. Δ_p depedends on the type of precoder and respective implementation type. The MRT precoder can be processed point-by-point, allowing for a high degree of pipelining. Similarly, channel estimation can be performed point-by-point. The highest latency configuration will be that for the ZF precoder due to the matrix inverse, matrix-matrix multiplications, and its serial-to-parallel conversions.

2.6 Synchronization

A massive MIMO basestation requires time synchronization and phase coherency between the RF chains. This is achieved using a reference clock and timing/trigger distribution network. This synchronization network consists of eight OctoClock modules in a tree structure with a master OctoClock feeding seven secondary OctoClocks. Low skew buffering circuits and matched-length transmission cables ensure that there is low skew between the reference clock input at each SDR. The source clock for the system is an oven-controlled crystal oscillator within an NI PXIe-6674T timing module. Triggering is achieved by generating a start pulse within the *Master* SDR via a software trigger. This trigger is then fed from an output port on the master to the NI PXIe-6674T timing module, which conditions and amplifies the trigger. The trigger is propagated to the master OctoClock and distributed down the tree to each SDR in the system (including the master itself). This signal sets the reference clock edge to use for start of acquisition for the transmitter (TX) and receiver (RX) within each channel. Initial results show that reference clock skew is within 100 ps and trigger skew is within 1.5 ns, which is well below the sampling period of 33 ns.

2.7 Antenna Array

The three different stages of the array building process are described below.

Material and characterization

We choose Diclad 880 with thickness of 3.2 mm as the printed circuit board substrate. The dielectric constant and dissipation factor were confirmed using a trapped waveguide characterization method [8]. To verify the substrate characterization, a six element patch array with slightly different element sizes was built, measured, and compared with the simulated data. To fit the final results, a final re-characterization of the substrate was performed, and the simulated and measured bandwidth matched within 1 MHz.

Design

A planar "T"-shaped antenna array was built with 160 dual polarized $\lambda/2$ shorted patch elements. The "T" upper horizontal rectangle has 4×25 elements and the central square 10×10 elements, (see Fig. 7). This yields 320 possible antenna ports that can be used to explore different antenna array arrangements. All antenna elements are center shorted which improves isolation, bandwidth, and reduces risk of static shock traveling into the active components if the elements encounter a static electric discharge. The feed placement shifts by 0.52 mm from the center of the array elements to the outer edge elements in order to maintain match with changing array effects which impact individual elements differently. The size of the element changes by 0.28 mm from the center of the array to the outer elements this maintains constant center frequency of 3.7 GHz throughout the entire array.

Measurements

The final 160 element array was simulated at 3.7 GHz. Results showed an average match of -51 dB, and an average 10dB-bandwidth of 185 MHz. Similar tests were done to the manufactured array which yielded an average 10dB-bandwidth of 183 MHz centered at 3.696 GHz and the average antenna match was found to be -28 dB.

2.8 Mechanical structure and electrical characteristics of BS

Two rack mounts assemble all BS components with combined measures of $0.8 \times 1.2 \times 1$ m shown in Fig. 7. They were attached on top of a four-wheel trolley not to compromise its mobility when testing different scenarios. Approximate combined weight and average power consumption are 300 kg and 2.5 kW, respectively.

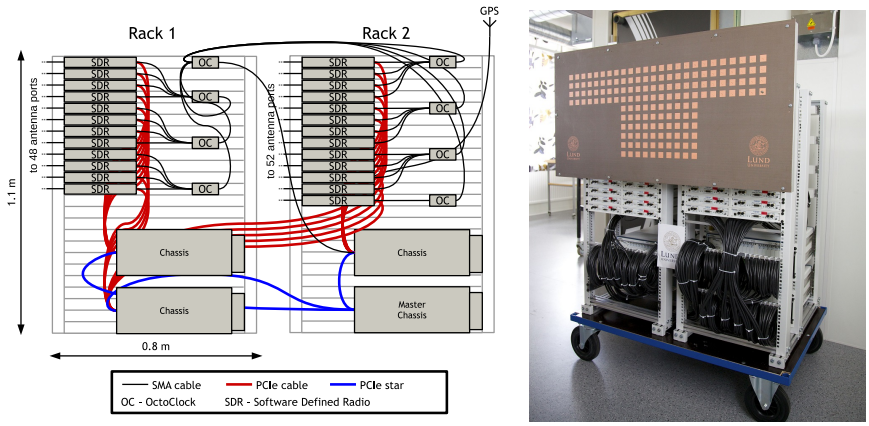

Figure 3: Left: Side view of the mechanical assembly the BS. The two racks sit side-by-side (not as shown) with the SDRs facing the same direction (towards the antenna array). Two columns of URSPs are mounted in each rack, totaling 50 of them. Right: Picture of the assembled BS, with mounted antenna array.

2.9 User Equipment

Five SDRs (NI 2953Rs/USRP-RIOs) are used at the terminal ends to emulate the UEs. They yield similar properties as the ones at the BS with the additional feature of their internal clocks can be locked to a GPS reference signal. This provides a reliable timing reference for sampling purposes, and a frequency offset of less than 1 ppb.

3 System specifications

3.1 General parameters

In the current setting, the testbed operates with many parameters similar to LTE-like cellular systems, as shown in Table 1.

3.2 Supported precoders

The heavy real-time processing requirements for massive MIMO have, in general, been restricting the attention mostly to linear precoders/equalizers. For a proof-of-concept of massive MIMO, we focus on the implementation of two standard linear precoders:

Table 2: High-level system parameters

Parameter	Variable	Value
Bandwidth	W	20 MHz
Carrier frequency	f_c	3.7 GHz
Sampling Rate	F_s	30.72 MS/s
FFT Size	N_{FFT}	2048
# Used subcarriers	N_{used}	1200
Slot time	T_S	0.5 ms
Sub-Frame time	T_{sf}	1 ms
Frame time	T_f	10 ms
# UEs	K	10
# BS antennas	M	100

Maximum Ratio Transmission (MRT)

The MRT precoder maximizes the signal-to-noise ratio (SNR) at the terminal side and precoding-weights simply consist of the complex conjugate of the estimated channels[1]. Thus, precoding is both of low complexity and can, in principle, be performed independently close to each antenna, i.e., in a non-centralized fashion.

Power scaling of precoding weights is still needed if an average transmit power level is to be met. This requires a centralized control structure with relatively low signaling overhead.

Zero Forcing (ZF)

The ZF precoder forces interfence among users to zero and precoding weights are obtained from inverting the inner Gram matrix of the full channel matrix, which contains all estimated channels. This implies a more complex precoder calculation and leads to a centralized architecture where all processing *typically* happens at a central controller.

3.3 Frame structure

The transmission of massive MIMO data is divided into 10 ms radio frames as shown in Fig. 6. The frame consists of 10 subframes, each containing two 0.5 ms slots. The radio frame starts with a special down-link broadcasting subframe

[1] The MRT precoding process is also known as conjugate beamforming.

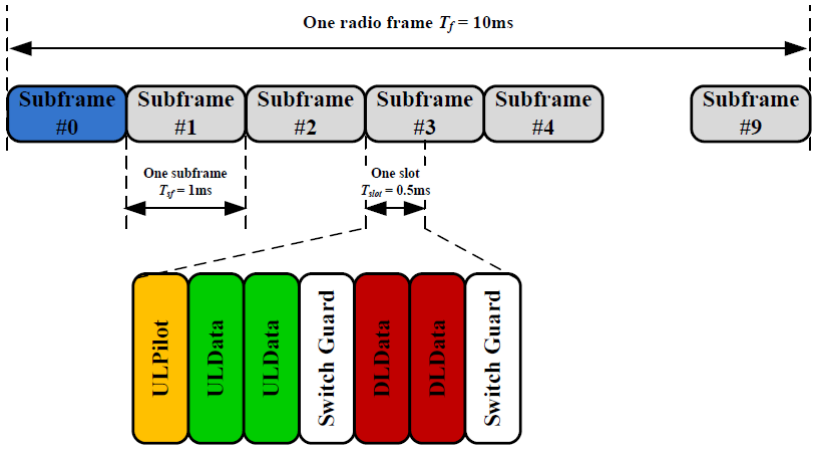

Figure 4: Frame structure.

(may consist of PN sequences) to setup the initial synchronization of the network, e.g., UEs can synchronize their frequencies (both carrier frequency and sampling frequency) and align the time offset due to their variable distance to the BS. The remaining 9 subframes are used for UL and DL data transmission.

As also demonstrated in Fig. 6, one slot consist of 7 OFDM symbols, where the 1st is used entirely for UL pilots, followed by 2 UL data symbols, a guard period for UL→DL switching, and 2 DL data symbols, followed by a guard period for DL→UL switching.

3.4 Pilot allocation

The frequency domain uplink pilots are sequentially interleaved to each of the 10 users in the system, as shown in Fig. 5, where $P_{i,j}$ is the pilot for user i and subcarrier chunk j, where each subcarrier chunk consist of 10 subcarriers. For a particular user, non-trained subcarrier channels can be estimated through an interpolation/extrapolation scheme using the trained ones. At the downlink, since users are spatially multiplexed, pre-coded pilots are inserted every 10^{th} subcarrier in the first DL OFDM symbol to allow compensation for the RF chain responses of the terminals.

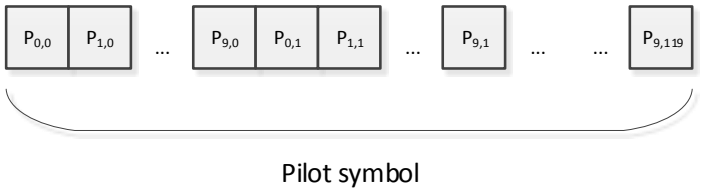

Figure 5: Frequency-domain pilot-symbol allocation.

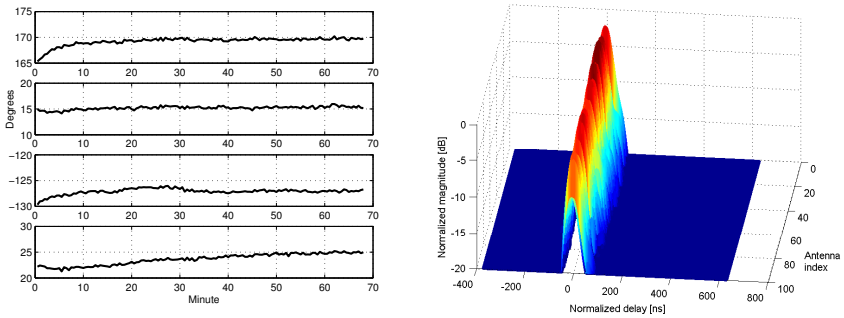

Figure 6: Left: RX RF-chain phase evolution of four different SDRs. Right:Impulse response of 100 simultaneously measured channels.

3.5 Throughput

The total amount of aggregated baseband traffic that can be handled by the testbed both in uplink and downlink directions is given by

$$R_{Bs} = \#\text{BS antennas} \times 2\ I/Q_{bits} \times ADC_{SR} = 384\ \text{Gbps} \tag{2}$$

where $I/Q_{bits} = 16$ is the maximum number of quantization bits per I/Q sample and $ADC_{SR} = 120$ MS/s is the ADC sampling rate.

An example of the data rate per user per direction is given by

$$R_{ue,ul/dl} = \frac{N_{used} \times N_{ul/dl}}{0.5\ ms} \times \frac{N_{sf} - N_{bf}}{N_{sf}} \times R_c \times N_{mod}, \tag{3}$$

where $N_{ul/dl}$ is the number of UL or DL OFDM symbols within one slot, N_{bf} is the number of broadcasting subframes within one radio frame, R_c is the coding rate, and N_{mod} is the number of bit per modulated symbol. In case of 16-QAM modulation with rate 2/3 channel coding, the system provides 11.52 Mbps data rate per user per direction, which can be enhanced to 17.28 Mbps if 64-QAM is used.

4 Initial results

In this section, the synchronization capabilities of the BS RF front ends are verified, and as a proof-of-concept, we realize an indoor uplink massive MIMO transmission with 100 BS antennas and four single antenna users and show equalized signal constellation points.

4.1 Phase coherence

We measured the phase drift of different RX RF-chains. A tone transmitted by one SDR is split into four signals, and input to four SDRs spanning four different OctoClocks and two switches. SMA cables and RF splitters were used as the channel for this experiment. Since all four channels share the same TX RF chain, and the cables/RF splitter have static responses, the phase drift is solely due to the RF chains of the receivers. Fig. 6 shows the phases of the measured signal phases which remain within 5 degrees across 1 hour of measurements. The largest change in phase is observed within the first 10 minutes, as the devices are coming up to temperature. After that warm-up period, phases are stable to within a few degrees over a one-hour period. The results suggest that reciprocity calibration can be performed on an hourly basis, without severe performance degradation [9].

4.2 Time Synchronization

An 800-sample 30.72 MHz Gaussian PN sequence is repeatedly transmitted by a single antenna. The transmitter is positioned about a meter in front of a 4×25 antenna array arrangement. All 100 receiving antennas are roughly at the same distance from the transmitter and their respective RF-chains share the same reference clock signals. This setup yields a strong LOS channel that can be used to verify the sampling synchronization capabilities of the RF chains. For each channel, the impulse response (IR) is obtained by performing a circular cross-correlation of the received signals with the original PN sequence. Fig. 6 shows that the measured channels yield a distinctive planar wavefront with a small delay spread. These results indicate that the received samples are well time aligned within one 30.72 MS/s sample, i.e. within 33 ns.

4.3 Uplink Massive MIMO transmission test

As proof-of-concept, we performed an uplink Massive MIMO transmission from four single-antenna UEs to 100 BS-antennas in our lab. Each UE is equipped with SkyCross UWB antennas (SMT-3TO10M-A) and radiate 0 dBm of power.

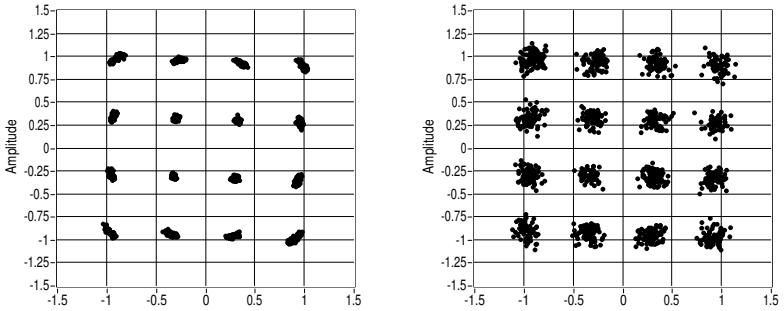

Figure 7: Equalized signal constellation points for one user, for the case when four users are spaced two meters from each other under LOS conditions. Left: ZF decoder; Right: MRC decoder.

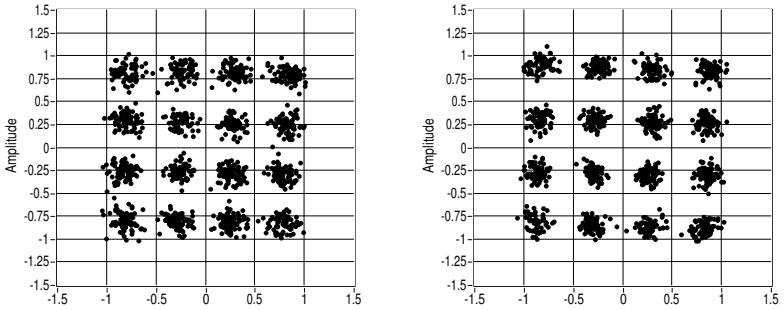

Figure 8: ZF equalized signal constellation points for two out of four closely spaced users (all four within a 15 cm-radius sphere) under NLOS conditions.

The uplink transmission mode was chosen since it can be realized without performing reciprocity calibration and implementation of the uplink/downlink frame structure, i.e., the base station simply equalizes the data symbols using their respective channel estimates. This also allows all baseband processing to be implemented solely at the CC if no real-time constraints are to be met. We took this provisional approach to be able to showcase a massive MIMO transmission. All baseband processing will, however, be moved to the FPGAs in subsequent work to meet the testbed description given in Sec. 2.2. Note that the parameters specified in Table 1 are still valid for this experiment, but slots are transmitted at a rate which can be handled by the CC. We used the same slot structure as Fig. 6 but no downlink data symbols were transmitted during this test.

Fig. 7 and Fig. 8 show the equalized signal constellation points of an received OFDM data symbol under different channel conditions, users separations and MIMO decoders. For a given user, we used zero-order hold to interpolate between trained subchannels. This explains the small rotation that can be observed for the measured signal constellation points.

Overall, ZF outperformed MRC in all experiments, and showed to be possible to separate both: (i) closely spaced users, and (ii) users at different distances to the BS, if enough power is transmitted. For the MRC case, its interference limited performance constrains user scenarios yielding acceptable performance to those where users are being spaced rather far apart with some sort of power control.

5 Conclusions and Future work

In this paper we detail our solution for realizing massive MIMO in a practical testbed. The testbed is operating with a 20MHz bandwidth and 100 antennas at the BS, entirely made of off-the-shelf hardware. To tackle the main hardware bottlenecks, we propose a hierarchical hardware architecture, baseband processing partitioning and a communication protocol that allows the processing to meet real-time requirements. To unveil key performance trade-offs for different system settings, it is of particular interest to be able to operate with flexible communication parameters and antenna array configurations. Synchronization tests between the BS RF chains show small and slow relative phase drifts and tight time alignment of received samples. As proof-of-concept, an over-the-air uplink massive MIMO transmission with spatial multiplexing of four users was performed with all baseband processing being conducted at the CC.

In future work, we intend to move the baseband processing to the SDRs FP-

GAs, such that both uplink and downlink transmissions can be realized under full real-time requirements. In addition to the distributed processing architecture presented in Fig. 2, we also intent to investigate alternative architectures based on a more centralized processing scheme.

Acknowledgments

This project has partially been funded by grants from the Swedish Foundation for Strategic Research, the Swedish Research Council and the Strategic Research Area ELLIIT. The authors would also like to thank Xilinx for donating LabVIEW compatible IP blocks.

References

[1] H. Q. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *Communications, IEEE Transactions on*, vol. 61, no. 4, pp. 1436–1449, April 2013.

[2] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, February 2014.

[3] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *Signal Processing Magazine, IEEE*, 2013.

[4] R. Thoma, D. Hampicke, A. Richter, G. Sommerkorn, A. Schneider, and U. Trautwein, "Identification of time-variant directional mobile radio channels," in *Instrumentation and Measurement Technology Conference, 1999. IMTC/99. Proceedings of the 16th IEEE*, vol. 1, 1999, pp. 176–181 vol.1.

[5] C. Shepard, H. Yu, N. Anand, E. Li, T. L. Marzetta, R. Yang, and Z. L., "Argos: Practical many-antenna base stations," *Proc. ACM Int. Conf. Mobile Computing and Networking (MobiCom)*, 2012.

[6] Samsung. (2013) Samsung takes first 5G steps with advanced antenna. [Online]. Available: {http://www.pcworld.idg.com.au/article/461656/}

[7] E. Luther. (2014) 5G massive MIMO testbed: From theory to reality. [Online]. Available: {http://www.ni.com/white-paper/52382/en/}

[8] A. Namba, O. Wada, Y. Toyota, Y. Fukumoto, Z. L. Wang, R. Koga, T. Miyashita, and T. Watanabe, "A simple method for measuring the relative permittivity of printed circuit board materials," *Electromagnetic Compatibility, IEEE Trans on*, vol. 43, no. 4, pp. 515–519, Nov 2001.

[9] J. Vieira, F. Rusek, and F. Tufvesson, "Reciprocity calibration methods for massive MIMO based on antenna coupling," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, Dec 2014.

Paper II

The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation

This paper sets up a framework for designing a massive multiple-input multiple-output (MIMO) testbed by investigating hardware (HW) and system-level requirements such as processing complexity, duplexing mode and frame structure. Taking these into account, a generic system and processing partitioning is proposed which allows flexible scaling and processing distribution onto a multitude of physically separated devices. Based on the given HW constraints such as maximum number of links and maximum throughput for peer-to-peer interconnections combined with processing capabilities, the framework allows to evaluate modular HW components. To verify our design approach, we present the LuMaMi (Lund University Massive MIMO) testbed which constitutes the first reconfigurable real-time HW platform for prototyping massive MIMO. Utilizing up to 100 BS antennas and more than 50 FPGA, up to 12 UE are served on the same time/frequency resource using an LTE-like orthogonal frequency division multiplexing (OFDM) time division duplex (TDD)-based transmission scheme. Proof-of-concept tests with this system show that massive MIMO can simultaneously serve a multitude of users in a static indoor and static outdoor environment utilizing the same time/frequency resource.

1 Introduction

In massive MIMO (MaMi) an unconventionally high number of base station (BS) antennas (hundreds or even higher) is employed to serve e.g., a factor of ten less user equipments (UEs). Due to the excess number of BS antennas, linear signal processing may be used to spatially focus energy with high precision, allowing to separate a multitude of UEs in the spatial domain while using the same time/frequency resource [1]. MaMi theory promises a variety of gains, e.g., increase in spectral and energy efficiencies as compared with single antenna and traditional MU-MIMO systems [2, 3], thereby tackling the key challenges defined for 5G.

Although MaMi is a promising theoretical concept, further development requires prototype systems for proof-of-concept and performance evaluation under real-world conditions to identify any further challenges in practice. Because of its importance, both industry and academia are making efforts in building MaMi testbeds, including the Argos testbed with 96-antennas [4], Eurecom's 64-antenna LTE compatible testbed, Samsung's Full-Dimension (FD) MIMO testbed and Facebook's Project Aries. Nevertheless, publications systematically describing the design considerations and methodology of a MaMi testbed are missing and real-time real-scenario performance evaluation of MaMi systems using testbeds have not been reported yet. At Lund University, the first real-time MaMi testbed, the LUMAMI testbed, showing successful MaMi transmission on the uplink UL, was built [5]. Ever since, many testbeds have been constructed based on identical HW utilizing the same generic design principle, e.g., the MaMi testbeds at the University of Bristol [6], Norwegian University of Science and Technology in Trondheim and University of Leuven in Belgium. The LUMAMI testbed provides a fully reconfigurable platform for testing MaMi under real-life conditions. To build a real-time MaMi testbed many challenges have to be coped with. For example, shuffling data from 100 or more antennas, processing large-scale matrices and synchronizing a huge number of physically separated devices. All this has to be managed while still ensuring an overall reconfigurability of the system allowing experimental hardware and software solutions to be tested rapidly.

This paper discusses how implementation challenges are addressed by first evaluating high-level HW and system requirements, and then setting up a generic framework to distribute the data shuffling and processing complexity in a MaMi system based on the given HW constraints for interconnection network and processing capabilities. Taking into account the framework and requirements, a suitable modular HW platform is selected and evaluated. Thereafter, a thorough description of the LUMAMI testbed is provided including system parameters, base-band processing features, synchronization scheme and other details. The LUMAMI testbed constitutes a flexible platform that supports prototyping of up to 100-antenna 20MHz bandwidth MaMi, simultaneously serving 12 UEs in real-time using OFDM modulation in TDD transmission mode. Bit-error-rate (BER) and constellations for real-time UL and downlink (DL) uncoded transmission in a static indoor and static outdoor scenario are presented. Our first real-life proof-of-concept measurement campaigns show, that MaMi is capable of serving up to 12 UEs in the same time/frequency resource even

for high user density per unit area. The gathered results suggest a significant increase in spectral efficiency compared to traditional point-to-point MIMO systems. By building the LUMAMI testbed we now have a tool which supports accelerated design of algorithms [7] and their validation based on real measurement data, with the additional benefit of real-world verification of digital base-band solutions.

Our main contributions can be summarized as follows:

- We provide overall and thorough analysis for MaMi systems, especially from a signal processing perspective, and identify design requirements as well as considerations on building up a MaMi testbed.
- We propose signal processing breakdown and distribution strategy to master the tremendous computational complexity in a MaMi system and introduce general hardware architecture for a MaMi testbed.
- We present the world's first real-time 100-antenna MaMi testbed, built upon software defined radio (SDR) technology.
- We validate the MaMi concept and its spatial multiplexing capability in real-life scenarios (both indoor and outdoor) with over-the-air transmission and real-time processing.

The paper is organized as follows. Sec. 2 shortly introduces MaMi basic theory, detection and precoding methods before Sec. 3 details specifications and requirements for a MaMi testbed. Sec. 4 describes a generic hardware partitioning for a modular scalable testbed to overcome implementation challenges. Sec. 5 presents the selected HW platform and provides details about the LUMAMI testbed. Results from different field trials based on real-time measured BER and constellations from different scenarios are presented in Sec. 6. Finally, Sec. 7 concludes the paper. (*i*) scaling data rates, interfaces to 128 antennas, (*ii*) providing low-latency channel state acquisition, and (*iii*) synchronizing time and frequency over 128 antennas, all using commercial modular components. The use of such modular components enables the sharing of code and IP and encourages other researchers to reproduce and validate the experimental results.

2 Massive MIMO Basics

In this section, the basic key detection and precoding algorithms utilized in MaMi are presented. Implementation specific details required to apply these algorithms, such as channel state information (CSI) estimation, are discussed in Sec. 5. A simplified model of a MaMi BS using M antennas while simultaneously serving K single antenna UEs in TDD operation in a propagation channel $\boldsymbol{B}$ is shown in Figure 1. To simplify notation, this discussion assumes a base-band equivalent channel and expressions are given per subcarrier, with subcarrier indexing suppressed throughout.

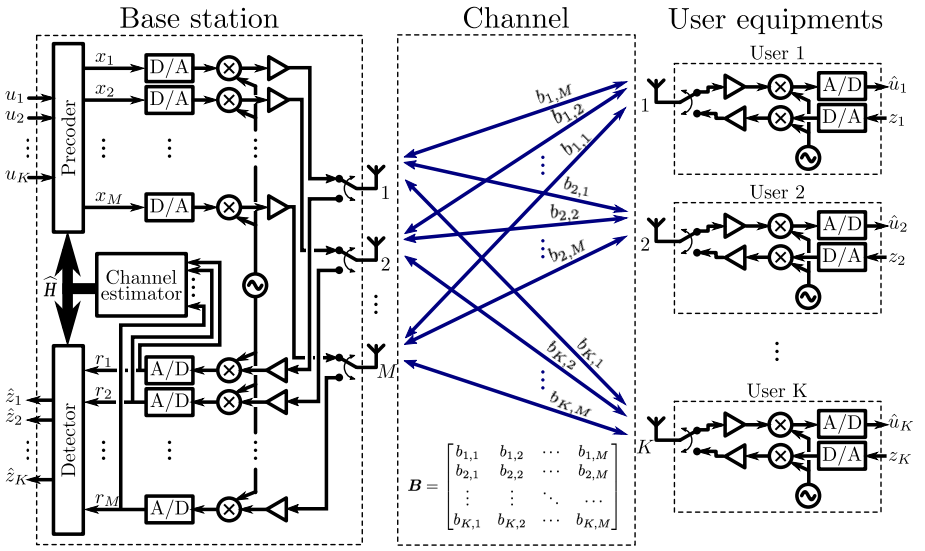

Figure 1: A MaMi system model. Each antenna at the BS (left side) transmits a linear combination of K user-intended data symbols $u_k{}_{k=1}^{K}$. After propagation through the DL wireless channel $\boldsymbol{B}$, each user antenna receives a linear combination of the signals transmitted by the M BS antennas. Finally, each of the K users, say user k, produces an estimate of its own intended data symbol, i.e., u_k. Similar operation is employed for UL data transmission. Here, reciprocity for the propagation channel is assumed, i.e., $\boldsymbol{B} = \boldsymbol{B}^{\mathsf{T}}$.

2.1 Up-link

UEs in MaMi are non-cooperative, thus, the only adjustable parameter for transmission is their power levels. The UL power levels used by the K UEs during transmission build the $K \times K$ diagonal matrix $\boldsymbol{P}_{\text{ul}}$. By collecting the transmitted UE symbols in a vector $\boldsymbol{z} \triangleq (z_1, \ldots, z_K)^{\mathsf{T}}$, the received signals $\boldsymbol{r} \triangleq (r_1, \ldots, r_M)^{\mathsf{T}}$ at the BS are described as

$$\boldsymbol{r} = \boldsymbol{G}\sqrt{\boldsymbol{P}_{\text{ul}}}\boldsymbol{z} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{G}$ is the $M \times K$ UL channel matrix[2], $\sqrt{\boldsymbol{P}_{\text{ul}}}$ an elementwise square-root, and $\boldsymbol{w} \sim \mathcal{CN}(0, \mathbf{I}_M)$ is independent and identically distributed (IID) circularly-symmetric zero-mean complex Gaussian noise. The estimated user symbols $\widehat{\boldsymbol{z}} \triangleq (\hat{z_1}, \ldots, \hat{z_K})^{\mathsf{T}}$

[2] $\boldsymbol{G}$ is the up-link radio channel capturing both, the propagation channel $\boldsymbol{B}^{\mathsf{T}}$ and the up-link hardware transfer functions.

Table 1: Linear Precoding/Detection Matrices

	MRT/MRC	**ZF**	**RZF**
DL	$\boldsymbol{C}\boldsymbol{G}^{*}$	$\boldsymbol{C}\boldsymbol{G}^{*}(\boldsymbol{G}^{\mathsf{H}}\boldsymbol{G})^{-\mathsf{T}}$	$\boldsymbol{C}\boldsymbol{G}^{*}(\boldsymbol{G}^{\mathsf{H}}\boldsymbol{G}+\beta_{\mathrm{reg}_{\mathrm{pre}}}\boldsymbol{I}_K)^{-\mathsf{T}}$
UL	$\boldsymbol{G}^{\mathsf{H}}$	$(\boldsymbol{G}^{\mathsf{H}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathsf{H}}$	$(\boldsymbol{G}^{\mathsf{H}}\boldsymbol{G}+\beta_{\mathrm{reg}_{\mathrm{dec}}}\mathbf{I}_K)^{-1}\boldsymbol{G}^{\mathsf{H}}$

from the K UEs are obtained by linear filtering of the received vector $\boldsymbol{r}$ as

$$\widehat{\boldsymbol{z}} = f_{\mathrm{eq}}(\boldsymbol{G})\boldsymbol{r}, \tag{2}$$

where $f_{\mathrm{eq}}(\cdot)$ constructs an appropriate equalization matrix.

2.2 Down-link

On the DL, each UE receives its corresponding symbol $\hat{u}_k$ which are collected in a vector $\widehat{\boldsymbol{u}} \triangleq (\hat{u}_1, \ldots, \hat{u}_K)^{\mathsf{T}}$, representing the symbols received by all UEs. With this notation, the received signal becomes

$$\widehat{\boldsymbol{u}} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{w}' \tag{3}$$

where the $K \times M$ matrix $\boldsymbol{H}$ is the DL radio channel[3], $\boldsymbol{w}' \sim \mathcal{CN}(0, \mathbf{I}_K)$ is an IID circularly-symmetric zero-mean complex Gaussian receive noise vector with covariance matrix $\mathbf{I}_K$, and $\boldsymbol{x} \triangleq (x_1, \ldots, x_M)^{\mathsf{T}}$ is the transmit vector.

As explicit DL channel estimation is very resource consuming, it is not considered practical in a MaMi setup [1]. Taking into account that the propagation channel $\boldsymbol{B}$ is generally agreed on to be reciprocal [7], the estimated UL channel matrix $\boldsymbol{G}$ can be utilized to transmit on the DL. However, differences due to analog circuitry in the UL and DL channels, $\boldsymbol{G}$ and $\boldsymbol{H}$, need to be compensated. Thus, a possible construction for $\boldsymbol{x}$ is of the form

$$\boldsymbol{x} = f_{\mathrm{cal}}(f_{\mathrm{pre}}(\boldsymbol{G}))\boldsymbol{u}, \tag{4}$$

where $\boldsymbol{u} \triangleq (u_1, \ldots, u_K)^{\mathsf{T}}$ is a vector containing the symbols intended for the K UEs, $f_{\mathrm{pre}}(\cdot)$ is some precoding function, and $f_{\mathrm{cal}}(\cdot)$ is a reciprocity calibration function to be discussed next.

2.3 Reciprocity Calibration

In most practical systems, the UL and DL channels are not reciprocal, i.e. $\boldsymbol{G} \neq \boldsymbol{H}^{T}$. This is easily seen by factorizing $\boldsymbol{G}$ and $\boldsymbol{H}$ as

$$\boldsymbol{G} = \boldsymbol{R}_{\mathrm{B}}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{T}_{\mathrm{U}}, \quad \text{and} \quad \boldsymbol{H} = \boldsymbol{R}_{\mathrm{U}}\boldsymbol{B}\boldsymbol{T}_{\mathrm{B}}, \tag{5}$$

[3] $\boldsymbol{H}$ is the down-link radio channel capturing both, the propagation channel $\boldsymbol{B}$ and the down-link hardware transfer functions.

where the two $M \times M$ and $K \times K$ diagonal matrices $\boldsymbol{R}_{\mathrm{B}}$ and $\boldsymbol{R}_{\mathrm{U}}$ model the non-reciprocal hardware responses of BS and UE receivers (RXs), respectively, and the two $M \times M$ and $K \times K$ diagonal matrices $\boldsymbol{T}_{\mathrm{B}}$ and $\boldsymbol{T}_{\mathrm{U}}$ similarly model hardware responses of their transmitters (TXs). Thus, in order to construct a precoder based on the UL channel estimates, the non-reciprocal components of the channel have to be calibrated. Previous calibration work showed that this is possible by using

$$\boldsymbol{C} f_{\mathrm{pre}}(\boldsymbol{G}) = f_{\mathrm{cal}}(f_{\mathrm{pre}}(\boldsymbol{G})), \tag{6}$$

where $\boldsymbol{C} = \boldsymbol{R}_{\mathrm{B}}\boldsymbol{T}_{\mathrm{B}}^{-1}$ is the, so-called, calibration matrix which can be estimated internally at the BS [7]. Such calibration is sufficient to cancel inter-user interference stemming from non-reciprocity [8].

2.4 Linear Detection & Precoding Schemes

Table 1 shows a selection of weighting matrices used in linear precoding and detection schemes, with non-reciprocity compensation included in the form of the $M \times M$ diagonal matrix $\boldsymbol{C}$ as defined above. The maximum ratio transmission (MRT) precoder and the maximum ratio combining (MRC) decoder maximize array gain without active suppression of interference among the UEs [1]. The zero forcing (ZF) precoder and ZF combiner employ the pseudo-inverse, which provides inter-user interference suppression with the penalty of lowering the achievable array gain. A scheme that allows trade-off between array gain and interference suppression is the regularized ZF (RZF) precoder and RZF combiner. This is achieved by properly selecting the regularization constants $\beta_{\mathrm{reg}_{\mathrm{pre}}}$ and $\beta_{\mathrm{reg}_{\mathrm{dec}}}$. If $\beta_{\mathrm{reg}_{\mathrm{pre}}}$ and $\beta_{\mathrm{reg}_{\mathrm{dec}}}$ are selected to minimize mean-squared error (MSE) $E\|\boldsymbol{u} - \frac{1}{\sqrt{\rho}}\hat{\boldsymbol{u}}\|^2$, where ρ is a scaling constant, we obtain the minimum MSE (MMSE) precoder/detector [9].

3 System Design Aspects

Having discussed the MaMi basics, we move on to system design aspects. These include modulation scheme, frame structure and hardware requirements.

3.1 Modulation Scheme

While many different modulation schemes can be used with MaMi, this paper focuses on OFDM, employed in many modern wireless communication systems. Properly designed OFDM renders frequency-flat narrowband subcarriers, facilitating the single channel equalization strategy used here.

For ease of comparison and simplicity, LTE-like OFDM parameters, as shown in Table 1, are used throughout this discussion. The more common parameters with LTE, the easier it is to evaluate how MaMi as an add-on would influence current cellular systems.

Table 2: High-level system parameters

Parameter	Variable	Value
Bandwidth	W	20 MHz
Sampling Rate	F_s	30.72 MS/s
FFT Size	N_FFT	2048
# Used subcarriers	N_used	1200
Cyclic prefix	N_cp	144 samples
OFDM symbol length	t_OFDM	71.4 µs

3.2 TDD versus FDD

Current cellular systems either operate in division duplex (FDD) or TDD mode. FDD is, however, considered impractical for MaMi due to excessive resources needed for DL pilots and CSI feedback. TDD operation relying on reciprocity only requires orthogonal pilots in the UL from the K UEs, making it the feasible choice [10]. For this reason, we focus entirely on TDD below.

3.3 Reciprocity

To allow operation in TDD mode, differences in the TX and RX transfer functions on both, the BS and UEs have to be calibrated as discussed in Sec. 2.3. Drifts over time are mainly caused by HW temperature and voltage changes, and thus, the calibration interval depends on the operating environment of the BS.

3.4 Frame Structure

The frame structure defines among other things, the pilot rate which determines how well channel variations can be tracked and, indirectly, the largest supported UE speed.

Mobility

The maximum supportable mobility, e.g., the maximum speed of the UEs is defined by the UL pilot transmission interval. In order to determine this constraint, a 2D wide-sense stationary channel with uncorrelated isotropic scattering is assumed. For the contributions from the different BS antennas to add up coherently high channel correlation is required and, as an approximation to formulate the final requirement, a correlation of 0.9 was used to ensure sufficient channel coherency. Further discussions on such modeling assumption are found in [11]. Although these assumptions may not be completely valid for MaMi channels, they allow an initial evaluation based on a maximum supported Doppler frequency, ν_max, by solving

$$J_0(2\pi\nu_\mathrm{max}T_\mathrm{p}) = 0.9, \tag{7}$$

for $\nu_{\max}$, where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind, stemming from a standard Jakes' fading assumption, and T_p the distance between pilots in time. Hence, the maximum supportable speed of any UE may be evaluated using

$$\mathrm{v}_{\max} = \frac{c\nu_{\max}}{f_\mathrm{c}}, \tag{8}$$

once a specific frame structure is provided. In (8) $\mathrm{v}_{\max}$ is the maximum supported speed of a UE, c the speed of light and f_c the chosen carrier frequency.

Processing latency

The frame structure has to be designed for the highest speed of UEs to be supported which requires a high pilot rate for high mobility scenarios. Within two consecutive UL pilot symbols, all UL data, DL data and guard symbols have to be accommodated which in turn decreases the available time between UL pilot reception and DL transmission. In a high mobility scenario this poses tight latency requirements for TDD transmission as CSI has to be estimated in order to produce the precoding matrix to beamform the DL data.

To formulate the TDD precoder turnaround time, Δ, all HW units introducing a delay must be taken into account. This includes the analog front-end delays for the TX $\Delta^{\mathrm{rf,TX}}$ and RX $\Delta^{\mathrm{rf,RX}}$, the processing latency for OFDM modulation/demodulation (including cyclic prefix (CP) and guard band operation) Δ^{OFDM}, the time for processing UL pilots to estimate CSI Δ^{CSI}, and the processing latency for precoding $\Delta^{\mathrm{precode}}$ including reciprocity compensation. Additional sources of latency include overhead in data routing, packing, and unpacking, i.e., Δ^{rout} such that the overall TDD precoder turnaround time may be formulated as

$$\Delta = \Delta^{\mathrm{rf,TX}} + \Delta^{\mathrm{rf,RX}} + \Delta^{\mathrm{OFDM}} + \Delta^{\mathrm{CSI}} + \Delta^{\mathrm{precode}} + \Delta^{\mathrm{rout}}. \tag{9}$$

Depending on the specific arrangement of the OFDM symbols and the pilot repetition pattern in the frame structure, base-band processing solutions, especially Δ^{CSI} and $\Delta^{\mathrm{precode}}$, have to be optimized to not violate the given constraint, i.e., Δ.

Pilot pattern

In general, to acquire CSI at the BS, the K UEs transmit orthogonal pilots on the UL. Different approaches are, e.g., distributed pilots over orthogonal subcarriers [12] or sending orthogonal pilot sequences over multiple subcarriers [13–15] but also semi-blind and blind techniques have been proposed [16].

Figure 2 shows a generic frame structure capturing the aforementioned aspects in a hierarchical manner assuming all UEs transmit their pilots within one dedicated pilot symbol. At the beginning of each BS reciprocity cycle, reciprocity calibration at the BS is performed and within these a certain number of DL pilot cycles are encapsulated where precoded DL pilot symbols are transmitted. The length of the BS reciprocity cycle is determined by the stability of the transceiver chains in the BS. As the reciprocity calibration at the BS side only compensates for BS transceivers, DL

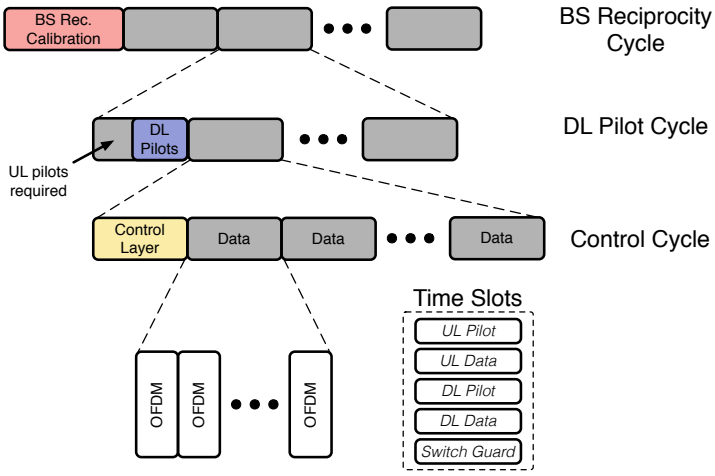

Figure 2: Generic frame structure of a LTE like TDD-based MaMi system. Within one BS reciprocity cycle the BS operates using the same reciprocity calibration coefficients. A certain number of DL pilot cycles are integrated as UEs suffer from faster changing environments. Each control cycle contains a control layer to perform, for example over-the-air synchronization and within these the data transmission slots are encapsulated.

pilots are necessary to compensate for transceiver differences at the UE side. Their frequency depends on the stability at the UE side and can be considered significantly smaller than for the BS as UEs are subject to faster changes in their operational environment, e.g., thermal differences when having the UE in a pocket or using it indoors or outdoors. To be able to send precoded pilots on the DL, transmission of UL pilots is required beforehand. Several control cycles are embedded inside each DL pilot cycle carrying a certain number of data time slots. Time slots contain five different OFDM symbol types for physical layer implementation. These are *(i)* UL Pilot where the UEs transmit orthogonal pilots to the BS, *(ii)* UL Data where all UEs simultaneously send data to the BS, *(iii)* DL Pilot where the BS sends precoded pilots to all UEs, *(iv)* DL Data where the BS transmits data to all UEs and *(v)* Switch Guard, which idles the RF chains to allow switching from RX to TX or vice versa.

Table 3: Processing Requirements in a MaMi system

Function	**General**	**Specific**
	Gops/s	Gops/s
FFT/IFFT	$4M \log_2(N_{\mathrm{FFT}}) N_{\mathrm{FFT}} / t_{\mathrm{OFDM}}$	126
Detection	$4MKN_{\mathrm{used}} / t_{\mathrm{OFDM}}$	80
Precoding	$4MKN_{\mathrm{used}} / t_{\mathrm{OFDM}}$	80
Recip. Cal.	$4MKN_{\mathrm{used}} / t_{\mathrm{OFDM}}$	80
Pseudo-inv.	$4N_{\mathrm{used}} \left(2MK^2 + K^3\right) / \left(2t_{\mathrm{OFDM}}\right)$	1080

3.5 Hardware Requirements

To illustrate the required HW capabilities for the testbed, the values from Table 1 are used to estimate the Gops/s [4] and the data shuffling on a per OFDM symbol basis for the general case and a specific case assuming $M = 100$ and $K = 12$.

Processing Capabilites

Table 3 summarizes the overall number of real-valued arithmetic operations. For the processing estimates, it is assumed that each complex multiplication requires four real multiplications. Close to the antennas, M fast-Fourier transforms (FFTs) or inverse fast-Fourier transforms (IFFTs) are needed equating to 126 Gops/s. Data precoding and detection as well as reciprocity compensation require large matrix and vector multiplications, for instance, an $M \times K$ matrix with a $K \times 1$ vector leading to up to 80 Gops/s.

Finally, when using ZF, the pseudo-inverse matrix is required which includes the calculation of the Gram matrix requiring MK^2 multiplications with the $K \times K$ matrix inversion adding another K^3 in complexity assuming a Neumann-Series approximation [17] or a QR decomposition. The last multiplication of the inverse with the Hermitian of the channel matrix $\boldsymbol{H}$ needs another MK^2 multiplications which combined with a requirement of finishing within two OFDM symbols leads to approximately 1 Tops/s for the overall pseudo-inverse calculation.

Data Shuffling Capabilities

Table 4 summarizes required interconnect bandwidth and number of links. Communication paths to each antenna transfer at the sampling rate of $F_{\mathrm{s}} = 30.72$ MS/s which is decreased to the subcarrier rate $F_{\mathrm{sub}} = 16.8$ MB/s by performing OFDM processing $(F_{\mathrm{s}} \cdot N_{\mathrm{used}}/(N_{\mathrm{FFT}} + N_{\mathrm{cp}}))$. Considering M antennas, the overall subcarrier data rate is $M \cdot w \cdot 16.8$ MB/s, with w being the combined wordlength for the in-phase and quadrature components in bytes. The information rate in an OFDM symbol carrying data is $K \cdot 16.8$ MB/s assuming 8 bit per sample, i.e., 256−QAM as

[4] Gops/s is used here, but these can be seen as GMACs/s, i.e., the number of multiply-accumulate operations, as almost all operations involve matrix-matrix and matrix-vector calculations.

Table 4: Data Shuffling Requirements in a MaMi system

Purpose	**General**	**Specific**
	#	#
Links to cent. proc	$2M$	200
	MB/s	MB/s
Antenna Rate	$w_{\text{ant}} M F_{\text{s}}$	w_{ant} 3,072
Subcarrier Rate	$w M F_{\text{sub}}$	w 1,680
Information rate	$K \cdot F_{\text{sub}}$	201.6

highest modulation. Assuming separate links between centralized processing and the antenna units on UL and DL, $2M$ peer-to-peer (P2P) links[5] are needed between the antennas and the centralized MIMO processing.

Reconfigurability

The testbed has to be reconfigurable and scalable, to support different system parameters, different processing algorithms and adaptive processing. It is also crucial to have the possibility to integrate in-house developed HW designs for validation and performance comparison of algorithms. Variable center frequencies, run-time adjustable RX and TX gains as well as configurable sampling rates are highly desirable to be able to adapt to other parameters than the ones presented in Table 1.

4 Generic Hardware and Processing Partitioning

In this section a generic HW and processing partitioning is presented to explore the parallelism in MaMi, which needs consideration of processing together with data transfer requirements (throughput, latency, # of P2P links), and at the same time provides scalability.

4.1 Hierarchical Overview

To be able to build a MaMi testbed with modular HW components, a hierarchical distribution as shown in Figure 3 is proposed. The main blocks are detailed as follows:

SDR

SDRs provide the interface between the digital and radio-frequency (RF) domain as well as local processing capabilities.

[5] In this discussion, each interconnection transferring data between physically separated devices is denoted a P2P link.

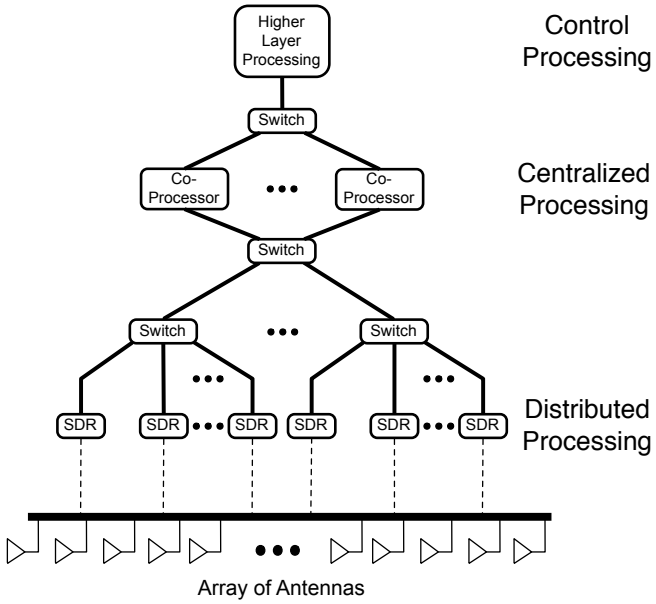

Figure 3: Hierarchical overview of a MaMi BS built from modular HW components.

Switches

Switches aggregate/disaggregate data between different parts of the system, e.g., between SDRs and the co-processors.

Co-processing modules

Co-processing modules provide a centralized node to perform MIMO processing.

Higher Layer Processing

Higher layer processing controls the system, configures the radios, and provides run-time status metrics of the system.

4.2 Processing and Data Distribution

For proper base-band processing partitioning, throughput constraints of HW components have to be taken into account. Assuming each SDR supports n_{ant} antennas, the required number of SDRs becomes $\lceil M/n_{\text{ant}} \rceil$ for an M-antenna system.

Subsystems

As shown in Figure 4, RF-Front End, OFDM processing and reciprocity compensation are performed on a per-antenna basis using the SDRs. This distributes a large fraction of the overall processing and reduces the data rate before transferring the acquired samples over the bus. Still, the number of direct devices on a bus is limited, and thus, setting up $2M$ P2P links directly to the co-processors would most likely exceed the number of maximum P2P links for any reasonable number of MaMi antennas. To reduce this number, data can be aggregated using the concept of grouping. The different data streams from several SDRs are interleaved on one common SDR and then sent via one P2P link. Therefore, subsystems are defined, each containing n_{sub} SDRs. Data from all antennas within a subsystem is aggregated/disaggregated on the outer two SDRs and distributed to the n_{co} co-processors using high-speed routers.

At closer look, Figure 4 reveals that the SDRs on the outer edges which realize the $(n_{\text{ant}}n_{\text{sub}})$ to (n_{co}) and (n_{co}) to $(n_{\text{ant}}n_{\text{sub}})$ router functionalities, require the highest number of P2P links, and thus have to deliver the highest throughput. Hence, the following inequalities have to be fulfilled for the subsystems not to exceed the constraints for maximum number of P2P links ($\text{P2P}_{\text{SDR,max}}$) and maximum bidirectional throughput ($R_{\text{SDR}_{\text{max}}}$):

$$R_{\text{SDR}_{\text{max}}} > R_{\text{SDR}_{\text{out}}} = R_{\text{SDR}_{\text{in}}} = n_{\text{ant}} \cdot n_{\text{sub}} \cdot w \cdot F_{\text{sub}} \tag{10}$$

$$\text{P2P}_{\text{SDR,max}} > \text{P2P}_{\text{SDR}} = n_{\text{co}} + n_{\text{sub}} \tag{11}$$

where it is assumed that if an SDR employs more than one antenna, the data is interleaved before it is sent to the router on the outer SDRs. The constraints given in equation (10)-(11) can be used to determine the maximum number of SDRs per subsystem (n_{sub}) such that hardware constraints are not exceeded.

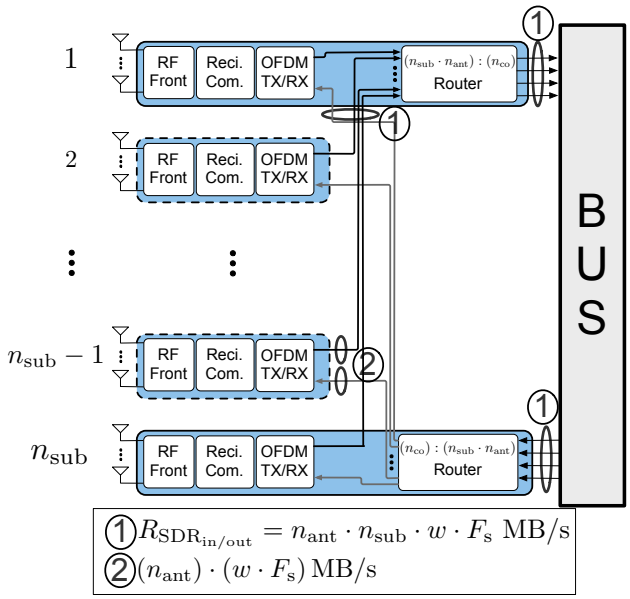

Figure 4: A subsystem consisting of n_{sub} SDRs where the two outer SDRs implement an antenna combiner / BW splitter and an antenna splitter / BW combiner, both implemented using high-speed FPGAs routers. Inter-SDR and SDR to central processor connections utilize a bus for transferring the samples.

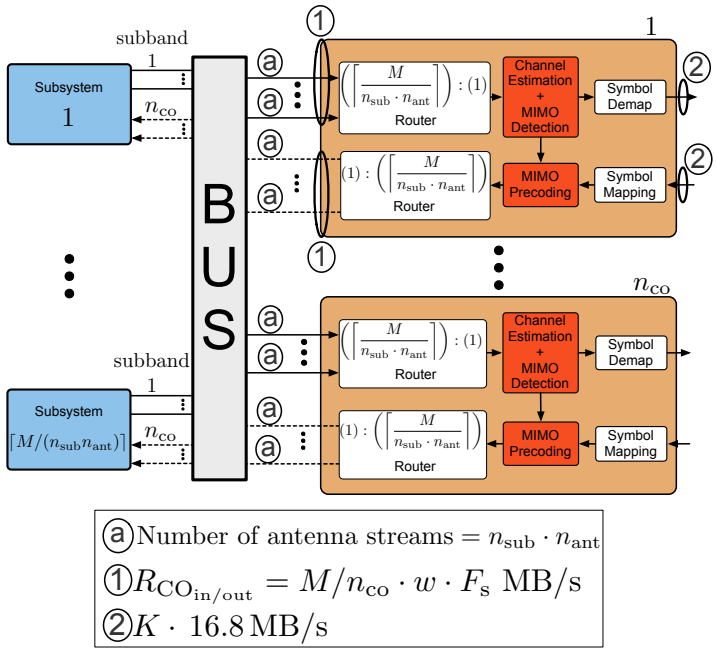

Figure 5: Shuffling data from the $\lceil M/(n_{\text{sub}}n_{\text{ant}})\rceil$ subsystems to the n_{co} co-processors. The routers use a simple round robin scheme to combine/distribute the data from/to corresponding subsystems.

Co-processors

As shown in Figure 5, detection, precoding, CSI acquisition, symbol mapping and symbol demapping are integrated in the centrally localized co-processor modules which collect data from all SDRs. Using CSI estimated from UL pilots, MIMO processing as discussed in Sec. 2 and symbol mapping/de-mapping is performed.

Based on the selected OFDM modulation scheme the subcarrier independence can be exploited allowing each of the n_{co} co-processors to work on a sub-band of the overall 20 MHz bandwidth. This efficiently circumvents issues with throughput and latency constraints in the MIMO signal processing chain. The co-processors aggregate/disaggregate data from all the antennas in the system using reconfigurable high-speed routers, as shown in Figure 5 for a system having $\lceil M/(n_{\text{sub}}n_{\text{ant}})\rceil$ subsystems and n_{co} co-processors.

Similarly to the SDRs, the two main constraints for the co-processors are the maximum number of P2P links denoted $\text{P2P}_{\text{CO,max}}$ and the maximum throughput denoted $R_{\text{CO}_{\text{max}}}$.

Table 5: Selected Hardware from National Instruments

Type	Model	Features
Host	PXIe-8135	• 2.3 GHz Quad-Core PXI Express Controller • Up to 8 GB/s system and 4 GB/s slot bandwidth
SDR	USRP RIO 294xR / 295xR	• 2 RF Front Ends and 1 Xilinx • Kintex-7 FPGA • Center frequency variable from 1.2 GHz to 6 GHz • 830 MB/s bidirectional throughput on up to 15 DMA channels
Co-Processor	FlexRIO 7976R	• 1 Xilinx Kintex-7 410T FPGA • 2.4 GB/s bidirectional throughput on up to 32 DMA channels
Switch	PXIe-1085	• Industrial form factor 18-slot chassis • 7 GB/s bidirectional throughput per slot • 2 switches per chassis with inter-switch traffic up to 3.2 GB/s • Links between chassis bound to 7 GB/s bidirectional
Expansion Module	PXIe-8374	• PXI Express (x4) Chassis Expansion Module • Software-transparent link without programming • Star, tree, or daisy-chain configuration
Reference Clock Source	PXIe-6674T	• 10 MHz reference clock source with < 5 ppb clock accuracy • 6 configurable I/O connections
Ref. Clock Distribution Distribution	OctoClock	• 10 MHz 8-channel clock and • timing distribution network

The following inequalities have to hold for the co-processor not to exceed these constraints:

$$\begin{aligned} R_{\mathrm{CO}_{\max}} &> R_{\mathrm{CO}_{\mathrm{out}}} = R_{\mathrm{CO}_{\mathrm{in}}} = \\ &= \left(\frac{M \cdot w + K}{n_{\mathrm{co}}}\right) \cdot F_{\mathrm{sub}} \qquad (12) \\ \mathrm{P2P}_{\mathrm{CO,max}} &> \mathrm{P2P}_{\mathrm{CO}} = 2 \cdot \lceil M/n_{\mathrm{sub}} \rceil + 2. \qquad (13) \end{aligned}$$

Using this modular and generic system partitioning, HW platforms built using modular components can be evaluated. Note, that expressions (10) - (13) may also be used with other system parameters, e.g., by redefining F_{s} and F_{sub}.

5 LUMAMI Testbed Implementation

In this section the LUMAMI specific implementation details are discussed based on the aforementioned general architecture. The LUMAMI system was designed with 100 BS antennas and can serve up to 12 UEs simultaneously. Based on these parameters, the selected modular HW platform is presented and given constraints are evaluated. Consequently, the specific frame structure and other features of the system including base-band processing, antenna array, mechanical structure and synchronization are briefly described. Before providing details, the authors would like to emphasize, that this is the initial version of the LUMAMI testbed and that add-ons and further improvements are planned for the future.

5.1 Selected Hardware Platform

The hardware platform was selected based on requirements discussed in Sec. 3. Table 5 shows the selected off-the-shelf modular hardware from National Instruments used to implement the LUMAMI testbed. The SDRs [18] allow up to 15 P2P links ($\mathrm{P2P}_{\mathrm{SDR,max}} = 15$) with a bidirectional throughput of $R_{\mathrm{SDR}_{\mathrm{max}}} = 830\,\mathrm{MB/s}$, support a variable center frequency from 1.2 GHz to 6 GHz and have a TX power of 15 dBm. Each SDR contains two RF chains, i.e., $n_{\mathrm{ant}} = 2$, and a Kintex-7 FPGA. Selected co-processors [19] allow a bidirectional P2P rate of $R_{\mathrm{CO}_{\mathrm{max}}} = 2.4\,\mathrm{GB/s}$ with up to $\mathrm{P2P}_{\mathrm{CO,max}} = 32$ P2P links and employ a powerful Kintex-7 FPGA with a reported performance of up to 2.845 GMACs/s [20]. This is sufficient for a 100 BS antenna MaMi testbed due to the fact that n_{co} co-processors can be utilized in parallel. Interconnection among devices is achieved using 18-slot chassis [21] combined with per-slot expansion modules [22]. Each chassis integrates two switches based on Peripheral Component Interconnect Express (PCIE) using direct memory access (DMA) channels which allow inter-chassis traffic up to 7 GB/s and intra-chassis traffic up to 3.2 GB/s.

The host [23] is an integrated controller, running LabVIEW on a standard Windows operating system and is used to configure and control the system. The integrated hardware/software stack provided by LabVIEW provides the needed reconfigurability as it abstracts the P2P link setup, communication among all devices and allows FPGA programming as well as host processing using a single programming language. An additional feature of LabVIEW is the possibility to seamlessly integrate intellectual property (IP) blocks generated via Xilinx Vivado platform paving a way to test in-house developed IP.

To be able to synchronize the full BS, a Reference Clock Source [24] and Reference clock distribution network [25] are required. Their functionalities will be later discussed when presenting the overall synchronization method.

5.2 Subsystems and Number of Co-processors

To build the LUMAMI testbed with $M = 100$ antennas, 50 SDRs are necessary. The maximum possible subsystem size is chosen to minimize the utilization of available

Table 6: System Parameters and validation of constraints in the LuMaMi testbed.

Parameters		**Rates** MB/s
M	100	$R_{\mathrm{SDR}_{\mathrm{max}}} = 830 > R_{\mathrm{SDR}_{\mathrm{out}}} = R_{\mathrm{SDR}_{\mathrm{in}}} = 806.4$
K	12	$R_{\mathrm{CO}_{\mathrm{max}}} = 2,400 > R_{\mathrm{CO}_{\mathrm{out}}} = R_{\mathrm{CO}_{\mathrm{in}}} = 1,460$
n_{ant}	2	**P2P Links**
n_{sub}	8[a]	$\mathrm{P2P}_{\mathrm{SDR,max}} = 15 > \mathrm{P2P}_{\mathrm{SDR}} = 12$
n_{co}	4	$\mathrm{P2P}_{\mathrm{CO,max}} = 32 > \mathrm{P2P}_{\mathrm{CO}} = 18$

[a] Note, that the last subsystem only consists of two SDRs.

P2P links at the co-processors. By using (10) and an internal fixed-point wordlength of $w = 3$ corresponding to a 12-bit resolution on the I- and Q-components, n_{sub} is found to be 8. As this is not an integer divider of 50, the last subsystem only contains two SDRs.

Based on Table 4, the combined subcarrier rate for all antennas is $wMF_{\mathrm{sub}} = 5\,\mathrm{GB/s}$ and another $K \cdot F_{\mathrm{sub}} = 200\,\mathrm{MB/s}$ are needed for information symbols. To not exceed $R_{\mathrm{CO}_{\mathrm{max}}}$ at least three co-processors must be utilized. To further lower the burden on the design of the low-latency MIMO signal processing chain, $n_{\mathrm{co}} = 4$ is chosen such that each co-processor processes 300 of the overall 1200 subcarriers.

Table 6 summarizes the LUMAMI testbed parameters and shows that constraints are met according to (10)-(13). It can also be seen that the design is still within the constraints if scaling up the number of BS antennas to $M = 128$, which has been done in subsequent designs based on the same hardware, e.g., [6].

5.3 Frame Structure

The default frame structure for the LUMAMI testbed is shown in Figure 6. One frame is $T_{\mathrm{f}} = 10\,\mathrm{ms}$ and is divided in ten subframes of length $T_{\mathrm{sf}} = 1\,\mathrm{ms}$. Each subframe consists of two slots having length $T_{\mathrm{slot}} = 0.5\,\mathrm{ms}$, where the first subframe is used for control signals, e.g., to implement over-the-air synchronization, UL power control and other control signaling. The 18 slots in the other nine subframes encapsulate seven OFDM symbols each. Comparing to Figure 2, a reciprocity calibration cycle is defined over the whole run-time of the BS for simplicity and due to the fact that there is no large drift after warming up the system in a controlled environment [5]. The DL pilot cycles and control cycles are both set to be the length of one frame. Each frame starts with one control subframe followed by one subframe with one DL pilot and one DL data symbol whereas all others use two DL data symbols.

5.4 Mobility

The pilot distance in time in the default frame structure given in Figure 2 is $T_{\mathrm{p}} \approx 430\,\mu\mathrm{s}$ or six OFDM symbols. Thus, $\nu_{\mathrm{max}} \approx 240\,\mathrm{Hz}$ for a correlation of 0.9. Due to

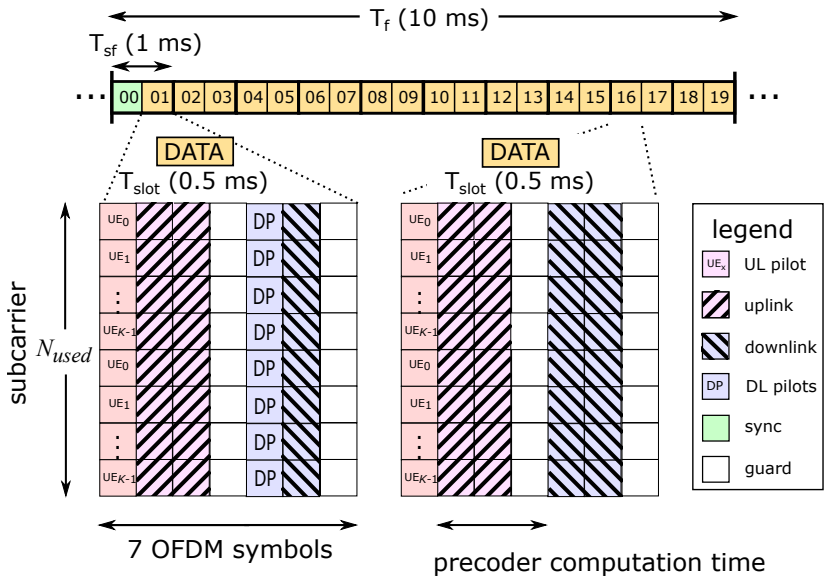

Figure 6: The default frame structure used in the LUMAMI testbed.

availability from a network operator, a carrier frequency of $f_c = 3.7\,\text{GHz}$ is selected. Using (8), $v_{max} = 70\,\text{km/h}$ is found as maximum supported speed.

5.5 TDD Turnaround Time

The pre-coding turnaround time requirement for the implementation can be analyzed based on (9). The analog front-end delay of the SDRs was measured to be about 2.25 µs. Taking the frame structure in Figure 6 (assuming $\Delta^{\text{rf,TX}} = \Delta^{\text{rf,RX}}$ which is not necessarily true), the latency budget for base-band processing is as follows: Overall time for pre-coding after receiving the UL pilots is 214 µs (3 OFDM symbols). The 2048 point FFT/IFFT (assuming a clock frequency of 200 MHz) requires around $35\,\mu\text{s} \times 2 = 70\,\mu\text{s}$ in total for TX and RX (including sample reordering). As a result, the remaining time for channel estimation, MIMO processing, and data routing is around 140 µs, which is the design constraint for this specific frame structure.

An analysis of the implemented design showed that the latency is far below the requirement for the default frame structure which makes it possible to use the testbed for higher mobility scenarios from this point of view [26].

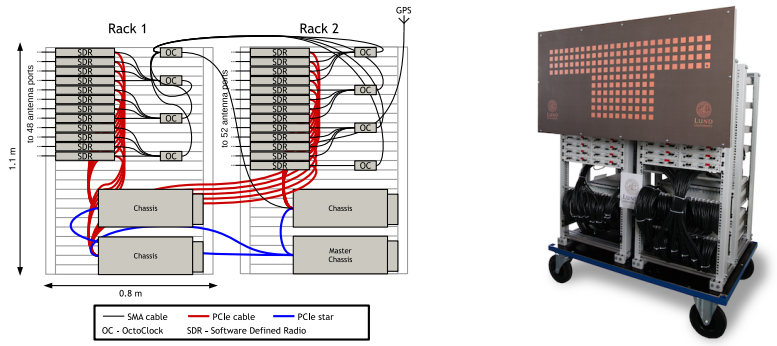

Figure 7: Left: Side view of the mechanical assembly of the BS. The two racks sit side by side (not as shown) with the SDRs facing the same direction (towards the antenna array). Two columns of USRP SDRs are mounted in each rack, totaling 50 of them. Right: The assembled LUMAMI testbed at Lund University, Sweden.

5.6 Implementation Features

Base-band Processing

On the LUMAMI testbed, each UE sends pilots on orthogonal subcarriers, i.e., each UE uses every K-th subcarrier with the first UE starting at subcarrier 0, the second at subcarrier 1 etc., overall utilizing a full OFDM symbol. It was shown that performance does not suffer significantly compared to a full detector calculated for each subcarrier using this method [12]. Moreover, it efficiently remedies processing requirements and reduces the required memory for storing estimated CSI matrices by a factor of K. A least-square CSI estimation algorithm with zeroth-order hold over $K = 12$ subcarriers was implemented, however, better estimates could be obtained by on-the-fly interpolation between the estimated subcarriers. Overall, utilizing this approach reduces the required detection matrix throughput to one matrix every 12 subcarriers, i.e., 16.8×10^6 subcarriers/s/12 $= 1.4 \times 10^6$ DetectionMatrices/s.

Two versions for detection were implemented. The first one based on a QR decomposition of the channel matrix augmented with the regularizations factors to a matrix of size $2M \times K$. This is then formulated into a partial parallel implementation employing a systolic array [27]. The latter one based on a Neumann-series [17]. In the QR decomposition, each column is processed using the discrete steps of the modified Gram-Schmidt algorithm. The logic on the co-processors can be reconfigured so that the same hardware resources that provide the RZF decoder can also provide the ZF and MRC decoders, i.e., the detection / precoding schemes discussed in Sec. 2 are supported with run-time switching. The Neumann-series based ZF detector utilizes the unique property that in MaMi, the Gramian matrix shows dominant diagonal elements if UEs use UL power control, or if scheduling is performed to serve UEs

Table 7: FPGA Utilization for two different MIMO processing implementations

Implementation	Registers	LUT	RAMs	DSP48
QRD	46470 (9.1%)	49315 (20.3%)	171 (21.5%)	596 (38.7%)
Neumann-Series	16000 (3.1%)	28700 (11.8%)	6 (0.75%)	176 (11.4%)

with similar power levels in the same time/frequency block to mitigate the influence of path loss differences. This, allows the matrix inversion to be approximated with low overall error [17]. The utilizations for the two FPGA designs are shown in Table 7. Clearly, overall processing complexity and resource utilization can be significantly reduced by exploiting the special properties of MaMi.

At this point, the regularizations factors $\beta_{\mathrm{reg}_{\mathrm{pre}}}$ and $\beta_{\mathrm{reg}_{\mathrm{dec}}}$ are not run-time optimized but set manually, however, implementation of this feature is planned in future. For a more detailed discussion of the low-latency signal processing implementation on the testbed we refer to [26].

Host-based visualization and data capturing

The available margin of 1 GB/s and 14 P2P links to the corresponding maximum values on the co-processors are used for visualization and system performance metrics. The host receives decimated equalized constellations and raw subcarriers for one UL pilot and one UL data symbol per frame. These features add another

$$\frac{300 \cdot 2\mathrm{bytes} + 2 \cdot 300 \cdot 4\mathrm{bytes}}{10\mathrm{ms}} = 300\,\mathrm{MB/s}$$

of data flowing in and out of the co-processor. The raw subcarriers are used to perform channel estimation and UL data detection on the host computer with floating point precision and allow fast implementation of different metrics, like constellation, channel impulse response, power level per antenna and user. Another 12 P2P links available are utilized to transmit and store real-time BER for all 12 UEs.

Moreover, to be able to capture dynamics in the channel for mobile UEs, CSI can be stored on a ms basis. An integrated 2 GB DRAM buffer on each of the co-processors was utilized for this since direct streaming to disk would exceed the P2P bandwidth limits. Snapshots can either be taken for 60 s in a 5 ms interval or over 12 s in a 1 ms interval, both corresponding to 2 GB of data for 300 subcarriers per co-processor.

Scalability/Reconfigurability

Before startup, the number of deployed BS antennas can be arbitrarily set between 4 and 100. This is achieved by introducing zeros for non-existing antennas within

the lookup-table (LUT)-based reconfigurable high-speed routers on the co-processors, thereby allowing to evaluate effects of scaling the BS antennas in real environments [26]. Additionally, all 140 OFDM symbols in a frame can be rearranged arbitrarily before start-up while each frame always repeats itself. For instance, we can choose to set the first symbol as UL pilots and all others as UL data in a static UL only scenario.

Reciprocity Calibration

Estimation of the reciprocity calibration coefficients was implemented on the host, mainly for two reasons: *(i)* the host can perform all operations in floating-point which increases precision and *(ii)* the drift of the hardware is not significant once the system reached operating temperature [5]. Estimated reciprocity coefficients are applied in a distributed manner on the SDRs [26].

5.7 Mechanical structure and electrical characteristics

Two computer racks containing all components measuring $0.8 \times 1.2 \times 1\,\text{m}$ were used, as shown Figure 7. An essential requirement for the LUMAMI testbed is to allow tests in different scenarios, e.g., indoor and outdoor. Therefore, the rack mount is attached on top of a 4-wheel trolley.

5.8 Antenna Array

The planar T-shaped antenna array with 160 dual polarized $\lambda/2$ patch elements was developed in-house. A 3.2 mm Diclad 880 was chosen for the printed circuit board substrate. The T upper horizontal rectangle has 4×25 elements and the central square has 10×10 elements (see Figure 7 right). This yields 320 possible antenna ports that can be used to explore different antenna array arrangements, for example 10×10 or 4×25 with the latter one being the default configuration. All antenna elements are center shorted, which improves isolation and bandwidth. The manufactured array yielded an average 10 dB-bandwidth of 183 MHz centered at 3.7 GHz with isolation between antenna ports varying between 18 dB and 28 dB depending on location in the array.

5.9 User Equipment

Each UE represents a phone or other wireless device with single antenna capabilities. One SDR serves as two independent UEs such that overall six SDRs are required for the 12 UEs. The base-band processing, i.e., OFDM modulation/demodulation and symbol mapping/demapping are essentially identical to the BS implementation. A least-square CSI acquisition is performed on precoded DL pilot followed by a ZF-equalizer. The DL pilots occupy a full OFDM symbol. The UEs may be equipped with any type of antenna using SMA connectors.

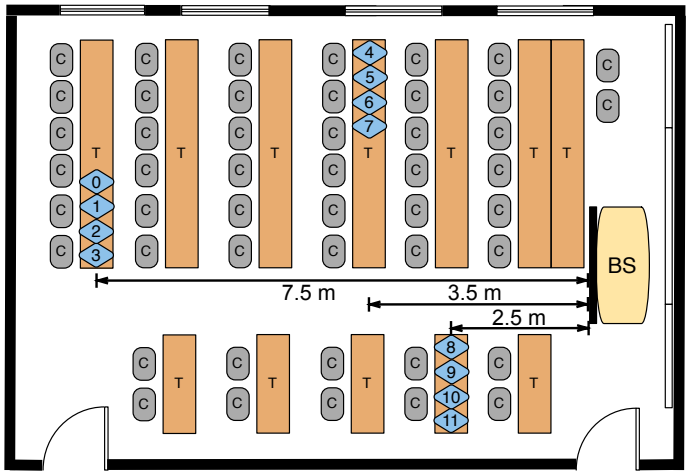

Figure 8: The indoor measurement setup in a lecture room including the positions of the 12 UEs. The BS is shown at the right-hand side and is situated at the front of the lecture hall. The terminals are placed in groups of four on three different tables and distances to the BS.

5.10 Synchronization

A MaMi BS requires time synchronization and phase coherence between each RF chain. This is achieved using the 10 MHz reference clock source and the reference clock and trigger distribution network (see Table 5). The reference clock is used as the source of each radio local oscillator, providing phase coherence among devices. The trigger signal is used to provide a time reference to all the radios in the system. A master provides an output digital trigger that is amplified and divided among all the radios. Upon receipt of the rising edge of the event trigger, all SDRs are started. The basic structure can be identified in Figure 7 on the left.

To synchronize the UEs with the BS over-the-air (OTA), the LTE Zadoff-Chu Primary Synchronisation Signal (PSS) is used, which occupies the center 1.2 MHz of the overall bandwidth. OTA synchronization and frequency offset compensation are achieved by employing a frequency-shifted bank of replica filters. The process follows a two step procedure: finding a coarse candidate position by scanning over the whole radio frame followed by tracking the PSS in a narrowed window located around the coarse candidate position. Additionally, by disciplining the UE SDRs with Global Positioning System (GPS), frequency offset compensation may be avoided by lowering the frequency offset to < 300 Hz.

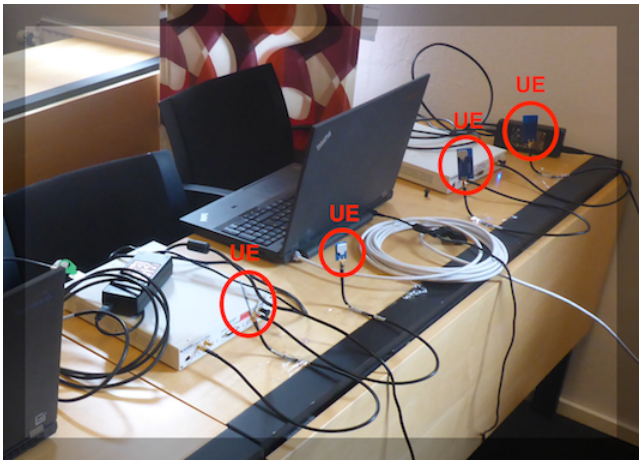

Figure 9: One group of four UEs with a high user density per unit area to validate the spatial multiplexing capabilities of MaMi.

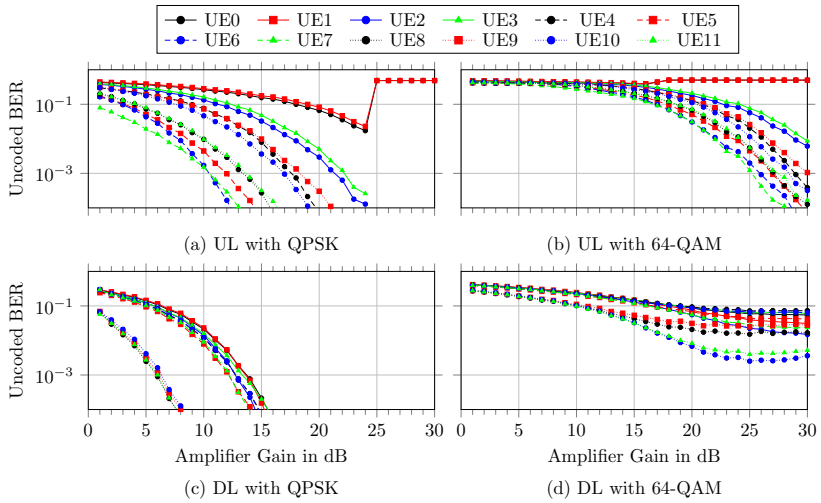

Figure 10: UL and DL BER for 12 UEs with ZF decoder/precoder.

6 Proof-of-concept Results

This section describes two experiments performed to validate our testbed design, the MaMi concept and its performance. The first test is performed indoors with high

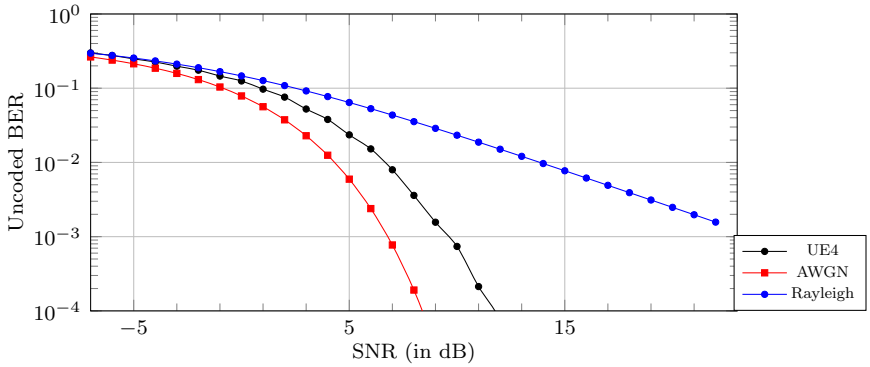

Figure 11: Comparing the BER of UE4 to AWGN and Rayleigh fading channels.

density of users per area unit to stress the spatial multiplexing capabilities of the system. The second test is conducted outdoors with less dense deployment of UEs and is primarily designed to test the range and multiplexing capabilities outdoors. For all tests, the default antenna configuration, i.e., 4×25 was used on the BS side whereas the UEs were equipped with linear polarized ultra-wideband antennas. It has to be noted that all results shown in this section are obtained from real-time operation without UL power control.

6.1 Indoor Test

In this test real-time uncoded BER curves are measured, employing MRC/MRT and ZF as decoders/precoders. The UL BER curves are obtained by sweeping all UE TX power amplifier (PA) gains synchronously, and for the DL BER curves the PA gains of the BS TX chains while keeping other system parameters constant. Note that the initial parameterization of the system is chosen empirically, so it allows smooth BER curves starting at about 0.5. Each gain step is held constant for about 4 s corresponding to about 36×10^6 and 108×10^6 transmitted bits per step for QPSK and 64-QAM modulation, respectively.

Scenario

Twelve UEs are set up in a lecture hall at Lund University with the BS at the front as shown in Figure 8 including the respective UE placements. All UEs are packed in groups of four resulting in a high density of UEs per area unit. One of these groups can be seen in Figure 9.

UL BER

Figure 10, (a) and (b), show the BER for all 12 UEs using ZF detector for QPSK and 64-QAM modulation, respectively. For both constellation sizes, the UEs furthest away, UE0 to UE3 show highest BER. UE0 and UE1 even show a sudden increase for the BER to 0.5 which was diagnosed to be due to saturation of their respective PAs. Moreover, their performance shows severe limitation compared to the other UEs, giving a clear indication that their performance is interference rather than power limited. The group closest to the BS, UE9-UE12, shows best performance although the variation within the group is still quite significant. Overall, the expected trend, increasing performance with increased transmit gain is clearly noticeable with the BER curve shapes resembling those of AWGN channels. Comparing the amplifier gain settings for QPSK and 64-QAM to achieve the same BER the differences are found to be in the range of 10 dB to 16 dB whereas a difference of 9 dB is expected for AWGN. Overall, it can be seen that all UEs except UE0 and UE1 achieve BER below 10 % at an amplifier gain of 15 dB for QPSK and 25 dB for 64-QAM, respectively.

DL BER

Figure 10, (c) and (d), show the DL BER using ZF precoder for QPSK- and 64-QAM modulation, respectively. Using QPSK modulation, the group closest to the BS, UE9-UE12, achieves a considerably better performance than the other two groups. Using 64-QAM, all UEs show an error-floor towards higher TX gain values which is likely a result of imperfect reciprocity calibration combined with leakage among UEs due to non-perfect channel knowledge resulting in interference among UEs. However, for the QPSK modulation case all UEs experience better BER rates which can be explained by the significantly higher available transmit power on the BS side, utilizing 100 active RF-chains. Comparing again the difference in amplifier gain setting for QPSK and 64-QAM, their differences are about 12 dB to 16 dB. The tests performed were mainly to prove functionality, and thus, no special care was taken to achieve best possible accuracy for the reciprocity calibration. However, individual parts are continuously tested to be improved.

Performance Evaluation

While the BER plots in Fig. 10 nicely show the trend with increasing transmit power, they do not provide a real performance indication against signal-to-noise ratio (SNR). The current implementation of the testbed does not provide SNR estimates in real-time such that the data presented in Fig. 10 can be seen as the raw data provided during measurements. To provide an indication of the system performance the SNR of UE4 was estimated based on the received UL channel estimates. Estimated sub-carriers at different time instances (about 200 ms apart) were subtracted / added to extract the noise / signal plus noise level which was then used to calculate the SNR value. However, this practice has limits as for close users interference may be stronger than the noise whereas for far away users the signal level may be too low. Therefore,

UE4 was chosen which due to its placement during the measurement allowed a relatively good SNR estimation. Fig. 11 shows the BER of UE4 in comparison with the theoretical performance in AWGN and Rayleigh fading channels. It is visible that due to the excess amount of BS antennas the performance is close to the AWGN channel. To be more specific, due to the channel hardening the performance is only about 3 dB worse than for a AWGN channel which would be achieved for perfect channel hardening. On the DL the SNRs are affected by several factors including the higher overall transmit power from the 100 active RF-chains and possible inaccuracies in the reciprocity calibration coefficients. As DL precoding is performed based on UL channel estimates, SNR estimation is practically not feasible.

As all shown BER curves closely resemble the shape of an AWGN channel it can be claimed that the MaMi concept works and is capable of serving 12 UEs on the same time/frequency resource even with a high UE density which in turn significantly improves the spectral efficiency compared to current cellular standards.

MRC/MRT versus ZF

To compare the performance of MRC/MRT and ZF it is beneficial to isolate the analysis to one UE. Figure 12a and Figure 12b show the BER for UE7 for QPSK, 16-QAM and 64-QAM modulations while the BS employs either MRC/MRT or ZF on the UL and DL, respectively.

Overall, ZF shows an superior performance trend with increasing PA gains, while the performance of MRC appears to level off[6]. Looking in more detail, ZF is capable of achieving more than an order of magnitude lower BER, compared to MRC. Using higher constellation sizes, 16-QAM or 64-QAM, the results for MRC show an even more significant deterioration. On the DL, ZF also outperforms MRT by far, the latter shows a significant error floor towards higher gains as in the UL case.

Unfortunately, direct comparison between UL and DL results shown here is not easy to perform. This is due to the fact that on the UL, the performance is isolated to the UL transmit power only whereas on the DL a combination of UL channel estimate quality, DL transmit power and reciprocity accuracy determines overall performance.

6.2 Outdoor Test

For the outdoor test, the testbed was placed on the rooftop of one of the wings of the department building while the UEs where placed on the opposite wing utilizing scaffolding mounted to the building. Up to eight UEs were served simultaneously in a distance of about 18 to 22 meters, six on the second floor and two on the first floor while the testbed was situated on the third floor (rooftop). The scenario is shown in Figure 13.

Figure 14 shows the BS placed on the rooftop of the department building facing towards the opposite wing. The placement for UEs 0 and 1 is also marked.

[6] This is expected from theory, as inter-user interference is the main source of error during data detection. The high density users setup adopted in this experiment highly contributes to this phenomena.

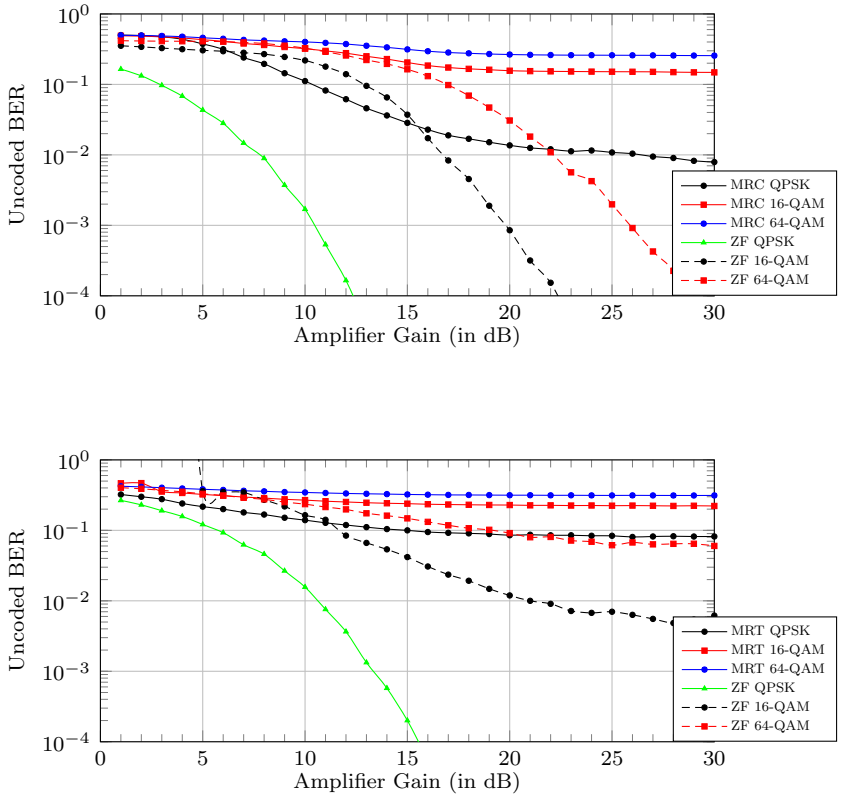

Figure 12: BER for UEs7 using QPSK, 16-QAM and 64-QAM modulation. Up–UL for ZF and MRC detector; Down–DL for ZF and MRT precoder.

Figure 15 shows a screenshot of the received UL QPSK constellations for this test setup when using MRC and ZF, respectively. Using MRC without error-correcting code (ECC) for this test, the six UEs show significant interference. Therefore, focus is put on the results obtained with ZF which is capable of separating up to eight UEs and shows very clear constellations, due to the interference suppression.

Considering ZF on the DL, the constellations for all 8 UEs can be seen in Figure 16. Although in-detail analysis is not provided for this test, it is clearly visible that ZF outperforms MRC which is often claimed to be sufficient in literature when analyzing

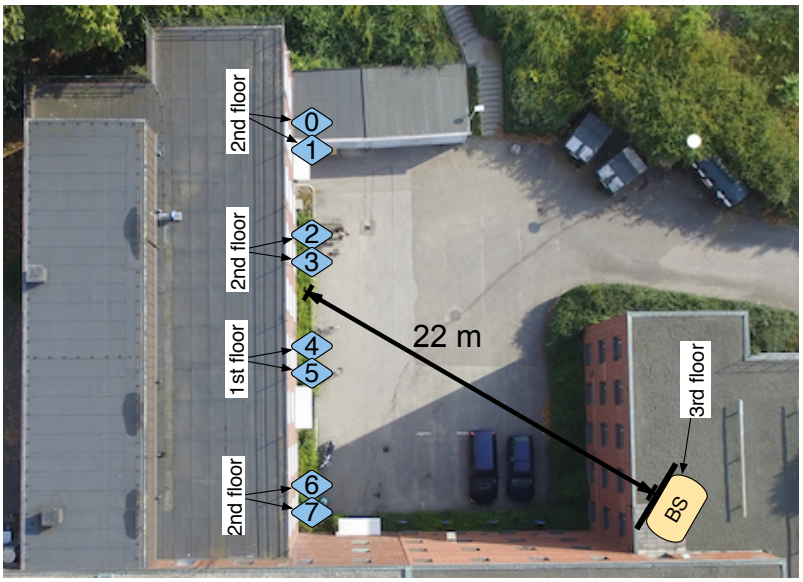

Figure 13: Scenario for the outdoor tests. BS placed on the rooftop of the building (third floor) serving eight UEs on the opposite wing, with six UEs on second floor and two UEs on first floor.

performance based on IID channel models [1]. The results observed in this experiment are representative for most tests performed so far, i.e., DL always showed to be the more challenging duplex case.

The LUMAMI testbed was also utilized to perform the first MaMi outdoor mobility measurements involving moving pedestrians and cars as UEs, however, a discussion of this is out of scope of this paper. Results and analysis from the mobility tests can be found in [28].

7 Conclusion

This paper presented the LUMAMI testbed, which is the first fully operational real-time testbed for prototyping massive MIMO. Based on massive MIMO system requirements, system parameters were discussed and defined. Further, a detailed generic hardware partitioning to overcome challenges for data shuffling and peer-to-peer link limitations while still allowing scalability, was proposed. By grouping software defined radios and splitting overall bandwidth, implementation of massive MIMO signal processing was simplified to cope with challenges like time division duplex precoding turnaround time and limited peer-to-peer bandwidth enforcing strict design require-

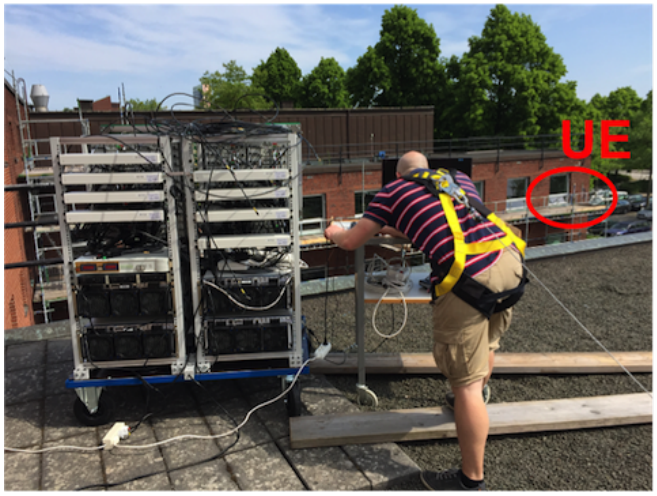

Figure 14: The outdoor test scenario setup with the BS deployed on the rooftop of the department building marked with two UEs on the opposite building wing.

ments when scaling the number of base station antennas up to 100 or higher. Based on the generic system partitioning and system requirements, a hardware platform was selected and evaluated. It was shown that internal system configuration is within throughput and processing capabilities before the complete LuMaMi testbed parameters were described. Finally, field trial results including Bit Error Rate performance measurements and constellations were presented from both indoor and outdoor measurement campaigns. The results showed that it is possible to separate up to 12 user equipments on the same time/frequency resource when using massive MIMO. Having established a flexible platform for testing new algorithms and digital base-band solutions we are able to take massive MIMO from theory to real-world tests and standardization for next generation wireless systems.

Acknowledgment

This work was funded by the Swedish foundation for strategic research SSF, VR, the strategic research area ELLIIT, and the EU Seventh Framework Programme (FP7/2007-2013) under grant agreement n 619086 (MAMMOET).

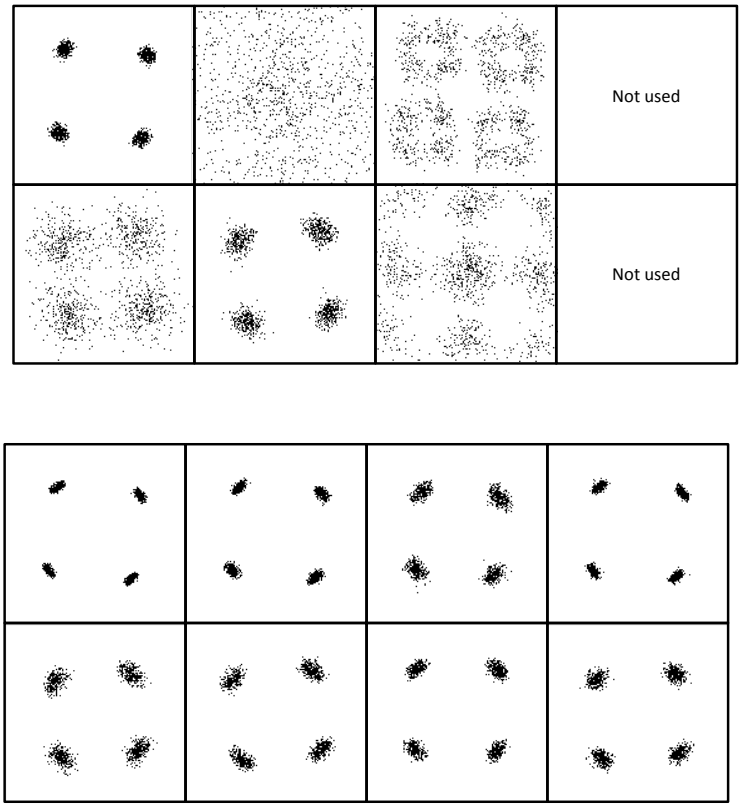

Figure 15: UL constellations for the outdoor experiment. Up–when using MRC with 6 UEs; Down–when using ZF to serve 8 UEs.

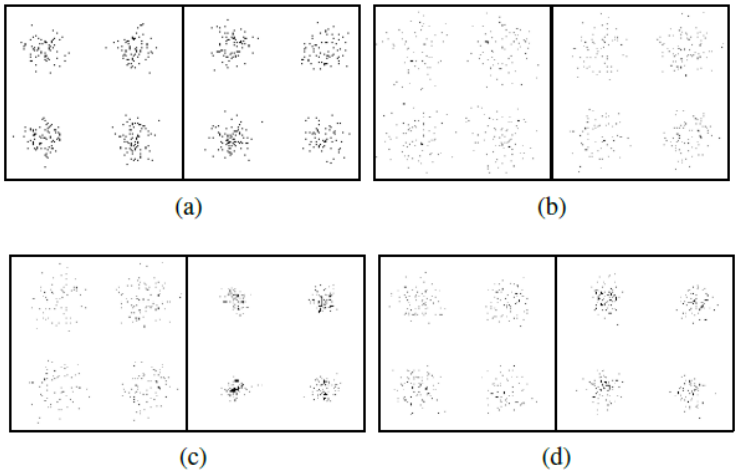

Figure 16: Received DL constellations using ZF: (a) UE0 & UE1; (b) UE2 & UE3; (c) UE5 & UE8; (d) UE9 & UE10.

References

[1] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser mimo systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, Apr 2013.

[4] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 53–64. [Online]. Available: http://doi.acm.org/10.1145/2348543.2348553

[5] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Owall, O. Edfors, and F. Tufvesson, "A flexible 100-antenna testbed for massive MIMO," in *Globecom Workshops (GC Wkshps)*, 2014, pp. 287–293.

[6] P. Harris, S. Zang, A. Nix, M. Beach, S. Armour, and A. Doufexi, "A distributed massive mimo testbed to assess real-world performance and feasibility," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–2.

[7] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson, "Reciprocity calibration for massive MIMO: proposal, modeling and validation," *CoRR*, vol. abs/1606.05156, 2016. [Online]. Available: http://arxiv.org/abs/1606.05156

[8] R. Rogalin, O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser mimo," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1815–1831, April 2014.

[9] E. Bjrnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, 2014.

[10] E. Bjrnson, E. G. Larsson, and T. L. Marzetta, "Massive mimo: ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, February 2016.

[11] A. Molisch, *Wireless Communications*, ser. Wiley IEEE. Wiley, 2010.

[12] MAMMOET (Massive MIMO for Efficient Transmission), "Ict-619086-d3.2: Distributed and centralized baseband processing algorithms, architectures, and platforms," EU-project Deliverable, Jan 2016, https://mammoet-project.eu/publications-deliverables.

[13] N. Shariati, E. Bjrnson, M. Bengtsson, and M. Debbah, "Low-complexity polynomial channel estimation in large-scale mimo with arbitrary statistics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 815–830, Oct 2014.

[14] Y. Han and J. Lee, "Uplink pilot design for multi-cell massive mimo networks," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1619–1622, Aug 2016.

[15] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 264–273, February 2013.

[16] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive mimo—5g systes," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 905–923, Secondquarter 2016.

[17] H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for very-large MIMO and applications to linear precoding systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 2710–2715.

[18] National Instruments. (2014) USRP-2943R Data Sheet. http://www.ni.com/datasheet/pdf/en/ds-538 (visited on 4 Oct. 2016).

[19] ——. (2014, Jul.) FlexRIO 7976R Data Sheet. http://www.ni.com/pdf/manuals/374546a.pdf (visited on 4 Oct. 2016).

[20] Xilinx. (2016) 7 series fpgas overview: Ds180 (v2.0) product specification. http://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf (visited on 4 Oct. 2016).

[21] National Instruments. (2015) PXIe 1085 Manual. http://www.ni.com/pdf/manuals/373712f.pdf (visited on 4 Oct. 2016).

[22] ——. (2011) MXI-Express x4 Series User Manual. http:/0/www.ni.com/pdf/manuals/371977c.pdf (visited on 4 Oct. 2016).

[23] ——. (2013) PXIe 8135 Manual. http://www.ni.com/pdf/manuals/373716b.pdf (visited on 4 Oct. 2016).

[24] ——. (2015) PXIe-6674T User Manual: Timing and Synchronization Module for PXI Express.

[25] Ettus Research. USRP Hardware Driver and USRP Manual: OctoClock. http://files.ettus.com/manual/page_octoclock.html (visited on 4 Oct. 2016).

[26] S. Malkowsky, J. Vieira, K. Nieman, N. Kundargi, I. Wong, V. wall, O. Edfors, F. Tufvesson, and L. Liu, "Implementation of low-latency signal processing and data shuffling for tdd massive mimo systems," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, Oct 2016, pp. 260–265.

[27] D. Wubben, R. Bohnke, V. Kuhn, and K. D. Kammeyer, "Mmse extension of v-blast based on sorted qr decomposition," in *IEEE 58th Vehicular Technology Conference*, vol. 1, Oct 2003, pp. 508–512 Vol.1.

[28] P. Harris, S. Malkowsky, J. Vieira, F. Tufvesson Wael Boukley Hassan, L. Liu, M. Beach, S. Armour, and O. Edfors, "Performance Characterization of a Real-Time Massive MIMO System with LOS Mobile Channels," *ArXiv e-prints*, Jan. 2017.

Paper III

Reciprocity calibration methods for Massive MIMO based on antenna coupling

In this paper we consider time-division-duplex (TDD) reciprocity calibration of a massive MIMO system. The calibration of a massive MIMO system can be done entirely at the base station (BS) side by sounding the BS antennas one-by-one while receiving with the other BS antennas. With an M antenna BS, this generates $M(M-1)$ signals that can be used for calibration purposes. In this paper we study several least-squares (LS) based estimators, differing in the number of received signals that are being used. We compare the performance of the estimators, and we conclude that is possible to accurately calibrate an entire BS antenna array using the mutual coupling between antennas as the main propagation mechanism.

1 Introduction

Massive MIMO has gained a lot of interest in the later years as it has a potential to increase the energy efficiency significantly of cellular networks compared to current technologies, while still providing a good network capacity and using mobile terminals with limited complexity [1]. In order to realize the true potential of this technology there are several practical challenges that need to be investigated, one of them being the reciprocity calibration problem [2]. Basically, one can not afford to transmit pilot symbols from every antenna in the downlink channel, receive them at the terminal side, and feed back channel state information (CSI) to the BS so that it can calculate suitable pre-coding coefficients. Such a procedure would degrade the spectral efficiency significantly considering the amount of feedback information required, due to the large number of BS antennas. Instead, a common approach is to operate in time division duplex (TDD) mode, and rely on the reciprocity of the channel to compute proper pre-coding coeficients based on uplink CSI.

It is generally agreed in wireless systems that the propagation channel is reciprocal, but the different transceiver radio frequency (RF) chains are not. Hence, in order to use reciprocity and calculate the pre-coding coefficients, we have to know or estimate the differences in the (frequency) responses between the uplink and downlink parts of the hardware chains. Such an estimation procedure is called reciprocity calibration.

Reciprocity calibration was discussed generally in [3]. A calibration scheme was presented where the reciprocity parameters are estimated based on bi-directional channel measurements. This requires feedback from one side of the link, thus making this approach not suitable in a massive MIMO context.

A novel massive MIMO calibration approach was proposed and implemented in a test bed in [4]. In this setup one of the antenna elements in the base station is used as a reference element, which successively transmits and receives pilot signals to and from all other antennas. The reciprocity calibration weights are simply calculated as the ratio between the forward and reverse radio channels with respect to this reference element. This method works well as long as the reference element has a good channel to all the other antenna elements, but has shown to be sensitive to the exact placement of the reference antenna.

In [5] the authors generalize the method presented in [4] and apply it in a distributed large-scale MIMO setup to calibrate access points. A robust least squares (LS) framework is derived based on successive transmission and reception of pilots solely between these access points. The methodology presented in our paper can be seen as an extension of this framework back to the case of Massive MIMO to calibrate a BS antenna array and its multiple RF-chains. Thus, instead of a random (often Rayleigh distributed) wireless channel between access points we have in our case a deterministic, often strong, component due to the antenna coupling. In this paper we *use* the mutual coupling between antennas to be able to estimate the reciprocity calibration coefficients.

The remainder of this paper is structured as follows: in Sec. 2 we introduce the system models used and the reciprocity calibration concept; in Sec. 3 we present the

different calibration methods studied; in Sec. 4 the impact of the calibration error in the capacity of a massive MIMO system is analyzed for different precoders; and finally Sec. 5 wraps up the paper.

2 System model

2.1 Channel Reciprocity

Due to the internal electronics of the BS and the single-antenna mobile stations (MS), the measured uplink/downlink channels are not only determined by the propagation channels, but those are also influenced by the RF chains. Let the uplink and downlink radio channels between the BS and MS be denoted as

$$\begin{aligned} g^{\mathrm{U}}_{m,k} &= r^{\mathrm{B}}_{m}\ \tilde{g}^{\mathrm{U}}_{m,k}\ t^{\mathrm{M}}_{k} \\ g^{\mathrm{D}}_{k,m} &= r^{\mathrm{M}}_{k}\ \tilde{g}^{\mathrm{D}}_{k,m}\ t^{\mathrm{B}}_{m}, \end{aligned} \tag{1}$$

where $m \in [0, ..., M-1]$ is the BS antenna index, $k \in [0, ..., K-1]$ is the MS antenna index, r^{B} and r^{M} represent the BS and MS receiver RF chains, t^{B} and t^{M} represent the BS and MS transmitter RF chains, and $\tilde{g}^{\mathrm{U}}$ and $\tilde{g}^{\mathrm{D}}$ are the uplink and the downlink propagation channels, respectively.

A relation between the uplink and downlink radio channels can be established as

$$g^{\mathrm{D}}_{k,m} = b_{m,k}\ g^{\mathrm{U}}_{m,k}. \tag{2}$$

Here we denote $b_{m,k}$ as the *calibration coefficient* between radios m and k, since if obtained, it allows to compute the downlink channel based on the uplink channel estimate. Assuming perfect reciprocity of the propagation channel, $b_{m,k}$ can be expanded as

$$b_{m,k} = \frac{r^{\mathrm{M}}_{k}\ \tilde{g}^{\mathrm{D}}_{k,m}\ t^{\mathrm{B}}_{m}}{r^{\mathrm{B}}_{m}\ \tilde{g}^{\mathrm{U}}_{m,k}\ t^{\mathrm{M}}_{k}} = \frac{r^{\mathrm{M}}_{k}\ t^{\mathrm{B}}_{m}}{r^{\mathrm{B}}_{m}\ t^{\mathrm{M}}_{k}}. \tag{3}$$

Hence, it can be seen that the non-reciprocity between radio channels can be calibrated externally, i.e., by feeding back the downlink channel. Such approach is unfeasible in a massive MIMO context, since for each terminal, the number of channel estimates to feedback to the BS scales with M [2].

2.2 Internal Calibration

Let us now introduce the channel between two BS radios as

$$h_{\ell,m} = r^{\mathrm{B}}_{\ell}\ \tilde{h}_{\ell,m}\ t^{\mathrm{B}}_{m} \tag{4}$$

where $\ell \neq m$, $\ell \in [0, ..., M-1]$, and $\tilde{h}_{\ell,m}$ is the propagation channel between the BS antennas ℓ and m. We introduce the calibration coefficient between BS radios as

$$h_{\ell,m} = b_{m\to\ell}\ h_{m,\ell}, \tag{5}$$

which by assuming perfect reciprocity yields[7]

$$b_{m\to\ell} = \frac{h_{\ell,m}}{h_{m,\ell}} = \frac{r_\ell^{\mathrm{B}} t_m^{\mathrm{B}}}{r_m^{\mathrm{B}} t_\ell^{\mathrm{B}}} = \frac{1}{b_{\ell\to m}}. \tag{6}$$

One of the main contributions from [4] was an internal reciprocity calibration method for a massive MIMO base station. The method has two main points as basis:

1. $$b_{m,k} = \frac{t_m^{\mathrm{B}}}{r_m^{\mathrm{B}}} \frac{r_k^{\mathrm{M}}}{t_k^{\mathrm{M}}} = \frac{r_n^{\mathrm{B}}}{r_m^{\mathrm{B}}} \frac{t_m^{\mathrm{B}}}{t_n^{\mathrm{B}}} \frac{r_k^{\mathrm{M}}}{r_n^{\mathrm{B}}} \frac{t_n^{\mathrm{B}}}{t_k^{\mathrm{M}}} = b_{m\to n} b_{n,k}. \tag{7}$$

   i.e., calibration between radios m and k can also be achieved if their forward and reverse channels to another BS radio n are jointly processed. Throughout the paper we set $n = 0$ for convenience and denote this radio as the reference radio.

2. As long as each downlink channel estimate from all BS antennas deviate from the real ones by the same complex factor, the resulting downlink beam pattern shape does not change. Thus, since the transceiver response of any terminal shows up as a constant factor to all BS antennas, its contribution can be omitted from the calibration procedure.

Combining (2) with the previous two points yields

$$g_{k,m}^{\mathrm{D}} = b_{m,k}\, g_{m,k}^{\mathrm{U}} \tag{8}$$

$$\overset{1)}{=} b_{m\to 0}\, b_{0,k}\, g_{m,k}^{\mathrm{U}} \tag{9}$$

$$\overset{2)}{\Leftrightarrow} g_{k,m}^{'\mathrm{D}} = b_{m\to 0} g_{m,k}^{\mathrm{U}} \tag{10}$$

where $g_{k,m}^{'\mathrm{D}}$ is a relative downlink channel that absorbs $b_{0,k}$. Thus relative downlink channels can be obtained by multiplying the respective uplink channels with their respective calibration coefficients to a reference radio. The authors in [5] took this approach one step forward in order to calibrate access points of a distributed MIMO network. A novelty in their approach was

$$g_{k,m}^{'\mathrm{D}} = b_{m\to 0}\, g_{m,k}^{\mathrm{U}} \tag{11}$$

$$\Leftrightarrow g_{k,m}^{''\mathrm{D}} = b_m\, g_{m,k}^{\mathrm{U}} \tag{12}$$

where $b_m = \frac{r_m^B}{t_m^B} = \frac{1}{b_{m\to 0}} \frac{t_0^B}{r_0^B}$, and $g_{k,m}^{''\mathrm{D}}$ is another relative downlink channel. This relative equivalence not relaxes the double-indexing overhead, but allows different calibration coefficients to be treated as mutually independent (!).

Note that the absolute reference to the terminals was lost in the derivation step 2), which makes $b_{m\to 0}$ or b_m valid calibration coefficients up to a complex factor. Thus,

[7] Note that we denote the calibration coefficients between two BS radios using "→" to distinguish from the calibration coefficient between a BS radio and an MS which uses ",".

downlink pilots still need to be broadcast through the beam to compensate for this uncertainty, as well as for the RF chain responses of the terminals. The overhead of these supplementary pilots is reported as very small [2]. Also note that the calibration coefficients are valid over long periods of time (compared to the channel coherence interval) since BS radios share the same synchronization references.

2.3 System Model for BS-BS Signals

As shown in Sec. 2.2, reciprocity calibration can be carried out without the need of any feedback from the MSs. To estimate the calibration coeficients b_m we sound the M antennas one-by-one by transmitting a pilot symbol from each one and receiving on the other $M-1$ silent antennas. For simplicity, we use a pilot symbol $p=1$. Let $y_{m,\ell}$ denote the signal received at antenna m when transmitting at antenna ℓ. It follows that the received signals between any pair of antennas can be written as

$$\begin{aligned}\begin{bmatrix} y_{\ell,m} \\ y_{m,\ell} \end{bmatrix} &= \tilde{h}_{\ell,m} \begin{bmatrix} r_\ell^{\mathrm{B}} t_m^{\mathrm{B}} \\ r_m^{\mathrm{B}} t_\ell^{\mathrm{B}} \end{bmatrix} + \begin{bmatrix} n_{\ell,m} \\ n_{m,\ell} \end{bmatrix} \\ &= \alpha_{\ell,m} \begin{bmatrix} b_\ell \\ b_m \end{bmatrix} + \begin{bmatrix} n_{\ell,m} \\ n_{m,\ell} \end{bmatrix},\end{aligned} \tag{13}$$

where $\alpha_{\ell,m} = t_\ell^B t_m^B \tilde{h}_{\ell,m} = t_\ell^B t_m^B \tilde{h}_{m,\ell}$ due to reciprocity, and $[n_{\ell,m}\ n_{m,\ell}]^T$ is a vector of independent zero-mean circularly symmetric complex Gaussian distributed random variables, each one with variance N_0.

2.4 Statistical Model of BS-BS Channels

We next put forth the statistical models for the channel between antennas that we have used in this work. The channel between two antennas ℓ and m is modeled as

$$\tilde{h}_{\ell,m} = \beta_{\ell,m} \exp(\jmath\phi_{\ell,m}) + w_{\ell,m}, \tag{14}$$

where $\beta_{\ell,m}$ is assumed known and models the channel gain due to antenna coupling, the channel phase $\phi_{\ell,m}$ is uniformly distributed between 0 and 2π, and $w_{\ell,m} \sim \mathcal{CN}(0, N_{\mathrm{w}})$ models multipath propagation with no dominant component.

To model the antenna coupling $\beta_{\ell,m}$, we measured channel gains between $\frac{\lambda}{2}$ spaced antennas of a 25x4 dual polarized antenna array, a custom made massive MIMO antenna array for our testbed [6], in an anechoic chamber. We averaged the frequency response magnitude over a 20 MHz bandwidth centered at 3.7 GHz which the array was originally designed to operate at. Fig. 1 shows the measured results. Only the E-plane orientation field was measured. This explains the difference between measured channel gains for same measured distances since antenna elements oriented in the E-plane orientation are more strongly coupled than others [7].

As a rough estimate a $0.03d^{-3.7}$ curve match our measurements well. This simplified fit will be used in our simulations which allows for reproducible results.

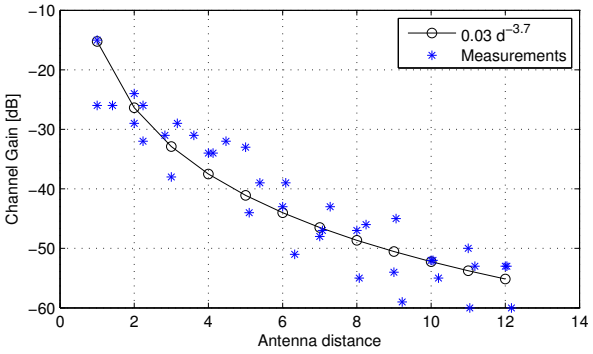

Figure 1: Measured antenna coupling. The horizontal axis represents the antenna spacing in units of $\frac{\lambda}{2}$ between measured antennas.

3 Reciprocity calibration methods

3.1 Direct-path based LS [4]

Here we estimate $\mathbf{b} = [b_0, b_1, ..., b_{M-1}]^T$ solely using the signals $y_{0,m}$ and $y_{m,0}$. Since b_m can be estimated up to a multiplicative constant, we set $b_0 = 1$ with no loss of generality and solve for the remaining $[b_1, ..., b_{M-1}]^T$. A least-squares approach can be pursued which seeks to jointly optimize b_m and $\alpha_{\ell,m}$ according to

$$(\hat{b}_m, \hat{\alpha}) = \arg\min_{b_m,\alpha} \left\| \begin{bmatrix} y_{0,m} \\ y_{m,0} \end{bmatrix} - \alpha_{\ell,m} \begin{bmatrix} 1 \\ b_m \end{bmatrix} \right\|^2. \tag{15}$$

It is easy to verify that the solution to (15) is given by

$$\hat{b}_m = \frac{y_{0,m}}{y_{m,0}} \quad \text{and} \quad \alpha_{\ell,m} = y_{0,m}. \tag{16}$$

Note that this ratio has unbounded second moment.

3.2 Generalized LS [5]

This approach generalizes the *Direct-path based LS* estimator by considering the full set of signals in (13). An LS cost function can be formulated as

$$J(\mathbf{b})_{\text{LS}} = \sum_{m,\ell \neq m} |b_m y_{m,\ell} - b_\ell y_{\ell,m}|^2. \tag{17}$$

To minimize (17) one can set its gradient $\nabla J(\mathbf{b})$ to zero and solve for $\mathbf{b}$. To exclude the trivial solution $\mathbf{b} = \mathbf{0}$, we set $b_0 = 1$ as previously mentioned. This yields

$$\hat{\mathbf{b}} = -\left(\mathbf{A}_1^H \mathbf{A}_1\right)^{-1} \mathbf{A}_1^H \mathbf{a}_1 b_0 \tag{18}$$

where $\mathbf{A} = (\mathbf{a}_1\ \mathbf{A}_1)$ (i.e., $\mathbf{a}_1$ is the first column of $\mathbf{A}$, $\mathbf{A}_1$ is a matrix made of the $M-1$ last columns of $\mathbf{A}$) and $\mathbf{A}$ is structured as

$$A_{m,\ell} = \begin{cases} \sum_{i=1}^{M} |y_{m,\ell}|^2, & m = \ell \\ -y_{m,\ell}^{*}\, y_{\ell,m} & m \neq \ell \end{cases}. \tag{19}$$

3.3 Generalized weighted LS

All sets of double directional measurements are given the same weights in (17). If one still maintains an LS formulation, it is intuitive that the estimator's performance can be improved if any statistical information of $\alpha_{\ell,m} = t_\ell^B t_m^B(\beta_{\ell,m}\exp(\jmath\phi_{\ell,m}) + w_{\ell,m})$ is known. In a practical (massive MIMO) antenna array, knowledge of the coupling gains $\beta_{m,\ell}$ is indeed at hand, see see Sec.2.4. Thus the cost function can be empirically re-defined to

$$J(\mathbf{b})_{\mathrm{WLS}} = \sum_{m,\ell} |\beta_{m,\ell} b_m y_{m,\ell} - \beta_{\ell,m} b_\ell y_{\ell,m}|^2. \tag{20}$$

It can be shown that weighting the cost function with the complex coupling gains $\beta_{m,\ell}\exp(\jmath\phi_{m,\ell})$ yields the same estimator as (20), thus making phase information irrelevant for the current problem formulation.

3.4 Generalized Neighboor LS

In Sec. 3.2 and Sec. 3.3 we addressed performance improvements to (15) by jointly processing $M(M-1)$ signals. In this subsection we investigate if an entire BS antenna array can be accurately calibrated solely based on signals to/from neighbor antennas, thus using less than $4M$ signals for the case of a planar array. The cost function in this case is given by

$$J(\mathbf{b})_{\mathrm{NLS}} = \sum_{m} \sum_{\ell \in \mathcal{A}_m} |b_m y_{m,\ell} - b_\ell y_{\ell,m}|^2. \tag{21}$$

where $\mathcal{A}_m$ is the set of indexes of adjacent antennas to antenna m. Besides the obvious reduced number of multiplications needed to generate $\mathbf{A}_1$, the advantages of such neighbor based calibration are manifold: (i) with proper antenna indexing, the final estimator inversion $(\mathbf{A}_1^H\mathbf{A}_1)^{-1}$ is potentially performed faster since $\mathbf{A}_1^H\mathbf{A}_1$ can be arranged as an L-banded Hermitian matrix with $L \ll M$ [8]; (ii) the received signal power level is approximately the same for all neighbor receiving antennas. This simplifies post-compensation due to hardware adaptations, e.g., automatic gain control (AGC), or non-linear dependencies, e.g., amplifiers; (iii) it allows distant antennas to measure their neighbor channel simultaneously with (almost) no interference, speeding up the calibration process.

3.5 Simulated calibration accuracy

We simulated reciprocity calibration for the case of a 5x20 planar patch array. We used the antenna coupling loss model established in Sec. 2.4 and set the variance of

the channel Rayleigh component to $N_{\mathrm{w}} = -50$ dB. One of the center antenna elements of the array was defined as the reference. For the general case, modeling the statistics of RF chains responses is a hard task, thus we follow the same approach as [5], where both transmitter and receiver (i.e., t_m^{B} and r_m^{B}) have uniformly distributed phase between $[-\pi, \pi[$ and uniformly distributed magnitude between $[1-\epsilon, 1+\epsilon]$ with ϵ such that $\sqrt{\mathbb{E}\{(|t_m^{\mathrm{B}}|-1)^2\}} = \sqrt{\mathbb{E}\{(|r_m^{\mathrm{B}}|-1)^2\}} = 0.1$.

We focus on the distinct cases of neighbor antennas and furthest away antennas from the reference one. The latter are positioned at the array edges where coupling to the reference is practically null, thus being the the hardest calibration case. Results for others antennas should, in principle, fall within these bounds.

For all approaches, we choose to normalize all results with respect to the (calibration) signal-to-noise ratio $\mathrm{SNR}_{\mathrm{Cal}}$ of the neighbor antenna channel. With this normalization it is straightforward to see how different calibration methods "close the gap" between the best and worst calibration scenarios.

At low $\mathrm{SNR}_{\mathrm{Cal}}$ values, its visible from Fig. 2 that the direct-path (DP) based estimator do not possess finite second moment, i.e., the simulated MSE do not converge as the number of simulation runs increases. As for the generalized estimators, the LS estimator (Sec. 3.2) shows the worst performance at low $\mathrm{SNR}_{\mathrm{Cal}}$. This is justified by the weak received signals being equally weighted in the cost function. The weighted LS estimator (Sec. 3.3) compensates for this, but has worst performance at high $\mathrm{SNR}_{\mathrm{Cal}}$ (by a small margin) since weights are not optimized in an MSE sense. Overall, the neighbor LS (Sec. 3.4) scheme works fairly well.

A rough estimate of the calibration $\mathrm{SNR}_{\mathrm{Cal}}$ regime where a massive MIMO basestation as our testbed [6] operates is given by

$$\mathrm{SNR}_{\mathrm{cal}} = P_{RX} - N \approx 80\mathrm{dB}, \tag{22}$$

where $P_{RX} = -15$ dBm is the maximum allowed receive power per RF-chain, $N = 10\log_{10}(kBT_0) + N_F + G \approx -95$ dBm is the receiver noise power, k is Boltzmann's constant, $B = 20$ MHz is the channel bandwidth, $T_0 = 290K$ is the standardized room temperature, $N_F = 6$ dB is the noise figure of the receiver chain, and $G = 0$ dB is a normalized amplifier gain. In practice, hardware limitations as ADC resolution and frequency harmonics will degrade the calibration performance. However, a margin of tens of dBs is still available to compensate for such impairments while still achieving acceptable performance for the applications we target, as will be discussed in further detail in Sec. 4.

4 Performance analysis of a reciprocity calibrated massive MIMO system

In this section we verify the impact of the reciprocity calibration error on the capacity/sum-rate of a massive MIMO downlink transmission with perfect (uplink) channel state information (CSI). We generated the set of calibration signals

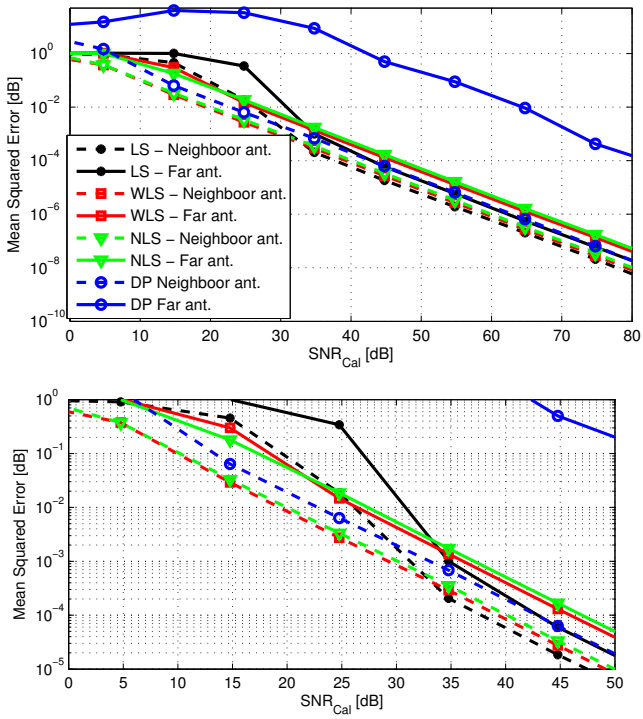

Figure 2: Mean squared error (MSE) of the calibration coefficients computed for the neighbor and the furthermost antenna from the reference.

$[y_{\ell,m}\ y_{m,\ell}]^T$ according to Sec. 3.5 and used the neighbor based calibration approach (i.e., see Sec. 3.4) to estimate the calibration coefficients.

The BS is equipped with $M = \{100, 400\}$ antennas and serves $K = 10$ single antenna mobile users in the same time/frequency resource. The composite received symbol vector at the user side for the case of a narrow-band MIMO channel is described as

$$\boldsymbol{y} = \sqrt{\frac{\rho}{K}}\boldsymbol{H}\boldsymbol{s} + \boldsymbol{n}, \tag{23}$$

where $\boldsymbol{H}$ and $\boldsymbol{n}$ are the K x M channel matrix and the K x 1 noise vector, respectively, with i.i.d. unit-norm zero-mean circularly symmetric complex Gaussian distributed random elements, $\boldsymbol{s} = f(\boldsymbol{x})$ subject to $\mathbb{E}\{||\boldsymbol{s}||^2\} = 1$ is the transmit precoded version of $\boldsymbol{x}$ with calibration errors, and ρ/K is the transmit power.

Fig. 3 shows the calibration error-free capacities/sum-rates of three precoders, i.e., maximum-ratio transmission (MRT), zero-forcing (ZF) and dirty paper coding (DPC)

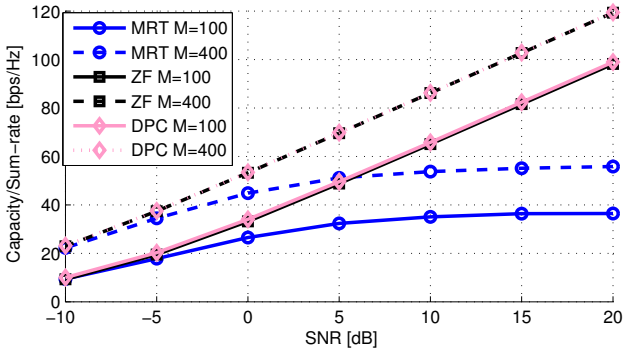

Figure 3: Capacity/sum-rates for different precoders in i.i.d. massive MIMO downlink channels.

scheme. In the high SNR regime, inter-user interference with total power I, upper-bounds the MRC precoder sum-rate to $C_{\text{MRT}}^{\text{UP}} = K \log_2(1 + \frac{M\rho}{KI})$, while the ZF sum-rate and DPC capacity converge to the interference-free case $C_{\text{IF}} = K \log_2(1 + \frac{M\rho}{KN})$.

Fig. 4 shows the downlink sum-rate loss due to reciprocity calibration errors for MRT and ZF precoders using the neighbor LS estimator to compute the calibration coefficients. For a given precoder, the sum-rate loss was obtained by normalizing the obtained sum-rates by their respective error-free ones (Fig. 3). Overall, the sum-rate loss using the MRT precoding scheme shows to be more robust to calibration errors compared to the ZF case: (i) less calibration SNR_{Cal} is needed to achieve similar capacity losses, (ii) capacity losses are less sensitive to the current communications SNR. Significant capacity losses happen for SNR_{Cal} values smaller than 35dB and 20dB for the ZF and MRT precoders, respectively. These SNR_{Cal} values provide reference levels for achieving "good enough" calibration performance. Noticeably, the extended estimators introduced in Sec. 3.3 and Sec. 3.4, improve the calibration performance within $0\text{dB} < \text{SNR}_{\text{Cal}} < 35\text{dB}$ compared to current state-of-art methods [5], where significant capacity losses occur due to calibration errors, see Fig. 2. Note that, for the considered channel model between BS antennas, calibration accuracy is reduced as the number of BS antennas M grows.

5 Conclusions

In this paper we extended a reciprocity calibration framework which was originally developed for calibrating the access points of a distributed MIMO system, in order to calibrate a massive MIMO BS antenna array.

Inter-BS antenna channels exhibit strong deterministic characteristics which can be incorporated in the calibration model to enhance performance. The performance of the studied estimators indicates that is possible to calibrate an entire massive

MIMO BS antenna array using antenna coupling as the main propagation mechanism. From the specifications of a massive MIMO base station testbed as [6], we verified a calibration accuracy margin of tens of dBs better than a calibration accuracy leading to significant capacity losses. The downlink capacity loss of a massive MIMO system using the MRT precoding scheme was shown to be more robust to calibration errors compared to the ZF case.

Acknowledgments

The work has been funded by grants from the Swedish foundation for strategic research SSF, the Swedish research council and the Excellence center at Linköping - Lund in Information Technology.

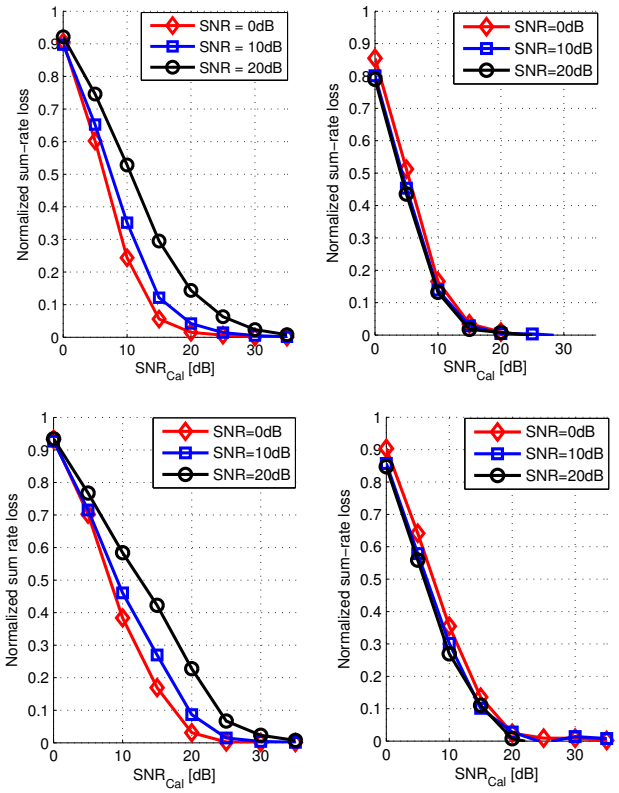

Figure 4: Sum-rate loss due to calibration errors for different precoders using the neighbor LS calibration method at different communications SNRs. Up) M=100; Down) M=400 left) Zero-forcing precoder; right) Maximum ratio transmission.

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[3] F. Kaltenberger, H. Jiang, M. Guillaud, and R. Knopp, "Relative channel reciprocity calibration in MIMO/TDD systems," in *Future Network and Mobile Summit, 2010*, June 2010, pp. 1–10.

[4] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 53–64.

[5] R. Rogalin, O. Bursalioglu, H. Papadopoulos, G. Caire, A. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *submitted to Wireless Communications, IEEE Transactions on*, vol. PP, no. 99, pp. 1–17, 2014. [Online]. Available: http://arxiv.org/abs/1310.7001

[6] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong and V. Öwall and O. Edfors and F. Tufvesson, "A flexible 100-antenna testbed for Massive MIMO," in *IEEE GLOBECOM 2014 Workshop on Massive MIMO: from theory to practice, 2014-12-08*. IEEE, 2014.

[7] R. Jedlicka, M. Poe, and K. Carver, "Measured mutual coupling between microstrip antennas," *Antennas and Propagation, IEEE Transactions on*, vol. 29, no. 1, pp. 147–149, Jan 1981.

[8] A. Asif and J. Moura, "Fast inversion of L-block banded matrices and their inverses," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, May 2002, pp. II–1369–II–1372.

Paper IV

Reciprocity Calibration for Massive MIMO: Proposal, Modeling and Validation

This paper presents a mutual coupling based calibration method for time-division-duplex massive MIMO systems, which enables downlink precoding based on uplink channel estimates. The entire calibration procedure is carried out solely at the base station (BS) side by sounding all BS antenna pairs. An Expectation-Maximization (EM) algorithm is derived, which processes the measured channels in order to estimate calibration coefficients. The EM algorithm outperforms current state-of-the-art narrow-band calibration schemes in a mean squared error (MSE) and sum-rate capacity sense. Like its predecessors, the EM algorithm is general in the sense that it is not only suitable to calibrate a co-located massive MIMO BS, but also very suitable for calibrating multiple BSs in distributed MIMO systems. The proposed method is validated with experimental evidence obtained from a massive MIMO testbed. In addition, we address the estimated narrow-band calibration coefficients as a stochastic process across frequency, and study the subspace of this process based on measurement data. With the insights of this study, we propose an estimator which exploits the structure of the process in order to reduce the calibration error across frequency. A model for the calibration error is also proposed based on the asymptotic properties of the estimator, and is validated with measurement results.

1 Introduction

Massive Multiple-input Multiple-output (massive MIMO) is an emerging technology with the potential to be included in next generation wireless systems, such as fifth-generation (5G) cellular systems. Massive MIMO departs from traditional multi-user MIMO approaches by operating with a large number of base station (BS) antennas, typically in the order of hundreds or even thousands, to serve a relatively small number of mobile terminals [1]. Such a system setup results in a multitude of BS antennas that can be used in an advantageous manner from multiple points of view [2].

One major challenge of operating with a large number BS antennas is that it renders explicit channel estimation in the downlink impractical. Basically, the overhead of channel estimation in the downlink and feeding back the channel estimate to the BS, scales linearly with the number of BS antennas, and quickly becomes unsupportable in mobile time-varying channels [3]. To deal with this challenge, the approach adopted is to operate in time-division-duplex (TDD) mode, rely on channel reciprocity, and use uplink channel state information (CSI) for downlink precoding purposes [4]. However, the presence of the analog front-end circuitry in practical radio units complicates the situation and makes the baseband-to-baseband channel non-reciprocal. Explained briefly, the baseband representation of the received signals [5] experience channels that are not only determined by the propagation conditions, but also by the transceiver front-ends at both sides of the radio link. While it is generally agreed that the propagation channel is reciprocal [6], the transceiver radio frequency (RF) chains at both ends of the link are generally not [7]. Hence, in order to make use of the reciprocity assumption and rely on the uplink CSI to compute precoding coefficients, the non-reciprocal transceiver responses need to be calibrated. Such a procedure is often termed reciprocity calibration, and contains two steps: (i) estimation of calibration coefficients, and (ii) compensation by applying those to the uplink channel estimates.[8]

Reciprocity calibration of small scale TDD MIMO channels has been a matter of study in recent years. Depending on the system setup and requirements, the approach adopted can take many forms. For example, [7] proposed a methodology based on bi-directional measurements between the two ends of a MIMO link to estimate suitable reciprocity calibration coefficients. This calibration approach falls in the class of "over-the-air" calibration schemes where users are involved in the calibration process. A different approach is to rely on dedicated hardware circuitry for calibration purposes, see [8,9]. Despite the possibilities of extending both mentioned calibration approaches to a massive MIMO context, e.g., [10,11], recent calibration works suggest this is more difficult than previously thought. For example, [12] questions the feasibility of having dedicated circuits for calibration when the number of transceivers to be calibrated grows large, and [13] argues that the calibration protocols should preferably not rely on mobile units. It thus appears that an increasing trend in massive

[8] However, with the term reciprocity calibration, we will interchangeably refer to the estimation step, compensation step, or both. The context will, hopefully, make clear which of the previous cases is being addressed.

MIMO systems is to carry out the calibration entirely at the BS side only through over-the-air measurements.

The first proposal in this vein was presented in [14]. The work proposes an estimator for the calibration coefficients, which only makes use of channel measurements between BS antennas. More specifically, bi-directional channel measurements between a given BS antenna, so-called reference antenna, and all other antennas. This estimator was later generalized in order to calibrate large-scale distributed MIMO networks [13, 15]. The estimation problem is formulated as constrained least-squares (LS) problem where the objective function uses channel measurements from a set of arbitrary antenna pairs of the network. The generality of this approach spurred many publications dealing with particular cases [16–18]. Parallel work in mutual coupling based calibration was also conducted in [12]. An estimator for the calibration coefficients, which enables maximum ratio transmission (MRT), was proposed for BS antenna arrays with special properties.

Although it appears that over-the-air reciprocity calibration only involving the BS side is feasible, some matters need further investigation. Firstly, the approaches available in the literature for co-located BSs are not of great practical convenience. They either rely on antenna elements that need to be (carefully) placed in front of the BS antenna array solely for calibration purposes [14], or are only available for a restrictive case of antenna arrays [12]. Secondly, most estimators for calibration have been derived from empirical standpoints, e.g., [12, 14], and respective extensions [15, 17, 18]. It is not clear how far from fundamental estimation performance bounds, or how close to Maximum likelihood (ML) performance, such estimators are. Thirdly, most available calibration approaches are proposed for narrow-band systems. Such systems bandwidths are usually defined by the frequency selectivity of the propagation channel, which is typically much smaller than the frequency selectivity of the transceiver responses. This results in similar calibration coefficients for adjacent narrowband channels. Thus, it is of interest to model the statistical dependency of such calibration coefficients, and provide means to exploit this dependency in order to reduce the calibration error across frequency. Lastly, there is little publicly available work on validation of massive MIMO calibration schemes. The need for validation is high, as it helps answering many questions of practical nature. For example, [19] raises the question whether the channel reciprocity assumption holds when strong coupling between BS antennas exist, and [20] questions if calibration assumptions similar to the ones used in this work, hold for massive MIMO arrays.

1.1 Main Contributions of the Paper

Below, we summarize the main contributions of this work.

- We propose a convenient calibration method mainly relying on mutual coupling between BS antennas to calibrate its non-reciprocal analog front-ends. We make no assumptions other than channels due to mutual coupling being reciprocal.
- We show that the narrow-band calibration coefficients can be estimated by solving a joint penalized-ML estimation problem. We provide an asymptotically

efficient algorithm to compute the joint solution, which is a particular case of the EM algorithm.

- We validate our calibration method experimentally using a software-defined radio massive MIMO testbed. More specifically, we verify how the measured Error-Vector-Magnitude (EVM) of the downlink equalized signals decreases as the calibration accuracy increases, in a setup where three closely spaced single-antenna users are spatially multiplexed by one hundred BS antennas.
- We propose a non-white Gaussian model for the narrow-band calibration error based on the properties of the proposed estimator, and partially validate this model with measurements.

1.2 Notation

The operators $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^\dagger$ denote element-wise complex conjugate, transpose, Hermitian transpose, and Moore-Penrose pseudo-inverse, respectively. The element in the nth row and mth column of matrix $\mathbf{A}$ is denoted by $\left[\mathbf{A}\right]_{n,m}$. The operator $\mathrm{E}\{\cdot\}$ denotes the expected value. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ return the real and imaginary part of their arguments. The matrix $\mathbf{I}$ denotes the identity matrix, and $\mathrm{diag}\{a_1, a_2, \ldots a_M\}$ denotes an $M \times M$ diagonal matrix with diagonal entries given by $a_1, a_2, \ldots, a_M$. The operator ln denotes the natural logarithm. The set of the complex numbers and the set containing zero and the real positive numbers are denoted by $\mathbb{C}$ and $\mathbb{R}_{\geq 0}$, respectively. The operator $\setminus$ denotes the relative set complement. Finally, $||\cdot||$ denotes the Frobenius norm.

1.3 Paper Outline

The remaining sections of the paper are as follows. Section 2 presents the signal models. Section 3 introduces the state-of-the-art estimator for the calibration coefficients, proposes a novel estimator, and provides a comparative analysis by means of MSE and downlink sum-rate capacities. Section 4 validates the proposed calibration method experimentally. Using the estimated calibration coefficients obtained from the experiments, the purpose of Section 5 is twofold: *i)* it studies several aspects of the calibration coefficients across 4.5 MHz of transceiver bandwidth, *ii)* it proposes a model for the calibration error of a narrowband system. Lastly, Section 5 summarizes the key takeaways from this work.

2 Signal Models

This section starts by introducing the uplink and downlink signal models, and shows how downlink precoding can be performed using calibrated uplink channel estimates. Finally, it models the channels between BS antennas which we use for calibration purposes.

2.1 Uplink and Downlink Signal models

Let K single-antenna users simultaneously transmit a pilot symbol in the uplink of a narrow-band MIMO system (e.g., a particular sub-carrier of an OFDM-MIMO system). Collecting the pilot symbols in the vector $\mathbf{p} = [p_1 \cdots p_K]^T$, the received signal by an M-antenna base station can be written as

$$\begin{aligned}\mathbf{y}_{\mathrm{UP}} &= \mathbf{H}_{\mathrm{UP}}\,\mathbf{p} + \mathbf{w} \\ &= \mathbf{R}_{\mathrm{B}}\mathbf{H}_{\mathrm{P}}\mathbf{T}_{\mathrm{U}}\,\mathbf{p} + \mathbf{w}. \end{aligned} \tag{1}$$

In (1), the matrix $\mathbf{R}_{\mathrm{B}} = \mathrm{diag}\left\{r_1^{\mathrm{B}}, \cdots, r_M^{\mathrm{B}}\right\}$ models the hardware response of M BS receive RF chains (one RF chain per antenna), and the matrix $\mathbf{T}_{\mathrm{U}} = \mathrm{diag}\left\{t_1^{\mathrm{U}}, \cdots, t_K^{\mathrm{U}}\right\}$ models the hardware response of K transmit RF chains (one chain per user). $\mathbf{H}_{\mathrm{P}}$ is the propagation channel matrix, $\mathbf{H}_{\mathrm{UP}}$ is the, so-called, uplink radio channel, and $\mathbf{w}$ is a vector modeling uplink noise. Under the reciprocal assumption of the propagation channel, the received downlink signal can be written as

$$\begin{aligned}\mathbf{y}_{\mathrm{DL}} =& \mathbf{H}_{\mathrm{DL}}\,\mathbf{z}' + \mathbf{w}' \\ =& \mathbf{R}_{\mathrm{U}}\mathbf{H}_{\mathrm{P}}^T\mathbf{T}_{\mathrm{B}}\,\mathbf{z}' + \mathbf{w}'. \end{aligned} \tag{2}$$

In (2), the matrix $\mathbf{R}_{\mathrm{U}} = \mathrm{diag}\left\{r_1^{\mathrm{U}}, \cdots, r_K^{\mathrm{U}}\right\}$ models the hardware response of the receive RF chains of the K users, and the matrix $\mathbf{T}_{\mathrm{B}} = \mathrm{diag}\left\{t_1^{\mathrm{B}}, \cdots, t_M^{\mathrm{B}}\right\}$ models the hardware response of M BS transmit RF chains. The entries of $\mathbf{w}'$ model downlink noise, $\mathbf{H}_{\mathrm{DL}}$ is the downlink radio channel, and $\mathbf{z}'$ is a vector with linearly precoded QAM symbols. In particular, $\mathbf{z}' = \mathbf{P}\mathbf{x}$, where $\mathbf{P}$ is the precoding matrix, and the entries of $\mathbf{x}$ contain QAM symbols.

2.2 Calibration Coefficients

Assume that an error free version of the uplink radio channel, $\mathbf{H}_{\mathrm{UP}}$, is available at the BS. The transpose of the result of pre-multiplying $\mathbf{H}_{\mathrm{UP}}$ with the matrix $\alpha\mathbf{T}_{\mathrm{B}}\mathbf{R}_{\mathrm{B}}^{-1}$, where $\alpha \in \mathbb{C} \setminus 0$ and $r_m \neq 0, \forall\, m$, is a matrix $\mathbf{G}$ that, if used for precoding purposes by means of a linear filtering, is sufficient for spatially multiplexing terminals in the downlink with reduced crosstalk. This can be visualized by expanding $\mathbf{G}$ as

$$\begin{aligned}\mathbf{G} &= \left(\left(\alpha\mathbf{T}_{\mathrm{B}}\mathbf{R}_{\mathrm{B}}^{-1}\right)\mathbf{H}_{\mathrm{UP}}\right)^T \\ &= \alpha\mathbf{T}_{\mathrm{U}}\mathbf{H}_{\mathrm{P}}^T\mathbf{T}_{\mathrm{B}} \\ &= \alpha\mathbf{T}_{\mathrm{U}}\mathbf{R}_{\mathrm{U}}^{-1}\mathbf{H}_{\mathrm{DL}}. \end{aligned} \tag{3}$$

From (3) we have that $\mathbf{G}$ is effectively the *true* downlink radio channel $\mathbf{H}_{\mathrm{DL}}$ pre-multiplied with a diagonal matrix with unknown entries accounting for the user terminals responses $\mathbf{T}_{\mathrm{U}}\mathbf{R}_{\mathrm{U}}^{-1}$, and α. The row space of $\mathbf{G}$ is thus the same as of the downlink radio channel $\mathbf{H}_{\mathrm{DL}}$. This is a sufficient condition to cancel inter-user interference if, for example, ZF precoding is used (i.e., $\mathbf{H}_{\mathrm{DL}}\mathbf{G}^{\dagger}$ is a diagonal matrix).

From (3), it can also be seen that any non-zero complex scalar α provides equally good calibration.[9] Thus, the matrix

$$\begin{aligned}\mathbf{C} =& \text{diag}\{\text{c}_1, \cdots, \text{c}_\text{M}\} \\ =& \mathbf{T}_\text{B}\mathbf{R}_\text{B}^{-1}\end{aligned} \tag{4}$$

is the, so-called, calibration matrix, and $\{c_m\}$ are the calibration coefficients which can be estimated up to a common complex scalar α. We remark that, although not strictly necessary to build estimators, the concept of a reference transceiver [14] can be used to deal with the ambiguity of estimating $\{c_m\}$ up to α.[10] The remainder of the paper deals with estimation aspects of $c_m = t_m^\text{B}/r_m^\text{B}$. Thus, for notational simplicity, we write $t_m = t_m^\text{B}$, $r_m = r_m^\text{B}$, $\mathbf{R} = \mathbf{R}_\text{B}$, and $\mathbf{T} = \mathbf{T}_\text{B}$. Also, we stack $\{c_m\}$ in the vector $\mathbf{c} = [c_1 \cdots c_M]^T$, for later use.

2.3 Inter-BS Antennas Signal model

To estimate the calibration coefficients c_m we sound the M antennas one-by-one by transmitting a sounding signal from each one and receiving on the other $M-1$ silent antennas. Let the sounding signal transmitted by antenna m be $s_m = 1, \forall\, m$, unless explicitly said otherwise. Also, let $y_{n,m}$ denote the signal received at antenna n when transmitting at antenna m. It follows that the received signals between any pair of antennas can be written as

$$\begin{bmatrix} y_{n,m} \\ y_{m,n} \end{bmatrix} = h_{n,m} \begin{bmatrix} r_n t_m & 0 \\ 0 & r_m t_n \end{bmatrix} \begin{bmatrix} s_m \\ s_n \end{bmatrix} + \begin{bmatrix} n_{n,m} \\ n_{m,n} \end{bmatrix}, \tag{5}$$

where

$$h_{n,m} = \bar{h}_{n,m} + \tilde{h}_{n,m} \tag{6}$$

$$= |\bar{h}_{n,m}| \exp(j2\pi\phi_{n,m}) + \tilde{h}_{n,m} \tag{7}$$

models the (reciprocal) channels between BS antennas. The first term $\bar{h}_{n,m}$ describes a channel component due to mutual coupling between antenna elements, often stronger for closely spaced antennas, which we lay down a model for in Sec. 2.4. The terms $|\bar{h}_{n,m}|$ and $\phi_{n,m}$ denote the magnitude and phase of $\bar{h}_{n,m}$, respectively. The term $\tilde{h}_{n,m}$, which absorbs all other channel multipath contributions except for the mutual coupling (e.g., reflections by scatterers in front of the BS) is modeled by an i.i.d. zero-mean circularly symmetric complex Gaussian random variable with variance σ^2. Non-reciprocal channel components are modeled by r_m and t_m which

[9] This follows since both magnitude and phase of α are not relevant in this calibration setup. The former holds since any real scaled channel estimate provides the same precoder matrix $\mathbf{P}$, if the precoder has a fixed norm. The latter follows from (3), since the (uniform phases of the) diagonal entries of $\mathbf{T}_\text{U}\mathbf{R}_\text{U}^{-1}$ are unknown to the precoder in this calibration setup.

[10] Explained briefly, assuming $c_{ref} = 1$ and solving for $\{c_m\} \backslash c_{ref}$, where c_{ref} is the calibration coefficient associated with a reference transceiver.

materially map to the cascade of hardware components, mainly in the analog front-end stage of the receiver and transmitter, respectively. We assume i.i.d. circularly symmetric zero-mean complex Gaussian noise contributions $n_{m,n}$ with variance N_0. Letting $\left[\mathbf{Y}\right]_{m,n} = y_{m,n}$, the received signals can be expressed more compactly as

$$\mathbf{Y} = \mathbf{RHT} + \mathbf{N}. \tag{8}$$

Note that $\mathbf{H} = \mathbf{H}^T$ is assumed, and the diagonal entries in the $M \times M$ matrix $\mathbf{Y}$ are undefined.

2.4 Modeling Mutual Coupling

The purpose of this section is to provide a model for the mutual coupling between antenna elements, i.e. $\bar{h}_{m,n}$, as a function of their distance. Instead of pursuing a circuit theory based approach to model the effect of mutual coupling [19], our modeling approach uses S-parameter measurements from a massive MIMO BS antenna array [21]. We note that this model is used only for simulation purposes, and not to derive any of the upcoming estimators of $\mathbf{c}$.

Test Array Description

The antenna array considered for modeling is a 2-dimensional planar structure with dual-polarized patch elements spaced by half a wavelength. More information about the antenna array can be found in [22]. The dimensional layout of the array adopted for this work corresponds to the 4×25 rectangular grid in the upper part of the array shown in Fig. 1. Only one antenna port is used per antenna element. For a given antenna, the polarization port is chosen such that its adjacent antennas - the antennas spaced by half wavelength - are cross-polarized. This setting provides, so-called, polarization diversity, and reduces mutual coupling effects between adjacent antennas since co-polarized antennas couple stronger [21].

Modeling coupling gains between antennas

The channel magnitude $|\bar{h}_{n,m}|$ between several pairs of cross and co-polarized antennas were measured in an anechoic chamber using a Vector Network Analyzer, at 3.7 GHz - the center frequency of the array. Fig. 2 shows the measured channel magnitudes. Different channel magnitudes for the very same measured distance and polarization cases, are due mostly to the relative orientation of the antenna pair with respect to their polarization setup. For example, vertically (co-)polarized antennas couple more strongly when they are oriented horizontally. A linear LS fit was performed to model the coupling gain $|\bar{h}_{n,m}|$ as a function of antenna distance. The phase $\phi_{m,n} = \phi_{n,m}$ is modeled uniformly in $[0, 1]$, as a clear dependence with distance was not found.

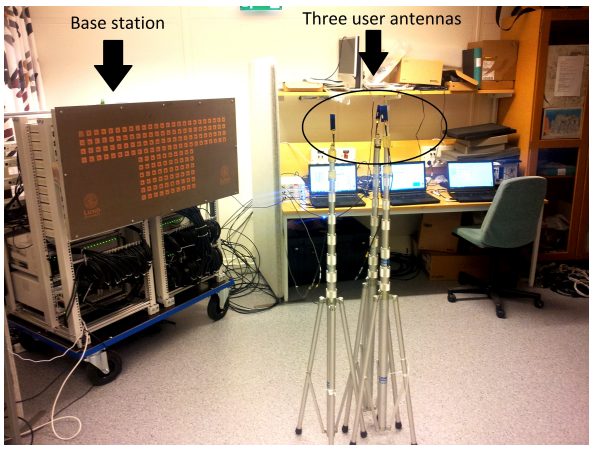

Figure 1: The massive MIMO lab setup used throughout this work. The BS is on the left side where a "T" shaped antenna array can be seen. Three closely spaced user antennas stand the middle of the picture.

3 Estimation of the Calibration Coefficients

In this section we deal with estimation aspects of the calibration matrix $\mathbf{C} = \mathbf{T}\mathbf{R}^{-1}$. We introduce the state-of-art estimator of $\mathbf{C}$ [13, 15], and propose a novel iterative penalized-ML estimator.[11] A comparative numerical analysis is made by means of MSE and sum-rate capacity. We conclude the section with two interesting remarks.

3.1 The Generalized Method of Moments estimator

Calibration of large-scale distributed MIMO systems using a similar system model to (8) was performed in [13] and [15].[12] Based on the structure of the system model, the authors identified that

$$\mathrm{E}\left\{y_{n,m}c_n - y_{m,n}c_m\right\} = 0. \tag{9}$$

Define $g_{m,n} \triangleq y_{n,m}c_n - y_{m,n}c_m$, and $\boldsymbol{g}(\mathbf{c}) = [g_{1,2} \dots g_{1,M}\, g_{2,3} \dots g_{2,M} \dots g_{M-1,M}]^T$.[13] An estimator for $\mathbf{c}$ was proposed by solving

$$\hat{\boldsymbol{c}}_{\mathrm{GMM}} = \arg \min_{\substack{\mathbf{c} \\ s.t.\ f_c(\mathbf{c})=1}} \boldsymbol{g}^H(\mathbf{c})\boldsymbol{W}\boldsymbol{g}(\mathbf{c}) \tag{10}$$

[11] We note that the only assumption used to derive the estimators is $\mathbf{H} = \mathbf{H}^T$. The generality of this assumption allows the estimators to be used in other calibration setups than those of co-located MIMO systems, as it will be pointed out later. [12] In their work, $h_{m,n}$ denotes the propagation channel between antennas of different BSs. The reciprocal model adopted for $h_{m,n}$ accounts for large-scale and small-scale fading. [13] The dependency of $\boldsymbol{g}(\mathbf{c})$ on $y_{n,m}$ is explicitly left out, for notational convenience.

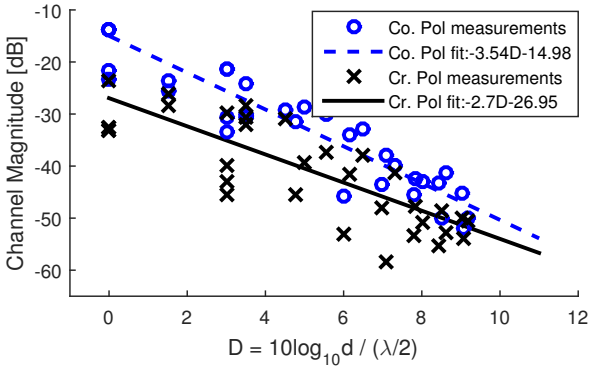

Figure 2: Measured coupling magnitudes $|\bar{h}_{n,m}|$ between different antenna pairs. The circles corresponds to measurements between co-polarized antenna elements, and the crosses between cross polarized antenna elements. The variable d corresponds to the physical distance between antenna elements. The straight lines represent the corresponding linear LS fits.

with $\mathbf{W} = \mathbf{I}$. Two constraints were suggested to avoid the all-zero solution, namely $f_c(\boldsymbol{c}) = c_1$ or $f_c(\boldsymbol{c}) = ||\mathbf{c}||^2$. By setting the gradient with respect to $\mathbf{c}$ to zero, an estimator in closed-form was given. Next, we provide a few remarks on this estimation approach.

A fact not identified in [13] and [15], is that this estimator is an instance of a estimation framework widely used for statistical inference in econometrics, namely the generalized method of moments (GMM). The variable $g_{m,n}$ - whose expectation is zero - is termed a moment condition within GMM literature [23]. With a proper setting of the weighting matrix $\mathbf{W}$, it can be shown that the solution to (4) provides an estimator that is asymptotically efficient [23]. However, no such claim can be made in the low signal-to-noise (SNR) regime, where an optimal form of $\mathbf{W}$ is not available in the literature. This typically leads to empirical settings of $\mathbf{W}$, e.g., $\mathbf{W} = \mathbf{I}$. As a result, moment conditions comprising measurements with low SNR constrain the performance since they are weighted equally. It thus appears that an inherent problem of the GMM estimator is the selection of $\mathbf{W}$. Nevertheless, it provides a closed-form estimator based on a cost function where nuisance parameters for calibration, as $h_{m,n}$, are conveniently left out.

3.2 Joint Maximum Penalized-Likelihood estimation

Here we address joint maximum penalized-likelihood estimation for $\mathbf{c}$ and for the equivalent channel $\mathbf{\Psi} \triangleq \mathbf{RHR}$. Noting that (8) can be written as

$$\begin{aligned}\mathbf{Y} &= \mathbf{RHRC} + \mathbf{N} \\ &= \mathbf{\Psi C} + \mathbf{N},\end{aligned} \tag{11}$$

the optimization problem can be put as

$$\begin{aligned}[\hat{\mathbf{c}},\ \hat{\mathbf{\Psi}}] &= \arg\max_{\mathbf{c},\mathbf{\Psi}} \ln p(\mathbf{Y}|\mathbf{C},\mathbf{\Psi}) + \mathrm{Pen}(\mathbf{C},\mathbf{\Psi},\epsilon') \\ &= \arg\min_{\mathbf{c},\mathbf{\Psi}} J_{\mathrm{ML}}(\mathbf{Y},\mathbf{C},\mathbf{\Psi},\epsilon)\end{aligned} \tag{12}$$

with $J_{\mathrm{ML}}(\mathbf{Y},\mathbf{C},\mathbf{\Psi},\epsilon) = ||\mathbf{Y} - \mathbf{\Psi C}||^2 + \mathrm{Pen}(\mathbf{C},\mathbf{\Psi},\epsilon)$. Here, $p(\mathbf{Y}|\mathbf{C},\mathbf{\Psi})$ denotes the probability density function (PDF) of $\mathbf{Y}$ conditioned on $\mathbf{C}$ and $\mathbf{\Psi}$, and $\mathrm{Pen}(\mathbf{C},\mathbf{\Psi},\epsilon)$ is a penalty term parametrized by $\epsilon = \epsilon' N_0$ with $\epsilon \in \mathbb{R}_{\geq 0}$.

There are many uses for the penalty term in ML formulations [24]. Here, we use it mainly to control the convergence rate of the algorithm (presented in Sec. 3.3), and use ϵ as a tuning parameter. With this in mind, we pursue Ridge Regression and set the penalty term as[14]

$$\mathrm{Pen}(\mathbf{C},\mathbf{\Psi},\epsilon) = \epsilon(||\mathbf{C}||^2 + ||\mathbf{\Psi}||^2). \tag{13}$$

After some re-modeling, a vectorized version of (11) can be written as

$$\widetilde{\mathbf{Y}} = \mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}})\mathbf{c} + \widetilde{\mathbf{N}}, \tag{14}$$

or as

$$\mathbf{Y}' = \mathbf{C}_{\mathrm{eq}}(\mathbf{c})\tilde{\mathbf{\Psi}} + \mathbf{N}', \tag{15}$$

where $\tilde{\mathbf{\Psi}}$ stacks all $\psi_{n,m} = [\mathbf{\Psi}]_{n,m}$ into an $(M^2 - M)/2 \times 1$ vector, and $\mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}})$ and $\mathbf{C}_{\mathrm{eq}}(\mathbf{c})$ are equivalent observation matrices which are constructed from $\tilde{\mathbf{\Psi}}$ and $\mathbf{c}$, respectively. The structure of these matrices is shown in Appendix A, but it can be pointed out that $\mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}})$ and $\mathbf{C}_{\mathrm{eq}}(\mathbf{c})$ are a block diagonal, where each block is a column vector.

From (15), it is seen that for a given $\mathbf{C}_{\mathrm{eq}}(\mathbf{c})$, the penalized-ML estimator of $\tilde{\mathbf{\Psi}}$ is given by[15]

$$\tilde{\mathbf{\Psi}}_{\mathrm{ML}} = \left(\mathbf{C}_{\mathrm{eq}}^H(\mathbf{c})\mathbf{C}_{\mathrm{eq}}(\mathbf{c}) + 2\epsilon\mathbf{I}\right)^{-1}\mathbf{C}_{\mathrm{eq}}^H(\mathbf{c})\mathbf{Y}', \tag{16}$$

[14] Ridge Regression [25] is an empirical regression approach widely used in many practical fields, e.g., Machine Learning [24], as it provides estimation robustness when the model is subject to a number of degeneracies. This turns out to the case in this work, and we point out why this occurs later. However, we emphasize that the main reason of adding the penalty terms is to control the convergence of the algorithm, which we also point out later why this is the case. To finalize, we parametrize the penalty term (13) with a single parameter in order simplify the convergence analysis and be able to extract meaningful insights. [15] The factor 2 in the regularization term of (16) appears since $\psi_{m,n} = \psi_{n,m}$. Note that ϵ is considered as a constant during the optimization, otherwise it is obvious that $\epsilon = 0$ minimizes (13).

If in (15), we replace $\tilde{\mathbf{\Psi}}$ by its estimate $\tilde{\mathbf{\Psi}}_{\mathrm{ML}}$, then the penalized ML solution for $\mathbf{c}$ is

$$\hat{\mathbf{c}}_{\mathrm{ML}} = \arg\min_{\mathbf{c}} ||\mathbf{Y}' - \mathbf{C}_{\mathrm{eq}}(\mathbf{c})\Big(\mathbf{C}_{\mathrm{eq}}^H(\mathbf{c})\mathbf{C}_{\mathrm{eq}}(\mathbf{c}) + 2\epsilon\mathbf{I}\Big)^{-1} \times \mathbf{C}_{\mathrm{eq}}^H(\mathbf{c})\mathbf{Y}'||^2, \tag{17}$$

It is possible to further simplify (17) for the case of unpenalized ML estimation ($\epsilon = 0$) and attack the optimization problem with gradient-based methods [26]. We have implemented the conjugate gradient method in a Fletcher-Reeves setting with an optimized step-size through a line-search. However, this turns out to be far less robust than, and computationally more expensive to, the method provided next. Therefore we omit to provide the gradient in closed form.

3.3 An EM Algorithm to find the joint Penalized-ML Estimate

Here we provide a robust and computational efficient algorithm to find the joint penalized-ML estimate of $\mathbf{c}$ and $\mathbf{\Psi}$. Instead of pursuing an approach similar to the one used to reach (17), the algorithm has its roots in the joint solution found by setting the gradient of $J_{\mathrm{ML}}(\mathbf{Y}, \mathbf{C}, \mathbf{\Psi}, \epsilon)$ to zero. Before presenting the algorithm, we therefore briefly address this gradient approach.

Each entry of (11) is given by $y_{n,m} = \psi_{n,m}c_m + n_{n,m}$. The derivative of $J_{\mathrm{ML}}(\mathbf{Y}, \mathbf{C}, \mathbf{\Psi}, \epsilon)$ with respect to c_m^* is given by

$$\frac{\partial J_{\mathrm{ML}}(\mathbf{Y}, \mathbf{C}, \Psi, \epsilon)}{\partial c_m^*} = \epsilon c_m + \sum_{\substack{n=1 \\ n\neq m}}^{M} |\psi_{n,m}|^2 c_m - y_{n,m}\psi_{n,m}^*. \tag{18}$$

Setting (18) to zero and solving for c_m yields

$$c_m = \left(\epsilon + \sum_{\substack{n=1 \\ n\neq m}}^{M} |\psi_{n,m}|^2\right)^{-1} \sum_{\substack{n=1 \\ n\neq m}}^{M} \psi_{n,m}^* y_{n,m}, \tag{19}$$

which can be expressed in a vector form as

$$\hat{\mathbf{c}}_{\mathrm{ML}} = \Big(\mathbf{\Psi}_{\mathrm{eq}}^H(\tilde{\mathbf{\Psi}})\mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}}) + \epsilon\mathbf{I}\Big)^{-1}\mathbf{\Psi}_{\mathrm{eq}}^H(\tilde{\mathbf{\Psi}})\tilde{\mathbf{Y}}. \tag{20}$$

In a similar fashion, setting the derivative of $J_{\mathrm{ML}}(\mathbf{Y}, \mathbf{C}, \mathbf{\Psi}, \epsilon)$ with respect to $\psi_{n,m}^*$ to zero and solving for $\psi_{n,m}$ provides

$$\psi_{n,m} = \left(|c_n|^2 + |c_m|^2 + 2\epsilon\right)^{-1}\left(y_{m,n}c_n^* + y_{n,m}c_m^*\right), \tag{21}$$

which can be expressed in a vector form as (16). Equations (19) and (21) show the analytical form for each entry of the penalized-ML vector estimates, which will prove

Algorithm 1 Expectation-Maximization

Require: Measurement matrix $\mathbf{Y}$, convergence threshold Δ_{ML}, penalty parameter ϵ, initial guess $\hat{\mathbf{c}}$

1: **Initialization**: set $\Delta = \delta$ where $\delta > \Delta_{\mathrm{ML}}$
2: **while** $\Delta \geq \Delta_{\mathrm{ML}}$ **do**
3: $\quad \tilde{\mathbf{\Psi}}_{\mathrm{ML}} = \left(\mathbf{C}_{\mathrm{eq}}^{H}(\hat{\mathbf{c}})\mathbf{C}_{\mathrm{eq}}(\hat{\mathbf{c}}) + 2\epsilon\mathbf{I}\right)^{-1}\mathbf{C}_{\mathrm{eq}}^{H}(\hat{\mathbf{c}})\mathbf{Y}'$
4: $\quad \hat{\mathbf{c}}_{\mathrm{ML}} = \left(\mathbf{\Psi}_{\mathrm{eq}}^{H}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}})\mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}}) + \epsilon\mathbf{I}\right)^{-1}\mathbf{\Psi}_{\mathrm{eq}}^{H}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}})\widetilde{\mathbf{Y}}$
5: $\quad \Delta = ||\hat{\mathbf{c}}_{\mathrm{ML}} - \hat{\mathbf{c}}||^2$
6: $\quad \hat{\mathbf{c}} = \hat{\mathbf{c}}_{\mathrm{ML}}$
7: **end while**
8: **Output:** Calibration coefficients estimate $\hat{\mathbf{c}}_{\mathrm{ML}}$.

to be useful during the complexity analysis. Combining the results from (20) and (16) yield the joint solution

$$\begin{bmatrix}\hat{\mathbf{c}}_{\mathrm{ML}} \\ \tilde{\mathbf{\Psi}}_{\mathrm{ML}}\end{bmatrix} = \begin{bmatrix}\left(\mathbf{\Psi}_{\mathrm{eq}}^{H}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}})\mathbf{\Psi}_{\mathrm{eq}}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}}) + \epsilon\mathbf{I}\right)^{-1}\mathbf{\Psi}_{\mathrm{eq}}^{H}(\tilde{\mathbf{\Psi}}_{\mathrm{ML}})\widetilde{\mathbf{Y}} \\ \left(\mathbf{C}_{\mathrm{eq}}^{H}(\hat{\mathbf{c}}_{\mathrm{ML}})\mathbf{C}_{\mathrm{eq}}(\hat{\mathbf{c}}_{\mathrm{ML}}) + 2\epsilon\mathbf{I}\right)^{-1}\mathbf{C}_{\mathrm{eq}}^{H}(\hat{\mathbf{c}}_{\mathrm{ML}})\mathbf{Y}'\end{bmatrix} \tag{22}$$

The particular structure of (22) suggests that a pragmatic approach for solving can be pursued. More specifically, (22) can be separated into two sub-problems, i.e., solving for $\hat{\mathbf{c}}_{\mathrm{ML}}$ and $\tilde{\mathbf{\Psi}}_{\mathrm{ML}}$ separately. Since each of the solutions depend on previous estimates, the joint solution can be computed iteratively, by sequentially solving two separate regularized LS problems, given an initial guess. Since each iteration estimates $\mathbf{c}$ and $\tilde{\mathbf{\Psi}}$ separately, this approach can be seen as an instance of the EM algorithm [27], where the - often challenging - *Expectation step* is performed by estimating only the first moment of the nuisance parameters $\{\psi_{m,n}\}$. The convergence of the algorithm can be analyzed using standard methods, such as a distance between consecutive point estimates. The GMM estimator can be used to compute a reliable initial guess for iteration - in contrast to a purely random initialization. This is often good practice to ensure convergence to a *suitable* local optimum since $J_{\mathrm{ML}}(\mathbf{Y}, \mathbf{C}, \mathbf{\Psi}, \epsilon)$ is not a convex function of its joint parameter space. For sake of clarity, Algorithm 1 summarizes the proposed iterative procedure.

Observe that ϵ, i.e. the penalty term parameter in (13), ends up regularizing both matrix inversions. This is of notable importance from two points-of-view: *i)* from an estimation (robustness) point-of-view, since the matrices to be inverted are constructed from parameter estimates (and thus are subject to estimation errors) and no favorable guarantee exists on their condition number, e.g., see (35). *ii)* from a convergence point-of-view, as it is well-known that the convergence rate of regularized LS adaptive filters is inversely proportional to their eigenvalue spread [28]; This property combo justifies why Ridge Regression was pursued in the first place.

A side remark regarding an application of the EM algorithm follows. We highlight that the calibration coefficients $\mathbf{c}$ and the equivalent channels $\psi_{m,n} = r_m h_{m,n} r_n$

are jointly estimated. As previously mentioned, this a feature is not present in the GMM estimator. Noticeably, this feature makes the EM algorithm robust and hence very suitable to calibrate distributed MIMO systems since channel fading (i.e., high variations of $|h_{m,n}|$) often occurs [13]. As mentioned in Sec. 3.1, the system model used can be also representative to that of distributed systems.

3.4 Complexity Analysis

The complexity of each iteration of Algorithm 1 is dominated by steps 3 and 4. Fortunately the block diagonal structure of the equivalent matrices allows for the inversions to be of reduced complexity, as detailed next. From (21), each calculation of $\psi_{m,n}$ requires a few multiplications and additions. Since $\big(M^2 - M\big)/2$ such calculations are needed to compute (16), the complexity order of step 3 is $\mathcal{O}(M^2)$. Similarly, the complexity of step 4 is $\mathcal{O}(M^2)$ which can be seen directly from (19). The explanation of the $\mathcal{O}(M^2)$ behavior is that the complexity of each calibration coefficient c_m is $\mathcal{O}(M)$, and M such calibration coefficients need to be computed. Overall, each iteration of the EM algorithm is of complexity $\mathcal{O}(M^2)$, and the algorithm's complexity is $\mathcal{O}(N_{\text{ite}}\, M^2)$, with N_{ite} being the number of iterations needed for convergence. The number of iterations needed for convergence is studied in Sec. 3.5.

As for the GMM estimator, the closed-form solutions presented in [13] and [15] have complexity orders of $\mathcal{O}(M^3)$, as they consist of an inverse of a Hermitian matrix of size $M-1$, and of the eigenvector associated with the smallest eigenvalue of a Hermitian matrix of size M.

On a practical note, we remark that the computational complexity of both approaches does not stand as a prohibitive factor for BS arrays using hundreds or even several thousands of antennas. This is because calibration typically needs to be performed on a hourly basis [14, 22].

3.5 Performance Assessment

Simulation setup for the MSE analysis

We simulate reciprocity calibration over a 4×25 rectangular array as the one in Fig. 1. The linear regression parameters obtained in Sec. 2.2 are used to model the coupling gains $\bar{h}_{m,n}$. The mth transceiver maps to the antenna in row a_{row} and column a_{col} of the array as $m = 25(a_{\text{row}} - 1) + a_{\text{col}}$. The reference transceiver index is set to $ref = 38$, as it is associated with one of the most central antenna elements of the 2-D array.

The Cramér-Rao Lower Bound (CRLB) is computed to verify the asymptotical properties of the estimators' error [27]. From (6) and (8), it can be seen that if $\bar{h}_{m,n}$ is assumed to be known, the PDF of $\mathbf{Y}$ conditioned on $\mathbf{R}$ and $\mathbf{T}$ is a multivariate Gaussian PDF. This makes the CRLB of $\mathbf{c}$ to have a well known closed-form, which is computed in Appendix B.

The transmitter t_m and receiver r_m gains are set to $t_m = (0.9 + \frac{0.2m}{M} \exp(-j2\pi m/M))$ $/t_{ref}$ and $r_m = (0.9 + \frac{0.2(M-m)}{M} \exp(j2\pi m/M))/r_{ref}$, respectively. We used this de-

terministic setting for the transceivers, as it allows for a direct comparison of the parameter estimates' MSE with the CRLB. Moreover, this setting incorporates eventual mismatches within the transceivers complex amplitude which are in line with the magnitude variations measured from the transmitters/receivers of our testbed, i.e., spread of around 10-percent around the mean magnitude (and uniform phase). This spread is in line with transceiver models adopted in other calibration works [13].

The variance σ^2 of the multipath propagation contribution during calibration is set to -60 dB. Our motivation for this value is as follows. If the closest physical scatter to the BS is situated, say, 15 meters away, then by Friis' law [29] we have a path loss of around $10\log_{10}(\frac{4\pi d}{\lambda}) = 10\log_{10}(\frac{4\pi(2\times 15m)}{3\times 10^8/(3.7\times 10^9)}) = 73$ dB per path. This number does not account for further losses due to reflections and scattering. Based on this, we use -60 dB as the power (variance) of the resulting channel stemming from a large number of such uncorrelated paths.

For consistency with the reference antenna concept used in the CRLB computations, the MSE of the EM algorithm output $\hat{\mathbf{c}}_{\mathrm{ML}}$, is defined as

$$\mathrm{MSE}_m = \mathrm{E}\left\{ \left| c_m - [\hat{\mathbf{c}}_{\mathrm{ML}}]_{m,1} \,/\, [\hat{\mathbf{c}}_{\mathrm{ML}}]_{ref,1} \right|^2 \right\}, \tag{23}$$

since the estimated "reference" coefficient $[\hat{\mathbf{c}}_{\mathrm{ML}}]_{ref,1}$ is not necessarily equal to 1. This is because the concept of reference antenna is not used by the EM algorithm. As for the GMM estimator, the constraint provided in [15] is adopted, i.e., $c_{ref} = 1$ in (4), which is already coherent with the computed CRLB. The results are averaged over 1000 Monte-Carlo simulations, and the threshold Δ_{ML} is set to 10^{-6} which, based on our experience, ensures that convergence is reached in many parameter settings. The initial guess for the EM algorithm is produced by the GMM estimator.

Estimators' MSE vs CRLB

Fig. 3 compares the MSE of the estimators with the CRLB for two transceiver cases. Both estimators appear to be asymptotically efficient. Noticeably, the performance gains of the EM algorithm can be grossly superior to the GMM (up to 10 dB), as it approaches the CRLB at much smaller values of N_0. As mentioned previously, this is mainly because the GMM estimator does not appropriately weight moment conditions with less quality.

Two remarks about the CRLB itself are now in place. *i)* As mentioned in Appendix B, the assumptions used during the CRLB computations, could result in an underestimated CRLB. Indeed, the results in Fig. 3 suggest that the assumptions used during the CRLB computations do not affect its final value since the estimators' MSE asymptotically converges to the computed CRLB. This is convenient since (asymptotically) efficient estimators can still be built with limited information. *ii)* It was assumed that $\phi_{m,n}$ - the phase of $\bar{h}_{m,n}$ - is known during the CRLB computations, although it is originally modeled as a random variable in Sec.2.2. However, if $\phi_{m,n}$ is assumed to be known, the CRLB is independent of the value of $\phi_{m,n}$. This is because a phase rotation in $\boldsymbol{\mu}_{n,m}$, does not influence (14), due to the structure of

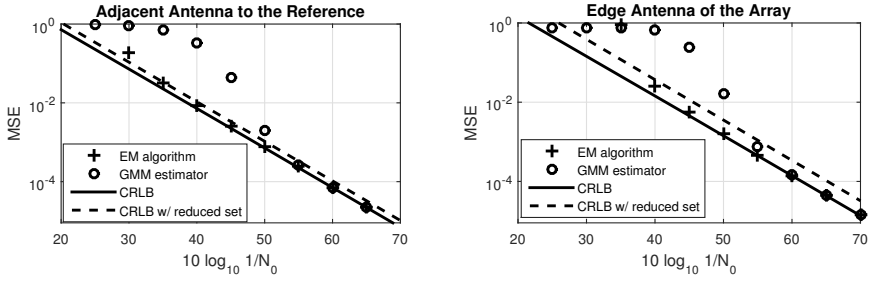

Figure 3: MSE of the GMM estimator and the EM algorithm (with $\epsilon = 0$), versus their CRLB (solid line), for 2 extreme transceiver cases. Namely, a transceiver associated with an antenna at the edge of the array, and a transceiver associated with an antenna adjacent to the reference. The CRLB plotted by a dashed line is discussed in Sec. 3.6.

$\boldsymbol{\Sigma}^{-1}$. Thus, any realization of $h_{m,n}$ - from the model proposed in Sec. 2.4 - provides the same CRLB result.

From the previous two remarks and standard estimation theory [27], it follows that the (narrowband) calibration error - in the high SNR regime - produced by the studied estimators can be well modeled as a multivariate zero-mean Gaussian distribution with covariance matrix given by the transformed inverse Fisher information matrix, found in (38). The Gaussianity of the calibration error is further verified (experimentally) in Sec. 5.4.

Convergence of the EM algorithm

The convergence is analyzed for $N_0 = -40$ dB, which from Fig. 3 appears to be a region where EM-based estimation provides significant gains compared to GMM. Fig. 4 illustrates the role played by the regularization constant ϵ in terms of convergence rate and MSE. Noticeably, the higher ϵ the faster the algorithm appears to converge. The number of iterations until convergence N_{ite} is seen to be much smaller than M with large enough ϵ (i.e., around 5 iterations when $\epsilon = 0.1$).[16] However, increasing ϵ indefinitely is not an option as it degrades the performance. Moreover, the results also indicate that proper tuning of ϵ can provide MSE gains compared to the unregularized case which is asymptotically efficient (notice that this does not conflict with the CRLB theorem, as an estimator built with $\epsilon \neq 0$ is not necessarily unbiased). This was - to some extent - expected due the benefits of Ridge Regression as discussed in Sec.3.3.

With that, we identify that a fine tuning of ϵ can provide many-fold improvements. We note that in the literature there is a number of approaches available that deal with

[16] If, instead, the initial guess is chosen randomly (e.g., calibration coefficients with unit-norm and i.i.d. uniform phases) then our simulations indicate that the order of N_{ite} is $o(M)$.

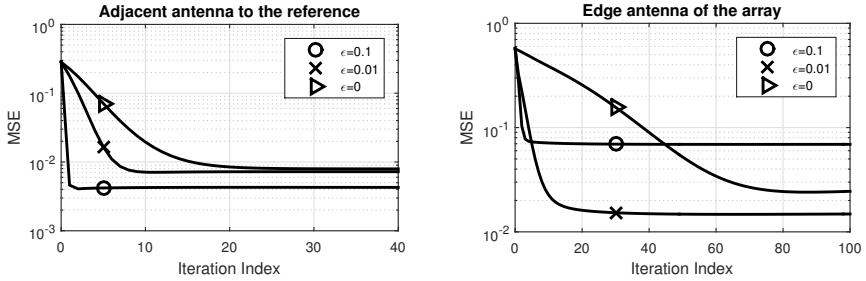

Figure 4: MSE per iteration of the EM algorithm, for different regularization constants ϵ. The plots are for $N_0 = -40$ dB, and the remaining simulation settings are the same as Fig. 3. Note the different scales of the plots.

optimization of regularization constants in standard (non-iterative) LS problems [24]. However, they are not directly applicable to this work as they typically optimize single error metrics, and are in general computationally expensive. Here, our main use for ϵ is to accelerate the convergence and provide estimation robustness to the algorithm, all achieved at no complexity cost. For this matter, we treat ϵ as a hyperparameter (an approach widely adopted in regularized LS adaptive filtering [28]). Further investigation on fully automatizing the EM algorithm is an interesting matter of future work.

For the remainder of the paper, we set $\epsilon = 0$ and proceed accordingly, for simplicity.

Simulation Setup for Sum-rate Capacity Analysis

The same parameter setting as in Sec. 3.5 is kept in this setup, and the remaining simulation framework is defined next.

We assume that the uplink channel $\mathbf{H}_{\text{UP}}$ is perfectly know to the BS, and that there are two noise sources in the system. The first noise source is downlink additive noise modeled by $\mathbf{w}'$, see (2). Here, $\mathbf{w}'$ have i.i.d. zero-mean circularly symmetric complex Gaussian distributed random entries with variance N_w equal to 1. The same model is used for the entries of the downlink channel matrix $\mathbf{H}_{\text{DL}}$. The second noise source is the error during estimation of $\mathbf{c}$ (i.e., calibration error). With that, the precoded signal $\mathbf{z}' = \mathbf{P}\mathbf{x}$ is subject to calibration errors. The transmit power constraint $\mathbb{E}\{||\mathbf{z}'||^2\} = K$ is used. Also, we set $K = 10$ single antenna users, and assume $t_k^U = t_k^B$ and $r_k^U = r_k^B$ for sake of simplicity.

The sum-rate capacities [30] are evaluated for different calibration cases. More specifically, when no calibration is employed (i.e., $\hat{c}_m = 1$), when calibration is performed with the GMM or the EM algorithm, for the case of perfect calibration (i.e., $\hat{c}_m = c_m$), and as a baseline, when precoding is performed using the *true* down-

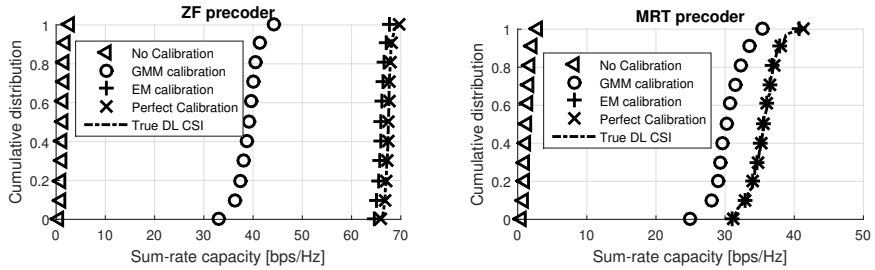

Figure 5: CDFs of the sum-rates capacities for different calibration cases. Left) ZF precoder; Right) MRT precoder.

link channel $\mathbf{H}_{\mathrm{DL}}$. The analysis is performed with $N_0 = -40$ dB, for the reasons mentioned during the convergence analysis.

Sum-rate Capacity Results

Fig. 5 shows the obtained sum-rates cumulative distribution functions (CDFs) for different precoding schemes [2]. Similarly to the MSE results, EM-based calibration provides significant gains compared to the GMM case. The magnitude of these gains obviously depend on both the calibration (and communication system) setup. For example, there are no sum-rate differences when $N_0 \to 0$ or $N_0 \to \infty$, as both GMM and EM approaches converge to that of perfect calibration, or to the uncalibrated case, respectively. Thus, it in only in a certain region of N_0 values that EM based calibration provides gains.[17]

It is interesting that - for this setup - there is no fundamental loss in capacity between this calibration approach (i.e., precoding with perfectly calibrated uplink CSI) and precoding with the true downlink CSI. Quantifying this loss is out of scope of this work, however, the interested reader is referred to [32] for an overview on the loss of different types of reciprocity calibration. We now finalize the section with two interesting remarks.

3.6 Remark 1: Calibration with Reduced Measurement Sets

There are several benefits of using a reduced measurement set for calibration (e.g., by only relying on high quality measurements). This is possible as long as (11) is not under-determined. As an illustrative example, the dashed line in Fig. 3 shows

[17] Our analysis based on a wide range of parameter values also indicates that, in general, stricter calibration requirements need to be met in order to release the full potential of ZF compared to MRT precoding (i.e., no sum-rate difference compared to the perfect calibrated case). Noticeably, this observation is in line with previous calibration studies [31].

the CRLB when a reduced measurements set - comprising the measurements between antenna pairs whose elements are distanced by at most $1/\sqrt{2}$ wavelengths - is used. The number of measurement signals in this case drops from $M(M-1)$ to less than $8M$, since one antenna signals to, at most, 8 other antennas. The performance loss turns out to be insignificant, i.e. 2 dB for the neighbor case and 4 dB for the edge case, considering the number of signals discarded. This indicates that the channels between neighbor antennas, which are dominated by mutual coupling, are the most important for calibration. Thus, there is an interesting trade-off between the asymptotic performance of an estimator and its computational complexity (proportional to the number of measurements).

Another benefit of using reduced measurement sets is a possible reduction of resource overhead dedicated for calibration. This can be very important from a system deployment point-of-view. To finalize, we remark that ML closed form estimators can be also reached when reduced measurement sets are used. This can be the case for the current (general) calibration setup when a reduced set of measurements is used, or for the case of working with a full set of measurements when the calibration setup is a special case. An example of the latter is given next.

3.7 Remark 2: Closed-form Unpenalized ML Calibration for Linear arrays

Consider an M-antenna linear array, and let m index the antennas in ascending order starting at one edge of the linear array. Assume that mutual coupling only exists between adjacent antenna elements, and that the channel between any other antenna pairs is weak enough so that it can be neglected without any noticeable impact on performance. We summarize our findings in Proposition 1.

Proposition 1: Using a reference antenna as a starting point, say $c_1 = 1$, the unpenalized ML solution for any $c_{\ell+1}$, with $1 \leq \ell \leq M-1$, can be obtained sequentially by

$$\hat{c}_{\ell+1} = \hat{c}_\ell \frac{y^*_{\ell+1,\ell} y_{\ell,\ell+1}}{|y_{\ell+1,\ell}|^2}. \tag{24}$$

Proof: See Appendix C. ■

We can also deduce the following interesting corollary.

Corollary 1: For any of the two constraints considered in (4), the GMM (vector) estimator coincides with (24) up to a common complex scalar.

Proof: See Appendix C. ■

4 Validation of the calibration method in a massive MIMO testbed

In this section, we detail the experiment performed to validate the proposed mutual coupling based calibration method. More specifically, we implemented it in a software-

Table 1: High-level OFDM parameters

Parameter	**Variable**	**Value**
Carrier frequency	f_{c}	3.7 GHz
Sampling Rate	F_{s}	7.68 MS/s
FFT Size	N_{FFT}	2048
# Used sub-carriers	N_{SUB}	1200

defined radio testbed, and performed a TDD transmission from 100 BS antennas to 3 single antenna terminals.

Note that the analysis conducted in this section and in Sec. 5 is measurement based. As stationarity is assumed in the analysis, we monitored the system temperature throughout the measurements and verified no significant changes. We also made an effort to keep static propagation conditions, and performed the experiments at late hours in our lab with no people around.

4.1 Brief Description of the Testbed

Here we briefly outline the relevant features of the testbed for this work. Further information can be found in [22].

Antenna/Transceiver setup

The BS operates with 100 antennas, each antenna connected to one distinct transceiver. For simplicity, the same transceiver settings (e.g., power amplifier gain and automatic gain control) are used in both calibration and data communication stages for all radio units. This ensures that the analog front-ends yield the same response during both stages, thus the estimated calibration coefficients are valid during the communication stage.

Synchronization of the radios

Time and Frequency synchronization is achieved by distributing reference signals to all radio units. However, this does not guarantee phase alignment between all BS transceiver radio chains which motivates reciprocity calibration.

4.2 Communication Protocol used

Once the measurements to construct the observation matrix $\mathbf{Y}$ are performed, $\mathbf{c}$ is estimated using the unpenalized EM algorithm. The following sequence of events is then performed periodically:

Uplink Channel Estimation and Calibration

Users simultaneously transmit frequency orthogonal pilot symbols. The BS performs LS-based channel estimation, and interpolates the estimates between pilot symbols. Reciprocity calibration is then performed independently per subcarrier, i.e. as in (3), for coherence purposes with Sec. 2. This calibrated version of the downlink channel is then used to construct a ZF precoder.

Downlink channel estimation and data transmission

Downlink pilot symbols are precoded in the downlink and each user performs LS-based channel estimation. Using the estimates, each user recovers the payload data using a one-tap equalizer.

We note that 4-QAM signaling per OFDM sub-carrier is used for uplink channel estimation and data transmission. The main parameters are shown in Table 1. Further information on the signaling protocol (e.g., uplink/downlink frame structure or uplink pilot design) is found on [22].

4.3 Measurement Description

The setup used in our experiments is shown in Figure 1. Although not being a typical propagation scenario found in cellular systems, this extreme setup - closely located users under strong line-of-sight conditions - requires high calibration requirements to be met if spatial separation of users is to be achieved. In addition, we use ZF precoding as it is known to be very sensitive to calibration errors [32].

The EVM [33] of the downlink equalized received samples at each mobile station was evaluated, and used as performance metric for validation purposes. The rationale is that, with multiple mobile terminals, calibration errors are translated into downlink inter-user interference (and loss of array gain), which increases the EVM. Letting r be the downlink equalized received sample when symbol s is transmitted, the EVM is defined as

$$\mathrm{EVM} = \mathrm{E}\left\{ \frac{|r-s|^2}{|s|^2} \right\}, \tag{25}$$

where the expectation is taken over all system noise sources (e.g., hardware impairments and thermal noise). Our estimate of (25) was obtained by averaging realizations of $|r-s|^2/|s|^2$ over all OFDM sub-carriers and over received OFDM symbols.

We estimated the EVM for different energy values of the uplink pilots and calibration signals. We do so in order to be able to extract insightful remarks for

the analysis of the results. In particular, letting $E_{\text{Pilot}} = \text{E}\{p_k p_k^*\}$ in (1) denote the energy of the uplink pilot, which, for simplicity, is the same for all users, and let E_{Cal} denote the energy of the sounding signal s_m in (5), we estimated the EVM for a 2-dimensional grid of E_{Pilot} and E_{Cal}. The results reported next are given with respect to the relative energies $Er_{\text{Pilot}} = E_{\text{Pilot}}/E_{\text{Pilot}}^{\max}$ and $Er_{\text{Cal}} = E_{\text{Cal}}/E_{\text{Cal}}^{\max}$, where $E_{\text{Pilot}}^{\max}$ and $E_{\text{Cal}}^{\max}$ are the maximum energies of the uplink pilot and calibration signal used in the experiments. Other systems parameters (e.g., transmit power in the downlink) were empirically set and kept constant throughout the experiment.

4.4 Validation Results

Fig. 6 shows the measured EVMs for the 3 user terminals in our experiment. Before discussing the results, we remark that analyzing the EVM when Er_{Cal} is reduced beyond -30 dB is not of fundamental interest, as it approaches the uncalibrated case (where high EVMs are to be expected). Overall, a positive trend is observed with increasing Er_{Cal} until -10 dB. This reflects the BS ability of spatially separating users which increases with increasing the calibration quality. The fact that downlink EVMs down to -10 dB are achieved, which are much smaller than the EVMs when $Er_{\text{Cal}} = -30$ dB, i.e. close to the uncalibrated case, motivates our validation claim.

It is possible to observe a saturation of the EVMs at high enough Er_{Cal} and Er_{Pilot} for all user cases. This is an expected effect in practical systems. Explained briefly, system impairments other than the calibration or the uplink channel estimation error, become the dominant error sources that bound the EVM performance[18]. Remarkably, this saturation effect implies that the calibration SNR - available in a practical array as ours - is sufficiently large not to be the main impairment to constrain the system performance. Mutual coupling channels are thus reliable (and reciprocal enough), so that they can be used for signaling in order to calibrate the system.[19]

[18] Mobile terminals error sources (e.g., in-phase and quadrature imbalance or thermal noise) qualify for such impairments. For a given downlink transmit power, it is straightforward to understand how such impairments bound the downlink EVMs regardless of the calibration and uplink estimation quality. [19] We note there exists an interesting theoretical trade-off between the calibration quality and the capacity of downlink channels with respect to the strength of mutual coupling. In practice, the proposed calibration method can be used in compact antenna arrays with very low coupling (say -30 dB between adjacent elements) provided that the transmit power during calibration is sufficient to provide good enough estimation SNR. In such a setup, the impact of coupling in the capacity is negligible.

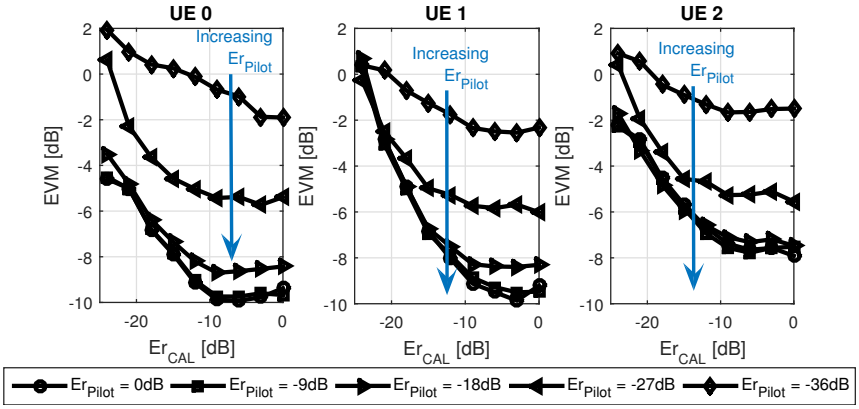

Figure 6: Measured EVM at each of the three user terminals during a massive MIMO downlink transmission.

5 Aspects of Wideband Calibration and Error Modeling

A short summary of this section follows. Using the measurements from the Sec. 4, we treat the estimated calibration coefficients across OFDM sub-carriers as realizations of a discrete stochastic process. Using low rank approximation theory, we propose a parametrized low dimensional basis that characterizes the subspace spanned by this process accurately. Based on the reduced basis, we propose a wideband estimator that averages out the calibration error across frequency. Using the wideband estimator results, we validate the narrowband calibration error model proposed in Sec. 3.5. We remark that our experiment makes use of a bandwidth of $F_s N_{\text{sub}}/N_{\text{FFT}} = 4.5\text{MHz}$.

5.1 Wideband Remarks for the Calibration Coefficients

Denote the calibration coefficient of BS antenna m at the kth OFDM sub-carrier as $C_m[k] = t_m^k / r_m^k$. The variable $\hat{C}_m[k]$ is the estimate of $C_m[k]$ at sub-carrier k - obtained, e.g., with the EM algorithm - and is modeled as

$$\begin{aligned}\hat{C}_m[k] =& C_m[k] + E_m[k] \\ =& |C_m[k]| \exp(j2\pi\zeta_m[k]) + E_m[k]\end{aligned} \tag{26}$$

where $E_m[k]$ is an i.i.d. random process representing the calibration error which is assumed zero-mean and independent of $C_m[k]$. Let the random phasor process $\exp(j2\pi\zeta_m[k])$ in (26) absorb the phase shift stemming from the arbitrary time that a local oscillator needs to lock to a reference signal. Such phase shift is often modeled as uniformly distributed, and thus

$$\mathrm{E}\left\{\exp(j2\pi\zeta_m[k])\right\} = 0. \tag{27}$$

Moreover, since local oscillators associated with different transceivers lock at arbitrary times, it is safe to assume

$$\mathrm{E}\left\{\exp(j2\pi\zeta_m[k_1])\exp(-j2\pi\zeta_n[k_2])\right\} = 0,\ m \neq n. \tag{28}$$

Not making further assumptions on the statistics of $\hat{C}_m[k]$, we now proceed with a series expansion, but before doing so we make one last remark. The series expansion conducted next is performed based on measurements from the 100 testbed transceivers, and serves as an example approach to obtain a suitable basis for $\hat{C}_m[k]$. This can well apply to mass-production transceiver manufactures that can reliably estimate the statistical properties of the hardware produced. However, as our testbed operates with relatively high-end transceivers - compared to the ones expected to integrate commercial massive MIMO BSs - the dimensionality of the subspace verified in our analysis might be underestimated. Intuitively, the higher transceiver quality, the less basis functions are needed to accurately describe $\hat{C}_m[k]$. Nevertheless, the upcoming remarks apply for smaller bandwidths - than 4.5MHz - depending on the properties of the transceivers.

5.2 Principal Component Analysis

From the assumption (27), it follows that the element at the v_1th row and v_2th column of the covariance matrix $\mathbf{K}_m$ of $\hat{C}_m[k]$ is defined as

$$[\mathbf{K}_m]_{[v_1,v_2]} = \mathrm{E}\left\{\hat{C}_m[v_1]\ \hat{C}_m^*[v_2]\right\}. \tag{29}$$

From the assumption (28), it follows that the principal components of $\hat{C}_m[k]$ are obtained by singular value decomposition (SVD) of $\mathbf{K}_m$ only [34]. Let the SVD of $\mathbf{K}_m$ be written as

$$\mathbf{K}_m = \sum_{i=1}^{\mathrm{N_{SUB}}} \mathbf{u}_i^m \lambda_i^m (\mathbf{u}_i^m)^H, \tag{30}$$

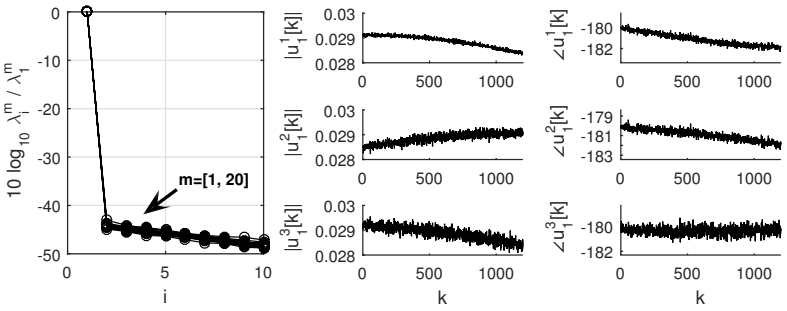

Figure 7: Principal component and coefficients of $\hat{C}_m[k]$. *Left)* The 10 strongest normalized singular values for 20 transceivers; *Middle)* Magnitude of the principal component for 3 transceivers; *Right)* Phase of the principal component for 3 transceivers.

where $\{\mathbf{u}_i^m\}_{i=1}^{\mathrm{N_{SUB}}}$ are the principal components, and λ_i^m is the power (variance) of the coefficient obtained from projecting $\hat{C}_m[k]$ into $\mathbf{u}_i^m$. We use the convention $\lambda_1^m \geq \lambda_2^m \cdots \geq \lambda_{N_{\mathrm{SUB}}}^m$, and $\mathbf{u}_i^m = \left[[u_i^m[1], \cdots, u_i^m[N_{\mathrm{SUB}}]\right]^{\mathrm{T}}$. Fig. 7 shows several coefficients and basis functions of the expansion, that were estimated based on 100 realizations of $\hat{C}_m[k]$, each measured with $Er_{\mathrm{Cal}} = 5$ dB (which from Fig. 6 provides a relatively high calibration SNR). Noticeably, it appears that all processes (one per transceiver) live mostly in a one-dimensional sub-space and thus can be well described by their first principal component $\mathbf{u}_1^m$. This fact also indicates that the contribution of the calibration error in the expansion is small, and thus the first principal component of $\hat{C}_m[k]$ is also representative for the true coefficients $C_m[k]$.

Visual inspection indicates that both magnitude and phase of the first principal component can be well approximated with a linear slope across frequency. The inherent error of this approximation is very small compared to the magnitude of the process itself. We note that this linear trend holds for any transceiver of the array (not only for the ones shown in Fig. 7).

5.3 Wideband Modeling and Estimation

The previous analysis indicates that any first principal component can be well described by a linear magnitude slope γ_m, and a linear phase ξ_m across frequency. Such properties are well captured by the Laplace kernel $\exp((\gamma_m + j2\pi\xi_m)k)$, for small values of $|\gamma_m|$ (since the range of k is finite). The final

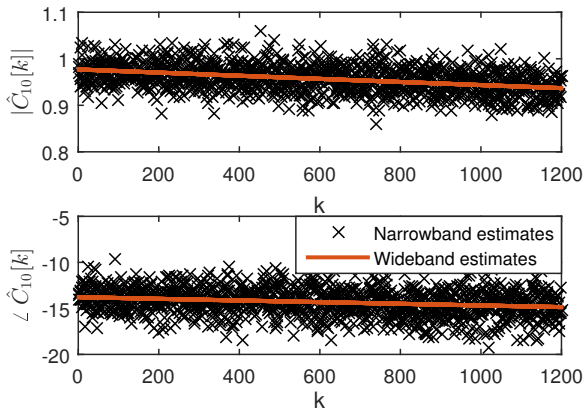

Figure 8: A realization of the narrow-band estimator $\hat{C}_m[k]$, and the proposed wideband estimator $\hat{C}_i[k]^{\mathrm{WB}}$.

parameter to model a realization of the process is the complex offset A_m. With that, the general model (26) can thus be re-written as

$$\hat{C}_m[k] = A_m \exp((\gamma_m + j2\pi\xi_m)k) + w_m[k], \tag{31}$$

where $w_m[k]$ is a random process that absorbs: the calibration error $E_m[k]$, the error due to the low rank approximation, and the error due to the linear modeling of the first principal component $\mathbf{u}_1^m$. Given an observation $\{\hat{C}_m[k]\}_{k=1}^{N_{\mathrm{SUB}}}$, the ML estimator of A_m, ξ_m and γ_m, namely, $\hat{A}_m$, $\hat{\xi}_m$ and $\hat{\gamma}_m$ is straightforward to derive [27]. Thus, we define the wideband estimator of $\hat{C}_m[k]$ as

$$\hat{C}_m[k]^{\mathrm{WB}} = \hat{A}_m \exp((\hat{\gamma}_m + j2\pi\hat{\xi}_m)k). \tag{32}$$

For illustration purposes, a realization of the ML wideband estimator $\hat{C}_m[k]^{\mathrm{WB}}$ is contrasted with that of the narrow-band estimator $\hat{C}_m[k]$ in Fig. 8. The obtained error reduction is evident.

5.4 A Model for the Calibration Error

Here, we use the wideband estimator results to verify the Gaussianity of the narrow-band calibration error proposed in Sec. 3.5. This is done under the two following main assumptions.

1) The residual process $E_m[k] = \hat{C}_m[k] - C_m[k]$ *is well described by* $\hat{E}_m[k] = \hat{C}_m[k] - \hat{C}_m[k]^{\mathrm{WB}}$. This is reasonable if $\mathrm{E}\left\{|\hat{C}_m[k]^{\mathrm{WB}} - C_m[k]|^2\right\} \ll$

$\mathrm{E}\left\{|\hat{C}_m[k] - C_m[k]|^2\right\}$. To justify, the estimation gains scale linearly in the number of realizations [27], which is $N_{\mathrm{SUB}} = 1200$ in this case. Assuming that: the estimation error is independent across realizations, the underlying model (31) describes the first principal component well, and the low rank approximation error is minuscule, there are gains of $10 \log_{10} N_{\mathrm{SUB}} \approx 30$ dB which justify the first main assumption.

2) The residual process $E_m[k]$ is ergodic.[20] This is met if $E_m[k]$ is stationary and the ensemble of N_{SUB} samples is representative for statistical modeling. The former holds for small OFDM bandwidths (e.g., 4.5 MHz) as the hardware impairments do not vary significantly across the band. The latter is also met, as we have $N_{\mathrm{SUB}} = 1200$ narrow-band estimators whose estimated errors $\{\hat{E}_m[k]\}_{k=1}^{N_{\mathrm{SUB}}}$ were found to be mutually uncorrelated.

Fig. 9 shows the empirical CDF of both real and imaginary parts of $\{\hat{E}_m[k]\}_{k=1}^{N_{\mathrm{SUB}}}$ - which we found to the uncorrelated - for two transceiver cases. Each of the empirical CDFs is contrasted with a zero-mean Gaussian distribution of equal variance. Overall, the empirical CDFs for both transceivers resemble a Gaussian CDF extremely well. The Gaussianity of the calibration error was further verified by passing a Kolmogorov-Smirnov test with 0.05 significance level [35]. We note that these observations hold not only for the two transceivers in Fig. 9, but for all transceivers of the array. Noticeably, the empirical distribution of the calibration error is in line with the asymptotic properties of ML estimators, i.e. the error can be modeled by an additive zero-mean Gaussian multivariate. The final element for a full characterization is its covariance matrix, relating the errors across antennas. A good approximation (at high SNR) is the inverse of the transformed Fisher Information matrix in (38). Noticeably, future calibration works can benefit from the convenience of safely assuming a non-white Gaussian calibration error.

6 Conclusions

We have proposed and validated a convenient calibration method which rely on mutual coupling to enable the reciprocity assumption in TDD massive MIMO systems. We verified that in a practical antenna array, the channels due to mutual coupling are reliable and reciprocal enough, so that they can be used for signaling in order to calibrate the array.

[20] Ergodicity is necessary since each (independent) measurement of $\hat{C}_m[k]$ takes about ten minutes with our test system (due to the locking time of the local oscillator to the reference signal). As potential system temperature drifts during the measurements can result in varying statistical properties, it is safer to perform the analysis based on one solely realization of $E_m[k]$.

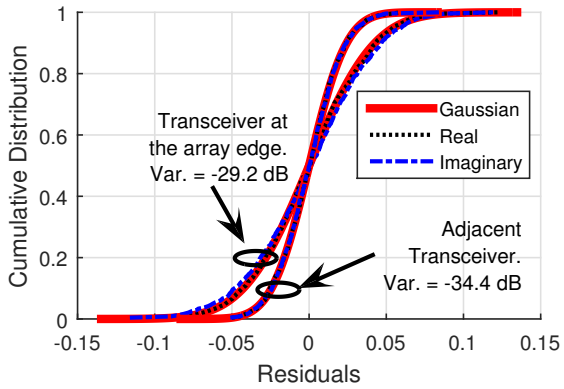

Figure 9: Empirical CDFs for the real and imaginary parts of the calibration error, for a transceiver at the edge of the array, and for an adjacent transceiver to the reference antenna. A Gaussian CDF of equal variance is plotted for both cases for comparison.

The iterative ML algorithm is asymptotically efficient and outperforms current state-of-the-art estimators in an MSE and sum-rate capacity sense. Further improvements - in terms of MSE and convergence rate - can be harvested by proper tuning of its regularization hyperparameter.

The calibration error can be further reduced by proper averaging over the radio bandwidth. More importantly, it did not stand as the main impairment to constraint the performance of the system, from our experiments. Our measurements also verified that the narrow-band calibration error (at high SNR) is Gaussian distributed, which is coherent with the theory of the estimator proposed. The convenience of safely assuming a non-white Gaussian calibration error can, hopefully, open the door for future analytical studies of calibrated TDD massive MIMO systems.

Acknowledgments

This work was funded by the Swedish foundation for strategic research SSF, VR, the strategic research area ELLIIT, and the E.U. Seventh Framework Programme (FP7/2007-2013) under grant agreement n 619086 (MAMMOET). We also thank the Comm. Systems group in Bristol University, for letting us replicate several of our results in their testbed.

Appendix A: Equivalent Channel Matrices

Here we show the structure of the equivalent models. Define the column vector $\boldsymbol{\Psi}_m = [\psi_{1,m} \dots \psi_{m-1,m}\ \psi_{m+1,m} \dots \psi_{M,m}]^T$. The equivalent channel matrix in (14) is written as

$$\boldsymbol{\Psi}_{\text{eq}}(\tilde{\boldsymbol{\Psi}}) = \text{diag}\left\{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \dots, \boldsymbol{\Psi}_M\right\}. \tag{33}$$

Now define

$$\bar{\mathbf{c}}_{n,m} = [c_n\ c_m]^T. \tag{34}$$

Noting that $\psi_{m,n} = \psi_{n,m}$, the equivalent matrix and the parameter vector in (15) are written as

$$\mathbf{C}_{\text{eq}}(\mathbf{c}) = \text{diag}\left\{\bar{\mathbf{c}}_{1,2}, \cdots, \bar{\mathbf{c}}_{1,M}, \bar{\mathbf{c}}_{2,3}, \cdots, \bar{\mathbf{c}}_{2,M}, \cdots\right\}, \tag{35}$$

and

$$\tilde{\boldsymbol{\Psi}} = \left[\psi_{2,1} \dots \psi_{M,1}\ \psi_{3,2} \dots \psi_{M,2} \dots \psi_{M,M-1}\right]^T. \tag{36}$$

Appendix B: The Cramér-Rao Lower Bound

Here we compute the CRLB for the calibration coefficients $\{c_m\} \setminus c_{ref}$. The exclusion of c_{ref} is justified in the end of the calculations. This is achieved by assuming $t_{ref} = r_{ref} = 1$, and treating $c_{ref} = t_{ref}/r_{ref}$ as known for estimation purposes. Define the $(4M-4) \times 1$ vector

$$\grave{} = [\text{Re}\{t_1\}\ \text{Im}\{t_1\}\ \text{Re}\{r_1\}\ \text{Im}\{r_1\}\ \text{Re}\{t_2\} \dots \text{Im}\{r_M\}]^T, \tag{37}$$

where t_{ref} and r_{ref} do not enter. The CRLB for $\{c_m\} \setminus c_{ref}$ is given by the diagonal entries of the transformed inverse Fisher information matrix [27]

$$\text{var}(\hat{c}_m) \geq \left[\frac{q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\text{H}}\right]_{m,m}, \ m \neq ref, \tag{38}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix of $\boldsymbol{\theta}$. The transformation of $\boldsymbol{\theta}$ into the calibration coefficients is given by

$$q(\boldsymbol{\theta}) = \left[\frac{\text{Re}\{t_1\} + j\,\text{Im}\{t_1\}}{\text{Re}\{r_1\} + j\,\text{Im}\{r_1\}} \cdots \frac{\text{Re}\{t_M\} + j\,\text{Im}\{t_M\}}{\text{Re}\{r_M\} + j\,\text{Im}\{r_M\}}\right]^T.$$

We now compute $\mathbf{I}(\boldsymbol{\theta})$. Assuming that $\bar{h}_{m,n}$, σ^2 and N_0 are at hand,[21] the mean $\boldsymbol{\mu}_{n,m}$ and the covariance matrix $\boldsymbol{\Sigma}_{n,m}$ of $\mathbf{y}_{n,m} = [y_{n,m}\ y_{m,n}]^T$ are given

[21] These assumptions are only used for the CRLB calculations, and were not used to derive any of the estimators. A possible implication is that the CRLB can be underestimated, but we will see that this is not the case from the simulations' results.

by

$$\boldsymbol{\mu}_{n,m} = \mathrm{E}\left\{\mathbf{y}_{n,m}\right\} = \bar{h}_{n,m}\left[r_n t_m \; r_m t_n\right]^T, \tag{39}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{n,m} &= \mathrm{E}\left\{(\mathbf{y}_{n,m} - \boldsymbol{\mu}_{n,m})(\mathbf{y}_{n,m} - \boldsymbol{\mu}_{n,m})^H\right\} \\ &= \begin{bmatrix} |r_n|^2|t_m|^2\sigma^2 + N_0 & r_n t_m r_m^* t_n^* \sigma^2 \\ r_m t_n r_n^* t_m^* \sigma^2 & |r_m|^2|t_n|^2\sigma^2 + N_0 \end{bmatrix}.\end{aligned} \tag{40}$$

We can observe that the PDF of $\mathbf{Y}''$, where

$$\mathbf{Y}'' = \left[\mathbf{y}_{1,2}^T \cdots \mathbf{y}_{1,M}^T \; \mathbf{y}_{2,3}^T \cdots \mathbf{y}_{2,M}^T \cdots \mathbf{y}_{M-1,M}^T\right]^T,$$

conditioned on $\grave{}$, follows a multivariate Gaussian distribution, i.e., $p(\mathbf{Y}''|\boldsymbol{\theta}) \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_{1,2}^T \cdots \boldsymbol{\mu}_{1,M}^T \boldsymbol{\mu}_{2,3}^T \cdots \boldsymbol{\mu}_{2,M}^T \cdots \boldsymbol{\mu}_{M-1,M}^T\right]^T$ and block diagonal covariance

$$\boldsymbol{\Sigma} = \mathrm{diag}\left\{\boldsymbol{\Sigma}_{1,2}, \cdots, \boldsymbol{\Sigma}_{1,M}, \boldsymbol{\Sigma}_{2,3}, \cdots, \boldsymbol{\Sigma}_{2,M}, \cdots, \boldsymbol{\Sigma}_{M-1,M}\right\}. \tag{41}$$

With that, we have

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{i,j} = \; \mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j}\right\} + 2\,\mathrm{Re}\left\{\frac{\partial \boldsymbol{\mu}^H}{\partial \theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_j}\right\}, \tag{42}$$

with $1 \leq i \leq (4M-4)$ and $1 \leq j \leq (4M-4)$. The remaining computations of $\left[\mathbf{I}(\boldsymbol{\theta})\right]_{i,j}$ are straightforward and thus omitted. We note that without the convention of $t_{ref} = r_{ref} = 1$ - and thus $\boldsymbol{\theta}$ is a $4M \times 1$ vector instead - it can be shown that the map $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}$ is not injective which renders $\mathbf{I}(\boldsymbol{\theta})$ not invertible. Thus, the convention of reference antenna is needed to be able to compute the CRLB.

Appendix C - Closed-form Unpenalized ML estimator for Linear Arrays

Here we derive the closed-form unpenalized (i.e. $\epsilon = 0$) ML estimator for the linear array setup described in Sec. 3.7. By leaving out the terms that do not depend on $\mathbf{c}$, it follows that, after a few manipulations, the optimization problem of (17) can be written as

$$\begin{aligned}\{\hat{c}_m\} &= \arg\max_{\mathbf{c}} \mathbf{Y'}^H \mathbf{C}_{\mathrm{eq}}(\mathbf{c})\mathbf{C}_{\mathrm{eq}}^{\dagger}(\mathbf{c})\mathbf{Y'} \\ &= \arg\max_{\{c_m\}} \sum_{\ell=1}^{M-1} f_{\mathrm{L}}(c_\ell, c_{\ell+1}, \mathbf{y}_{\ell+1,\ell}),\end{aligned} \tag{43}$$

with

$$f_{\mathrm{L}}(c_\ell, c_{\ell+1}, \mathbf{y}_{\ell+1,\ell}) = \mathbf{y}_{\ell+1,\ell}^H \overline{\mathbf{c}}_{\ell,\ell+1} \overline{\mathbf{c}}_{\ell,\ell+1}^H \mathbf{y}_{\ell+1,\ell} / \overline{\mathbf{c}}_{\ell,\ell+1}^H \overline{\mathbf{c}}_{\ell,\ell+1}.$$

See (34) for structure of $\overline{\boldsymbol{c}}_{\ell,\ell+1}$, and (12) for structure of $\mathbf{y}_{m,n}$.

Our ability to solve (43) is due to the following property.

Property 1: For the function $f_{\mathrm{L}}(c_\ell, c_{\ell+1}, \mathbf{y}_{\ell+1,\ell})$, the maximum over $c_{\ell+1}$ equals $||\mathbf{y}_{\ell+1,\ell}||^2$, and thus it does not depend on c_ℓ.

Hence, the ML estimate of $c_{\ell+1}$, i.e. $\hat{c}_{\ell+1}$, can be found for a given c_ℓ. With that, the joint maximization problem (43) can be split into

$$\hat{c}_{\ell+1} = \arg\max_x f_{\mathrm{L}}(\hat{c}_\ell, x, \mathbf{y}_{\ell+1,\ell}).$$

This optimization is a particular case of the Rayleigh quotient problem, and the solution is given in (24) when the reference element (i.e., the starting point) is chosen to be c_1.

We now provide a short proof for Corollary 1. For the case of linear arrays with coupling solely between adjacent antennas, the optimization problem in (4) can be written - ignoring any constraint for now - as

$$\hat{\boldsymbol{c}}_{\mathrm{GMM}} = \arg\min_{\boldsymbol{c}} \sum_{\ell=1}^{M-1} f_{\mathrm{G}}(c_\ell, c_{\ell+1}, \mathbf{y}_{\ell+1,\ell}) \tag{44}$$

where $f_{\mathrm{G}}(c_\ell, c_{\ell+1}, \mathbf{y}_{\ell+1,\ell}) = |y_{\ell+1,\ell} c_{\ell+1} - y_{\ell,\ell+1} c_\ell|^2$. We solve (44) using the following property.

Property 2: Letting $\hat{c}_\ell$ be the ML estimator from (24), it follows that

$$f_{\mathrm{G}}(\hat{c}_\ell, \hat{c}_{\ell+1}, \mathbf{y}_{\ell+1,\ell}) = 0, \ \forall \ell. \tag{45}$$

Thus, the GMM solution (under any of the 2 constraints) coincides with that of the ML up to a common complex scalar. Uniqueness follows since the GMM cost function is quadratic.

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[3] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, Feb 2002.

[4] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[5] M. Salehi and J. Proakis, *Digital Communications*. McGraw-Hill Education, 2007.

[6] C. A. Balanis, *Antenna Theory: Analysis and Design*. Wiley-Interscience, 2005.

[7] F. Kaltenberger, H. Jiang, M. Guillaud, and R. Knopp, "Relative channel reciprocity calibration in MIMO/TDD systems," in *2010 Future Network and Mobile Summit*, June 2010.

[8] M. Petermann et al., "Multi-User Pre-Processing in Multi-Antenna OFDM TDD Systems with Non-Reciprocal Transceivers," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3781–3793, September 2013.

[9] K. Nishimori, K. Cho, Y. Takatori, and T. Hori, "Automatic calibration method using transmitting signals of an adaptive array for TDD systems," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 6, pp. 1636–1640, Nov 2001.

[10] K. Nishimori, T. Hiraguri, T. Ogawa, and H. Yamada, "Effectiveness of implicit beamforming using calibration technique in massive MIMO system," in *2014 IEEE International Workshop on Electromagnetics (iWEM)*, Aug 2014, pp. 117–118.

[11] X. Luo, "Robust Large Scale Calibration for Massive MIMO," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015.

[12] H. Wei, D. Wang, H. Zhu, J. Wang, S. Sun, and X. You, "Mutual Coupling Calibration for Multiuser Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 606–619, Jan 2016.

[13] R. Rogalin et al., "Scalable Synchronization and Reciprocity Calibration for Distributed Multiuser MIMO," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, April 2014.

[14] C. Shepard et al., "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 53–64.

[15] R. Rogalin, O. Y. Bursalioglu, H. C. Papadopoulos, G. Caire, and A. F. Molisch, "Hardware-impairment compensation for enabling distributed large-scale mimo," in *Information Theory and Applications Workshop (ITA), 2013*, Feb 2013.

[16] H. Wei, D. Wang, J. Wang, and X. You, "TDD reciprocity calibration for multiuser massive MIMO systems with iterative coordinate descent," *Science China Information Sciences*, vol. 59, no. 10, p. 102306, 2015.

[17] H. Papadopoulos, O. Bursalioglu, and G. Caire, "Avalanche: Fast RF calibration of massive arrays," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 607–611.

[18] J. Vieira, F. Rusek, and F. Tufvesson, "Reciprocity calibration methods for massive MIMO based on antenna coupling," in *2014 IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 3708–3712.

[19] H. Wei, D. Wang, and X. You, "Reciprocity of mutual coupling for TDD massive MIMO systems," in *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2015.

[20] J. Xiwen et al., "MIMO-TDD reciprocity under hardware imbalances: Experimental results," in *2015 IEEE International Conference on Communications ICC, 8-12 June 2015, London, United Kingdom*, London, U.K., 2015.

[21] R. Jedlicka, M. Poe, and K. Carver, "Measured mutual coupling between microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 29, no. 1, pp. 147–149, Jan 1981.

[22] J. Vieira et al., "A flexible 100-antenna testbed for Massive MIMO," in *IEEE GLOBECOM 2014 Workshop on Massive MIMO: from theory to practice*, Dec 2014.

[23] A. Hall, *Generalized Method of Moments*, ser. Advanced Texts in Econometrics. OUP Oxford, 2004.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, Feb. 2000.

[26] L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, ser. Addison-Wesley series in electrical and computer engineering. Addison-Wesley Publishing Company, 1991.

[27] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[28] S. Haykin, *Adaptive Filter Theory (3rd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[29] A. Molisch, *Wireless Communications*, ser. Wiley - IEEE. Wiley, 2010.

[30] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*, 1st ed. New York, NY, USA: Cambridge University Press, 2008.

[31] X. Luo, "Multi-User Massive MIMO Performance with Calibration Errors," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4521–4534, Jul 2016.

[32] W. Zhang et al., "Large-Scale Antenna Systems With UL/DL Hardware Mismatch: Achievable Rates Analysis and Calibration," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1216–1229, April 2015.

[33] T. Schenk, *RF Imperfections in High-rate Wireless Systems: Impact and Digital Compensation*. Springer, 2008.

[34] H. Van Trees, *Detection, Estimation, and Modulation Theory*. Wiley, 2004.

[35] W. Daniel, *Applied nonparametric statistics*, ser. The Duxbury advanced series in statistics and decision sciences. PWS-Kent Publ., 1990.

Paper V

A Receive/Transmit Calibration Technique based on Mutual Coupling for Massive MIMO Base Stations

This paper presents a calibration technique for massive MIMO base stations, where the frequency responses of the transmit and/or receive analog front-ends are individually estimated and compensated for. Calibration is achieved by a first-round of channel sounding between base station antennas, followed by post-processing and a compensation stage. The proposed technique is general in the sense that is does not use external sources, nor internal dedicated circuits for calibration purposes. The only requirement of the technique is that mutual coupling between all pairs of sounded base station antennas exists and is known. Our analysis suggests that mutual coupling can be conveniently used for calibration purposes, and that multipath propagation during calibration is the most prominent source for calibration inaccuracies.

1 Introduction

Massive multiple-input multiple output (MIMO) systems have attracted a lot of attention in the wireless research community. In massive MIMO, base stations (BSs) equipped with hundreds of antennas serve a relatively low number of terminals in the same time/frequency resource. This approach holds great promises in terms of energy efficiency, spectral efficiency, etc [1]. However, BSs operating with (very large) antenna arrays usually require some type of calibration to compensate for non-ideal characteristics in the system. These non-idealities are often related to hardware aspects that in theory are assumed ideal for sake of simplicity, but need to be compensated in real systems.

Research efforts to calibrate massive MIMO BSs in order to enable time division duplex (TDD) operation were presented in [2, 3]. These efforts enabled downlink precoding based on non-reciprocal uplink channel estimates. However uplink channel estimates can also be useful for other purposes. For example, the estimates can be used for real time positioning of terminals, for codebook based precoding, or other applications where explicit transmit/receive beamforming are necessary. Indeed, this can only be achieved if the transmit/receive radio-frequency (RF) chains and cables of the BS, are calibrated to yield aligned responses with respect to both phase and magnitude.

The problem setup addressed in this work is as follows. Consider a massive MIMO BS facing an open area, as a typical BS in cellular systems. We envision a calibration procedure that can conveniently be performed on-the-fly, by exchanging signals to-and-from all pairs of BS antennas without the need for external sources nor internal calibration circuits. The received signals are then processed in order to estimate and compensate for the differences between the analog front-ends and cables associated with different antennas. Channels between nearby BS antennas are strongly dominated by mutual coupling effects but nearby reflections can also contribute. No assumptions are made on the array structure nor on the propagation conditions, other than knowing the complex channel gains between all pairs of antennas due to mutual coupling. For a fixed array structure, these quantities can be measured once after array manufacturing and can be considered constant for an arbitrarily long period of time. Note that no user terminals are involved in the calibration process.

For the remainder of this paper, we address calibration as the process of estimating the transmitters/receivers frequency responses, since the compensation stage is straightforward. The measurements in [4] suggest that, with high-end analog components, these estimates are valid for time periods of at least one hour.

A method for transceiver calibration based on signals exchanged among antenna elements of the own array was proposed in [5]. Array shape estimation is performed in a first stage, followed by estimation of the transceiver responses. However, the method proposed in [5] rely on the assumption of spherical wave propagation between antenna elements, which is not suitable for our case due to the strong mutual coupling between BS antennas.

A transceiver calibration method using bi-directional mutual coupling based mea-

surements between adjacent antennas of a 2-dimensional array, was given in [6]. However, all circuitry was made of passive components, e.g. phase shifters and power dividers, yielding a reciprocal channel even including the hardware circuitry. This work generated attention regarding mutual coupling based calibration and spurred many publications, but none fitting the problem setup of this paper. To the best of the authors' knowledge, no calibration work addressing our particular setup is available in the literature.

The remainder of this paper is organized as follows: in Sec. 2 introduces the signal models. Sec. 3 presents our proposed estimator and the Cramér-Rao lower bound for the transceiver responses estimates. In Sec. 4 the estimator performance is accessed by means of numerical simulations, and finally Sec. 5 concludes the paper.

2 System models

2.1 Inter-radio model

Consider an M antenna array where BS transceiver calibration is performed on a flat bandwidth, e.g., a particular subcarrier in an OFDM system. Let for simplicity $x_p = 1$ be the transmitted signal for channel sounding purposes. All antennas are sounded 1-by-1. The vector with the measured forward and reverse channels between radio units n and m with $1 < n < M$, $1 < m < M$ and $m \neq n$ is modeled as

$$\boldsymbol{y}_{n,m} = \begin{bmatrix} y_{n,m} \\ y_{m,n} \end{bmatrix} = h_{n,m} \begin{bmatrix} r_n\, t_m \\ r_m\, t_n \end{bmatrix} + \begin{bmatrix} z_{n,m} \\ z_{m,n} \end{bmatrix}, \tag{1}$$

where the reciprocal propagation channels between antenna elements n and m are described by

$$h_{n,m} = c_{n,m} + \tilde{h}_{n,m}. \tag{2}$$

The term $c_{n,m}$ describes a deterministic and known component due to mutual coupling, which often is stronger for closely spaced antennas. The sum of the remaining multipath contributions are modeled by a zero-mean circularly-symmetric complex Gaussian (ZMCSCG) random variable $\tilde{h}_{n,m}$ with variance σ^2. The non-reciprocal receiver and transmitter frequency responses, that we want to estimate, are modeled by r_m and t_m, respectively, which materially map to the cascade of antenna responses, SMA cables, and all hardware circuitry in the analog front-end stage of the radios. Finally, independent and identically distributed (IID) ZMCSCG noise contributions $z_{n,m}$, each with variance N_0, are assumed.

A few remarks on the modeling assumptions of (1) follow: ***(i)*** The transceiver responses are modeled linearly, although it is well known that front-ends exhibit non-linear behavior in general. The non-linear effects occur mostly due to amplifiers operating close to their saturation point and can be modeled by a sum of one linear and other non-linear terms [7]. Two arguments can be pointed out to justify our pure linearity assumption, the first being that with well behaved amplifiers operating below the compression point, the linear term dominates over the other terms. Secondly, the main goal of the paper is to find a simple way to calibrate the transmitter

and receiver responses, and linear modeling simplifies the approach. ***(ii)*** For a fixed antenna array structure, both magnitude and phase of the coupling component $c_{n,m}$ are known. They are can measured once after antenna array manufacturing, and can be considered constant for an arbitrarily long period of time due to the time-invariant properties of the dielectric materials of the array. ***(iii)*** If $\tilde{h}_{n,m}$ is seen as a self-interference channel during calibration[22], then averaging over several realizations of (1) may not necessarily improve the quality of the observations. This is especially true in static scenarios where $\tilde{h}_{n,m}$ has a time-invariant behavior, if $\sigma^2 \gg N_0$. ***(iv)*** Uncorrelated scattering contributions between different antennas are assumed, i.e., $\mathrm{E}\left\{\tilde{h}_{n,m}\tilde{h}^*_{k,p}\right\} = \sigma^2\delta[n-k]\delta[m-p]$. This is a rule-of-thumb in wireless propagation for rich scattering environments for antenna spacings of $\frac{\lambda}{2}$ that we use as an approximation. ***(v)*** For generality purposes, no probabilistic models are assumed for t_m and r_m, i.e., they will be treated as deterministic but unknown parameters for estimation.

2.2 Modeling the coupling gains between antennas

To allow reproducibility of our simulation results later on, we now give a simple measurement based model for the coupling magnitudes $|c_{m,n}|$. The phases $\angle c_{m,n}$ are drawn from an uniform distribution between $-\pi$ and π. To model $|c_{n,m}|$ as a function of antenna spacing, the channel magnitudes between several pairs of antennas were measured in an anechoic chamber from a 2-dimensional 25x4 dual polarized patch antenna array with $\frac{\lambda}{2}$ spaced elements [4]. The frequency response magnitude over a 20 MHz bandwidth centered at 3.7 GHz - which the array is designed to operate at - was averaged. Fig. 1 shows the measured channel magnitudes as a function of antenna spacing. Different channel magnitudes for equidistant antennas occur due to the relative orientation of the respective antenna pair with respect to the measured E-field polarization. In general, antenna elements placed in the same orientation as the measured E-field polarization couple stronger than others [8]. A linear least square fit has been performed to model the coupling magnitude as a function of antenna distance.

3 Transmitter/Receiver Calibration

3.1 The Generalized Method of Moments estimator

Introduced originally for statistical inference in econometrics, the Generalized Method of Moments (GMM) is an estimation approach which exploits a particular structure of the signal model, more specifically the *moment conditions* [9]. In our case, a vector of moment conditions $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{t})$ that satisfies

$$\mathrm{E}\{\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{t})\} = \mathrm{E}\{\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\phi})\} = \mathbf{0} \tag{3}$$

[22] This will be seen later in the paper.

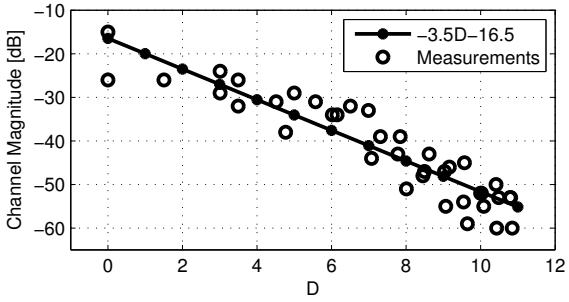

Figure 1: Measured channel magnitudes (circles) and respective linear regression (solid line). The horizontal axis variable $D = 10\log_{10}\left(\frac{d}{\lambda/2}\right)$ - with d denoting the physical distance between measured antenna pairs - represents the normalized antenna distance in terms of $\lambda/2$ in decibels. For example, the two measurements at $D = 0$ consist of channel magnitudes between two different pairs of adjacent $\lambda/2$ spaced antennas.

is required. In (3), $\boldsymbol{r} = [r_1 \cdots r_M]^T$, $\boldsymbol{t} = [t_1 \cdots t_M]^T$, $\boldsymbol{\phi} = \left[\boldsymbol{t}^T\ \boldsymbol{r}^T\right]^T$, $\mathbf{0}$ is an all zeros column vector, and

$$\boldsymbol{y} = \left[\boldsymbol{y}_{1,2}^T \ldots \boldsymbol{y}_{1,M}^T\ \boldsymbol{y}_{2,3}^T \cdots \boldsymbol{y}_{2,M}^T \cdots \boldsymbol{y}_{M-1,M}^T\right]^T. \tag{4}$$

Noting that all observations in (4) can also be paired as

$$\begin{bmatrix} y_{m,\ell} \\ y_{n,\ell} \end{bmatrix} = \begin{bmatrix} r_m h_{m,\ell} \\ r_n h_{n,\ell} \end{bmatrix} t_\ell + \begin{bmatrix} z_{m,\ell} \\ z_{n,\ell} \end{bmatrix}, \tag{5}$$

with $n \neq \ell \neq m$, inspection indicates a moment condition to be

$$\mathrm{E}\left\{f_{n,m,\ell}\right\} = \mathrm{E}\left\{y_{m,\ell} r_n c_{n,\ell} - y_{n,\ell} r_m c_{m,\ell}\right\} = 0\ \ \forall\, m, \ell, n. \tag{6}$$

A similar formulation of (5), where instead r_ℓ is the common factor, provides the moment condition

$$\mathrm{E}\left\{d_{n,m,\ell}\right\} = \mathrm{E}\left\{y_{\ell,m} c_{\ell,n} t_n - y_{\ell,n} c_{\ell,m} t_m\right\} = 0\ \forall\, m, \ell, n. \tag{7}$$

Stacking all useful[23] terms $f_{n,m,\ell}$ in $\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{r}) \in \mathbb{C}^{M(M-1)/2\times 1}$, and $d_{n,m,\ell}$ in $\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{t}) \in \mathbb{C}^{M(M-1)/2\times 1}$, and denoting $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\phi}) = \left[\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{r})^T\, \boldsymbol{d}(\boldsymbol{y}, \boldsymbol{t})^T\right]^T$, the GMM estimator is obtained by solving

$$\hat{\boldsymbol{\phi}} = \arg \min_{\substack{\boldsymbol{\phi} \\ ||\boldsymbol{t}||^2 = M \\ ||\boldsymbol{r}||^2 = M}} \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\phi})^H\, \hat{\boldsymbol{W}}\, \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\phi}), \tag{8}$$

[23] There are $M(M-1)$ moment conditions $f_{n,m,\ell}$, however half of them are negative counterparts of the other half and will not contribute for the final cost function.

where $\hat{\boldsymbol{W}}$ is a weighting matrix that generally needs to be optimized. Note the imposed constraint $||\boldsymbol{t}||^2 = ||\boldsymbol{r}||^2 = M$ which avoids the all-zero solution and normalizes the average energy per entry of $\hat{\boldsymbol{\phi}}$ to one. The solution to (4) can be found by standard numerical optimization methods. Newton's algorithm, or any other suitable method, is guaranteed to converge to the global optimum since the problem at hand is quadratic with quadratic constraints. Note that neither $\tilde{h}_{n,\ell}$ nor $\tilde{h}_{m,\ell}$ are included in any of the moment conditions, i.e. only the coupling components are included. Multipath propagation will thus show up as self-interference, as it will be seen later in Sec. 4.2.

We address now the choice of $\hat{\boldsymbol{W}}$. Most of the work within the GMM framework has been done under asymptotic assumptions, i.e., when an infinitely large record of finite signal-to-noise ratio (SNR) observations $\boldsymbol{y}$, or a finite record of infinite SNR observations is at hand. Under such conditions the optimal weighing matrix $\boldsymbol{W}_{\text{opt}}$ has a known form [9]. In our case, only one finite SNR observation is at hand. Depending on the SNR of such observation, the asymptotic regime under which $\boldsymbol{W}_{\text{opt}}$ was derived may not hold. Also, no estimators for $\boldsymbol{W}$ claiming any optimality criteria at the low SNR regime are available in the literature (at least to the best of the authors' knowledge). Claiming no optimality properties on our estimator, we set $\hat{\boldsymbol{W}} = \boldsymbol{I}$ for simplicity, and leave this optimization problem as future work. With $\hat{\boldsymbol{W}} = \boldsymbol{I}$, (4) can be seen as an instance of the Rayleigh quotient problem [10]. Also note that in a calibration setup resembling ours[24], the authors in [3] empirically defined an LS cost function based on their model's inherent structure, not being aware that it corresponds to the particular case of the GMM estimator with $\hat{\boldsymbol{W}} = \boldsymbol{I}$.

3.2 Cramér-Rao lower bound for the transmitters/receivers

In this section we compute the Cramér-Rao lower bound (CRLB) [11], a lower bound on the variance of any unbiased estimator, for $\hat{\boldsymbol{\phi}} = \left[\hat{\boldsymbol{t}}^T \hat{\boldsymbol{r}}^T\right]^T$. Assuming[25] $r_1 = t_1 = 1$, we denote the vector of real parameters

$$\boldsymbol{\theta} = [\text{Re}\{t_2\}\ \text{Im}\{t_2\}\ \text{Re}\{r_2\}\ \text{Im}\{r_2\}\ \text{Re}\{t_3\} \dots \text{Im}\{r_M\}]^T, \tag{9}$$

where $\text{Re}\{\}$ and $\text{Im}\{\}$ return real and imaginary part of their arguments, respectively. The CRLB is given by [11]

$$\begin{cases} CRLB(t_m) &= \left[\frac{\boldsymbol{\phi}_0}{\partial \boldsymbol{\theta}} \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\boldsymbol{\phi}_0^H}{\partial \boldsymbol{\theta}}\right]_{2m-1,2m-1} \\ CRLB(r_m) &= \left[\frac{\boldsymbol{\phi}_0}{\partial \boldsymbol{\theta}} \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\boldsymbol{\phi}_0^H}{\partial \boldsymbol{\theta}}\right]_{2m,2m} \end{cases} \tag{10}$$

where $\mathbf{I}(\boldsymbol{\theta}) \in \mathbb{C}^{(2M-2)\times(2M-2)}$ is the Fisher Information matrix of $\boldsymbol{\theta}$, and $\boldsymbol{\phi}_0 = [t_2\ r_2\ t_3\ r_3\ \cdots\ t_M\ r_M]^T$. Before computing $\mathbf{I}(\boldsymbol{\theta})$, note that the mean of (1) is given

[24] In their work, calibration between access points of a distributed MIMO system was performed. The estimated parameters were the ratios $\frac{r_m}{t_m}$, and no coupling between antennas was explored. Their moment conditions were also different. [25] See Sec. 3.3 for justification.

by

$$\boldsymbol{\mu}_{n,m} = \mathrm{E}\left\{\boldsymbol{y}_{n,m}\right\} = c_{n,m}\left[r_n t_m \; r_m t_n\right]^T, \tag{11}$$

and the covariance matrix of (1) is given by

$$\begin{aligned}\boldsymbol{\Sigma}_{n,m} &= \mathrm{Var}\left\{\boldsymbol{y}_{n,m}\boldsymbol{y}_{n,m}^H\right\} \\ &= \begin{bmatrix} |r_n|^2|t_m|^2\sigma^2 + N_0 & r_n t_m r_m^* t_n^* \sigma^2 \\ r_m t_n r_n^* t_m^* \sigma^2 & |r_m|^2|t_n|^2\sigma^2 + N_0 \end{bmatrix}. \end{aligned} \tag{12}$$

We can observe that the likelihood function for (4) is a multivariate Gaussian PDF, i.e., $p(\boldsymbol{y}|\boldsymbol{\theta}) \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_{1,2}^T \ldots \boldsymbol{\mu}_{1,M}^T \boldsymbol{\mu}_{2,3}^T \ldots \boldsymbol{\mu}_{2,M}^T \ldots \boldsymbol{\mu}_{M-1,M}^T\right]^T$ and block diagonal covariance

$$\boldsymbol{\Sigma} = \mathrm{diag}\left\{\boldsymbol{\Sigma}_{1,2}, \cdots, \boldsymbol{\Sigma}_{1,M}, \boldsymbol{\Sigma}_{2,3}, \cdots, \boldsymbol{\Sigma}_{2,M}, \cdots, \boldsymbol{\Sigma}_{M-1,M}\right\}. \tag{13}$$

For such likelihood form, the Fisher Information matrix entry at the ith row and jth column is given by

$$\begin{aligned}\left[\mathbf{I}(\boldsymbol{\theta})\right]_{i,j} = \; & \mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_j}\right\} \\ & + 2\;\mathrm{Re}\left\{\frac{\partial\boldsymbol{\mu}^H}{\partial\theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_j}\right\}. \end{aligned} \tag{14}$$

Note that i and j can also be expressed as $i = 4(m-1) + K_m$ and $j = 4(n-1) + K_n$, with $1 < K_m < 4$ and $1 < K_n < 4$.

Due to the symmetric property of $\mathbf{I}(\boldsymbol{\theta})$ and thus assuming $j \geq i$, (14) can be written as

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{i,j} = \begin{cases} A_1 + A_2, & m \neq n \\ B_1 + B_2, & m = n \end{cases} \tag{15}$$

with

$$
\begin{aligned}
A_1 &= \operatorname{Tr}\left\{ \boldsymbol{\Sigma}_{m,n}^{-1} \frac{\partial \boldsymbol{\Sigma}_{m,n}}{\partial \theta_i} \boldsymbol{\Sigma}_{m,n}^{-1} \frac{\partial \boldsymbol{\Sigma}_{m,n}}{\partial \theta_j} \right\}, \\
B_1 &= \sum_{\ell > m}^{M} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}_{m,\ell}^{-1} \frac{\partial \boldsymbol{\Sigma}_{m,\ell}}{\partial \theta_i} \boldsymbol{\Sigma}_{m,\ell}^{-1} \frac{\partial \boldsymbol{\Sigma}_{m,\ell}}{\partial \theta_j} \right\} \\
&\quad + \sum_{\ell=1}^{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}_{\ell,m}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\ell,m}}{\partial \theta_i} \boldsymbol{\Sigma}_{\ell,m}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\ell,m}}{\partial \theta_j} \right\}, \\
A_2 &= 2 \operatorname{Re}\left\{ \frac{\partial \boldsymbol{\mu}_{m,n}^H}{\partial \theta_i} \boldsymbol{\Sigma}_{m,n}^{-1} \frac{\partial \boldsymbol{\mu}_{m,n}}{\partial \theta_j} \right\}, \\
B_2 &= 2 \operatorname{Re}\left\{ \sum_{\ell > m}^{M} \frac{\partial \boldsymbol{\mu}_{m,\ell}^H}{\partial \theta_i} \boldsymbol{\Sigma}_{m,\ell}^{-1} \frac{\partial \boldsymbol{\mu}_{m,\ell}}{\partial \theta_j} \right\} \\
&\quad + 2 \operatorname{Re}\left\{ \sum_{\ell=1}^{m} \frac{\partial \boldsymbol{\mu}_{\ell,m}^H}{\partial \theta_i} \boldsymbol{\Sigma}_{\ell,m}^{-1} \frac{\partial \boldsymbol{\mu}_{\ell,m}}{\partial \theta_j} \right\}.
\end{aligned}
$$

An example of a derivative of (12) with respect to an entry of $\boldsymbol{\theta}$ is given in the appendix.

Note the general block diagonal structure of $\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$, e.g., for $\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_7}$ (where $m = 2$ and $K_m = 3$) we have

$$
\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_7} = \operatorname{diag}\left\{ \frac{\partial \boldsymbol{\Sigma}_{1,2}}{\partial \theta_7}, \boldsymbol{\emptyset}, \cdots, \boldsymbol{\emptyset}, \frac{\partial \boldsymbol{\Sigma}_{2,3}}{\partial \theta_7}, \cdots, \frac{\partial \boldsymbol{\Sigma}_{2,M}}{\partial \theta_7}, \boldsymbol{\emptyset}, \cdots \right\} \tag{16}
$$

where $\boldsymbol{\emptyset}$ is an all zero matrix. If $m \neq n$, there is only one matrix entry where the two block diagonal matrices $\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$ and $\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$ are non-zero. If $m = n$, then M of such matrix entries are shared which explains the summations in A_2 and B_2. To finalize, note that both N_0 and σ^2 were assumed to be known in the CRLB computation.

3.3 A Reference Element for Calibration

The CRLB computations for $\hat{t}_m$ and $\hat{r}_m$ made use of the concept of a reference antenna element for calibration purposes, see [2]. That is, since calibration can be performed up to a common complex factor among the transceivers, a convenient approach is to assume $r_1 = t_1 = 1$ where such transceiver is considered a reference[26]. The CRLB for the remaining elements are calculated accordingly. This ensures invertibility for the Fisher Information matrix. However, note that setting $r_1 = t_1 = 1$ needs not to be strictly met to derive an estimator, as done in Sec. 3.1.

[26] More generally, one can assume $r_m = q_1$ and $t_m = q_2$ with $\mathbf{0} \neq [q_1 \; q_2]^T \in \mathbb{C}^{2\times 1}$.

4 Performance Assessment

4.1 Simulation Setup

The main focus of our analysis contrasts the GMM estimator mean square error (MSE) against the CRLB. We define the following MSE metric for the GMM estimator performance:

$$\begin{aligned} MSE(r_m) &= \left[\, \mathrm{E}\left\{ |\boldsymbol{r} - \hat{\boldsymbol{r}}/\hat{r}_1|^2 \right\} \,\right]_m \\ MSE(t_m) &= \left[\, \mathrm{E}\left\{ |\boldsymbol{t} - \hat{\boldsymbol{t}}/\hat{t}_1|^2 \right\} \,\right]_m . \end{aligned} \tag{17}$$

This MSE definition is coherent with the reference element concept used in the CRLB calculations. Moreover, our numerical simulations for the GMM estimator and CRLB computations were all performed with $r_m = t_m = 1 \ \forall \ m$. The constraint in (4), namely $||\boldsymbol{t}||^2 = ||\boldsymbol{r}||^2 = M$, ensures that

$$\hat{\boldsymbol{\phi}}_0 \xrightarrow{p} \boldsymbol{\phi}_0 \ \text{ as } \begin{cases} N_0 \to 0 \\ \sigma^2 \to 0 \end{cases} \tag{18}$$

due to the asymptotically unbiasedness property of the GMM estimator [9]. The notation $\xrightarrow{p}$ denotes convergence in probability. The unbiasedness property in (18), and the MSE definition in (17), allow a coherent comparison between the estimator MSE and the CRLB. Also, having $r_m = t_m = 1$ makes MSE = 0 dB a reference point in the analysis.

Naturally, we select the reference element as one of the most central antenna elements of the 2-dimensional array. In average, this choice yields the strongest coupling channels to all other antenna elements of the array.

We limit our analysis to two extreme cases, namely, calibration of a transceiver which is associated to an adjacent (or neighbor) antenna element to the reference element, and the case of calibration of a transceiver that is associated with one of the antennas at the four corners (edges) of the 2D array.

Note that all upcoming results, and evaluated range of N_0 and σ^2, should be considered together with the derived coupling model in Fig. 1 where the strongest coupling channel, i.e. neighbor channel, yields a -16.5 dB gain.

4.2 Results and Analysis

The symmetry of the model in (1) for r_m and t_m together with the simulations settings adopted, makes the statistical performance of $\hat{r}_m$ similar to $\hat{t}_m$. Thus, our analysis holds for both cases.

Fig. 2 shows the CRLB for the transceiver associated with a neighbor antenna with respect to the reference. Overall, this lower bound of the variance decreases as M increases due to the increasing number of signal alternatives available to estimate the same parameter. At low enough N_0, the CRLB flattens out which happens due to the self-interference term $\tilde{h}_{m,n}$.

The CRLB for the edge antenna transceivers is shown in Fig. 3. An increase of M also decreases the CRLB as in the neighbor case. An interesting fact, since, although a larger signal set is available for estimation purposes, the distance (and hence the path loss) with respect to the reference element also increases. Compared to the neighbor case, a performance degradation of up to 4 dB is to be expected, as can be seen in Fig. 6 in Appendix B.

Fig. 4 and 5 show the MSE performance of the GMM estimator, and contrast this with the CRLB for $M = 100$ antenna elements. For the neighbor antenna case, performance very close to the theoretical optimum is achieved for some parameter regions, e.g. $N_0 \approx -35$ dB and $\sigma^2 = -50$ dB for $M = 36$ and $M = 100$.

Since we set $t_m = r_m = 1\ \forall m$ in our simulations, increasing N_0 or σ^2 by the same amount is equally degrading for the performance of our estimator. For example when $M = 400$ in Fig. 4, an MSE of -27 dB is obtained for $\sigma^2 = N_0 = -50$ dB, and increasing σ^2 or N_0 by 10 dB results in an MSE of -15 dB. It is worthwhile mentioning that in a practical scenario, increasing the transmit power beyond a certain level is not beneficial to improve the quality of the observations, since $\tilde{h}_{n,m}$ multiplies the transmit signal. This aspect together with a possible time-invariant behavior of $\tilde{h}_{n,m}$, which results in no quality improvements by averaging, makes multipath propagation a prominent source for calibration inaccuracies.

Overall, GMM calibration for neighboring antennas is more accurate than for edge antennas, as one would expect. However some calibration cases indicate that performance degradation seems to always increase with M. This is particularly true for the edge antenna element, contrarily with one would expect simple by analyzing their CRLBs for different M. The sub-optimality of our estimator may justify this behavior.

5 Conclusions

In this paper, a transceiver calibration technique using mutual coupling between antenna elements of a massive MIMO base station was proposed. For a 2-dimensional antenna array of a given size, transceivers associated with antennas at the edge of the array are the hardest to calibrate. Moreover, the results from our proposed estimator indicate that the calibration error associated with these transceivers grows for arrays with increasing number of antennas. This implies that in practice, stricter calibration requirements are needed to calibrate bigger arrays, while still maintaining the same error criterion. This is however contrary to what the CRLB indicates.

Overall, the proposed calibration method does not suffer from convergence problems, and it is of practical use, given that the mutual coupling gains between BS antennas are known, and multipath contributions during calibration are small compared to mutual coupling effects.

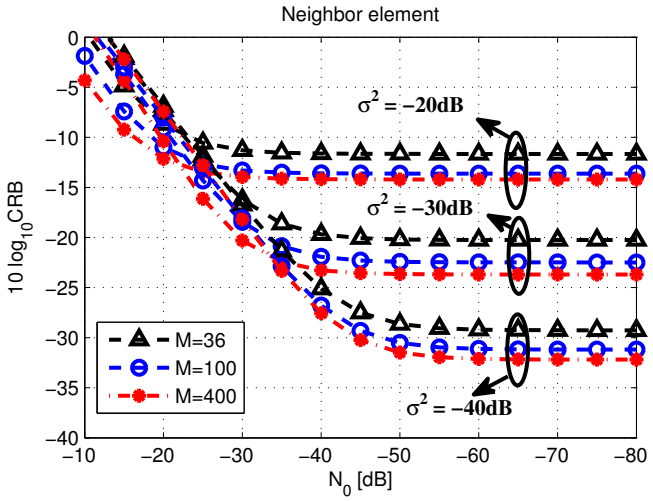

Figure 2: CRLB vs N_0 for the neighbor antenna element, for different number of base station antennas M, and parameter of the Rayleigh channel σ.

Acknowledgments

The research leading to these results has been funded by grants from the Swedish foundation for strategic research SSF, the Swedish research council, the Excellence center at Linkoping - Lund in Information Technology, and the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 619086 (MAMMOET).

6 Appendix A

Fig. 6 shows the difference between the CRLB of a transceiver estimate associated with edge antennas, and transceiver estimate associated with neighbor antennas to the reference antenna.

7 Appendix B

An example of a derivative of (12) with respect to an entry of $\boldsymbol{\theta}$, namely $\theta_{4(k-1)+K_m}$ with $K_m = 2$ given by

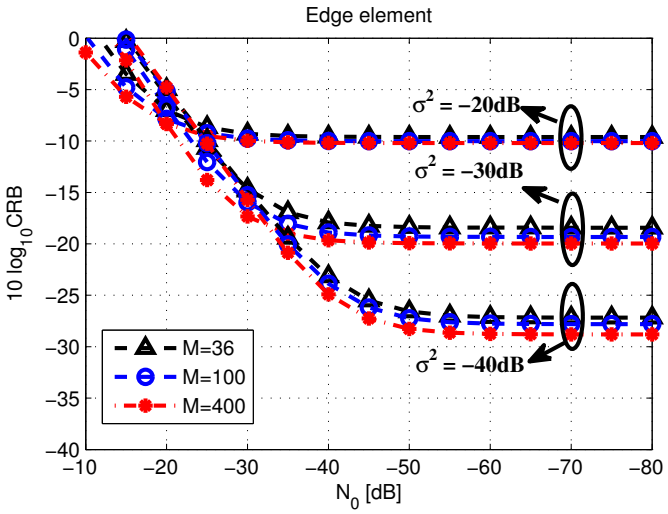

Figure 3: CRLB vs N_0 for the edge antenna element, for different number of base station antennas M, and parameter of the Rayleigh channel σ.

$$\frac{\partial \Sigma_{i,j}}{\partial \operatorname{Im}\{t_k\}} = \begin{cases} \begin{bmatrix} 0 & -jr_i t_j r_j^* \sigma^2 \\ jr_i^* t_j^* r_j \sigma^2 & 2\operatorname{Im}\{t_i\} |r_j|^2 \sigma^2 \end{bmatrix}, & k = i \\ \begin{bmatrix} 2\operatorname{Im}\{t_j\} |r_i|^2 \sigma^2 & jr_i r_j^* t_i^* \sigma^2 \\ -jr_j t_i r_i^* \sigma^2 & 0 \end{bmatrix}, & k = j \\ \mathbf{Ø}, & \text{otherwise.} \end{cases}$$

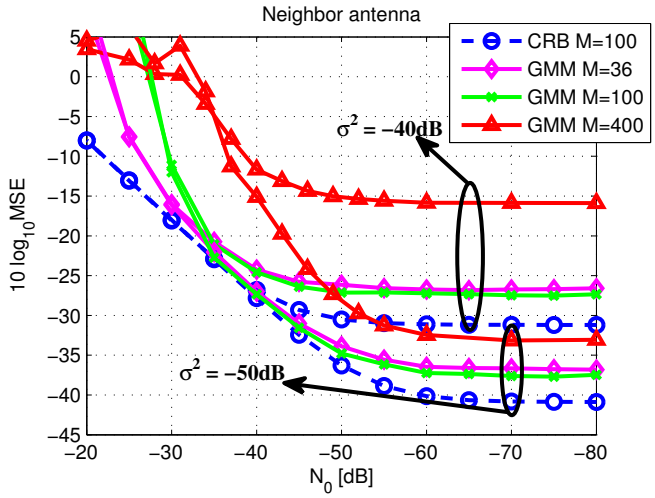

Figure 4: MSE of the GMM estimator vs N_0 for the neighbor antenna element, for different number of base station antennas M, and parameter of the Rayleigh channel σ.

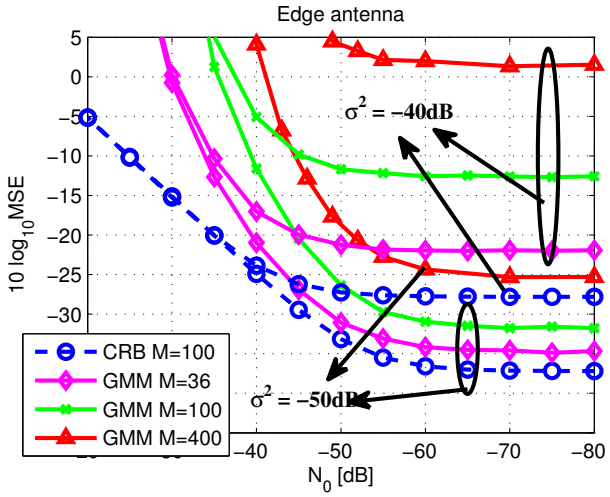

Figure 5: MSE of the GMM estimator vs N_0 for the edge antenna element, for different number of base station antennas M, and parameter of the Rayleigh channel σ.

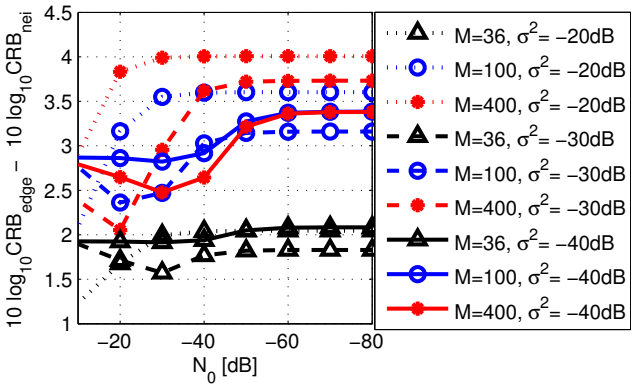

Figure 6: Difference between the CRLB of a transceiver estimate associated with edge antennas, and transceiver estimate associated with neighbor antennas to the reference antenna, for different number of BS antennas M, and squared parameter of the Rayleigh channel σ^2.

References

[1] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, February 2014.

[2] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 53–64.

[3] R. Rogalin, O. Bursalioglu, H. Papadopoulos, G. Caire, A. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 4, pp. 1815–1831, April 2014.

[4] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors and F. Tufvesson, "A flexible 100-antenna testbed for Massive MIMO," in *IEEE GLOBECOM 2014 Workshop on Massive MIMO: from theory to practice, 2014-12-08*. IEEE, 2014.

[5] M. Willerton, "Array auto-calibration," Ph.D. dissertation, Imperial College London, U.K., 2013.

[6] H. Aumann, A. Fenn, and F. Willwerth, "Phased array antenna calibration and pattern prediction using mutual coupling measurements," *Antennas and Propagation, IEEE Transactions on*, vol. 37, no. 7, pp. 844–850, Jul 1989.

[7] F. Ghannouchi and O. Hammi, "Behavioral modeling and predistortion," *Microwave Magazine, IEEE*, vol. 10, no. 7, pp. 52–64, Dec 2009.

[8] R. Jedlicka, M. Poe, and K. Carver, "Measured mutual coupling between microstrip antennas," *Antennas and Propagation, IEEE Transactions on*, vol. 29, no. 1, pp. 147–149, Jan 1981.

[9] A. Hall, *Generalized Method of Moments*, ser. Advanced Texts in Econometrics. OUP Oxford, 2004.

[10] P. Lancaster and M. Tismenetsky, *The Theory of Matrices: With Applications*, ser. Computer Science and Scientific Computing Series. Academic Press, 1985.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

Paper VI

A Generalized Method of Moments Detector for Block Fading SIMO Channels

In this letter we apply the Generalized Method of Moments (GMM), widely used in econometrics, to receivers operating with imperfect channel state information (CSI) of single-input-multiple-output (SIMO) block-fading channels where a single pilot symbol is used. The GMM results in the standard maximum ratio combining (MRC) receiver, but with an improved channel estimate. Although not our goal at the outset, this result reveals an inherent capability of the GMM to improve *any* channel estimate through filtering of the initial channel estimate with a matrix that is constructed from the received signals. The filtering involves a matrix inverse of size $\min\{T, M\}$, where M is the number of receive antennas, and $T + 1$ is the coherence interval of the channel. The gain over an MRC receiver, using a scaled version of the pilot observation as channel estimate, lies in the range 0.1-3 dB depending on the system configuration. A coherence interval of about 5 symbol intervals is sufficient to reach these gains.

1 Introduction

We study receiver design for Single-input Multiple-output (SIMO) transmissions in block fading channels in the face of additive Gaussian noise where the channel is a-priori unknown. Such systems have been widely studied and the interplay between the amount of training and payload data is today well understood [1,2]. In this paper we allocate one time slot of the coherence interval to training data. The literature on possible receiver designs is vast. The Generalized Likelihood Ratio Test optimal detector has been studied in [3] and has cubic complexity in the block length $T+1$. If the statistics of the channel are known, then the true Maximum-Likelihood (ML) detector operates on the basis of the conditional probability but has exponential complexity in $T+1$ [4]. Another method is iterative joint channel estimation and data detection [5] which can approach the performance of the ML detector but with much less computational complexity.

In this paper we investigate to what extent the Generalized Method of Moments (GMM) can be utilized for SIMO receiver design with imperfect channel state information (CSI). The GMM is a relatively new method that has been widely and successfully applied within econometrics [6]. However, applications of the GMM to the field of communications are scarce. To the best of the authors' knowledge, it was first applied, with great success, in [7] for reciprocity calibration of base stations in a large-scale multiuser Multiple-input Multiple-output (MIMO) setup. Noteworthy, [7] rediscovered the GMM as it did not identify the method as an instance of the GMM.

The contributions of this letter are as follows.

- We apply the GMM to SIMO systems with imperfect CSI and investigate its potential for receiver design.
- We find that the GMM improves the quality of *any* channel estimate by filtering it with a matrix that is constructed explicitly from the received signals.
- We show that the GMM receiver performs maximum ratio combining (MRC) using the improved channel estimate.
- We show that the complexity overhead compared with an MRC receiver that forms a channel estimate based on the pilot observation is small.

Altogether, the GMM detector has low complexity, yields signal-to-noise ratio (SNR) gains, and can be expressed in an easily understandable manner. Our current GMM detector cannot easily be extended to MIMO systems; we mention where the problem for MIMO occurs later. For a single-input single-output systems GMM detection does not offer any gain.

2 System Model

We consider a block fading SIMO system with coherence time $T+1$ symbol times in additive Gaussian noise. The received signal vector at time k can be expressed as

$$\boldsymbol{y}_k = \boldsymbol{h}x_k + \boldsymbol{n}_k, \quad 0 \leq k \leq T \tag{1}$$

where $x_0 = 1$ is a known inserted training symbol, $\{x_k\}_{k=1}^{T}$ are multi-level QAM data symbols with unit average energy, $\boldsymbol{h}$ is a random $M \times 1$ vector representing the communication channel, and $\boldsymbol{n}_k$ is complex Gaussian noise with covariance matrix $N_0\boldsymbol{I}$. The vector $\boldsymbol{h}$ is unknown to the receiver, and we also assume that the receiver does not know the joint density of the entries of $\boldsymbol{h}$. For later use, we assemble the vectors $\boldsymbol{y}_k$ into a matrix $\tilde{\boldsymbol{Y}} = [\boldsymbol{y}_0\ \boldsymbol{Y}]$ where $\boldsymbol{Y} = [\boldsymbol{y}_1\ \ldots\ \boldsymbol{y}_T]\ \mathbb{C}^{M\times T}$ is containing the data observations. Similarly, $\boldsymbol{x}$ denotes the vector of data symbols $\boldsymbol{x} = [x_1\ \ldots\ x_T]$ and $\tilde{\boldsymbol{x}} = [1\ \boldsymbol{x}]$. With that, we have $\boldsymbol{Y} = \boldsymbol{h}\boldsymbol{x} + \boldsymbol{N}$ where $\boldsymbol{N} = [\boldsymbol{n}_1\ \ldots\ \boldsymbol{n}_T]$.

We analyze two corner cases: ***(i)*** detection with known SNR $\gamma = \mathbb{E}[\|\boldsymbol{h}\|^2]/MN_0$ (measured per receive antenna)[27], and ***(ii)*** detection with unknown SNR. The former case corresponds to a scenario where the statistics of the channel do not change among the transmitted blocks, such that the SNR can be estimated with zero asymptotic error. The latter corresponds to a case where the transmitted blocks are sparse so that the SNR changes abruptly between two blocks. This occurs, e.g., often in machine-to-machine communications, where single antenna nodes transmit a small packet of control data at a very low periodicity [8]. Indeed, if the SNR is not known to the receiver, it can be estimated from the received data block. This is, however, out of scope for this letter, and we assume the SNR to be either unknown or perfectly known, which represents the two corner cases.

2.1 Benchmark Detectors

A standard way to implement a detector for signals of the form (1) is to estimate the channel as $\hat{\boldsymbol{h}} = \beta\boldsymbol{y}_0$. In the case that the SNR is known, then $\beta = \gamma(\gamma+1)^{-1}$ so that minimum mean square error (MMSE) channel estimation is obtained. If the SNR is not known, a least squares (LS) estimation is performed which implies that $\beta = 1$. Once the channel has been estimated, the detector proceeds as if the estimate is correct and performs MRC, i.e.

$$\hat{\boldsymbol{x}} = \frac{1}{\|\hat{\boldsymbol{h}}\|^2}\hat{\boldsymbol{h}}^{\mathrm{H}}\boldsymbol{Y}. \tag{2}$$

3 The Generalized Method of Moment Detector

The GMM explores a particular structure of the system model, more specifically the *moment conditions* [6]. Inspection of (1) indicates the moment condition

$$\mathbb{E}[\boldsymbol{f}_{k,\ell}] \triangleq \mathbb{E}[\boldsymbol{y}_k x_\ell - \boldsymbol{y}_\ell x_k] = \boldsymbol{0}, \tag{3}$$

where $0 \le k < \ell \le T$, and $\boldsymbol{0}$ has zeros in all of its entries. For MIMO, a condition similar to (3) is not available which limits the application of the GMM to SIMO.

[27] Here, $\mathbb{E}[\cdot]$ is the expectation operator, and $\|\cdot\|$ is the Frobenius norm.

With $\boldsymbol{g} \triangleq \left[\boldsymbol{f}_{0,1}^{\mathrm{T}} \cdots \boldsymbol{f}_{T-1,T}^{\mathrm{T}}\right]^{\mathrm{T}}$, the GMM detector is given by

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \boldsymbol{g}^H \boldsymbol{W} \boldsymbol{g}. \tag{4}$$

Since the optimal form of the weighting matrix $\boldsymbol{W}$ depends on the unknown second-order statistics of $\boldsymbol{h}$ [6], we proceed with $\boldsymbol{W} = \boldsymbol{I}$. The cost function can thus be rewritten as

$$\begin{aligned} \boldsymbol{g}^H \boldsymbol{g} &= \sum_{k=0}^{T} \sum_{\ell=k+1}^{T} \|\boldsymbol{y}_k x_\ell - \boldsymbol{y}_\ell x_k\|^2 \\ &= \boldsymbol{x}\boldsymbol{\Psi}\boldsymbol{x}^{\mathrm{H}} - 2\Re\{\boldsymbol{b}\boldsymbol{x}^{\mathrm{H}}\} + \|\boldsymbol{Y}\|^2, \end{aligned} \tag{5}$$

where $\boldsymbol{\Psi} = \|\tilde{\boldsymbol{Y}}\|^2 \boldsymbol{I} - \boldsymbol{Y}^{\mathrm{H}}\boldsymbol{Y}$ and $\boldsymbol{b} = \boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{Y}$. The minimizer to the quadratic form (5) is

$$\begin{aligned} \hat{\boldsymbol{x}} &= \boldsymbol{b}\boldsymbol{\Psi}^{-1} \\ &= \frac{1}{\|\tilde{\boldsymbol{Y}}\|^2} \boldsymbol{y}_0^{\mathrm{H}} \boldsymbol{Y} \left[\boldsymbol{I} - \frac{\boldsymbol{Y}^{\mathrm{H}}\boldsymbol{Y}}{\|\tilde{\boldsymbol{Y}}\|^2}\right]^{-1}. \end{aligned} \tag{6}$$

As can be seen from (6), the GMM detector operates over the entire block $\boldsymbol{Y}$. It is preferable to reach an expression where the GMM detector operates over one received vector per time. To get such a form, we rewrite (6) into

$$\begin{aligned} \hat{\boldsymbol{x}} &= \frac{1}{\|\tilde{\boldsymbol{Y}}\|^2} \boldsymbol{y}_0^{\mathrm{H}} \left[\boldsymbol{I} - \frac{\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{H}}}{\|\tilde{\boldsymbol{Y}}\|^2}\right]^{-1} \boldsymbol{Y} \\ &= \boldsymbol{y}_0^{\mathrm{H}} \boldsymbol{E} \boldsymbol{Y}, \end{aligned} \tag{7}$$

where

$$\boldsymbol{E} = \frac{1}{\|\tilde{\boldsymbol{Y}}\|^2} \left[\boldsymbol{I} - \frac{\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{H}}}{\|\tilde{\boldsymbol{Y}}\|^2}\right]^{-1}.$$

The particular form of the GMM detector (7) can, however, be computationally overwhelming if $M > T$, since the inversion required to establish $\boldsymbol{E}$ is of size $M \times M$. As a remedy we use the matrix inversion lemma and rewrite (7) as

$$\begin{aligned} \hat{\boldsymbol{x}} &= \frac{1}{\|\tilde{\boldsymbol{Y}}\|^2} \boldsymbol{y}_0^{\mathrm{H}} \left[\boldsymbol{I} + \frac{\boldsymbol{Y}}{\|\tilde{\boldsymbol{Y}}\|} \left[\boldsymbol{I} - \frac{\boldsymbol{Y}^{\mathrm{H}}\boldsymbol{Y}}{\|\tilde{\boldsymbol{Y}}\|^2}\right]^{-1} \frac{\boldsymbol{Y}^{\mathrm{H}}}{\|\tilde{\boldsymbol{Y}}\|}\right] \boldsymbol{Y} \\ &= \boldsymbol{y}_0^{\mathrm{H}} \boldsymbol{E} \boldsymbol{Y}, \end{aligned} \tag{8}$$

where in this case we have

$$\boldsymbol{E} = \frac{1}{\|\tilde{\boldsymbol{Y}}\|^2} \left[\boldsymbol{I} + \frac{\boldsymbol{Y}}{\|\tilde{\boldsymbol{Y}}\|} \left[\boldsymbol{I} - \frac{\boldsymbol{Y}^{\mathrm{H}}\boldsymbol{Y}}{\|\tilde{\boldsymbol{Y}}\|^2}\right]^{-1} \frac{\boldsymbol{Y}^{\mathrm{H}}}{\|\tilde{\boldsymbol{Y}}\|}\right],$$

so that only a $T \times T$ inversion is needed.

Let us now discuss the interpretation of the GMM detector $\hat{\boldsymbol{x}} = \boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{E}\boldsymbol{Y}$. The benchmark (2) applies MRC based on $\hat{\boldsymbol{h}}$. With unknown SNR, $\hat{\boldsymbol{h}}$ coincides with $\boldsymbol{y}_0$ so that the benchmark detector reads, omitting the scaling term in (2) for now, $\boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{Y}$. In view of this, we can see that the GMM detector $\hat{\boldsymbol{x}} = \boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{E}\boldsymbol{Y}$ takes the LS estimate $\boldsymbol{y}_0$, *purifies* it through the matrix $\boldsymbol{E}$, and then applies MRC, i.e.

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{h}}_{\mathrm{GMM}}^{\mathrm{H}}\boldsymbol{Y} = (\boldsymbol{E}\hat{\boldsymbol{h}})^{\mathrm{H}}\boldsymbol{Y}.$$

Thus, the GMM provides a simple method for improving the LS channel estimate. In the next section we will draw on this and show that the matrix $\boldsymbol{E}$ in fact improves any channel estimate, not only the LS one, especially as T grows large. The purification of the channel estimate is done once per received block. The additional complexity compared to the benchmark detector lies in computing the matrix $\boldsymbol{E}$ and the vector $(\boldsymbol{E}\hat{\boldsymbol{h}})^{\mathrm{H}}$ once per coherence time. Thus the complexity is dominated by a matrix inversion of size $\min\{T, M\} \times \min\{T, M\}$. The GMM detector $\hat{\boldsymbol{x}} = \boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{E}\boldsymbol{Y}$ is linear in $\boldsymbol{Y}$ once $\boldsymbol{E}$ has been computed. An important remark is that for $T = 1$, the GMM detector coincides with MRC based on $\hat{\boldsymbol{h}}_{\mathrm{LS}}$ and thus provides no additional benefit.

Finally, note that the matrix $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{H}}/\|\tilde{\boldsymbol{Y}}\|^2$ is not the sample covariance matrix of $\boldsymbol{Y}$ as it is normalized with its own energy, rather than with the block length. A problem with the GMM detector is that as the block length $T + 1$ grows large, the matrix $\boldsymbol{E}$ converges to the all zero matrix. In other words, it is a biased estimator. For constant modulus constellations this is not an issue, but it quickly renders the GMM detector unsuitable for multi-level QAM constellations. Before discussing the bias, we turn to an SNR analysis for which the bias effect is irrelevant.

4 SNR Asymptotic Analysis

In this section we analyze the asymptotic SNR of the GMM detector as the block length grow large. As scalings are irrelevant for SNR computations, we ignore the scaling of the matrix $\boldsymbol{E}$, so that we let the GMM detector for x_k be $\hat{x}_k = \boldsymbol{y}_0^{\mathrm{H}}(\boldsymbol{I} - \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{H}}/\|\tilde{\boldsymbol{Y}}\|^2)^{-1}\boldsymbol{y}_k$. However, recalling that $\boldsymbol{y}_0$ is the LS channel estimate of $\boldsymbol{h}$, we here replace $\boldsymbol{y}_0$ with any estimate $\hat{\boldsymbol{h}}$ that has the form $\hat{\boldsymbol{h}} = \alpha\boldsymbol{h} + \boldsymbol{w}$ where $\boldsymbol{w}$ is complex Gaussian noise with covariance matrix $N_{\mathrm{w}}\boldsymbol{I}$. We will now show that the matrix $\boldsymbol{E}$ will improve *any* channel estimate regardless of N_{w} and α. The LS and MMSE channel estimates are then the special cases $N_{\mathrm{w}} = \alpha^2 N_0$, with $\alpha = 1$ and $\alpha = \gamma(\gamma + 1)^{-1}$, respectively.

The post-processing SNR of the GMM detector is denoted by γ_{GMM} and is defined as follows. Define the random variable $z_k = \hat{x}_k|(x_k = 1)$. The SNR is then defined as $\gamma_{\mathrm{GMM}} = 2|\mathbb{E}[z_k]|^2/\mathrm{Var}[z_k]$, where $\mathrm{Var}[\cdot]$ denotes the variance of its input. We summarize our findings for the asymptotic post processing SNR $\gamma_{\mathrm{GMM}}^{\infty} = \lim_{T\to\infty}\gamma_{\mathrm{GMM}}$ in Proposition 1.

Proposition 1 *As $T \to \infty$, the post-processing SNR of the GMM detector $\hat{\boldsymbol{x}} =$*

$(\boldsymbol{E}\hat{\boldsymbol{h}})^{\mathrm{H}}\boldsymbol{Y}$, *is*

$$\gamma_{\mathrm{GMM}}^{\infty} = \frac{\alpha^2\lambda^2}{\alpha\lambda(N_0+N_{\mathrm{w}})+N_0N_{\mathrm{w}}+\frac{N_0^3N_{\mathrm{w}}(M-1)^3}{(\lambda+(M-1)N_0)^2}}.$$

Furthermore, the post-processing SNR of the benchmark detector $\hat{\boldsymbol{x}} = \hat{\boldsymbol{h}}^{\mathrm{H}}\boldsymbol{Y}$, γ_{BM}, *satisfies* $\gamma_{\mathrm{BM}} < \gamma_{\mathrm{GMM}}^{\infty}$.

Proof. Let $\boldsymbol{Q}[\sqrt{\lambda}\ 0\ \ldots\ 0]^{\mathrm{T}}$ denote the singular value decompostion of the channel vector $\boldsymbol{h}$. As $T\to\infty$, we have that $(\boldsymbol{I}-\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{H}}/\|\tilde{\boldsymbol{Y}}\|^2)^{-1}\to\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\mathrm{H}}$ where

$$\boldsymbol{D}=\mathrm{diag}\left(\left[\frac{\lambda+MN_0}{(M-1)N_0}\ \frac{\lambda+MN_0}{\lambda+(M-1)N_0}\cdots\frac{\lambda+MN_0}{\lambda+(M-1)N_0}\right]\right).$$

We now have,

$$\begin{aligned}\hat{x}_k &= ([\alpha\sqrt{\lambda}\,\mathbf{0}]\boldsymbol{Q}^{\mathrm{H}}+\boldsymbol{w}^{\mathrm{H}})\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\mathrm{H}}(\boldsymbol{Q}[\alpha\sqrt{\lambda}\,\mathbf{0}]^{\mathrm{T}}x_k+\boldsymbol{n}_k)\\ &= \lambda\frac{\lambda+MN_0}{(M-1)N_0}x_k+\eta, \qquad (9)\end{aligned}$$

where η is a noise variable and $\mathbf{0}$ is a $1\times(M-1)$ vector with all elements equal to zero. Based on (9) it is straightforward to derive the variance of η. By doing so, the SNR equals

$$\gamma_{\mathrm{GMM}}^{\infty} = \frac{\alpha^2\lambda^2}{\alpha\lambda(N_0+N_{\mathrm{w}})+N_0N_{\mathrm{w}}+\frac{N_0^3N_{\mathrm{w}}(M-1)^3}{(\lambda+(M-1)N_0)^2}}.$$

For the benchmark detector $\hat{x}_k=\hat{\boldsymbol{h}}^{\mathrm{H}}\boldsymbol{y}_k$ we obtain the SNR by replacing $\boldsymbol{D}$ with the identity matrix. This gives,

$$\gamma_{\mathrm{BM}} = \frac{\alpha^2\lambda^2}{\alpha\lambda(N_0+N_{\mathrm{w}})+MN_0N_{\mathrm{w}}}.$$

The relation $\gamma_{\mathrm{GMM}}^{\infty}>\gamma_{\mathrm{BM}}$ is shown as follows,

$$\begin{aligned}\gamma_{\mathrm{GMM}}^{\infty} &> \frac{\alpha^2\lambda^2}{\alpha\lambda(N_0+N_{\mathrm{w}})+N_0N_{\mathrm{w}}+\frac{N_0^3N_{\mathrm{w}}(M-1)^3}{(0+(M-1)N_0)^2}}\\ &= \gamma_{\mathrm{BM}}.\end{aligned}$$

■

From the proof of Proposition 1 we can observe that as $N_0\to 0$, we have asymptotically no SNR gain[28], i.e., $\gamma_{\mathrm{GMM}}^{\infty}/\gamma_{\mathrm{BM}}\to 1$. We can also deduce the following corollary.

Corollary 1 *As* $T\to\infty$, *the phase of the GMM channel estimate* $\hat{\boldsymbol{h}}_{\mathrm{GMM}}=\boldsymbol{E}\hat{\boldsymbol{h}}$ *is closer to the phase of the true channel than the phase of the initial estimate, that is,*

$$\left\|\frac{\hat{\boldsymbol{h}}_{\mathrm{GMM}}}{\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\|}-\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\right\| \le \left\|\frac{\hat{\boldsymbol{h}}}{\|\hat{\boldsymbol{h}}\|}-\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\right\|.$$

[28] This does not imply that error rates converge.

Proof. Since relative magnitudes between $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ and $\hat{\boldsymbol{h}}$ are irrelevant for SNR computations, the increased SNR implies that $|\hat{\boldsymbol{h}}_{\mathrm{GMM}}^{\mathrm{H}}\boldsymbol{h}|/\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\| \geq |\hat{\boldsymbol{h}}^{\mathrm{H}}\boldsymbol{h}|/\|\hat{\boldsymbol{h}}\|$. Thus, the phase of $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ is better aligned with the phase of $\boldsymbol{h}$ than the phase of $\hat{\boldsymbol{h}}$, which yields the inequality. ■

5 Bias Considerations of the GMM Detector

From Corollary 1 we know that the phase of $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ is better aligned to the phase of $\boldsymbol{h}$ than the phase of $\hat{\boldsymbol{h}}$ is. However, the magnitude $\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\|$ needs not to be closer to $\|\boldsymbol{h}\|$ than the magnitude $\|\hat{\boldsymbol{h}}\|$ is. In fact, the magnitude $\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\|$ is typically very small. The reason is that as T grows, the GMM detector converges to the all-zero solution. To resolve this, we constrain $\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\|$ to be equal to $\|\hat{\boldsymbol{h}}\|$. Thus, we construct $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ as

$$\hat{\boldsymbol{h}}_{\mathrm{GMM}} = \frac{\|\hat{\boldsymbol{h}}\|}{\|\boldsymbol{E}\hat{\boldsymbol{h}}\|}\boldsymbol{E}\hat{\boldsymbol{h}}.$$

The GMM detector performs MRC scaling similar to (2) and is given by

$$\begin{aligned}\hat{\boldsymbol{x}} &= \frac{1}{\|\hat{\boldsymbol{h}}_{\mathrm{GMM}}\|^2}\hat{\boldsymbol{h}}_{\mathrm{GMM}}^{\mathrm{H}}\boldsymbol{Y} \\ &= \frac{1}{\|\hat{\boldsymbol{h}}\|\|\boldsymbol{E}\hat{\boldsymbol{h}}\|}\hat{\boldsymbol{h}}^{\mathrm{H}}\boldsymbol{E}\boldsymbol{Y} \end{aligned} \tag{10}$$

where $\hat{\boldsymbol{h}} = \beta\boldsymbol{y}_0$. The GMM detector (10) can be reached without knowing the SNR, in which case $\beta = 1$.

The vector $\hat{\boldsymbol{x}}$ can be modeled as $\hat{\boldsymbol{x}} = \rho\boldsymbol{x} + \tilde{\boldsymbol{n}}$ where ρ is a bias which depends on the system parameters and on the unknown channel distribution. This bias is given by the expectation $\rho = \mathbb{E}[z_k]$, where z_k is defined shortly before Proposition 1. Notice that the expectation is not dependent on k, since all z_k are statistically equivalent. Since the phase of $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ is better aligned with the phase of $\boldsymbol{h}$ than the phase of $\hat{\boldsymbol{h}}$ is, the bias ρ is closer to unity than the corresponding bias for the benchmark detector is. Thus, we expect superior performance with the GMM detector also for QAM constellations.

Up to this point, we have not made use of any knowledge of the SNR. If the SNR is at hand, we can seek to compensate for the bias. However, the expectation $\rho = \mathbb{E}[z_k]$ is not feasible to find in closed form, and we resort to approximations. We point out that if the receiver has access to multiple received blocks, then the bias ρ can be estimated over time. For Gaussian channels, we empirically found that a good approximation of ρ is given by

$$\rho = \frac{1}{2}\left(1 + \tanh\left(\frac{k_1 + 10\log_{10}(\gamma)}{k_2}\right)\right). \tag{11}$$

The two parameters k_1 and k_2 depends on M and T, but can be tabulated off-line. Thus, for the case where the SNR is known to the receiver, the GMM detector becomes

$$\hat{\boldsymbol{x}} = \frac{1}{\rho}\frac{1}{\|\boldsymbol{y}_0\|\|\boldsymbol{E}\boldsymbol{y}_0\|}\boldsymbol{y}_0^{\mathrm{H}}\boldsymbol{E}\boldsymbol{Y}. \tag{12}$$

The reason for using $\boldsymbol{y}_0$ in (12) and not the more general notation of $\hat{\boldsymbol{h}}$ is that (11) has been computed for initial LS estimation so that $\hat{\boldsymbol{h}} = \boldsymbol{y}_0$. As MMSE estimation and LS estimation only differs with a constant, this is irrelevant for performance provided that ρ is computed for LS estimation.

6 Numerical Evaluations

We next provide numerical examples for the performance improvement attained by the GMM detectors over their linear counterparts, i.e., the benchmark detectors of Sec. 2.1. In all presented cases, the entries of $\boldsymbol{h}$ are chosen as zero mean independent and identically distributed complex Gaussian entries with unit variance. We set $\min\{T, M\} = 4$ for most simulations, and hence low additional complexity is required to perform GMM detection. Relaxing this complexity constraint trades off with better performance.

Fig. 1 shows the symbol error rate (SER) performance of different detectors. GMM detection provides SNR gains that increase with higher values of M. This dependency is shown explicitly in Fig. 2 for a given SER. Higher gains are harvested when the SNR is unknown. These gains seem linear in M which renders the method especially interesting for massive SIMO systems. A few dBs are harvested for rather small block length values. Fig. 3 shows the convergence of the post processing SNR of the GMM detector γ_{GMM} to its asymptotic bound found in Proposition 1. Estimating γ_{GMM} is done by means of Monte Carlo computations of $\mathbb{E}[z_k]$ and $\text{Var}[z_k]$. Most gains seem to be reached at moderate values of T, e.g. say $T + 1 = 10$ for the cases of Fig. 3, but this varies depending on the parameters setting chosen. Gains close to 2 dB are reached compared to the standard LS scheme, i.e. $T = 1$.

7 Conclusions

The GMM receiver explores the structure of block fading channels to improve the channel estimate quality, i.e. it filters the raw channel estimate with a matrix constructed from the received pilots and data. The computation overhead of the GMM receiver is relatively small compared to benchmark schemes as LS and MMSE. This overhead trades off with SNR gains, that can be harvested with small block lengths. For example, gains of 1.5 dB compared to benchmark schemes are attained with a block length of five and eight receive antennas. For a fixed system setup, these gains converge asymptotically with the block length, and appear linear with the number of receive antennas which renders the method especially interesting for massive SIMO systems.

Acknowledgments

This work was funded by the Swedish foundation for strategic research SSF, VR, the strategic research area ELLIIT, and the European Union Seventh Framework

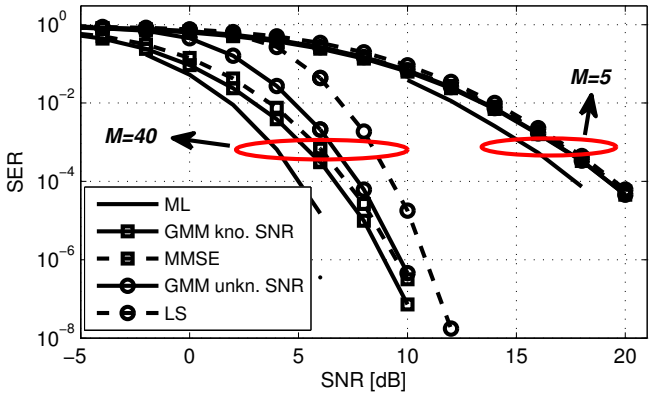

Figure 1: Symbol error rate with 16-QAM signaling, different values of M, and $T+1=5$. The dashed curves represent the LS and MMSE (linear) detectors, and the solid marked curves represent the GMM detector when the SNR is known and unknown. The solid curve represents the receiver where joint ML detection of $\boldsymbol{h}$ and $\boldsymbol{x}$ is performed [3, 4].

Programme (FP7/2007-2013) under grant agreement n 619086 (MAMMOET).

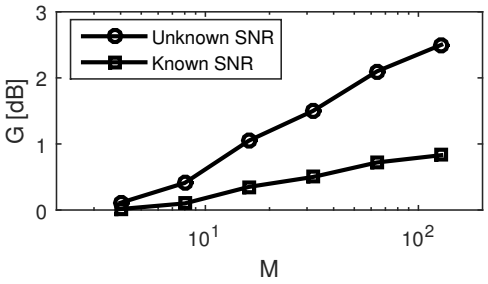

Figure 2: Gain of the GMM receiver without SNR knowledge over the LS scheme (upper curve), and gain of the GMM receiver with SNR knowledge over the MMSE scheme (lower curve), respectively. In both cases, we use 16-QAM symbols, $T = 4$, and SER $\approx 10^{-3}$.

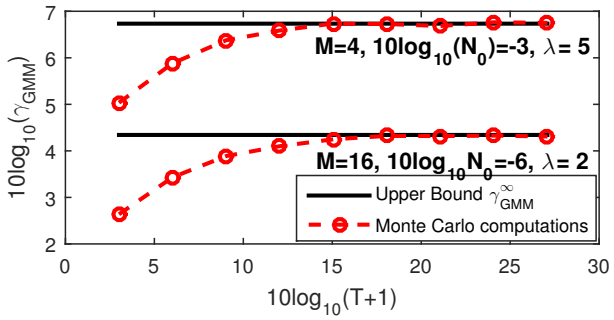

Figure 3: Monte Carlo computations of the SNR of the GMM detector, and respective asymptotic bound, for different system parameters combinations. Here $\{x_k\}_{k=1}^{T}$ are 4-QAM symbols.

References

[1] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *Information Theory, IEEE Transactions on*, vol. 49, no. 4, pp. 951–963, April 2003.

[2] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *Information Theory, IEEE Transactions on*, vol. 45, no. 1, pp. 139–157, Jan 1999.

[3] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, "GLRT-optimal noncoherent lattice decoding," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3773–3786, July 2007.

[4] F. Rusek, "Achievable rates of IID Gaussian symbols on the non-coherent block-fading channel without channel distribution knowledge at the receiver," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 4, pp. 1277–1282, Apr 2012.

[5] M. Lončar et al., "Iterative channel estimation and data detection in frequency-selective fading MIMO channels," *Eur. Trans. Telecomm.*, vol. 15, pp. 459–470, 2002.

[6] A. Hall, *Generalized Method of Moments*, ser. Advanced Texts in Econometrics. OUP Oxford, 2004.

[7] R. Rogalin et al., "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 4, pp. 1815–1831, April 2014.

[8] J. G. Andrews et al., "What will 5G be?" *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 6, pp. 1065–1082, June 2014.

Paper VII

Deep Convolutional Neural Networks for Massive MIMO Fingerprint-Based Positioning

This paper provides an initial investigation on the application of convolutional neural networks (CNNs) for fingerprint-based positioning using measured massive MIMO channels. When represented in appropriate domains, massive MIMO channels have a sparse structure which can be efficiently learned by CNNs for positioning purposes. We evaluate the positioning accuracy of state-of-the-art CNNs with channel fingerprints generated from a channel model with a rich clustered structure: the COST 2100 channel model. We find that moderately deep CNNs can achieve fractional-wavelength positioning accuracies, provided that an enough representative data set is available for training.

Published in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Oct. 2017*

1 Introduction

Massive MIMO is a candidate technology to integrate next generation cellular systems, such as 5G systems, and deliver manifold enhancements in the communications link [1]. In its conceived form [2], massive MIMO uses a large number of base station (BS) antennas together with *measured* channel state information (CSI) to multiplex user terminals spatially. Measured CSI is essential to yield spectrally efficient communications, but it can also be a key enabler to achieve highly-accurate terminal positioning, where down to centimeter order accuracy may be required in some 5G applications, e.g., autonomous driving [3]. Explained briefly, since positioning is a spatial inference problem, it makes sense to use large antenna arrays that oversample the spatial dimension of a wireless channel (thus benefiting from, e.g., increased angular resolution, resilience to small-scale fading, and array gain effects) to aid the positioning task.

Several approaches that make use of measured massive MIMO channels for positioning exist. For example, the approach proposed in [4] detects a terminal's position (from a grid of candidate positions) using line-of-sight (LOS) based triangulation from a terminal to several distributed massive MIMO BSs. In [5], positioning is performed using the phases of multipath components estimated from massive MIMO channels. Another positioning approach was proposed in [6], where received signal strength (RSS) based fingerprinting from one single-antenna terminal to N_{BS} M-antenna BSs is employed. Here, the challenge was to learn the inverse map

$$f_{\mathrm{RSS}}^{-1} : \{|\mathbf{Y}_i|^2\} \rightarrow \{\mathbf{x}_i\} \tag{1}$$

from a set of training observations, i.e. the training set $\{\mathbf{Y}_i, \mathbf{x}_i\}_{i=1}^{N_{\mathrm{Train}}}$. Here the label $\mathbf{x}_i \in \mathbb{R}^{2\mathrm{x}1}$ is the 2-dimensional terminal coordinate of training observation i, and $\mathbf{Y}_i \in \mathbb{C}^{M\mathrm{x}N_{\mathrm{BS}}}$ is its associated channel fingerprint. Gaussian Process-based Regression was used to learn $f_{\mathrm{RSS}}^{-1}(\cdot)$.

What the previous two mentioned positioning proposals and other proposals have in common (e.g., [7] and [8]), is that the structure of their solutions is typically composed by 2 distinct steps. In the first step, empirical feature extraction (from the measured channel snapshots) is performed (e.g., RSS), and in the second step, positioning of the terminal is done using the extracted features and a suitable algorithm-the algorithm typically being is the main contribution of the work. Although such 2 step solutions simplify the entire positioning task, they are inherently sub-optimal since they are constrained to use only partial–and typically not statistically sufficient [9]–channel statistics to solve the problem at hand. Thus, is it of interest to explore positioning frameworks that jointly *extract* and *process* channel features from measurements–under some joint optimality criterion.

In this work, fingerprinting-based positioning is performed using a framework that jointly extracts and processes channel features for positioning. More specifically, we are interested to learn

$$f^{-1} : \{s\left(\mathbf{Y}_i\right)\} \rightarrow \{\mathbf{x}_i\}, \tag{2}$$

i.e., the inverse of the underlying function $f(\cdot)$ that maps a set of single-antenna terminal coordinate vectors $\{\mathbf{x}_i\}$ to their respective *measured* but transformed channel

snapshots $s(\mathbf{Y}_i) \in \mathbb{C}^{d_1 \mathrm{x} \ldots \mathrm{x} d_D}\ \forall i$, where D is the dimensionality of the transformed snapshot. We note that the main point of the transformation $s(\cdot)$ is to obtain a sparse representation for $s(\mathbf{Y}_i)$. This is motivated in detail in Sec. 2.2. For now, we remark that the sparse transformations considered in this work are bijective, and thus yield no information loss.[29]

Our proposal to learn (2) is by means of deep convolutional neural networks (CNNs). Deep neural networks provide state-of-the-art learning machines that yield the most learning capacity from all machine learning approaches [10], and lately have been very successful in image classification tasks. Just like most relevant information for an image classification task is sparsely distributed at some locations of the image [10], measured channel snapshots $\mathbf{Y}_i$ have, when represented in appropriate domains, a sparse structure which–from a learning perspective–resemble that of images. This sparse channel structure can be learned by CNNs and therefore used for positioning purposes. To the best of the authors' knowledge there is no prior work on this matter.

The main contributions of this paper are summarized below.

- We investigate the feasibility of deep CNNs for fingerprint-based positioning with massive MIMO channels, and provide insights on how to design such networks based on machine learning and wireless propagation theory.
- As a proof-of-concept, we demonstrate the accuracy of our approach by performing fractional-wavelength positioning using channel realizations generated from a widely accepted cellular channel model: the COST 2100 MIMO channel model [11].

2 Channel Fingerprinting and Pre-Processing

In this section, we explain the fingerprinting scenario addressed in this work, and motivate why CNNs are learning machines suitable to perform positioning under such scenarios. To maximize insights, we focus most motivational remarks on the current case-study, but also provide several generalization remarks at the end of the section.

2.1 Channel Fingerprinting

In this work, we assume a BS equipped with a linear M-antenna array made of omnidirectional $\lambda/2$-spaced elements, and that narrowband channels sampled at N_F equidistant frequency points are used for positioning. With that, the dimensionality of each channel fingerprint $\mathbf{Y}_i$ (and, as it will be seen later, of each transformed fingerprint $\mathrm{s}(\mathbf{Y}_i)$) is $D = 3$ and

$$d_1 = M,\ d_2 = N_F,\ \text{and}\ d_3 = 2.$$

[29] We remark that this positioning approach is inherently designed for the single-user case. This fits well within a massive MIMO context since mutually orthogonal pilot sequences, which are seen as sounding sequences in the context of this work, are typically used by different users during uplink training [2]. The extention of this approach to a multi-user case is thus straightforward.

Given a terminal position, its associated fingerprint is generated through $f(\cdot)$, i.e. the inverse of the function we wish to learn. We implement $f(\cdot)$ using the COST 2100 channel model, the structure of which is illustrated by Fig. 1, under the parametrization proposed in [12]. This parametrization was performed for outdoor environments and is further detailed in Sec. 4. However, we note that our method is not restricted to work only in outdoor channels–we remark on the required channel properties in Sec. 2.2. It is important to note that, in this work, $f(\cdot)$ is implemented as a bijective deterministic map, i.e., there is only one unique fingerprint per position.[30]

2.2 Motivation for CNNs and Sparse Input Structures

Applying standard feed-forward neural networks to learn the structure of $\{\mathbf{Y}_i\}$ may be computationally intractable, specially when M grows very large. However, the structure of neural networks can be enhanced, both from a computational complexity and a learning point-of-view, if designed with sparse interaction and parameter sharing properties [10]. This is a widely used architecture for CNNs, suitable to process inputs with grid-like structures (e.g., an image can be thought as a two-dimensional grid of pixels) with minimal amounts of pre-processing.

CNNs are efficient learning machines given that their inputs meet the following two structural assumptions:

1. most relevant information features are sparsely distributed in the input space;
2. the shape of most relevant information features is invariant to their location in the input space, and are well captured by a finite number of kernels.

From a wireless channel point-of-view, these assumptions apply well when channels snapshots (i.e., the CNN inputs) are represented in domains that yield a sparse structure [13]. For example, in the current case study, sparsity is achieved by representing $\mathbf{Y}_i$ in its, so-called, angular-delay domains, see Fig. 1. Trivially, $s(\cdot)$ can take the form of a two-dimensional discrete Fourier transform, i.e.,

$$s(\mathbf{Y}_i) = \mathbf{F}\,\mathbf{Y}_i\,\mathbf{F}^H. \tag{3}$$

If specular components of the channel, which are typically modeled by Dirac delta functions [13], are seen as the information basis for positioning, then the two structural assumptions of the CNNs inputs listed above are met. The same applies, if instead, clusters of multipath components are seen as the information features for positioning.

[30] Bijectiveness of $f(\cdot)$ applies in most practical propagation scenarios with high probability (the probability typically approaches one as M increases). This is an important aspect to consider as it addresses the conditions needed to be able to use CNNs (or more generally, to solve the inverse problem). On a different note, regarding the deterministic structure for $f(\cdot)$, this is done by generating both training and test sets from the same given realization of the COST 2100 channel model stochastic parameters. This makes each fingerprint to be completely determined solely by the geometry of the propagation channel itself. Stochastic effects in the fingerprinting process, such as measurement and labeling noise, or even time–variant channel fading are interesting impairments to be considered in the design of CNN in future work. For now, we focus on the case of having unique fingerprints per position, due to simplicity.

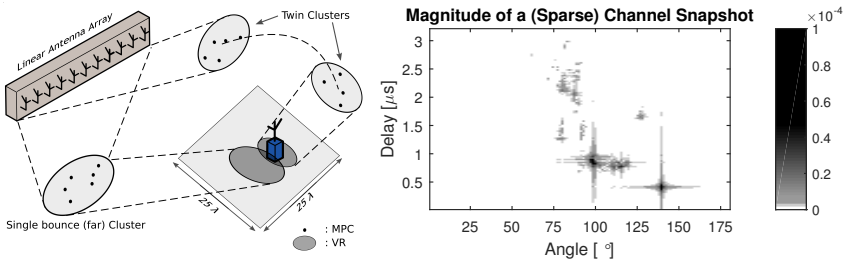

Figure 1: Left–Link setup considered in this work: an M-linear BS array positioning one single-antenna terminal in a confined square area. Channel realizations are generated through the COST 2100 MIMO channel model. This geometry-based stochastic channel model is composed by different types of clusters of multipath components (MPCs) that illuminate certain visibility regions (VRs) of an area. Right–Example of the magnitude of a channel snapshot represented in a sparse domain. Such channel has a rich structure that can be learned by a CNN for positioning purposes.

2.3 Generalizing the Current Case Study

The current case study can be extended to more generic/higher-dimensionality fingerprints. For example, if $\{\mathbf{Y}_i\}$ is comprised by snapshots measured from a multi-antenna terminal to a BS array, both with arbitrary array structures and non-omnidirectional antenna elements, the effective aperture distribution function (EADF) [14] may be accounted in $s(\cdot)$ in order to obtain a sparse fingerprint representation. However, we emphasize that, in contrast to most propagation studies, one does not need to necessarily de-embed the measurement system from the propagation channel in order to obtain valid fingerprints for positioning. Also, multi-antenna channels yielding phenomena such as violation of the plane wave assumption, or even the existence of cluster visibility regions [15], can be made sparse by means of proper transformations, e.g., generalized Fourier transforms. In any case, the key is the ability to obtain a sparse representation for $s(\mathbf{Y}_i)$.

3 Deep CNN Architecture

In this section, we describe the network architecture used for learning $f^{-1}(\cdot)$, and discuss some design aspects.

For notation convenience, we drop the dependence of the training sample index i, and write $\mathbf{Y} \triangleq \mathbf{Y}_i$ and $\mathbf{x} \triangleq \mathbf{x}_i$ until explicitly stated otherwise.

3.1 Convolutional-Activation-Pooling Layers

After the input layer, which takes the transformed snapshots, a typical structure of CNNs employs a cascade of L convolutional-activation-pooling (CAP) layers. Each CAP layer is composed by: *i)* a convolutional operation of its input with K convolutional Kernels, *ii)* a non-linear transformation, i.e., activation function, and *iii)* a pooling layer, respectively. A detailed description of the CAP layer structure used in this work follows below.

Let the tensor $\mathbf{H}^{\ell-1} \in \mathbb{R}^{M \mathtt{x} N_F \mathtt{x} S_3}$ be the input of the ℓth CAP layer, with $1 \leq \ell \leq L$. Also, let the jth convolutional Kernel of the ℓth layer be denoted by $\mathbf{w}_j^\ell \in \mathbb{R}^{S_1 \mathtt{x} S_2 \mathtt{x} S_3}$, with $1 \leq j \leq K$, and S_1 and S_2 denoting the sizes of the Kernels (which are CNN hyper-parameters). With $\mathbf{h}_{r,c}^{\ell-1} \in \mathbb{R}^{S_1 \mathtt{x} S_2 \mathtt{x} S_3}$ being a sub-tensor of a zero-padded version[31] of $\mathbf{H}^{\ell-1}$, an output entry of the ℓth convolutional layer can be written as

$$c_{r,c,j}^\ell = b_j^\ell + \mathbf{1}^T \left(\mathbf{w}_j^\ell \circ \mathbf{h}_{r,c}^{\ell-1} \right) \mathbf{1}. \tag{4}$$

Here $\circ$ denotes the Hadamard product, b_j^ℓ is a bias term, $\mathbf{1}$ denotes the all-ones column vector, and r and c are indices in the convolution output volume which are implicitly defined.

In the first CAP layer, the input $\mathbf{H}^0$ is a tensor made out of the complex-valued entries of (3), and thus we have $S_3 = 2$ (real dimensions). This is because, although channel snapshots are inherently complex-valued, we pursue an implementation of a real-valued CNN with real-value inputs (thus we have $S_3 = 2$ in the first CAP layer). We motivate why we do so in Sec. 3.4. In the remaining CAP layers, we have $S_3 = K$.

Each convolutional output entry (4) is fed to an activation function. We use the current default choice for activation functions in CNNs, namely, the rectified linear unit (RELU) [10], where the output can be written as

$$g_{r,c,j}^\ell = \max \left(c_{r,c,j}^\ell,\ 0 \right). \tag{5}$$

Finally, after each activation function follows a pooling operation which down samples the outputs of the activation functions. A standard option, also used here, is to forward propagate the maximum value within group of $N_1 \times N_2$ activation functions outputs. The pooling result can be written as

$$h_{r,c,j}^\ell = \max_{m=1}^{N_1} \max_{n=1}^{N_2} \left(g_{(r-1)N_1+m,\ (c-1)N_2+n,\ j}^\ell \right). \tag{6}$$

3.2 Fully-Connected Layer

A fully-connected layer, following the L CAP layers, finalizes the CNN. With that, the position estimate of the network, $\mathbf{t} \in \mathbb{R}^{2\mathtt{x}1}$, is given by

$$\mathbf{t} = \mathbf{W} \mathrm{vec} \left\{ \mathbf{H}^L \right\} + \mathbf{b}^L, \tag{7}$$

[31] The zero-padded version of $\mathbf{H}^{\ell-1}$ is obtained by padding the borders of the volume of $\mathbf{H}^{\ell-1}$ with zeros, such that, when convolved with any Kernel $\mathbf{w}_j^\ell$, the input and output volumes are the same [10].

where vec $\{\cdot\}$ vectorizes its argument, $\mathbf{b}^L = [b_1^L \; b_2^L]^T$ is a vector of biases, and $\mathbf{W}$ is a weight matrix whose structure is implicitly defined.

3.3 Network Optimization

The CNN network learns the weights $\mathbf{W}$ and $\left\{\mathbf{w}_j^\ell\right\}$, and biases $\left\{b_j^\ell\right\}$, in order to make $\mathbf{t}$ the best approximation of $\mathbf{x}$. Since we address positioning as a regression problem, we use the squared residuals averaged over the training set as the optimizing metric. Re-introducing the dependence on the sample index i, and defining the column vector $\boldsymbol{\theta}$ by stacking all network parameters as $\boldsymbol{\theta} = \left[\text{vec}\left\{[\mathbf{w}_1^1 \ldots \mathbf{w}_L^K]\right\}^T \text{vec}\left\{\mathbf{W}\right\}^T [b_1^1 \ldots b_2^L]\right]^T$, the optimum parameters are given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \;\; J(\boldsymbol{\theta}), \tag{8}$$

with

$$J(\boldsymbol{\theta}) = \frac{\beta}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} (\mathbf{x}_i - \mathbf{t}_i(\boldsymbol{\theta}))^2. \tag{9}$$

A Tikhonov penalty term is added to harvest the benefits of regularization in CNNs–β is its associate hyper-parameter. On a practical note, we minimize (9) using stochastic gradient-descent and back propagation [10].

3.4 Network Design Considerations

We now motivate our choice for the implementation of real-valued CNNs. From our experience, the main challenge of generalizing CNNs to the complex-valued case appears to be in finding a *suitable* generalization of (5), the RELU, in order to "activate" complex inputs. For example, the complex-valued CNN generalizations presented in [16] apply (5) to both real and imaginary parts separately, which from our experience appears not to perform well from a network optimization point-of-view. Explained briefly, such RELUs are non-continuous functions in $\mathbb{C}$, as opposed to only being non-differentiable in $\mathbb{R}$ as in real-valued network. This makes the network's optimization unstable with the current optimization method (i.e. gradient descent). Therefore, as real-valued CNNs have been very successful in image classification tasks, we conservatively choose to pursue this design option, and let the extension to the complex-valued case to be a matter of future work.

A practical remark regarding the choice of the number of CAP Kernels, K, follows. In practice, information features, addressed in Sec. 2.2, are subject to a number of variability factors. For example, the shape of a "*measured* Dirac delta function" can vary with its location due to discretization. Also, it is clear that, there is a higher variability in the informational features if they are seen as clusters of MPCs, rather than single MPCs themselves. In both cases, a practical approach to deal with such variability factors is to set a high number for K, and let the CNNs learn a set of kernels that span most possible variations.

To finalize, if the information features are seen as the clusters of MPCs, the sizes of the Kernels, S_1 and S_2, should cover their range in the angular and delay channel representations. This is because MPCs within a cluster may be statistically dependent, but different clusters are typically not.

3.5 Complexity Aspects

The most computationally challenging aspect of the entire approach is the optimization stage of the network. This is due to the large size of the training sets, network dimensionality, non-convexity of $J(\boldsymbol{\theta})$, etc. However, once the optimization stage is finalized, *real-time* positioning can easily be achieved due to the feed-forward structure of the network. This can be easily observed (in the current case-study) by looking at the overall complexity order of a CNN point-estimate: $\mathcal{O}(K^2MLN_FS_1S_2)$. The fact that the complexity does not depend on the training set size N_{train} is one main advantage of using CNNs for positioning.

4 Positioning Results

Next, we address the positioning capabilities of CNNs by means of numerical results. We omit showing optimization aspects of the network (e.g., convergence across epochs) as the main point of the paper is to analyse the positioning capabilities of an optimized CNN.

4.1 Simulation Setup

The setup used in our experiments is illustrated by Fig. 1: the terminal is constrained to be in a square area $\mathcal{A}$ of 25×25 wavelengths. Channel fingerprints are obtained in this area through the COST 2100 channel model under the 300 MHz parameterization (e.g., for path-loss and cluster-based parameters) established in [12]. The remaining parameters are shown in Table 1, and the other CNNs hyper-parameters, i.e. L and K, are varied during the simulations.

The closest and furthest coordinate points of $\mathcal{A}$ with respect to the first BS antenna are:

$$\mathbf{u}_c = [-12.5\lambda \ -12.5\lambda]^T \text{ and } \mathbf{u}_f = [12.5\lambda \ 12.5\lambda]^T,$$

respectively (i.e., the user is at least $||\mathbf{u}_c - \mathbf{B}_1||/\lambda$ wavelengths away from the first BS antenna). The coordinates of these two spatial points implicitly define the relative orientation of the linear array with the area $\mathcal{A}$. Similarly, the upcoming performance analysis is done by means of the normalized root mean-squared error (NRMSE) where the mean is calculated as the average over the

Table 1: Channel and CNNs parameters

Parameter	**Variable**	**Value**
Carrier frequency	f_c	300 MHz
Bandwidth	W	20 MHz
# Frequency points	N_F	128
# BS antennas	M	128
First BS antenna coordinate	$\mathbf{B}_1$	$[-200\lambda \quad -200\lambda]^T$
Last BS antenna coordinate	$\mathbf{B}_M$	$[-200\lambda \quad -200\lambda + \frac{(M-1)\lambda}{2}]^T$
Tikhonov hyper-parameter	β	10^{-3}
Kernel angular length [$^\circ$]	S_1	9.8
Kernel delay length [μs]	S_2	0.175
Pooling windows length	N_1 and N_2	2

test sets samples. Thus, we have

$$\text{NRMSE} = \frac{1}{\lambda}\sqrt{\frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}(\mathbf{x}_i - \mathbf{t}_i(\boldsymbol{\theta}))^2}.$$

This error metric has an understandable physical intuition as it shows how the error distance relates to the wavelength.

The CNN training and testing is described as follow:

1. First, the training set is obtained by fingerprinting a 2-dimensional uniformly-spaced (thus, deterministic) grid of positions spanning the totality of $\mathcal{A}$. The impact of the sampling density is discussed in Sec. 4.3.

2. For the test set, each fingerprint's position is obtained by sampling a random variable with a uniform distribution with support $\mathcal{A}$.

Note that, if the CNN cannot use the available fingerprints for training, then the position estimator is $\mathbb{E}\{\mathbf{x}\} = \mathbf{0}$, see (9). Its NRMSE, for the current case study, is given by

$$\text{NRMSE}^{\text{ref}}(\mathcal{A}) = \frac{1}{\lambda}\sqrt{\frac{1}{\int_{\mathcal{A}}\partial\mathbf{d}}\int_{\mathcal{A}}\left(\mathbf{d} - \mathbb{E}\{\mathbf{x}\}\right)^2\,\partial\mathbf{d}} \approx 10.2. \tag{10}$$

Obviously, this reference value increases when $\mathcal{A}$ is larger. Since an optimized CNN with a non-zero number of fingerprints should be able to do better or equal than (10), we use (10) as a reference level in the analysis.

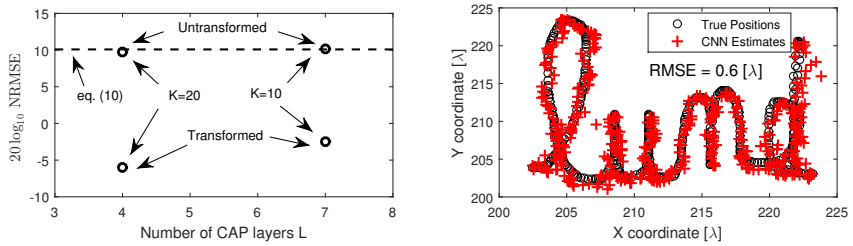

Figure 2: Left–NRMSE obtained by CNNs under different parameterizations. The upper horizontal line corresponds to the reference level (10). Here we only report the test error, since a similar error value was obtained during training (i.e., no overfitting exists). Right–An illustrative example of the point-estimates from a pre-defined set of positions by the optimized CNN (from the left figure) with $K = 20$ and $L = 4$.

For benchmarking purposes, we also contrast our CNN results against a standard non-parametric fingerprinting approach [17]. Seeing a training fingerprint as a function of its position, i.e. $\mathbf{Y}_i(\mathbf{x}_i)$, this approach computes the position from a new fingerprint $\mathbf{Y}_{\text{new}}$ through a grid-search over normalized correlations as

$$\hat{\mathbf{x}}_i = \underset{\mathbf{x}_i \in \{\mathbf{x}_i\}_{i=1}^{N_{\text{train}}}}{\text{argmax}} \frac{|\operatorname{Tr}\{\mathbf{Y}_i(\mathbf{x}_i)^H \mathbf{Y}_{\text{new}}\}|}{\sqrt{|\mathbf{Tr}\{\mathbf{Y}_i(\mathbf{x}_i)^H \mathbf{Y}_i(\mathbf{x}_i)\} \operatorname{Tr}\{\mathbf{Y}_{\text{new}}^H \mathbf{Y}_{\text{new}}\}|}}. \tag{11}$$

The following remarks can be made about this non-parametric approach:

1. Compared to the use of CNNs, a main drawback is its computational complexity order, $\mathcal{O}(MN_F^2 N_{\text{train}})$, as it depends on the size of the training set;

2. it has no inherent interpolation abilities, and thus its error can be lower bounded given the spatial density of the training set;

4.2 Proof-of-Concept and Accuracy for Different CNN Parametrizations

Here, we report the positioning results when the spacing between neighbor training fingerprints is $\lambda/4$. We consider this extreme case for now, in order to mitigate the impact of spatial undersampling from the results-here the focus is solely on the positioning capabilities of the network. The impact of the training fingerprints spacing is addressed later in Sec. 4.3.

Fig. 2 (left) illustrates the positioning accuracy for different cases of CNN parameterizations. First, and as a sanity check, we see that for the same parameterizations, a network fed with untransformed inputs (i.e., $s(\mathbf{Y}_i) = \mathbf{Y}_i$) cannot effectively learn the channel structure for positioning purposes– the order of magnitude of the positioning error is similar to (10). However, with transformed inputs, fractional-wavelength positioning can be achieved in both network settings, with the lowest achieved test NRMSE being of about $-6\text{dB} \approx 1/2$ of a wavelength. This showcases the capabilities of CNNs to learn the structure of the channel for positioning purposes. We remind that such positioning accuracies are attained with only 20 MHz of signaling bandwidth, see Table 1, which suggests that CNNs can efficiently trade-off signal bandwidth by BS antennas and still achieve very good practical performance. Decreasing the error further than fractional-wavelength ranges becomes increasingly harder due to the increased similarities of nearby fingerprints–such range approaches the coherence distance of the channel. Also, as an illustrative example of the positioning accuracy, Fig. 2 (right) shows the point-estimates of the CNN under the parameterization that attained the lowest test NRMSE.

4.3 Accuracy for Different Training Grids

To finalize, we analyze the impact of spatial sampling during training. For benchmarking, we contrast the CNN performance with the performance of the correlation-based classifier (11). We use the CNN hyper-parameters that attained the lowest MSE in Fig. 2, namely, the model with $L = 4$ and $K = 20$.[32] Fig. 3 contrasts the NRMSEs obtained from a CNN and the correlation-based classifier (11), against spatial sampling in the training set. Overall, both approaches are able to attain fractional-wavelength accuracies at smaller training densities. Noticeably, the CNN tend to behave better than (11) for less dense training sampling. Given that (11) does not have interpolation abilities, this result is closely connected with the inherent interpolation abilities of the CNNs. The fact that the CNNs achieve similar, or even superior performance compared to standard non-parametric approaches while having attractive implementation complexity further corroborates their use in fingerprint-based localization systems.

[32] Ideally, the CNN hyper-parameters should be tuned according to the current training set. However, we keep the same hyper-parameterization throughout this analysis, for simplicity.

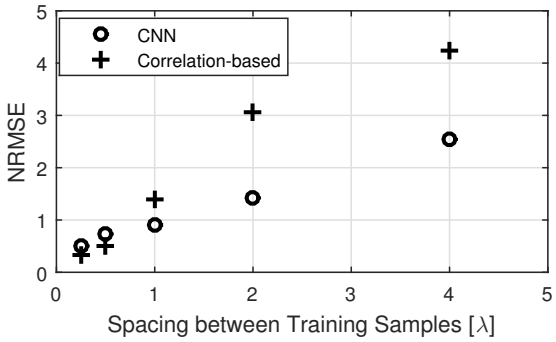

Figure 3: NRMSE obtained by different positioning approaches for different spacings between samples of the uniform training grid.

5 Takeaways and Further Work

We have investigated a novel approach for massive MIMO fingerprint-based positioning by means of CNNs and measured channel snapshots. CNNs have a feedforward structure that is able to compactly summarize relevant positioning information in large channel data sets. The positioning capabilities of CNNs tend to generalize well, e.g. in highly-clustered propagation scenarios with or without LOS, thanks to their inherent feature learning abilities. Proper design allows fractional-wavelength positioning to be obtained under real-time requirements, and with low signal bandwidths.

The current investigation showcased some of the potentials of CNNs for positioning using channels with a complex structure. However, the design of CNNs in this contexts should be a matter of further investigation, in order to be able to deal with real-world impairments during the fingerprinting process. In this vein, some questions raised during this study are, for example, *i)* how to achieve a robust CNN design that is able to deal with impairments such as measurement and labeling noise, or channel variations that are not represented in the training set, or *ii)* how to design complex-valued CNNs that perform well and are robust during optimization.

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] Z. Chaloupka, "Technology and Standardization Gaps for High Accuracy Positioning in 5G," *IEEE Communications Standards Magazine*, vol. 1, no. 1, pp. 59–65, March 2017.

[4] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich, and M. Coulon, "Direct Localization for Massive MIMO," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2475–2487, May 2017.

[5] X. Li et al., "Robust Phase-Based Positioning Using Massive MIMO with Limited Bandwidth," in *2017 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Oct 2017.

[6] V. Savic and E. G. Larsson, "Fingerprinting-Based Positioning in Distributed Massive MIMO Systems," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, Sept 2015, pp. 1–5.

[7] S. A. Shaikh and A. M. Tonello, "Localization based on angle of arrival in EM lens-focusing massive MIMO," in *2016 IEEE 6th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, Sept 2016, pp. 124–128.

[8] M. Zhu, J. Vieira, Y. Kuang, K.Astrom, A. F. Molisch, and F. Tufvesson, "Tracking and positioning using phase information from estimated multi-path components," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 712–717.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[11] L. Liu et al., "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, December 2012.

[12] M. Zhu, G. Eriksson, and F. Tufvesson, "The COST 2100 Channel Model: Parameterization and Validation Based on Outdoor MIMO Measurements at 300 MHz," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 888–897, February 2013.

[13] A. Molisch, *Wireless Communications*, ser. Wiley - IEEE. Wiley, 2010.

[14] T. Kaiser, H. B. Andre Bourdoux, J. B. A. Javier Rodriguez Fonollosa, and W. Utschick, *Smart Antennas - State of the Art*. Hindawi Publishing Corporation, 2005.

[15] X. Gao, F. Tufvesson, and O. Edfors, "Massive MIMO channels - Measurements and models," in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov 2013, pp. 280–284.

[16] N. Guberman, "On complex valued convolutional neural networks," *CoRR*, vol. abs/1602.09046, 2016. [Online]. Available: http://arxiv.org/abs/1602.09046

[17] Z. H. Wu, Y. Han, Y. Chen, and K. J. R. Liu, "A time-reversal paradigm for indoor positioning system," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1331–1339, April 2015.