D3.2

Distributed and centralized baseband processing algorithms, architectures, and platforms

Project number:	619086
Project acronym:	**MAMMOET**
Project title:	Massive MIMO for Efficient Transmission
Project Start Date:	1 January, 2014
Duration:	36 months
Programme:	FP7/2007-2013
Deliverable Type:	Report
Reference Number:	ICT-619086-D3.2
Workpackage:	WP 3
Due Date:	31 December, 2015
Actual Submission Date:	15 January, 2016
Responsible Organisation:	ULUND
Editor:	Liang Liu
Dissemination Level:	PU
Revision:	1.0
Abstract:	Baseband processing and the corresponding processing distribution for Massive MIMO systems are discussed. Digital signal processing algorithms are discussed in the context of practical deployment scenarios and in conjunction with hardware architecture, implementation cost, and power consumption.
Keywords:	Massive MIMO, digital baseband processing, processing distribution, hardware implementation, accelerator, power consumption

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 619086.

Editor

Liang Liu (ULUND)

Contributors (ordered according to beneficiary numbers)

Claude Desset (IMEC)
Emil Björnson, Salil Kashyap, Erik G. Larsson, Christopher Mollén (LIU)
Liang Liu, Steffen Malkowsky, Hemanth Prabhu, Yangxurui Liu, Joao Vieira, Ove Edfors (U-LUND)
Eleftherios Karipidis (EAB)
Franz Dielacher, Diana Vasilica Pop (IFAT)

Executive Summary

Massive MIMO (MaMi) is a promising technology to both increase the system capacity and reduce the power consumption for 5G network. It is well studied that the radiated power can be reduced in MaMi systems, while the total processing power may be increased due to the large number of antenna chains. Thereby, it is crucial to find a way for power-efficient implementation of MaMi processing to keep the overall power consumption low.

To achieve this target, extensive investigation and optimization is needed at different design stages. At the algorithm design level, low-complexity processing methods should be developed by exploring the unique features of MaMi systems while still providing good performance. At the platform design level, processing architectures should be proposed taking processing distribution, data movement, and data storage into consideration. The requirements of processing throughput and latency should be taken care of together with processing power and flexibility. At the hardware design level, advanced CMOS technology should be leveraged together with circuit-level optimization to achieve efficient implementation. More importantly, co-optimization at all the aforementioned design levels should be conducted, which requires a good system-level model covering processing components from analog to digital domain.

To set the stage for MAMMOET year 3 in WP3, this deliverable (D3.2) addresses the above topics by collecting available knowledge among partners and results from investigations performed in the second year of the MAMMOET project. Conclusions include:

- MaMi allows for low complexity baseband processing to achieve good performance and enable low-power implementation. Examples are approximative matrix inversions and interpolation-based matrix operations.
- The power consumption of all hardware components in MaMi systems can remain small enough to keep a large benefit from MaMi concept. Further processing power optimization is still required, especially for the analog components.
- Processing hardware power consumption should be analyzed together with link budget and real-life channel environment to get better understanding of the overall power consumption in MaMi systems.
- Due to the processing of a large number of data streams, processing distribution, data shuffling capacity, and memory requirements play an important role for efficient baseband implementation. It is important to develop processing algorithms keeping in mind their affection on these 3 aspects to achieve algorithm-architecture co-optimization.
- Key digital processing blocks, like the zero-forcing precoder, can be implemented efficiently and consumes relatively low power. This is achieved by exploring the unique features in MaMi systems, for instance the Gram matrix is diagonally dominated.
- It is also possible to leverage the digital signal processing and the corresponding low-power implementation to further reduce the analog processing requirements, for instance to use highly power-efficient non-linear amplifiers, low-cost mixers, and low-precision data converters.
- A list of main complements and improvements to D3.1 is summarized in Chapter 6.

Contents

1 Introduction **1**

2 Baseband processing algorithm **3**
2.1 Uplink detection with active multi-cell interference suppression . . . 3
2.1.1 System model and transceiver design . . . 3
2.1.2 Simulation results . . . 6
2.1.3 Extensions to multi-cell downlink precoding . . . 9
2.2 Reciprocity calibration . . . 9
2.2.1 System model . . . 9
2.2.2 Calibration procedure . . . 10
2.2.3 Implementation in the LuMaMi testbed . . . 10
2.3 Downlink precoding . . . 12
2.3.1 System model . . . 12
2.3.2 The constant-envelope MIMO channel . . . 13
2.3.3 Continuous-time constant-envelope precoding . . . 14
2.3.4 Numerical analysis of the CTCE precoder . . . 18
2.4 Frequency interpolation of detection and precoding . . . 21
2.4.1 System model . . . 21
2.4.2 Uplink Ergodic rate analysis . . . 23
2.4.3 Numerical results . . . 28
2.4.4 Summary of interpolation methods . . . 32
2.5 Hardware imperfection assessment . . . 32
2.5.1 System model . . . 33
2.5.2 Downlink performance analysis with hardware impairments . . . 35
2.5.3 Numerical results . . . 38
2.5.4 Summary of hardware imperfection analysis . . . 39

3 Signal, noise and interference power in Massive MIMO links **40**
3.1 System definition and link assumptions . . . 40
3.2 Link analysis with interference and channel estimation errors . . . 42
3.2.1 Interference . . . 43
3.2.2 Channel training . . . 44
3.2.3 Overall analysis . . . 44
3.3 Conclusions . . . 46

4 Baseband processing profile **48**
4.1 Computational complexity and power consumption . . . 48
4.1.1 Overall approach . . . 49
4.1.2 PA and output power . . . 51

4.1.3 Digital complexity . . . 52
4.1.4 Analog components . . . 53
4.1.5 Power trends and conclusions . . . 55
4.2 Processing distribution and impact . . . 56
4.2.1 Processing latency . . . 56
4.2.2 Core processing elements . . . 58
4.2.3 Data movement bandwidth and data storage requirement . . . 59
4.2.4 Core memory . . . 61
4.2.5 On-chip communication . . . 64

5 Hardware implementation of baseband processing 69
5.1 System model . . . 69
5.2 QRD based Zero-Forcing (ZF) precoder . . . 70
5.3 Peak-to-Average Power Ratio (PAPR) aware precoding . . . 70
5.3.1 Antenna reservation based on ZF . . . 71
5.3.2 Discrete-time constant envelope precoder . . . 71
5.4 IQ imbalance pre-compensation . . . 72
5.4.1 Effects of IQ imbalance in massive MIMO . . . 73
5.4.2 Pre-compensation architecture . . . 74
5.5 Analysis of processing energy-per-bit . . . 77
5.6 Conclusion . . . 77

6 Summary 79

List of Abbreviations 84

List of Figures

1.1 Simplified block diagram of MaMi base station baseband processing (with M base-station antennas serving K single-antenna UEs). An example of processing partition and distribution is also shown. 2

2.1 The 19-cell-wrap-around hexagonal network topology for $f = 1$, $f = 3$, and $f = 4$. 7
2.2 Achievable sum SE of Multi-Cell MMSE (M-MMSE) (squares), Full pilot-based Zero-Forcing (P-ZF) (triangles), Single-Cell MMSE (S-MMSE) (diamonds) and MR (circles) with $f = 1$, $K = 10$ and $K = 30$. 8
2.3 Achievable sum SE of M-MMSE (squares), P-ZF (triangles), S-MMSE (diamonds) and MR (circles) with $f = 4$, $K = 10$ and $K = 30$. 8
2.4 Achievable sum SE of M-MMSE (squares), P-ZF (triangles), S-MMSE (diamonds) and MR (circles) with $f = 7$, $K = 10$ and $K = 30$. 9
2.5 MaMi physical setup used to validate reciprocity calibration. Three very closely spaced single-antenna terminals yielding strong LoS propagation channels to the BS. 11
2.6 Left: Equalized downlink signals at one of three users for the case of when the precoding matrix is the identity matrix. Right: Equalized downlink signals at one of three users for the case of ZF precoding. 12
2.7 The power spectral density of a typical constant-envelope signal. Its 30 dB-bandwidth $BT = 1.8$ is indicated. 14
2.8 The 30 dB-bandwidth of Continuous-Time Constant-Envelope (CTCE) precoded transmit signals for different choices of the regularizing factors λ_1 and λ_2. 20
2.9 The ergodic sum-rate of the simulated system for different precoders. The thick lines represent the proposed CTCE precoder for different bandwidths. 20
2.10 System model: K single-antenna users communicating with an M-antenna BS. . 22
2.11 DFT-interpolation in four steps: I. Compute L_0 equally spaced ZF matrices $\hat{\mathbf{G}}(s)\left(\hat{\mathbf{G}}(s)^H\hat{\mathbf{G}}(s)\right)^{-1}$ at $s = 1, N/L_0 + 1, \ldots, (L-1)N/L_0 + 1$, II. L_0-point IDFT of L_0 equally spaced ZF matrices ($L_0 > L$), III. Pad $N - L_0$ zeros starting at $\frac{L_0+L}{2}$, IV. N-point DFT of the ZF impulse response above. 25
2.12 Piecewise constant in two steps: I. Compute L_0 equally spaced ZF detectors, II. The ZF detector computed at subcarrier $\left(\frac{N}{L_0} + 1\right)$ is used over a cluster of adjacent subcarriers. 26
2.13 Linear interpolation in two steps: I. Compute L_0 equally spaced ZF detectors, II. For any subcarrier $1 \leq s \leq \frac{N}{L_0} + 1$, with linear interpolation and imperfect Channel State Information (CSI), the ZF detector at subcarrier s is $\hat{\mathbf{A}}(s) = \frac{L_0}{N}\left(\frac{N}{L_0} + 1 - s\right)\hat{\mathbf{A}}(1) + \frac{L_0(s-1)}{N}\hat{\mathbf{A}}\left(\frac{N}{L_0} + 1\right)$. 27
2.14 DFT-interpolation: Average ergodic rate vs. L_0 ($K = 4$, $N = 1024$, $\rho = -10$ dB) 29

2.15 DFT-interpolation: Average ergodic rate vs. L_0 ($K = 16$, $N = 1024$, $\rho = -10$ dB) 30
2.16 DFT-interpolation: Loss in average ergodic rate vs. M ($L = 64$, $N = 1024$, $\rho = -10$ dB). Loss in average ergodic rate is the ratio of the difference between the average ergodic rate when $L_0 = N$ and the average ergodic rate when $L_0 = L$ to the average ergodic rate when $L_0 = N$. 30
2.17 Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L_0 = L = 16$, $N = 1024$, $\rho = -10$ dB) 31
2.18 Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L = 16$, $L_0 = 32$, $N = 1024$, $\rho = -10$ dB) 31
2.19 Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L = 16$, $L_0 = 64$, $N = 1024$, $\rho = -10$ dB) 31
2.20 Illustration of the TDD protocol where each coherence block consists of $T = \tau_{\mathrm{UL}} + \tau_{\mathrm{DL}} + B$ symbols. 33
2.21 Illustration of the multi-cell MaMi scenario with distributed arrays considered in the numerical evaluation. 38
2.22 Average DL spectral efficiency for distributed MaMi with fixed or increasing hardware impairments. 39

3.1 Comparison of MR transmission performance between the ideal bound (100×1 with ideal CSI), inter-user interference (100×15 with ideal CSI) and interference with channel estimation error (100×15 with channel estimation from one pilot every 15th subcarrier). The system uses 1200 subcarriers out of 2048 based on LTE specifications and the channel model is time-domain Rayleigh with 20 taps of equal expected energy. 46

4.1 Power consumption for the reference macro BS (Configuration 1) compared with the three MaMi scenarios of Table 4.1, based on technology year 2014 [9]. . . . 50
4.2 Impact of PA input back-off (IBO) on system performance: linear operation (+20 dB) leads the optimum performance, entering the saturation region (0 dB and below) leads a limited degradation and complete saturation (down to -30 dB back-off) a degradation around 1.5 dB. 51
4.3 Power breakdown in downlink, uplink and training phases for a 100×10 MaMi system using MRT precoding, QPSK and LDPC coding rate 3/4 [9]. 55
4.4 Power breakdown in downlink, uplink and training phases for a 100×25 MaMi system using ZF precoding, 16-QAM and LDPC coding rate 3/4 [9]. 56
4.5 Simplified timing diagram for MaMi to point out some major challenges. 57
4.6 Multiplication Count Ratio of ZF to Matched Filter (MF). 59
4.7 Number of samples interchanged between different blocks in a MaMi system for frame structure used in the Lund University Massive MIMO (LuMaMi) testbed. 60
4.8 Illustration of On-chip subsystem for MaMi. 62
4.9 Vector-wise data storage and retrieval breakdown of the MaMi application in 1 subframe, using Long Term Evolution (LTE) parameter, 1200 sub-carriers, H size 128x16, Single-Instruction Multiple-Date (SIMD) width 16. 63
4.10 Examples of Access Pattern in MaMi Application. 64
4.11 MaMi system using single BUS as interconnection network. 66
4.12 Memory access pattern for memories inside the PSP blocks. 66
4.13 MaMi system using a multi-level router and a NoC for efficient communication . 67
4.14 First-In-First-Out (FIFO) write order pattern for UP (left) and UD (right) . . . 68

5.1 Data-flow illustration of the low complexity PAPR reduction approach, where the dedicated set of compensation antennas χ^c counteracts the clipping based distortion. 71
5.2 Systolic array for CE precoder based on coordinate-descent algorithm, where each processing element solves phase for an antenna. 72
5.3 Transmitter IQ imbalance model, with ϵ and $\delta\phi$ the physical mismatch parameters, $x_L(t)$ time domain baseband IQ signal and $x_{Tx}(t)$ is transmitted signal. . . 73
5.4 Simulated IQ imbalance for $K = 10$ users massive Multiple-Input Multiple-Output (MIMO) system with 6% amplitude and 6° degree phase mismatch. . . . 74
5.5 Pre-compensation for $M = 20$, $K = 10$ system, with different IQ imbalance estimation accuracy. 75
5.6 IQ imbalance pre-compensation top level data flow. 75
5.7 Hardware architecture of pre-compensation based on Jacobi solver. 76

List of Tables

2.1 Simulation Setup 19
2.2 Computational Complexity of Different ZF Detectors 28

4.1 Definition of cellular BSs investigated for power consumption. The macro reference design suffers 3-dB feeder losses between PA and antenna. 49
4.2 Digital complexity for scenario 2 (100×10 MaMi), in GOPS during the corresponding phase. 54
4.3 Power consumption of analog MaMi components based on scaled traditional architecture vs. alternative digital RF implementation prospects, per antenna assuming scenario 2 from Table 4.1. 54
4.4 High-level system parameters 57
4.5 Number of complex multiplications for MaMi system. 58
4.6 Estimated BW requirements for MaMi processing 61
4.7 Detailing Uplink (UL) and Downlink (DL) of BaseBand Processing 63
4.8 Overall communication time for single BUS MaMi system 67

5.1 Hardware results for IQ imbalance pre-compensation in 28 nm FD-SOI technology. 76
5.2 Energy-per-bit comparison for different precoding techniques to tackle various hardware aspects. 78

Chapter 1

Introduction

This deliverable serves as an update to Deliverable 3.1, with the focus on the distributed and centralized baseband processing algorithms, architectures, and platforms. We approach the target by first providing an overview on baseband processing algorithms developed during MAMMOET's first two years , followed by a discussions on the computational complexity profiling and the corresponding power consumption analysis when mapped to hardware platforms. The signal, noise, and interference power model is then studied to facilitate the system performance analysis. Before summarizing the deliverable, we also demonstrate hardware implementation results of selected baseband processing algorithms, using advanced Complementary Metal-Oxide Semiconductor (CMOS) technology.

Figure 1.1 depicts a simplified block diagram of MaMi baseband processing at the base station side. We take an Orthogonal Frequency-Division Multiplexing (OFDM) modulated system here for our introduction, however we believe the discussion can be conveniently extended to other modulation schemes. As can be seen, the digital baseband processing is divided into three parts, namely per-antenna processing, MIMO processing, and per-User Equipment (UE) processing. Per-antenna processing can be implemented as a near-antenna processor and mainly contains digital front-end and FFT/IFFT for OFDM. Gathering (distributing) data from (to) these per-antenna processors, the MIMO processing is responsible for channel estimation, uplink detection, reciprocity calibration, and downlink precoding. Per-UE processing decodes (generates) information for each user equipment and mainly includes coding/decoding, interleaving/de-interleaving, as well as mapping/de-mapping. It is worthwhile to be mentioned here that the processing distribution can be different depending on the selected algorithms, which will affect the overall complexity, data movement, and memory requirement. To obtain a balanced trade-off, this deliverable will investigate and evaluate different MaMi baseband processing strategies, including algorithm choices, processing partition, hardware architectures, and accelerator designs.

In Deliverable 3.1 [27], we demonstrated that the properties of MaMi allows many of the processing algorithms to be linear rather than non-linear, which helps to balance the computational complexity from increased parallel processing chains. In this deliverable, we further explore the unique features provided by MaMi and focus more on processing algorithms tackling practical issues in real-life deployment and hardware implementation. In Chapter 2, we discuss uplink detection algorithm with active multi-cell interference suppression, continuous-time constant-envelop precoding to enable extensive use of low-cost power amplifiers, reciprocity calibration allowing efficient TDD operation, as well as the assessment of system performance in the presence of hardware imperfection.

In wireless communication systems, the selection of processing algorithms is highly depended

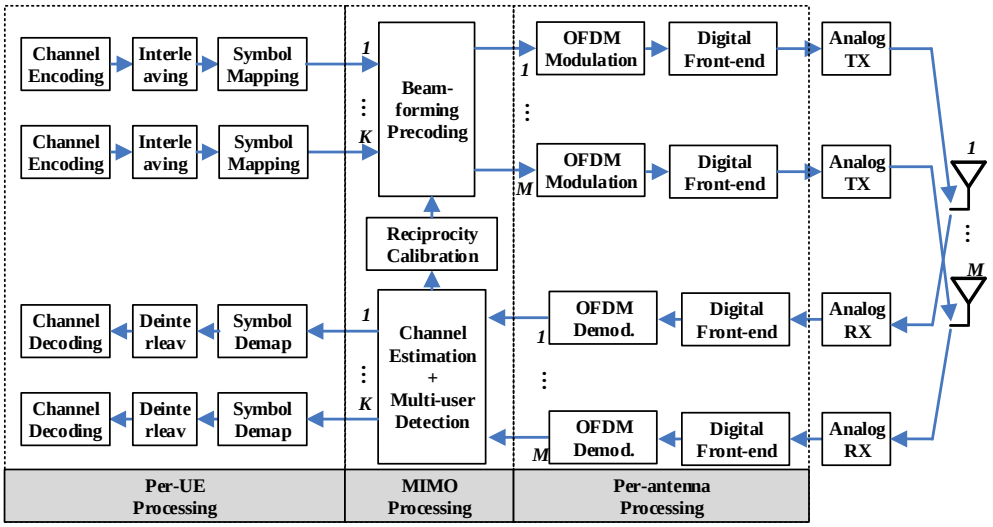


Figure 1.1: Simplified block diagram of MaMi base station baseband processing (with M base-station antennas serving K single-antenna UEs). An example of processing partition and distribution is also shown.

on the operating scenarios. This applies to MaMi systems as well. For example, initially the MaMi concept was proposed with Maximum Ratio (MR) processing, which has the benefit of low processing complexity. MR provides good performance in some use scenarios, while in others the more advanced ZF and Minimum Mean Square Error (MMSE) processing are needed. To facilitate the algorithm selection (or adaptation), Chapter 3 discusses the signal, noise, and interference power in different MaMi operating cases. Based on the model, we determine the bound on MR operation in terms of number of users and possible modulation and coding scheme.

MaMi has the potential of reducing the radiated power inversely proportionally to the square root of the number of base station antennas, or at an even faster pace, thanks to the coherent combination of all antennas. An important question is whether the processing power consumption related to the larger number of transceiver chains is not counterbalancing this benefit. Chapter 4 answers this question by profiling the MaMi processing in terms of computational complexity and hardware power consumption. Moreover, the impact (of different processing algorithms) on processing distribution strategy, data shuffling bandwidth, and memory requirement will also be discussed. The discussion can serve as a guideline to future hardware implementation of the MaMi baseband processor.

Chapter 5 presents the hardware implementation of key signal processing algorithms in MaMi base station. In this chapter, we further explore the MaMi channel matrix feature to implement low-cost and low-power precoders and detectors using algorithm-circuit co-optimization. In MaMi systems, low-cost RF chains can be employed to reduce the cost, however this may require additional baseband processing to handle induced distortions due to the hardware impairments. We analyze various such processing schemes and estimate the required processing energy per transmitted information bit. Simulation on gate-level show that the energy cost of performing pre-coding and tackling of hardware impairments are low.

Finally, a short summary of MaMi baseband processing profiling is given in Chapter 6.

In this deliverable MaMi has been analyzed with different focus in variable scenarios and system setups. Therefore, it is not very practical to have a unified system model covering all the aspects. Instead, we implement distributed system model descriptions localized in each section, which is a more convenient format when discussing particular problems.

Chapter 2

Baseband processing algorithm

2.1 Uplink detection with active multi-cell interference suppression

The MAMMOET D3.1 described the commonly studied methods for linear uplink detection, namely MR, ZF and MMSE. MR amplifies the desired signal, which is generally suboptimal in multi-user contexts but works relatively well in MaMi since the multi-user interference is suppressed by the favorable propagation that appears when having many antennas at the Base Station (BS) [4]. In contrast, ZF and MMSE suppress intra-cell interference *actively* by processing the received signals over the array to cancel interference in the spatial domain. Inter-cell interference is not actively suppressed in any of these detection schemes, but only suppressed by virtue of favorable propagation and the higher pathlosses to other cells.

In this section, we describe a new way to suppress also inter-cell interference actively. If the number of pilot sequences is τ_p, then each BS can locally estimate τ_p channel directions by listening to the pilot signalling from all cells instead of only from its own cell. Since the K users in a given cell only occupy K out of the τ_p channel directions, the serving BS can utilize the additional dimensions to select its detectors to also suppress inter-cell interference. The MAMMOET D1.1 described a new detection scheme called P-ZF, which exploits and orthogonalizes all τ_p directions to mitigate parts of the inter-cell interference. However, P-ZF only excels over conventional ZF in scenarios with very strong inter-cell interference; partly due to the loss in array gain of τ_p in P-ZF, instead of K as with conventional ZF, and partly because only cell-edge users of the neighboring cells need to be suppressed while more distant interferers already cause relatively weak interference.

In this section, we describe a new-state-of-the-art M-MMSE detection scheme that utilizes all τ_p pilots at each BS to actively suppress both intra-cell and inter-cell interference. It resembles the M-MMSE detector from [15] which was derived under perfect CSI, but we show how to utilize the pilot resources and suppress interference with M-MMSE detection under realistic considerations. A key property of the M-MMSE detector, as compared to the P-ZF detector, is that it provides soft interference suppression where cell-edge users of other cells are strongly suppressed and more distant interference sources are automatically less suppressed.

2.1.1 System model and transceiver design

To describe the M-MMSE scheme, we consider a synchronous MaMi network with multiple cells. Each cell is assigned with an index in the cell set $\mathcal{L}$, and the cardinality $|\mathcal{L}|$ is the number of cells. The BS in each cell is equipped with an antenna array of M antennas and serves

K single-antenna users within each coherence block. Assume that this time-frequency block consists of T_c seconds and W_c Hz, such that T_c is smaller than the coherence time of all users and W_c is smaller than the coherence bandwidth of all users. This leaves room for $\tau_c = T_c \times W_c$ transmission symbols per block, and the channels of all users remain constant within each block. Let $\mathbf{h}_{jlk} \in \mathbb{C}^M$ denote the channel response from user k in cell l to BS j within a block, and assume that it is a realization from a zero-mean circularly symmetric complex Gaussian distribution:

$$\mathbf{h}_{jlk} \sim \mathcal{CN}\left(\mathbf{0}, \beta_{jlk}\mathbf{I}_M\right), \tag{2.1}$$

where the variance β_{jlk} accounts for the channel attenuation (e.g., path loss and shadowing). The coherence block is divided into two parts: 1) uplink channel estimation, where each BS estimates the CSI from uplink pilot signalling which occupies τ_p out of τ_c symbols; 2) uplink payload data transmission phase, where each BS processes the received uplink signal from the remaining $\tau_c - \tau_p$ symbols.

Channel estimation

In the channel estimation phase, the collective received signal at BS j is denoted as $\mathbf{Y}_j \in \mathbb{C}^{M\times\tau_p}$ where τ_p is the length of the pilot sequences (it also equals to the number of orthogonal pilot sequences available in the network). Then $\mathbf{Y}_j$ can be expressed as

$$\mathbf{Y}_j = \sum_{l\in\mathcal{L}}\sum_{k=1}^{K}\sqrt{p_{lk}}\mathbf{h}_{jlk}\mathbf{v}_{i_{lk}}^T + \mathbf{N}_j, \tag{2.2}$$

where $\mathbf{h}_{jlk}$ is the channel response defined in (2.1), $p_{lk} \geq 0$ is the power control coefficient for the pilot of user k in cell l, and $\mathbf{N}_j \in \mathbb{C}^{M\times\tau_p}$ contains independent and identically distributed (i.i.d.) noise elements that follow $\mathcal{CN}(0,\sigma^2)$. We assume that all pilot sequences originate from a predefined orthogonal pilot book, defined as $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{\tau_p}\}$, where

$$\mathbf{v}_{b_1}^H\mathbf{v}_{b_2} = \begin{cases} \tau_p, & b_1 = b_2, \\ 0, & b_1 \neq b_2, \end{cases} \tag{2.3}$$

and let $i_{lk} \in \{1, \ldots, \tau_p\}$ denote the index of the pilot sequence used by user k in cell l. Arbitrary pilot reuse is supported here by denoting the relation between τ_p and K by $\tau_p = fK$, where $f \geq 1$ is the pilot reuse factor. If the pilots are allocated wisely in the network, a larger f brings a lower level of interference during pilot transmission, known as pilot contamination.

Based on the received signal in (2.2), the MMSE estimate of the uplink channel $\mathbf{h}_{jlk}$ is

$$\hat{\mathbf{h}}_{jlk} = \sqrt{p_{lk}}\beta_{jlk}\mathbf{Y}_j\left(\mathbf{\Psi}_j^*\right)^{-1}\mathbf{v}_{i_{lk}}^*, \tag{2.4}$$

where $\mathbf{\Psi}_j = \sum_{\ell\in\mathcal{L}}\sum_{m=1}^{K} p_{\ell m}\beta_{j\ell m}\mathbf{v}_{i_{\ell m}}\mathbf{v}_{i_{\ell m}}^H + \sigma^2\mathbf{I}_{\tau_p}$. We utilize that

$$\mathbf{v}_{i_{lk}}^H\mathbf{\Psi}_j^{-1} = \underbrace{\frac{1}{\sum_{\ell\in\mathcal{L}}\sum_{m=1}^{K} p_{\ell m}\beta_{j\ell m}\mathbf{v}_{i_{lk}}^H\mathbf{v}_{i_{\ell m}} + \sigma^2}}_{\alpha_{ji_{lk}}}\mathbf{v}_{i_{lk}}^H = \alpha_{ji_{lk}}\mathbf{v}_{i_{lk}}^H, \tag{2.5}$$

where $\alpha_{ji_{lk}}$ is a scalar, and according to the orthogonality principle of MMSE estimation, the covariance matrix of the estimation error $\tilde{\mathbf{h}}_{jlk} = \mathbf{h}_{jlk} - \hat{\mathbf{h}}_{jlk}$ is

$$\mathbf{C}_{jlk} = \mathbb{E}\left\{\tilde{\mathbf{h}}_{jlk}\tilde{\mathbf{h}}_{jlk}^H\right\} = \beta_{jlk}\left(1 - p_{lk}\beta_{jlk}\alpha_{ji_{lk}}\tau_p\right)\mathbf{I}_M. \tag{2.6}$$

As pointed out in [4], the part $\mathbf{Y}_j(\mathbf{\Psi}_j^*)^{-1}\mathbf{v}_{i_{lk}}^*$ of the MMSE channel estimate in (2.4) depends only on which pilot sequence that user k in cell l uses. Consequently, users who use the same pilot sequence have parallel estimated channels at each BS, while only the amplitudes are different in the estimates. To show this explicitly, define the $M \times \tau_p$ matrix

$$\hat{\mathbf{H}}_{\mathcal{V},j} = \left[\hat{\mathbf{h}}_{\mathcal{V},j1}, ..., \hat{\mathbf{h}}_{\mathcal{V},j\tau_p}\right] = \mathbf{Y}_j \left(\mathbf{\Psi}_j^*\right)^{-1} \left[\mathbf{v}_1^*, ..., \mathbf{v}_{\tau_p}^*\right], \tag{2.7}$$

which allows the channel estimate in (2.4) to be reformulated as

$$\hat{\mathbf{h}}_{jlk} = \sqrt{p_{lk}}\beta_{jlk}\hat{\mathbf{H}}_{\mathcal{V},j}\mathbf{e}_{i_{lk}}, \tag{2.8}$$

where $\mathbf{e}_i$ denotes the ith column of the identity matrix $\mathbf{I}_{\tau_p}$. The property that users with the same pilot have parallel estimated channels is the very essence of pilot contamination. Notice that the channel estimate $\hat{\mathbf{h}}_{jlk}$ is also a zero-mean complex Gaussian vector, with its covariance matrix being $\mathbf{\Phi}_{jlk} = p_{lk}\beta_{jlk}^2\alpha_{ji_{lk}}\tau_p\mathbf{I}_M \in \mathbb{C}^{M\times M}$. Define the covariance matrix of $\hat{\mathbf{h}}_{\mathcal{V},ji}$ as $\tilde{\mathbf{\Phi}}_{\mathcal{V},ji}$, we obtain $\tilde{\mathbf{\Phi}}_{\mathcal{V},ji} = \alpha_{ji}\tau_p\mathbf{I}_M$ according to (2.8).

Multi-cell MMSE detector

After the uplink channel estimation, during the uplink payload data transmission phase, the received signal $\mathbf{y}_j \in \mathbb{C}^{M\times 1}$ at BS j is

$$\mathbf{y}_j = \sum_{l\in\mathcal{L}}\sum_{k=1}^{K}\sqrt{p_{lk}}\mathbf{h}_{jlk}x_{lk} + \mathbf{n}_j, \tag{2.9}$$

where p_{lk} is the transmit power of the payload data from user k in cell l, $x_{lk} \sim \mathcal{CN}(0,1)$ is the transmitted signal from a Gaussian codebook, and $\mathbf{n}_j \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$ is additive white Gaussian noise (AWGN). Denote the linear detector used by BS j for an arbitrary user k in its cell as $\mathbf{g}_{jk} \in \mathbb{C}^M$, the detected signal $\hat{x}_{jk}$ is

$$\hat{x}_{jk} = \mathbf{g}_{jk}^H\mathbf{y}_j = \sqrt{p_{jk}}\mathbf{g}_{jk}^H\mathbf{h}_{jjk}x_{jk} + \mathbf{g}_{jk}^H\sum_{(l,m)\neq(j,k)}\sqrt{p_{lm}}\mathbf{h}_{jlm}x_{lm} + \mathbf{g}_{jk}^H\mathbf{n}_j. \tag{2.10}$$

By using (2.10), the following achievable ergodic Spectral Efficiency (SE) can be achieved for this user

$$R_{jk}^{\text{ul}} = \left(1 - \frac{\tau_p}{\tau_c}\right)\mathbb{E}_{\{\hat{\mathbf{h}}_{(j)}\}}\left\{\log_2\left(1+\eta_{jk}^{\text{ul}}\right)\right\}, \quad \text{[bit/s/Hz]} \tag{2.11}$$

where $\mathbb{E}_{\{\hat{\mathbf{h}}_{(j)}\}}$ denotes the expectation with respect to all the channel estimates obtained at BS j, and the Signal-to-interference-plus-noise ratio (SINR) η_{jk}^{ul} is given by

$$\eta_{jk}^{\text{ul}} = \frac{p_{jk}\mathbf{g}_{jk}^H\hat{\mathbf{h}}_{jjk}\hat{\mathbf{h}}_{jjk}^H\mathbf{g}_{jk}}{\mathbf{g}_{jk}^H\left(p_{jk}\mathbf{C}_{jjk} + \sum\limits_{(l,m)\neq(j,k)}p_{lm}\left(\hat{\mathbf{h}}_{jlm}\hat{\mathbf{h}}_{jlm}^H + \mathbf{C}_{jlm}\right) + \sigma^2\mathbf{I}_M\right)\mathbf{g}_{jk}}, \tag{2.12}$$

where $\mathbb{E}\{\cdot|\hat{\mathbf{h}}_{(j)}\}$ denotes the conditional expectation given all the estimated channels at BS j. Due to the fact that only the imperfectly estimated channels are available, the SE in (2.11) is achieved by treating $\mathbf{g}_{jk}^H\hat{\mathbf{h}}_{jjk}$ as the true channel, and treating uncorrelated interference and

channel uncertainty as worst-case Gaussian noise. Thus, R_{jk}^{ul} is a lower bound on the uplink ergodic capacity.

The second line of (2.12) shows that the uplink SINR takes the form of a generalized Rayleigh quotient. Therefore, the M-MMSE detector can be derived to maximize this SINR for given channel estimates:

$$\mathbf{g}_{jk}^{\mathrm{M-MMSE}} = \left(\hat{\mathbf{H}}_{\mathcal{V},j}\boldsymbol{\Lambda}_j\hat{\mathbf{H}}_{\mathcal{V},j}^H + \left(\sigma^2 + \varphi_j\right)\mathbf{I}_M\right)^{-1}\hat{\mathbf{h}}_{jjk}, \tag{2.13}$$

where $\boldsymbol{\Lambda}_j = \sum_{l\in L}\sum_{k=1}^{K} p_{lk}^2\beta_{jlk}^2\mathbf{e}_{i_{lk}}\mathbf{e}_{i_{lk}}^H$ is a diagonal matrix, and its ith diagonal element λ_{ji} depends on the large scale fading, the pilot and payload power of the users that use the ith pilot sequence in $\mathcal{V}$. The scalar φ_j is defined as

$$\varphi_j = \sum_{l\in\mathcal{L}}\sum_{k=1}^{K} p_{lk}\beta_{jlk}(1 - p_{lk}\beta_{jlk}\alpha_{ji_{lk}}\tau_p),$$

where $\alpha_{ji_{lk}}$ is defined in (2.5). As the name suggests, this detector also minimizes the Mean Square Error (MSE) in estimating x_{jk}, $\mathbb{E}\{|\hat{x}_{jk} - x_{jk}|^2|\hat{\mathbf{h}}_{(j)}\}$.

Compared with the M-MMSE detector proposed in [15], our detector seems similar to it at first glance, since both of them try to suppress the inter-cell interference. However, the difference is substantial. With perfect CSI, the detector in [15] is able to suppress the interference from all user channel directions, since the small scale fading realizations of the users are likely to be different. However, the promised performance is vastly over-optimistic. Firstly, the performance loss from the CSI estimation errors need to be taken into account in practice. Secondly, with limited pilot resources, the number of distinguishable channel directions that can be learned locally at a BS is much smaller than the number of users in the network. Thus, only part of the inter-user interference can be actively suppressed, and the performance loss from the inability to mitigate the remaining interference should also be modeled and minimized. Therefore, our detector is not a simple extension from a perfect CSI-based detector to a imperfect CSI-based one. It shows the way to optimally utilize the available resources and suppress interference under realistic and important considerations: the limitation of the pilot resources as well as the necessity of channel estimation.

Since τ_p instead of K directions need to be calculated in the M-MMSE detector, the complexity increase over the conventional S-MMSE detector is about $4(f-1)fMK^2$ real number multiplications and real number additions. Since in MaMi systems $M \gg K$ is often assumed, the complexity increase is not a big issue when K has a small or moderate value. The M-MMSE scheme can be seen as a coordinated beamforming scheme, but since there is no need for rapid signalling between the BSs (BS j estimates $\hat{\mathbf{H}}_{\mathcal{V},j}$ from the uplink pilots), the M-MMSE scheme is fully scalable. The pilot allocation can either be optimized across cells, which requires some inter-cell signaling, or the pilot sequences can be the reused across the cells in a fixed manner; see Fig. 2.1 for an example with a fixed pilot reuse patterns.

2.1.2 Simulation results

In this section, we illustrate the benefit of the M-MMSE detection scheme for a symmetric hexagonal network topology. We apply the classic 19-cell-wrap-around structure to avoid edge effects and guarantee consistent simulated performance for all cells. Each hexagonal cell has a radius of $r = 500$ meters, and is surrounded by 6 interfering cells in the first tier and 12 in the

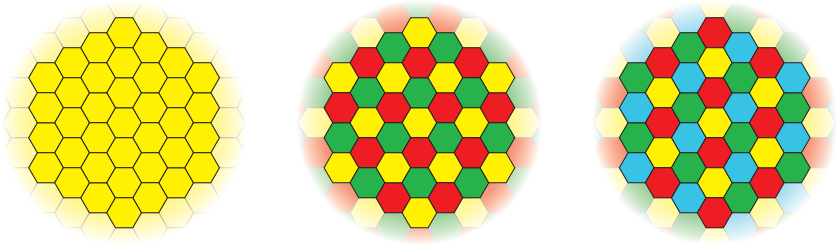


Figure 2.1: The 19-cell-wrap-around hexagonal network topology for $f = 1$, $f = 3$, and $f = 4$.

second tier. To achieve a symmetric pilot allocation in this network, the pilot reuse factor can be $f \in \{1, 3, 4, 7\}$; see Figure 2.1 for an example of different reuse factors. For each pilot reuse policy, the same subset of pilots are allocated to the cells with the same color, and pilots in each cell are allocated randomly to the users.

The user locations are generated independently and uniformly at random in the in cells, but the distance between each user and its serving BS is at least $0.14r$. For each user a simple pathloss model is considered, where the variance of the channel attenuation is computed as C divided by the propagation distance to the exponent κ, where $C > 0$ models independent shadow fading with $10 \log_{10}(C) \sim \mathcal{N}(0, \sigma_{sf}^2)$. In the simulation, we assume $\kappa = 3.7$, $\sigma_{sf}^2 = 5$ and the coherence block length $\tau_c = 1000$.

Benefits of the proposed M-MMSE scheme

Next, we compare the new M-MMSE detector with the conventional alternatives. Statistical channel inversion power control is applied to the pilot and payload data, i.e., $p_{lk} = \frac{\rho}{\beta_{llk}}$ [4]. Thus, the average effective channel gain between users and their serving BSs is constant: $\mathbb{E}\{p_{lk} \|\mathbf{h}_{llk}\|^2\} = M\rho$. Then the average uplink SNR per antenna and user at its serving B-S is ρ/σ^2. This is a simple but effective policy to avoid near-far blockage and, to some extent, guarantee a uniform user performance in the uplink. In our simulation, ρ/σ^2 is set to 0 dB to allow for decent channel estimation accuracy, and the SEs are all multiplied by 1/2 to model a 50% time portion of uplink transmission.

To show explicitly the advantages of our M-MMSE scheme, simulation results for the MR detector from [28], the P-ZF scheme from [4], and the S-MMSE scheme from [18] are provided for comparison. Notice that $M - fK > 0$ is needed for the P-ZF scheme, thus the minimum value of M for the P-ZF is $fK + 1$. Simulation results are shown in Figs. 2.2 – 2.4 for $f = 1$, $f = 4$ and $f = 7$, respectively. The MR detector always achieves the lowest performance since it does not suppress any interference. Compared to S-MMSE, our proposed M-MMSE always achieves a higher sum SE, and the advantage becomes more significant as f and/or K increases. For $f = 4$ and $M = 200$, the SE of M-MMSE are 31% and 53% higher than those of S-MMSE for $K = 10$ and $K = 30$, respectively. For $f = 7$, the gains increase to 42% and 82% for $K = 10$ and $K = 30$, respectively. The higher performance gain with a larger K or f comes from the fact that more residual directions can be learned and utilized for interference suppression by the M-MMSE, while the S-MMSE always uses K directions regardless of f. The advantage of the M-MMSE over the P-ZF is only minor for small f and small K, but the gain becomes notable as f and K grow. Since the complexity of our M-MMSE scheme is essentially the same as for the P-ZF, and the P-ZF can sometimes achieve very low SE for small M, in general our scheme is the better choice if high system SE is desirable.

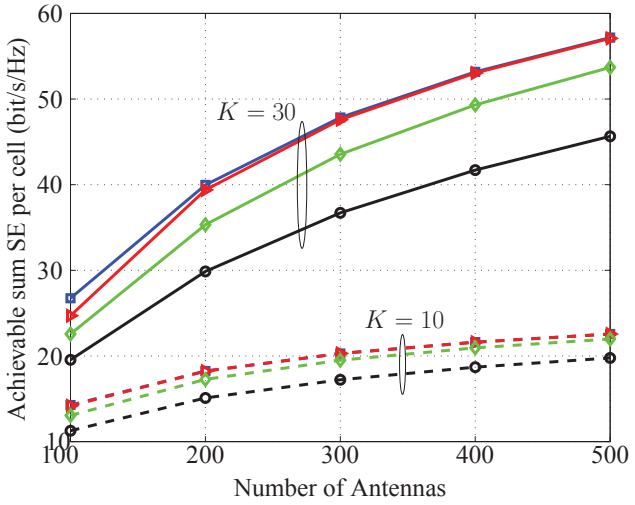


Figure 2.2: Achievable sum SE of M-MMSE (squares), P-ZF (triangles), S-MMSE (diamonds) and MR (circles) with $f = 1$, $K = 10$ and $K = 30$.

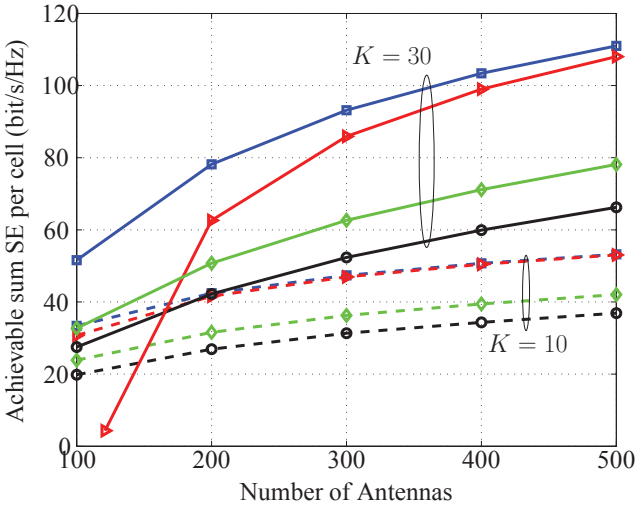


Figure 2.3: Achievable sum SE of M-MMSE (squares), P-ZF (triangles), S-MMSE (diamonds) and MR (circles) with $f = 4$, $K = 10$ and $K = 30$.

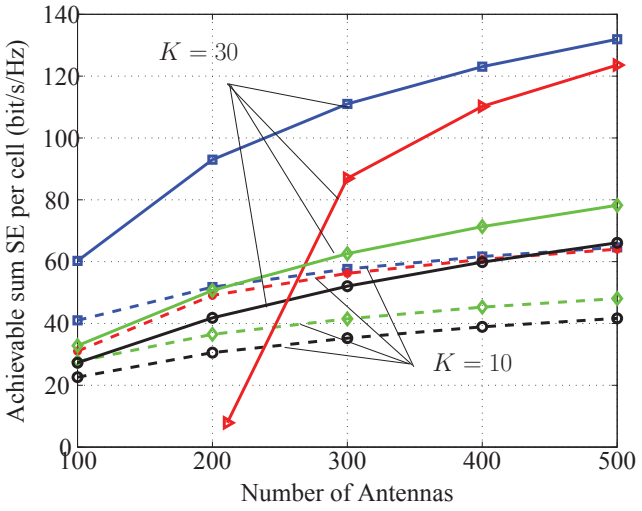


Figure 2.4: Achievable sum SE of M-MMSE (squares), P-ZF (triangles), S-MMSE (diamonds) and MR (circles) with $f = 7$, $K = 10$ and $K = 30$.

2.1.3 Extensions to multi-cell downlink precoding

It was shown in [4] that when each precoder is a scaled version of the corresponding detector, the same per user SEs as in the uplink can be achieved in the downlink by properly selecting the downlink payload power. This is known as uplink-downlink duality. Hence, the M-MMSE detector can also be used as a basis for creating a downlink M-MMSE precoder. This is analyzed in further detail in the MAMMOET publication [26], but is not included here since the uplink-downlink duality implies that the same performance can be achieved in the downlink.

2.2 Reciprocity calibration

The channel hardening in MaMi systems makes the effective precoded downlink channel gains very stable over the time and frequency domain. This may render explicit downlink channel estimation unnecessary [25], by operating in Time Division Duplex (TDD) mode and relying on the reciprocity of the propagation channel to compute proper precoding coefficients based on uplink channel estimates. However in most practical systems, the channel is composed by the the cascade of the transmitter analog front-end response, propagation channel, and receive analog front-end response. While the propagation channel is assumed to be reciprocal, the analog front-ends are not. Hence, in order to use reciprocity and calculate the precoding coefficients, one needs to estimate and compensate, i.e., calibrate, for the differences of the transceivers front-end responses.

2.2.1 System model

Let the estimated uplink radio channel from K users to M BS antennas for the case of a narrow-band MaMi transmission be modeled as

$$\tilde{\mathbf{H}}_{\mathrm{UP}} = \mathbf{H}_{\mathrm{UP}} + \mathbf{N} = \mathbf{R}_B \mathbf{H}_P \mathbf{T}_U + \mathbf{N}, \tag{2.14}$$

where $\mathbf{R}_B = \mathrm{diag}(r_1^{\mathrm{B}} \dots r_M^{\mathrm{B}})$ and $\mathbf{T}_U = \mathrm{diag}(t_1^{\mathrm{U}} \dots t_K^{\mathrm{U}})$ are diagonal matrices, with r_m^{B} and t_m^{U} denoting the response of the BS receiver $1 \leq m \leq M$ and terminal transmitter $1 \leq k \leq K$,

respectively, $\mathbf{H}_P$ is the propagation channel matrix with random entries which are assumed to share the same coherence time T_{coh}, and $\mathbf{N}$ is a matrix with random entries modelling uplink noise. The associated downlink radio channel to (2.14) can be written as

$$\tilde{\mathbf{H}}_{\mathrm{DL}} = \mathbf{H}_{\mathrm{DL}} + \mathbf{N}' = \mathbf{R}_U \mathbf{H}_P^T \mathbf{T}_B + \mathbf{N}', \tag{2.15}$$

where $\mathbf{R}_U = \mathrm{diag}(r_1^{\mathrm{U}} \dots r_K^{\mathrm{U}})$ and $\mathbf{T}_B = \mathrm{diag}(t_1^{\mathrm{B}} \dots t_M^{\mathrm{B}})$ where r_k^{U} and t_m^{B} denote the response of the terminal receiver k and BS transmitter m, respectively, and the entries of $\mathbf{N}'$ model downlink noise. The underlying assumption is that (2.14) and (2.15) model uplink and downlink channels that occur within a time interval much smaller than T_{coh}, such that the propagation conditions are essentially the same in both cases.

Assume that an error-free uplink channel estimate is at hand, and hence $\mathbf{H}_{\mathrm{UP}}$ available for precoding purposes. Multiplying the uplink channel at antenna $1 \leq m \leq M$ with the ratio $\alpha t_m^B (r_m^B)^{-1} = \alpha c_m$ where $\alpha \in \mathbb{C} \setminus \{0\}$, provides calibrated version of the radio downlink channel which can be written as

$$\begin{aligned} \mathbf{H}_{\mathrm{DL}}^{\mathrm{CAL}} =& \left(\left(\alpha \mathbf{T}_B \mathbf{R}_B^{-1} \right) \mathbf{H}_{\mathrm{UP}} \right)^T \\ =& \alpha \mathbf{T}_U \mathbf{H}_P^T \mathbf{T}_B, \end{aligned} \tag{2.16}$$

If (2.16) is used for linear precoding purposes, i.e., we build a tunable precoding matrix

$$\mathbf{W}(\rho)^{CAL} = \left(\mathbf{H}_{\mathrm{DL}}^{\mathrm{CAL}} \right)^H \left(\rho \, \mathbf{H}_{\mathrm{DL}}^{\mathrm{CAL}} \left(\mathbf{H}_{\mathrm{DL}}^{\mathrm{CAL}} \right)^H + (1-\rho)(N_0 \mathbf{I}) \right)^{-1}, \tag{2.17}$$

where $0 < \rho \leq 1$ is the tunable parameter and $0 \leq N_0 < \infty$. If ZF precoding is performed, i.e. $\rho = 1$, the equivalent error-free downlink channel (the cascade of precoder and an error-free version of the radio downlink channel) can be written as

$$\begin{aligned} \mathbf{H}_{\mathrm{DL}} \mathbf{W}(1)^{\mathrm{CAL}} =& \mathbf{R}_U \mathbf{H}_P^T \mathbf{T}_B \mathbf{W}(1)^{\mathrm{CAL}} \\ =& \alpha \mathbf{R}_U \mathbf{T}_U^{-1}, \end{aligned} \tag{2.18}$$

which achieves a desired diagonal form. This enables downlink spatial multiplexing with ideally no interference. However, noting that (2.18) is not the identity matrix but a diagonal matrix made of the hardware responses of the users' terminals, scarce downlink pilots that can be shared among users need to be broadcasted through the beam to equalize this uncertainty [25].

2.2.2 Calibration procedure

The problem of estimating the calibration coefficients $\{c_m\}$ with a prototyping MaMi platform was addressed in the MAMMOET D3.1. The proposed calibration process consisted in: (1) estimating the channels between all BS transceiver units, (2) processing the estimated channels in order to estimate and compensate for the differences between the transmitter and receiver analog front-ends. The analysis conducted in [37] revealed that mutual coupling between BS antennas can be conveniently exploited to estimate the calibration coefficients $\{c_m\}$. Moreover, its was verified that good performance at low signal-to-noise ratio during calibration is achieved if the set of signals is reduced to only measurements between adjacent antennas.

2.2.3 Implementation in the LuMaMi testbed

Validation of the reciprocity calibration procedure proposed in [37], using only adjacent channels estimates between BS antennas, was performed by implementing it in a MaMi testbed

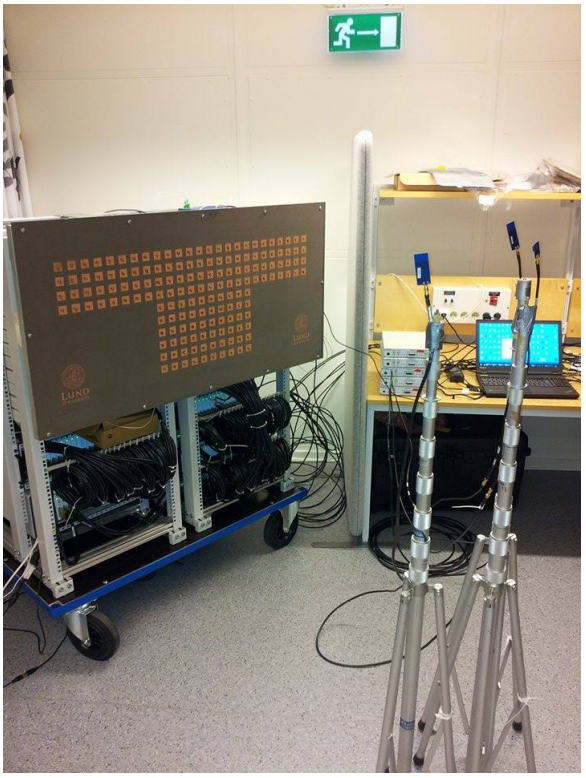


Figure 2.5: MaMi physical setup used to validate reciprocity calibration. Three very closely spaced single-antenna terminals yielding strong LoS propagation channels to the BS.

prototype, namely the LuMaMi testbed [20]. Once the calibration coefficients were attained, we performed a downlink MaMi transmission from 50 BS antennas to three single-antenna mobile stations in our lab, as proof-of-concept. The physical setup used is shown in Figure 2.5. It consists on one of the hardest propagation scenarios in terms of inter-user separability [14], which we emulate to showcase the validity our reciprocity calibration methodology.

Once the BS estimates the calibration coefficients, the transmission protocol is as follows. Using OFDM based signalling with similar parametrization as in [20], we let users transmit orthogonal pilots in frequency in the same time slot, see [20] for more details in the frequency structure of the pilots. The BS performs channel estimation per user, by interpolating on non-estimated subcarrier channels. Once uplink channel estimation of all users is complete, reciprocity calibration on a subcarrier basis is performed, by multiplying each channel estimate by the respective calibration coefficients, as in 2.16. The calibrated version of the downlink channel is then used for beamforming using ZF precoder.[1]

The baseband processing was implemented solely at the Central Controller (CC) of the BS, and in the CPU of the terminal, since no real-time constraints are to be met. We took this provisional approach to be able to showcase a massive downlink MIMO transmission. All baseband processing will, however, be moved to the FPGAs in subsequent work. Data symbols are transmitted at a rate which can be handled by the CC and terminals' CPUs. We made an effort to keep an invariant channel between the uplink pilots and the downlink data transmission. Figure 2.6 illustrates a realization of the downlink equalized signal points, for the cases of ZF precoding and identity precoding. For the case of identity precoding, inter-user interference constrains the performance. Noticeable, ZF precoding using calibrated uplink channel estimates allows users separability. Per-user Error Vector Magnitude (EVM)s down to

[1]The performance of ZF is known to be sensitive to calibration errors.

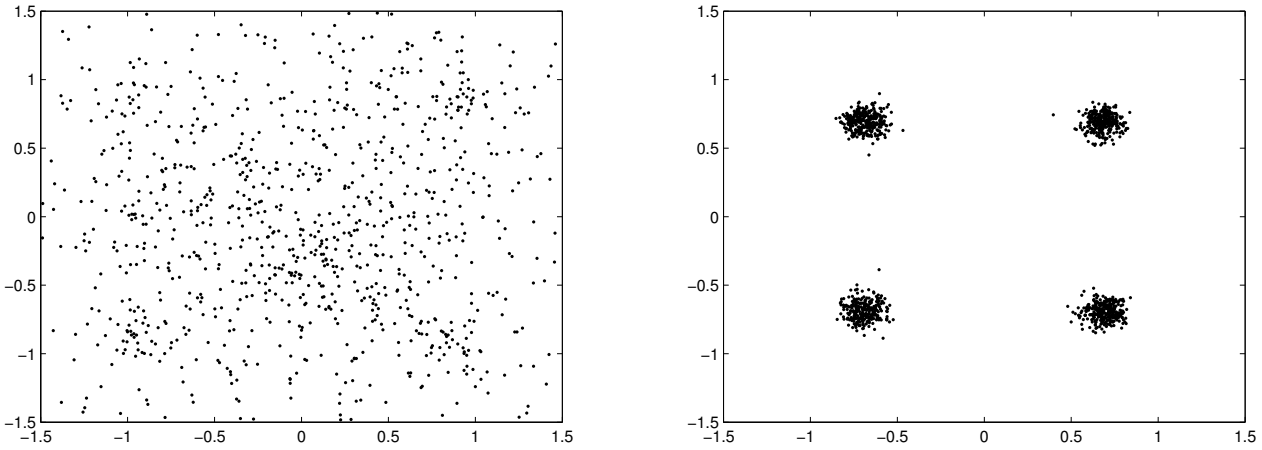


Figure 2.6: Left: Equalized downlink signals at one of three users for the case of when the precoding matrix is the identity matrix. Right: Equalized downlink signals at one of three users for the case of ZF precoding.

−10 dB were obtained.

Future work on this matter will quantize the previous experiment by studying EVM trade-offs between the calibration coefficients error and the error in the uplink channel estimate. Modelling on the calibration coefficients, based on measured data from our MaMi testbed prototype will also be performed.

2.3 Downlink precoding

The conventional precoding schemes are MR, ZF, and regularized ZF (also known as MMSE), which were previously described in the MAMMOET D3.1. In addition, the Discrete-Time Constant-Envelope (DTCE) precoding scheme was described in D3.1. This a convenient method to reduce the PAPR in the downlink transmission. The reduction in PAPR is achieved by requiring that the time-discrete signals emitted from each antenna has constant envelope. Although there is no power variations in the time-discrete signal, this signal is used to generate a continuous-time signal and this signal will generally have power variations. In this section we describe a new method to achieve CTCE precoding that deals with this issue, to achieve an even lower PAPR.

2.3.1 System model

The downlink single-cell transmission from a BS with M antennas to K single-antenna users is studied. Let $x_m(t)$ be the complex baseband transmit signal from antenna m and where t is a continuous time variable. Then the received signal $r_k(t)$ at user k is given by

$$r_k(t) = \sqrt{P} \sum_{m=1}^{M} \big(h_{km}(\tau) \star x_m(\tau)\big)(t) + w_k(t), \tag{2.19}$$

where $h_{km}(\tau)$ is the impulse response of the channel between antenna m and user k, $w_k(t)$ is a complex white Gaussian noise process with zero mean and spectral height N_0 that is independent

of the transmit signals $\{x_m(t)\}$ and the channel $\{h_{km}(\tau)\}$. The power of the transmit signals should fulfill

$$\mathbb{E}\Big[\,\big|x_m(t)\big|^2\,\Big] = \frac{1}{M}. \tag{2.20}$$

The factor P therefore represents the total radiated power.

Each user is equipped with a filter with impulse response $p(\tau)$ that is chosen such that

$$\int_{-\infty}^{\infty} \big|p(\tau)\big|^2 \mathrm{d}\tau = 1/T. \tag{2.21}$$

The received signal is filtered by $p(\tau)$ and uniformly sampled to produce the received samples

$$r_k[n] = y_k[n] + w_k[n], \quad n = 0, \ldots, N{-}1. \tag{2.22}$$

Each of these samples is the sum of two parts: the noise-free signal and a noise sample that is independent of the signal:

$$y_k[n] = \sqrt{P} \int_{-\infty}^{\infty} p(\tau) \sum_{m=1}^{M} \big(h_{km}(t) \star x_m(t)\big)(nT{-}\tau)\mathrm{d}\tau, \tag{2.23}$$

$$w_k[n] = \int_{-\infty}^{\infty} p(\tau) w_k(nT-\tau)\mathrm{d}\tau. \tag{2.24}$$

Each user thus observes N samples. The sampling period T will be referred to as the *symbol period*. It is assumed that the impulse response $p(\tau)$ is a root-Nyqvist pulse of period T; then the noise samples are i.i.d. $\mathcal{CN}(0, N_0/T)$.

2.3.2 The constant-envelope MIMO channel

Let the transmit signals $\{x_m(t)\}$ be stochastic processes and assume that they are of some *operational power spectral density* $S_x(f)$ [24].

Definition 1 *A* continuous-time constant-envelope signal *is a stochastic process that fulfills*

$$|x_m(t)|^2 = \frac{1}{M}, \quad \forall t, \tag{2.25}$$

almost surely.

Since the only strictly bandlimited signals that have property (2.25) are pure sinusoids, a relaxed measure of bandwidth will be used.

Definition 2 *The* δ-bandwidth *with respect to the symbol rate* $1/T$ *of the process* $x_m(t)$ *is*

$$B = \inf\{B' \geq 0 : S_x(f) < P_0/\delta, \quad \forall |f| > B'/2\}, \tag{2.26}$$

where the in-band power is given by

$$P_0 = T \int_{-1/(2T)}^{0^-} S_x(f)\mathrm{d}f + T \int_{0^+}^{1/(2T)} S_x(f)\mathrm{d}f. \tag{2.27}$$

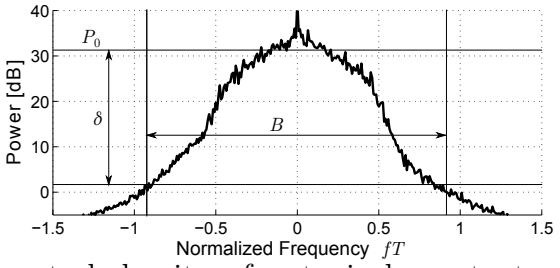


Figure 2.7: The power spectral density of a typical constant-envelope signal. Its 30 dB-bandwidth $BT = 1.8$ is indicated.

The power at $f = 0$ is excluded in the in-band power to allow for signals with a nonzero mean. The 30 dB-bandwidth measure is illustrated in Figure 2.7. The fraction of the bandwidth greater than the symbol rate

$$TB - 1, \tag{2.28}$$

will be referred to as the *excess δ-bandwidth*.

Now the channel that is studied in this section can be defined in terms of the two previous definitions.

Definition 3 *The $M \times K$* continuous-time constant-envelope MIMO broadcast channel *of δ-bandwidth B_{max} is the channel described in* (2.19) *between M antennas, which only emit continuous-time constant-envelope transmit signals $\{x_m(t)\}$ with δ-bandwidth smaller than or equal to B_{max}, and each of the K single-antenna users that receive the signals $\{r_k(t)\}$.*

2.3.3 Continuous-time constant-envelope precoding

To lower-bound the sum-capacity of the constant-envelope channel detailed in Section 2.3.2, a transmission scheme for a MaMI downlink channel that uses transmit signals with constant envelopes is presented here. The proposed precoder will be called the CTCE precoder.

CTCE precoding

The random symbol intended for user k at sample instant n is denoted by $u_k[n]$, for all $k = 1, \ldots, K$ and $n = 0, \ldots, N-1$. The symbols are required to have unit energy

$$\mathbb{E}\big[\,|u_k[n]|^2\,\big] = 1, \quad \forall n, k. \tag{2.29}$$

Given the parameters $\gamma, \lambda_1, \lambda_2 \in \mathbb{R}^+$, we choose the transmit signals, for each realization of the channel and the random symbols, to be a continuous solution to

$$\min_{\{x_m(t)\}} \left(\sum_{k=1}^{K} \sum_{n=0}^{N-1} \left| y_k[n] - \sqrt{\gamma P} u_k[n] \right|^2 + \lambda_1 \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left| \frac{\mathrm{d}}{\mathrm{d}t} x_m(t) \right|^2 \mathrm{d}t + \lambda_2 \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left| \frac{\mathrm{d}^2}{\mathrm{d}t^2} x_m(t) \right|^2 \mathrm{d}t \right) \tag{2.30}$$

subject to the modulus constraint $|x_m(t)| = 1/\sqrt{M}$, $\forall m, t$.

The precoder given by (2.30) minimizes the mismatch between the actual received sample and the desired symbol and lets each user receive a new symbol every instant $t = nT$. The

power of the desired symbols is determined by the parameter γ. The two latter terms in (2.30) serve the purpose of regularizing the first and second derivatives of the transmit signals in order to produce smooth signals. By choosing the regularizing factors λ_1 and λ_2 large enough, it has been observed numerically that the resulting solution has a limited δ-bandwidth. The parameters γ, λ_1 and λ_2 will be chosen to maximize the sum-rate and to fulfill a bandwidth requirement; see Section 2.3.3.

The optimization problem (2.30) can be approximately solved in discrete time by expressing each $y_k[n]$ in terms of sampled versions of the transmit signals $\{x_m(t)\}$. If the sampling rate is high enough, there are constant-envelope modulation schemes that produce continuous-time signals with limited bandwidth from the discrete-time solution, see Section 2.3.3.

The noise-free received sample in (2.23) can be rewritten as

$$y_k[n] = \sqrt{P}\sum_{m=1}^{M} \underbrace{\int\int p(\check{\tau})h_{km}(\tau-\check{\tau})\mathrm{d}\check{\tau}}_{=f_{km}(\tau)} x_m(nT-\tau)\mathrm{d}\tau. \tag{2.31}$$

The inner integral $f_{km}(\tau) = \big(p(t) \star h_{km}(t)\big)(\tau)$ could be estimated by letting the users send uplink pilots. Here, however, it is assumed that $f_{km}(\tau)$ is perfectly known by the BS.

Denote the κ-times oversampled (with respect to the symbol period T) signals

$$x_m[\nu] = x_m(\nu T/\kappa), \tag{2.32}$$
$$f_{km}[\nu] = \tfrac{T}{\kappa} f_{km}(\nu T/\kappa). \tag{2.33}$$

It is assumed that there exist integers $\ell_{\min}$ and $\ell_{\max}$ such that $f_{km}(\tau)$ practically is zero for τ outside $[\ell_{\min}T/\kappa, \ell_{\max}T/\kappa]$. In what follows, only the samples $f_{km}[\nu]$ with indices $\nu \in [\ell_{\min}, \ell_{\max}]$ will be considered.

The aggregate channel impulse response $f_{km}(\tau)$ is bandlimited to κ/T if the impulse response $p(\tau)$ is. Assume that the transmit signal $x_m(t)$ is bandlimited to κ/T too; then (2.31) can be written in terms of the two discrete-time signals $x_m[\nu]$ and $f_{km}[\nu]$. Even if the modulated transmit signal, being a constant-envelope signal with non-constant phase derivative, is not strictly bandlimited, it is practically bandlimited to κ/T for some κ when the parameters λ_1 and λ_2 are large enough, as will be shown in Section 2.3.4. Therefore by choosing κ big enough, the received signal is approximately given by

$$y_k[n] \approx \sqrt{P}\sum_{m=1}^{M}\sum_{\ell=\ell_{\min}}^{\ell_{\max}} f_{km}[\ell]x_m[n\kappa-\ell]. \tag{2.34}$$

Denote by $\zeta_k[n]$ the argument of the modulus operator in the first term in (2.30) (divided by $\sqrt{P}$):

$$\zeta_k[n] = \sqrt{\gamma}u_k[n] - \sum_{m=1}^{M}\sum_{\ell=\ell_{\min}}^{\ell_{\max}} f_{km}[\ell]x_m[n\kappa-\ell]. \tag{2.35}$$

By using the first-order approximations of the first and second derivatives

$$\frac{\mathrm{d}}{\mathrm{d}t}x_m(t)\bigg|_{t=\frac{\nu T}{\kappa}} \approx \frac{x_m[\nu-1]-x_m[\nu]}{T/\kappa} = x'_m[\nu], \tag{2.36}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}x_m(t)\bigg|_{t=\frac{\nu T}{\kappa}} \approx \frac{x_m[\nu-1]-2x_m[\nu]+x_m[\nu+1]}{(T/\kappa)^2} = x''_m[\nu],$$

the optimization problem (2.30) can be approximated as

$$\min_{|x_m[\nu]|=\frac{1}{\sqrt{M}}} \sum_{k,n} \left|\zeta_k[n]\right|^2 + \lambda_1 \sum_{m,\nu} \left|x'_m[\nu]\right|^2 + \lambda_2 \sum_{m,\nu} \left|x''_m[\nu]\right|^2. \tag{2.37}$$

Only the transmit samples $x_m[\nu]$ with indices

$$\nu = -\ell_{\max}, \ldots, (N-1)\kappa - \ell_{\min} \tag{2.38}$$

influence the received samples. These are the samples that are optimized with respect to in (2.37).

The objective function in (2.37) is non-convex and the optimization is hard to solve explicitly. By using a technique similar to the one used in [29], a solver that uses cyclic optimization can be devised by observing that the problem can be explicitly solved for one sample $x_{\tilde{m}}[\tilde{\nu}]$ by

$$x_{\tilde{m}}[\tilde{\nu}] = \frac{1}{\sqrt{M}} \frac{z^*_{\tilde{m}}[\tilde{\nu}]}{|z_{\tilde{m}}[\tilde{\nu}]|}, \tag{2.39}$$

where $z_{\tilde{m}}[\tilde{\nu}] = z_1 + z_2 + z_3$ is the sum of the three terms

$$z_1 = \sum_{k=1}^{K} \sum_{n=\underline{n}}^{\bar{n}} f^*_{k\tilde{m}}[n\kappa - \tilde{\nu}] \big(\zeta_k[n] + f_{k\tilde{m}}[n\kappa - \tilde{\nu}] x_{\tilde{m}}[\tilde{\nu}]\big), \tag{2.40}$$

$$z_2 = \tfrac{\lambda_1 \kappa^2}{T^2} \big(x^*_{\tilde{m}}[\tilde{\nu}-1] + x^*_{\tilde{m}}[\tilde{\nu}+1]\big), \tag{2.41}$$

$$z_3 = \tfrac{\lambda_2 \kappa^4}{T^4} (4x^*_m[\nu-1] - x^*_m[\nu-2] + 4x^*_m[\nu+1] - x^*_m[\nu+2]).$$

The limits in (2.40) are given by

$$\underline{n} = \max\big(\lceil (\ell_{\min} + \tilde{\nu})/\kappa \rceil, 0\big), \tag{2.42}$$

$$\bar{n} = \min\big(\lfloor (\ell_{\max} + \tilde{\nu})/\kappa \rfloor, N-1\big). \tag{2.43}$$

Since the objective function in (2.37) does not increase when a signal sample $x_{\tilde{m}}[\tilde{\nu}]$ is set to its optimum value (2.39), an algorithm that sets the signal samples one-by-one to their optimal values by letting the indices cyclically run through

$$\begin{aligned}(\tilde{m}, \tilde{\nu}) : (1, -\ell_{\max}) &\to (2, -\ell_{\max}) \to \cdots \to (M, -\ell_{\max}) \\ &\to (1, 1-\ell_{\max}) \to \cdots \to (M, 1-\ell_{\max}) \\ &\to \cdots\cdots \to (M, (N-1)\kappa - \ell_{\min})\end{aligned} \tag{2.44}$$

a couple of rounds will make the objective function (2.37) converge to a local minimum. How many cycles are needed depends on the parameters λ_1 and λ_2. For small λ_1 and λ_2, which correspond to larger bandwidth requirements, the optimization converges in 5-10 rounds. The greater λ_1 and λ_2, and narrower bandwidths, the more rounds are needed. To produce steep and narrow spectra for tough bandwidth requirements, as many as 100 rounds might be needed. How close the local minimum, which the algorithm converges to, is to the global optimum depends on the initialization of the samples.

Many initialization methods have been tested, such as: setting $x_m[\nu] = 0$ at first, to initialize $x_m[\nu]$ to have the same phase for all m and ν, to initialize with random phases, etc. The best method that was found, which is the method later used in Section 2.3.4, is to successively increase the oversampling factor κ. An initial optimization is done with oversampling factor

$\kappa_0 = 1$ over the signals $\{x_m^{(0)}[\nu] = x_m(\nu T/\kappa_0)\}$ by initializing $x_m^{(0)}[\nu] = 1/\sqrt{M}$. The so-obtained solution is used to initialize a second optimization over the $\kappa_1 = 2$-times oversampled signals $x_m^{(1)}[\nu] = x_m(\nu T/\kappa_1)$. The result is again used to initialize a further optimization with a higher oversampling factor. The i-th optimization is done over the $\kappa_i = 2^i$-times oversampled signals

$$x_m^{(i)}[\nu] = x_m(\nu T/\kappa_i) \tag{2.45}$$

that are initialized by

$$x_m^{(i)}[\nu] = \begin{cases} x_m^{(i-1)}[\nu/2], & \text{if } \nu \text{ is even,} \\ \frac{1}{\sqrt{M}} \frac{x_m^{(i-1)}[(\nu-1)/2] + x_m^{(i-1)}[(\nu+1)/2]}{\left| x_m^{(i-1)}[(\nu-1)/2] + x_m^{(i-1)}[(\nu+1)/2] \right|}, & \text{if } \nu \text{ is odd.} \end{cases}$$

The optimization procedure is terminated after a high enough oversampling factor is reached, e.g., $\kappa_3 = 8$. A high enough oversampling factor a) ensures that the signals maintain their limited bandwidths after constant-envelope modulation and b) does not improve the cost function much from the previous oversampling factor.

Constant-envelope modulation

A discrete-time constant-envelope signal can be modulated into a continuous-time constant-envelope signal by

$$x_m(t) = \tfrac{1}{\sqrt{M}} \exp\left(j \int_{-\infty}^{t} \sum_{\nu=-\ell_{\min}}^{(N-1)\kappa-\ell_{\min}} \arg(x_m^*[\nu-1] x_m[\nu]) p_f(\tau - \nu \tfrac{T}{\kappa}) \mathrm{d}\tau \right),$$

where $-\pi < \arg(z) \leq \pi$ is the principal argument of z, and $p_f(\tau)$ is an L^2-function called the *frequency shaping pulse* that satisfies

$$\int_{-\infty}^{\infty} p_f(\tau) \mathrm{d}\tau = 1. \tag{2.46}$$

For example, to get linear interpolation of the phase, the frequency shaping pulse shall be chosen as

$$p_f(\tau) = \begin{cases} \frac{\kappa}{T}, & -\frac{T}{\kappa} \leq \tau \leq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.47}$$

For other choices of frequency shaping pulses, see e.g. [2]. Let B_{disc} be the δ-bandwidth of the signal obtained from ideal pulse-amplitude modulation (with $\operatorname{sinc}(t\kappa/T)$) of the discrete-time signal $x_m[\nu]$. The modulation scheme in (2.46) ensures a continuous-phase constant-envelope signal, whose δ-bandwidth is approximately B_{disc}, if the oversampling factor $\kappa \gg B_{\text{disc}} T$ is big relative to the bandwidth.

Because the discrete-time signals of the CTCE precoder are highly oversampled however, no choice of the frequency shaping pulse results in significantly lower bandwidth than what is obtained from linear interpolation of the phase. Therefore, linear interpolation is good enough to use.

Achievable rates

The n-th received sample at user k, same as (2.22), can always be written as the sum of three terms:

$$r_k[n] = \sqrt{P} g_k u_k[n] + \sqrt{P} i_k[n] + w_k[n]. \tag{2.48}$$

The first term is the desired signal, scaled by some deterministic factor g_k. The second is an error term, that describes the mismatch between the desired signal and the noise-free received signal. The third is a noise term.

The deterministic factor is chosen to be

$$g_k = \tfrac{1}{\sqrt{P}} \, \mathbb{E}[\, u_k^*[n] r_k[n] \,] \,, \tag{2.49}$$

in order to make the interference and symbol terms uncorrelated, $\mathbb{E}[u_k^*[n] i_k[n]] = 0$. Because the interference is uncorrelated to the symbol, assuming that it is Gaussian distributed is to assume a worst-case scenario. The rate

$$R_k = \log_2\!\left(1 + \frac{P G_k}{P I_k + N_0/T}\right) \quad \text{[bit / channel use]} \tag{2.50}$$

$$G_k = |g_k|^2, \qquad I_k = \mathbb{E}\big[\, |i_k[n]|^2 \,\big] \,. \tag{2.51}$$

is thus achievable with Gaussian distributed symbols. The same bound is derived in [17] for point-to-point MIMO systems. The expectations in (2.49) and in (2.51) are taken with respect to all sources of randomness: channel, symbols and noise.

Both the gain G_k and the interference I_k depend on the parameters γ, λ_1 and λ_2 and on the distribution of the symbols. By maximizing (2.50) with respect to γ, λ_1, λ_2, an achievable sum-rate for CTCE precoding can be established:

$$R_{\text{CTCE}} = \max_{\{(\gamma,\lambda_1,\lambda_2): B \le B_{\max}\}} \sum_{k=1}^{K} R_k(\gamma, \lambda_1, \lambda_2), \tag{2.52}$$

where the optimization is over all choices of the parameters $(\gamma, \lambda_1, \lambda_2)$ that result in a δ-bandwidth less than a given $B_{\max}$.

2.3.4 Numerical analysis of the CTCE precoder

The performance of the CTCE precoder has been evaluated through extensive Monte-Carlo simulations. The channel is assumed to be block-fading and modeled in complex baseband by a tapped delay-line. The channel from antenna m to user k is described by the time-invariant impulse response

$$h_{km}(\tau) = \sum_{d=1}^{D} \sqrt{a(\tau_d)} \alpha_d \delta(\tau - \tau_d), \tag{2.53}$$

where D is the number of propagation paths of the channel, $a(\tau_d)$ a power delay profile, τ_d the delay of path d and $\alpha_d \in \mathbb{C}$ the phase rotation and small-scale fading of path d. The complex attenuations $\{\alpha_d\}$ are modeled as i.i.d. $\mathcal{CN}(0,1)$, and the delays $\{\tau_d\}$ as uniformly distributed between 0 and σ_τ, where σ_τ is the maximum excess delay. The power delay profile is chosen to

Table 2.1: Simulation Setup

Baudrate	$1/T = 5{\times}10^6\,\mathrm{Hz}$
Maximum excess delay	$\sigma_\tau = 3{\times}10^{-6}\,\mathrm{s} \quad (= 15T)$
No. propagation paths	$D = 10$
Receive filter $p(t)$	root-raised cosine, roll-off 0.22
Symbol constellation	i.i.d. Gaussian, $u_k[n] \sim \mathcal{CN}(0,1)$
Oversampling factor	$\kappa = 8$
No. antennas × users	$M \times K = 40 \times 4$
Bandwidth threshold	$\delta = 30\,\mathrm{dB}$

be $a(\tau) = Ae^{-\lambda\tau}$, where the decay rate is chosen such that $a(\sigma_\tau) = 0.1A$, and A is chosen such that $\sum_{d=1}^{D} \mathbb{E}[\,a(\tau_d)\,] = 1$, i.e.:

$$\lambda = \ln(10)/\sigma_\tau, \tag{2.54}$$

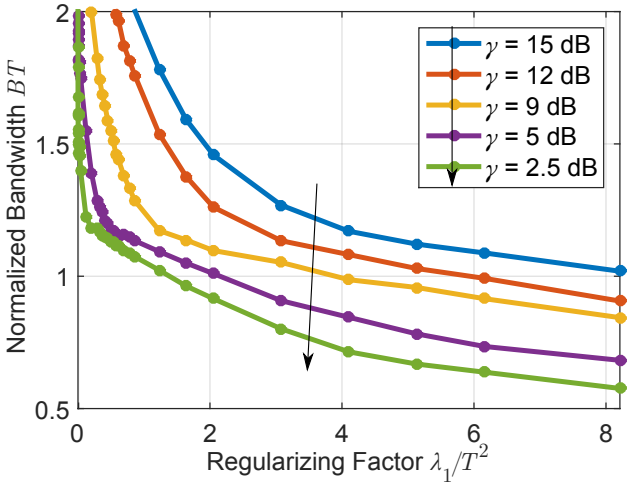
$$A = \frac{\lambda\sigma_\tau}{D(1 - e^{-\lambda\sigma_\tau})}. \tag{2.55}$$

The studied system is specified in Table 2.1. In Figures 2.8a and 2.8b, the estimated 30 dB-bandwidths of the transmit signals of the CTCE precoder are shown for different choices of the parameters γ, λ_1 and λ_2. It can be seen that, for any given γ, the bandwidth of the transmit signals decreases as the factors λ_1 or λ_2 increase. It can also be seen that when γ increases and λ_1, λ_2 are fixed, the precoder has to put more effort to make the mismatch term in (2.30) small, which means that the two regularizing terms, together with the bandwidth, will increase.

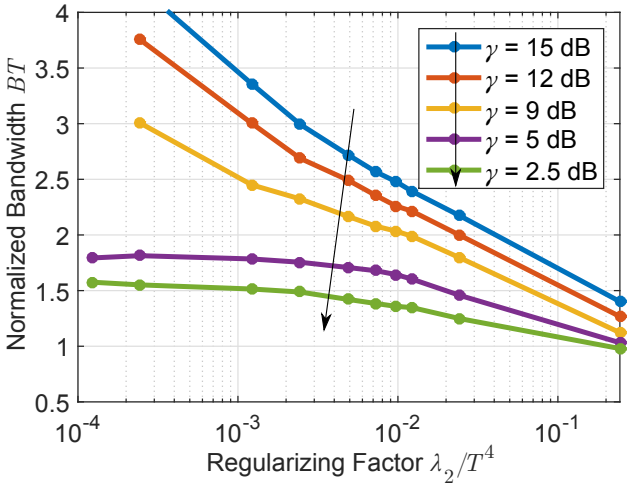
Compare the bandwidths of the CTCE precoder with the bandwidth B_{PAM} of a conventional system using pulse-amplitude modulation, which ideally would be that of the pulse shaping filter that is matched to the receive filter $p(t)$ (that is $TB_{\mathrm{PAM}} = 1.22$ for our choice of receive filter). However, MR and ZF precoding produce transmit signals with high peak-to-average ratio that will be subject to spectral regrowth in the power amplifier. Their actual bandwidth is therefore expected to be greater than that of the pulse shaping filter, $TB_{\mathrm{PAM}} > 1.22$. How much greater depends on the amount of back-off and the complexity of the amplifiers. Note that CTCE can produce signals with bandwidths narrower than B_{PAM}. However, as seen in Figure 2.8a, to make the bandwidth narrow, a big regularizing factor λ_1 has to be used, which rapidly reduces the performance.

The sum-rate of the studied system is shown in Figure 2.9 for different bandwidth requirements $B_{\max}$. The rate was computed by computing G_k and I_k for a mesh of γ, λ_1, λ_2 and maximizing (2.52) over the set of parameter values that resulted in a bandwidth smaller than $B_{\max}$. The proposed precoder is compared to conventional MR and ZF precoding, for which achievable rates were derived by [38]. If the BS were to use the CTCE precoder instead of these conventional precoders to deliver the same sum-rate, then the its radiated power has to be increased. For $TB_{\max} = 1.4$ and for low sum-rates, 2–4 bpcu, this increase is about 3 dB.

It was also noticed in the simulations that the third term in (2.30), which served the purpose of regularizing the second derivative of the transmit signals, did little to improve the performance. The sum-rate curves in Figure 2.9 would still be the same if the regularizing term λ_2 were set to zero in (2.52), with one exception: the sum-rate curve for $TB_{\max} = 2$ improves slightly when the second derivative is regularized. It thus seems, that adding a smoothing term that regularizes the second derivative to the optimization in (2.30) is only important if we allow



(a) λ_1 variable and $\lambda_2 = 0$



(b) λ_2 variable and $\lambda_1 = 0$

Figure 2.8: The 30 dB-bandwidth of CTCE precoded transmit signals for different choices of the regularizing factors λ_1 and λ_2.

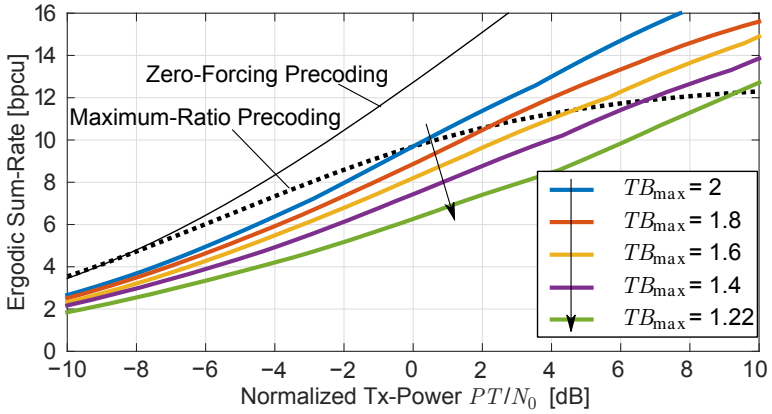


Figure 2.9: The ergodic sum-rate of the simulated system for different precoders. The thick lines represent the proposed CTCE precoder for different bandwidths.

for signals with large excess bandwidths.

2.4 Frequency interpolation of detection and precoding

In this section, we consider ways to reduce the computational complexity of detection and precoding in MaMi-OFDM systems; in particular, for systems that suppress interference using methods such as ZF, where large-dimensional pseudo-inverses need to be computed. Specifically, we ask the following question: How often do we need to compute the ZF pseudo-inverse over frequency? In other words, we investigate on how few subcarriers the ZF matrix has to be computed without incurring a loss in ergodic rate compared to the case where the ZF matrix is computed at all subcarriers. Note that the same ZF matrix can be used for uplink detection and downlink precoding, but for notational convenience we focus on the former case in this section. We propose Discrete Fourier Transform (DFT)-interpolation based low complexity ZF computation and derive a new expression for the achievable uplink ergodic rate with imperfect CSI. We claim and show numerically that by exploiting channel hardening in the MaMi regime it is enough to compute the ZF matrix at L equally spaced subcarriers, with L being the number of resolvable multipaths,[2] and then DFT-interpolate to obtain the detection/precoding matrices at all the N subcarriers.

As a baseline, we compare the ergodic rate obtained using the proposed interpolation method to the ergodic rate obtained based on the full inversion scheme where the ZF matrix is computed at every subcarrier. We also benchmark against the linear interpolation implementation where ZF matrices are computed at L equally spaced subcarriers and then linearly interpolated to get the detector/precoder over all the N subcarriers. We further compare the performance of the proposed DFT-interpolation against the piecewise constant ZF interpolation where as before, L ZF matrices are computed at equally spaced subcarriers and the detector/precoder computed at let's say $(N/L+1)^{\text{th}}$ subcarrier is used to decode/precode transmissions over a cluster of adjacent subcarriers.

2.4.1 System model

We consider the uplink of a single-cell MaMi-OFDM system, where the entire bandwidth is divided into N orthogonal subcarriers. The BS is equipped with an array of M antennas and there are K single-antenna users in the cell. The channel from the k^{th} user to the m^{th} antenna at the BS is denoted by $\tilde{\mathbf{g}}_k^m = \sqrt{\beta_k}\tilde{\mathbf{h}}_k^m = \sqrt{\beta_k}[\tilde{h}_k^m[0]\ \tilde{h}_k^m[1]\ \cdots\ \tilde{h}_k^m[L-1]]^T$, where L is the number of resolvable multipaths, $\tilde{\mathbf{h}}_k^m$ denotes small-scale fading, and β_k is the distance-dependent pathloss of the k^{th} user. For simplicity, we assume that L is known at the BS, but general the channel length and the tap positions will also have to be estimated.

We assume that the path loss from a user is the same to all the antennas at the BS, which is reasonable when the antenna array is much smaller than the distance between users and the BS and there are no dominant scatterers close to the array. Furthermore, we assume Rayleigh fading. Therefore, $\tilde{\mathbf{g}}_k^m \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{\Lambda}_k)$, where $\mathbf{\Lambda}_k$ is a diagonal matrix with the diagonal representing the channel power delay profile of the k^{th} user. We stress that the interpolation method described in this section is not limited to independent Rayleigh fading, but can be applied to any propagation scenario where channel hardening occurs.

[2]Please note that we will denote the number of resolvable multipaths by L and the number of ZF matrix computations or the number of pseudo-inverse computations by L_0.

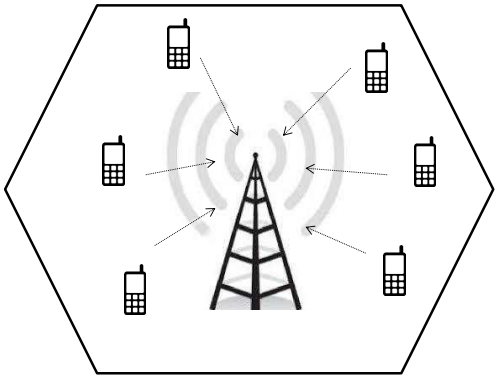


Figure 2.10: System model: K single-antenna users communicating with an M-antenna BS.

Uplink pilot signaling and channel estimation

The frequency-domain signal $\mathbf{y}_m \in \mathbb{C}^{N_p \times 1}$ received at the m^{th} antenna of the BS during uplink pilot signaling is given by

$$\mathbf{y}_m = \sum_{i=1}^{K} \sqrt{p_i} \boldsymbol{\Upsilon}_i^t \boldsymbol{\Omega}_r \tilde{\mathbf{g}}_i^m + \mathbf{w}_m, \tag{2.56}$$

where p_i is the average pilot power per subcarrier with which the i^{th} user transmits during uplink pilot signaling, $\boldsymbol{\Upsilon}_i^t \in \mathbb{C}^{N_p \times N_p}$ is a diagonal matrix with the diagonal comprising of the N_p-length pilot sequence $\mathbf{x}_i^t$ corresponding to user i, $\boldsymbol{\Omega}_r \in \mathbb{C}^{N_p \times L}$ consists of N_p rows of the N-point DFT matrix $\boldsymbol{\Omega}$. These rows correspond to the set of subcarriers on which the N_p pilots are sent out, and $\tilde{\mathbf{g}}_i^m \sim \mathcal{CN}(\mathbf{0}, \beta_i \boldsymbol{\Lambda}_i)$ is the L-tap channel from the i^{th} user to the m^{th} antenna at the BS. The thermal noise vector at the m^{th} antenna of the BS is denoted by $\mathbf{w}_m$. Furthermore, $\mathbf{w}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_p})$. If the pilot sequences are chosen such that[3]

$$\boldsymbol{\Omega}_r^{\text{H}} \boldsymbol{\Upsilon}_k^{t^{\text{H}}} \boldsymbol{\Upsilon}_i^t \boldsymbol{\Omega}_r = N_p I_L \delta_{ki}, \tag{2.57}$$

then a sufficient statistic for estimating $\tilde{\mathbf{g}}_k^m$ is given by

$$\tilde{\mathbf{y}}_m = \frac{1}{\sqrt{N_p}} \boldsymbol{\Omega}_r^{\text{H}} \boldsymbol{\Upsilon}_k^{t^{\text{H}}} \mathbf{y}_m = \sqrt{p_k N_p} \tilde{\mathbf{g}}_k^m + \tilde{\mathbf{w}}_m, \tag{2.58}$$

where $\tilde{\mathbf{w}}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$. Therefore, based on $\tilde{\mathbf{y}}_m$, the minimum mean square error (MMSE) estimate of the time-domain channel $\tilde{\mathbf{g}}_k^m$ from the k^{th} user to the m^{th} antenna at the BS is given by

$$\hat{\tilde{\mathbf{g}}}_k^m = \mathbb{E}\left[\tilde{\mathbf{g}}_k^m \mid \tilde{\mathbf{y}}_m\right] = \sqrt{p_k N_p} \beta_k \boldsymbol{\Lambda}_k \left(p_k N_p \beta_k \boldsymbol{\Lambda}_k + I_L\right)^{-1} \tilde{\mathbf{y}}_m. \tag{2.59}$$

Uplink data transmission

The data signal $\mathbf{y}(s) \in \mathbb{C}^{M \times 1}$ received on the uplink over the s^{th} subcarrier is given by

$$\mathbf{y}(s) = \mathbf{G}(s) \boldsymbol{\Upsilon}_d^{1/2} \mathbf{x}(s) + \mathbf{w}(s), \tag{2.60}$$

where $\mathbf{G}(s) \in \mathbb{C}^{M \times K}$ denotes the frequency-domain channel matrix over the s^{th} subcarrier. Furthermore, $\mathbf{G}(s) = \mathbf{H}(s)\mathbf{D}^{1/2}$, where $\mathbf{H}(s) \in \mathbb{C}^{M \times K}$ denotes small-scale fading over the s^{th}

[3] To ensure orthogonality among pilot sequences of different users, $N_p \geq KL$.

subcarrier and $\mathbf{D}$ is a $K \times K$ diagonal matrix denoting distance-dependent pathloss of the K users, where $[\mathbf{D}]_{k,k} = \beta_k$. Also, the frequency-domain subcarrier gain between the k^{th} user and the m^{th} antenna over subcarrier s is $[\mathbf{G}(s)]_{m,k} = G_k^m(s) = \omega_s^H \tilde{\mathbf{g}}_k^m$, where $\omega_s = \mathbf{\Omega}(s, 1:L)$, $\tilde{\mathbf{g}}_k^m \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{\Lambda}_k)$, and $\mathbf{\Upsilon}_d$ is a $K \times K$ diagonal matrix of average data power per subcarrier of the K users, where $[\mathbf{\Upsilon}_d]_{k,k} = p_k$. The data vector of K users over the s^{th} subcarrier is denoted by $\mathbf{x}(s)$ and the thermal noise vector at the BS over the s^{th} subcarrier is denoted by $\mathbf{w}(s)$. Furthermore, $\mathbf{x}(s) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ and $\mathbf{w}(s) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$.

2.4.2 Uplink Ergodic rate analysis

Let L_0 denote the number of ZF matrix computations or the number of pseudo-inverse computations. In this section, we discuss uplink data detection using ZF and analyze the achievable uplink ergodic rate with imperfect CSI. Specifically, we propose DFT-interpolation based low complexity implementation in which ZF matrices need to be computed only at L_0 equally spaced subcarriers which can then be DFT-interpolated to obtain the detector over all the N subcarriers. To this end, we analyze the achievable uplink ergodic rate for this heuristic scheme in this section and show numerically in Section 2.4.3 that in the MaMi regime due to channel hardening, it is enough to compute the ZF detector at $L_0 = L$ equally spaced subcarriers where L denotes the number of channel taps and then perform DFT-interpolation of the L_0 ZF matrices to get the detector matrices over all the N subcarrier.

As a baseline, we analyze the achievable uplink ergodic rate for the full inversion scheme where the ZF matrix is computed at every subcarrier, i.e., the case when $L_0 = N$. We next analyze the achievable uplink ergodic rate based on the piecewise constant scheme where L_0 ZF matrices are computed at equally spaced subcarriers and the same detector is used to decode transmissions over a cluster of adjacent subcarriers, for example, the detector computed at $(N/L_0 + 1)^{\text{th}}$ subcarrier is used to decode transmissions over $\pm N/(2L_0)$ adjacent subcarriers. Further, we also benchmark our results against the linear interpolation based ZF implementation, where as before, L_0 ZF matrices are computed at equally spaced subcarriers and these are then linearly interpolated to obtain the detector over all the N subcarriers. We let the detector matrix $\hat{\mathbf{A}}(s)$ be an $M \times K$ matrix which depends on the estimated frequency-domain channel matrix and also on whether we employ a DFT interpolation based ZF implementation, a full inversion ZF detector, a piecewise constant ZF detector or a linear interpolation based ZF detector. The received vector on the s^{th} subcarrier post the ZF detector is given by

$$\mathbf{r}(s) = \hat{\mathbf{A}}^{\text{H}}(s)\mathbf{y}(s) = \hat{\mathbf{A}}^{\text{H}}(s)\mathbf{G}(s)\mathbf{\Upsilon}_d^{1/2}\mathbf{x}(s) + \hat{\mathbf{A}}^{\text{H}}(s)\mathbf{w}(s). \tag{2.61}$$

Thus, the k^{th} element of $\mathbf{r}(s)$ is

$$r_k(s) = \sqrt{p_k}\hat{\mathbf{a}}_k^H(s)\mathbf{g}_k(s)x_k(s) + \sum_{i=1, i\neq k}^{K} \sqrt{p_i}\hat{\mathbf{a}}_k^H(s)\mathbf{g}_i(s)x_i(s) + \hat{\mathbf{a}}_k^H(s)\mathbf{w}(s), \tag{2.62}$$

where p_k is the average data power per subcarrier of the k^{th} user, $\hat{\mathbf{a}}_k(s) \in \mathbb{C}^{M\times 1}$ is the column of the detector matrix corresponding to the k^{th} user and is a function of the estimated channel, and $\mathbf{g}_k(s) \in \mathbb{C}^{M\times 1}$ is the frequency-domain channel vector of the k^{th} user over the s^{th} subcarrier.

Note that the MMSE estimate of $\mathbf{g}_k(s)$ is $\hat{\mathbf{g}}_k(s) = \mathbf{g}_k(s) - \mathbf{e}_k(s)$, where $\mathbf{e}_k(s) \in \mathbb{C}^{M\times 1}$ is the estimation error vector over the s^{th} subcarrier that is uncorrelated to $\hat{\mathbf{g}}_k(s)$. Furthermore, the m^{th} entry of $\mathbf{e}_k(s)$ is given by

$$e_k^m(s) = \omega_s^H \tilde{\mathbf{g}}_k^m - \sqrt{p_k N_p}\omega_s^H \mathbf{\Psi}_k \tilde{\mathbf{g}}_k^m - \omega_s^H \mathbf{\Psi}_k \tilde{\mathbf{w}}_m, \tag{2.63}$$

for all $m = 1, \ldots, M$, where $\mathbf{\Psi}_k = \sqrt{p_k^t N_p}\beta_k \mathbf{\Lambda}_k \left(p_k N_p \beta_k \mathbf{\Lambda}_k + \mathbf{I}_L\right)^{-1}$ and $\tilde{\mathbf{w}}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$ and is independent of $\tilde{\mathbf{g}}_k^m$. Thus, we can rewrite (2.62) as

$$r_k(s) = \sqrt{p_k}\hat{\mathbf{a}}_k^H(s)(\hat{\mathbf{g}}_k(s)+\mathbf{e}_k(s))x_k(s)+\sum_{i=1,i\neq k}^{K}\sqrt{p_i}\hat{\mathbf{a}}_k^H(s)(\hat{\mathbf{g}}_i(s)+\mathbf{e}_i(s))x_i(s)+\hat{\mathbf{a}}_k^H(s)\mathbf{w}(s), \quad (2.64)$$

Therefore, the achievable uplink ergodic rate for the k^{th} user over the s^{th} subcarrier with imperfect CSI is given by

$$R_k(s)=\mathbb{E}\left[\log_2\left(1+\frac{\frac{p_k\left|\mathbb{E}\left(\hat{\mathbf{a}}_k^H(s)(\hat{\mathbf{g}}_k(s)+\mathbf{e}_k(s))\middle|\hat{\mathbf{g}}_k(s)\ \forall\ k,s\right)\right|^2}{||\hat{\mathbf{a}}_k^H(s)||^2}}{\sum_{i=1}^{K}\frac{p_i\mathbb{E}\left(\left|\hat{\mathbf{a}}_k^H(s)(\hat{\mathbf{g}}_i(s)+\mathbf{e}_i(s))\right|^2\middle|\hat{\mathbf{g}}_k(s)\ \forall\ k,s\right)}{||\hat{\mathbf{a}}_k^H(s)||^2}-\frac{p_k\left|\mathbb{E}\left(\hat{\mathbf{a}}_k^H(s)(\hat{\mathbf{g}}_k(s)+\mathbf{e}_k(s))\middle|\hat{\mathbf{g}}_k(s)\ \forall\ k,s\right)\right|^2}{||\hat{\mathbf{a}}_k^H(s)||^2}+1}\right)\right] \quad (2.65)$$

The detector matrix $\hat{\mathbf{A}}(s)$ at the s^{th} subcarrier depend on the choice of detection scheme, as discussed below.

Proposed DFT interpolation based ZF detector

Note that the ZF detector over subcarrier s and with imperfect CSI is $\hat{\mathbf{G}}(s)\left(\hat{\mathbf{G}}(s)^H\hat{\mathbf{G}}(s)\right)^{-1}$ where $[\hat{\mathbf{G}}(s)]_{m,k} = \omega_s^H\hat{\tilde{\mathbf{g}}}_k^m$. Also, as defined above, L_0 denotes the number of pseudo-inverse computations or the number of ZF matrix computations. In this heuristic scheme, ZF matrices of dimension $M \times K$ each are computed at L_0 equally spaced subcarriers, i.e., at subcarriers which are N/L_0 apart based on the estimated channel matrix. For each m and k, as illustrated in Figure 2.11, DFT-interpolation basically involves an L_0-point inverse discrete Fourier transform (IDFT) of each element of these equally spaced ZF matrices. This is followed by padding of $N - L_0$ zeros starting at $(L_0 + L)/2$, since the ZF impulse response is symmetric around $L/2$. Thereafter, an element-wise N-point DFT of the ZF impulse response gives the detector over all the N subcarriers. Thus, $(N/L_0 - 1)$ new bins are obtained between each pair of L_0 original bins. This is done for each user-antenna pair to obtain N detector matrices of dimension $M \times K$ each.

In other words, in this scheme, L_0 equally spaced ZF detectors $\hat{\mathbf{G}}(s)(\hat{\mathbf{G}}(s)^H\hat{\mathbf{G}}(s))^{-1}$ of dimension $M\times K$ are computed at $s = 1, N/L_0+1, \ldots, (L-1)N/L_0+1$. For each m and k, an L_0-length vector $\mathbf{u}$ is obtained. Let $\tilde{\mathbf{u}} = \mathbf{\Omega}_{L0}^H\mathbf{u}$ denote the IDFT of $\mathbf{u}$, where $[\mathbf{\Omega}_{L0}]_{j,k} = \frac{1}{L_0}e^{j2\pi(j-1)(k-1)/L_0}$. Let $\tilde{\mathbf{v}} = \text{ZEROPAD}\{\tilde{\mathbf{u}}\}$. The last step involves taking the N-point DFT of $\tilde{\mathbf{v}}$ which gives $\mathbf{v} = \mathbf{\Omega}\tilde{\mathbf{v}}$. This is repeated for each m and k to obtain N detectors of dimension $M \times K$ each. Therefore, for this scheme and with imperfect CSI, the detector matrix $\hat{\mathbf{A}}(s) = \hat{\mathbf{G}}_{\text{DFT-intp}}(s)$, where $\hat{\mathbf{G}}_{\text{DFT-intp}}(s)$ is the DFT-interpolated detector matrix corresponding to the s^{th} subcarrier. Note that for $L_0 \leq L$, perfect reconstruction of the ZF impulse response in step III is not possible due to time-domain aliasing. Using (2.65), the achievable uplink ergodic rate of the k^{th} user over the s^{th} subcarrier with imperfect CSI and with the DFT-interpolation based ZF detector becomes

$$R_k(s) = \mathbb{E}\left[\log_2\left(1+\frac{p_k|\hat{\mathbf{g}}_{k_{\text{DFT-intp}}}^H(s)\hat{\mathbf{g}}_k(s)|^2}{\sum_{i=1,i\neq k}^{K}p_i|\hat{\mathbf{g}}_{k_{\text{DFT-intp}}}^H(s)\hat{\mathbf{g}}_i(s)|^2+||\hat{\mathbf{g}}_{k_{\text{DFT-intp}}}^H(s)||^2\left(1+\sum_{i=1}^{K}p_i^d\phi_i\right)}\right)\right], \quad (2.66)$$

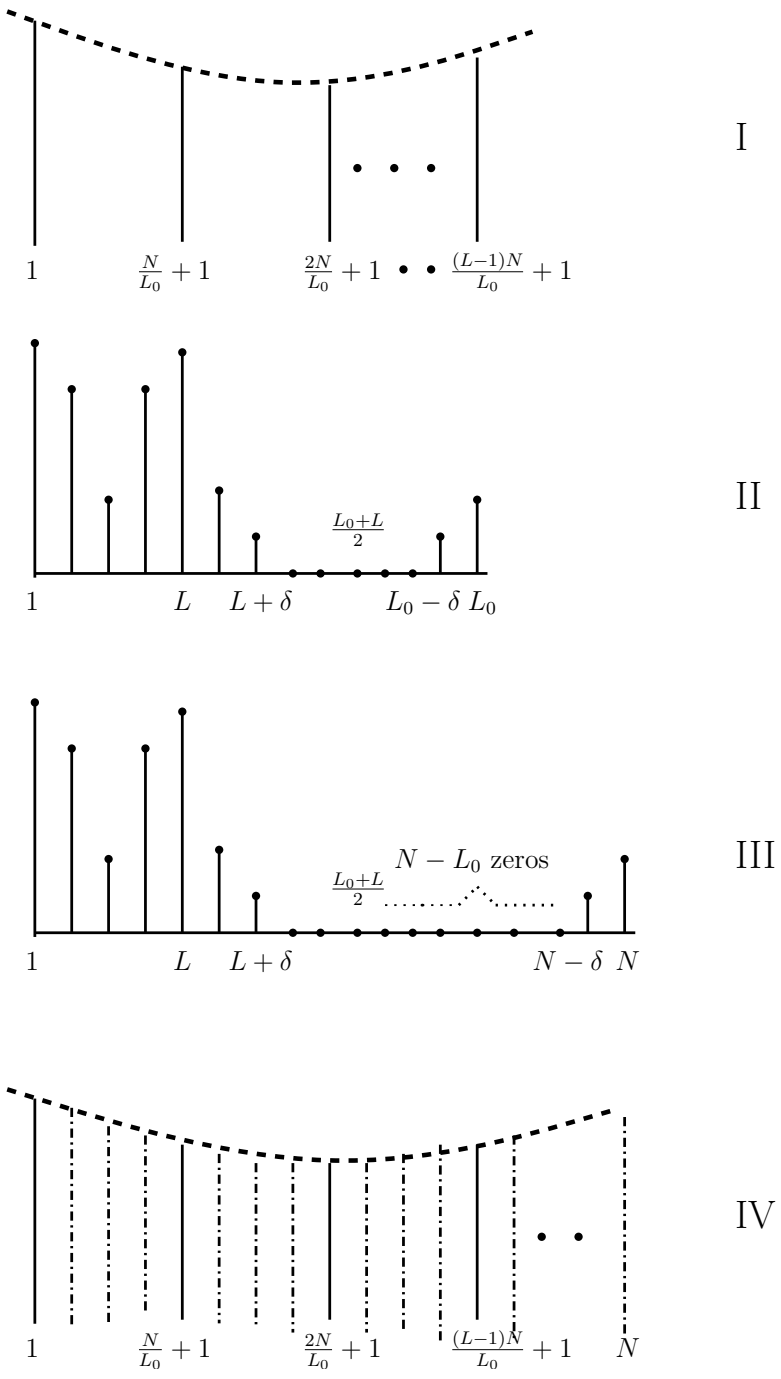


Figure 2.11: DFT-interpolation in four steps: I. Compute L_0 equally spaced ZF matrices $\hat{\mathbf{G}}(s)\left(\hat{\mathbf{G}}(s)^H\hat{\mathbf{G}}(s)\right)^{-1}$ at $s = 1, N/L_0+1, \ldots, (L-1)N/L_0+1$, II. L_0-point IDFT of L_0 equally spaced ZF matrices ($L_0 > L$), III. Pad $N - L_0$ zeros starting at $\frac{L_0+L}{2}$, IV. N-point DFT of the ZF impulse response above.

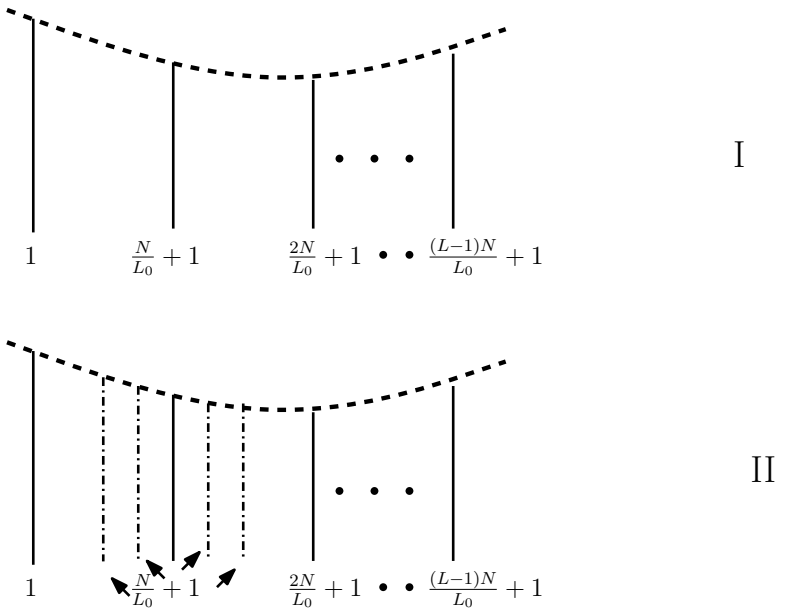


Figure 2.12: Piecewise constant in two steps: I. Compute L_0 equally spaced ZF detectors, II. The ZF detector computed at subcarrier $\left(\frac{N}{L_0}+1\right)$ is used over a cluster of adjacent subcarriers.

where

$$\phi_i = \sum_{l=1}^{L} \frac{\beta_i [\mathbf{\Lambda_i}]_{l,l}}{1 + p_i N_p \beta_i [\mathbf{\Lambda_i}]_{l,l}}. \tag{2.67}$$

Full inversion based ZF detector

For the full inversion scheme, the ZF matrix is computed over each of the N subcarriers based on the estimated channel matrix, i.e., $L_0 = N$. In other words, the detector computed over the s^{th} subcarrier using the estimated channel matrix is used to decode the transmissions over the s^{th} subcarrier. Therefore for this scheme, the detector matrix $\hat{\mathbf{A}}(s) = \hat{\mathbf{G}}(s)(\hat{\mathbf{G}}(s)^H \hat{\mathbf{G}}(s))^{-1}$ and the achievable uplink ergodic rate of the k^{th} user over the s^{th} subcarrier with imperfect CSI and with the full inversion based ZF detector is given by

$$R_k(s) = \mathbb{E}\left[\log_2\left(1 + \frac{p_k}{\left[\left(\hat{\mathbf{G}}(s)^H \hat{\mathbf{G}}(s)\right)^{-1}\right]_{k,k}\left(1 + \sum_{i=1}^{K} p_i^d \phi_i\right)}\right)\right], \tag{2.68}$$

where ϕ_i is given in (2.67).

Piecewise constant ZF detector

For the piecewise constant ZF matrix with imperfect CSI, L_0 ZF matrices are computed at equally spaced subcarriers using the estimated subcarrier gain matrix and the same detector is used to decode transmissions over a cluster of adjacent subcarriers as shown in Figure 2.12. For example, the noisy detector computed over subcarrier $n = N/L_0 + 1$ is used to decode transmissions over some adjacent subcarrier s, where $s = n \pm N/(2L_0)$.

Therefore, for this scheme, the detector matrix to decode transmissions over the s^{th} subcarrier is $\hat{\mathbf{A}}(s) = \hat{\mathbf{G}}_{\text{const}}(n) = \hat{\mathbf{G}}(n)(\hat{\mathbf{G}}(n)^H \hat{\mathbf{G}}(n))^{-1}$ and the achievable uplink ergodic rate of the k^{th}

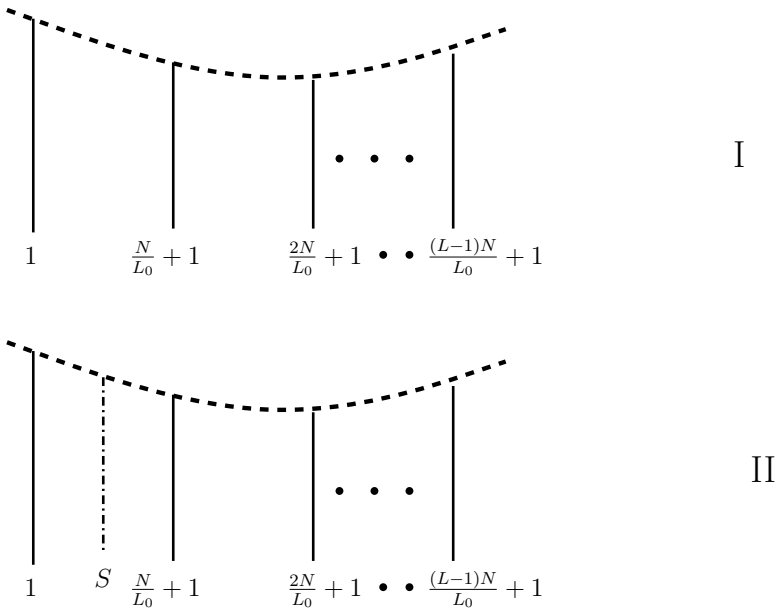


Figure 2.13: Linear interpolation in two steps: I. Compute L_0 equally spaced ZF detectors, II. For any subcarrier $1 \leq s \leq \frac{N}{L_0}+1$, with linear interpolation and imperfect CSI, the ZF detector at subcarrier s is $\hat{\mathbf{A}}(s) = \frac{L_0}{N}\left(\frac{N}{L_0}+1-s\right)\hat{\mathbf{A}}(1) + \frac{L_0(s-1)}{N}\hat{\mathbf{A}}\left(\frac{N}{L_0}+1\right)$.

user over the s^{th} subcarrier with imperfect CSI and for the piecewise constant ZF detector is given by

$$R_k(s) = \mathbb{E}\left[\log_2\left(1 + \frac{p_k|\hat{\mathbf{g}}_{k_{\text{const}}}^H(n)\hat{\mathbf{g}}_k(s)|^2}{\sum\limits_{i=1,i\neq k}^{K} p_i|\hat{\mathbf{g}}_{k_{\text{const}}}^H(n)\hat{\mathbf{g}}_i(s)|^2 + ||\hat{\mathbf{g}}_{k_{\text{const}}}^H(n)||^2\left(1+\sum\limits_{i=1}^{K} p_i^d\phi_i\right)}\right)\right], \quad (2.69)$$

where ϕ_i is given in (2.67).

Linear interpolation based ZF detector

In this scheme, as before, L_0 ZF matrices are computed at equally spaced subcarriers. The linearly interpolated ZF matrix at any subcarrier s such that $1 \leq s \leq \frac{N}{L_0}+1$ is given by $\hat{\mathbf{A}}(s) = \hat{\mathbf{G}}_{\text{lin-intp}}(s) = \frac{L_0}{N}\left(\frac{N}{L_0}+1-s\right)\hat{\mathbf{A}}(1) + \frac{L_0(s-1)}{N}\hat{\mathbf{A}}\left(\frac{N}{L_0}+1\right)$, where $\hat{\mathbf{A}}(1) = \hat{\mathbf{G}}(1)(\hat{\mathbf{G}}(1)^H\hat{\mathbf{G}}(1))^{-1}$ and $\hat{\mathbf{A}}(N/L_0+1) = \hat{\mathbf{G}}(N/L_0+1)(\hat{\mathbf{G}}(N/L_0+1)^H\hat{\mathbf{G}}(N/L_0+1))^{-1}$.

The achievable uplink ergodic rate of the k^{th} user over the s^{th} subcarrier with imperfect CSI and with the linear interpolation based ZF detector is given by

$$R_k(s) = \mathbb{E}\left[\log_2\left(1 + \frac{p_k|\hat{\mathbf{g}}_{k_{\text{lin-intp}}}^H(s)\hat{\mathbf{g}}_k(s)|^2}{\sum\limits_{i=1,i\neq k}^{K} p_i|\hat{\mathbf{g}}_{k_{\text{lin-intp}}}^H(s)\hat{\mathbf{g}}_i(s)|^2 + ||\hat{\mathbf{g}}_{k_{\text{lin-intp}}}^H(s)||^2\left(1+\sum\limits_{i=1}^{K} p_i^d\phi_i\right)}\right)\right], \quad (2.70)$$

where ϕ_i is given by (2.67).

The computational complexity of different ZF matrix computations discussed above is given in Table 2.2. There are clearly multiple ways to reduce the number of pseudo-inverses that

Table 2.2: Computational Complexity of Different ZF Detectors

Scheme	No. of pseudo-inverse computations	No. of computations in interpolation
Full inversion	N	0
DFT-interpolation	L	$\mathcal{O}(N \log N)$
Piecewise constant	L	0
Linear interpolation	L	$N - L$ complex multiplications and $2(N - L)$ complex additions

are computed, each attached with a certain additional interpolation complexity. Next, we will compare the performance of theses interpolation schemes.

2.4.3 Numerical results

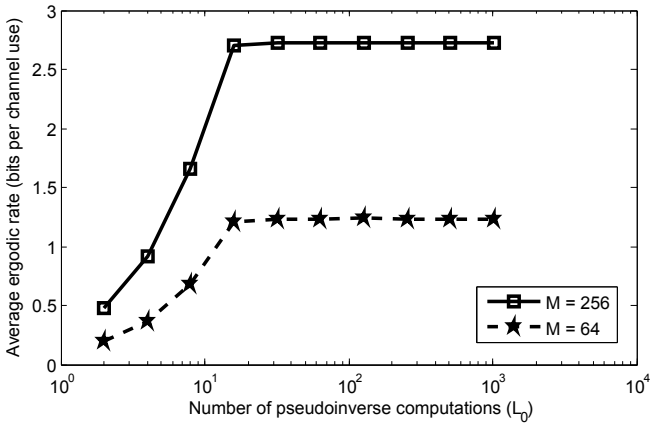
In this section, we present numerical results to investigate on how few subcarriers the ZF detector needs to be computed without incurring a rate loss compared to the full inversion scheme. For simplicity we let the SNR $\rho = p_k\beta_k$ be the same for all users, which for instance can be achieved by uplink power control. We consider a frequency-selective channel with uniform power delay profile, i.e., we take $\mathbf{\Lambda}_k = \frac{1}{L}\mathbf{I}_L$ for all $k = 1, \ldots, K$. Also, we take the number of pilot subcarriers $N_p = KL$.

Figure 2.14a plots the average ergodic rate (sum rate divided by the total number of subcarriers) for the DFT-interpolation based ZF detector for $K = 4$ and $L = 16$ as a function of the number of pseudo-inverse computations (or the number of ZF detector computations) L_0 for imperfect CSI and for two different values of M. It can be observed that it is enough to compute $L_0 = L$ equally spaced pseudo-inverses or ZF matrices and then DFT-interpolate without incurring any visible loss in the average ergodic rate compared to the case when $L_0 = N$ which corresponds to full inversion. Figure 2.14b plots the same for a more frequency-selective channel, $L = 64$ and similar conclusions are obtained. This is because if K is small relative to M, then applying classical random matrix results, we can conclude that in the MaMi regime, the empirical distribution of the singular values of the pseudo-inverse or the ZF detector converges to the same deterministic limiting distribution across all subcarriers.

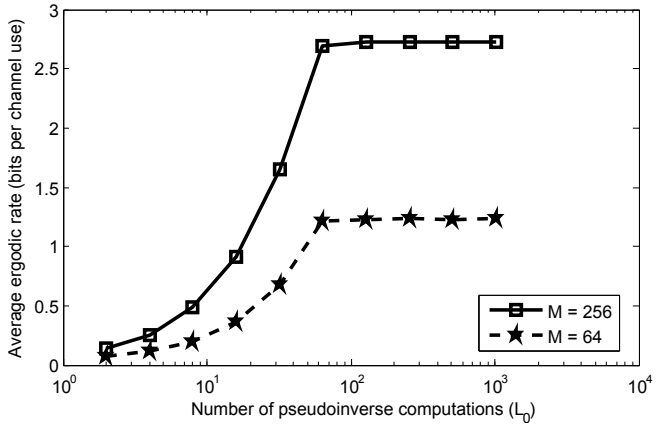
Figure 2.15a plots the average ergodic rate for relatively larger number of users, i.e., $K = 16$ and a less frequency-selective channel $L = 16$ as a function of L_0 for two different values of M. It can be observed that even with larger K, there is a marginal loss in the average ergodic rate of about 9.8% for $M = 64$ and 6.6% for $M = 256$ when $L_0 = L$ compared to when $L_0 = N$. Figure 2.15b plots the same for a relatively more frequency-selective channel with $L = 64$. Similar conclusions are obtained from this case, thus illustrating the generality of the results.

Figure 2.16 plots the loss in average ergodic rate as a function of the number of antennas M at the BS. We define the loss in average ergodic rate as the ratio of the difference between the average ergodic rate when $L_0 = N$ and the average ergodic rate when $L_0 = L$ to the average ergodic rate when $L_0 = N$. It can be observed that the loss in the average ergodic rate reduces as M increases due to channel hardening. The loss in rate, however, is higher for larger K.

Figure 2.17 plots the ergodic rate as a function of the subcarrier index with imperfect CSI for $L_0 = L = 16$. We benchmark the ergodic rates obtained using DFT-interpolation based ZF detector against the full inversion, the piecewise constant, and the linear interpolation based ZF detectors. We observe that for the computationally less expensive DFT-interpolation based ZF detector, the loss in ergodic rate is marginal when compared to the full inversion based ZF

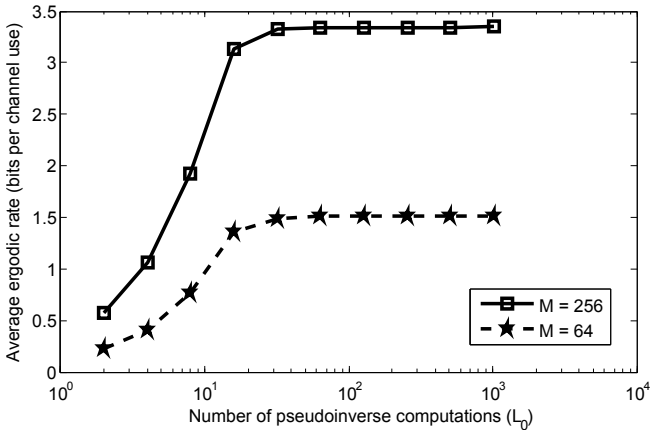


(a) $L = 16$

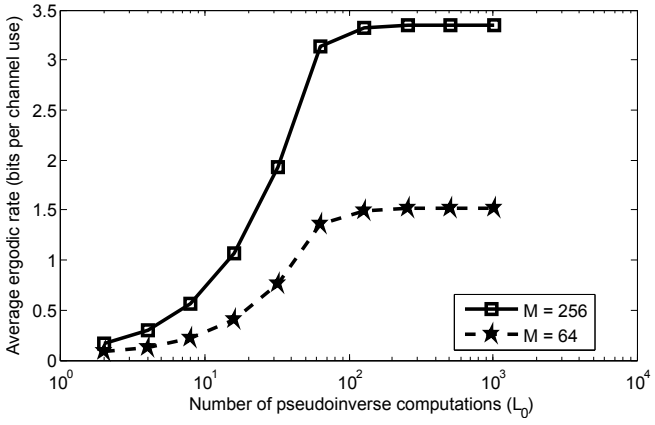


(b) $L = 64$

Figure 2.14: DFT-interpolation: Average ergodic rate vs. L_0 ($K = 4$, $N = 1024$, $\rho = -10$ dB)



(a) $L = 16$



(b) $L = 64$

Figure 2.15: DFT-interpolation: Average ergodic rate vs. L_0 ($K = 16$, $N = 1024$, $\rho = -10$ dB)

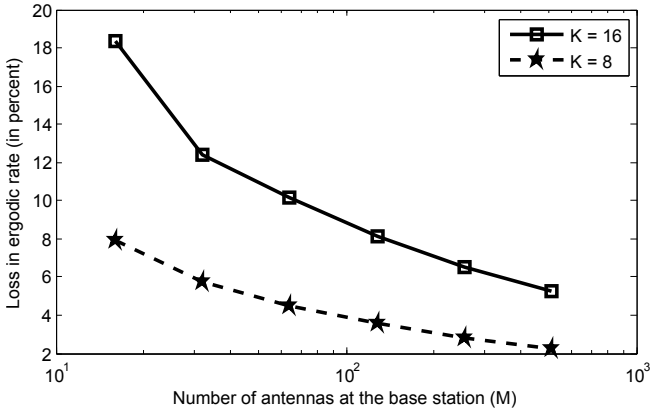


Figure 2.16: DFT-interpolation: Loss in average ergodic rate vs. M ($L = 64$, $N = 1024$, $\rho = -10$ dB). Loss in average ergodic rate is the ratio of the difference between the average ergodic rate when $L_0 = N$ and the average ergodic rate when $L_0 = L$ to the average ergodic rate when $L_0 = N$.

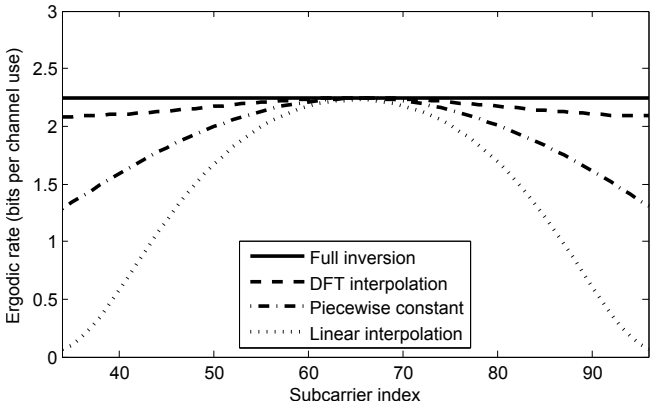


Figure 2.17: Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L_0 = L = 16$, $N = 1024$, $\rho = -10$ dB)

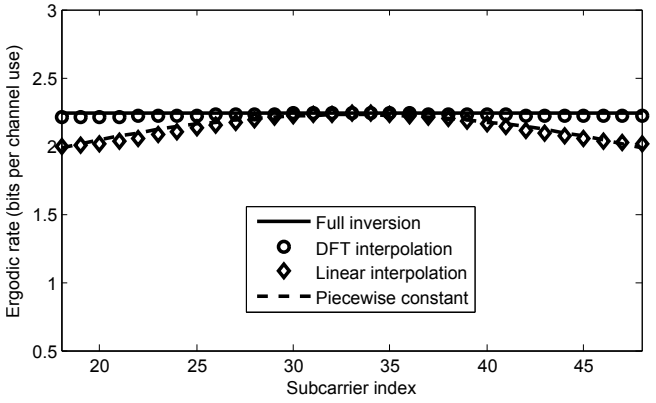


Figure 2.18: Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L = 16$, $L_0 = 32$, $N = 1024$, $\rho = -10$ dB)

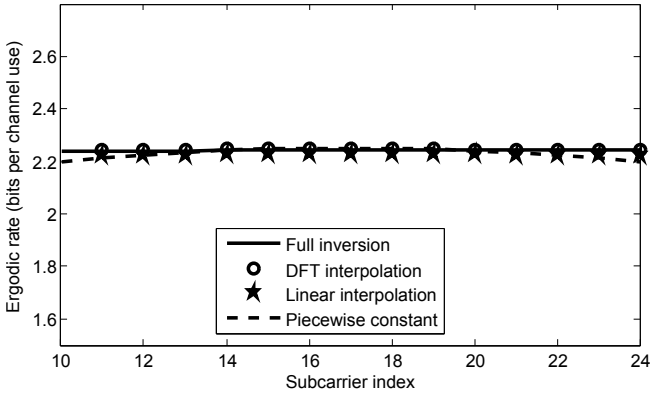


Figure 2.19: Imperfect CSI: Ergodic rate vs. subcarrier index ($M = 128$, $K = 8$, $L = 16$, $L_0 = 64$, $N = 1024$, $\rho = -10$ dB)

detector. It also gives substantially better performance compared to the piecewise constant and the linear interpolation based ZF detector. Note that the linear interpolation based detector performs poorly when $L_0 = L$.

Figure 2.18 plots the same for the case when $L_0 = 32 > L$. In this case, DFT-interpolation based detector performs as well as full inversion. Also, the linear interpolation based detector works as well as the piecewise constant detector. However, both of these give inferior performance when compared to DFT-interpolation. Figure 2.19 plots the same for the case when $L_0 = 64 > L$. For this scenario, linear interpolation gives a marginally better ergodic rate performance when compared to piecewise constant detector.

2.4.4 Summary of interpolation methods

We investigated on how few subcarriers do we need to compute the ZF matrix or the pseudo-inverse in a MaMi-OFDM system without incurring a visible rate loss compared to the full inversion scheme. We showed numerically that by exploiting channel hardening it is enough to compute the pseudo-inverse or the ZF matrix at L equally spaced subcarriers and then DFT-interpolate to get the detector/precoder at all the N subcarriers. This is explained by classical random matrix results where in the large antenna regime, the empirical distribution of the singular values of the pseudo-inverse or the ZF matrix converges to the same deterministic limiting distribution across all subcarriers. We compared the proposed DFT-interpolation based ZF implementation to full inversion, piecewise constant and linear interpolation and showed that it achieves a splendid trade-off between computational complexity and the ergodic rate performance.

2.5 Hardware imperfection assessment

A potential showstopper for MaMi would be that the technology is too sensitive to transceiver hardware impairments; for example, phase noise in Local Oscillator (LO), amplifier non-linearities, non-ideal analog filters, and finite-precision analog/digital converters. The impact of hardware impairments on MaMi has therefore received considerable attention in recent years [3, 5, 16, 21, 23, 32, 33]. The paper [3] showed that it is of fundamental importance to include hardware impairments in the performance analysis, since this can be a main limiting factor in systems with many antennas. Nevertheless, [3, 5] showed that MaMi is resilient to additive distortions originating from the BS. Multiplicative distortions such as phase noise can, however, reduce the system performance. These works use analytically tractable stochastic impairment models, but the validity of the results has been confirmed in [16] by simulations based on sophisticated and realistic models.

For physically large and distributed antenna array, an important question is whether the antennas should share a common LO (CLO) or if each antenna should be equipped with a separate LO (SLO). In the CLO case the clock-drift is the basically same for all antennas, while in the SLO case the system would try to lock all LOs to a common clock but the drifts will be independent. A number of recent works have looked into how this design choice impacts the severeness of the phase noise [5, 21, 23, 32, 33]. The papers [5, 23, 32, 33] establish the consensus that a setup with SLOs is preferable in the uplink (UL), since the independent phase rotations average out over the BS antennas. However, the answer is still open when it comes to the DL; [21] showed that a CLO is preferable for non-fading channels, while [23] considered fading single-cell systems and claimed that CLO prevails for few BS antennas (per user) or high SNR, and SLOs are desirable in the opposite cases.

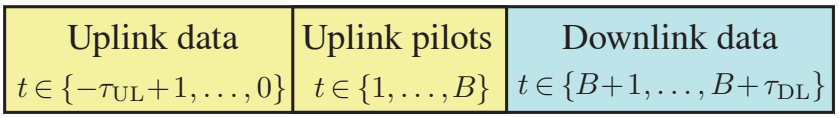


Figure 2.20: Illustration of the TDD protocol where each coherence block consists of $T = \tau_{\mathrm{UL}} + \tau_{\mathrm{DL}} + B$ symbols.

In this section, the previous UL results from the MAMMOET publication [5] are extended to the DL. We consider a multi-cell MaMi system with three kinds of hardware impairments: phase noise, additive distortion noise, and noise amplification. We derive new spectral efficiency expressions for MR precoding, which establish a performance baseline in hardware-impaired multi-cell scenarios. These expressions are used to prove how the hardware quality may scale with the number of antennas. The analysis shows that SLOs is systematically a better choice than CLO also in the DL.

2.5.1 System model

We consider a cellular network with L cells that operate in a synchronized TDD mode. Each cell serves K single-antenna UEs using a BS equipped with M antennas. The TDD protocol divides the time-frequency resources into coherence blocks, as illustrated in Figure 2.20. Each block consists of T symbols with time indices $t = -\tau_{\mathrm{UL}} + 1, \ldots, B + \tau_{\mathrm{DL}}$, whereof τ_{UL} are UL data symbols, B are UL pilots, and τ_{DL} are DL data symbols. Note that $T = \tau_{\mathrm{UL}} + \tau_{\mathrm{DL}} + B$.

Let $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote the transpose and conjugate transpose, respectively. The channel response between UE k in cell l and BS j is a constant vector $\mathbf{h}_{jlk} = [h_{jlk}^{(1)} \ldots h_{jlk}^{(M)}]^{\mathrm{T}} \in \mathbb{C}^M$ within each block, where $h_{jlk}^{(n)}$ is the channel response for the nth BS antenna. The channels are assumed to be Rayleigh fading as

$$\mathbf{h}_{jlk} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda}_{jlk}), \tag{2.71}$$

where the covariance matrix is $\mathbf{\Lambda}_{jlk} = \mathrm{diag}(\lambda_{jlk}^{(1)}, \ldots, \lambda_{jlk}^{(M)})$. The average channel attenuation $\lambda_{jlk}^{(n)} \geq 0$ is different for each combination of cell indices, UE index, and BS antenna index n. It depends, for example, on how the BS antennas are distributed in the cell and on the UE positions. This model supports both co-located and distributed arrays.

Uplink model with hardware impairments

A main goal of this section is to investigate how transceiver hardware impairments impact the DL spectral efficiency. We mainly consider impairments at the BSs, since MaMi systems can operate with reduced BS hardware precision by capitalizing on the averaging effect that occurs when processing the signals over the array [3]. Reduced BS hardware precision can lead to lower hardware cost, higher energy efficiency, reduced BS size, and relaxed requirements on synchronization.

To this end, we adopt the UL system model from [5] and generalize it to also cover the DL. Since the BSs in MaMi use channel estimates from the UL to perform transmit precoding in the DL, we need to model both directions of the links. As in [5], the received UL signal $\mathbf{y}_j(t) \in \mathbb{C}^M$ in cell j at symbol time $t \in \{-\tau_{\mathrm{UL}} + 1, \ldots, B\}$ is modeled as

$$\mathbf{y}_j(t) = \mathbf{D}_{\boldsymbol{\phi}_j(t)} \sum_{l=1}^{L} \mathbf{H}_{jl}\mathbf{x}_l(t) + \boldsymbol{\upsilon}_j(t) + \boldsymbol{\eta}_j(t) \tag{2.72}$$

where $\mathbf{x}_l(t) = [x_{l1}(t) \ldots x_{lK}(t)]^{\mathrm{T}} \in \mathbb{C}^K$ contains pilot/data symbols from UEs in cell l and the channel matrix from these UEs to BS j is $\mathbf{H}_{jl} = [\mathbf{h}_{jl1} \ldots \mathbf{h}_{jlK}] \in \mathbb{C}^{M\times K}$. The symbols from UE k in cell j have power $p_{jk}^{\mathrm{UL}} = \mathbb{E}\{|x_{jk}(t)|^2\}$, where $\mathbb{E}\{\cdot\}$ denotes the expected value of a random variable.

The matrix $\mathbf{D}_{\boldsymbol{\phi}_j(t)} = \mathrm{diag}\left(e^{\imath\phi_{j1}(t)}, \ldots, e^{\imath\phi_{jM}(t)}\right)$ models the multiplicative effect of phase noise (with $\imath = \sqrt{-1}$). The variable $\phi_{jn}(t)$ is the phase rotation at the nth BS antenna in cell j at time t, and it is modeled as a Wiener process [31]: $\phi_{jn}(t) \sim \mathcal{M}(\phi_{jn}(t-1), \delta)$ where $\delta \geq 0$ is the variance of the phase-noise increments. We consider two implementations:

1. Common LO (CLO): $\phi_{j1}(t) = \ldots = \phi_{jM}(t)$ within a cell.

2. Separate LOs (SLOs): All $\phi_{jn}(t)$ are independent.

The above represent having one LO that feeds all antennas at BS j or one separate LO connected to each of the M antennas.

Moreover, $\boldsymbol{\upsilon}_j(t) \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Upsilon}_j(t))$ is additive distortion noise (e.g., from finite-precision quantization, non-linearities, and interference leakage in the frequency domain). It is proportional to the received signal power at the antenna and uncorrelated between antennas [5, 39]:

$$\boldsymbol{\Upsilon}_j(t) = \kappa_{\mathrm{UL}}^2 \sum_{l=1}^{L} \sum_{k=1}^{K} p_{lk}^{\mathrm{UL}} \mathrm{diag}\left(|h_{jlk}^{(1)}|^2, \ldots, |h_{jlk}^{(M)}|^2\right) \tag{2.73}$$

where $\kappa_{\mathrm{UL}} \geq 0$ is the proportionality coefficient.

Finally, $\boldsymbol{\eta}_j(t) \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathrm{BS}}^2 \mathbf{I}_M)$ is the receiver noise with variance σ_{BS}^2 (including noise amplification in circuits).

Downlink model with hardware impairments

Similar to the UL, we model the received DL signal $z_{jk}(t) \in \mathbb{C}$ at UE k in cell j at time $t \in \{B+1, \ldots, B+\tau_{\mathrm{DL}}\}$ as

$$z_{jk}(t) = \sum_{l=1}^{L} \mathbf{h}_{ljk}^{\mathrm{H}} \left(\mathbf{D}_{\boldsymbol{\phi}_l(t)} \sum_{m=1}^{K} \mathbf{w}_{lm}(t) s_{lm}(t) + \boldsymbol{\psi}_l(t) \right) + \eta_{jk}(t) \tag{2.74}$$

where $s_{lm}(t)$ is a DL data symbol with power $p_{jk}^{\mathrm{DL}} = \mathbb{E}\{|s_{lm}(t)|^2\}$ and $\mathbf{w}_{lm}(t) = [w_{lm}^{(1)}(t) \ldots w_{lm}^{(M)}(t)]^{\mathrm{T}} \in \mathbb{C}^M$ is the corresponding linear precoding vector. The receiver noise is $\eta_{jk}(t) \sim \mathcal{CN}(0, \sigma_{\mathrm{UE}}^2)$, where σ_{UE}^2 is the variance (including noise amplification). The phase-noise matrix $\mathbf{D}_{\boldsymbol{\phi}_j(t)}$ was defined earlier, while $\boldsymbol{\psi}_j(t) \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Psi}_j)$ is the additive distortion in the DL (e.g., due to non-linearities and leakage in the frequency domain). Similar to (2.73), the distortion at a certain antenna is proportional to the transmit power at this antenna and uncorrelated with the distortions at other antennas:

$$\boldsymbol{\Psi}_j = \kappa_{\mathrm{DL}}^2 \sum_{k=1}^{K} p_{jk}^{\mathrm{DL}} \mathrm{diag}\left(|w_{jk}^{(1)}(t)|^2, \ldots, |w_{jk}^{(M)}(t)|^2\right)$$

where $\kappa_{\mathrm{DL}} \geq 0$ is the proportionality coefficient.

This analytically tractable system model is used in the next section to compute achievable DL spectral efficiencies. These depend on the level of hardware impairments, as characterized by the variance of the phase-noise increments δ, the distortion noise proportionality coefficients $\kappa_{\mathrm{UL}}, \kappa_{\mathrm{DL}}$, and the receiver noise variances $\sigma_{\mathrm{BS}}^2, \sigma_{\mathrm{UE}}^2$. The results are applicable for any p_{jk}^{DL} and p_{jk}^{UL}, for each j and k, thus under arbitrary power control.

2.5.2 Downlink performance analysis with hardware impairments

In this section, we derive the DL spectral efficiency per UE and study its asymptotic behavior (when M is large) to understand the impact of hardware impairments.

Uplink channel estimation

In order to perform coherent transmit precoding in the DL, each BS acquires the channels to its UEs by using the UL pilots. The pilot sequence of UE k in cell j is defined as $\tilde{\mathbf{x}}_{jk} = [x_{jk}(1) \ldots x_{jk}(B)]^{\mathrm{T}} \in \mathbb{C}^{B\times 1}$. The analysis in this section holds for arbitrary pilot sequences (with $|x_{jk}(b)|^2 = p_{jk}^{\mathrm{UL}}$ for $b = 1, \ldots, B$), while we consider columns from a Fourier matrix in Sec. 2.5.3 (to achieve mutual orthogonality and constant energy per symbol). Since the effective channels

$$\mathbf{h}_{jlk}(t) = \mathbf{D}_{\boldsymbol{\phi}_j(t)}\mathbf{h}_{jlk} \tag{2.75}$$

depend on the phase-noise and are different at every symbol time t, we need a channel estimator that provides new estimates at each t. Such a linear MMSE estimator can be derived as follows:

Let $\boldsymbol{\psi}_j = \left[\mathbf{y}_j^{\mathrm{T}}(1) \ldots \mathbf{y}_j^{\mathrm{T}}(B)\right]^{\mathrm{T}} \in \mathbb{C}^{BM}$ denote the combined received signal in cell j from the pilot transmission. The linear MMSE estimate of $\mathbf{h}_{jlk}(t)$ at any symbol time $t \in \{-\tau_{\mathrm{UL}} + 1, \ldots, B + \tau_{\mathrm{DL}}\}$ for any l and k is

$$\hat{\mathbf{h}}_{jlk}(t) = \left(\tilde{\mathbf{x}}_{lk}^{\mathrm{H}}\mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{jlk}\right)\boldsymbol{\Phi}_j^{-1}\boldsymbol{\psi}_j \tag{2.76}$$

where $\otimes$ denotes the Kronecker product,

$$\mathbf{D}_{\boldsymbol{\delta}(t)} = \mathrm{diag}\left(e^{-\frac{\delta}{2}|t-1|}, \ldots, e^{-\frac{\delta}{2}|t-B|}\right), \tag{2.77}$$

$$\boldsymbol{\Phi}_j = \sum_{\ell=1}^{L}\sum_{m=1}^{K}\mathbf{X}_{\ell m} \otimes \boldsymbol{\Lambda}_{j\ell m} + \sigma_{\mathrm{BS}}^2\mathbf{I}_{BM}, \tag{2.78}$$

and the element at position (b_1, b_2) in $\mathbf{X}_{\ell m} \in \mathbb{C}^{B\times B}$ is

$$[\mathbf{X}_{\ell m}]_{b_1,b_2} = \begin{cases} p_{\ell m}^{\mathrm{UL}}(1+\kappa_{\mathrm{UL}}^2), & b_1 = b_2, \\ x_{\ell m}(\tau_{b_1})x_{\ell m}^*(\tau_{b_2})e^{-\frac{\delta}{2}|\tau_{b_1}-\tau_{b_2}|}, & b_1 \neq b_2. \end{cases} \tag{2.79}$$

The corresponding error covariance matrix is

$$\mathbf{C}_{jlk}(t) = \boldsymbol{\Lambda}_{jlk} - (\tilde{\mathbf{x}}_{lk}^{\mathrm{H}}\mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{jlk})\boldsymbol{\Phi}_j^{-1}(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}}\tilde{\mathbf{x}}_{lk} \otimes \boldsymbol{\Lambda}_{jlk}).$$

Downlink spectral efficiency

Next, we derive achievable DL spectral efficiencies, using normalized linear precoding vectors of the general form

$$\mathbf{w}_{jk}(t) = \frac{\boldsymbol{\omega}_{jk}(t)}{\sqrt{\mathbb{E}\{\|\boldsymbol{\omega}_{jk}(t)\|^2\}}}. \tag{2.80}$$

With this notation, MR precoding is given by $\boldsymbol{\omega}_{jk}(t) = \hat{\mathbf{h}}_{jjk}(t)$.

Suppose that UE k in cell j knows the channel and interference statistics, but not the channel realizations. An achievable lower bound on the ergodic capacity of this UE is

$$R_{jk} = \frac{1}{T}\sum_{t=B+1}^{B+\tau_{\mathrm{DL}}}\log_2\left(1 + \mathrm{SINR}_{jk}(t)\right) \quad [\text{bit/symbol}] \tag{2.81}$$

where the signal-to-interference-and-noise ratio (SINR) is

$$\mathrm{SINR}_{jk}(t) = \frac{p_{jk}^{\mathrm{DL}} \frac{|\mathbb{E}\{\mathbf{h}_{jjk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{jk}(t)\}|^2}{\mathbb{E}\{\|\boldsymbol{\omega}_{jk}(t)\|^2\}}}{\sum_{l=1}^{L}\sum_{m=1}^{K} p_{lm}^{\mathrm{DL}} \left(\frac{\mathbb{E}\{|\mathbf{h}_{ljk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{lm}(t)|^2\} + \kappa_{\mathrm{DL}}^2 \sum_{n=1}^{M} \mathbb{E}\{|\mathbf{h}_{ljk}^{(n)}|^2 |\boldsymbol{\omega}_{lm}^{(n)}(t)|^2\}}{\mathbb{E}\{\|\boldsymbol{\omega}_{lm}(t)\|^2\}} \right) - p_{jk}^{\mathrm{DL}} \frac{|\mathbb{E}\{\mathbf{h}_{jjk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{jk}(t)\}|^2}{\mathbb{E}\{\|\boldsymbol{\omega}_{jk}(t)\|^2\}} + \sigma_{\mathrm{UE}}^2} \tag{13}$$

This expression is obtained by using the signal received over the average channel $\mathbb{E}\{\mathbf{h}_{jjk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{jk}(t)\}$ for decoding, while treating the signal received over the uncorrelated deviation from this average value, the inter-user interference and distortion noise as worst-case Gaussian noise in the decoder. The expression in (2.81) is a reasonable bound on the practical performance that can be achieved using simple signal processing at the UE (i.e., detect the useful signal and treat everything unknown as Gaussian noise). The SINR expression in (13) contains a number of expectations that can be computed numerically for any choice of precoding vectors. Next, we provide closed-form expressions for MR precoding.

If MR is used, then the expectations in $\mathrm{SINR}_{jk}(t)$ are computed as follows (where $\mathbf{e}_n$ denotes the nth column of $\mathbf{I}_M$):

$$\mathbb{E}\{\|\boldsymbol{\omega}_{jk}(t)\|^2\} = \mathbb{E}\{\mathbf{h}_{jjk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{jk}(t)\} = \mathrm{tr}\left(\left(\tilde{\mathbf{x}}_{jk}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{jjk} \right) \boldsymbol{\Phi}_j^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{jk} \otimes \boldsymbol{\Lambda}_{jjk} \right) \right) \tag{2.82}$$

$$\begin{aligned} \mathbb{E}\{|\mathbf{h}_{ljk}^{\mathrm{H}}(t)\boldsymbol{\omega}_{lm}(t)|^2\} &= \kappa_{\mathrm{DL}}^2 \sum_{n=1}^{M} \mathbb{E}\{|\mathbf{h}_{ljk}^{(n)}|^2 |\boldsymbol{\omega}_{lm}^{(n)}(t)|^2\} \\ &\quad + (1+\kappa_{\mathrm{DL}}^2)\mathrm{tr}\left(\boldsymbol{\Lambda}_{ljk} \left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{llm} \right) \boldsymbol{\Phi}_l^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{lm} \otimes \boldsymbol{\Lambda}_{llm} \right) \right) \end{aligned} \tag{2.83}$$

$$+ \begin{cases} \sum_{n_1=1}^{M}\sum_{n_2=1}^{M} \lambda_{llm}^{(n_1)} \lambda_{ljk}^{(n_1)} \lambda_{llm}^{(n_2)} \lambda_{ljk}^{(n_2)} \left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \mathbf{e}_{n_1}^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left((\mathbf{X}_{jk} - \kappa_{\mathrm{UL}}^2 p_{jk}^{\mathrm{UL}} \mathbf{I}_B) \otimes \mathbf{e}_{n_1} \mathbf{e}_{n_2}^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{lm} \otimes \mathbf{e}_{n_2} \right) & \text{if CLO} \\ \left(\mathrm{tr}\left(\left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{llm} \right) \boldsymbol{\Phi}_l^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{jk} \otimes \boldsymbol{\Lambda}_{ljk} \right) \right) \right)^2 & \text{if SLOs} \end{cases}$$

$$+ \begin{cases} \sum_{n=1}^{M} \left(\lambda_{llm}^{(n)} \lambda_{ljk}^{(n)} \right)^2 \left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \mathbf{e}_n^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left((\kappa_{\mathrm{UL}}^2 p_{jk}^{\mathrm{UL}} \mathbf{I}_B + \kappa_{\mathrm{DL}}^2 \mathbf{X}_{jk}) \otimes \mathbf{e}_n \mathbf{e}_n^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{lm} \otimes \mathbf{e}_n \right) & \text{if CLO} \\ \sum_{n=1}^{M} \left(\lambda_{llm}^{(n)} \lambda_{ljk}^{(n)} \right)^2 \left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \mathbf{e}_n^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left(((1+\kappa_{\mathrm{DL}}^2) \mathbf{X}_{jk} - \mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{jk} \tilde{\mathbf{x}}_{jk}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)}) \otimes \mathbf{e}_n \mathbf{e}_n^{\mathrm{H}} \right) \boldsymbol{\Phi}_l^{-1} \left(\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{lm} \otimes \mathbf{e}_n \right) & \text{if SLOs} \end{cases}$$

Asymptotic behavior and scaling laws

Next, we investigate the behavior at large M. For tractability, we consider $A < \infty$ spatially separated subarrays each with $\frac{M}{A}$ antennas. Recall that these antennas are either controled by a common LO that sends clock signals or separate LOs at each antenna. The channel covariance matrices then factorize as

$$\boldsymbol{\Lambda}_{jlk} = \tilde{\boldsymbol{\Lambda}}_{jlk}^{(A)} \otimes \mathbf{I}_{\frac{M}{A}} \tag{2.84}$$

where $\tilde{\boldsymbol{\Lambda}}_{jlk}^{(A)} = \mathrm{diag}(\tilde{\lambda}_{jlk}^{(1)}, \ldots, \tilde{\lambda}_{jlk}^{(A)}) \in \mathbb{C}^{A \times A}$ and $\tilde{\lambda}_{jlk}^{(a)}$ is the average channel attenuation between subarray a in cell j and UE k in cell l. By letting the number of antennas in each subarray grow large, we obtain the following property:

If MR precoding is used and the channel covariance matrices can be factorized as in (2.84), then

$$\mathrm{SINR}_{jk}(t) = \frac{p_{jk}^{\mathrm{DL}} \mathcal{S}_{jk}}{\sum_{l=1}^{L}\sum_{m=1}^{K} p_{lm}^{\mathrm{DL}} \mathcal{I}_{lmjk} - p_{jk}^{\mathrm{DL}} \mathcal{S}_{jk} + \mathcal{O}(\frac{1}{M})} \tag{2.85}$$

where the signal part is

$$\mathcal{S}_{jk} = \mathrm{tr}\left((\tilde{\mathbf{x}}_{jk}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \tilde{\boldsymbol{\Lambda}}_{jjk}^{(A)}) \widetilde{\boldsymbol{\Phi}}_j^{-1} (\mathbf{D}_{\boldsymbol{\delta}(t)} \tilde{\mathbf{x}}_{jk} \otimes \tilde{\boldsymbol{\Lambda}}_{jjk}^{(A)}) \right)$$

with $\widetilde{\boldsymbol{\Phi}}_j = \sum_{\ell=1}^{L} \sum_{m=1}^{K} \mathbf{X}_{\ell m} \otimes \tilde{\boldsymbol{\Lambda}}_{j\ell m}^{(A)} + \sigma_{\mathrm{BS}}^2 \mathbf{I}_{AB}$, where the interference terms $\mathcal{I}_{lmjk}$ with a CLO are

$$\mathcal{I}_{lmjk}^{\mathrm{CLO}} = \frac{\sum_{a_1=1}^{A} \sum_{a_2=1}^{A} \tilde{\lambda}_{llm}^{(a_1)} \tilde{\lambda}_{ljk}^{(a_1)} \tilde{\lambda}_{llm}^{(a_2)} \tilde{\lambda}_{ljk}^{(a_2)} \left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \mathbf{e}_{a_1}^{\mathrm{H}} \right)}{\mathrm{tr}\left(\left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{llm} \right) \widetilde{\boldsymbol{\Phi}}_l^{-1} (\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{jk} \otimes \boldsymbol{\Lambda}_{ljk}) \right)}$$
$$\times \widetilde{\boldsymbol{\Phi}}_l^{-1} \left((\mathbf{X}_{jk} - \kappa_{\mathrm{UL}}^2 p_{jk}^{\mathrm{UL}} \mathbf{I}_B) \otimes \mathbf{e}_{a_1} \mathbf{e}_{a_2}^{\mathrm{H}} \right) \widetilde{\boldsymbol{\Phi}}_l^{-1} (\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{lm} \otimes \mathbf{e}_{a_2})$$

and the interference terms with SLOs are

$$\mathcal{I}_{lmjk}^{\mathrm{SLOs}} = \mathrm{tr}\left(\left(\tilde{\mathbf{x}}_{lm}^{\mathrm{H}} \mathbf{D}_{\boldsymbol{\delta}(t)} \otimes \boldsymbol{\Lambda}_{llm} \right) \widetilde{\boldsymbol{\Phi}}_l^{-1} (\mathbf{D}_{\boldsymbol{\delta}(t)}^{\mathrm{T}} \tilde{\mathbf{x}}_{jk} \otimes \boldsymbol{\Lambda}_{ljk}) \right).$$

The notation $\mathcal{O}(\frac{1}{M})$ is used for terms that go to zero as $\frac{1}{M}$ or faster when $M \to \infty$, while $\mathbf{e}_a$ is the ath column of $\mathbf{I}_A$.

These asymptotic expressions do not contain κ_{UL}, κ_{DL}, σ_{BS}^2, or σ_{UE}^2, thus it shows that the impact of distortion noise and receiver noise vanishes as $M \to \infty$. The asymptotic SINRs are only limited by the channel distributions, pilot-contaminated interference, and phase noise. This means that DL MaMi systems can handle larger additive distortions than conventional systems, which can also be posed as a scaling law for the hardware quality:

Suppose that $\kappa_{\mathrm{UL}}^2 = \kappa_{\mathrm{UL},0}^2 M^{z_1}$, $\kappa_{\mathrm{DL}}^2 = \kappa_{\mathrm{DL},0}^2 M^{z_1}$, $\sigma_{\mathrm{BS}}^2 = \sigma_{\mathrm{BS},0}^2 M^{z_2}$, $\sigma_{\mathrm{UE}}^2 = \sigma_{\mathrm{UE},0}^2 M^{z_2}$, and $\delta = \delta_0(1+\ln(M^{z_3}))$, for some scaling exponents $z_1, z_2, z_3 \geq 0$ and constants $\kappa_{\mathrm{UL},0}, \kappa_{\mathrm{DL},0}, \sigma_{\mathrm{BS},0}^2, \sigma_{\mathrm{UE},0}^2, \delta_0 \geq 0$. The SINRs, $\mathrm{SINR}_{jk}(t)$, with MR converge to non-zero limits as $M \to \infty$ if

$$\begin{cases} \max(z_1, z_2) \leq \frac{1}{2} \text{ and } z_3 = 0 & \text{for a CLO} \\ \max(z_1, z_2) + z_3 \frac{\delta_0 |\tau_{\mathrm{DL}} - B|}{2} \leq \frac{1}{2} & \text{for SLOs.} \end{cases} \tag{2.86}$$

This results shows that the DL can handle additive distortions with variances that scale as $\sqrt{M}$ (i.e., $z_1 = z_2 = \frac{1}{2}$), while achieving decent performance. The scaling law also shows that the phase noise variance with SLOs can increase logarithmically with M, while this is not allowed with a CLO. This proves that MaMi with SLOs are preferable in the DL, at least when the number of antennas is large. The scaling law holds also for any judicious precoder that performs better than MR precoding.

The system model that was used to obtain these results is very general, in order to capture the joint effect of various sources of hardware impairments. The flat-fading channel model can describe either single-carrier transmission over the full available flat-fading bandwidth or one of the subcarriers in a system based on multi-carrier modulation; for example, OFDM. To some extent, it can also describe single-carrier transmission over frequency-selective channels as in [33]. The mapping from the impairments in a certain circuit to the three categories of distortions depends on the modulation scheme; for example, the multiplicative distortions caused by phase-noise leads also to inter-carrier interference in OFDM which is an additive noise-like distortion.

To exemplify how the hardware scaling law can be interpreted, we now consider single-carrier transmission over flat-fading channels for clarity of presentation. In this case, an ADC with b bit resolution causes uncorrelated quantization noise with approximately the power $2^{-2b} P_{\mathrm{signal}}$, where P_{signal} denotes the power of the received signal. The scaling law in (2.86) allows us to increase the variance κ_{UL}^2 and κ_{DL}^2 as M^{z_1} for $z_1 \leq \frac{1}{2}$. This corresponds to reducing the ADC

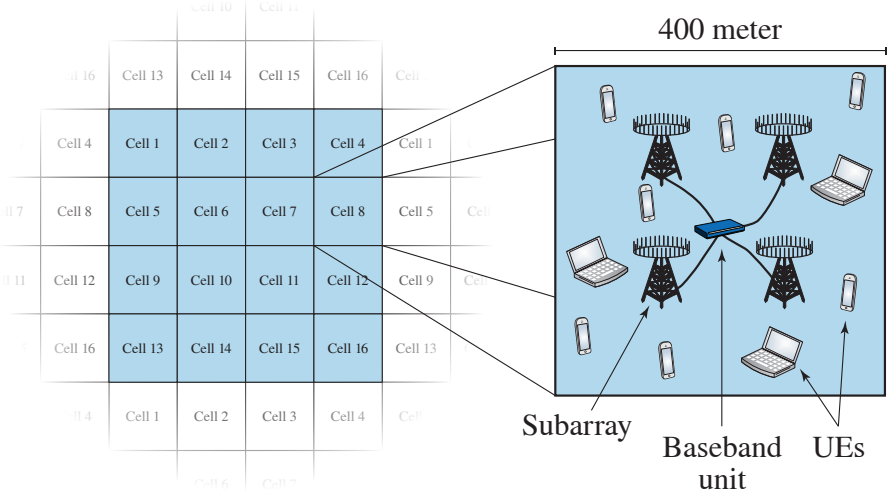


Figure 2.21: Illustration of the multi-cell MaMi scenario with distributed arrays considered in the numerical evaluation.

resolution by around $\frac{z_1}{2}\log_2(M)$ bits. Hence, we can reduce the ADC resolution per antenna by at least 2 bits if we deploy 256 antennas instead of one.

Similarly, the scaling law in (2.86) allows us to increase to increase the noise figure as M^{z_2} for $z_2 \leq \frac{1}{2}$. The noise figure can thus be increased by $z_2 10\log_{10}(M)$ dB. For example, at $z_2 = \frac{1}{2}$ we can allow an increase by 10 dB if we deploy 100 antennas instead of one.

2.5.3 Numerical results

The analytic results are corroborated for the distributed MaMi setup in Figure 2.21. This is a wrap-around topology with 16 cells of 400×400 meters, each consisting of $A = 4$ subarrays with $\frac{M}{A}$ antennas located 100 meters from the cell center. The $K = 15$ UEs per cell are uniformly distributed, with a minimum distance of 25 meters from the subarrays. The transmit powers are $p_{jk}^{\mathrm{DL}} = p_{jk}^{\mathrm{UL}} = -50$ dBm/Hz for all j and k (e.g., 100 mW over 10 MHz). The channel attenuations are modeled as in [5]: $\lambda_{jlk}^{(n)} = 10^{s_{jlk}^{(n)}-1.53}/(d_{jlk}^{(n)})^{3.76}$, where $d_{jlk}^{(n)}$ is the distance in meters between BS antenna n in cell j and UE k in cell l and $s_{jlk}^{(n)} \sim \mathcal{N}(0, 3.16)$ is shadow-fading.

The hardware impairments are characterized by the distortion proportionality coefficients $\kappa_{\mathrm{UL}} = \kappa_{\mathrm{DL}} = 0.03$, the variance of phase noise increments $\delta = 1 \cdot 10^{-5}$, and the receiver noise powers $\sigma_{\mathrm{BS}}^2 = \sigma_{\mathrm{UE}}^2 = -169$ dBm/Hz (with 5 dB noise amplification). These are also the initial constants when we scale the hardware quality based on the scaling law described in (2.86).

Figure 2.22 shows the average spectral efficiency per UE. The coherence block contains $T = 300$ symbols, whereof $B = 15$ symbols are used for pilot sequences and $\tau_{\mathrm{DL}} = 285$ for DL payload data. Hardware impairments incur a performance loss as compared to ideal hardware. The gap is small with SLOs, but larger with a CLO. This validates the analytic observation that SLOs is the better choice in MaMi.

The figure also illustrates the scaling law established in (2.86). The middle curves show the behavior when satisfying the scaling law ($z_1 = z_2 = 0.48$ with a CLO and adding also $z_3 = 0.48$ with SLOs). By gradually degrading the hardware with M, there is a performance loss at every M, but the curves are still increasing with M. The performance loss is small for SLOs, but very

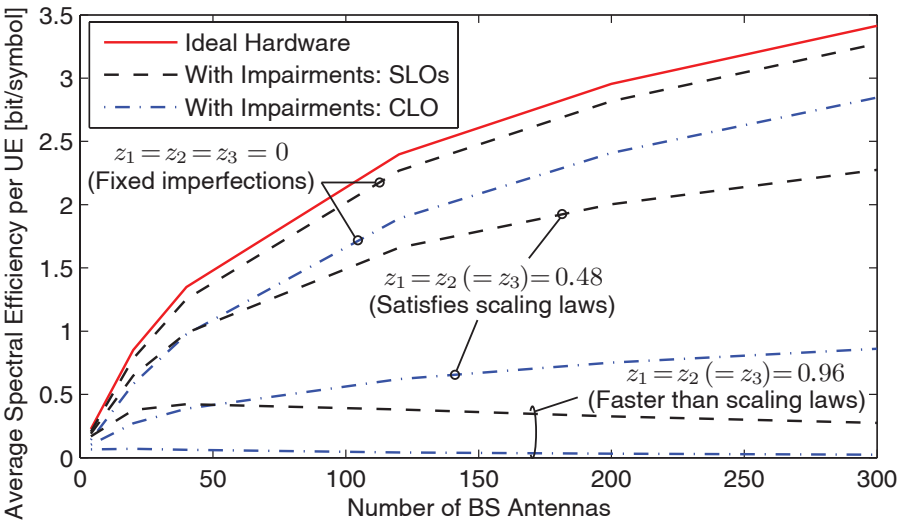


Figure 2.22: Average DL spectral efficiency for distributed MaMi with fixed or increasing hardware impairments.

large for a CLO. The curves at the bottom are for a case when the scaling law is not satisfied, which gives a performance that goes to zero as $M \to \infty$.

2.5.4 Summary of hardware imperfection analysis

We have analyzed the DL performance of MaMi systems, with focus on the impact of hardware impairments. We have proved that additive distortions have smaller impact on MaMi than conventional networks, since the variance may increase as $\sqrt{M}$ with little performance loss. Multiplicative phase noise can be more severe, but the performance is better if each BS antenna has a separate oscillator.

The DL analytic results in this section are in line with previous UL results in [5, 23, 32, 33]. This is natural since the UL-DL duality for systems with linear processing implies that the same performance is achievable in both directions (if the power allocation is optimized). However, our results stand in contrast to the recent works [21, 23] where the DL behave differently than the UL when it comes to phase noise. This is due to different system models: [21] considers high SNRs in non-fading single-user cases, while [23] considers a single cell with relatively good CSI. In comparison, we consider a generalized multi-cell setup with more inter-user interference and thus lower SINRs.

Chapter 3

Signal, noise and interference power in Massive MIMO links

In the last few years, many papers have characterized different performance bounds of MaMi, such as the capacity with either fixed or asymptotically large numbers of antennas and users, under various channel conditions [4, 28, 30]. From a more practical point of view, each user in the system should experience a sufficient signal-to-noise ratio in order to achieve successful communication for the selected modulation and coding scheme. Depending on scenarios, the limiting factor may not be noise only, but also interference from other user streams, certainly when many users are simultaneously active.

For example, initially the MaMi concept was proposed with the so-called conjugate beamforming, i.e., MR precoding [28]. This concept has the benefit of extreme simplicity. In some scenarios it provides very good performance, while in others the more complex ZF precoding is needed to reduce the inter-user interference. In this chapter, we determine the bound on MR operation in terms of number of users and possible modulation and coding scheme.

Section 3.1 briefly describes the selected system and related variables. It characterizes the basic downlink and uplink operation. Section 3.2 computes the signal, noise and interference terms expected under different scenarios. Section 3.3 concludes this chapter and links the results to system power consumption. Indeed, power models such as [11] or as presented in Chapter 4 illustrate the specific system power breakdown for Massive MIMO as compared to traditional systems. Results from this chapter can help refining the requirements on output power based on the equivalent SINR (signal to noise and interference ratio), and hence dimension the system.

3.1 System definition and link assumptions

Let us consider a Massive MIMO system with M antennas at the BS side and K single-antenna UEs. The system is based on time-domain duplexing, with phases of UL training for channel estimation, UL data from the UEs to the BS and DL data towards the UEs.

The analysis is performed assuming OFDM, such that flat-fading is observed on individual subcarriers, and after averaging over subcarriers such that the frequency dimension is omitted. As compared to the case of a narrowband single-carrier system undergoing flat fading, selecting OFDM and averaging power levels over the subcarriers allows for simplifying statistical assumptions, leading after averaging to power levels that are almost constant over independent channel realizations and signal distributions that are Gaussian.

All signals are complex values in baseband representation. The terms signal power, signal energy or signal variance are interchangeably used, and defined as the expected value of the

square of the corresponding signal magnitude, as all considered signals are zero-mean and assuming a normalized time scale of digital samples (sampling period equal to one). For a signal s the corresponding signal power is denoted by σ_s^2. The following signals are used in this analysis: s_i for the precoded signal at one BS antenna sent towards UE i, p_i for the pilot signal sent by UE i in UL training phase, u_i for the UL data signal sent by UE i, and n_i for the received additive noise at UE i during DL. In UL training and data phases, the received additive noise vector over all antennas is denoted N_{UL}. $\mathbf{H}$ is the $K \times M$ channel matrix, assuming i.i.d. Rayleigh fading statistics.

In order to assess the performance of Massive MIMO fairly as compared to traditional systems, we define a reference SISO link with the following assumptions[1]:

1. Output power P_{SISO} at the transmitter side;
2. Expected channel energy $E\{|H|^2\} = 1$, H being scalar a scalar in the SISO case;
3. Additive white Gaussian noise of variance $\sigma_n^2 = 1$;
4. Required SNR at the UE side defined as SNR_{Rx} for successful reception; symmetrically the same SNR is assumed to be required at the BS side in UL, i.e., BS and UE transceivers have similar quality.

Under those assumptions, the requirements on received SNR directly translate into a power specification for the transmitter:

$$\text{SNR}_{Rx} = \frac{P_{SISO} E\{|H|^2\}}{\sigma_n^2} \tag{3.1}$$

$$\Leftrightarrow P_{SISO} = \text{SNR}_{Rx} \tag{3.2}$$

This conclusion is trivial, but the Massive MIMO case is more interesting. It builds on the following assumptions:

1. Output power σ_s^2 per BS antenna and per user, or for the whole BS $P_{total,DL} = MK\sigma_s^2$;
2. Expected channel path energy $E\{|H_{i,j}|^2\} = 1$, where $H_{i,j}$ is the (i,j)th element of $\mathbf{H}$, i denotes the UE and j the BS antenna;
3. Additive white Gaussian noise of variance $\sigma_n^2 = 1$ per receiver antenna, i.e., per UE or per BS antenna;
4. Required SINR[2] equal to SNR_{Rx} at the UE side in DL;
5. UL transmit power σ_u^2 for data symbols, such that after precoder-based combination, corresponding streams at BS side also experience SNR_{Rx} as SINR;
6. UL transmit power for pilots same as for UL data after averaging over subcarriers.

[1]When comparing to fully loaded multi-layer SU-MIMO reference instead of SISO, the assumption is that the $M \times M$ MIMO system transmits the same power as a SISO system on each of its antennas, i.e., in total M times more power. This keeps output energy per bit constant.

[2]SINR is used given that depending on the selected Massive MIMO scheme noise as well as inter-user interference are present and can be modelled as independent additive Gaussian processes.

The required transmit power levels σ_s^2 and σ_u^2 based on those assumptions are computed in Section 3.2. Given that unlike data, pilots may not be present on all subcarriers, the last assumption leads to a pilot power boosting effect ensuring fairness, i.e., the pilot energy per subcarrier grows inversely proportionally to the density of pilot subcarriers used by the UE, assuming other subcarriers are void. The default pilot scheme in this chapter assumes loading every Pth subcarrier with a pilot symbol (comb P approach for LTE Sounding Reference Signal), leading $E\{|p_i|^2\} = P\sigma_u^2$, while the $P-1$ subcarriers in-between are not used. Different UEs use different pilot combs in order to ensure independent estimation for each UE. This approach is used in Massive MIMO as well as in LTE, to exploit the frequency coherence of the channel over multiple subcarriers in order to reduce the amount of pilot symbols [13].

In total, if n_P training OFDM symbols are used in order to be able to estimate channels for all users, the constraint for orthogonal estimation is to have at least $n_P P \geq K$. The assumption that channel coherence in frequency is sufficient for one channel estimate to be valid over P subcarriers allows an energy gain factor P in channel estimation SNR as compared to the uplink data SNR received on a single antenna. This is important in order to partially compensate for the fact that precoding gain is not available during the training phase. The same gain would come equivalently from using orthogonal sequences using all subcarriers but no power boosting. Indeed, sending a pilot over each subcarrier with power σ_u^2 and using a code of length P leads to a recombination gain P thanks the the assumption of a fixed channel over P subcarriers, leading coherent recombination of the P subcarriers and non-coherent noise addition.

3.2 Link analysis with interference and channel estimation errors

In this section, we estimate the expected energy coming from the different signal, noise and interference components, and use it to predict the system performance. Thanks to the large number of antennas, users and subcarriers, most signals are summed or averaged over many components, leading to accurate Gaussian approximations and also to values with limited fluctuations for non-zero-mean variables. For instance, the coherent precoder-based combination of the M antenna streams leads to an average useful signal energy distributed with little variation, despite the fact that channel coefficients themselves are Rayleigh fading and hence show large energy fluctuations.

Pilot p_i is first sent by user i in order to estimate its channel, leading to the estimate $\hat{\mathbf{H}}_i$ of the ith row $\mathbf{H}_i$ of the channel matrix $\mathbf{H}$. The corresponding transmission used for channel estimation is subject to a received noise vector $\mathbf{N}_i$ of length M at the base station side. Once the channel has been estimated for all users, it can be used to derive for each user the precoder $\tilde{\mathbf{H}}_i$ used in downlink, or equivalently used as decoder in uplink. In order to guarantee that each BS antenna transmits at a constant power of σ_s^2 over different channel realization, the only additional constraint is that the total energy of the precoder should be normalized to one per component of the vector[3]:

$$\tilde{\mathbf{H}}_i = \frac{\hat{\mathbf{H}}_i}{\|\hat{\mathbf{H}}_i\|}, \tag{3.3}$$

where $\|\hat{\mathbf{H}}_i\| = \frac{\sum_j |H_{i,j}|^2}{M}$.

[3]The normalization factor tends to 1 as the number of antennas increases; without this factor the expected received power factor becomes $M^2 + M$ instead of M^2 in (3.7), according to the Lemma 2 of [5].

Assuming an MR (maximum ratio) transmission precoder, the steps corresponding to channel estimation, downlink communication, and uplink communication for user i are the following, respectively, where $()^T$ denotes the transpose and $()^H$ the conjugate transpose:

$$\mathbf{y}_{est,i} = \mathbf{H}_i^T p_i + \mathbf{N}_i \tag{3.4}$$

$$y_{DL,i} = \mathbf{H}_i \tilde{\mathbf{H}}_i^H s_i + \sum_{k \neq i}^{K} \mathbf{H}_i \tilde{\mathbf{H}}_k^H s_k + n_i \tag{3.5}$$

$$\mathbf{y}_{UL,i} = \mathbf{H}_i^T u_i + \sum_{k \neq i}^{K} \mathbf{H}_k^T u_k + \mathbf{N}_{UL} \tag{3.6}$$

In the ideal MR case, $\hat{\mathbf{H}}_i = \mathbf{H}_i$ and the signal s_i corresponding to user i is precoded by the conjugate transpose of the corresponding row of the channel matrix, multiplied by the normalization constant. The corresponding received signal by the UE has the following expected energy, while the corresponding noise power is σ_n^2:

$$P_{DL} = E\left\{ |\sum_j H_{i,j} \tilde{H}_{i,j}^*|^2 \right\} \sigma_s^2 = M^2 \sigma_s^2 \tag{3.7}$$

Due to the symmetry of precoder used in DL for precoding or in UL for coherent signal combination, the total transmit power corresponding to one user should be identical in uplink and in downlink, i.e., $\sigma_u^2 = M\sigma_s^2$, given that BS has M antennas and UE only one antenna. This can be seen by computing the expected SNR in UL (after combination) and in DL, where in (3.9) the coherent combination of M antennas leads a gain M^2 on the useful signal and M on the noise:

$$\text{SNR}_{DL} = \frac{M^2 \sigma_s^2}{\sigma_n^2} \tag{3.8}$$

$$\text{SNR}_{UL} = \frac{M^2 \sigma_u^2}{M \sigma_n^2} \tag{3.9}$$

$$\text{SNR}_{DL} = \text{SNR}_{UL} \Leftrightarrow M\sigma_s^2 = \sigma_u^2 \tag{3.10}$$

3.2.1 Interference

Given the presence of multiple users in the system, the MR precoded streams create some inter-user interference. The related signal comes from the second term in (3.5) and its energy is the following when using the un-normalized precoder $\tilde{H}_i = \hat{H}_i$:

$$\begin{aligned} P_{I,DL} &= E\left\{ \left| \sum_{k \neq i}^{K} \sum_j H_{i,j} H_{k,j}^* \right|^2 \right\} \sigma_s^2 \\ &= (K-1) M \sigma_s^2, \end{aligned} \tag{3.11}$$

where we have used the fact that the product of two independent complex Gaussian variables of variance 1 has a variance 1, as well as channel independence over different antennas. Adding the normalization constraint leads to an additional factor $M/(M-1)$ in this term, based on the inverse chi-square distribution with $2M$ degrees of freedom of the normalization factor. This factor is neglected in the remainder of this analysis for simplicity, given its very limited impact (0.04 dB with 100 antennas).

3.2.2 Channel training

Due to channel training, channel estimates for both the user under consideration and the interfering users from the same BS are noisy. This leads to several additional interference terms. Based on (3.4) the channel is estimated as:

$$\hat{\mathbf{H}}_i = \frac{\mathbf{y}_{est,i}^T}{p_i} \tag{3.12}$$

This leads to an error $\hat{\mathbf{H}}_i - \mathbf{H}_i$ of variance σ_n^2/σ_p^2 or $\sigma_n^2/(P\sigma_u^2)$ on each BS antenna. This error creates interference at the receiver side by affecting the first term of (3.5) due to the non-ideal precoding, which can be treated as an additional noise source. This noise term $\mathbf{H}_i(\hat{\mathbf{H}}_i - \mathbf{H}_i)^H s_i$ is independent of the signal s_i due to the independence of the channel estimation error term. It has the following energy obtained by adding its M components:

$$\begin{aligned} P_{Chest,U,DL} &= \frac{M\sigma_n^2}{P\sigma_u^2}\sigma_s^2 \\ &= \frac{\sigma_n^2}{P} \end{aligned} \tag{3.13}$$

Thanks to the pilot boosting effect enabled by channel coherence in frequency, we can see that the impact of the channel estimation noise is P times smaller than the impact of the direct thermal noise in the DL direction. However, channel estimation errors corresponding to the other users also lead to similar additional noise sources from the second term in (3.5). Due to the independence of channel estimation error terms, those additional noise sources have an equivalent energy for each user, leading in total for the $(K-1)$ interfering users to the following contribution:

$$P_{Chest,I,DL} = \frac{(K-1)\sigma_n^2}{P} \tag{3.14}$$

3.2.3 Overall analysis

The energy of the different signal, noise and interference terms has been computed at the receiver side in DL. By combining them, the equivalent SINR determining system performance can be computed and constrained to be equal to SNR_{Rx} for successful reception:

$$\mathrm{SNR}_{Rx} = \frac{M^2\sigma_s^2}{\sigma_n^2 + (K-1)M\sigma_s^2 + \sigma_n^2/P + (K-1)\sigma_n^2/P} \tag{3.15}$$

The denominator terms are responsible for the thermal noise, the inter-user interference from (3.11), the channel estimation error on the useful user from (3.13) and channel estimation error on the other users from (3.14), respectively. Depending on the selected scenario, we can use its result to validate the system operational range.

Let us first consider an ideal case in the absence of any impairment, i.e., having a single user and using ideal CSI. Then only the first term remains in the denominator and the requirement in transmit power illustrates the Massive MIMO gain:

$$\mathrm{SNR}_{Rx} = \frac{M^2\sigma_s^2}{\sigma_n^2} \tag{3.16}$$

$$\Leftrightarrow \sigma_s^2 = \frac{P_{SISO}}{M^2} \tag{3.17}$$

$$\Leftrightarrow P_{total,DL} = \frac{KP_{SISO}}{M} \tag{3.18}$$

In the last equation we have re-introduced K users. It is hence a bound on the minimal power requirement for MR in downlink, given that the impact inter-user interference is not considered. It is actually a bound on any linear precoder, given that other precoders such as ZF that can remove inter-user interference will be sub-optimal as compared to MR with respect to the useful signal energy, due to the use of some degrees of freedom for interference cancellation instead of energy maximization. This bound illustrates the well-known relationship between number of BS antennas and total output power requirement.

Considering the UL direction in this ideal case, i.e., combining (3.10) and (3.17), we can see that the Massive MIMO gain also enables a linear power reduction at the UE side.

In the general case including impairments from interference and channel estimation, the required transmit power is computed by transforming (3.15) into the following equivalent equation, noting that $P_{SISO} = \mathrm{SNR}_{Rx}$ and $\sigma_n^2 = 1$:

$$\begin{aligned} \left(1 + \frac{K}{P}\right) P_{SISO} &= \left(M^2 - \mathrm{SNR}_{Rx}(K-1)M\right)\sigma_s^2 \\ \Leftrightarrow \sigma_s^2 &= \frac{\left(1 + \frac{K}{P}\right) P_{SISO}}{M^2\left(1 - \frac{(K-1)\mathrm{SNR}_{Rx}}{M}\right)} \end{aligned} \tag{3.19}$$

As compared to (3.17), this leads to an interesting interpretation of the system requirements under non-ideal conditions. First, the channel estimation noise on the considered user and more importantly on the other users lead to the factor $(1 + K/P)$ on the required transmit power. When $P = K$ which is expected to be typical in order to minimize the pilot overhead, this directly translates into a 3-dB shift in SNR in order to compensate for the estimation noise. In the unlikely case of a scenario having a larger number of users than the coherence bandwidth of the channel, multiple training symbols would be required and the degradation would increase as $K/P > 1$. This comes from the fact that the pilot boosting factor P does not compensate for UEs being silent during $(n_P - 1)$ training symbols. Hence, in order to keep the performance loss to 3 dB while not exceeding the pilot boosting factor, each UE has to send energy during the complete training phase, using a training code of length $n_P P$ spanning all training symbols.

On the other hand, when the number of users is smaller than the channel coherence bandwidth, keeping a number of pilot sequences $P > K$ enables to reduce the performance loss, thanks to the averaging of the channel estimate over the full coherence bandwidth. In practical systems, the block-fading approach used in frequency-domain is only approximately valid. There can hence be relevant trade-offs to explore depending on selected channel interpolators in the frequency domain, between minimizing channel estimation noise by averaging over more subcarriers and minimizing channel deviation from the block-fading assumption by using finer-resolution estimation.

The second corrective factor $(1 - (K-1)\mathrm{SNR}_{Rx}/M)$ in the denominator of (3.19) represents the effect of MR inter-user interference. It depends on the system load K/M and the required SNR_{Rx} which is function of the selected modulation and coding scheme. It also puts a bound on the maximum load of an MR system, for a given modulation and coding scheme. For example, considering QPSK and LDPC encoding (based on IEEE 802.11ac) at rate 3/4, an error-free performance is possible around $\mathrm{SNR}_{Rx} = 5.5$ dB [12]. This means that for $M = 100$ antennas, the number of users may never exceed 29 users (at infinite transmit power) and practically in order to limit the output power increase due to inter-user interference to 3 dB, only half of this value is allowed, i.e., 15 users.

Based on i.i.d. Rayleigh channels of equal expected power, Figure 3.1 validates those conclusions: a 100×1 system using the selected modulation and coding should benefit from a 20 dB

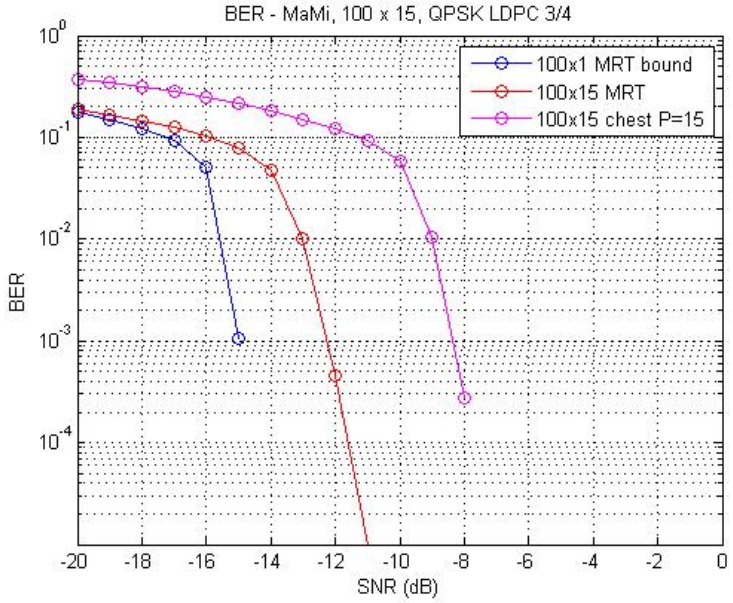


Figure 3.1: Comparison of MR transmission performance between the ideal bound (100×1 with ideal CSI), inter-user interference (100×15 with ideal CSI) and interference with channel estimation error (100×15 with channel estimation from one pilot every 15th subcarrier). The system uses 1200 subcarriers out of 2048 based on LTE specifications and the channel model is time-domain Rayleigh with 20 taps of equal expected energy.

gain and hence work without errors at an SNR of $\sigma_s^2 = -14.5$ dB, which is the case. When 15 users are present, the related inter-user interference leads to some 3 dB degradation. Moreover, when using as many pilots as users the additional degradation is again around 3 dB as expected. It is actually slightly more, due to the implicit assumption of constant channel response over groups of 15 subcarriers while the actual channels show some limited fluctuations within a coherence band. The figure also validates the statistical averaging in the system, providing performance close to the AWGN-based reference SISO operation at 5.5 dB despite the use of Rayleigh-fading channels, thanks to the large diversity in Massive MIMO systems.

3.3 Conclusions

This chapter has derived relative energy levels of signal, noise and interference components in Massive MIMO systems using MR transmission. In the ideal case it reproduces the well-known Massive MIMO gain leading a reduction in total output power proportional to the number of antennas, in downlink as well as in uplink, and justifying the operation in the negative SNR region. More importantly, the impact of both multi-user interference and channel estimation error has been derived. For a number of pilot sequences equal to the number of users, the channel estimation error causes a performance loss of 3 dB. When more than one training symbol is needed, orthogonal pilot sequences have to span multiple symbols in order to keep this amount of degradation. On the other hand, when few users are present, the excess coherence bandwidth of the channel can be used to reduce this performance loss, or the corresponding resources can be used for UL data transmission.

Concerning the multi-user interference, the MR operation was shown to support a maximum number of users, function of the selected modulation and coding scheme. For example, QPSK with LDPC coding rate 3/4 supports a number of users which is 15% of the number of BS

antennas if we allow 3 dB of related degradation, and never more than 29%.

A shift in required output power is critical to dimension the system, given the trade-off between output power (reduced by adding more antennas) and overhead analog and digital power (increased when adding more antennas) [11]. This approach can further be extended. For example we could include the impact of pilot contamination, of different path loss and power control requirements for the different UEs, of different precoders or of additional impairments coming from the hardware implementation.

Chapter 4

Baseband processing profile

In this chapter, we analyze and profile some of the presented baseband processing algorithms, with the focus on computational complexity and potential power consumption. The impact (of different algorithms) on processing distribution strategy, data shuffling bandwidth, and memory requirement will also be discussed.

4.1 Computational complexity and power consumption

MaMi is a promising technology in order to both increase capacity and reduce power consumption for 5G systems [25]. Concerning the power consumption, the required output power is reduced inversely proportionally to the square root of number of BS antennas, or even linearly in operating regimes with good channel estimation quality, thanks to the coherent combination of all antennas using channel-based precoding. An important question is whether the hardware power consumption related to the larger number of antenna chains is not counterbalancing this benefit. Fortunately, the power consumption of all components can remain small enough to keep a large benefit from the MaMi approach [6, 11].

As an illustration, Table 4.1 specifies a reference macro BS scenario as well as three MaMi alternative scenarios dimensioned for a similar coverage, based in each case on 1 PA per antenna, i.e., M PAs in total. but providing a full-buffer throughput being smaller, similar or larger than the reference macro case. The three scenarios mostly differ in number of antennas and number of simultaneous users. All those scenarios are based on LTE parameters in a 20 MHz band. The reference macro uses FDD while MaMi uses TDD. Figure 4.1 illustrates the power consumption of the four scenarios defined in Table 4.1, based on using [9]. Especially, Scenario 3 which is providing the same throughput as the reference BS, illustrates the gain of a factor 35 in power consumption. The scenarios of Table 4.1 have been seleted in order to target a typical throughput similar to traditional macro base stations, while being used to illustrate the large potential for power savings. Alternatively, operators may install one full Massive MIMO system per each sector. This will end up with a Massive MIMO solution offering significantly more throughput than traditional base stations, while still consuming less power, i.e., typically 3 times more thoughput from 10 times less total power consumption.

This power modeling effort validates the concept of MaMi from the point of view of power efficiency when compared to traditional BSs. However, in order to better understand implementation challenges of MaMi and design such systems for optimal efficiency, a number of elements should be added or revisited in this model. For example, only the simplest MR (maximum ratio) precoder was modeled, while the more complex ZF or RZF precoding is required in many scenarios. Moreover, the digital complexity of some DSP blocks was not accurately modeled

Table 4.1: Definition of cellular BSs investigated for power consumption. The macro reference design suffers 3-dB feeder losses between PA and antenna.

Scenario	1	2	3	4
Type	Reference macro	Small Massive	Medium Massive	Large Massive
Antennas[1] $M \times K$	4×4	100×10	100×25	400×100
Output per PA	46 dBm	8 dBm	17 dBm	11 dBm
Sectors	3	1	1	1
Total radiated	49 dBm	28 dBm	37 dBm	37 dBm
Precoder		MR	ZF	ZF
MCS	16-QAM 3/4	QPSK 3/4	16-QAM 3/4	16-QAM 3/4
Frame structure[2]		(14, 1)	(14, 2)	(28, 7)
Throughput	1100 Mbps	200 Mbps	1000 Mbps	3400 Mbps

[1] Multi-layer SU-MIMO in the macro case and single-layer MU-MIMO with single-antenna users in MaMi scenarios

[2] Total number of OFDM symbols in one frame and number of those symbols being used for channel acquisition.

in the initial version of the power model. Finally, alternative analog architectures could be considered in order to further reduce the system power consumption.

Subsection 4.1.1 presents the general power modeling framework inspired from [11] and lists the elements being updated in this chapter. Subsections 4.1.2, 4.1.3, and 4.1.4 describe the updated models of PA, digital components and analog components, respectively. Subsection 4.1.5 presents power consumption results based on the updated model, illustrating the relative importance of the different components.

4.1.1 Overall approach

The model presented in this chapter takes its roots in the GreenTouch project [1]. The objective of GreenTouch was to pave the way for an ambitious 1000x increase in energy efficiency of cellular networks. Many ingredients were proposed, one of them being Large-Scale Antenna Systems (LSAS) which is another name for MaMi [28]. The power model developed in the GreenTouch project targets comparisons between different power-saving techniques, for example comparing traditional large cell as well as small cell designs to MaMi solutions [8, 10]. This model is available at [9]. Based on best-effort estimation of the power consumption of the different components, it enables quantitative comparisons that illustrate, e.g., the large benefit of MaMi over traditional architectures.

The model splits the BSs into five main components. Digital baseband (physical layer), digital control and backhauling, and analog front-end are the first three components. Those are modeled based on reference values of complexity (for digital) or power (for analog), in well-defined scenarios. In order to scale the power consumption to any arbitrary scenario, the dependency of each subcomponent with respect to system parameters (bandwidth, number of antennas, system load...) has been investigated and implemented into the model. The impact of silicon technology has also been incorporated, given that it leads to a reduction of power consumption of digital and analog components over the years.

The fourth component is the power amplifier (PA). Depending on scenarios, its power consumption is derived from power efficiency characteristics, which are function of the type of

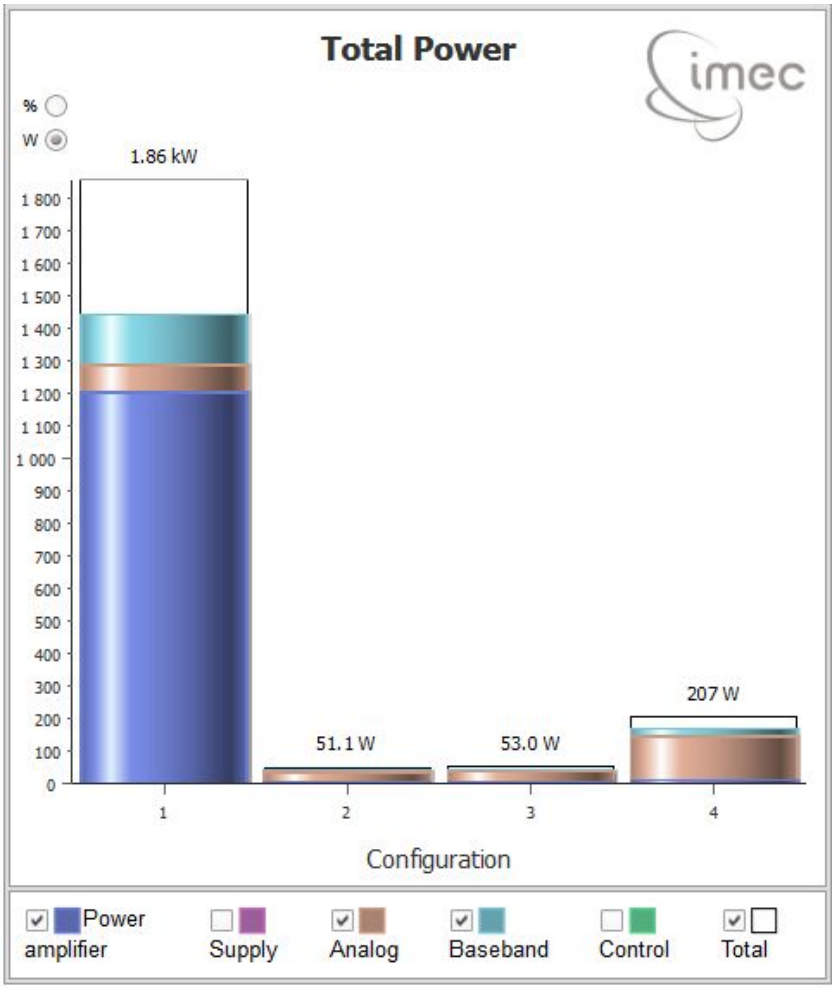


Figure 4.1: Power consumption for the reference macro BS (Configuration 1) compared with the three MaMi scenarios of Table 4.1, based on technology year 2014 [9].

PA architecture and requirements in linearity. The fifth component represents power supply (such as AC/DC conversion) and BS cooling when needed. Those terms are proportional to the power of the other components. In this chapter, the digital control part is neglected as it is independent of the BS type. Moreover, the power supply overhead is kept according to [11]. The other three components are updated according to Subsections 4.1.2 to 4.1.4.

The system uses M antennas at the BS side and serves K users. A frame consists in F_{Total} OFDM symbols out of which F_{Chest} are used for channel training, while the others are shared between uplink and DL data. Assuming a 20-MHz bandwidth based on LTE parameters, data at the antenna side is sampled at $f_s = 30.72$ MHz and assembled into OFDM symbols of 2048 subcarriers and 144-point cyclic prefix. After removing margins, there are 1200 used subcarriers, corresponding to a constellation symbol rate of $f_c = 18$ MHz.

Channel estimation in MaMi exploits the frequency-domain coherence of the channel in order to reduce the number of training symbols. A value of $N_c = 15$ subcarriers is used for channel coherence, i.e., a channel estimate from one subcarrier can be used over 15 neighboring subcarriers with almost no performance degradation, as shown in 2.

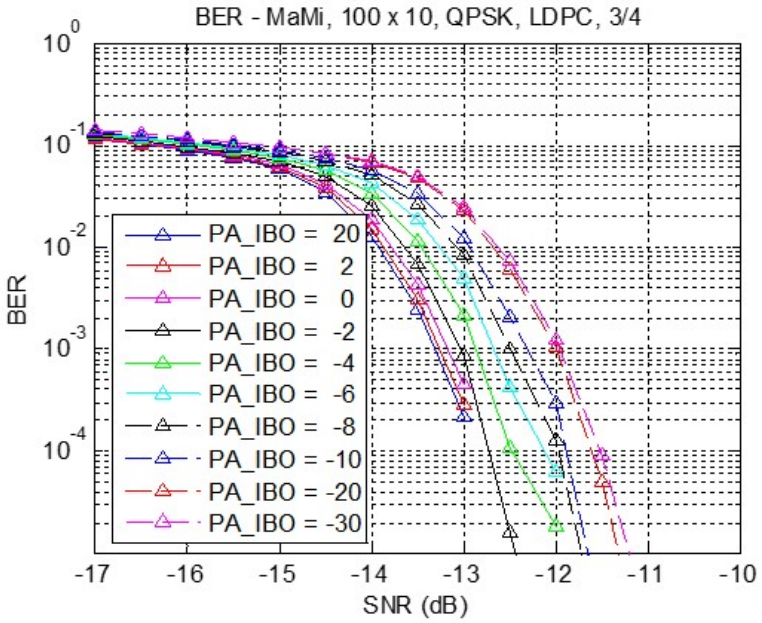


Figure 4.2: Impact of PA input back-off (IBO) on system performance: linear operation (+20 dB) leads the optimum performance, entering the saturation region (0 dB and below) leads a limited degradation and complete saturation (down to -30 dB back-off) a degradation around 1.5 dB.

4.1.2 PA and output power

Modeling the PA efficiency can be done with reasonable accuracy. PAs for MaMi systems are expected to be simple and will not contain the complex feedback and predistortion architectures used in macro BSs in order to provide high efficiency at low distortion. The distortion can be sacrificed while focusing only on high-efficiency non-linear PAs. In [11] an efficiency of 50% was assumed. As illustrated on Figure 4.2, even a completely saturated PA only leads to 1.5 dB degradation on the performance. A power efficiency close to 60% is hence expected to be feasible and taken as assumption in this chapter.

Accurately modeling the output power requirement in different scenarios is more difficult. The classical assumption is based on independent channel coefficients and coherent combining, with a power reduction proportionally to M in regimes where the channel estimation quality is good. This gain is definitely present and is central to the MaMi concept, but its exact value might differ from the asymptotic theory. This assumption is used in Table 4.1: based on a radiated power of $P_{Ref} = 43$ dBm/stream, the requirement on output power is computed as follows, with $P_{Antenna}$ provided in Table 4.1 and R_{Mod} accounting for the difference in SNR requirement between different modulations, i.e., 5 dB lower when using QPSK instead of 16-QAM:

$$P_{Total} = \frac{KP_{Ref}}{MR_{Mod}}, \tag{4.1}$$

$$P_{Antenna} = \frac{KP_{Ref}}{M^2R_{Mod}}. \tag{4.2}$$

The following elements will influence the power requirements. The first one plays in favour of MaMi while the others are detrimental. Future studies based on actual propagation models and simulating the corresponding system performance will be needed in order to refine the gain value:

- MaMi benefits from additional diversity, leading to channel hardening.
- The actual channel components might not be independent over the array.
- Multi-user interference reduces the gain, but the interference is also smaller on correlated channels.
- Channel estimation is not perfect.
- Antennas might not be omni-directional.

More details on the link performance based dimensioning are provided in Chapter 3, especially covering the channel estimation impact.

4.1.3 Digital complexity

Based on [11], digital power consumption is estimated by assessing the number of arithmetic operations and converting it into power consumption. The conversion efficiency is based on technology year, type of digital components and quantization level. For dedicated implementations, the conversion factor for 2015 based on 4-bit quantization is computed to be 400 GOPS/W (Giga complex arithmetic OPeration per Second, per Watt) at the selected resolution (4 bits real and 4 bits imaginary). The successful operation at such a low resolution was validated for MaMi [12]. A factor 2 of overhead is taken in order to account for memory operations, typically as costly as arithmetic operations.

The complexity of the different operations is reviewed: channel estimation, precoder computation, baseband filtering, FFT, mapping, and channel coding. Channel estimation can either be done by allocating different subcarriers to different users and estimating the channel on each individual N_c'th subcarrier as in [11], or use orthogonal codes spanning the coherence bandwidth equivalent to N_c subcarriers. In the second case accumulation of the corresponding values leads to a larger complexity:

$$C_{Chest,individual} = \frac{K}{N_c} M f_c, \tag{4.3}$$

$$C_{Chest,orthogonal} = K M f_c. \tag{4.4}$$

Channel estimation that is carried out over N_c subcarriers leads to an N_c times higher effective SNR in the channel estimation, but an equivalent gain is obtained in the other approach by increasing the power of used pilot subcarriers by a factor N_c, given that other subcarriers are not loaded, hence the total transmitted power in pilot symbols remains the same.

The implementation of the ZF precoder is performed assuming first the computation of the $\mathbf{H}\mathbf{H}^H$ matrix product, from the $K \times M$ multi-user channel matrix $\mathbf{H}$, and secondly inversion of the obtained $K \times K$ matrix. For the product computation, each element requires an adder (ADD) and a multiplier (MUL), hence the factor 2 in the numerator of (4.5). However, thanks to the Hermitian symmetry, only half of the elements have to be computed, hence the factor 2 at the denominator. The Gauss-Jordan inversion has complexity as $3K^3$. Both operations only have to be performed every N_c subcarrier thanks to the channel coherence bandwidth. Optionally, interpolation techniques can be applied on the channel or its inverse in order to improve the system performance, as proposed in Chapter 2. The complexity of ZF precoder

computation operations are:

$$C_{HH} = \frac{2}{2N_c} MK^2 f_c, \tag{4.5}$$

$$C_{Inv} = \frac{3}{N_c} K^3 f_c. \tag{4.6}$$

More specific inversion algorithms in order to reduce the complexity are generally not needed, the computation of the matrix product being generally dominant as $M > 3K$ in typical MaMi scenarios.

The baseband filter from [11] was assuming 40 taps at baseband and an oversampling factor of 2. This specification comes from large BSs and is not needed for MaMi; 10 taps are expected to be sufficient and a polyphase implementation prevents a doubling of the complexity with oversampling. Using fewer taps in MaMI is possible given that out-of-band specifications are relaxed thanks to the lower total power level. The corresponding complexity is obtained as follows, noting that each tap implies one ADD and one MUL:

$$C_{Filter} = 2 \cdot 10 M f_s. \tag{4.7}$$

FFTs require 3 complex operations per butterfly and per stage (10 if real operations are counted). Given that 12 stages are needed (one more than $\log_2(2048)$) while one butterfly processes 2 symbols, this leads the following complexity:

$$C_{FFT} = \frac{12 \cdot 3}{2} M f_s. \tag{4.8}$$

Symbol mapping and demapping onto constellation points has a low complexity, taken from [11] and function of the spectral efficiency s (in bps/Hz):

$$C_{Mapping} = s^{1.5} K f_c. \tag{4.9}$$

Finally, channel coding is based on LDPC codes for high performance. A number of 35/2 operations per stage per bit are required, where the division by 2 is in order to get complex operations. As 5 iterations present a good trade-off between complexity and performance, this leads the complexity in (4.11). Encoding is simpler, 14 operations (the check node degree of the selected code) are required for encoding 3 bits. However, those operations are only binary, hence their complexity is assumed to be 1/8 of a low-resolution complex arithmetic operation as used elsewhere in this analysis. This gives the encoding and decoding complexities

$$C_{LDPC,enc} = \frac{14/3}{8} sK f_c, \tag{4.10}$$

$$C_{LDPC,dec} = \frac{5 \cdot 35}{2} sK f_c. \tag{4.11}$$

Table 4.2 summarizes the complexity after scaling it to our scenario 2. Filtering and FFT dominate as are active on all OFDM symbols and more importantly all antennas, while most other components are only active during some phases and scale only with the number of users.

4.1.4 Analog components

Based on [11], the power consumption of analog components can be significantly larger than digital components for MaMi. The expected order of magnitude of analog power consumption

Table 4.2: Digital complexity for scenario 2 (100×10 MaMi), in GOPS during the corresponding phase.

Component	Downlink data	Uplink data	Training
Baseband filter	123	123	123
FFT/IFFT	111	111	111
Precoding/decoding	72	72	
Direct channel est.			2
Orthogonal channel est.			36
$\mathbf{H}\mathbf{H}^H$ product			24
Gauss-Jordan inversion			7
Mapping/demapping	1	1	
Channel coding	3	47	

Table 4.3: Power consumption of analog MaMi components based on scaled traditional architecture vs. alternative digital RF implementation prospects, per antenna assuming scenario 2 from Table 4.1.

Subcomponent	Downlink [mW]		Uplink [mW]	
	Traditional	Digital RF	Traditional	Digital RF
Predriver	68	17		
Modulator	119	25		
Frequency synthesis	74	10	74	10
Clock generation	4.5	1.5	4.5	1.5
DAC (Digital-to-Analog Converter)	22	3.5		
LNA (Low-Noise Amplifier)			74	4
Mixer			119	25
VGA (Variable-Gain Amplifier)			37	
ADC (Analog-to-Digital Converter)			17	1.5

for 2015 is around 300 mW per antenna in downlink as well as in uplink. Unlike digital components where reduced resolution leads to quantization with fewer bits and hence directly to reduced power consumption, analog components do not have such a direct relationship between resolution and power consumption. However, considering completely different analog architectures more suited to MaMi systems, such as those of WP2, could help bringing the power down.

One such architecture trend is digital RF. The idea is to replace several analog components by digital components performing equivalent operations. The main advantage is to benefit from deeply-scaled CMOS technology, which enables a strong reduction in digital power consumption [19]. Essentially, the upconversion is not performed by an analog mixer anymore, but through a digital upsampling filter, as well as a direct digital RF modulator in downlink. This removes the power consumption of mixers. This also relaxes the power consumption of clock and frequency synthesis, given that no pure carrier tone is required but only an equivalent clock at the same frequency. Low-resolution DACs/ADCs are sufficient and consume significantly less, given that those signals do not need to drive as much current in the absence of mixers. This also holds for LNAs.

Table 4.3 compares the per-antenna power figures derived from [11] with the power figures estimated for a digital RF architecture. All figures are scaled to 2015, i.e., the effect of technology evolution on the intrinsic power consumption of the different components is taken into

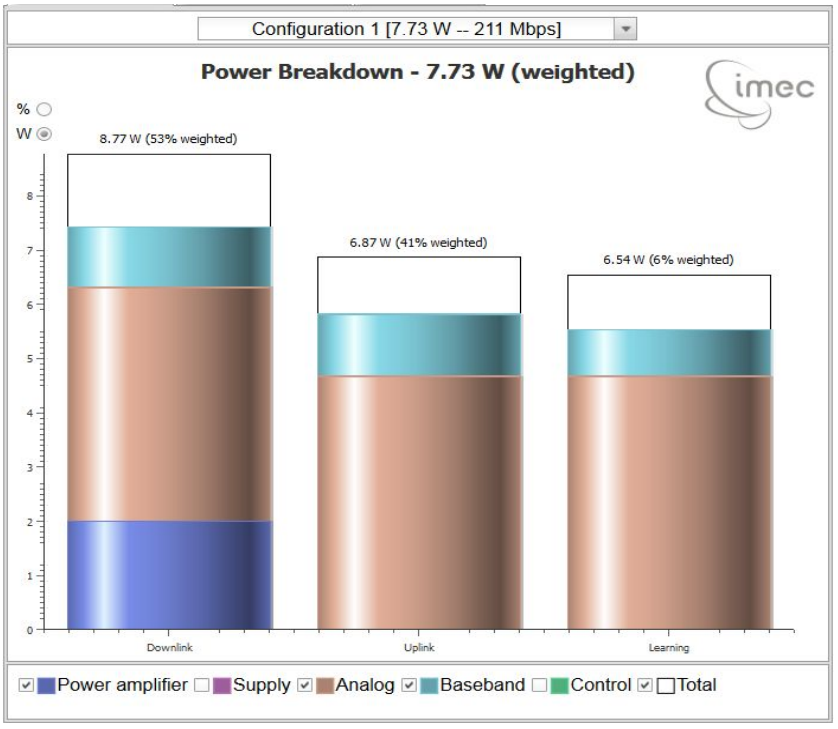


Figure 4.3: Power breakdown in downlink, uplink and training phases for a 100×10 MaMi system using MRT precoding, QPSK and LDPC coding rate 3/4 [9].

account. Those projections are still subject to the condition that digital RF architecture can be demonstrated, especially in uplink where potential show-stoppers such as saturation through out-of-band blockers requires further investigation.

4.1.5 Power trends and conclusions

Figure 4.3 illustrates the power breakdown for our Scenario 2. It shows that anticipating advanced analog architectures with optimistic power figures can reduce the total power by a factor 6 as compared to Figure 4.1, but the optimized analog components still dominate the total power. For Scenario 3 with higher spectral efficiency, which is the scenario closest to the reference macro BS in terms of maximum throughput, the share related to output power increases due to the larger number of users and higher-order modulation, which both translate into a larger SNR requirement at the receiver.

The MaMi power model presented in this chapter improves state-of-the-art versions by introducing new elements such as ZF precoding or advanced analog architectures. It shows a path towards further reduction of the power consumption and illustrates the dominant elements. A number of points will benefit from further validation in the last year of the project. First, the link budget analysis and determination of output power should be investigated based on real MaMi channel models, such as the measurements reported in MAMMOET Deliverable D1.2, and not only theoretical channel models. Secondly, although the complexity of digital computations has become more accurate, the conversion factor into power consumption should be revisited based on digital platform expectations, especially given the fact that different DSP blocks could be implemented to different hardware components with very different energy efficiencies. Finally, further validation should support the anticipation of digital RF architectures.

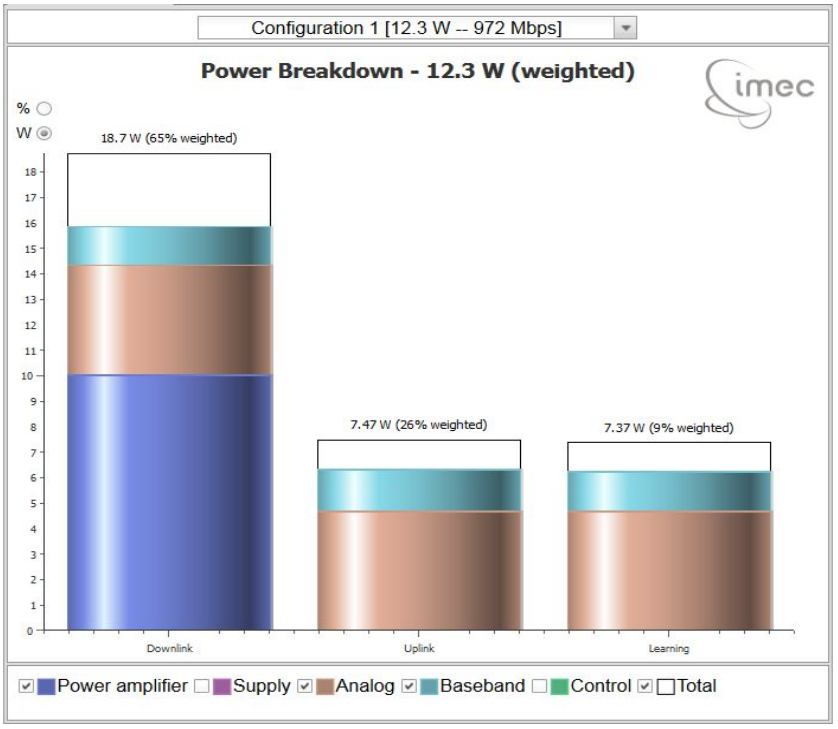


Figure 4.4: Power breakdown in downlink, uplink and training phases for a 100×25 MaMi system using ZF precoding, 16-QAM and LDPC coding rate 3/4 [9].

4.2 Processing distribution and impact

This section considers challenges and analyzes possible solutions to processing distribution in a MaMi system and their impact on overall functionality.

Considering the baseband processing blocks from Figure 1.1 the processing in MaMi system is mainly divided in three different groups, i.e., per-antenna processing, per-subcarrier processing and per-user processing. It is quite natural to distribute the processing in the same manner. Per-antenna processing, meaning the Digital Front End (DFE), OFDM modulation and demodulation are most efficiently implemented in accelerators rather than a processor. Accelerators are in common faster and more energy efficient whereas processors are highly reconfigurable. MaMi channel estimation, detection, precoding and reciprocity calibration which define the per-subcarrier calculation domain are candidates for processors and Processing Elements (PEs) to enable reconfigurability and alternative algorithm usage. Finally, symbol-mapping/demapping, (de)interleaving and channel coding/decoding define the per-user processing. As these are memory and computation extensive but do not require much reconfigurability they are best suited as accelerators.

MaMi poses many challenges regarding implementation of efficient circuit design in the digital baseband processing, some of which will be discussed shortly here. Whenever we give examples throughout this text, we will use the PHY parameters, borrowed from current LTE systems as given in Table 4.4.

4.2.1 Processing latency

One major challenge is the low-latency high-throughput processing. Figure 4.5 is a simplified timing diagram for a MaMi TDD system including UL pilot, UL data, guard and DL data slots. Shaded boxes show data that is received at the antenna according to frame structure whereas

Table 4.4: High-level system parameters

Parameter	Variable	Value
Carrier frequency	f_c	3.7 GHz
Sampling Rate	f_s	30.72 MS/s
FFT Size	N_{FFT}	2048
# Used subcarriers	N_{SUB}	1200
Slot time	T_S	0.5 ms
Sub-Frame time	T_{sf}	1 ms
Frame time	T_f	10 ms
Cyclic Prefix in Samples	N_{CP}	144
# UEs	K	16
# BS antennas	M	128

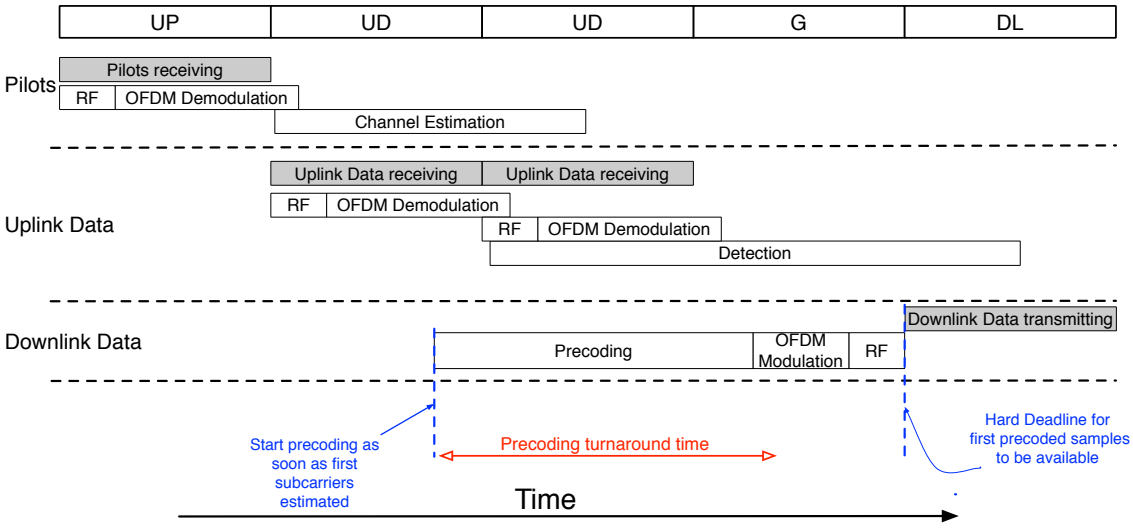


Figure 4.5: Simplified timing diagram for MaMi to point out some major challenges.

the other boxes show processing blocks. First the K users transmit pilots, orthogonal due to the usage of different subcarriers per user (or orthogonal pilot sequences that span multiple subcarriers). The Radio Frequency (RF) front end introduces some delay before data is fed to OFDM demodulation. Data is received with the sampling rate f_s which is much less than the actual clock rate of the OFDM processing blocks. As soon as all pilots are received the first subcarriers will be available for channel estimation while received UL data samples propagate through RF and OFDM Demodulation blocks for the consecutive UL data symbols. Detection of UL data is typically not time critical and could be buffered until Hardware (HW) resources are available, however, buffering increases requirements on memory on the processing platform.

The critical part to point out is the precoding turnaround time as samples need to be available at the RF at a hard deadline, shown in blue in Figure 4.5. This puts a hard constraint on the precoding of the user symbols; therefore, precoding has to be initiated as early as possible when first channel estimates are available.

The guard time allows the RF chains to switch from receive to transmit and vice versa and also gives time to flush the Fast Fourier Transform (FFT)/Inverse Fast Fourier Transform (IFFT) blocks so that same hardware resources are usable here (guard time might be decreased if hardware supports faster switching). If the time to switch the RF front ends is less than the cyclic prefix, guard interval might be removed to allow for higher data rates. However, due to

Table 4.5: Number of complex multiplications for MaMi system.

Block	# used per subframe	Algorithm	Multiplication amount
FFT	3	butterfly	$M \cdot N_{FFT} \cdot \log_2(N_{FFT})$
Channel Estimation	1		$M \cdot N_{SUB}$
Channel Interpolation	1	lin. interp.	$2M \cdot N_{SUB}$
Detection Matrix	1	Zero-forcing	$2 \cdot N_{SUB} \cdot K^2 \cdot M + 2 \cdot K^3$
	1	MRC	$N_{SUB} \cdot K \cdot (M + 5)$
Data Detection	2		$N_{SUB} \cdot K \cdot M$
Reciprocity Cal.	1		$N_{SUB} \cdot K \cdot M$
Precoding Matrix	1	Zero-forcing	Uplink estimate is used
	1	MRT	Normalization done in MRC
Data Precoding	2		$N_{SUB} \cdot K \cdot M$
IFFT	2	butterfly	$M \cdot N_{FFT} \cdot \log_2(N_{FFT})$
In total MR			$\approx 2.44 \cdot 10^7$
In total ZF			$\approx 1,06 \cdot 10^8$

the inherent latency of the FFT/IFFT implementation, FFT and IFFT blocks would have to be both implemented, i.e., sharing the same hardware is not possible. Furthermore, removing of the guard symbol increases overall bandwidth and timing requirements of the system.

Using OFDM multi-carrier modulation makes it possible to remedy actual latency requirements. As all subcarriers (after carrier-frequency offset correction) are orthogonal, overall bandwidth can be splitted in bandwidth chunks processed in parallel in the hardware.

Moreover, note that a frame structure having UL pilots and directly afterwards DL data is not possible, as there is no CSI to actually precode the data.

4.2.2 Core processing elements

To cope with the high operation count, existing architectural solutions utilize the potential of parallelizing the operations. This is usually done by employing the concept of Very-large Instruction Word (VLIW) or Single-Instruction Multiple-Date (SIMD) architectures. Both of these architectures aim to exploit the high data level parallelism. While a SIMD architecture allows multiple processing elements to perform the same operation on multiple data points concurrently, a VLIW architecture allows different operations to be executed simultaneously. This may increase processing throughput, i.e., the number of instructions executed over a period of time. For even further improvement, combination of these two architectures are possible.

As most of the operations in the digital baseband processing domain are operations on matrices and vectors, another potential candidate, from the architectural perspective, is to exploit Graphical Processing Units (GPUs) [40]. Exploiting GPUs in baseband processing is a relatively new approach, and recent research shows the benefits of utilizing GPU. Important consideration when addressing the challenge of high operation count is to utilize hardware accelerators as aforementioned, possibly in parallel with use of GPUs. Some heavy computational blocks can be extracted and implemented as hardware accelerators.

The number of operations scales linearly with to the number of antennas. Assuming, that complexity of addition is negligible compared to the complex multiplications, especially since those may be efficiently implemented in Multiply-Accumulate (MAC), Table 4.5 lists the required number of multiplications for latency critical blocks in MaMi baseband processing. Basically, the number of complex multiplications is defined by the matrix and vector operations required. For matrix inversion, an approximate inverse using Neumann-series was assumed [34].

Using the system parameter given in Table 4.4, the overall count for MF and ZF MaMi system is also given. Expanding these numbers to complex multiplications per second results in a range of $44 \cdot 10^9 \frac{muls}{sec}$ and $190 \cdot 10^9 \frac{muls}{sec}$ for MF and ZF, respectively. These number are based on the example frame structure which is also employed in the LuMaMi testbed. Since all of these operations are vector and matrix operations the number of complex additions will be within the same range.

Calculating the ratio of both, it is found that ZF is about four times as complex as MF, however this is only true if the number of users is kept constant. Figure 4.6 shows the ratio of both multiplication counts for different number of users as well as BS antennas. While changing the number of BS antennas keeps this ratio constant, it grows with increasing number of users; thus, for many user scenarios usage of MF algorithm could reduce the operations needed and thereby also energy consumption enormous compared to ZF. Furthermore, for many user cases, throughput of ZF will be higher compared to MF due to interference. ZF and MF are both linear

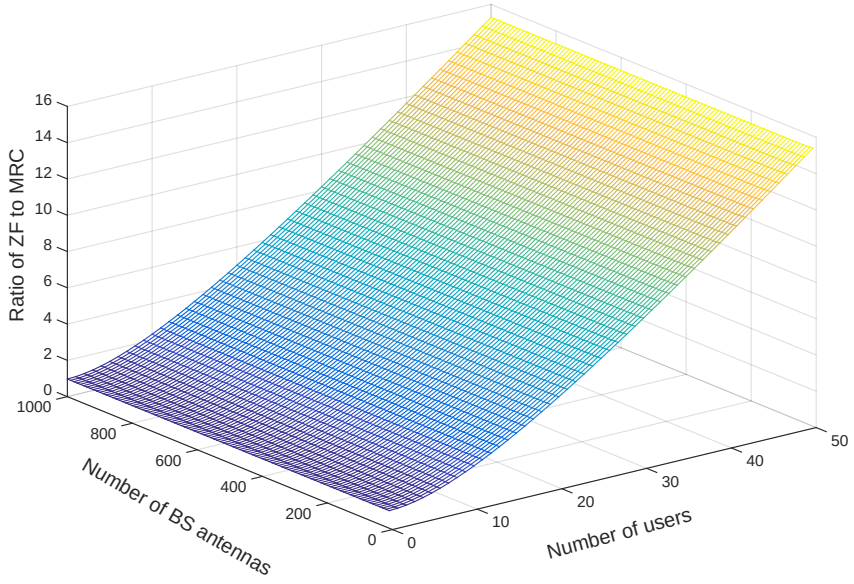


Figure 4.6: Multiplication Count Ratio of ZF to MF.

detector/precoding schemes and if certain users show high channel correlation (e.g., for line-of-sight users with similar angles and distances to the array), non-linear interference cancellation might be required to actually separate them.

4.2.3 Data movement bandwidth and data storage requirement

The fact that many operations are needed, also puts high requirements on the on-chip communication as well as memory bandwidth. Considering the high-level dataflow for MaMi given in Figure 1.1 an approximation for the overall bandwidth requirements in the system may be calculated.

Figure 4.7 shows a high-level overview of the different MaMi blocks and the required communication among them. Starting from the antenna side, as the number of BS antennas scales, so does the combined number of samples to and from all antennas which is $5 \cdot M \cdot (N_{FFT} + N_{CP})$ per slot (UL pilot + 2*UL data + 2*DL data, see Figure 4.5). Samples to and from OFDM are $5 \cdot M \cdot N_{SUB}$ and will be written into an input/output buffer.

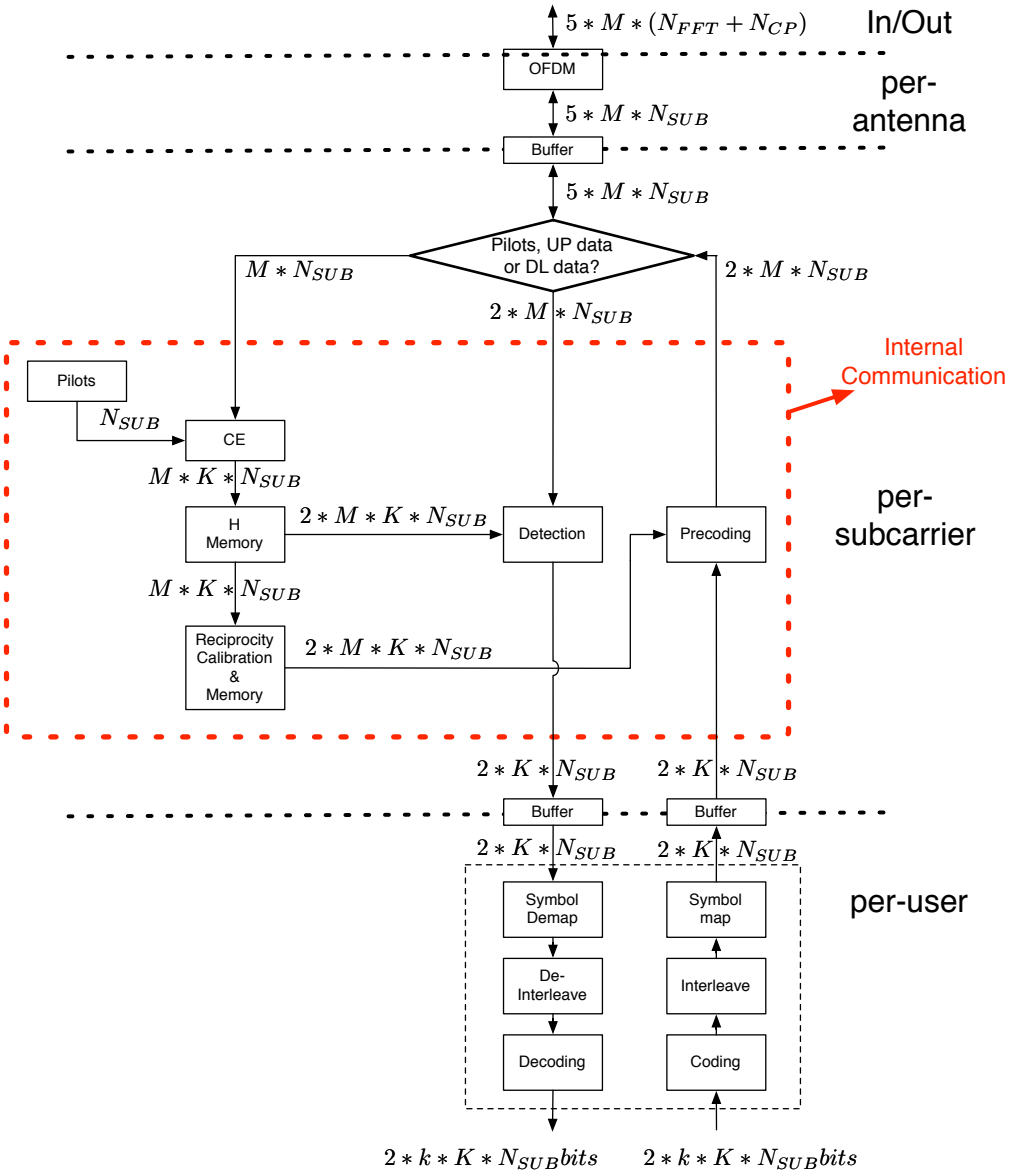


Figure 4.7: Number of samples interchanged between different blocks in a MaMi system for frame structure used in the LuMaMi testbed.

Consequently, data is processed either upstream or downstream depending on if it is UL pilots, UL data or DL data. First, for UL pilots, channel estimation is performed which requires to read the received vector from input buffer and the corresponding pilot leading to $M \cdot N_{SUB} + N_{SUB}$ and to write the resulting (to simplify we assume that interpolation is performed) channel estimates back into a memory, giving $M \cdot K \cdot N_{SUB}$ write operations. Second, two UL data symbols are received and processed in the MIMO detection block requiring to read received vector and the CSI which corresponds to $2 \cdot M \cdot N_{SUB} + 2 \cdot M \cdot K \cdot N_{SUB}$ accesses. The results are $2 \cdot K \cdot N_{SUB}$ samples stored for further processing (for ZF more accesses might be required, but those will be handled internally in the processing block and not put pressure on the system bus). Lastly, two DL data symbols are processed reading in $M \cdot K \cdot N_{SUB}$ estimated CSI samples, perform reciprocity calibration and save them in a memory. $2 \cdot K \cdot N_{SUB}$ information symbols and $2 \cdot M \cdot K \cdot N_{SUB}$ reciprocity calibrated CSI samples are read and mapped to $2 \cdot M \cdot N_{SUB}$ transmit samples to be sent to the OFDM modulation blocks.

After detection or before precoding, respectively, data is processed per user, i.e., processing is performed on the bit level, e.g. channel encoding and decoding. These blocks are most efficiently implemented as hardware accelerators; thus, assume data is read at the symbol demap block and then internally processed through de-interleaving and channel decoding. On the DL side, reversed processing is performed. The worst case scenario for bandwidth at the input and outputs of these blocks is a high code rate (≈ 1) and a high-order Quadrature-Amplitude Modulation (QAM) constellation. Assuming 64-QAM and a code rate 1, the input and output rate of the symbol demap and symbol map blocks, respectively, are $2 \cdot K \cdot N_{SUB}$ symbols which are buffered for further processing in each direction. After decoding or before coding, these correspond to $2 \cdot k \cdot K \cdot N_{SUB}$ information bits, where k is the number of bits for the QAM constellation, for instance $k = 6$ for 64-QAM.

Note that samples within the dashed box in per-subcarrier processing are communicated internally in the core processing units and therefore do not add to the bandwidth required to interconnect the different blocks.

Adding up all these numbers and dividing it by the length of seven OFDM symbols (0.5 ms), the estimated overall data rate on the system is given in Table 4.6 for 8-bit and 16-bit wordlength per I and Q sample. Core processing bandwidths are also added for completeness

Table 4.6: Estimated BW requirements for MaMi processing

Stage	Rate [GB/s]	
	8-bit	16-bit
In/Out	3.6	7.21
Per-antenna	1.98	3.95
Per-subcarrier	2.17	4.35
Per-user	0.086	0.086
Total	**7.84**	**15.6**
Core processing	37.92	75.84

but do not include rates required for storing and loading of intermediate results, e.g. when calculating pseudo-inverse matrix for ZF. These data rates are not tremendously high, but one has to keep in mind that all data has to be collected, communicated and transferred among hundred or more different HW processing blocks in the system.

Using Figure 4.7 a minimum memory requirement for the overall system can also be roughly estimated. Neglecting any possible optimization which might be possible and assuming that each of the buffer should be able to hold at least the overall samples of one OFDM symbol, overall memory necessary is $5.26 \cdot 10^6$ samples or 10.5 MB and 21 MB for 8-bit and 16-bit per I and Q sample, respectively.

4.2.4 Core memory

Memory requirement analysis

The algorithm properties will be used as guideline for system level and detailed core architecture design. The specific requirement of on-chip memory subsystem is evaluated based on the known overall structure of MaMi application.

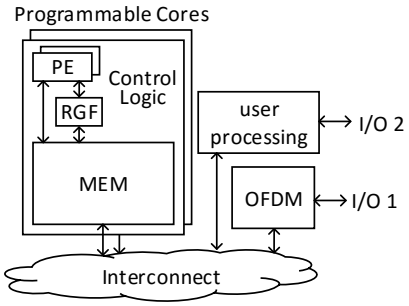


Figure 4.8: Illustration of On-chip subsystem for MaMi.

To assist analysis, the data flow using MMSE processing is preliminary categorized to kernel operations and further decomposed to atomic operations. Kernel operations are coarse-grained operations derived from algorithms, as shown in Table 4.7. The only difference between MMSE and ZF is on Kernel OP.II. The digital baseband processing part of MaMi processing is dominated by matrix-matrix/matrix-vector operation and consists of a large amount of Data Level Parallelism (DLP). In order to efficiently utilize processing elements(PE) in a SIMD style processor, the on-chip memory subsystem has to provide data to PEs in a way to maximize the executed data and thereby the performance of the overall system. Atom operations are vector level operations that regarded as instructions in hardware and executed in one clock cycle using SIMD technique. Most atom operations are either vector multiplication or vector addition. For example, the Gram matrix multiplication $\mathbf{H}^H\mathbf{H}$ in Kernel OP.I is decomposed into $\frac{M(M+1)}{2}$ K-length vector multiplications and further decomposed into several segments as atom operations with the SIMD width of the underlying hardware platform.

The on-chip memory subsystem, shown in Figure 4.8, exchanges data with external modules through Network-on-Chip (NoC) and supplies operands for the PEs. The register file (RGF) can substantially improve the access time and reduce the amount of memory access by storing intermediate results. However, the area consumption of memories storing the entire CSI matrix is unbearable. To assist memory bandwidth evaluation, we made two assumptions.

- infinite RGF size that requires memory access only at input or output of a kernel operation. (best case)
- limited RGF size, only scalar elements are stored in the RGF which requires memory access for each atomic operation. (worst case)

The vector-wise data storage and retrieval breakdown in these two cases is plotted in Figure 4.9. The bandwidth requirement for memory is about 300 GB/s in worst case for whole system, including read and write operation. As can be seen, the detection part dominates in the data storage and retrieval and occupies 64% and 53% in best case and worst case, separately. It is worth emphasizing that the data access of pre-processing is shrunk by a factor of 16 for duplicating channel information in 16 neighbor sub-carriers. The data access of pre-processing grows dramatically when the register file size is limited. That is effected by matrix-inverse in pre-processing. Matrix-inverse operation is a complex operation and generates large sum of intermediate results.

The operands and results of SIMD operations will be vectors and is part of operation matrices. The location of these vectors among input or output matrices determines the access pattern requirements. As shown in Figure 4.10, there exist different access modes towards same data

Table 4.7: Detailing UL and DL of BaseBand Processing

Stages	No.	Kernel Operations	Details
Pre-processing	I	$\mathbf{H}^H \cdot \mathbf{H}$	Matrix Mul
	II	$\mathbf{H}^H\mathbf{H}+\alpha\mathbf{I}$	Vector Add
	III	$(\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I})^{-1}$	Matrix Inv*
Detection	IV	$\mathbf{H}^H \cdot \mathbf{y}$	Matrix-Vector Mul
	V	$(\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I})^{-1} \cdot \mathbf{H}^H\mathbf{y}$	Matrix-Vector Mul
Precoding	VI	$(\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I})^{-1} \cdot \mathbf{x}$	Matrix-Vector Mul
	VII	$\mathbf{H} \cdot (\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I})^{-1}\mathbf{x}$	Matrix-Vector Mul

*Using Neumann series approximation.

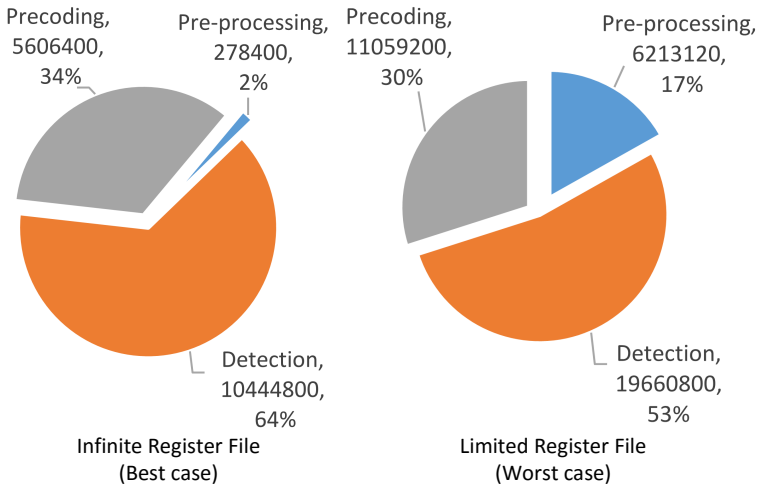


Figure 4.9: Vector-wise data storage and retrieval breakdown of the MaMi application in 1 subframe, using LTE parameter, 1200 sub-carriers, H size 128x16, SIMD width 16.

area. $\mathbf{H}^H\mathbf{H}$, $\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I}$, and $(\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I})^{-1}$ are sharing same part of memory as there is no operation requiring any two of them as input at same time and is referred to as General Gram Matrix (GGM). For $\mathbf{H}$, column-wise access is required for matrix multiplication from the right and vice versa. For GGM, diagonal access in $\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I}$, row-wise access and column-wise access in matrix inverse are both necessary.

The number of users is changing due to different standards and algorithms. That means operand matrix is scaling during run-time or pre-determined due to software. Despite this, a scaling operand matrix might not match the hardware and requires memory sub-system to adapt in a reasonable range.

Potential methods

1. **Dedicated Register Array**: Dedicated register array has the capability to access data with any access pattern. Besides that, register files have the advantage of fast storage and retrieval, usually within one clock cycle. To fully benefit from dedicated register files, the entire matrix or a matrix block (for example 16×16) should be stored within the register

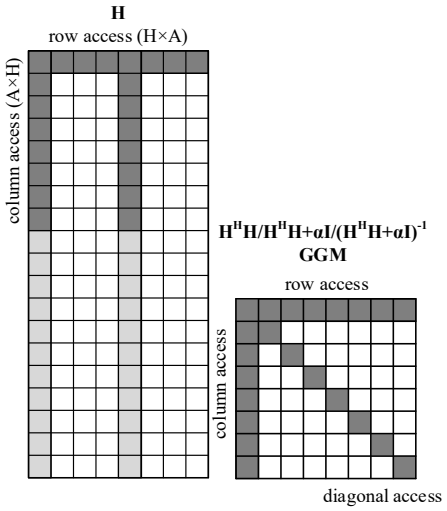


Figure 4.10: Examples of Access Pattern in MaMi Application.

file. Thus the capacity of register file must exceed the size of operand matrix, however, the area consumption of a matrix register file is prohibitive.

2. **Matrix Rearrangement Accelerator**: The matrix transposition or rearrangement accelerator attached to the MEM unit is easily programmed through software with one instruction. Most of the accelerators aimed for matrix transposition lack of flexible access pattern. The rearrangement using these accelerators requires fetching entire operand matrix and re-write into core memory, which leads to high latency and increased memory bandwidth.

3. **Parallel Memory (PLM)**: Memory banks are conjuncted together in PLM architecture and execute storage and retrieval instruction simultaneously in parallel to obtain higher throughput and flexibility. Through dedicated data allocation scheme, matrix entries are distributed over different memory banks to avoid multiple data within one access range to appear in the same memory module. The PLM architecture is suitable for fixed access pattern applications and has a higher data density by using Static Random Access Memory (RAM) (SRAM) or Embedded Dynamic RAM (DRAM) (eDRAM) comparing with registers.

4. **Matrix Transposition During Transferring**: There are several procedures that can be utilized to decrease complexity. For example, during UL processing, data is organized by antenna in OFDM demodulation and in subcarriers during channel estimation and detection. Thus, there will be a matrix transpose converting a $M \times N_{FFT}$ matrix to a $N_{SUB} \times M$ matrix in every OFDM symbol. The output of FFT is a sequence in the order of sub-carriers. Collecting the output of different antennas and organizing and shuffling them will be an efficient method.

4.2.5 On-chip communication

Well designed on-chip communication network is substantial for successful implementation of a MaMi processor and to ensure low latency communication between all different processor cores, accelerators and reconfigurable blocks. In a MaMi system, the number of different HW blocks

to be connected may go up to hundreds, introducing numerous challenges to be addressed. On-chip communication targets to maximize throughput, minimize latency while keeping energy-efficiency high. Moreover, the overall power envelope for the chip puts constraints on possible implementations. Lastly, nano-scale technology forces advanced circuit-design techniques as dynamic voltage scaling (DVS) and clock-gating to not exceed maximum instantaneous power limits. Asynchronous connections for global wires might be necessary, as wires do not scale likewise than transistors, which may cause signals traveling over the entire chip having a delay of several clock cycles.

Requirements

In a MaMi Baseband Processor system several key features must be targeted by the interconnection network:

Heterogeneity: Homogeneous networks will not be able to fulfill processing requirements in MaMi. Heterogeneous networks incorporating a bigger control processor, smaller Reduced Instruction Set Computer (RISC) processor cores, dedicated accelerators and reconfigurable processing elements are indispensable.

Hard Deadlines: The used frame schedule used for the wireless networks, puts hard deadlines onto the processing scheduling, for example, precoding has to be performed within the same slot as otherwise channel estimates would become invalid for changing channels.

Fast Reconfigurability: Low turnaround time for reconfiguration of PEs in changing conditions, for example, change of locked-on users or different detector/precoder schemes for changing Signal-to-Noise ratio (SNR) regions. This requires dedicated, low-latency links from the main controller to configuration interfaces of the reconfigurable blocks.

Scalability: Ability of interconnecting arbitrary number of accelerators, PE and processors (within a certain range) without penalty in overall performance, for instance, latency, to ensure functionality with arbitrary antenna and user configurations.

On-chip communication candidates

Many processors optimized for MIMO baseband processing nowadays are using standard on-chip buses. As buses are quite simple to implement they might be also a viable option for MaMi baseband processing. Figure 4.11 shows how such a system could look like. The four main blocks are: the Bus Interface (BI) including buffers, per-antenna processing (PAP), per-subcarrier processing (PSP) and per-user processing (PUP) (see also Figure 1.1). Using a RISC processor to configure a Direct Memory Access (DMA) controller responsible for transfer among the three different types of blocks. For the PSP part, we assume that the overall subcarrier $N_{SUB} = 1200$ are split and fed into $N_{core} = 8$ different PSP blocks. It is worth mentioning that this bus with our system parameters would count 153 nodes connected. To keep the number of arbitrations within a reasonable limit each PAP that is arbitrated sends its demodulated antenna data for $\frac{N_{SUB}}{N_{core}} = 150$ subcarriers into the Matrix Memory of the respective core. PSP blocks start calculating as soon as enough data is available. For UL pilot that is data of one antenna and several subcarriers and for UL and DL data when data for all antennas or users is available for one subcarrier, respectively. Writing data from OFDM modulation and reading

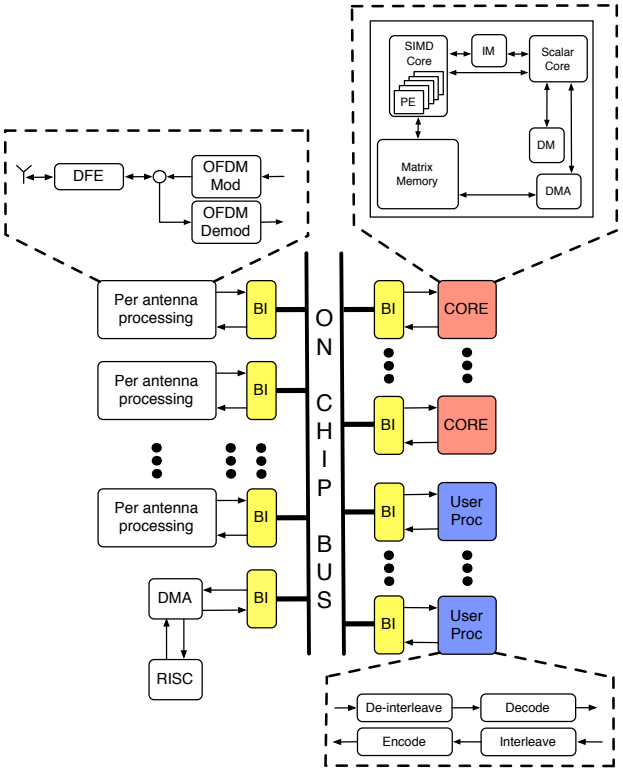


Figure 4.11: MaMi system using single BUS as interconnection network.

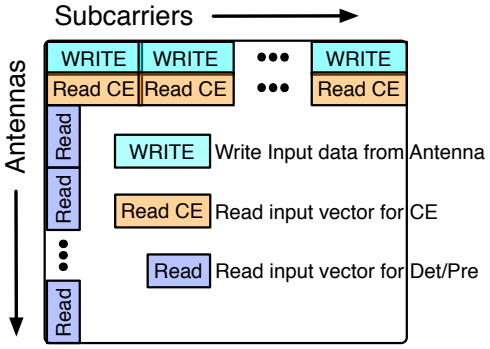


Figure 4.12: Memory access pattern for memories inside the PSP blocks.

data for channel estimation has to be performed row-wise from the matrix memory whereas reading for UL pilot detection has to be performed column-wise as shown in Figure 4.12. For DL data, write and read access will be reversed. Analyzing the communication time assuming burst transfers for all 150 subcarriers per PAP, a bus-width of 128-bit and a penalty of 2 clock cycles per arbitration, overall communication time for different I/Q-data wordlengths and clock frequencies is given in Table 4.8. Overall time for one slot in the frame structure is 0.5 ms; thus only one configuration, i.e., 500 MHz at 16-bit sample size seems to be really tight. Notice, that different cores have to share information about some of the subcarriers among each other to be able to do interpolation which is not included in this analysis.

Although a rough timing analysis may suggest that a single bus would work and buses are usually quite straightforward to implement, there are certain drawbacks. (1) Many buses do not support such a high number of nodes, (2) Scalability to higher number of BS antennas

Table 4.8: Overall communication time for single BUS MaMi system

Clock Frequency	I/Q sample wordlength	
	8-bit	16-bit
500 MHz	$\approx 240\mu s$	$\approx 480\mu s$
1 GHz	$\approx 120\mu s$	$\approx 240\mu s$

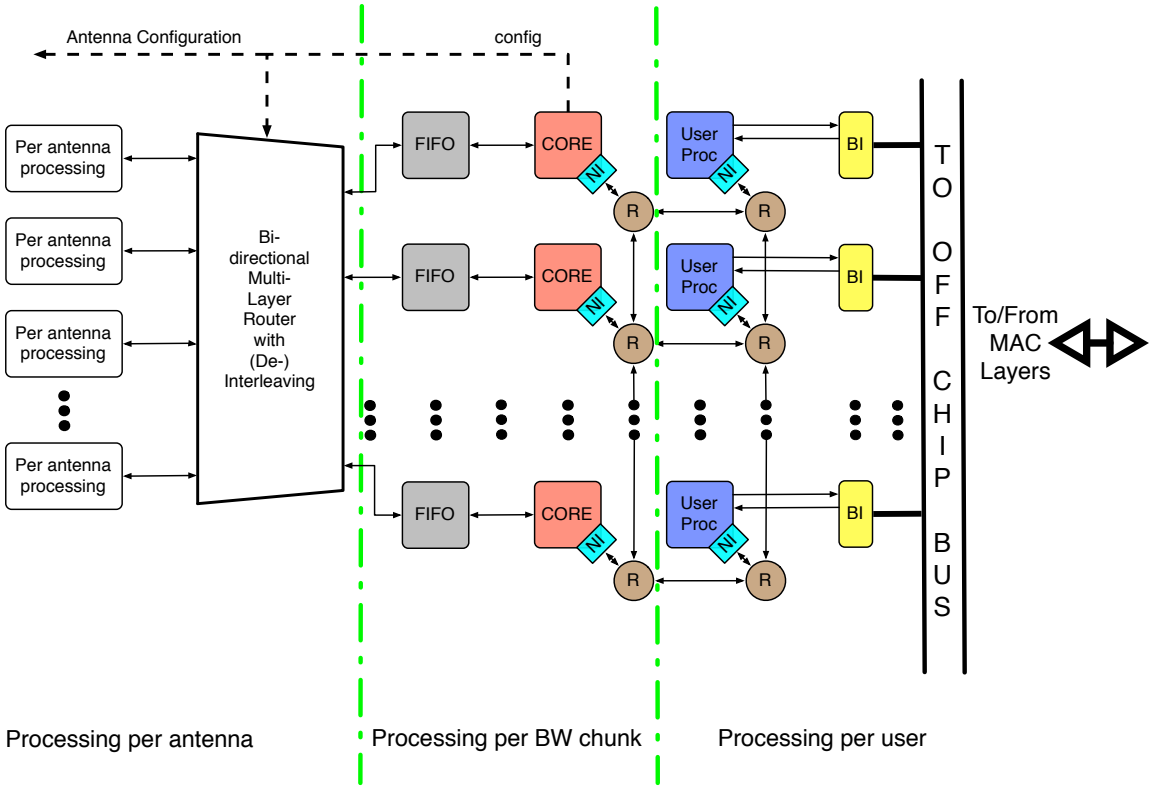


Figure 4.13: MaMi system using a multi-level router and a NoC for efficient communication

and users is quite limited, (3) channel estimation interpolation requires communication among PSP cores which will force them to be masters in many bus architectures and (4) Bus access is exclusive and scheduling might become hard even though overall timing is met.

To remedy the problems coming with a single bus connecting all processing blocks, we propose a hybrid interconnection scheme as shown in Figure 4.13. The PAP blocks are connected to a configurable bi-directional multi-layer router which also aggregates samples over several subcarrier or antennas, respectively. For UP transmission, the scheme shown in the left side of Figure 4.14 is used. One antenna per time with several subcarriers aggregated is written into the FIFO. The example shows how this pattern would work when using two FIFOs and therefore 300 subcarriers. It is advantageous, since channel estimation has to be performed on a per-antenna basis and as users are orthogonalized in the frequency domain interpolation techniques have to be employed. Moreover, this order allows for the fastest calculation of CSI, thereby tackling efficiently the precoding turnaround time, however, it increases buffer size in the routing block.

During UD, it is most beneficial to fill the FIFOs on a per-subcarrier basis, i.e., write the data of all antennas for one subcarrier for each core and FIFO and iterate through this pattern

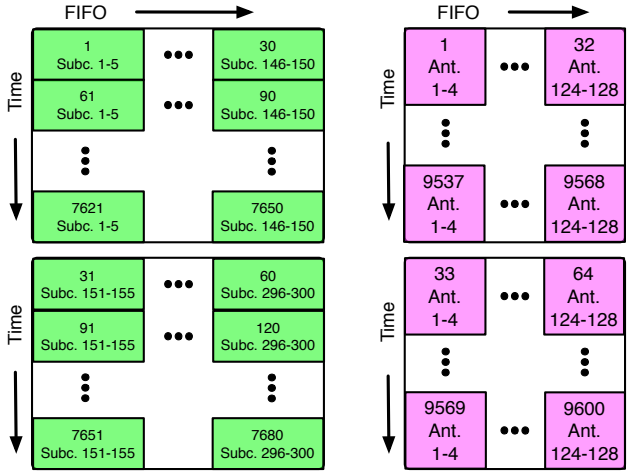


Figure 4.14: FIFO write order pattern for UP (left) and UD (right)

until all data has been processed as shown in right side of Figure 4.14.

One PSP is selected to be Master and controls the configuration of the router and RF-chains so run-time antenna selection is possible to reduce overall power consumption. PSP and PUP are connected through a circuit-switched (CS) NoC using a static Time Division Multiplexing (TDM) scheme with free reserved slots to capture non-deterministic traffic. Non-deterministic traffic occurs (1) due to different selected channel estimation algorithms and interpolation techniques which may require more or less adjacent subcarrier information or (2) through usage of iterative detection and precoding algorithms requiring for instance feedback from decoder to the detection. A static CS NoC with TDM scheme allows to properly analyze and schedule all the traffic offline and guarantees bandwidths among processing cores and accelerators which is essential in a system having hard deadlines.

Interface between software implemented Medium-Access Control Layer (MACL) protocols is handled by a bus as data rates are not too high. As information bits to be sent and received always have to go through the PUPs, direct connection of these with the bus is proposed in our scheme with the master of the bus being the MACL software environment.

For further development of the interconnection network several topics have to be analyzed. (1) A proper wordlength analysis has to be performed to find optimal wordlength for I/Q samples and the interal processing blocks. Additionally, usage of a floating point unit due to high dynamic range has to be evaluated, (2) maximum number of supported BS antennas and users for the system has to be specified and evaluated and (3) different algorithms have to be profiled to map a TDM schedule which can handle the required traffic patterns.

Chapter 5

Hardware implementation of baseband processing

This chapter presents the hardware implementation of MaMi baseband processing, mainly using Applicatin Specific Integarted Circuit (ASIC) design methodology for key processing accelerators. Special focus will be on the tradeoff between digital processing complexity and analog transceivers quality. With a massive number of antennas, low-cost RF chains are needed to reduce the overall cost. In MaMi systems, most of the hardware impairments are shown to cause an additive distortion that is substantially uncorrelated with the desired signals and, hence, vanish asymptotically with an increasing number of antennas. As result, it is expected that much lower hardware precision can be adopted in MaMi systems than in traditional systems [5,12]. Nevertheless, for a practical massive MIMO system with a limited number of antennas, effects of hardware impairments like IQ imbalance will not completely disappear. Also, highly linear PAs are inefficient and consume more power than those with lower requirements on linearity. It is therefore of interest to reduce the PAPR of transmitted signals to be able to use more efficient PA without causing in-band and out-of-band distortions.

In this chapter two approaches of tackling PAR are compared, i.e., a single-carrier discrete-time Constant Envelope (CE) modulation and an OFDM-based antenna reservation technique. The CE precoding has stringent constraints on amplitude and utilizes the high degree-of-freedom available in massive MIMO systems to provide almost 0 dB PAR in the discrete-time domain. Conversion to continuous-time will increase the PAR, but leave it at a tolerable level. The antenna reservation technique is based on ZF and OFDM modulation, and adds a 15% complexity overhead. Also, the effects of IQ imbalance in massive MIMO and its pre-compensation are described. We analyze various processing schemes and implement (synthesis and/or layout) them using state-of-the-art CMOS technology. The circuit architectures for efficient implementation are described. Moreover, the required processing energy per transmitted information bit is simulated on gate-level. The results show that the energy cost of performing precoding and tackling of hardware impairments are low.

5.1 System model

Let M be the number of antennas at the BS and K the number of single antenna users. The channel matrix to all users at the n-th tone is denoted as $\mathbf{H}_n \in \mathbb{C}^{K\times M}$, and the subscript is dropped when a single-carrier system is considered. Let $\mathbf{x}_n = [x_{1,n}, x_{2,n}, ..., x_{M,n}]^{\mathrm{T}}$ denote the transmitted vector from the M BS antennas, which is normalized to satisfy $\mathbb{E}[\mathbf{x}_n^{\mathrm{H}}\mathbf{x}_n] = 1$, and $()^{\mathrm{H}}$ is the Hermitian transpose. The overall symbol vector received by the K autonomous users

is

$$\mathbf{y}_n = \sqrt{\frac{P_{\mathrm{T}}}{M}}\mathbf{H}_n\mathbf{x}_n + \mathbf{w}_n\,, \tag{5.1}$$

where P_{T} is the total transmit power, and $\mathbf{w}_n$ is a $K\times 1$ vector i.i.d complex Gaussian variables with variance $\sigma^2\mathbf{I}_{K\times K}$.

To fully exploit a large antenna array, the user symbols/information at the BS needs to be translated or mapped to correct signals in the antennas, so that each user receives the information with low (zero) interference from signals intended for other users. For linear precoding schemes, this mapping is expressed as

$$\mathbf{x}_n = \mathbf{F}_n\mathbf{s}_n\,, \tag{5.2}$$

where $\mathbf{s}_n$ is a $K\times 1$ vector containing the symbols intended for the K users on n-th tone, and $\mathbf{F}_n$ is the $M\times K$ precoding matrix mapping user symbols to antenna signals $\mathbf{x}_n$. Two well known linear precoding schemes in massive MIMO are, MR and ZF, with $\mathbf{F}_{\mathrm{MR}} \propto \mathbf{H}^{\mathrm{H}}$ and $\mathbf{F}_{\mathrm{ZF}} \propto \mathbf{H}^{\mathrm{H}}(\mathbf{H}\mathbf{H}^{\mathrm{H}})^{-1}$, respectively [34]. The ZF precoder is basically a constrained least-squares solution for an under-determined system, i.e., ZF cancels all inter-user interference with least transmit energy ($\min ||\mathbf{x}||_2,$ subject to $\mathbf{s} = \mathbf{H}\mathbf{x}$).

5.2 QRD based ZF precoder

Several methods can be used to realize low-complexity ZF operation by leveraging the unique feature of MaMi channel matrix. In Deliverable 3.1 [27] we introduced the Neumann series based approximations. Here, we describe another efficient way of matrix inversion using approximative QR-decomposition. The processing of the QR based ZF precoder is split into four parts, namely, matrix multiplication ($\mathbf{H}^H$ same as in MR), generation of a Gram matrix ($\mathbf{H}\mathbf{H}^{\mathrm{H}}$), performing a QR-decomposition, and applying the corresponding solver. The Hermitian matrix multiplication is processed per-antenna, and each instance implements a simple vector-dot-product based on MAC units. For the Gram matrix generation the computational complexity is $\mathcal{O}(\frac{1}{2}MK^2)$, and implemented using a triangular systolic array.

There is a plethora of highly optimized QR decomposition implementations in traditional MIMO systems. Unfortunately, scaling-up these implementations for massive MIMO is quite expensive in terms of hardware. However, under favourable conditions and high ratios between number of antennas at BS and **MS!** (**MS!**) (β), $\mathbf{Z}$ becomes diagonally dominant. This property is also extensively used in [34] as an initial condition for Neumann series. In case of a QR decomposition the diagonal dominance eases the computations resulting in a complexity $\mathcal{O}(K^2(K-1)+3K)$, around 50% lower than for traditional QR algorithms.

After the QR decomposition, the user data is precoded by performing $\mathbf{R}^{-1}\mathbf{Q}^H$ implicitly. This computation is performed to reduce hardware cost, and also compared to an explicit computation requires lower latency. The hardware for the precoder is implemented in 28 nm FD-SOI (Fully Depleted Silicon On Insulator) technology, and in this chapter we use this technology as a reference for power consumption evaluation. The power consumption for performing the QR decomposition and running the solver are 29 mW and 26 mW, respectively.

5.3 PAPR aware precoding

Power amplifiers may contribute a large portion of the total power consumption in the BS. This is mainly due to the high linearity constraints on the Power Amplifier (PA) over a large dynamic

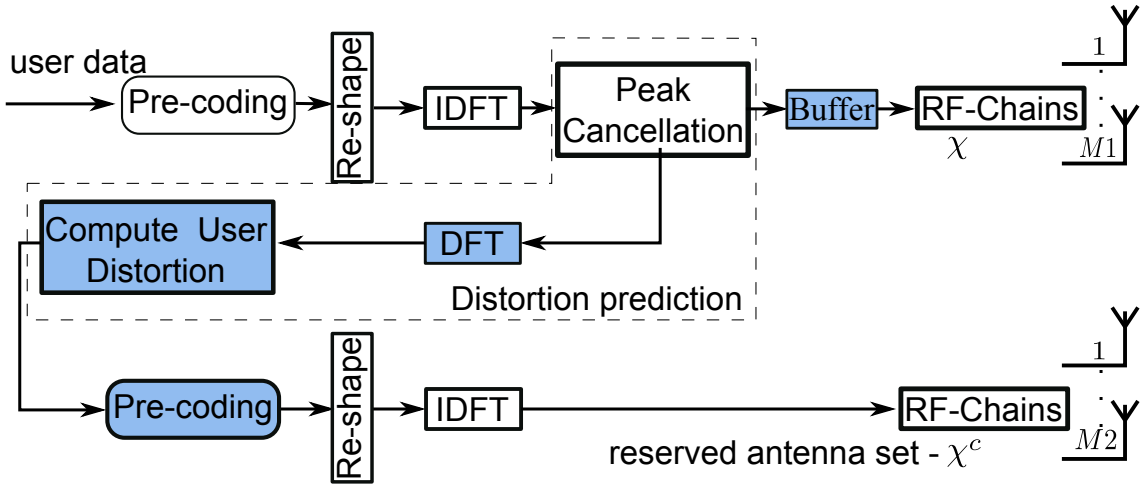


Figure 5.1: Data-flow illustration of the low complexity PAPR reduction approach, where the dedicated set of compensation antennas χ^c counteracts the clipping based distortion.

range, which translates into an inefficient operation. Non-linear PAs are highly efficient and employing them requires the PAPR of the transmitted signal to be low. In this section some digital signal processing approaches to lower PAPR of transmitted signal in MaMi systems is described. Firstly, the antenna reservation technique combined with ZF precoding for OFDM modulated MaMi systems is described. This is followed by narrow band discrete-time constant envelope precoder implementation.

5.3.1 Antenna reservation based on ZF

Clipping in the digital domain is a very simple technique to reduce PAPR, but suffers from in-band distortion. An approach to compensate this in-band (not including the guard-band) distortion is to dedicate a subset of the antennas which transmit signals used to mitigate the resulting distortion. This technique adheres with the availability of large number of antennas in massive MIMO, and is coined as "antennas reservation" similar to the "tone-reservation" in an OFDM system. Unlike reserving tones in OFDM, which lowers capacity linearly, reserving antennas reduces capacity logarithmically (due to the reduction of the antenna gain, hence SINR).

Figure 5.1 describes the top-level data flow, with additional modules (shaded) required to perform distortion mitigation. For a system with $M = 100$ antennas, where $M2 = 20$ are reserved, about 4 dB of PAPR improvement is achieved (at the cost of $10\log_{10}^{20}$-dB loss in SNR). The overhead in terms of baseband processing has been highlighted using dashed-line. The complexity overhead compared to a system without antenna reservation is about 15% [35].

5.3.2 Discrete-time constant envelope precoder

To employ a highly efficient non-linear PA, a very strict constraint on the amplitude of the transmitted signal is enforced, resulting in nearly 0 dB PAPR. This strict amplitude constraint downlink transmission scheme is known as "discrete-time constant envelope" and has been described in Deliverable 3.1. The information is carried on the phase and exploits the large degree of freedom available in massive MIMO to provide high sum-rates [29].

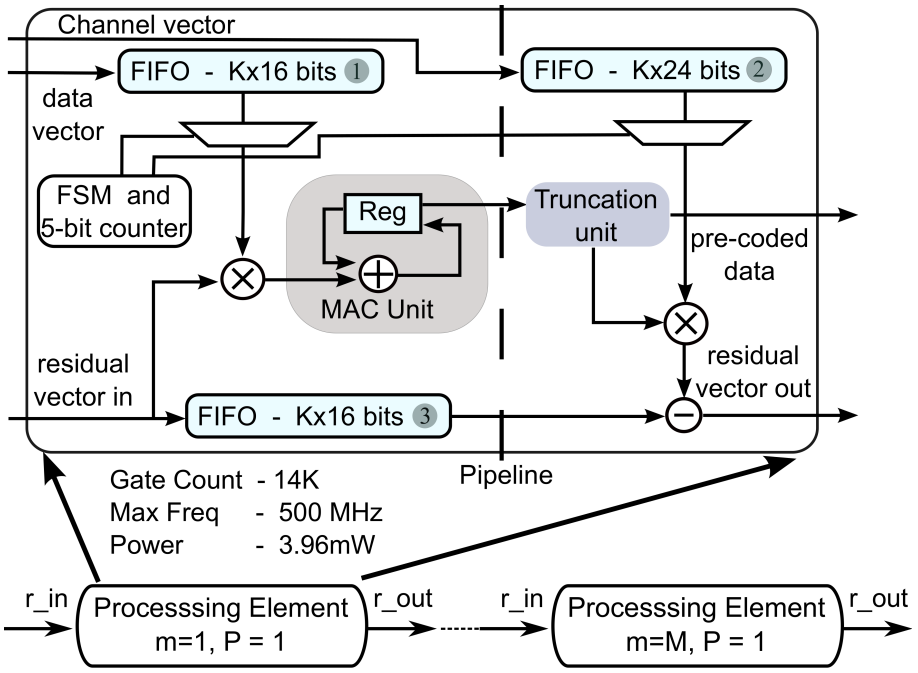


Figure 5.2: Systolic array for CE precoder based on coordinate-descent algorithm, where each processing element solves phase for an antenna.

The CE precoder can be viewed similar to ZF [36], i.e., suppression of inter-user interference, but with an additional constraint on the amplitude as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\,\mathbf{s} - \mathbf{H}\mathbf{x}||_2 \\ \text{subject to} \quad & |x_m|^2 = 1, \text{where } m = 1, \cdots M. \end{aligned} \tag{5.3}$$

The solution of (5.3) has multiple local-minima, but in a massive MIMO system, even the local minima tend to be close to optimal. To solve the CE precoder the coordinate-descent algorithm is employed, which is similar to gradient-descent, barring that the optimization is performed on one coordinate (variable) at a time. The complexity is $\mathcal{O}((9K+5)MP)$, where P is the number of iterations. It should be noted that this is valid for a single tap (narrow-band) channel. For wide-band channels we expect the complexity to scale linearly with the number of channel taps.

The proposed optimization is very suitable for a systolic array implementation, where each processing element computes the phase for an antenna see Figure 5.2. The processing element needs to store the channel vector of the corresponding antenna. After computing the phase the residual vector is streamed to the next processing element for computation. We have implemented the architecture in Figure 5.2 with Register Transfer Level (RTL) description and synthesized using 65nm CMOS technology. Each element takes 14.1 K gates (one gate corresponding to a 2-input NAND gate in the standard cell library) and the hardware cost scales linearly with the number of antennas and iterations.

5.4 IQ imbalance pre-compensation

Direct-conversion transceivers have an in-phase (I branch) and quadrature (Q-branch) which are passed through two mixers with a phase difference of 90°. IQ imbalance arises when there

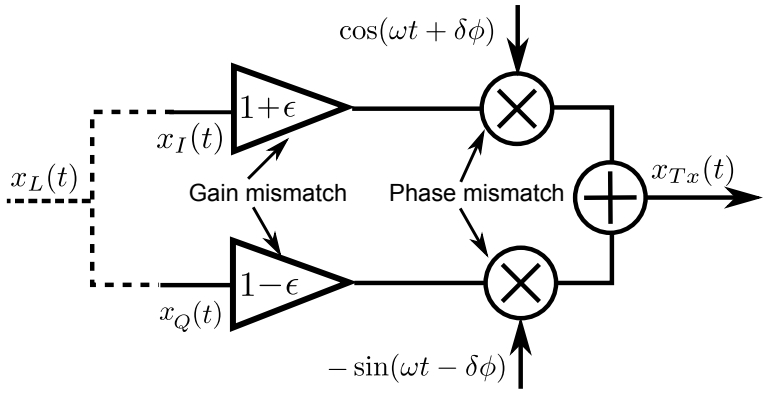


Figure 5.3: Transmitter IQ imbalance model, with ϵ and $\delta\phi$ the physical mismatch parameters, $x_L(t)$ time domain baseband IQ signal and $x_{Tx}(t)$ is transmitted signal.

is a mismatch in amplitude or phase between the mixers. This effect can be modeled by two parameters, i.e., ϵ amplitude and $\delta\phi$ phase mismatch, as shown in Figure 5.3. The effects and compensation of IQ imbalance are well studied [22]. In-line with these works, we define two variables, a and b, which are calculated from the physical parameters as

$$\begin{aligned} a &= \cos(\delta\phi) + j\epsilon\sin(\delta\phi) \\ b &= \epsilon\cos(\delta\phi) + j\sin(\delta\phi), \end{aligned} \tag{5.4}$$

where $a \to 1$ and $b \to 0$ with decreasing ϵ and $\delta\phi$. The signal received at a perfect receiver when there is frequency independent IQ imbalance at a transmitter, becomes

$$x_{\mathrm{Rx}}(t) = a x_{\mathrm{Tx}}(t) + b x^*_{\mathrm{Tx}}(t), \tag{5.5}$$

which, in the corresponding frequency domain is expressed as

$$X_{\mathrm{Rx}}(f) = a x_{\mathrm{Tx}}(f) + b x^*_{\mathrm{Tx}}(-f), \tag{5.6}$$

indicating a dual effect. There is both an attenuation of the correct signal and interference from a frequency mirrored copy of the signal.

Various studies on the effects of hardware impairments for massive MIMO systems were performed [22]. In the following section an analysis of IQ imbalance in the downlink is performed, which shows that there is a need for pre-compensation. Based on the result, an IQ imbalance pre-compensation circuitry is introduced and the corresponding hardware cost is evaluated.

5.4.1 Effects of IQ imbalance in massive MIMO

To evaluate the effects of IQ imbalance, we look at the Signal-to-noise-plus-distortion ratio (SNDR) at the user terminals

$$\mathrm{SNDR} = 10\log_{10}\left(\frac{P_{\mathrm{s}}}{P_{\mathrm{d}} + \sigma_w^2}\right), \tag{5.7}$$

where P_{s} is the signal strength, P_{d} is the distortion due to IQ imbalance at the transmitter and σ_w^2 is the additive noise variance at the receiver. For a fixed transmission power budget, signal power increases linearly with the number of antennas, due to the array gain. However, the IQ

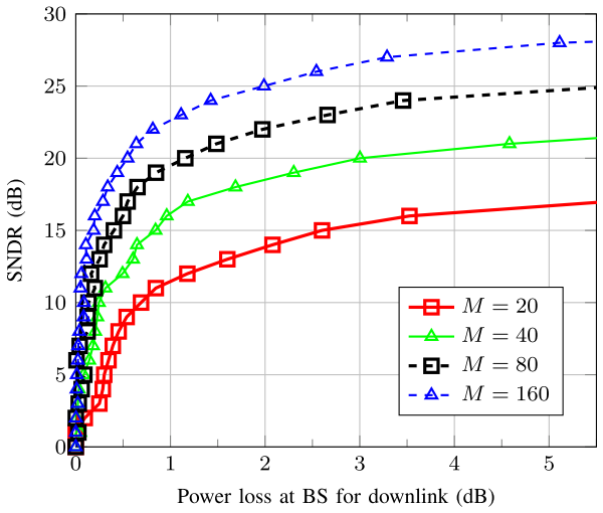


Figure 5.4: Simulated IQ imbalance for $K = 10$ users massive MIMO system with 6% amplitude and 6° degree phase mismatch.

distortion increases at a much slower rate, mainly due the fact that the phase of distortion is negated (x^*_{Tx} in (5.5)), and rotated (multiplying by b). Hence, the IQ distortion is unlikely to add-up constructively at the receiver. This effect can be seen in Figure 5.4, where the horizontal axis is the loss in power compared to a system with no IQ imbalance. For a fixed configuration, the SNDR will saturate if the distortion dominates over noise, and further increasing transmission power has very little effect on the SNDR. One way to improve the SNDR is to increase the number of antennas, as seen in Figure 5.4. The improvement is, however, rather limited and digital pre-compensation may be a better option to limit this particular effect.

5.4.2 Pre-compensation architecture

Increasing the number of antennas is a robust approach to tackle IQ imbalance, since no knowledge of the IQ imbalance parameters is required. However, increasing the number of antennas only for this purpose may not be the most cost effective. In Figure 5.5 we show how the achieved SNDR of $M = 20$ antennas system increases with digital pre-compensation and different quality of the estimated IQ imbalance parameters. Drastic improvements are achieved for fairly low estimation accuracies and low-energy digital pre-compensation can be a very viable alternative.

The IQ imbalance pre-compensation is performed after precoding as shown in Figure 5.6. The main idea of pre-compensation is to transmit the signal w such that after the mixer with IQ imbalance the transmitted signal is the desired signal x. As described in (5.6), mirroring affects the n-th and $-n$-th tone, which needs to be considered during pre-compensation. We therefore group the two sets of linear equations, and express them in the real domain as

$$\begin{pmatrix} a_r^n & -a_i^n & b_r^{-n} & b_i^{-n} \\ a_i^n & a_r^n & b_i^{-n} & -b_r^{-n} \\ b_r^n & b_i^n & a_r^{-n} & -a_i^{-n} \\ b_i^n & -b_r^n & a_i^{-n} & a_r^{-n} \end{pmatrix} \begin{pmatrix} w_r^n \\ w_i^n \\ w_r^{-n} \\ w_i^{-n} \end{pmatrix} = \begin{pmatrix} x_r^n \\ x_i^n \\ x_r^{-n} \\ x_i^{-n} \end{pmatrix}, \tag{5.8}$$

where the subscripts r, i indicate real and imaginary parts of complex signals.

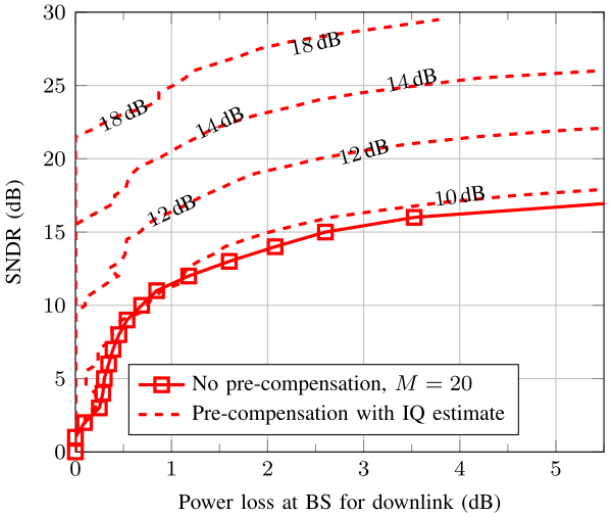


Figure 5.5: Pre-compensation for $M = 20$, $K = 10$ system, with different IQ imbalance estimation accuracy.

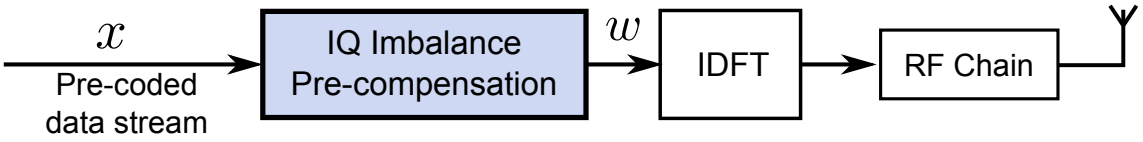


Figure 5.6: IQ imbalance pre-compensation top level data flow.

The pre-compensation scheme basically involves solving (5.8). One technique is to perform a brute force inversion and a matrix vector multiplication. However, since a and b are close to 1 and 0, respectively, an iterative method of solving linear equations is favorable. This approach is more hardware friendly and Figure 5.7 shows a Jacobi iterative approach.

To illustrate Figure 5.7, we define the 4×4 matrix in (5.8) as $\mathbf{A}$, the 4×1 vectors $\mathbf{w}$ and $\mathbf{x}$. The matrix $\mathbf{A}$ is split into two matrices $\mathbf{A} = \mathbf{D} + \mathbf{R}$, where $\mathbf{D}$ contains only diagonal elements of $\mathbf{A}$. The initial value of $\mathbf{w}$ is set with values of $\mathbf{x}$. The 12 multipliers in Figure 5.7 are used to perform matrix vector ($\mathbf{Rw}$) multiplications. The resulting vector is subtracted with input vector using 4 subtracters ($\mathbf{x} - \mathbf{Rw}$). The residual vector is then divided by the diagonal elements i.e., $\mathbf{D}^{-1}(\mathbf{x} - \mathbf{Rw})$. Division is performed when updating the estimates by using Newton-Raphson method and 4 multipliers. The hardware has a flexible iterative path, and the input vector can be loaded with the residual vector for the next iterations. For a low IQ mismatch parameters, the numerical accuracy of the solver is around 27 dB and 38 dB with just one and two iterations respectively. The pre-compensation was implemented in 28nm Fully Depleted Silicon On Insulator (FD-SOI) technology and the power simulations are performed on a gate level netlist with back annotated timing and toggle information. The corresponding hardware results are shown in Table 5.1. In the next section a comparison of all the aforementioned techniques to perform precoding and tackling of hardware impairment are compared.

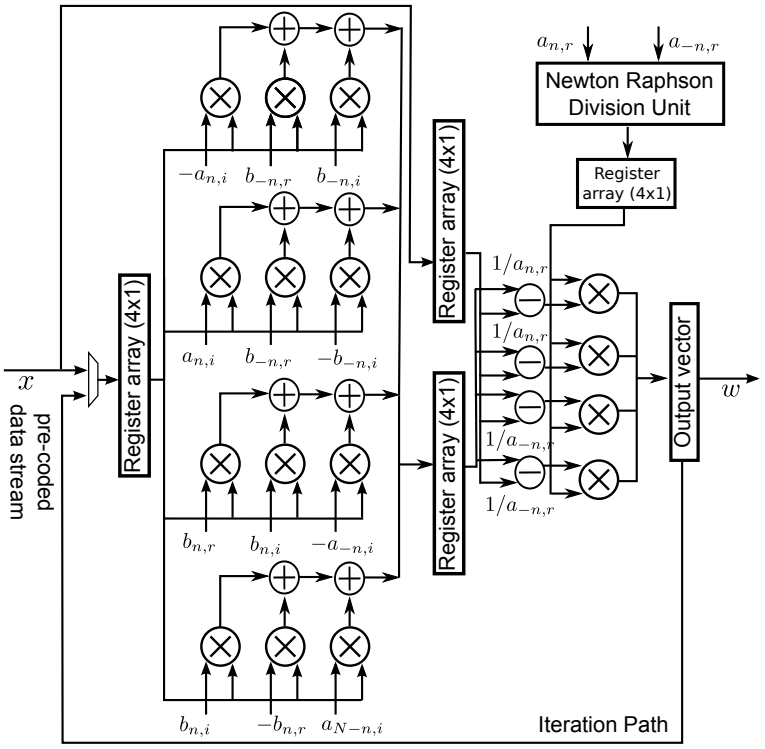


Figure 5.7: Hardware architecture of pre-compensation based on Jacobi solver.

Table 5.1: Hardware results for IQ imbalance pre-compensation in 28 nm FD-SOI technology.

	Per Instance	For $M = 100$
Area [mm^2]$^{\#}$	.008	0.8
Gate Count [10^3]	27	2700
Max. Clock [MHz]	200	200
Latency*[cycles]	2	2
Power [mW]	0.6	60

$^{\#}$ Only synthesis
* Latency is for 1 pair of tones per iteration.

5.5 Analysis of processing energy-per-bit

To perform a comparison of all the aforementioned baseband processing algorithms, we estimate the required processing energy per transmitted information bit. For relatively fair and reasonable comparison, the gate-level simulation results have been normalized to 28 nm FD-SOI technology with normal voltage supply. The corresponding results have been tabulated in Table5.2. The metric is evaluated for an LTE like 100×10 massive MIMO system with 16-QAM modulation.

The energy-per-bit for MR is around 50 pJ/bits, which is the lowest energy consumption among the investigated precoding schemes. This is in-line with the computational complexity, since MR only requires one matrix-vector multiplication. Furthermore, the operations are distributed per-antenna, reducing data-shuffling and power consumption of the system bus. Compared to MR, the ZF precoding is more complex and has a higher energy consumption. However, the performance of ZF is superior to that of MR for the same number of antennas, due to better inter-user interference suppression.

The discrete-time constant envelope precoding has lower energy requirements than ZF. Furthermore, since the PAR is low, extremely efficient PAs can be used. However, the implemented CE is for single-carrier narrow band system. For wideband systems, the computational complexity and energy consumption is expected to increase linearly with the number of taps in the channel. As an example, an LTE like system with FFTs required for OFDM modulation along with ZF, requires a total energy-per-bit of 580 pJ (FFT 180 pJ/bit + ZF 400 pJ/bit). Such an OFDM-based system can handle up to 144 taps, which would result in very high energy consumption for a corresponding single-carrier system with CE precoding. An alternative low complexity approach to tackle the PAPR issue is to use "antenna reservation" techniques. It is based on ZF in a OFDM system, with a complexity overhead of 15% of the total complexity, which when translated to estimated energy is 667 pJ/bit (1.15×580 pJ). This is a reasonable overhead considering that it provides around 4 dB of PAPR improvement. The performance improvement due to pre-compensation of IQ imbalance is very high with a relatively low energy consumption.

5.6 Conclusion

This chapter shows various implementations and estimates energy consumption of key processing blocks for MaMi systems. Several linear and non-linear precoding schemes, with and without reduction of PAPR to allow energy efficient PAs, are being compared. A scheme for IQ imbalance compensation is also analyzed. All comparisons show that digital baseband processing in a 100-antenna MaMi system can be done at reasonable hardware cost and processing energy consumption levels.

Table 5.2: Energy-per-bit comparison for different precoding techniques to tackle various hardware aspects.

	Gate count [K] [1]	Throughput [MSamples/sec]	Power [mW]	Technology	Energy-per-bit [pJ/bit]@28 nm [3]
Maximum ratio precoding	3.9	25	0.42	28nm	50
Zero Forcing (regularized) precoding	400	31.25	29	28nm	338 [4]
Single-carrier (Narrow band) constant envelope	14.1 [2]	50	3.96	65nm	175
Antenna reservation PAR aware precoding based on Zero forcing	-	-	-	-	+15% [5]
IQ imbalance pre-compensation	24	100	0.61	28nm	9
OFDM modulation 2048-FFT [7]	180		117	90nm	243

[1] Per instance cost, depending on throughput rates and implementation, multiple instances will be required.
[2] Require one instance per antenna and iteration.
[3] Energy-per-bit = (Power) $*$ $(28\,nm/\text{Tech}) * (1/\text{V}_{\text{DD}})^2/(\text{data-rate})$
[4] Includes Gram matrix generation and matrix inversion and MR filter, however, updated once every 10 sub-carrier and symbols.
[5] Antenna reservation has 15% more computational complexity compared to OFDM based ZF.

Chapter 6

Summary

Developing and understanding baseband signal processing is crucial to harvest the potential benefit provided by the MaMi concept and enable efficient hardware implementation for future practical deployment of this technology. In this deliverable, we have discussed key signal processing kernels for MaMi system. They focus on different aspects, including inter-cell/inter-user interference cancellation, reciprocity calibration, and hardware-aware precoding, which together promise good system performance with reasonable computational complexity. We also presented implementation-oriented processing profile based on system-level model containing processing complexity, components power consumption, and signal-noise-interference power. How different algorithms affect the implementation strategy is further discussed with detailed analysis on processing distribution, data shuffling bandwidth and latency, and memory requirement. Finally, the hardware implementation results of different precoder designs have been demonstrated. Using state-of-the-art CMOS technology, we are able to realize MaMi processing with low hardware cost and reasonable power consumption.

The extension and update in this deliverable, comparing to MAMMOET D3.1 [27], can be summarized as:

- Apply MaMi in multi-cell scenarios and extend linear detection methods, like M-MMSE, to actively suppress both intra-cell and inter-cell interference.
- Further reduce the computational complexity of linear detection schemes using the concept of interpolation without incurring a loss in ergodic rate.
- Extend the constant-envelope precoding to continuous-time waveforms, further reducing linearity requirements on the hardware and maximizing the power amplifier efficiency.
- Verified reciprocity calibration scheme with LuMaMi testbed and demonstrate real-life downlink MaMi transmission.
- Conduct hardware imperfection assessment to downlink system and in multi-cell cases.
- Provide more sophisticated system-level power modeling, validating the concept of MaMi from the point of view of power efficiency when compared to traditional base stations.
- Target on efficient baseband processor implementation and provide much detailed processing profile for optimized processing element, on-chip network, and memory subsystem.
- Validate via ASIC design and gate-level simulation that MaMi baseband digital signal processing can be extensively leveraged to achieve low-cost and low-power system realization.

We believe the MAMMOET project efforts collected in this deliverable is capable of serving as a solid basis and an appropriate guideline for efficient realizing MaMi baseband processing. It also opens future important research questions in this area. We expect that more development, analysis, and validation on the baseband processing will become available as the MAMMOET project progresses.

Bibliography

[1] GreenTouch consortium. http://www.greentouch.org/.

[2] John B. Anderson, Tor Aulin, and Carl-Erik Sundberg. *Digital Phase Modulation*. Springer, 1986.

[3] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah. Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits. *IEEE Trans. Inf. Theory*, 60(11):7112–7139, 2014.

[4] E. Björnson, E. G. Larsson, and M. Debbah. Massive MIMO for maximal spectral efficiency: how many users and pilots should be allocated? *IEEE Trans. Wireless Commun.* to appear.

[5] E. Björnson, M. Matthaiou, and M. Debbah. Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design. *IEEE Trans. Wireless Commun.*, 14(8):4353–4368, 2015.

[6] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah. Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer? *IEEE Trans. Wireless Commun.*, 14(8):4353–4368, 2015.

[7] Yuan Chen, Yu-Chi Tsao, Yu-Wei Lin, Chin-Hung Lin, and Chen-Yi Lee. An indexed-scaling pipelined fft processor for ofdm-based wpan applications. *Circuits and Systems II: Express Briefs, IEEE Transactions on*, 55(2):146–150, 2008.

[8] Björn Debaillie, Claude Desset, and Filip Louagie. A flexible and future-proof power model for cellular base stations. In *VTC Spring*, Glasgow, Scotland, May 2015.

[9] Claude Desset, Eddy De Greef, and Björn Debaillie. Power model for today's and future base stations. Available at http://www.imec.be/powermodel, 2015.

[10] Claude Desset, Björn Debaillie, and Filip Louagie. Towards a flexible and future-proof power model for cellular base stations. In *Invited at TIWDC*, Genoa, Italy, September 2013.

[11] Claude Desset, Björn Debaillie, and Filip Louagie. Modeling the hardware power consumption of large scale antenna systems. In *Invited at IEEE OnlineGreenComm*, November 2014.

[12] Claude Desset and Liesbet Van der Perre. Validation of low-accuracy quantization in massive MIMO and constellation EVM analysis. In *EUCNC*, Paris, France, June 2015.

[13] F. Fernandes, A. Ashikhmin, and T. L. Marzetta. Inter-cell interference in noncooperative TDD large scale antenna systems. *IEEE Journal on Selected Areas in Communications*, 31(2):192–201, February 2013.

[14] Jose Flordelis, Xiang Gao, Ghassan Dahman, Fredrik Rusek, Ove Edfors, and Fredrik Tufvesson. Spatial Separation of Closely-Spaced Users in Measured Massive Multi-User MIMO Channels. 2015.

[15] K. F. Guo and G. Ascheid. Performance analysis of multi-cell MMSE based receivers in MU-MIMO systems with very large antenna arrays. In *Proc. IEEE WCNC*, pages 3175–3179, Apr. 2013.

[16] U. Gustavsson et al. On the impact of hardware impairments on massive MIMO. In *Proc. IEEE GLOBECOM*, 2014.

[17] Babak Hassibi and Bertrand M. Hochwald. How much training is needed in multiple-antenna wireless links? *IEEE Trans. Inf. Theory*, 49(4):951–963, 2003.

[18] J. Hoydis, S. ten Brink, and M. Debbah. Massive MIMO in the UL/DL of cellular networks: How many antennas do we need? *IEEE J. Sel. Areas Commun.*, 31(2):160–171, 2013.

[19] Mark Ingels, Xiaoqiang Zhang, Kuba Raczkowski, Sungwoo Cha, Pieter Palmers, and Jan Craninckx. A linear 28nm CMOS digital transmitter with 2x12bit up to LO baseband sampling and -58dbc C-IM3. In *ESSCIRC*, pages 379–382, Venezia Lido, Italy, September 2014.

[20] 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.

[21] M. R. Khanzadi, G. Durisi, and T. Eriksson. Capacity of SIMO and MISO phase-noise channels with common/separate oscillators. *IEEE Trans. Commun.*, 63(9):3218–3231, 2015.

[22] Nikolaos Kolomvakis, Michail Matthaiou, Jingya Li, Mikael Coldrey, and Tommy Svensson. Massive mimo with iq imbalance: Performance analysis and compensation. In *Communications (ICC), 2015 IEEE International Conference on*, pages 1703–1709. IEEE, 2015.

[23] R. Krishnan et al. Linear massive MIMO precoders in the presence of phase noise—a large-scale analysis. *IEEE Trans. Veh. Technol.* To appear.

[24] Amos Lapidoth. *A Foundation in Digital Communication.* Cambridge University Press, 2009.

[25] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, February 2014.

[26] X. Li, E. Björnson, E. G. Larsson, S. Zhou, and J. Wang. Massive MIMO with multi-cell MMSE processing: Exploiting all pilots for interference suppression,. *IEEE Trans. Wireless Commun.* submitted.

[27] MAMMOET. Mammoet deliverable d3.1, first assessment of baseband processing. http://mammoet-project.eu/downloads/publications/deliverables/MAMMOET-D3.1-Assessment-PU-M12.pdf, Jan. 2015.

[28] T. L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. 9(11):3590–3600, November 2010.

[29] Saif Mohammed and Erik G. Larsson. Constant-envelope multi-user precoding for frequency-selective massive MIMO systems. 2(5):547–550, 2013.

[30] H. Q. Ngo, E. G. Larsson, and T. L. Marzerra. Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Transactions on Communications*, 61(4):1436–1449, April 2013.

[31] D. Petrovic, W. Rave, and G. Fettweis. Effects of phase noise on OFDM systems with and without PLL: Characterization and compensation. *IEEE Trans. Commun.*, 55(8):1607–1616, 2007.

[32] A. Pitarokoilis, E. Björnson, and E. G. Larsson. Optimal detection in training assisted SIMO systems with phase noise impairments. In *Proc. IEEE ICC*, 2015.

[33] A. Pitarokoilis, S. K. Mohammed, and E. G. Larsson. Uplink performance of time-reversal MRC in massive MIMO systems subject to phase noise. *IEEE Trans. Wireless Commun.*, 14(2):711–723, 2015.

[34] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek. Hardware efficient approximative matrix inversion for linear pre-coding in massive mimo. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1700–1703, June 2014.

[35] Hemanth Prabhu, Ove Edfors, Jose Rodrigues, Liang Liu, and Fredrik Rusek. A low-complex peak-to-average power reduction scheme for ofdm based massive mimo systems. In *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*, pages 114–117. IEEE, 2014.

[36] Hemanth Prabhu, Fredrik Rusek, Joachim Rodrigues, and Ove Edfors. High throughput constant envelope pre-coder for massive mimo systems. In *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2015.

[37] Joao Vieira, Fredrik Rusek, and Fredrik Tufvesson. Reciprocity calibration methods for massive MIMO based on antenna coupling. In *Global Communications Conference (GLOBECOM), 2014 IEEE*, Dec 2014.

[38] Hong Yang and Thomas L. Marzetta. Performance of conjugate and zero-forcing beamforming in large-scale antenna systems. 31(2):172–179, 2013.

[39] W. Zhang. A general framework for transmission with transceiver distortion and some applications. *IEEE Trans. Commun.*, 60(2):384–399, 2012.

[40] Qi Zheng, Yajing Chen, Hyunseok Lee, Ronald Dreslinski, Chaitali Chakrabarti, Achilleas Anastasopoulos, Scott Mahlke, and Trevor Mudge. Using graphics processing units in an lte base station. *Journal of Signal Processing Systems*, 78(1):35–47, 2014.

List of Abbreviations

ASIC Applicatin Specific Integarted Circuit

BER Bit Error Rate

BI Bus Interface

BIST Built-In Self-Test

BS Base Station

BSU backward substitution unit

CC Central Controller

CCDF Complementary Cumulative Distribution Function

CE Constant Envelope

CMOS Complementary Metal-Oxide Semiconductor

CORDIC COordinate Rotation DIgital Computer

CS circuit-switched

CSI Channel State Information

CTCE Continuous-Time Constant-Envelope

DFE Digital Front End

DFT Discrete Fourier Transform

DL Downlink

DLP Data Level Parallelism

DMA Direct Memory Access

DNS Diagonal Neumann Series

DPC Dirty Paper Coding

DRAM Dynamic RAM

DSP Digital Signal Processor

DTCE Discrete-Time Constant-Envelope

DUT Design Under Test

DVS dynamic voltage scaling

eDRAM Embedded DRAM

EVM Error Vector Magnitude

FD-SOI Fully Depleted Silicon On Insulator

FF Folding Factor

FFT Fast Fourier Transform

FIFO First-In-First-Out

GGM General Gram Matrix

GPU Graphical Processing Unit

HW Hardware

IBO Input-Back-off

IDFT Inverse Discrete Fourier Transform

IFFT Inverse Fast Fourier Transform

IO Input Output

ISI Inter-Symbol Interference

JTAG Joint Test Action Group

LO Local Oscillator

LTE Long Term Evolution

LuMaMi Lund University Massive MIMO

LUT Look-Up-Table

M-MMSE Multi-Cell MMSE

MAC Multiply-Accumulate

MACL Medium-Access Control Layer

MaMi Massive MIMO

MF Matched Filter

MIMO Multiple-Input Multiple-Output

MMSE Minimum Mean Square Error

MR Maximum Ratio

MSB Most Significant Bit

MSE Mean Square Error

MU-MIMO Multi-User MIMO

MUI Multi-User Interference

MUI Multi-user interference

NoC Network-on-Chip

NS Neumann Series

OBO Output-Back-off

OBR Out-of-Band (ratio) Power

OFDM Orthogonal Frequency-Division Multiplexing

P-ZF Full pilot-based Zero-Forcing

PA Power Amplifier

PAP per-antenna processing

PAPR Peak-to-Average Power Ratio

PE Processing Element

PLM Parallel Memory

PSP per-subcarrier processing

PUP per-user processing

QAM Quadrature-Amplitude Modulation

RAM Random Access Memory

RF Radio Frequency

RGF register file

RISC Reduced Instruction Set Computer

RNG Random Number Generator

RTL Register Transfer Level

S-MMSE Single-Cell MMSE

SDNR Signal-to-distortion-plus-noise ratio

SE Spectral Efficiency

SIMD Single-Instruction Multiple-Date

SINR Signal-to-interference-plus-noise ratio

SNDR Signal-to-noise-plus-distortion ratio

SNR Signal-to-Noise ratio

SRAM Static RAM

TAP Test Access Port

TDD Time Division Duplex

TDM Time Division Multiplexing

TNS Tri-diagonal Neumann series

UE User Equipment

UL Uplink

VLIW Very-large Instruction Word

VPU vector projection unit

WL Word Length

ZF Zero-Forcing