D3.3

Hardware-aware signal processing for MaMi systems

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-D3.3
Workpackage:	WP 3
Due Date:	31 December, 2016
Actual Submission Date:	22 December, 2016
Responsible Organisation:	LIU
Editor:	Emil Björnson
Dissemination Level:	PU
Revision:	1.0
Abstract:	Hardware-aware baseband processing and hardware implementation are developed and evaluated from a performance-complexity tradeoff perspective. The deliverable covers channel estimation, robust and efficient detection and precoding, novel MaMi hardware architectures and implementations, out-of-band radiation, link-level evaluation, and energy consumption profiling.
Keywords:	Massive MIMO, digital baseband processing, channel estimation, hardware implementation, performance-complexity tradeoff, accelerator, energy 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

Emil Björnson (LIU)

Contributors (ordered according to beneficiary numbers)

Claude Desset (IMEC)
Ali Zaidi (EAB)
Franz Dielacher, Christos Thomos (IFAT)
Ove Edfors, Liang Liu, Yangxurui Liu, Steffen Malkowsky, Hemanth Prabhu, Muris Sarajlic, Joao Vieira (ULUND)
Emil Björnson, Hei Victor Cheng, Salil Kashyap, Erik G. Larsson, Christopher Mollén (LIU)

Executive Summary

Massive MIMO (MaMi) is the next generation multi-user MIMO technology, which has been evolved to deliver the theoretical gains also under practical conditions. A key difference from previous generations is the large number of base station (BS) antennas and ability to spatially multiplex many tens user equipments (UEs) on the same time/frequency resource. The LuMaMi testbed has demonstrated that the MaMi technology can be implemented using off-the-shelf hardware, which shows the maturity of the technology, but also leads to a greatly over-designed system with huge computational capability, high-grade hardware resolution, and high energy consumption. These issues can be circumvented by tailoring the signal processing algorithms, processing architecture, and hardware implementation to the specific characteristics of MaMi, including channel hardening and the robustness towards hardware distortion. In particular, the MAMMOET project has demonstrated that a well-designed system implementation can achieve performance close to the theoretical limits, using a substantially reduced computational complexity, hardware resolution, and energy consumption.

Major developments on the algorithmic, architectural, and hardware design levels have been carried out in WP3 during the first two years of MAMMOET. In this deliverable, we consolidate these efforts by evaluating selected algorithms from a system and implementation perspective and by developing new algorithms to address key bottlenecks that remain. Conclusions include:

- The propagation channel has a greater predictability in MaMi systems than conventionally, which allows for estimating the channel relatively sparsely over time and frequency. Interpolation schemes over the frequency domain can be implemented efficiently and prediction schemes can prolong the time interval over which coherent downlink transmission is possible, thus increasing the interval between pilot transmissions.

- The resolution of the analog-to-digital converters (ADCs) at the BS can be reduced substantially, from the around 15 bit per real dimension in legacy systems to 3-4 bit in MaMi. The loss in data rate is negligible and the energy efficiency is maximized by doing so. The reason is that the total number of bits obtained by a BS is large, when having many antennas, while the hardware implementation is greatly simplified for every reduction in bit-width. MaMi can even be operated with 1 bit ADC resolution, but at a noticeable performance loss.

- Man-made interference, such as pilot contamination, is not a fundamental limiting factor in MaMi, as sometimes claimed in the literature. The M-MMSE detector, developed in this project, can reject any type of interference, at the price of the increased complexity of estimating the channel to each such interferer. Hence, there is a complexity-interference design tradeoff in practical implementation.

- Power control is key in both uplink and downlink, to determine how the high sum rate of a MaMi cell is divided between the UEs. Thanks to the channel hardening properties, the sum rate or max-min fairness power control problems can be solved efficiently and the solutions utilize only depend on the large-scale fading characteristics. Since the large-scale fading is fixed for a substantial time period, the power control coefficients can be updated every second instead of every millisecond, which allows for the use of the proposed power control algorithms in practice.

- The reduced hardware resolution may affect system operating in adjacent bands, due to the out-of-band (OOB) radiation. We have proved that the OOB radiation is basically the

same in MaMi as with legacy systems, using the same total transmit power and hardware resolution. Hence, the transmit power needs to be reduced along with the hardware resolution, to keep the OOB radiation fixed in practice. This is an important result, but not a showstopper, since MaMi systems are anyway intended to use their array gain to operate at lower transmit power levels.

- The main computational complexity in the baseband originates from OFDM modulation and ZF detection/precoding. These operations have been successfully implemented in CMOS in 28nm CMOS, for a typical MaMi setup with 128 BS antennas and 8 UEs. The conclusion from the implementation is that the complexity and energy consumption over these baseband processing tasks are highly feasible in practice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The key difference between massive MIMO (MaMi) and legacy systems is not the number of base station (BS) antennas; indeed, each panels in an LTE site contains tens of antennas which are interconnected to form a fixed beam tilted towards the coverage area. The main difference is instead the number of transceiver chains; each antenna is digitally steerable in MaMi while legacy systems typically have an order-of-magnitude fewer transceiver chains than physical antenna elements. In principle, legacy transceiver technology can be used along with MaMi, but resulting in a substantially higher cost, energy consumption, and computational complexity. This deliverable takes on the challenge of developing hardware-aware algorithms that can greatly reduce the implementation complexity of MaMi, by exploiting the spatial resolution and multiplexing capabilities while not over-dimensioning the hardware resolution and algorithmic accuracy.

This deliverable serves as a continuation of the line of work presented in Deliverable 3.1 [38] and Deliverable 3.2 [39]. This deliverable focuses on evaluation of hardware-aware signal processing strategies for MaMi, robust and efficient detectors, and hardware-implementation of selected algorithms.

In Chapter 2, we provide new algorithmic development and evaluation related to the baseband processing. Idealized block-fading models, where each channel is fixed within a coherence time-frequency interval, have dominated the MaMi literature. In Deliverable 3.2, we considered realistic frequency-selective fading and analyzed how densely the channels need to be estimated in the frequency domain. In this deliverable, we validate previous algorithms from a system-level perspective and extend the analysis to also capture time-variations of the channels. Furthermore, we show that the uplink detection is robust towards the distortion caused by low-resolution quantization. The number of quantization bits required for operating the system efficiently, from a rate or energy efficiency perspective, are quantified. We also show that the multi-cell minimum mean-squared error (M-MMSE) detector, which was developed as an optimal detector in Deliverable 3.2, is robust with respect to pilot contamination and other types of man-made interference.

In Chapter 3, we consider the interaction between hardware and signal processing beyond the baseband processing. The complexity of uplink and downlink power control algorithms can be greatly reduced due to the channel hardening, which alleviates the need to adapt the power control to small-scale or frequency-selective fading variations. We also address the important area of out-of-band (OOB) radiation, to determine if and how systems operating in adjacent bands are affected by the beamforming that takes place in MaMi.

The practical feasibility of the MaMi baseband processing is validated in Chapter 4, by chip implementation of the key processing tasks, including OFDM modulation, uplink detection,

and downlink precoding. The energy consumption of this dedicated implementation is showed to be low. A processing architecture that divided processing tasks between per-antenna, per-subcarrier, and per-UE is also presented.

In this deliverable, MaMi has been analyzed with different focus in different sections. To keep the analysis and notation simple, each section uses its own dedicated system model covering all the aspects that are important for that particular analysis and discussion, while leaving out unimportant aspects. Finally, we note that the development of algorithms for per-antenna constant envelope precoding are disseminated in Deliverable 3.1 and Deliverable 3.2, and not in this deliverable, due to very successful advances in this area during first two years of the project.

Chapter 2

Hardware-Aware Baseband Processing

This chapter describes hardware-aware refinements of the baseband processing algorithms for downlink precoding, uplink detection, and channel acquisition in MaMi. The main observation is that the computational and hardware complexities can be greatly reduced in MaMi, as compared to what has previously been studied in the literature, with only a negligible loss in end performance. This chapter both provides new algorithms and deeper analysis of algorithms that have been proposed in previous deliverables in the MAMMOET project.

2.1 Time-Frequency Channel Acquisition Schemes

The MaMi physical layer concept was initially developed under the simplifying assumption of block-fading channels, where each channel takes a random realization within a time/frequency block called the channel coherence interval and then another independent random realization in the next such interval. In practice, the channel variations in the time and frequency domains are continuous, thus the dimensionality of a coherence interval cannot be characterized exactly and there will always be some channel variations within each interval.

In this section, we present analysis and results on the estimation, interpolation, and prediction of channels over the time and frequency domain, using practical models for the time and frequency channel variations. Initial work on this topic was presented in Section 2.4 in Deliverable 3.2 [39], with focus on frequency interpolation. We extend this analysis to time-domain prediction in Section 2.1.1. In Section 2.1.2, the previously developed frequency interpolation schemes are analyzed in terms of system-level performance and complexity.

2.1.1 Prediction of Prediction of Channel Response

In this subsection, we present a framework for prediction of the channel variations in MaMi systems, based on Kalman filtering. This framework is particularly relevant for the downlink precoding, which must rely on the uplink pilots that were transmitted earlier in time. In particular, we want to answer the following practical questions:

- Can channel prediction be used to reduce the training overhead in MaMi? And what is the loss in performance if the zero forcing (ZF) matrix is based on the predicted channel compared to that based on the true channel (estimated using pilots)?
- How robust is the predictor to imperfect knowledge of the temporal channel statistics? What is the loss in performance incurred, for example, due to a mismatch in the Doppler spread value?

- Can we use simple predictor models at the base station (BS) to approximate a given channel spectrum?

To this end, we consider both an auto-regressive (AR) process as well as an auto-regressive moving average (ARMA) process to model at the BS the time-variations of the channel whose true temporal spectrum can be either the Jakes spectrum or a rectangular spectrum. We derive both the uplink and the downlink achievable rates for imperfect CSI acquired using Kalman estimation or prediction. We present extensive numerical results to quantify the loss in rate incurred due to prediction errors and due to imperfect knowledge of the channel statistics.

System Model

We consider the uplink and downlink of a single-cell massive MIMO-OFDM system, where the bandwidth is divided into S orthogonal subcarriers and there are a total of N OFDM symbols. The BS is equipped with an array of M antennas and there are K single-antenna users in the cell. The L-tap channel from the kth user to the mth antenna at the BS over the nth OFDM symbol is denoted by

$$\tilde{\mathbf{g}}_k^m[n] = [\tilde{g}_k^m[n,0]\ \tilde{g}_k^m[n,1]\ \cdots\ \tilde{g}_k^m[n,L-1]]^T. \tag{2.1}$$

Moreover, for a user-antenna pair, the taps are assumed to be independent, but they need not be identically distributed. Each entry $\tilde{g}_k^m[n,l]$ consists of both small scale fading and distance-dependent path loss of the kth user. We assume that the path loss from a user is the same to all the antennas at the BS. This assumption is justified because the size of a co-located MaMi antenna array is much smaller than the distance between the users and the BS. Furthermore, we assume uncorrelated Rayleigh fading and the path loss constant across time. Therefore, $\tilde{\mathbf{g}}_k^m[n] \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda}_k)$, where $\mathbf{\Lambda}_k$ is a diagonal matrix with the diagonal representing the channel power delay profile (PDP) of the kth user and that includes the path loss as well.

Uplink Pilot Signaling and Channel Estimation: The frequency-domain signal $\mathbf{y}_m[n] \in \mathbb{C}^{N_p \times 1}$ received over the nth OFDM symbol at the mth antenna of the BS during uplink pilot signaling is

$$\mathbf{y}_m[n] = \sum_{k=1}^{K} \sqrt{p_k^u}\mathbf{\Upsilon}_k^t \mathbf{\Omega}_r \tilde{\mathbf{g}}_k^m[n] + \mathbf{w}_m[n], \tag{2.2}$$

where p_k^u is the uplink pilot SNR per subcarrier and per OFDM symbol of the kth user, $\mathbf{\Upsilon}_k^t \in \mathbb{C}^{N_p \times N_p}$ is a diagonal matrix with the N_p-length pilot sequence $\mathbf{x}_k^t$ corresponding to user k, $\mathbf{\Omega}_r \in \mathbb{C}^{N_p \times L}$ consists of the first L columns and N_p rows of the S-point discrete Fourier transform (DFT) matrix $\mathbf{\Omega} \in \mathbb{C}^{S \times S}$ where $[\mathbf{\Omega}]_{m,n} = e^{-j2\pi(m-1)(n-1)/S}$. These rows correspond to the set of subcarriers on which the N_p pilots are sent. The pilots are assumed to be equally spaced in frequency and $N_p \leq S$. The noise vector at the mth antenna of the BS over the nth OFDM symbol is denoted by $\mathbf{w}_m[n]$. Furthermore, $\mathbf{w}_m[n] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_p})$, independent and identically distributed (i.i.d.) across antennas m and time n. If the pilot sequences are chosen such that[1] $\mathbf{\Omega}_r^{\mathrm{H}} \mathbf{\Upsilon}_k^{t^{\mathrm{H}}} \mathbf{\Upsilon}_i^t \mathbf{\Omega}_r = N_p \mathbf{I}_L \delta_{ki}$, where $\delta_{ki} = 1$ if $k = i$ and $\delta_{ki} = 0$ otherwise. Then a sufficient statistics for estimating $\tilde{\mathbf{g}}_k^m[n]$ is

$$\tilde{\mathbf{y}}_k^m[n] = \frac{1}{\sqrt{N_p}} \mathbf{\Omega}_r^{\mathrm{H}} \mathbf{\Upsilon}_k^{t^{\mathrm{H}}} \mathbf{y}_m[n] = \sqrt{p_k^u N_p}\tilde{\mathbf{g}}_k^m[n] + \tilde{\mathbf{w}}_k^m[n], \tag{2.3}$$

[1] To ensure orthogonality among pilot sequences of different users over any OFDM symbol, it is necessary to have $N_p \geq KL$.

where $\tilde{\mathbf{w}}_k^m[n] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$, i.i.d. across k and m.

The autocorrelation function (ACF) captures the variability of a wireless channel over time. The ACF, which is a second-order temporal statistic, depends on the propagation geometry, the velocity with which the user moves and the antenna characteristics. A common assumption is that the propagation path from the transmitter to the receiver consists of two-dimensional scattering with a vertical monopole antenna at the receiver. In that case, the theoretical power spectral density (PSD) of either the in-phase or the quadrature component of the received fading signal has the U-shaped band-limited form [27]

$$S(f) = \frac{1}{\pi f_d \sqrt{1 - \left(\frac{f}{f_d}\right)^2}}, \quad |f| \leq f_d \tag{2.4}$$

and is 0 otherwise, where $f_d = v/\lambda$ is the maximum Doppler frequency in Hz, v is the mobile speed, and λ is the wavelength of the received carrier wave. The corresponding continuous-time normalized autocorrelation of the received signal is $r_{gg}(\tau) = J_0(2\pi f_d T_s \tau)$, where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind.

Another variation on the PSD that is based on a three-dimensional model with isotropic scattering in all the three directions is examined in [13]. In this model, the PSD has flat bandlimited characteristics with a normalized ACF $r_{gg}(\tau) = \text{sinc}(2f_d T_s \tau)$.

Our objective is to estimate the channel taps at different time instants. To this end, we consider both an ARMA process as well as an AR process to approximately model the time variations of the channel taps $\tilde{g}_k^m[n, l]$. In this work, we consider the channels that can either have Bessel ACF or sinc ACF as described above. A pth order AR model for $\tilde{g}_k^m[n, l]$ is presented as

$$\tilde{g}_k^m[n, l] = \sum_{i=1}^{p} \phi_i \tilde{g}_k^m[n - i, l] + \psi_0 \tilde{\eta}_k^m[n, l], \tag{2.5}$$

where ϕ_is and ψ_0 are obtained using the Yule-Walker equations as follows [29]:

$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} r_{gg}[0] & r_{gg}[1] & \cdots & r_{gg}[p-1] \\ r_{gg}[1] & r_{gg}[0] & \ddots & r_{gg}[p-2] \\ \vdots & \ddots & \ddots & \vdots \\ r_{gg}[p-1] & \cdots & \cdots & r_{gg}[0] \end{bmatrix}^{-1} \begin{bmatrix} r_{gg}[1] \\ r_{gg}[2] \\ \vdots \\ r_{gg}[p] \end{bmatrix} \tag{2.6}$$

where $r_{gg}[\tau] = \mathbb{E}\left[\tilde{g}_k^m[n, l]\tilde{g}_k^m[n - \tau, l]^*\right]$ and

$$\psi_0 = \Lambda_k[l, l] \left(r_{gg}[0] - \sum_{i=1}^{p} \phi_i r_{gg}[i] \right), \tag{2.7}$$

where $\Lambda_k[l, l]$ is the power of the lth channel tap. We use the AR model to capture the time-variations of the channel that has a Bessel ACF, i.e., channels for which $r_{gg}[\tau] = J_0(2\pi f_d T_s \tau)$. Similarly, an ARMA($p$,$p$) model for $\tilde{g}_k^m[n, l]$ can be presented as [29]

$$\tilde{g}_k^m[n, l] = \sum_{i=1}^{p} \phi_i \tilde{g}_k^m[n - i, l] + \sum_{i=0}^{p} \psi_i \tilde{\eta}_k^m[n - i, l], \tag{2.8}$$

where the coefficients ϕ_i and ψ_i, for $i = 0, \ldots, p$, are obtained from the transfer function of a Butterworth low pass filter of order p designed with the cutoff frequency equal to the

normalized Doppler frequency f_dT_s, where f_d is the maximum Doppler frequency and T_s is the OFDM symbol duration.[2] We use the ARMA model to capture the time-variations of the channel that has a sinc ACF, i.e., channels for which $r_{gg}[\tau] = \text{sinc}(2f_dT_s\tau)$. Since the ARMA coefficients are taken from the transfer function of a Butterworth low pass filter that has a rectangular spectrum, it therefore can closely model such channels.

The AR or the ARMA models used to capture the time-variations of the channel can be equivalently given as a state-space model. Specifically, the state transitions can be modeled as

$$\tilde{\mathbf{X}}_k^m[n+1,l] = \mathbf{A}\tilde{\mathbf{X}}_k^m[n,l] + \mathbf{B}\tilde{\mathbf{u}}_k^m[n+1,l], \tag{2.10}$$

where $\tilde{\mathbf{X}}_k^m[n,l] \triangleq [\tilde{g}_k^m[n,l],\ldots,\tilde{g}_k^m[n-p+1,l]]^T$ is the state of the system of the lth channel tap at time n, p is the order of the model, and $\tilde{\mathbf{u}}_k^m[n+1,l]$ is the white Gaussian process noise.

The matrices $\mathbf{A}$ and $\mathbf{B}$ are given as follows:

$$\mathbf{A} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{C}^{p\times p}, \tag{2.11}$$

$$\mathbf{B} = \begin{pmatrix} \psi_0 & \psi_1 & \cdots & \psi_p \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{C}^{p\times(p+1)}, \text{for ARMA model,} \tag{2.12}$$

$$\mathbf{B} = \begin{pmatrix} \psi_0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{C}^{p\times 1}, \text{for AR model.} \tag{2.13}$$

From (2.3), the observations of the state (each channel tap) at time n can be represented by a linear equation of the form

$$\tilde{y}_k^m[n,l] = \mathbf{S}\tilde{\mathbf{X}}_k^m[n,l] + \tilde{w}_k^m[n,l], \tag{2.14}$$

where

$$\mathbf{S} = [\sqrt{p_k^u N_p}, 0, \ldots, 0] \tag{2.15}$$

and $\tilde{w}_k^m[n,l]$ is the additive measurement noise. Furthermore, $\tilde{w}_k^m[n,l] \sim \mathcal{CN}(0,1)$.

Now, given a set of observations $\tilde{y}_k^m[1,l]$, $\tilde{y}_k^m[2,l]$, …, $\tilde{y}_k^m[n+1,l]$, the task is to determine the estimation filter that at the $(n+1)$th time instant would generate an optimal estimate $\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]$ of the state $\tilde{\mathbf{X}}_k^m[n+1,l]$. We present below the steps to obtain the Kalman estimate of the lth channel tap:

1. *Initialization:* We begin by initializing $\hat{\tilde{\mathbf{X}}}_k^m[0,l]\big|[0] = \mathbf{0}$ and the prediction error covariance

[2]The transfer function $T(z)$ of a Butterworth low pass filter of order p designed with a cutoff frequency equal to the normalized Doppler frequency f_dT_s is given by

$$T(z) = \frac{\sum_{i=0}^{p}\psi_i z^{-i}}{1+\sum_{i=1}^{p}\phi_i z^{-i}}, \tag{2.9}$$

where $z = \exp(j2\pi f_d/f_s)$.

matrix $\mathbf{P}_{0|0} = \Lambda_k[l,l]\mathbf{R}$, where

$$\mathbf{R} = \begin{bmatrix} r_{gg}[0] & r_{gg}[1] & \cdots & r_{gg}[p-1] \\ r_{gg}[1] & r_{gg}[0] & \ddots & r_{gg}[p-2] \\ \vdots & \ddots & \ddots & \vdots \\ r_{gg}[p-1] & \cdots & \cdots & r_{gg}[0] \end{bmatrix} \tag{2.16}$$

As mentioned before, the Jakes channel model has a Bessel ACF, i.e., $r_{gg}[\tau] = J_0(2\pi f_d T_s \tau)$ while the channel model with rectangular spectrum has sinc ACF, i.e., $r_{gg}[\tau] = \text{sinc}(2f_d T_s \tau)$.

2. *One-step-ahead prediction:* This involves estimating the state at $n+1$ based on observations up to time instant n:

$$\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n] \triangleq \mathbb{E}[\tilde{\mathbf{X}}_k^m[n+1,l]\big|(\tilde{y}_k^m)^n] = \mathbf{A}\hat{\tilde{\mathbf{X}}}_k^m[n,l]\big|[n], \tag{2.17}$$

where $(\tilde{y}_k^m)^n = \tilde{y}_k^m[1,l],\ldots,\tilde{y}_k^m[n,l]$.

3. *Computing the prediction error covariance matrix:* The prediction error covariance matrix is given by

$$\begin{aligned} \mathbf{P}_{n+1|n} &\triangleq \mathbb{E}[(\tilde{\mathbf{X}}_k^m[n+1,l] - \hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n])(\tilde{\mathbf{X}}_k^m[n+1,l] - \hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n])^H\big|(\tilde{y}_k^m)^n] \\ &= \mathbf{A}\mathbf{P}_{n|n}\mathbf{A}^H + \mathbf{B}\mathbf{B}^H. \end{aligned} \tag{2.18}$$

4. *Kalman update:* Having obtained the predictive estimate of the state, $\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n]$, suppose now that we take another observation $\tilde{y}_k^m[n+1,l]$, then this can be used to update the predictive estimate, i.e. to obtain $\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n+1]$. The Kalman estimate of the state of the system is given by

$$\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n+1] = \hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n] + \mathbf{K}_{n+1}\left(\tilde{y}_k^m[n+1,l] - \mathbf{S}\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n]\right), \tag{2.19}$$

where $\tilde{y}_k^m[n+1,l]$ denotes the observation at the current time instant $n+1$ and $\mathbf{S}\hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n]$ denotes the predicted observation. Thus, the Kalman estimate of the channel tap can be interpreted as the sum of the prediction and a fraction of the difference between the predicted and the actual observation. The Kalman gain matrix $\mathbf{K}_{n+1}$ which is chosen so as to minimize the mean square error is given by

$$\mathbf{K}_{n+1} = \mathbf{P}_{n+1|n}\mathbf{S}^H\left(\mathbf{S}\mathbf{P}_{n+1|n}\mathbf{S}^H + 1\right)^{-1}. \tag{2.20}$$

Next we compute the updated error covariance matrix through Riccati recursion.

5. *Updated error covariance matrix:* The updated error covariance matrix is given by:

$$\begin{aligned} &\mathbf{P}_{n+1|n+1} \\ &\triangleq \mathbb{E}[(\tilde{\mathbf{X}}_k^m[n+1,l] - \hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n+1])(\tilde{\mathbf{X}}_k^m[n+1,l] - \hat{\tilde{\mathbf{X}}}_k^m[n+1,l]\big|[n+1])^H\big|(\tilde{y}_k^m)^{n+1}] \\ &= (\mathbf{I}_P - \mathbf{K}_{n+1}\mathbf{S})\,\mathbf{P}_{n+1|n}\,(\mathbf{I}_P - \mathbf{K}_{n+1}\mathbf{S})^H + \mathbf{K}_{n+1}\mathbf{K}_{n+1}^H. \end{aligned} \tag{2.21}$$

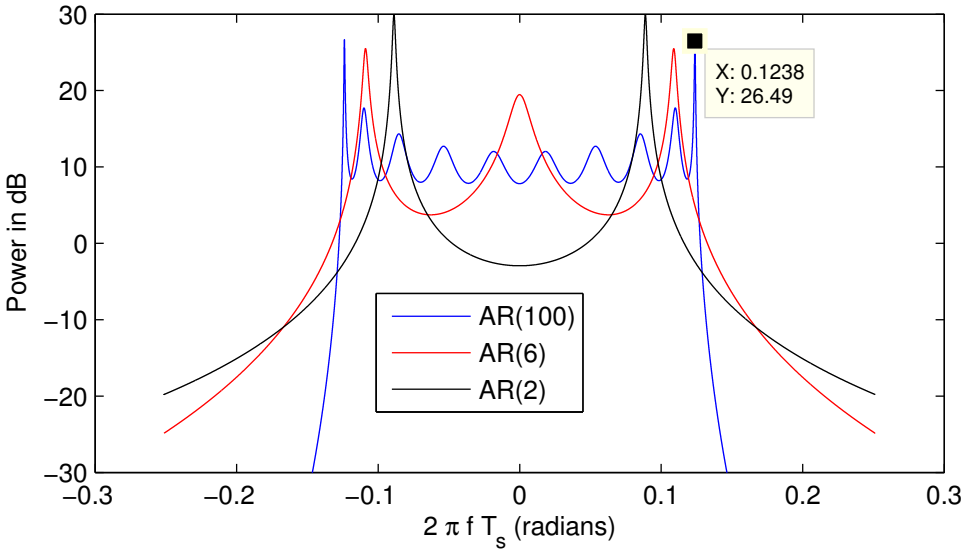


Figure 2.1: Power spectral density of AR predictors of different orders, $f_d T_s = 0.02$

Next, we will briefly illustrate the PSDs of the two types of spectrum, and the ability to approximate them by AR or ARMA predictors. Figure 2.1 plots the PSDs of AR predictors of different orders, whose AR coefficients are obtained using Yule-Walker equations based on the Jakes spectrum. These are used as approximate models to characterize the Jakes spectrum. We can observe that as the model order increases, the spectrum becomes sharper at the edges at the cost of increased ripples in the passband.

Figure 2.2 plots the PSDs of ARMA predictors of different orders, whose ARMA coefficients are obtained from the transfer function of a Butterworth low pass filter designed with a cutoff frequency of $f_d T_s$. These are used as approximate models to characterize the rectangular spectrum with sinc ACF. We can observe that as the model order increases, the spectrum falls off sharply at the transition from the passband to the stopband.

Uplink Data Transmission: The data signal $\mathbf{y}_u[n,s] \in \mathbb{C}^{M\times 1}$ received on the uplink over the sth subcarrier and the nth OFDM symbol is given by

$$\mathbf{y}_u[n,s] = \mathbf{G}[n,s]\mathbf{\Phi}_u[n,s]^{1/2}\mathbf{x}_u[n,s] + \mathbf{w}_u[n,s], \tag{2.22}$$

where $\mathbf{G}[n,s] \in \mathbb{C}^{M\times K}$ denotes the frequency-domain channel matrix over the nth OFDM symbol and the sth subcarrier such that $\mathbf{G}[n,s] = [\mathbf{g}_1[n,s] \dots \mathbf{g}_K[n,s]]$ and $\mathbf{g}_k[n,s] \in \mathbb{C}^{M\times 1}$ is the frequency-domain channel vector of the kth user over the nth OFDM symbol and the sth subcarrier. Furthermore, $[\mathbf{G}[n,s]]_{m,k} = G_k^m[n,s] = \omega_s^H \tilde{\mathbf{g}}_k^m[n]$, where ω_s^H denotes the sth row consisting of only the first L columns of the N-point DFT matrix $\mathbf{\Omega}$. Also, $\mathbf{\Phi}_u[n,s]$ is a $K \times K$ diagonal matrix of the uplink data SNR of the K users such that $[\mathbf{\Phi}_u[n,s]]_{k,k} = p_k^u$. The data vector of the K users over the nth OFDM symbol and the sth subcarrier is denoted by $\mathbf{x}_u[n,s]$ and the noise vector at the BS over the nth OFDM symbol and the sth subcarrier is denoted by $\mathbf{w}_u[n,s]$. Furthermore, $\mathbf{x}_u[n,s] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ and $\mathbf{w}_u[n,s] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$.

Downlink Data Transmission: The signal $\mathbf{x}_d[n,s] \in \mathbb{C}^{M\times 1}$ transmitted by the BS on the downlink over the sth subcarrier and the nth OFDM symbol is given by

$$\mathbf{x}_d[n,s] = \sqrt{p_d}\hat{\mathbf{A}}[n,s]\mathbf{q}[n,s], \tag{2.23}$$

where p_d is the downlink SNR, $\hat{\mathbf{A}}[n,s] \in \mathbb{C}^{M\times K}$ is the precoding matrix that depends on the estimated or the predicted CSI at the nth OFDM symbol index and the sth subcarrier, and

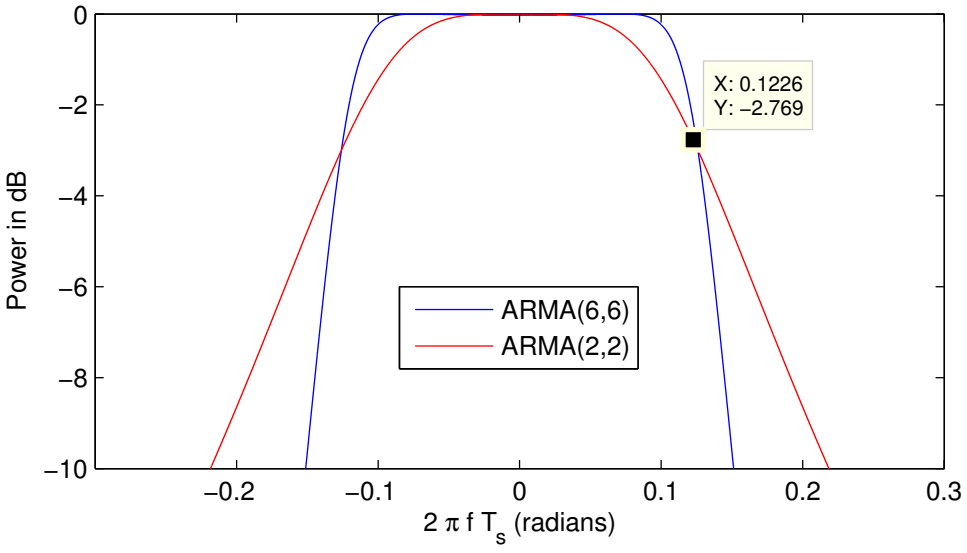


Figure 2.2: Power spectral density of ARMA predictors of different orders, $f_d T_s = 0.02$

$\mathbf{q}[n,s] \in \mathbb{C}^{K\times 1}$ is the vector of information bearing symbols of the K users. To satisfy the power constraint at the BS, the transmission symbols and the precoding matrix $\hat{\mathbf{A}}[n,s]$ are chosen such that $\mathbb{E}[\mathbf{q}[n,s]] = \mathbf{0}$, $\mathbb{E}[\mathbf{q}[n,s]\mathbf{q}[n,s]^H] = \mathbf{I}_K$ and $\mathrm{tr}(\hat{\mathbf{A}}[n,s]\hat{\mathbf{A}}[n,s]^H) = 1$, where $\mathrm{tr}(\cdot)$ is the trace. This will imply that $\mathbb{E}[\|\mathbf{x}[n,s]\|^2] = p_d$.

The signal vector $\mathbf{y}_d[n,s] \in \mathbb{C}^{K\times 1}$ received collectively at the K users is given by

$$\mathbf{y}_d[n,s] = \mathbf{G}^H[n,s]\mathbf{x}_d[n,s] + \mathbf{w}_d[n,s], \tag{2.24}$$

where $\mathbf{w}_d[n,s] \in \mathbb{C}^{K\times 1}$ and whose kth entry denotes the additive white Gaussian noise at the kth user. Furthermore, we assume that $w_{d_k}[n,s] \sim \mathcal{CN}(0,1)$. Then, the signal $y_{d_k}[n,s]$ received on the downlink at the kth user over the nth OFDM symbol and the sth subcarrier can be written as

$$y_{d_k}[n,s] = \sqrt{p_d}\mathbf{g}_k^H(n,s)\hat{\mathbf{a}}_k[n,s]q_k[n,s] + \sqrt{p_d}\sum_{i\neq k}\mathbf{g}_k^H[n,s]\hat{\mathbf{a}}_i[n,s]q_i[n,s] + w_{d_k}[n,s]. \tag{2.25}$$

Achievable Rate Analysis

We now present expressions for the achievable uplink and the downlink rates.

Uplink Rate Analysis: We let the detector matrix $\widehat{\mathbf{A}}[n,s]$ be an $M \times K$ matrix which depends on the estimated frequency-domain channel matrix and on the choice of detection method. The received vector on the nth OFDM symbol and the sth subcarrier after using the detector is given by

$$\mathbf{r}_u[n,s] = \hat{\mathbf{A}}^{\mathrm{H}}[n,s]\mathbf{y}_u[n,s] = \hat{\mathbf{A}}^{\mathrm{H}}[n,s]\mathbf{G}[n,s]\mathbf{\Phi}_u[n,s]^{1/2}\mathbf{x}_u[n,s] + \hat{\mathbf{A}}^{\mathrm{H}}[n,s]\mathbf{w}_u[n,s]. \tag{2.26}$$

Thus, the kth element of $\mathbf{r}_u[n,s]$ is

$$r_{u_k}[n,s] = \sqrt{p_k^u}\hat{\mathbf{a}}_k^H[n,s]\mathbf{g}_k[n,s]x_{u_k}[n,s] + \sum_{i=1,i\neq k}^{K}\sqrt{p_i^u}\hat{\mathbf{a}}_k^H[n,s]\mathbf{g}_i[n,s]x_{u_i}[n,s] + \hat{\mathbf{a}}_k^H[n,s]\mathbf{w}_u[n,s], \tag{2.27}$$

where $\hat{\mathbf{a}}_k[n,s] \in \mathbb{C}^{M\times 1}$ is the kth column of $\hat{\mathbf{A}}(n,s)$ corresponding to the kth user and is a function of the estimated channel. Let us define $\widehat{V}_{ki} = \hat{\mathbf{a}}_k^H[n,s]\mathbf{g}_i[n,s]$. Then, we can write (2.27) as

$$r_{u_k}[n,s] = \sqrt{p_k^u}\widehat{V}_{kk}x_{u_k}[n,s] + \sum_{i\neq k}\sqrt{p_i^u}\widehat{V}_{ki}x_{u_i}[n,s] + \hat{\mathbf{a}}_k^H[n,s]\mathbf{w}_u[n,s]. \tag{2.28}$$

Using the standard technique in [28], an achievable ergodic uplink rate over subcarrier s and OFDM symbol n is given by

$$R_k(n,s) = \log_2\left(1 + \frac{p_k^u\left|\mathbb{E}\left[\widehat{V}_{kk}\right]\right|^2}{\mathbb{E}\left[||\hat{\mathbf{a}}_k^H[n,s]||^2\right] + p_k^u\text{var}\left[\widehat{V}_{kk}\right] + \sum_{i\neq k}p_i^u\mathbb{E}\left[\left|\widehat{V}_{ki}\right|^2\right]}\right), \tag{2.29}$$

Downlink Rate Analysis: We let the precoder matrix $\widehat{\mathbf{A}}[n,s]$ be an $M\times K$ matrix which depends on the estimated or the predicted frequency-domain channel matrix and on the choice of precoding method. Let us define $\widehat{W}_{ki} = \mathbf{g}_k^H[n,s]\hat{\mathbf{a}}_i[n,s]$. Then, from (2.25), the signal received at the kth user can also be written as

$$y_{d_k}[n,s] = \sqrt{p_d}\widehat{W}_{kk}q_k[n,s] + \sqrt{p_d}\sum_{i\neq k}\widehat{W}_{ki}q_i[n,s] + w_{d_k}[n,s]. \tag{2.30}$$

Using the standard technique in [28], an achievable ergodic downlink rate over subcarrier s and OFDM symbol n, assuming that the user has only statistical CSI is given by

$$R_k(n,s) = \log_2\left(1 + \frac{p_d\left|\mathbb{E}\left[\widehat{W}_{kk}\right]\right|^2}{1 + p_d\text{var}\left[\widehat{W}_{kk}\right] + p_d\sum_{i\neq k}\mathbb{E}\left[\left|\widehat{W}_{ki}\right|^2\right]}\right), \tag{2.31}$$

Note that we compute a separate rate for each (n,s). On an average, the rate over any time-frequency grid will be

$$\frac{1}{SN_d}\sum_n\sum_s R_k(n,s), \tag{2.32}$$

where N_d is the number of OFDM symbols used for downlink data transmission. Furthermore, with ZF, the precoding matrix is given by $\widehat{\mathbf{A}}[n,s] = \alpha_{\text{ZF}}\widehat{\mathbf{G}}[n,s]\left(\widehat{\mathbf{G}}[n,s]^H\widehat{\mathbf{G}}[n,s]\right)^{-1}$, where the normalization constant α_{ZF} is chosen to satisfy the power constant $\text{tr}(\hat{\mathbf{A}}[n,s]\hat{\mathbf{A}}[n,s]^H) = 1$.

Numerical Results

In this section, we present numerical results to understand how often do we need to predict the channel and compute the detector or the precoder in time without a significant reduction in the achievable rate. Unless mentioned otherwise, we take $M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, and $p_d = p_k^u = -5$ dB. We consider a frequency-selective channel with uniform power delay profile[3] and we take the number of pilot subcarriers $N_p = KL$. We further assume that the pilots are equally spaced in frequency and are located over OFDM symbols $1-7$ as shown in Figure 2.3. We consider normalized Doppler spread values in the range 0.01 to 0.03, which corresponds to speeds in the range 80 to 240 km/h at a carrier frequency of 2 GHz and OFDM symbol duration $T_s = 66.67$ μs.

[3]Note that channels with uniform power delay profile represent the worst case scenario [57]. Therefore, study of such channels gives us an insight into the performance under the worst case conditions.

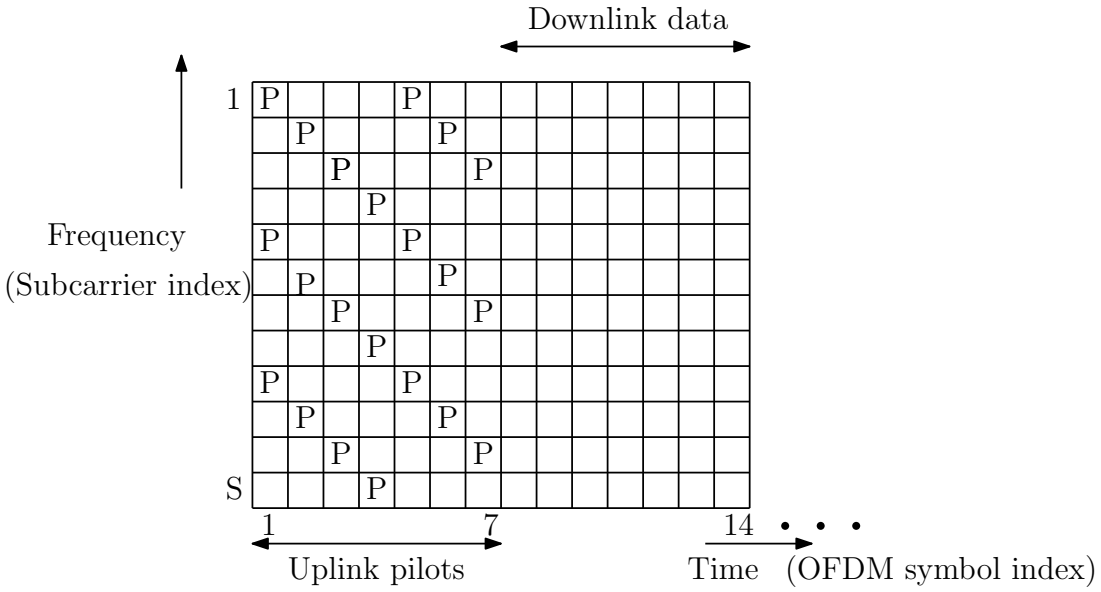


Figure 2.3: Uplink pilots and downlink data transmission.

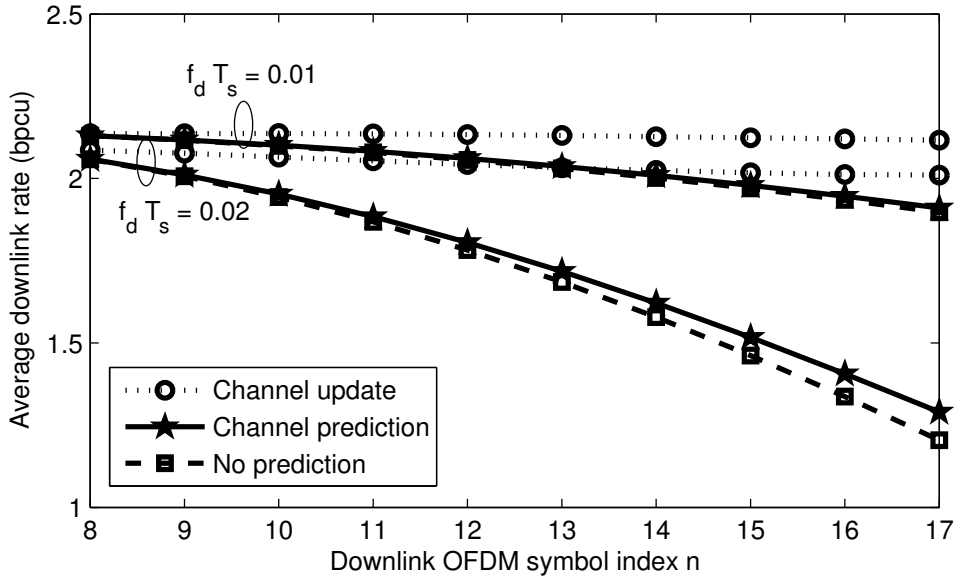


Figure 2.4: Channel with rectangular spectrum ans sinc ACF, ARMA(2, 2) predictor ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).

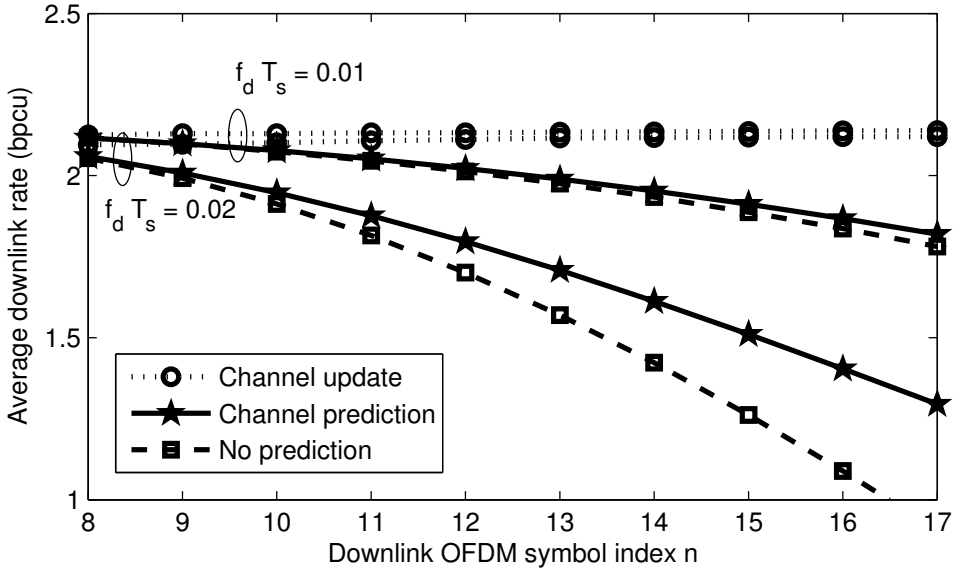


Figure 2.5: Jakes channel with Bessel ACF, AR(2) predictor ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).

Figure 2.4 plots the average downlink rate $\left(\frac{1}{S}\sum_{s=1}^{S} R_k(n,s)\right)$ as a function of the OFDM symbol index for the case when the channel has sinc ACF while an ARMA(2,2) predictor with coefficients taken from a Butterworth low pass filter of order 2 is used to capture the time variations in the channel. The pilots are located over symbols 1 to 7 and they are equally spaced in frequency as shown in Figure 2.3. Over these symbols, based on the observations, Kalman estimates of the channel matrices are obtained. For symbol indices 8 to 17 channel prediction is performed as only downlink data is transmitted over these symbols and there are no uplink pilots. The ZF precoder computations over symbols 8 to 17 are based on the predicted channel matrices. It can be observed that the downlink rate decreases as f_dT_s increases or as time elapses with the increase in OFDM symbol index, since the channel becomes more and more outdated. We also plot the case of no prediction, where instead of predicting the channel from 8th-17th OFDM symbol, we just continue to use the ZF matrix computed at the 7th OFDM symbol index over subsequent symbols while precoding. While no prediction performs as well as ARMA(2,2) prediction at lower values of the normalized Doppler frequency f_dT_s, the gain in rate due to prediction increases as f_dT_s increases. Also plotted is the channel update approach where ZF matrix computations are obtained assuming uplink pilot transmissions over all the symbols 1 to 17.

This plot gives us an idea about how often do we need to send uplink pilots without incurring a rate reduction by a certain percentage. For example, the pilots can be sent every 4th OFDM symbol in case of channel prediction and at $f_dT_s = 0.02$ if the system can tolerate a rate reduction by about 8%. In other words, we can send 4 downlink symbols before turning to uplink if we can tolerate a rate reduction by 8%.

Figure 2.5 plots the average downlink rate as a function of the OFDM symbol index for the case when the channel has Bessel ACF while an AR(2) predictor is used to capture the time variations in the channel. As before, observations are obtained over symbols 1 to 7 and channel prediction is performed over symbols 8 to 17. For the case of channel prediction, the ZF precoder computations over 8 to 17 are based on the predicted channel matrices. We also plot the case of no prediction, where instead of predicting the channel from 8th-17th OFDM

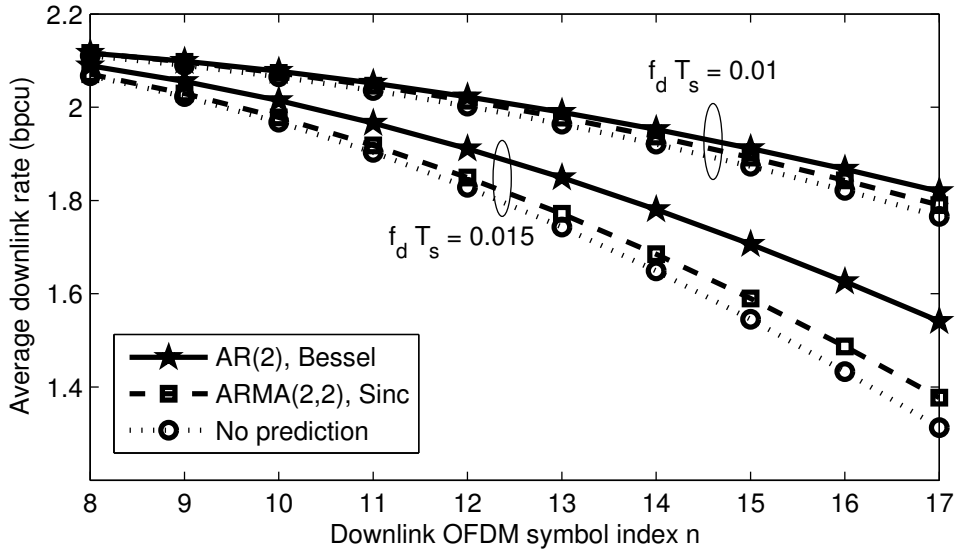


Figure 2.6: Study robustness of predictor on Jakes channel with Bessel ACF ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).

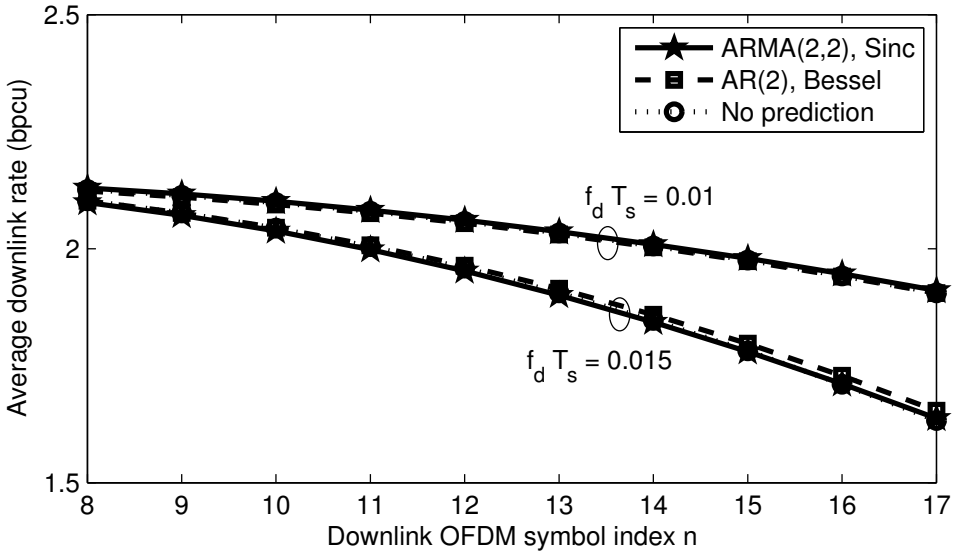
symbol, we just reuse the ZF matrix computed at the 7th OFDM symbol index over subsequent symbols while precoding. Also, plotted is the channel update approach where the entire frame of 17 OFDM symbols has uplink pilots so that the estimated channel matrices are used for ZF matrix computations. This gives the best possible rate. Note that we do not account for the pilot overhead in the rate calculations in the channel update approach. Similar conclusions are obtained as in Figure 2.4.

Figure 2.6 plots the average downlink rate as a function of the OFDM symbol index for the case when the channel has Bessel ACF while different predictors are used to model the time variations of the channel matrix. It can be observed that an AR(2) predictor for which the time-correlation takes the form of a Bessel function captures the dynamics of the Jakes channel slightly better than an ARMA(2,2) predictor for which the time-correlation takes the form of a sinc function. The gain in performance of the AR channel model is higher at higher Doppler spreads. As expected, no prediction gives the worst downlink rate.

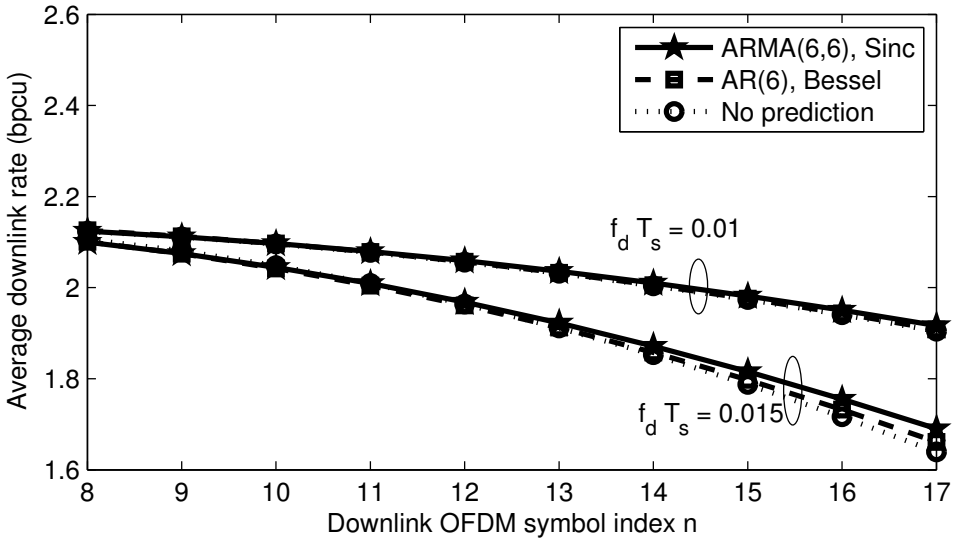
Figure 2.7a plot the average downlink rate as a function of the OFDM symbol index for the case when the channel has a rectangular spectrum and sinc ACF while different predictors are used to model the time variations of the channel matrix. It can be observed that an AR(2) predictor with Bessel ACF is almost as good as an ARMA(2,2) predictor with sinc ACF and in fact slightly better in capturing the dynamics of the channel with rectangular spectrum. As expected, no prediction gives the worst downlink rate. Figure 2.7b plots the same for order 6 predictors and it can be observed that a higher order ARMA(6,6) predictor is better at capturing the time variations of the channel matrix with sinc ACF than an AR(6) model. From these figures, it can be concluded that a channel with sinc autocorrelation is fairly robust to predictor mismatch.

Figure 2.8 plots the average downlink rate vs. the OFDM symbol index for the case when there is a mismatch between the actual Doppler spread of the channel and the Doppler spread with which the predictor is designed. The channel has Bessel ACF and a normalized Doppler frequency of 0.02 while an AR(2) predictor is used. There is a slight reduction in rate in case of a mismatch, both when $f_d T_s$ is smaller or larger than the true value.

Figures 2.9a and 2.9b plot the average downlink rate as a function of the OFDM symbol



(a) Compare predictors of order 2



(b) Compare predictors of order 6

Figure 2.7: Study robustness of predictor on a channel with rectangular spectrum and sinc ACF ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).

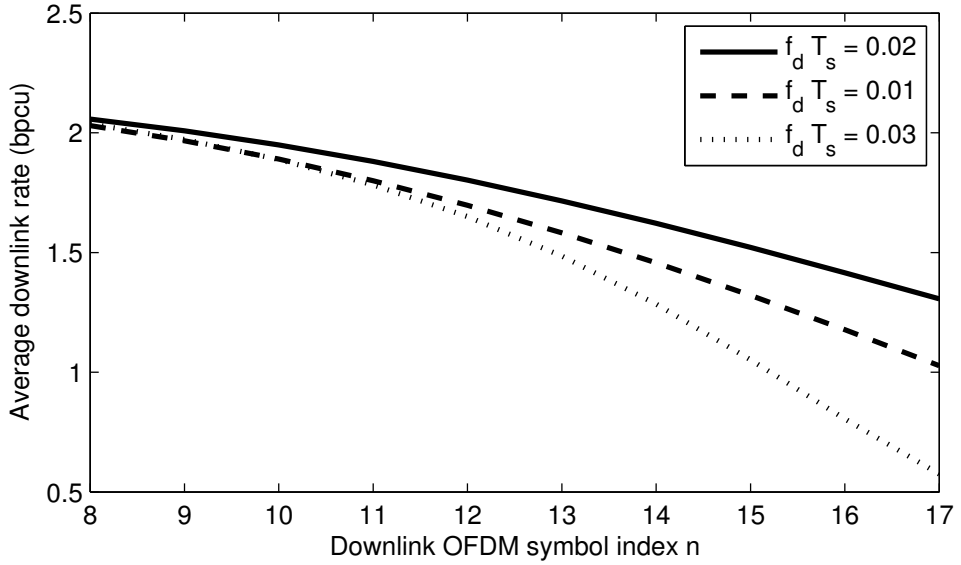


Figure 2.8: Effect of mismatch of f_dT_s, Jakes channel with Bessel ACF and $f_dT_s = 0.02$ ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).

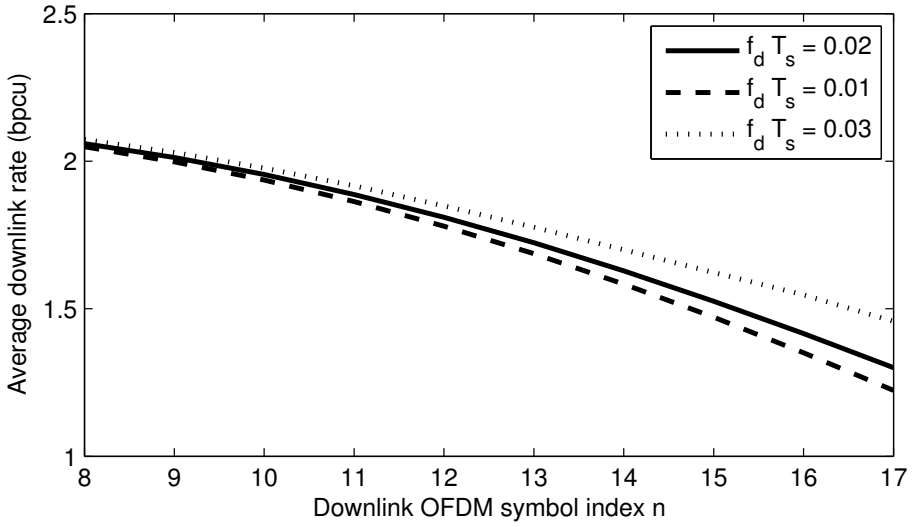
index for the case when there is a mismatch between the Doppler spread of the channel and the Doppler spread with which the predictor is designed for a channel with rectangular spectrum and for ARMA(2,2) and ARMA(6,6) predictors respectively. As can be seen in Figure 2.8, the ARMA predictor seems more robust to f_dT_s mismatch than the AR predictor, since a sinc function has a slower rate of decay compared to a Bessel function. Note that for a channel with rectangular spectrum, the ACF takes the form of a sinc function while that for a Jakes channel it takes the form of a Bessel function. Therefore, the channel decorrelates relatively faster for a Jakes channel than it does for a channel with rectangular spectrum.

Figures 2.10a and 2.10b plot the average downlink rate over the 14th OFDM symbol as a function of the model order for the AR and the ARMA models respectively. It can be observed that the downlink rate increases marginally with the increase in the model order. It is, therefore, justified to use predictor models of order 2 which are computationally less expensive without compromising on the performance.

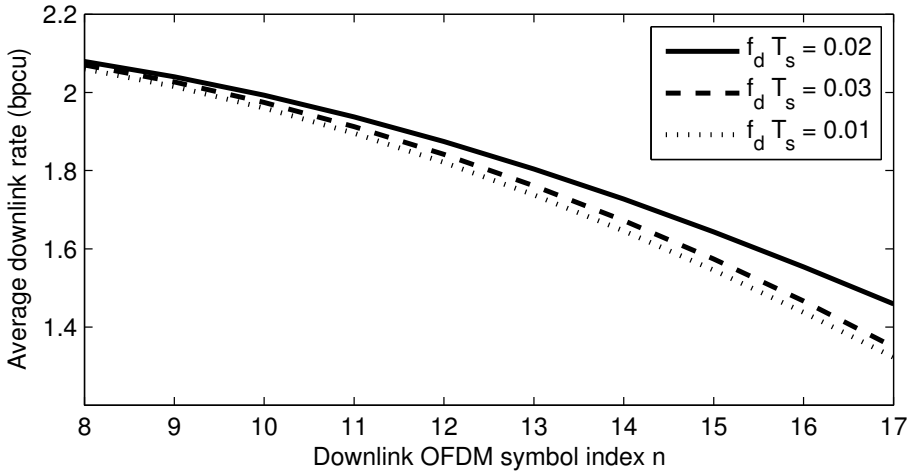
In Figure 2.11, we investigate the tradeoff between increase in rate due to reduced pilot overhead and the decrease in rate due to channel outdatedness. We plot the average uplink rate as a function of the inter-pilot spacing. Note that an inter-pilot spacing of one implies that we have pilots located over every OFDM symbol, an inter-pilot spacing of two implies that we have pilots located every 2nd OFDM symbol and so on. We observe that a larger inter-pilot spacing can be tolerated for smaller Doppler spreads. For example, for a normalized Doppler spread value of 0.01 which corresponds to a user speed of 80 km/h, the optimal inter-pilot spacing can be 3 OFDM symbols, while for a normalized Doppler spread 0.02, the optimal spacing is 1 OFDM symbol.

Conclusions

We investigated how often we need to predict the channel matrix and compute the ZF detector or precoder over time without incurring a significant loss in rate. To this end, we considered different predictors to model the time-variations of the channel. We observed that AR or ARMA prediction gives substantial gains over no prediction particularly at higher Doppler

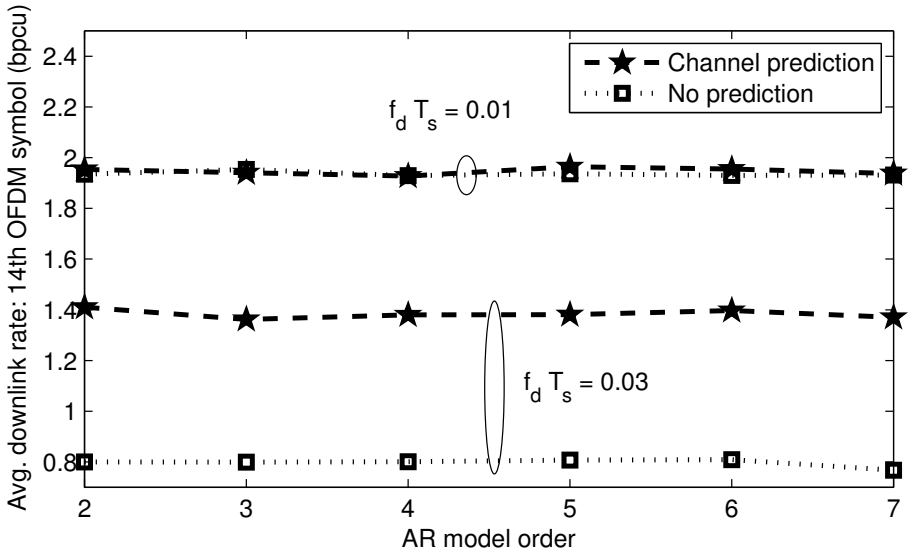


(a) ARMA(2, 2) predictor

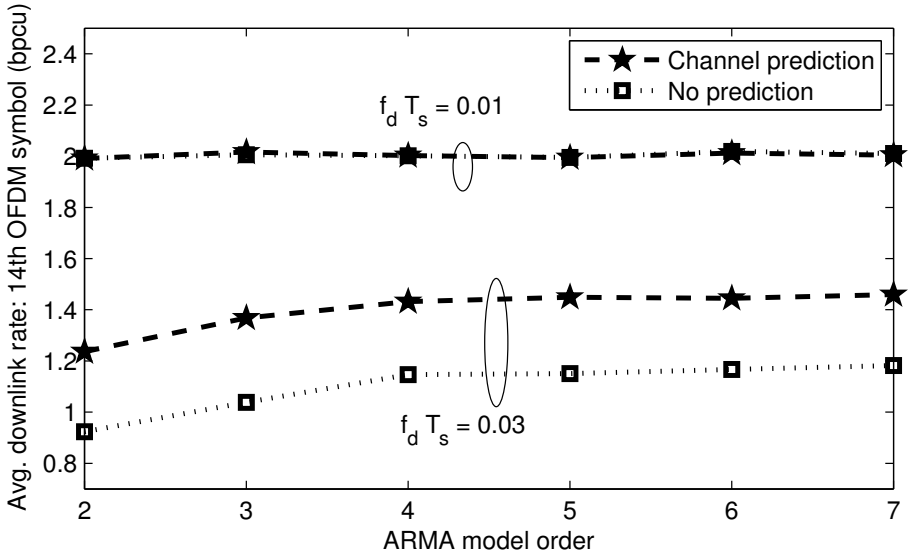


(b) ARMA(6, 6) predictor

Figure 2.9: Effect of mismatch of $f_d T_s$ for a channel with rectangular spectrum and sinc ACF, with $f_d T_s = 0.02$, ($M = 100$, $K = 8$, $L = 8$, $S = 256$, $N = 17$, $p_d = p_k^u = -5$ dB).



(a) Rate vs. AR model order



(b) Rate vs. ARMA model order

Figure 2.10: Rate vs. model order ($M = 100$, $K = 8$, $L = 4$, $S = 256$, $N = 14$, $p_d = p_k^u = -5$ dB).

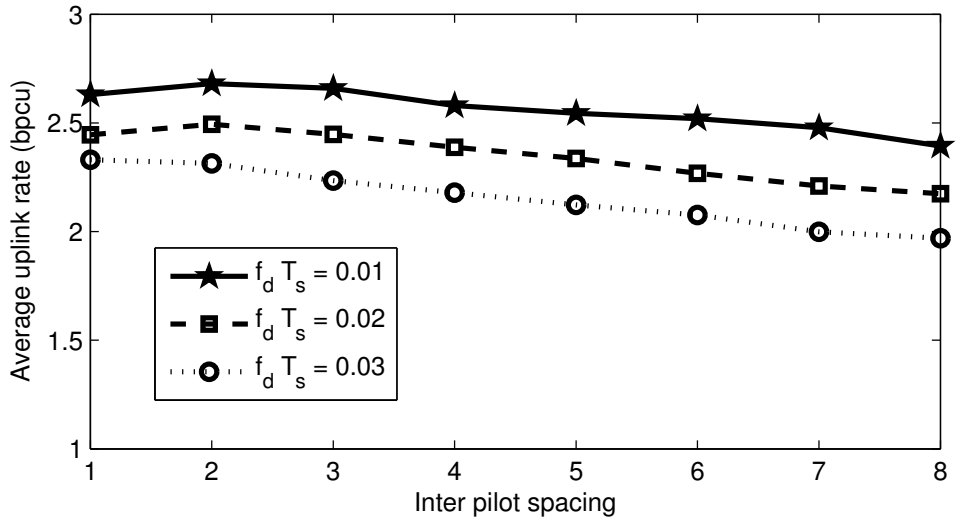


Figure 2.11: Average uplink rate vs inter-pilot spacing, Jakes channel with Bessel ACF, AR(2) predictor ($M = 100$, $K = 8$, $L = 4$, $S = 256$, $N = 100$, and $\rho = -10$ dB).

spread values, while at low Doppler spreads, no prediction does reasonably well and performs poorer than channel prediction only when the channel becomes highly outdated. We also looked at how robust the predictors are to mismatches in the Doppler spreads. We found that for the channel with rectangular spectrum and the ARMA predictor the loss in rate is marginal in the presence of a mismatch because the channel decorrelates slowly over time, while that for the Jakes channel and an AR predictor, the loss can be substantial for higher degree of mismatch.

2.1.2 System Validation of Frequency-domain Interpolation

In this section, we revisit the frequency-domain interpolation problem that was previously addressed in Section 2.4 of Deliverable 3.2 [39]. The purpose is to study the performance-complexity tradeoff for different algorithms, from a system-level perspective, and thereby draw conclusion on their practical feasibility. We simulate the MaMi system performance for different channel estimation and smoothing/interpolation algorithms over the frequency domain and investigate the benefits from the advanced solutions developed in MAMMOET, taking into account the related complexity increase.

Reference Configuration

The baseline scenario offers the lowest channel estimation performance but also the lowest complexity. System parameters are aligned with the scenarios defined in Chapter 4 of Deliverable D4.1. Concerning channel estimation, assuming the number of users, K, is smaller than the channel coherence bandwidth expressed in number of subcarriers, the pilot interval is set to $P = K$ subcarriers. It enables channel estimation for all UEs based on a single uplink OFDM symbol, where UE k sends pilots on subcarriers k, $k + K$, $k + 2K$, etc.

The channel estimated on a given subcarrier is reused over K neighboring subcarriers, in such a way that the same channel matrix and precoder is used over each sub-block of K subcarriers. This implies that blocks of K subcarriers are aligned over all users, such that only one precoder computation every K subcarriers is required. As compared to this baseline solution, any improved channel estimator will have the drawback that the same channel matrix

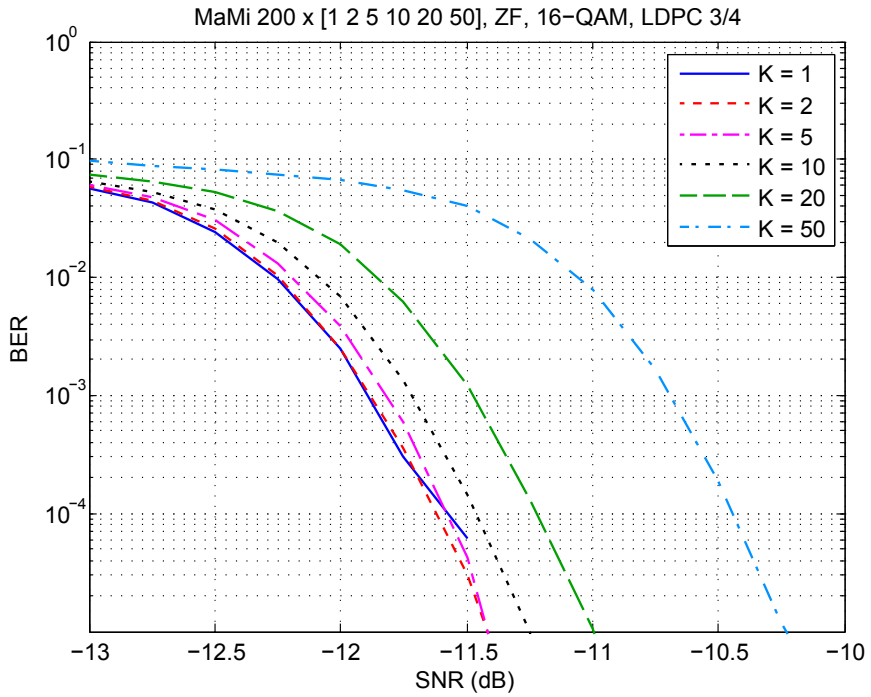


Figure 2.12: Performance of a MaMi system under perfect CSI when increasing the number of users.

and precoder cannot be reused anymore over neighboring subcarriers and hence K times more precoder computations will be required.

When the number of users is relatively low as compared to the coherence bandwidth of the channel, the duration of a channel estimation block could be made larger than K. It has the benefit of accumulating more energy in the estimation phase, hence improving the channel estimate SNR, while the deviation from the actual frequency-selective channel remains limited as soon as the coherence bandwidth is sufficient. The corresponding trade-off is explored in the next section. On the other hand, when the number of users grows too large, a denser pilot scheme can still be used, at the cost of using multiple training OFDM symbols each addressing a subset of users.

Benefits of Interpolation/Smoothing on System Performance

Before simulating the impact of frequency interpolation/smoothing, let us consider the difference between perfect CSI operation and the baseline channel estimator working on a per-subcarrier basis. Figure 2.12 illustrates the progressive degradation under perfect CSI when increasing the number of users. It uses a configuration fixed to M=200 antennas, ZF precoding, 16-QAM, LDPC coding rate 3/4 and time/frequency parameters from 20-MHz LTE. The channel is Rayleigh 72-tap i.i.d. A single-user configuration operates at BER 10^{-5} around -11.5 dB SNR, which corresponds to a required SNR of 11.5 dB for the selected modulation and coding scheme, given the 23-dB gain coming from the 200 antennas. With a few users the degradation is limited to a fraction of dB and reaches 0.5 dB around 20 users, which is consistent with the theory (only 181 degrees of freedom left instead of 200, leading to a loss of $10 \cdot \log_{10}(200/181) = 0.43$ dB in array gain).

Let us consider the baseline channel estimation instead of perfect CSI. When considering the default $P = K$ pilot density, allowing to estimate channels from all users from a single OFDM

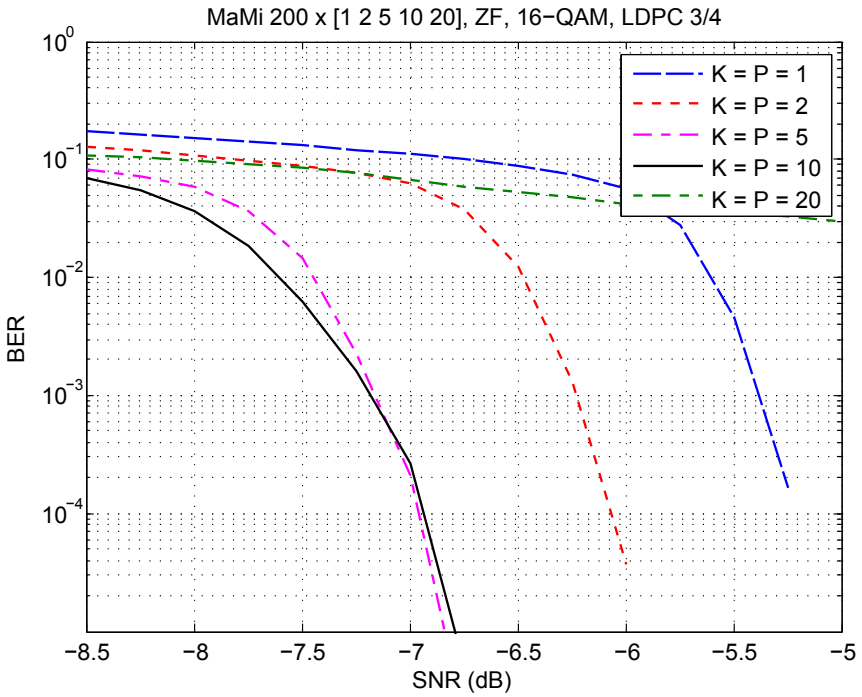


Figure 2.13: Performance of a MaMi system similar to Figure 2.12 but including a baseline per-subcarrier channel estimation with as many pilot sequences as users in the system.

symbol, a trade-off appears between channel coherence and estimation SNR, as can be seen in Figure 2.13. When having few users and hence a dense channel estimation scheme such as $P = 1$ or 2, the frequency-domain representation of the channel has the best expected accuracy with respect to the actual frequency-selective channel. However, individual subcarriers do not benefit from a sufficient power level as UEs have to transmit on almost all subcarriers even during training phase and do not have margin to boost the power on pilot tones. Hence, the channel estimates suffer from a poor SNR which leads to a larger degradation of the operating SNR. On the other hand, when using many users and a large spacing between pilot subcarriers, such as $P = 20$ on Figure 2.13, the channel estimation misses a sufficient representation of the frequency-domain channel characteristics, leading to a complete flooring of the system performance. Hence, for the selected configuration and channel model, supporting 5 or 10 users with the baseline estimation scheme provides the best trade-off between both effects. The degradation with respect to perfect CSI operation is around 4.5 dB.

In order to enable a working 200×20 MaMi configuration, different pilot densities have been tested by reducing P (Figure 2.14). In such scenarios, any configuration denser than $P = 20$ subcarriers implies that multiple OFDM symbols are required in order to perform CSI estimation for all users, e.g., up to 20 symbols when training on every subcarrier ($P = 1$), 10 symbols when training on every second subcarrier ($P = 2$), and so on. Besides a reduced estimation SNR, this excessive resource consumption is an additional reason to avoid performing channel estimation on almost every subcarrier when the channel coherence is sufficient. Whatever the value of P, the performance with the baseline channel estimation algorithm is not satisfactory. The best value ($P = 10$) still leads a degradation of around 7 dB for the 200×20 system, as compared to the perfect CSI case.

Let us introduce a frequency-domain interpolation/smoothing based on Section 2.4 of Deliverable 3.2 [39]. It exploits the constraint on channel delay spread staying below the cyclic

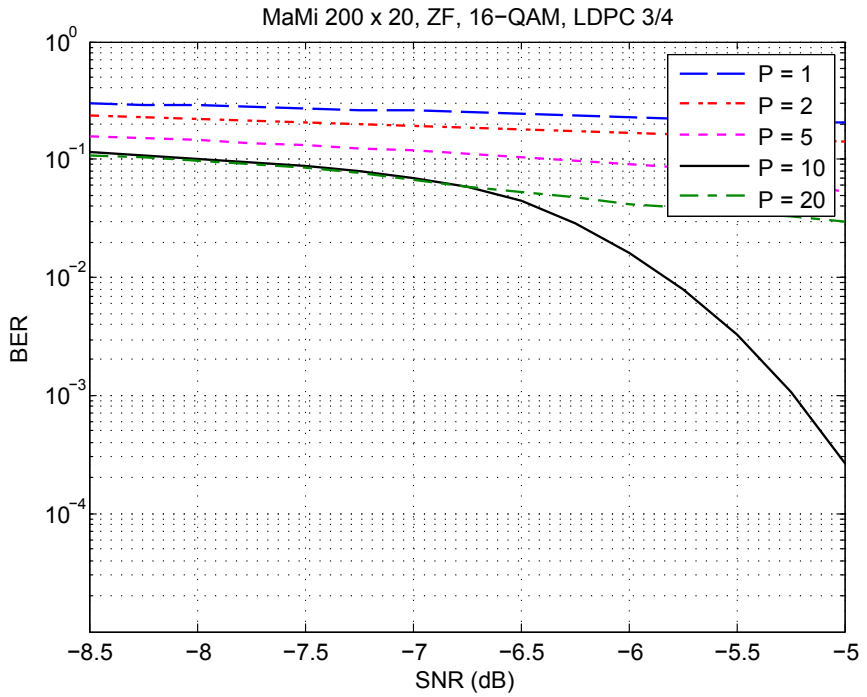


Figure 2.14: Performance of a 200×20 MaMi system including a baseline per-subcarrier channel estimation with different values of P.

prefix by performing an FFT/IFFT-based interpolation. As can be seen on Figure 2.15, there is no more penalty from using dense pilot schemes, because the related noise is averaged in the smoothing/interpolation step. Only $P = 20$ leads to bad performance because the pilot subsampling period is getting larger than the channel coherence bandwidth and starts missing information. Any other value works and leads to a gain of 2 dB as compared to the baseline estimator, although still 4.5 dB away from perfect-CSI performance. A recommendation could be in this case $P = 10$, limiting the training overhead to 2 OFDM symbols.

For fewer users, the use of smoothing (Figure 2.16) also bring significant benefits, allowing configurations with few users ($K = 1$ or 2) to operate only 2 dB away from the perfect CSI case and bringing 1.5 to 2 dB of gain for average configurations ($K = 5$ or 10).

Complexity of the Interpolation Algorithm

Let us assume a system with N subcarriers (1200 out of 2048 in the LTE 20 MHz case) while the channel is estimated for each user from a subset of $S = N/P$ equally-spaced subcarriers spanning the whole band. For instance, assuming a channel coherence over 15 neighboring subcarriers, $S = 80$ covers the full band. This channel coherence is similar to the bound obtained when using a cyclic prefix of 144 for 2048 subcarriers, hence it is a realistic assumption. The precoder interpolation algorithm starts from precoders computed over the S subcarriers and interpolates the precoder in-between, which is simpler than interpolating the channel first and having to compute the precoder on all N subcarriers.

Based on Deliverable 3.2 [39], the interpolation can be implemented by applying an S-point IFFT, zero-padding in order to up-sample by a factor P, and applying an N-point FFT. This has to be performed on each of the $M \times K$ entries of the channel precoder. Hence, a training symbol containing uplink pilots will require MK additional N-point and S-point FFTs, while the baseline processing of each (data or training) symbol requires M N-point FFTs only.

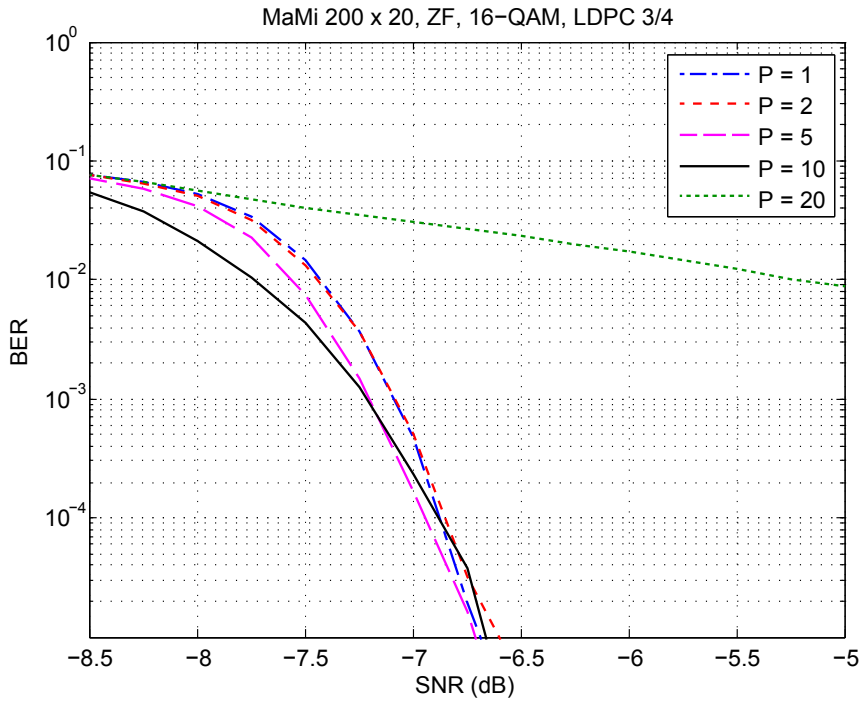


Figure 2.15: Performance of a 200×20 MaMi system similar to Figure 2.14 but including FFT-based frequency-domain smoothing, for different values of P.

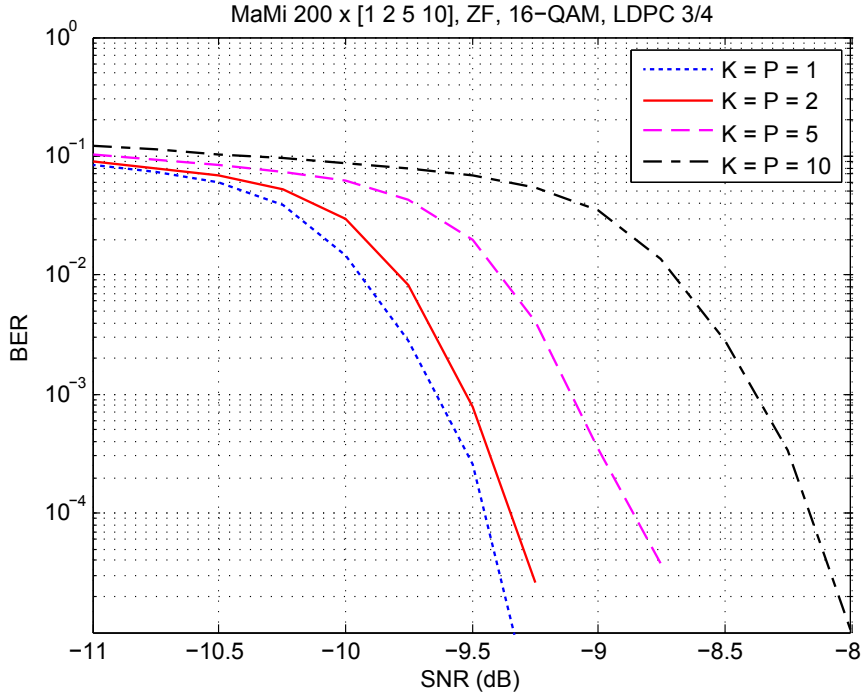


Figure 2.16: Performance of a MaMi system with different numbers of users but including FFT-based frequency-domain smoothing, for $P = K$.

Table 2.1: Relative complexity of different DSP components for a 100×25 scenario.

Phase	Downlink	Uplink	Training	Frame average
Number OFDM symbols	7	5	2	14
Digital front-end [GOPS]	233	233	233	233
Inner modem [GOPS]	180	180	176	179
Interpolator [GOPS]	0	0	1440	206
Outer modem [GOPS]	16	236	0	108

Given that the complexity of an N-point FFT scales as $N\log_2(N)$, we can neglect the role of the S-point FFTs in the overall complexity analysis as $S \ll N$. Based on the assumptions and complexity models presented in Chapter 4 of Deliverable 3.2 [39] and from [15], Table 2.1 illustrates the relative complexity of the different DSP components, in giga complex arithmetic operations per second (GOPS). Digital front-end refers to antenna-level processing (FFT and baseband filter); inner modem refers to user-level processing (precoding); interpolator refers to precoder interpolation; outer modem refers to channel coding.

After averaging over the full frame, the impact of the interpolator is an increase in the total complexity by 40%. This is certainly acceptable from the point of view of digital power consumption, given the low share of digital computations in the total power budget. However, the more critical element could be system dimensioning. Indeed, in order to allow real-time operation, the precoder interpolation should take place during the training phase. This means that even if this interpolator is not activated continuously, the peak complexity of the system during training symbols should be increased by a factor 4.5 when including the interpolator functionality $((1440 + 233 + 176)/(233 + 176))$. This could be considered in view of the total system cost.

If the size and complexity of the required DSP components become an issue, intermediate solutions could offer a relevant trade-off. For instance, interpolating not to every subcarrier but to every second subcarrier would reduce the overall complexity by at least a factor 2 based on the FFT complexity, while still providing a very close approximation of the channel response on every subcarrier and hence a gain close to the simulated 2 dB.

2.2 Receiver Processing Architectures with Low Bit-Width

The large number of transceiver chains required in MaMi base stations make the hardware complexity and cost a challenge that has to be overcome in the MaMi implementation [4]. It has been proposed to build each transceiver chain from low-end hardware to reduce the complexity [10], since MaMi appears to have an inherent robustness towards the distortion caused by non-ideal transceivers. In this section, we quantify what this means for the resolution of the analog-to-digital converters (ADCs) in MaMi deployments. While legacy systems require around 15 bit quantization resolution per transceiver chain, we will show that MaMi systems operate well using a substantially lower resolution. Section 2.2.1 analyzes the impact that low-resolution ADCs have on the channel estimation and achievable rates. Since there is a tradeoff between performance and energy consumption when having low-resolution ADCs, Section 2.2.2 further determines which resolution maximizes the energy efficiency.

2.2.1 MaMi Base Stations with Low-Resolution ADCs

In this section, we perform an information theoretical analysis of a MaMi system with arbitrary ADCs and present an achievable rate, which takes quantization into account, for a linear combiner that uses low-complexity channel estimation. The achievable rate is used to draw the conclusion that ADCs with 3 bits are sufficient to achieve a rate close to that of an unquantized system.

System Model

The uplink transmission from K single-antenna users to a MaMi base station with M antennas is studied. The transmission is based on pulse-amplitude modulation and, for the reception, a matched filter is used for demodulation. It is assumed that the matched filter is implemented as an analog filter and that its output is sampled at symbol rate by an ADC with finite resolution. Because the nonlinear quantization of the ADC comes after the matched filter, the transmission can be studied in symbol-sampled discrete time.

Each user k transmits the signal $\sqrt{P_k}x_k[n]$, which is normalized,

$$\mathbb{E}\big[|x_k[n]|^2\big] = 1, \tag{2.33}$$

so that P_k denotes the transmit power. The channel from user k to antenna m at the base station is described by its impulse response $\sqrt{\beta_k}h_{mk}[\ell]$, which can be factorized into a large-scale fading coefficient β_k and a small-scale fading impulse response $h_{mk}[\ell]$. The large-scale fading varies slowly in comparison to the symbol rate and can be accurately estimated with little overhead by both the user and the base station. It is therefore assumed to be known throughout the system. The small-scale fading, in contrast, is *a priori* unknown to everybody. It is independent across ℓ and follows the power delay profile

$$\sigma_k^2[\ell] \triangleq \mathbb{E}\left[|h_{mk}[\ell]|^2\right], \tag{2.34}$$

however, is assumed to be known. It is also assumed that $\sigma_k^2[\ell] = 0$ for all $\ell \notin [0, \ldots, L{-}1]$. Since variations in received power should be described by the large-scale fading only, the power delay profile is normalized such that

$$\sum_{\ell=0}^{L-1} \sigma_k^2[\ell] = 1, \quad \forall k. \tag{2.35}$$

Base station antenna m receives the signal

$$y_m[n] = \sum_{k=1}^{K} \sqrt{\beta_k P_k} \sum_{\ell=0}^{L-1} h_{mk}[\ell]x_k[n-\ell] + z_m[n]. \tag{2.36}$$

The thermal noise of the receiver $z_m[n]$ is modeled as a white stochastic process, for which $z_m[n] \sim \mathcal{CN}(0, N_0)$. The received power is denoted

$$P_{\text{rx}} \triangleq \mathbb{E}\left[|y_m[n]|^2\right] = \sum_{k=1}^{K} \beta_k P_k + N_0. \tag{2.37}$$

Transmission is assumed to be done with a cyclic prefix in blocks of N symbols. The received signal can than be given in the frequency domain as

$$\mathsf{y}_m[\nu] \triangleq \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y_m[n]e^{-j2\pi n\nu/N} = \sum_{k=1}^{K} \mathsf{h}_{mk}[\nu]\mathsf{x}_k[\nu] + \mathsf{z}_m[\nu], \tag{2.38}$$

The Fourier transforms $\mathsf{x}_k[\nu]$ and $\mathsf{z}_k[\nu]$ of the transmit signal $x_k[n]$ and noise $z_m[n]$ are defined in the same way as $\mathsf{y}_m[\nu]$. The frequency response of the channel is defined as

$$\mathsf{h}_{mk}[\nu] \triangleq \sum_{\ell=0}^{L-1} h_{mk}[\ell] e^{-j2\pi\ell\nu/N}. \tag{2.39}$$

Quantization

The inphase and quadrature signals are assumed to be quantized separately by two identical ADCs with quantization levels given by $\mathcal{Q}_{\Re\mathfrak{e}} \subseteq \mathbb{R}$. The set of quantization points is denoted $\mathcal{Q} \triangleq \{a + jb : a, b \in \mathcal{Q}_{\Re\mathfrak{e}}\}$ and the quantization by

$$[y]_{\mathcal{Q}} \triangleq \arg\min_{q\in\mathcal{Q}} |y - q| . \tag{2.40}$$

To adjust the input signal to the dynamic range of the ADC, an automatic gain control scales the input power by A. The ADC outputs

$$q_m[n] \triangleq \left[\sqrt{A} y_m[n]\right]_{\mathcal{Q}} . \tag{2.41}$$

To analyze the effect of the quantization, the quantized signal is partitioned into one part $\rho y_m[n]$ that is correlated to the transmit signal and one part $e_m[n]$ that is uncorrelated:

$$q_m[n] = \rho y_m[n] + e_m[n] \tag{2.42}$$

where the constant ρ and the variance of the uncorrelated part can be derived through the orthogonality principle:

$$\rho = \frac{\mathbb{E}\left[q_m[n] y_m^*[n]\right]}{\mathbb{E}\left[|y_m[n]|^2\right]}, \tag{2.43}$$

$$\mathbb{E}\left[|e_m[n]|^2\right] = \mathbb{E}\left[|q_m[n]|^2\right] - \frac{\left|\mathbb{E}\left[q_m[n] y_m^*[n]\right]\right|^2}{\mathbb{E}\left[|y_m[n]|^2\right]}. \tag{2.44}$$

The normalized mean-square error (MSE) of the quantization is denoted by

$$Q \triangleq \frac{1}{|\rho|^2} \mathbb{E}\left[|e_m[n]|^2\right] \tag{2.45}$$

$$= P_{\text{rx}} \left(\frac{\mathbb{E}\left[|q_m[n]|^2\right] \mathbb{E}\left[|y_m[n]|^2\right]}{\left|\mathbb{E}\left[q_m[n] y_m^*[n]\right]\right|^2} - 1 \right). \tag{2.46}$$

An ADC with b-bit resolution has $|\mathcal{Q}_{\Re\mathfrak{e}}| = 2^b$ quantization levels. In [41], the quantization levels that minimize the MSE for a Gaussian input signal with unit variance are derived numerically for 1–5 bit ADCs, both with arbitrarily and uniformly spaced quantization levels. The normalized MSE of the quantization has been computed numerically and is given in Table 2.2 for the optimized quantizers. To obtain the MSE in Table 2.2 with the quantization levels from [41], the input power has to be unity and the automatic gain control $A = A^\star \triangleq 1/P_{\text{rx}}$. Figure 2.17 shows how the quantization MSE in a four-bit ADC changes with imperfect gain control. Even if the gain control varies between -8 dB and 5 dB from the optimal value, the MSE is still better than that of a three-bit ADC.

Table 2.2: Normalized quantization mean square-error Q/P_{rx}

resolution b	1	2	3	4	5
optimal levels	0.5708	0.1331	0.03576	0.009573	0.002492
uniform levels	0.5708	0.1349	0.03889	0.01166	0.003506

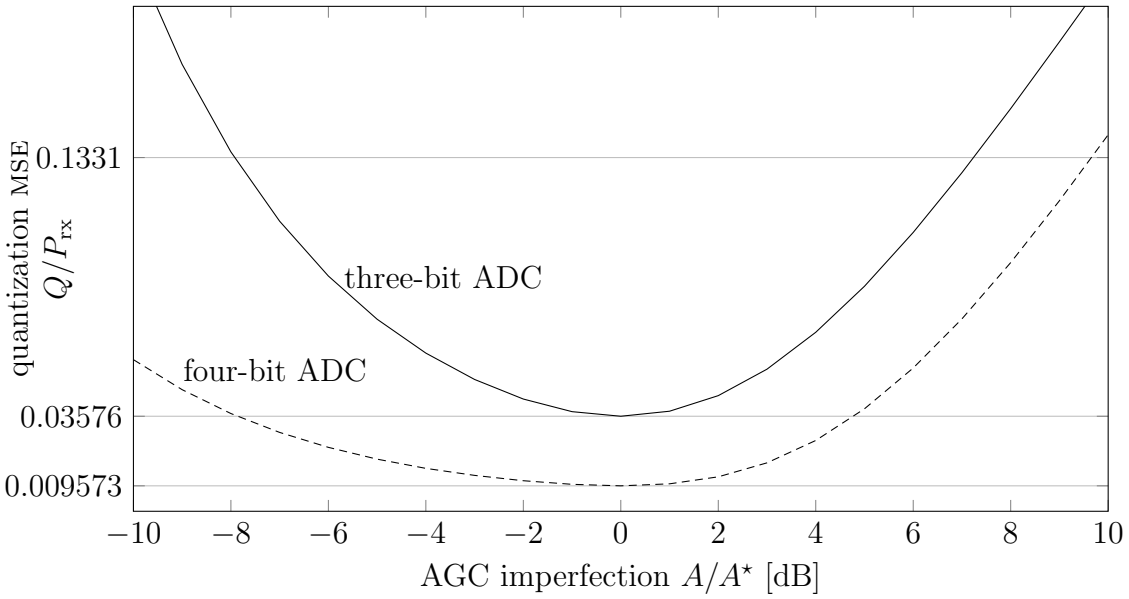


Figure 2.17: Quantization MSE for optimal four-bit ADC with imperfect AGC.

Channel Estimation

Channel estimation is done by receiving $N = N_{\text{p}}$-symbol long orthogonal pilots from the users, i.e., pilots $x_k[n]$ such that:

$$\sum_{n=0}^{N_{\text{p}}-1} x_k[n]x_{k'}^*[n+\ell] = \begin{cases} N_{\text{p}}, & \text{if } k = k', \ell = 0, \\ 0, & \text{if } k \neq k', \ell = 1, \ldots, L-1, \end{cases} \tag{2.47}$$

where the indices are taken modulo N_{p}. To fulfill (2.47), $N_{\text{p}} \geq KL$. We will call the factor of extra pilots $\mu \triangleq N_{\text{p}}/(KL)$ the *pilot excess factor*. As remarked upon in [42], not all sequences fulfilling (2.47) result in the same performance. Here we use the pilots proposed in [42]. Using (2.42) and (2.47), an observation of the channel is obtained by correlation:

$$r_{mk}[\ell] = \frac{1}{\rho\sqrt{N_{\text{p}}}} \sum_{n=0}^{N_{\text{p}}-1} q_m[n]x_k^*[n+\ell] \tag{2.48}$$

$$= \sqrt{\beta_k P_k N_{\text{p}}} h_{mk}[\ell] + e'_{mk}[\ell] + z'_{mk}[\ell], \tag{2.49}$$

where

$$e'_{mk}[\ell] \triangleq \frac{1}{\rho\sqrt{N_{\text{p}}}} \sum_{n=0}^{N_{\text{p}}-1} e_m[n]x_k^*[n+\ell], \tag{2.50}$$

$$z'_{mk}[\ell] \triangleq \frac{1}{\sqrt{N_{\text{p}}}} \sum_{n=0}^{N_{\text{p}}-1} z_m[n]x_k^*[n+\ell] \sim \mathcal{CN}(0, N_0). \tag{2.51}$$

The linear minimum MSE estimate of the frequency response of the channel is thus

$$\hat{h}_{mk}[\nu] = \sum_{\ell=0}^{L-1} \frac{\sqrt{\beta_k P_k}\sigma_k^2[\ell]}{\beta_k P_k N_{\mathrm{p}} \sigma_k^2[\ell] + Q + N_0} r_{mk}[\ell] e^{-j2\pi\ell\nu/N} \tag{2.52}$$

and the error $\epsilon_{mk}[\nu] \triangleq \hat{h}_{mk}[\nu] - h_{mk}[\nu]$ has the variance $1 - c_k$, where the channel estimation variance is given by

$$c_k \triangleq \mathbb{E}\left[|\hat{h}_{mk}[\nu]|^2\right] = \sum_{\ell=0}^{L-1} \frac{\sigma_k^4[\ell]\beta_k P_k N_{\mathrm{p}}}{\sigma_k^2[\ell]\beta_k P_k N_{\mathrm{p}} + Q + N_0}. \tag{2.53}$$

Figure 2.18 shows the channel estimation variance. A resolution of 2 bits is enough to obtain a channel estimation variance only 0.5 dB worse than in an unquantized system. With a resolution of 3 bits or higher, the channel estimation variance is practically the same as that of the unquantized system. Increasing the pilot length, increases the channel estimation variance in all systems. The improvement is, however, the largest when going from $\mu = 1$ to $\mu = 2$; thereafter the improvement gets smaller.

Data Transmission

The uplink data is transmitted in a block of length $N = N_{\mathrm{u}}$, which is separated from the pilot block in time. The received signal is processed by a linear combiner and an estimate of the transmitted signal is obtained by

$$\hat{x}_k[\nu] = \frac{1}{\rho} \sum_{m=1}^{M} w_{mk}[\nu] q_m[\nu], \tag{2.54}$$

where the Fourier transform $q_m[\nu]$ of $q_m[n]$ is defined in the same way as $y_m[\nu]$ in (2.38) and the combiner weights $w_{mk}[\nu]$ are chosen as functions of the channel estimate. For example, the MR and ZF combiners can be used.

If we code over many channel realizations, an achievable rate, independent of ν, is given by [42]:

$$R_k = \log_2\left(1 + \frac{\left|\mathbb{E}\left[\hat{x}_k^*[\nu] x_k[\nu]\right]\right|^2}{\mathbb{E}\left[|\hat{x}_k[\nu]|^2\right] - \left|\mathbb{E}\left[\hat{x}_k^*[\nu] x_k[\nu]\right]\right|^2}\right). \tag{2.55}$$

To compute the expected values in (2.55), the estimate of the transmit signal in (2.54) can be

expanded by using the relation in (2.42) and writing the channel as $h_{mk}[\nu] = \hat{h}_{mk}[\nu] - \epsilon_{mk}[\nu]$:

$$\begin{aligned}\hat{x}_k[\nu] = {} & x_k[\nu]\sqrt{\beta_k P_k}\sum_{m=1}^{M}\mathbb{E}\left[w_{mk}[\nu]\hat{h}_{mk}[\nu]\right] \\ & + x_k[\nu]\sqrt{\beta_k P_k}\sum_{m=1}^{M}\left(w_{mk}[\nu]\hat{h}_{mk}[\nu] - \mathbb{E}\left[w_{mk}[\nu]\hat{h}_{mk}[\nu]\right]\right) \\ & + \sum_{k'\neq k} x_{k'}[\nu]\sqrt{\beta_{k'} P_{k'}}\sum_{m=1}^{M} w_{mk}[\nu]\hat{h}_{mk'}[\nu] \\ & - \sum_{k'=1}^{K} x_{k'}[\nu]\sqrt{\beta_{k'} P_{k'}}\sum_{m=1}^{M} w_{mk}[\nu]\epsilon_{mk}[\nu] \\ & + \sum_{m=1}^{M} w_{mk}[\nu]z_m[\nu] + \frac{1}{\rho}\sum_{m=1}^{M} w_{mk}[\nu]e_m[\nu], \end{aligned} \tag{2.56}$$

where the Fourier transform $e_m[\nu]$ of $e_m[n]$ is defined as in (2.38). Note that only the first term is correlated to the desired signal. By assuming that the channel is i.i.d. Rayleigh fading, it can be shown that the other terms in (2.56)—channel gain uncertainty, interference, channel estimation error, thermal noise, quantization error—are mutually uncorrelated and the variance of each term can be evaluated. In [42], for example, it is shown, for one-bit ADCs, that the variance of the last term in (2.56) asymptotically equals

$$\mathbb{E}\left[\left|\frac{1}{\rho}\sum_{m=1}^{M} w_{mk}[\nu]e_m[\nu]\right|^2\right] \to Q, \quad L \to \infty, \tag{2.57}$$

if the combiner is normalized such that $\sum_{m=1}^{M}\mathbb{E}\left[\left|w_{mk}[\nu]\right|^2\right] = 1$, which will be assumed here. This can be generalized to general quantization in a similar way. The rate in (2.55) can then be written as

$$R_k \to \log_2\left(1 + \frac{\beta_k P_k c_k G}{\sum_{k'=1}^{K}\beta_{k'}P_{k'}(1 - c_{k'}(1-I)) + Q + N_0}\right), \tag{2.58}$$

as $L \to \infty$, where the *array gain* and *interference* terms are defined as

$$G \triangleq \left|\sum_{m=1}^{M}\mathbb{E}\left[w_{mk}[\nu]\hat{h}_{mk}[\nu]\right]\right|^2, \tag{2.59}$$

$$I \triangleq \mathrm{Var}\left(\sum_{m=1}^{M} w_{mk}[\nu]\hat{h}_{mk'}[\nu]\right), \tag{2.60}$$

where

$$G = \begin{cases} M \\ M-K \end{cases}, \quad I = \begin{cases} 1, & \text{for MR combining,} \\ 0, & \text{for ZF combining.} \end{cases} \tag{2.61}$$

It is shown in [42] that the limit in (2.58) can approximate the rate with negligible error also for practical delay spreads L. The approximation can even be good for some frequency-flat channels ($L = 1$) when the received power $\sum_{k=1}^{K}\beta_k P_k$ is small relative to the noise power N_0 or when the number of users is large and there is no dominant user, i.e., no user k for which

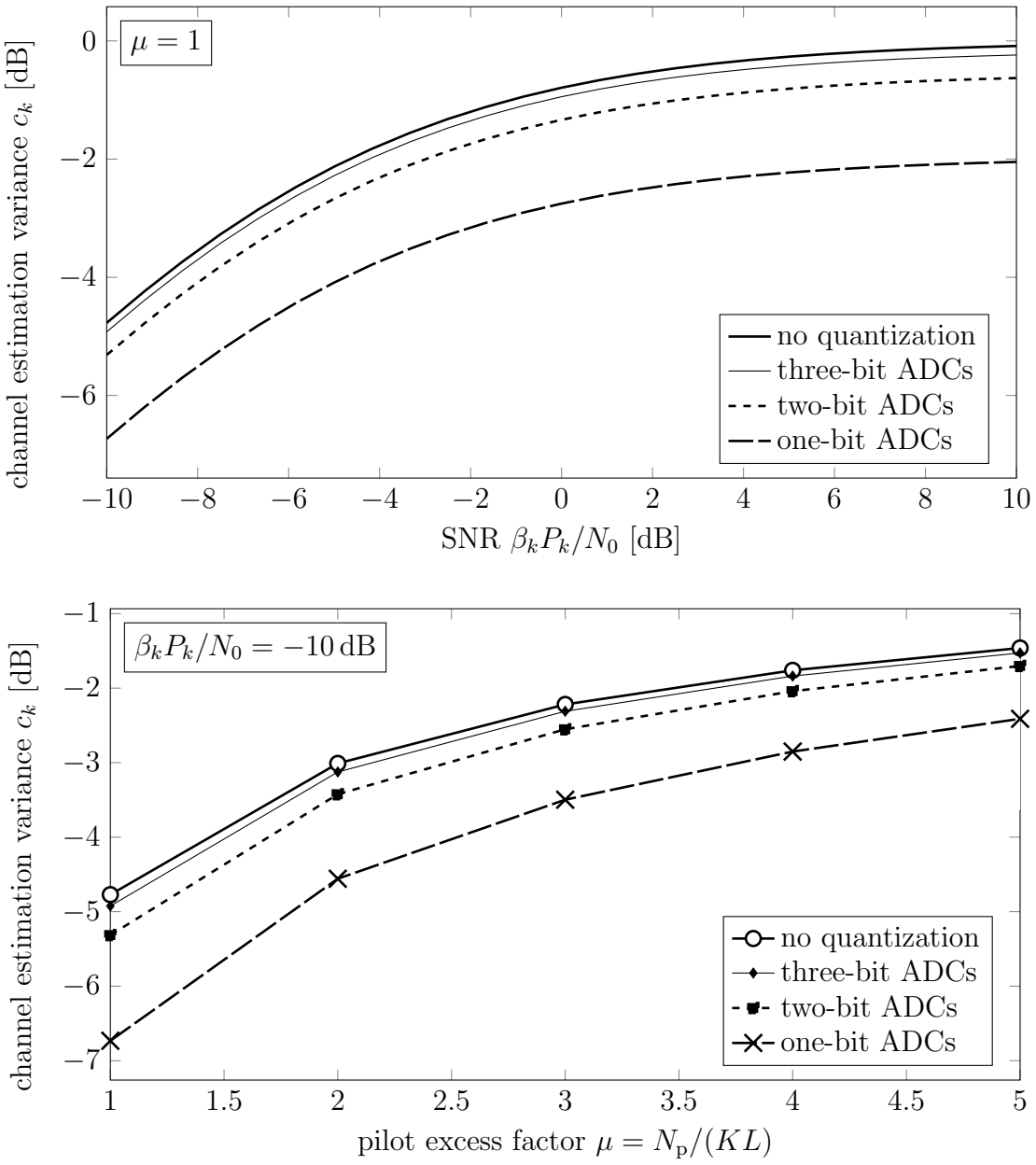
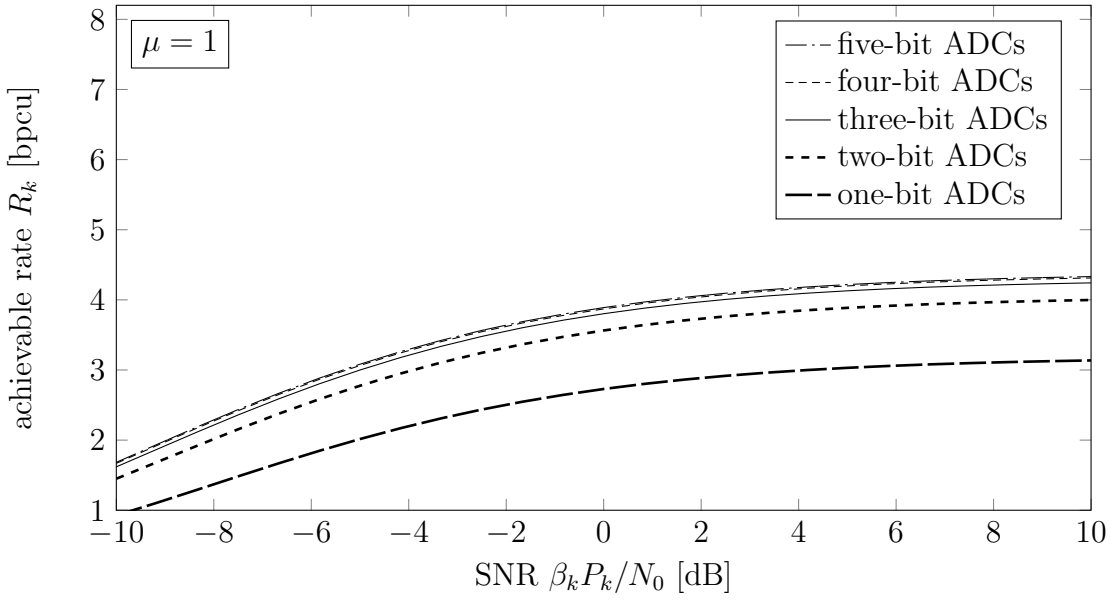


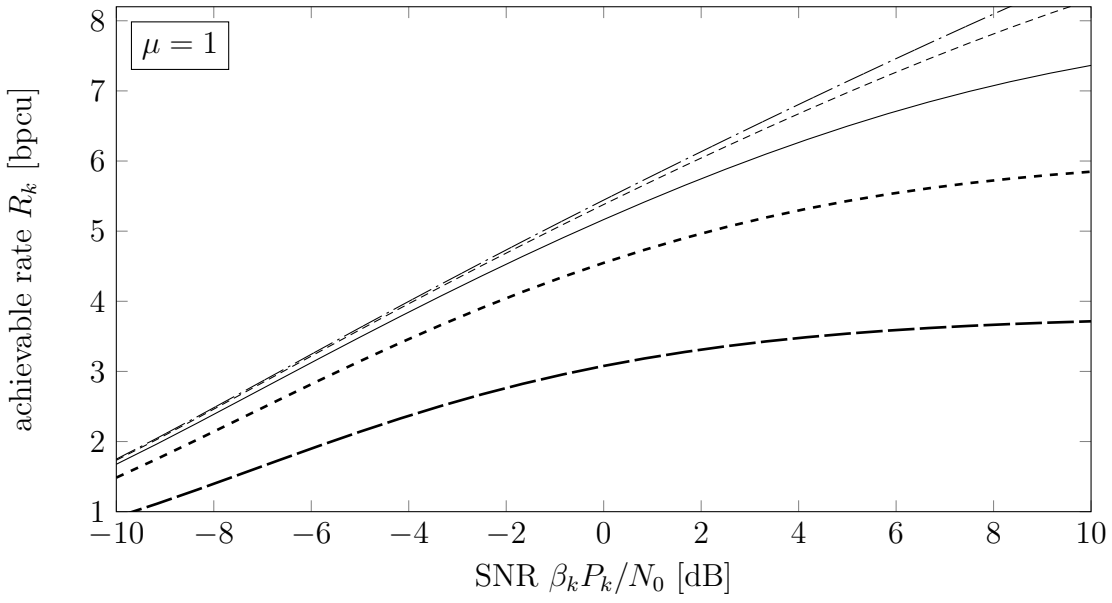
Figure 2.18: The channel estimation variance with 5 users and a uniform power delay profile $\sigma_k^2[\ell] = 1/L$, for all k, ℓ, and with equal received power from all users $\beta_k P_k = \beta_1 P_1$, for all k. The optimal quantization levels derived in [41] are used. Only integer pilot excess factors are considered.

$\beta_k P_k \gg \sum_{k' \neq k} \beta_{k'} P_{k'}$. For general frequency-flat channels, however, it is *not* true that the quantization error variance vanishes with increasing number of antennas, as it does for large L in (2.58); this seems to be overlooked in some of the literature [19, 35, 36, 62].

The rate R_k is plotted in Figure 2.19 for MR and ZF combining. The transmit powers are allocated proportionally to $1/\beta_k$ and channel estimation is done with $N_\mathrm{p} = KL$ pilots, i.e., the pilot excess factor $\mu = 1$. It can be seen that low-resolution ADCs cause very little performance degradation at spectral efficiencies below 4 bpcu. One-bit ADCs deliver approximately 40 % lower rates than the equivalent unquantized system and the performance degradation becomes practically negligible with ADCs with as few as 3 bit resolution. Assuming that the power dissipation in an ADC is proportional to 2^b, the use of one-bit ADCs thus reduces the



(a) MR combining



(b) ZF combining

Figure 2.19: Rate of a system with 100 antennas and 10 users, where the power is proportional to $1/\beta_k$ and training is done with $N_{\mathrm{p}} = KL$ pilots. The channel is i.i.d. Rayleigh fading with uniform power delay profile $h_{mk}[\ell] \sim \mathcal{CN}(0, 1/L)$. The optimal quantization levels derived in [41] are used.

ADC power consumption by approximately 6 dB at the price of 40 % performance degradation compared to the use of three-bit ADCs, which deliver almost all the performance of an unquantized system.

In [25], it is pointed out that low-resolution ADCs create a *near–far* problem, where users with relatively weak received power drown in the interference from stronger users. This is illustrated with a ZF combiner in Figure 2.20, where it can be seen how the performance of the weak users degrades if there is a stronger user in the system. Note that the performance degrades also in the unquantized system, where the imperfect channel estimates prevent perfect

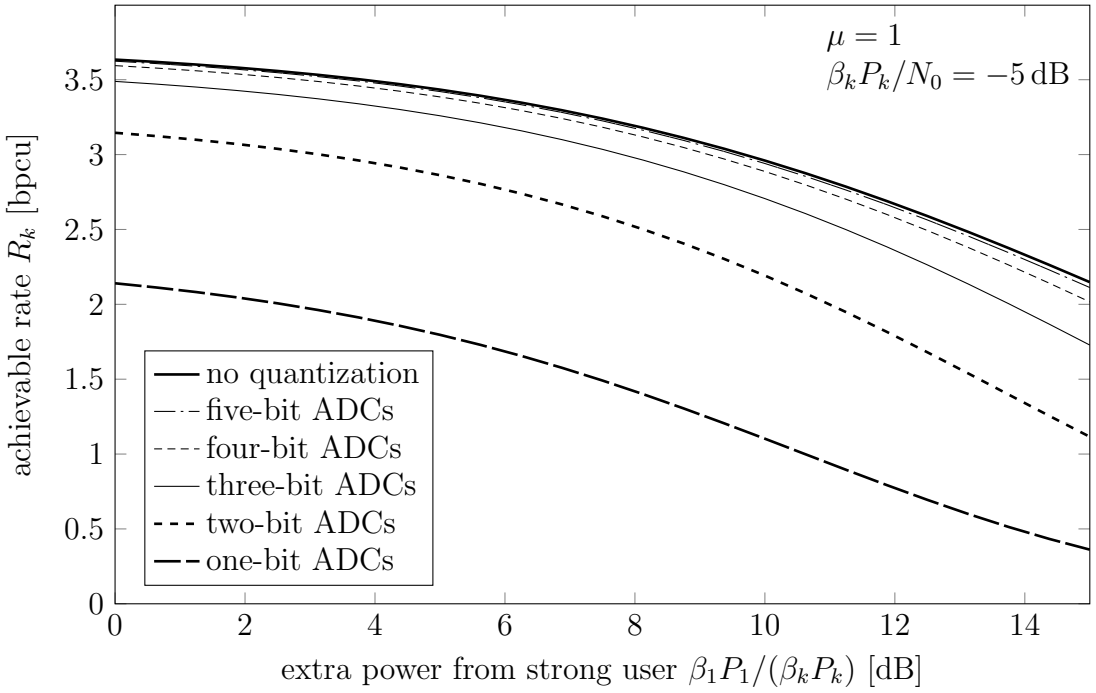


Figure 2.20: Per-user rate R_k for users $k = 2, \ldots, K$ when user $k = 1$ has a different receive SNR. The system has 100 antennas and $K = 10$ users, the channel is i.i.d. Rayleigh with uniform power delay profile and is estimated with $N_\text{p} = KL$ pilots. The optimal quantization levels derived in [41] are used.

suppression of the interference from the strong user. In the quantized systems, there is a second cause of the performance degradation: With quantization, the pilots are no longer perfectly orthogonal and the quality of the channel estimates is negatively affected by interference from the strong user. This effect can be seen in (2.53), where Q scales with the received power P_rx and thus with the power of the interferer.

Figure 2.20, however, shows that the near–far problem does not become prominent until the received power from the strong user is around 10 dB higher than that of the weak users, where the data rate is degraded by approximately 15 % in the unquantized system. The degradation is larger in the quantized systems but the additional degradation due to quantization is almost negligible when the resolution is 3 bits or higher. With one-bit ADCs and one strong user with 10 dB larger received power, the degradation of the data rate increases to almost 50 %. Power control among users, however, can eliminate the near–far problem altogether, but at the cost of reducing the flexibility to use power control to optimize the system performance. Power control that eliminates received power difference is suitable for maximizing the minimum rate in the cell, but less suitable for maximizing the sum rate.

Conclusion

We have derived an achievable rate for a single-cell MaMi system that takes quantization into account. The derived rate shows that ADCs with as low resolution as 3 bits can be used with negligible performance loss compared to an unquantized system, also with interference from stronger users. For example, with three-bit ADCs, the data rate is decreased by 4 % at spectral efficiency of 3.5 bpcu in a system with 100 antennas that serves 10 users. It also shows that four-bit ADCs can be used to accommodate for imperfect automatic gain control—imperfections up to 5 dB still result in better performance than the three-bit ADCs. One-bit ADCs can be built

from a single comparator and do not need a complex gain control (which ADCs with more than one-bit resolution need), which simplifies the hardware design of the base station receiver and reduce its power consumption. The derived rate, however, shows that one-bit ADCs lead to a significant rate reduction. For example, one-bit ADCs lead to a 40 % rate reduction in a system with 100 antennas that serves 10 users at spectral efficiencies of 3.5 bpcu. In the light of the good performance of three-bit ADCs, whose power consumption should already be small in comparison to other hardware components, the primary reason for the use of one-bit ADCs would be the simplified hardware design, not the lower power consumption.

2.2.2 Optimized Bit-Width for Energy Efficiency

In this section, we perform a parametric energy efficiency analysis of the MaMi uplink for the entire base station receiver system with varying ADC resolutions. The analysis shows that, for a wide variety of system parameters, ADCs with intermediate bit resolutions (4 - 10 bits) are optimum in the energy efficiency sense, and that using very low bit resolutions actually results in degradation of energy efficiency.

Energy Efficiency Metric

Energy efficiency η, as a function of ADC bit resolution b, for the uplink of a MaMi system is defined as

$$\eta(b) = \frac{C(b)}{P_c(b)} \quad \left[\frac{\text{bit}}{\text{Joule}}\right], \tag{2.62}$$

where C [bit/s] is the uplink system sum rate and P_c [W] is the total power consumption of the MaMi base station (ADCs together with all other receiver blocks, analog and digital).

Dependencies of sum rate and power consumption on ADC resolution b need to be resolved separately. To this end, we first turn to finding an appropriate model for describing the effects that ADCs have on system performance.

ADC Performance Modeling

The ADC used in this analysis is instantaneous, memoryless and uniform with a finite number of quantization levels ($N_q = 2^b$ in total). Additionally, sampling is assumed to be performed at Nyquist rate. Although uniform quantization is not optimum in the MMSE sense (unless the ADC input is uniformly distributed), it was nevertheless chosen because it is both close to hardware implementation reality [22] and gives way to simple and tractable modeling.

After sampling, the ADC performs a nonlinear mapping of values from $\mathbb{R}$ to a discrete set of quantization levels, resulting in distortion. The nature of this distortion is twofold:

- If the amplitude of input discrete-time signal x is larger than some predefined overload level X_{ol}, the sample is represented by one of the "outer" quantization levels, resulting in *overload distortion*.
- If $|x| < X_{ol}$, samples are rounded to the nearest quantization level, yielding *granular noise*.

In practical systems, an ADC is usually preceded by an automatic gain control (AGC) variable gain amplifier that is used to conveniently set the dynamic range of the input signal to the ADC. The primary purpose of AGC is to minimize *overload distortion*. A welcome consequence of a properly controlled dynamic range of the input signal is a particularly convenient model for

the ADC distortion. Namely, it was shown in [56] that, for a uniform quantizer with normally distributed x, the output can be modeled as

$$y = x + q, \tag{2.63}$$

where the additive noise term q can very well be approximated as being

- uniformly distributed,
- uncorrelated with the input,
- white.

Additionally, the variance of q can be shown to be

$$\mathbb{E}\{q^2\} = \frac{1}{3}\, X_{ol}^2\, 2^{-2b}. \tag{2.64}$$

This commonly used model is usually referred to as the pseudoquantization noise (PQN) model. The quality of the approximation in the PQN model is determined by properly setting the dynamic range of x.

A commonly used design parameter for the AGC is the *input backoff*, defined as

$$\mu = \frac{X_{ol}^2}{\mathbb{E}\{x^2\}}. \tag{2.65}$$

In this work, μ is set so that the variance of overload distortion is 20 dB below the variance of granular noise. Corresponding values of μ, for $X_{ol} = 1$ and bit resolutions $b \in [2, 25]$, are shown in Figure 2.21, together with a linear fit (chord) that can be alternatively used in place of the actual values.

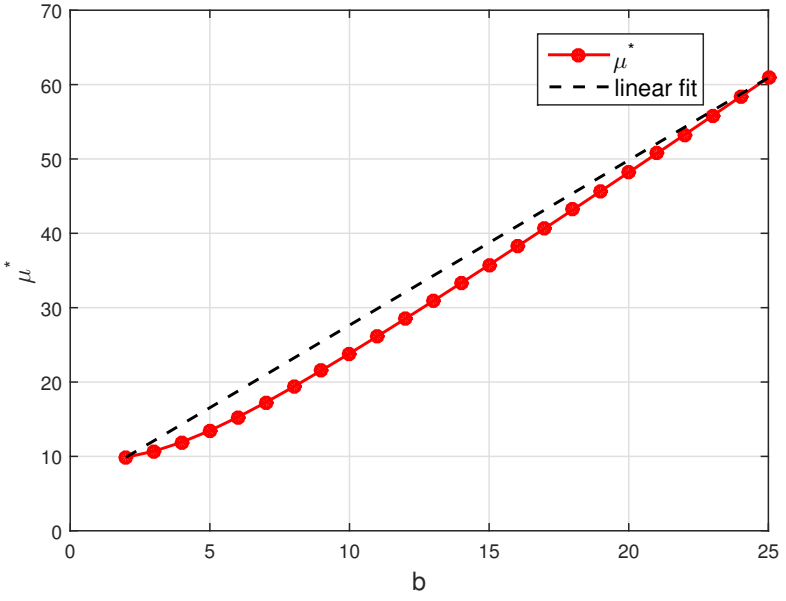


Figure 2.21: Input backoff μ values that result in overload distortion power 20 dBs below granular noise power.

One metric—correlation between x and q–.is important for the subsequent analysis. It was found that, using the presented values of μ for the AGC, the crosscorrelation coefficent between

x and q was always below 0.018, which essentially means that in this setup (uniform ADC with Gaussian inputs and a properly set up AGC), x and q can be considered *uncorrelated*. This result will prove to be important in the system model analysis.

System Model and Sum rate Calculation

This work analyzes the following setup:

- Uplink of a single-cell MaMi system with M antennas and K users;
- Uncorrelated Rayleigh block fading over T symbols;
- Channel estimation is performed using orthogonal pilot sequences of length τ in the uplink. Estimation is performed using the least-squares (LS) approach;
- Channel estimates used for linear receiver processing. Maximum ratio (MR) and zero-forcing (ZF) combining receivers are considered.

A system model of the uplink, where ADCs are substituted by quantization noise sources following the PQN model and AGCs precede ADCs, is illustrated in Figure 2.22.

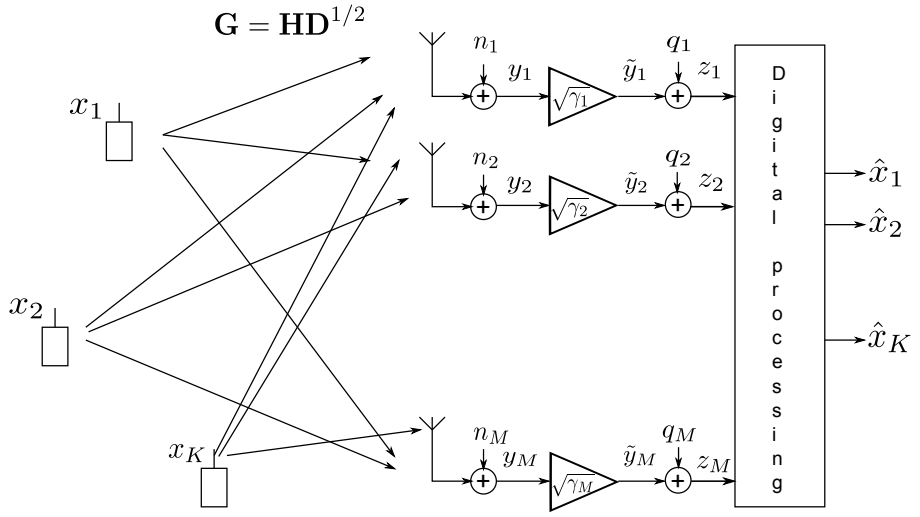


Figure 2.22: Uplink system model with quantization noise.

User k sends a data symbol x_k. User symbols are collected in a vector $\boldsymbol{x} = (x_1\ x_2\ \ldots\ x_k)^T$, with $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] = \boldsymbol{I}_K$. Single-carrier, narrowband transmission is assumed, and thus the propagation channel is represented by the $M \times K$ matrix $\boldsymbol{G} = \boldsymbol{H}\boldsymbol{D}^{1/2}$, where the $M \times K$ $\boldsymbol{H}$ contains small-scale fading coefficients. The elements of $\boldsymbol{H}$ are zero-mean complex Gaussian distributed with unit variance.

The $K \times K$ matrix $\boldsymbol{D}^{1/2}$ is a diagonal matrix of combined amplitude path gains and large-scale fading. The (m, k) element of $\boldsymbol{G}$ can be written as $g_{mk} = h_{mk}\sqrt{\beta_k}$, with h_{mk} being the narrowband small-scale fading coefficient between the kth user and mth antenna and β_k the *power* path gain and large-scale fading, taken jointly. It should be noted that, if some uplink power control is employed, its effects will also be modeled by β. In the case of ideal uplink power control, all $\beta_k = 1$.

Assuming that every user transmits with equal transmit power p_u, the signal model at the receive antennas is

$$\boldsymbol{y} = \sqrt{p_u}\boldsymbol{G}\boldsymbol{x} + \boldsymbol{n} = \sqrt{p_u}\boldsymbol{H}\boldsymbol{D}^{1/2}\boldsymbol{x} + \boldsymbol{n}, \tag{2.66}$$

where $\boldsymbol{n}$ is the vector of input-referred thermal noise at each antenna: $\boldsymbol{n} = (n_1\ n_2\ \ldots\ n_m)^T$, with thermal noise powers at each antenna assumed equal with value p_n.

The received signal y_i will experience variations of average power due to LS and fading, and its power, averaged over both small-scale fading and large-scale fading is determined so that optimum gains of AGCs can be determined. By doing this, the optimum gain of AGC in receiver branch i is found as

$$\gamma_i = \frac{2}{\mu^* \left(p_u \sum_{k=1}^{K} \beta_k + p_n \right)}. \tag{2.67}$$

Amplitude AGC gains $\sqrt{\gamma_i}$ can be conveniently collected in a diagonal matrix $\boldsymbol{\Gamma}^{1/2}$.

The signal after the AGC is

$$\tilde{\boldsymbol{y}} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{y} = \sqrt{p_u}\ \boldsymbol{\Gamma}^{1/2}\boldsymbol{H}\boldsymbol{D}^{1/2}\boldsymbol{x} + \boldsymbol{\Gamma}^{1/2}\boldsymbol{n} = \sqrt{p_u}\ \widetilde{\boldsymbol{H}}\boldsymbol{x} + \tilde{\boldsymbol{n}}. \tag{2.68}$$

Finally, quantization noise is added. With the assumption that $X_{ol} = 1$, variance of quantization noise in the ith chain is

$$p_{q,i} = \mathbb{E}\left[|q_i|^2\right] = \frac{2}{3}\ 2^{-2b_i}. \tag{2.69}$$

Signal model after the ADC:

$$\boldsymbol{z} = \tilde{\boldsymbol{y}} + \boldsymbol{q} = \sqrt{p_u}\ \widetilde{\boldsymbol{H}}\boldsymbol{x} + \tilde{\boldsymbol{n}} + \boldsymbol{q}. \tag{2.70}$$

The vector $\boldsymbol{q}$ contains the complex quantization noise samples from all the antennas.

Channel estimation in the uplink is performed using pilot sequences that are orthogonal in space and time and τ symbols long. More precisely, pilot sequences for all K users are represented by a $K \times \tau$ matrix $\boldsymbol{\Phi} = \sqrt{p_u \tau}\boldsymbol{\Psi}$, where in turn $\boldsymbol{\Psi}$ is a $K \times \tau$ matrix with orthonormal rows: $\boldsymbol{\Psi}\boldsymbol{\Psi}^H = \boldsymbol{I}_{K\times K}$. This type of matrices is optimal for LS pilot-based channel estimation [5].

When a block of pilot symbols $\boldsymbol{\Phi}$ is transmitted, the received signal is

$$\boldsymbol{Z} = \widetilde{\boldsymbol{H}}\boldsymbol{\Phi} + \widetilde{\boldsymbol{N}} + \boldsymbol{\Xi}, \tag{2.71}$$

where $\widetilde{\boldsymbol{N}} = [\tilde{\boldsymbol{n}}_1\ \tilde{\boldsymbol{n}}_2 \ldots \tilde{\boldsymbol{n}}_\tau]$ and $\boldsymbol{\Xi} = [\boldsymbol{q}_1\ \boldsymbol{q}_2 \ldots \boldsymbol{q}_\tau]$ are thermal and quantization noise vectors for each channel use (symbol), conveniently stacked in matrix form. The channel estimate is then

$$\hat{\widetilde{\boldsymbol{H}}} = \boldsymbol{Z}\boldsymbol{\Phi}^\dagger = \widetilde{\boldsymbol{H}} + \left(\widetilde{\boldsymbol{N}} + \boldsymbol{\Xi}\right)\boldsymbol{\Phi}^\dagger = \widetilde{\boldsymbol{H}} + \hat{\boldsymbol{H}}_{\boldsymbol{\epsilon}}. \tag{2.72}$$

Linear processing matrices for the uplink are formulated from the channel estimates:

- MR: $\hat{\boldsymbol{A}}_{\mathrm{MR}} = \hat{\widetilde{\boldsymbol{H}}}$,
- ZF: $\hat{\boldsymbol{A}}_{\mathrm{ZF}} = \hat{\widetilde{\boldsymbol{H}}}\left(\hat{\widetilde{\boldsymbol{H}}}^H\hat{\widetilde{\boldsymbol{H}}}\right)^{-1}$.

The MIMO receiver applies the processing matrix to estimate the vector of symbols sent by the users:

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{A}}^H\boldsymbol{z} = \sqrt{p_u}\ \hat{\boldsymbol{A}}^H\widetilde{\boldsymbol{H}}\boldsymbol{x} + \hat{\boldsymbol{A}}^H\tilde{\boldsymbol{n}} + \hat{\boldsymbol{A}}^H\boldsymbol{q}. \tag{2.73}$$

It can be shown that $\hat{\boldsymbol{A}}$ can be split into a sum of terms, one being the "true" processing matrix (based solely on the actual channel $\widetilde{\boldsymbol{H}}$) and the other an error term that is a consequence of channel estimation errors, namely:

- MR: $\hat{\boldsymbol{A}}_{\mathrm{MR}} = \boldsymbol{A}_{\mathrm{MR}} + \boldsymbol{A}_{\mathrm{MR},\epsilon} = \widetilde{\boldsymbol{H}} + \boldsymbol{A}_{\mathrm{MR},\epsilon}$,
- ZF: $\hat{\boldsymbol{A}}_{\mathrm{ZF}} = \boldsymbol{A}_{\mathrm{ZF}} + \boldsymbol{A}_{\mathrm{ZF},\epsilon} = \widetilde{\boldsymbol{H}} \left(\widetilde{\boldsymbol{H}}^H \widetilde{\boldsymbol{H}} \right)^{-1} + \boldsymbol{A}_{\mathrm{ZF},\epsilon}$.

This simple decomposition allows for splitting the estimate of user data symbol x_k, pertaining to kth user, into a wanted signal term and a noise term:

$$\hat{x}_k = x_k^{(w)} + w_k = \sqrt{p_u} \boldsymbol{a}_k^H \tilde{\boldsymbol{h}}_k x_k + w_k, \tag{2.74}$$

where $\boldsymbol{a}_k$ is the kth column of $\boldsymbol{A}$. The additive noise term w_k contains residual interuser interference and effects of thermal and quantization noise during channel estimation and data transmission phases.

One important observation (the proof of which is omitted here for brevity) is that the constituent terms of w_k are all *uncorrelated and Gaussian*. This is a consequence of several factors, namely: quantization noise being uncorrelated with the input to the ADC, noise in channel estimation phase being independent from the one in data transmission phase, and a large number of antennas (so that the central limit theorem applies).

The signal-to-interference-thermal-and-quantization-noise ratio (SINQR) for kth user is then calculated as

$$\mathrm{SINQR}_k = \frac{\mathbb{E}_{x,n,q}\{|x_k^{(w)}|^2\}}{\mathbb{E}_{x,n,q}\{|w_k|^2\}}. \tag{2.75}$$

The ergodic sum rate of the system is the sum of achievable rates for each user, averaged over channel realizations:

$$C = B \frac{T - \tau}{T} \sum_{k=1}^{K} \mathbb{E}\left\{\log_2(1 + \mathrm{SINQR}_k)\right\}, \tag{2.76}$$

with B being the bandwidth of the system.

Power Consumption Model

In this section, system setup choices and models are aimed to be as close to hardware reality as possible. To this end, we focus on a particular type of ADCs - namely, the pipeline ADC. This type of ADCs is typically designed for mid-range sampling rates and bit resolutions and has power consumption that is comparatively superior to other types of ADCs [34], [59].

For the power consumption model of the ADCs, this work adopts a theoretical model described in [58]. This model is a theoretical bound on power dissipation of pipeline ADCs that was nevertheless shown to correctly predict the trends observed in actual pipeline ADC designs. As such, it can be of use in a parametric power consumption model, where the *character of functional dependency* between b and power consumption is of primary interest.

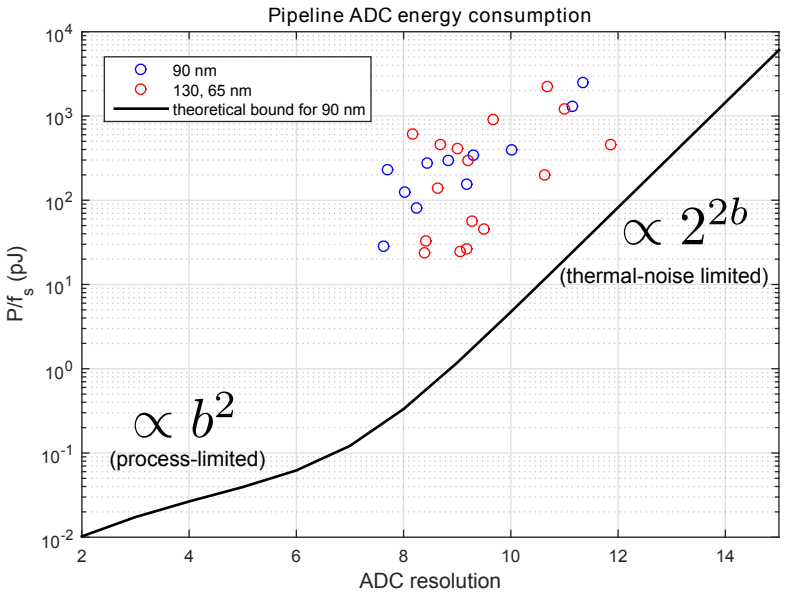


Figure 2.23: ADC power consumption model, compared with actual ADC designs.

As it can be seen in Figure 2.23, where the model from [58] is compared with actual pipeline ADC designs collected in [46], the functional dependency in the model matches the trend exemplified by best ADC designs. However, there is a gap (about two orders of magnitude wide) between the model and the designs. This implies that a correction factor Ω can be used if the model is to be matched to state-of-the-art designs. Therefore, the model can be given as

$$P_{ADC} = \Omega \left(c_1 b + c_2 b^2 + c_3 2^{2b} + c_4 b 2^{2b}\right) f_s, \tag{2.77}$$

where f_s is the Nyquist sampling rate of the ADC and coefficients c_1 through c_4 depend on ADC circuit parameters [58, Eq. (27)].

Another characteristic of this model that is worth pointing out is that the power consumption is linear with sampling rate f_s. This trend is also shown to be correct by analyzing the actual ADC designs in [46]; it only breaks down for high sampling rates (on the order of 100 MHz).

In order to show the complete picture regarding the energy efficiency of a MaMi base station in the uplink, power consumption of the remaining blocks (analog and digital) needs to be taken into account. This proves to be an extremely challenging task due to wide variability of available system designs and apparent lack of unifying theoretical information. Therefore, this work adapts a *parametric approach* to model the total power consumption.

Namely, power consumption of the blocks excluding the ADCs, denoted by P_{rest}, is normalized by $P_{ADC,ref} = \Omega \left(c_1 b_{ref} + c_2 b_{ref}^2 + c_3 2^{2b_{ref}} + c_4 b_{ref} 2^{2b_{ref}}\right) f_s$, where b_{ref} is an arbitrarily chosen bit resolution. This yields the architecture factor α:

$$\alpha = \frac{P_{rest}}{2M P_{ADC,ref}}. \tag{2.78}$$

Total power consumption of the BS in the uplink can then be expressed as

$$P_{tot} = 2M P_{ADC} + P_{rest} = 2M(P_{ADC} + \alpha P_{ADC,ref}). \tag{2.79}$$

Results

The aim of this section was to provide an initial overview of the energy efficiency trends as various system parameters change. To provide this initial insight, system performance simulations have been performed across a wide variety of system parameters.

Alongside the primary system parameter b, several other important system parameters have been considered, namely M, K, T, τ and preprocessing SNR $= p_u/p_n$ (defined with large-scale fading normalized to the level of thermal noise). In order to reduce the dimensionality of the analysis, two auxiliary system parameters have been introduced, namely *spatial loading* (K/M) and *temporal loading* (K/T).

In addition to all the assumptions on system setup listed before, it was assumed that perfect power control was performed in the uplink (so all $\beta_k = 1$). In all the analyses, reference bit resolution b_{ref} was set to 2.

For the first set of results, α and SNR were swept together with b. Additionally, $M = 100$, $\tau = K$, $K/T = 0.01$ [users/coherence time], $K/M = 0.1$ [users/antenna]. Results are shown in Figure 2.24.

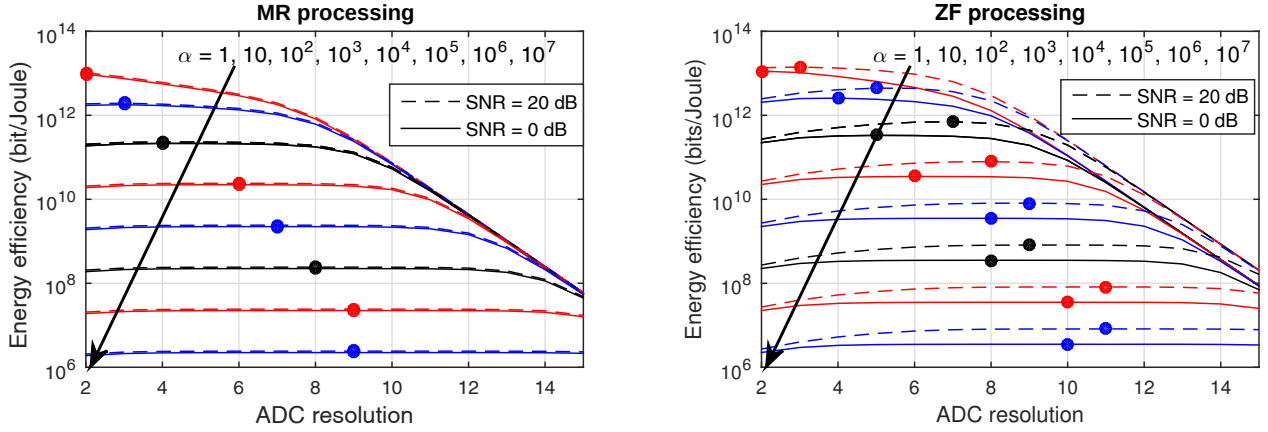


Figure 2.24: Energy efficiency as a function of architecture parameter α and SNR. Left: MR, right: ZF

Optimum energy efficiency points are denoted by the circular marker. Results indicate that, as the power consumption of ADCs becomes comparable to the power consumption of all the other blocks, from energy efficiency point of view it is beneficial to use smaller bit resolutions. However, in practical system designs it can be expected that ADC power consumption is only a small fraction of the total power consumption when the ADC resolution is low.

Just to provide an illustrative example, the BS power model presented in [16] was used with the parameters listed above and yielded $P_{rest} = 43.3W$. On the other hand, at $b_{ref} = 2$, using a correction factor $\Omega = 100$, the ADC power consumption model described above gave $2MP_{ADC} = 3\ mW$, resulting in $\alpha = 1.5 \times 10^4$. While this is by no means a definite power number, it serves to illustrate what are reasonable orders of magnitude for α.

Some other interesting insights can be drawn from this result, for example: a system using MR proves to be quite insensitive to changes in SNR and b, indicating that an overwhelmingly dominant impairment is the inter-user interference and that playing with ADC resolutions will not yield a considerable impact on the energy efficiency; if ZF is used, the dynamics are much more pronounced and show that by going from a system design with a large SNR and large α ("wasteful" system) to a system where SNR and α are low (a more "economical" system) allows

for choosing ADCs with smaller resolutions. Nevertheless, all systems with a "reasonable" α (say 10 - 10^5) should use ADCs with resolutions in the range 4-10 bits.

In order to focus more on what are the improvements and degradations of energy efficiency when using different ADC resolutions, we turn to a different analysis where spatial load K/M and M are swept together with b, and additionally SNR $= 0$ dB, $K/T = 0.01$ [users/coherence time] with $\tau = K$ and $\alpha = 10^4$, results shown in Figure 2.25.

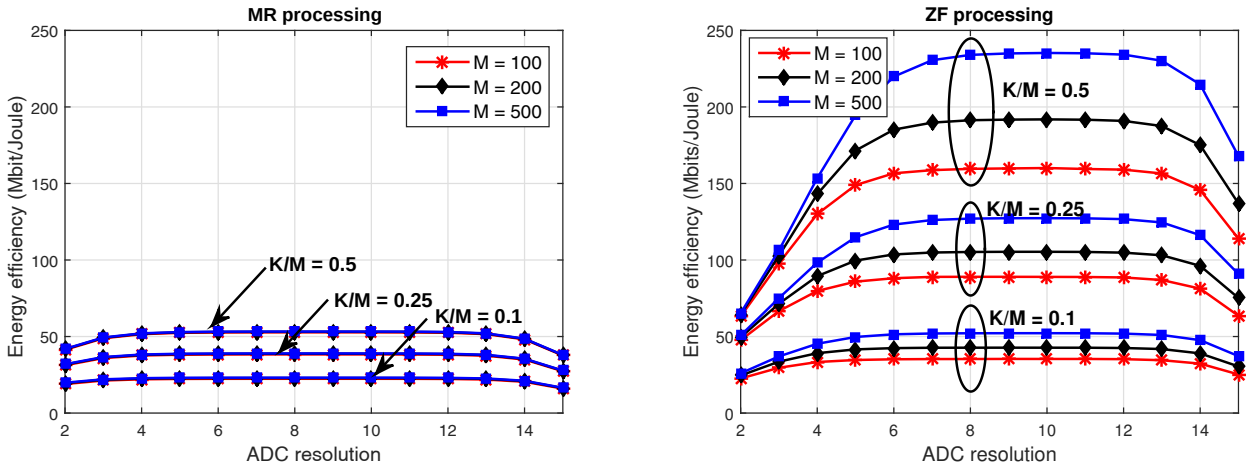


Figure 2.25: Energy efficiency as a function of M, K and b. Left: MR, right: ZF.

What these results show is that going from optimum ADC resolution to a very low one incurs a substantial degradation of the energy efficiency (up to 2.6 times in case of ZF!). This is due to sum rate being degraded while the overwhelming power consumption of the other blocks "drowns" the savings in power consumption of the ADCs. Another interesting observation is that, in the ZF case, increasing the number of antennas can help recover the energy efficiency lost by going to lower bit resolutions.

Finally, we take a look at the interplay between the channel estimation length and b in the context of energy efficiency. We analyze a system with $K/M = 0.1$ [users/antenna], while varying b, SNR and training length. Architecture parameter α is again fixed to 10^4. What is plotted is the normalized training length τ/T that maximizes the energy efficiency, results shown in Figure 2.26.

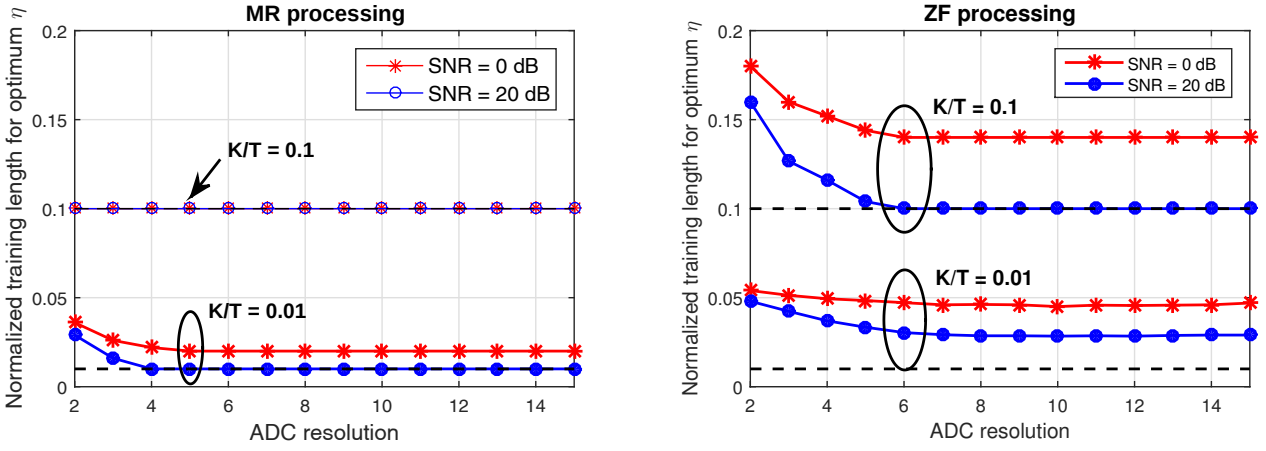


Figure 2.26: Normalized training length that maximizes energy efficiency. Left: MR, right: ZF.

The conclusions from here are that ZF is much more sensitive to quantization noise during training; even in the case of high temporal loading (indicating fast fading), when there is little room to afford for channel estimation, it is beneficial to train the system longer than the minimum required time in order to compensate for the effects of quantization. The effect becomes more pronounced as the fading becomes slower and channel estimation is not so costly in terms of time. On the other hand, we see that the system using MR is so overwhelmed by inter-user interference that additional training does little to improve the sum rate (and through it also energy efficiency).

2.3 Detection with Robustness Against Uplink Interference

The man-made interference is one of the major limiting factors in multi-user communications. The array gain that is obtained in MaMi makes the system robust to non-coherent interference sources, which includes conventional multi-user interference, hardware distortion, and receiver noise. However, the necessary reuse of pilot sequences across cells create a pilot contamination effect, where it seems impossible for the BS to coherently combine the desired signals from a UE without also coherently combining the interfering signal from the UEs that use the same pilot. In particular, T. Marzetta proved this result in his seminal paper [40] under the assumptions of maximum ratio combining (MRC) and independent Rayleigh fading channels. The implication was that pilot contamination creates a finite upper limit on the SE, as the number M of antennas goes to infinity. The same result has been proved for ZF and single-cell MMSE (S-MMSE) detection. Hence, it appears that MaMi cannot be made robust to pilot contamination.

In this section, we show that this is not the case in general, but only for independent Rayleigh fading channels and when heuristic detection schemes are used. The optimal M-MMSE detection scheme, presented in Section 2.1 of Deliverable 3.2 [39], is robust also to pilot contamination in all practical cases, meaning that it can reject also the coherent interference caused by pilot contamination and, in theory, make the SE grow without bound as we increase the number of antennas. This result is proved analytically in the MAMMOET publication [6], while we only provide an explanation and numerical validation in this section.

2.3.1 Definition of Optimal M-MMSE Detection

Consider a multi-cell scenario with L cells, each comprising a BS with M antennas and K UEs. There are K pilot sequences and the kth UE in each cell uses the same pilot. The received signal $\mathbf{y}_j \in \mathbb{C}^M$ at BS j is

$$\mathbf{y}_j = \sum_{l=1}^{L} \sum_{i=1}^{K} \sqrt{\rho} \mathbf{h}_{jli} x_{li} + \mathbf{n}_j \tag{2.80}$$

where ρ is the transmit power, x_{li} is the unit-power signal from UE i in cell l, $\mathbf{h}_{jli} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{jli})$ is the channel from this UE to BS j, $\mathbf{R}_{jli} \in \mathbb{C}^{M \times M}$ is the channel covariance matrix, and $\mathbf{n}_j \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ is the independent noise at BS j.

Using a total uplink pilot power of ρ^{tr} per UE and standard MMSE channel estimation techniques, BS j obtains the estimate

$$\hat{\mathbf{h}}_{jli} = \mathbf{R}_{jli} \mathbf{Q}_{ji}^{-1} \left(\sum_{l'=1}^{L} \mathbf{h}_{jl'i} + \frac{1}{\sqrt{\rho^{\text{tr}}}} \mathbf{n}_{ji} \right) \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{\Phi}_{jli}\right) \tag{2.81}$$

of $\mathbf{h}_{jli}$, where

$$\mathbf{Q}_{ji} = \sum_{l'=1}^{L} \mathbf{R}_{jl'i} + \frac{1}{\rho^{\text{tr}}} \mathbf{I}_M, \quad \mathbf{\Phi}_{jli} = \mathbf{R}_{jli} \mathbf{Q}_{li}^{-1} \mathbf{R}_{jli}. \tag{2.82}$$

The estimation error $\tilde{\mathbf{h}}_{jli} = \mathbf{h}_{jli} - \hat{\mathbf{h}}_{jli} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{jli} - \mathbf{\Phi}_{jli})$ is independent of $\hat{\mathbf{h}}_{jli}$.

We denote by $\mathbf{v}_{jk} \in \mathbb{C}^M$ the detection vector associated with UE k in cell j. Using standard techniques, the ergodic capacity is lower bounded by

$$\text{SE}_{jk} = \mathbb{E}\left\{\log_2\left(1 + \gamma_{jk}\right)\right\} \quad \text{[bit/s/Hz]} \tag{2.83}$$

where the expectation is with respect to the channel estimates is different coherence interval and the instantaneous SINR in a coherence interval is

$$\gamma_{jk} = \frac{|\mathbf{v}_{jk}^{\text{H}} \hat{\mathbf{h}}_{jjk}|^2}{\mathbf{v}_{jk}^{\text{H}} \left(\sum\limits_{(l,i) \neq (j,k)} \hat{\mathbf{h}}_{jli} \hat{\mathbf{h}}_{jli}^{\text{H}} + \mathbf{Z}_j \right) \mathbf{v}_{jk}}$$

where $\mathbf{Z}_j$ depends on the channel estimation error as

$$\mathbf{Z}_j = \sum_{l=1}^{L} \sum_{i=1}^{K} (\mathbf{R}_{jli} - \mathbf{\Phi}_{jli}) + \frac{1}{\rho} \mathbf{I}_M. \tag{2.84}$$

As proved in Section 2.1 of Deliverable 3.2 [39], in a given channel coherence interval, γ_{jk} is maximized by

$$\mathbf{v}_{jk} = \left(\sum_{l=1}^{L} \sum_{i=1}^{K} \hat{\mathbf{h}}_{jli} \hat{\mathbf{h}}_{jli}^{\text{H}} + \mathbf{Z}_j \right)^{-1} \hat{\mathbf{h}}_{jjk}. \tag{2.85}$$

This detection scheme is called multi-cell MMSE (M-MMSE) detection. The "multi-cell" notion is used to differentiate it from the single-cell MMSE (S-MMSE) detection scheme, which is widely used in the literature and is defined as

$$\mathbf{v}_{jk} = \left(\sum_{i=1}^{K} \hat{\mathbf{h}}_{jji} \hat{\mathbf{h}}_{jji}^{\text{H}} + \bar{\mathbf{Z}}_j \right)^{-1} \hat{\mathbf{h}}_{jjk}$$

with $\bar{\mathbf{Z}}_j$ being given by

$$\bar{\mathbf{Z}}_j = \sum_{i=1}^{K} \mathbf{R}_{jji} - \mathbf{\Phi}_{jji} + \sum_{\substack{l=1 \\ l \neq j}}^{L} \sum_{i=1}^{K} \mathbf{R}_{jli} + \frac{1}{\rho} \mathbf{I}_M. \tag{2.86}$$

The main difference from (2.85) is that only channel estimates in the own cell are computed in S-MMSE, while $\hat{\mathbf{h}}_{jli} \hat{\mathbf{h}}_{jli}^{\text{H}} - \mathbf{\Phi}_{jli}$ is replaced with its average (i.e., zero) for $l \neq j$. The computational complexity of S-MMSE is thus lower compared with M-MMSE, but the pilot overhead is identical since the same pilots are used to estimate both intra-cell and inter-cell channels.

The S-MMSE scheme coincides with M-MMSE when there is only one isolated cell, but it is generally different and lacks the ability to suppress interference from strongly interfering UEs in other cells (e.g., located at the cell edge). This might seem as a marginal difference, but it makes a fundamental difference when it comes to robustness to pilot contamination.

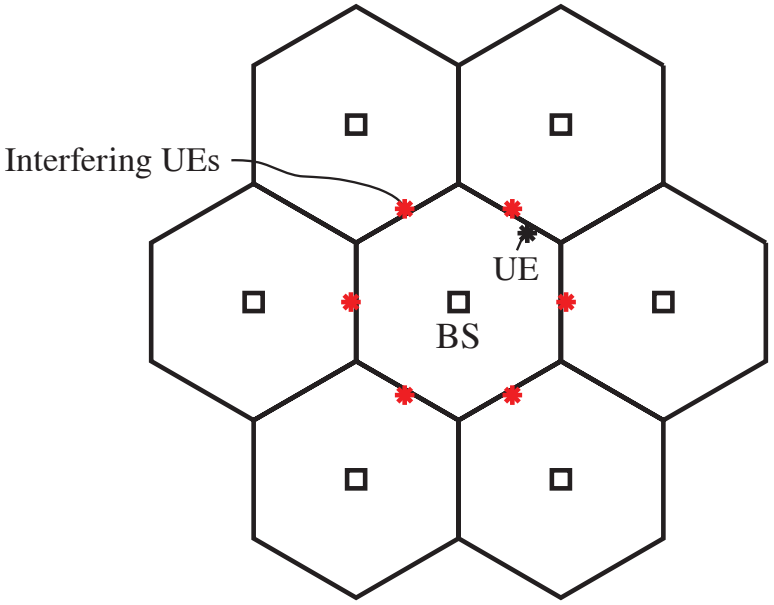


Figure 2.27: Multi-cell setup with one cell-edge UE in the center cell and one cell-edge UE in each of the neighboring cells, all using the same pilot sequence.

Consider UE i in cell j and UE i in cell n, which use the same pilot sequence. From (2.81) we can see that the estimates $\hat{\mathbf{h}}_{jji}$ and $\hat{\mathbf{h}}_{jni}$ of their channels to BS j are correlated as

$$\mathbb{E}\{\hat{\mathbf{h}}_{jni}\hat{\mathbf{h}}_{jji}^{\mathrm{H}}\} = \mathbf{R}_{jni}\mathbf{Q}_{li}^{-1}\mathbf{R}_{jji}. \tag{2.87}$$

This relationship can also be expressed as

$$\hat{\mathbf{h}}_{jni} = \mathbf{R}_{jni}\mathbf{R}_{jji}^{-1}\hat{\mathbf{h}}_{jji}^{\mathrm{H}} \tag{2.88}$$

if $\mathbf{R}_{jji}$ is invertible. If $\mathbf{R}_{jni}\mathbf{R}_{jji}^{-1}$ is a scaled identity matrix, then the two estimates are parallel. This implies that the covariance matrices are equal up to a scaling factor, which holds under independent Rayleigh fading, but not in general. For example, for a given pair of identical covariance matrices, it is sufficient that the elements are perturbed in a small independent random fashion to break the condition of being identical.

In general, the channel estimates of two UEs that use the same pilot are pointing in different directions and it is then possible for BS j to reject the interference from the interfering UE in cell n, while keeping a non-zero fraction of the desired signal from its own UE. This is exactly what the optimal M-MMSE combining does and which makes it robust to both conventional interference and pilot contamination.

2.3.2 Numerical Validations

To illustrate the fact that pilot contamination generally does not limit the asymptotic SE, when M-MMSE detection is being used, we numerically evaluate the multi-cell scenario in Figure 2.27 with $K = 1$ one UE per cell and $L = 7$ cells. All UEs use the same pilot sequence and are at the cell edge near the center cell. This is a challenging setup with very high pilot contamination, and it will show the robustness result very clearly.

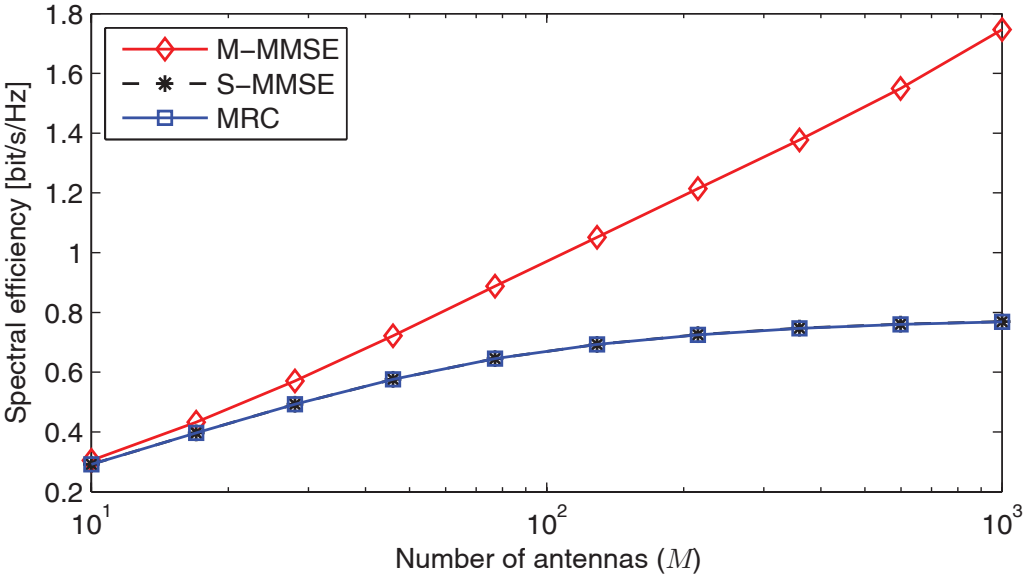


Figure 2.28: SE as a function of the number of BS antennas, for covariance matrices based on the exponential correlation model in (2.89).

The asymptotic SE behavior is considered in Figure 2.28 using the exponential correlation model where the m, nth entry of the covariance matrix $\mathbf{R}$ between an arbitrary UE and BS is

$$[\mathbf{R}]_{m,n} = \beta r^{|n-m|} e^{\imath(n-m)\theta} \tag{2.89}$$

where β is the average pathloss, $r \in [0, 1]$ is the correlation factor between adjacent antennas and θ is the angle-of-arrival, as seen from the BS. This model is selected since it provides full-rank covariance matrices whenever $r < 1$, thus the results shown in this section are not due to any spatial sparseness. We consider $r = 0.5$ and the detection schemes M-MMSE, S-MMSE, and MRC. The average SNR observed at a BS antenna in the center cell is set equal for the pilot and data transmission: $\rho \mathrm{tr}(\mathbf{R}_{jli})/M = \rho^{\mathrm{tr}} \mathrm{tr}(\mathbf{R}_{jli})/M$. It is -7.0 dB for the desired UE and -8.6 dB for each of the interfering UEs. Figure 2.28 shows that S-MMSE provides slightly higher SE than MRC, but both converge to an asymptotic limit of around 0.8 bit/s/Hz as the number of antennas grows. In contrast, M-MMSE provides an SE that clearly grows without bound. This means that the effective SINR grows linearly with M, as seen from the fact that the SE grows linearly with a logarithmic horizontal scale. This validates the robustness towards pilot contamination that M-MMSE achieves: it provides an array gain (i.e., signal gain proportional to M) for the desired UE, while the interference from the contaminating UEs does not grow with M.

Next, we consider an alternative channel covariance model, which builds upon independent Rayleigh fading, but includes independent log-normal large-scale fading over the array. The covariance matrix between an arbitrary UE and BS is

$$\mathbf{R} = \beta \mathrm{diag}\left(10^{f_1/10}, \ldots, 10^{f_M/10}\right) \tag{2.90}$$

where β is the average pathloss, $f_m \sim \mathcal{N}(0, \sigma^2)$ and σ is the standard deviation. The SE with $M = 1000$ and varying standard deviation σ of the large-scale fading variations is shown in Figure 2.29. M-MMSE provides no benefit over S-MMSE or MRC in the special case of $\sigma = 0$, where all covariance matrices are linearly dependent (scaled identity matrices). This

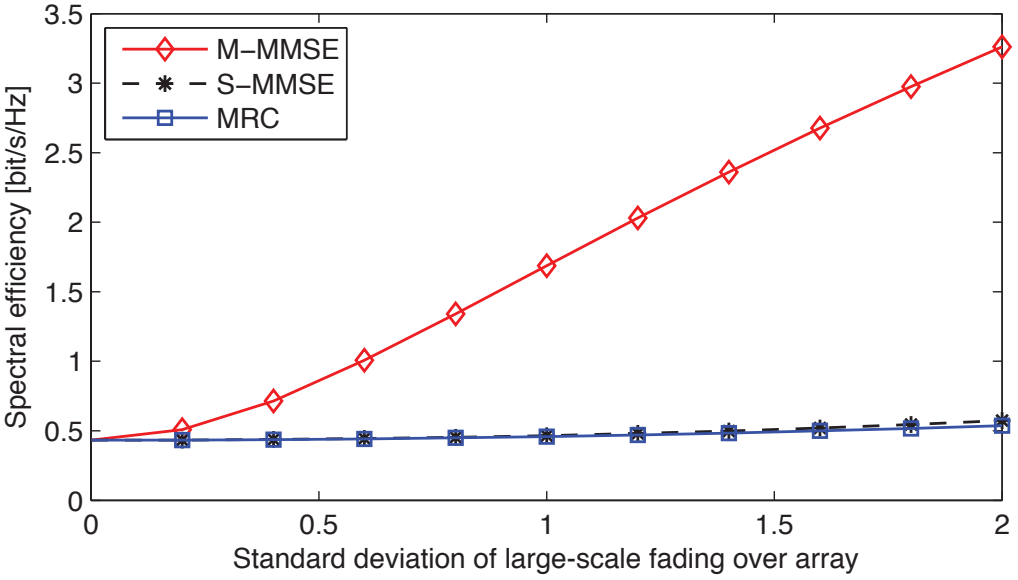


Figure 2.29: SE as a function of the standard deviation of the independent large-scale fading variations, for covariance matrices modeled by (2.90).

is a special case that has received massive attention from academic researchers. However, M-MMSE provides substantial gains as soon as there are some minor variations in channel gain over the array, which effectively make the covariance matrices linearly independent. The range of fading variations in this simulation can be compared with the measurements in [20], which show large-scale variations of around 4 dB over a MaMi array.

Conclusion

This study shows that M-MMSE combining is robust to man-made interference, such as pilot contamination, in the sense that it generally does not cause a fundamental upper limit on the SE in MaMi, despite previous studies that have pointed towards that direction. There are indeed special cases where the channel covariance matrices are linearly dependent, which make the channel estimates of the desired and interfering UEs parallel such that linear detection cannot remove the interference. In general, the covariance matrices and the channel estimates are not linearly dependent, thus M-MMSE detection can extract the desired signal while rejecting the pilot contamination. There is a power loss, as compared to the contamination-free case, but the SE still grows without bound as $M \to \infty$. Importantly, this means that MRC is generally not asymptotically optimal in MaMi.

Chapter 3

Cross Layer and System Operation

This chapter consider the impact that MaMi signal processing algorithms have, beyond the physical baseband processing. Section 3.1 and Section 3.2 describe and analyze power control algorithms that exploit the channel hardening properties to reduce the complexity. The impact that MaMi has on other systems, in terms of OOB radiation, is analyzed in Section 3.3.

3.1 Uplink Pilot and Payload Power Control: Throughput-Fairness Trade-Offs

This section deals with pilot and data power allocation in uplink single-cell MaMi systems, under ideal link adaptation. Compared to conventional power control in single-antenna systems, power control in MaMi networks is a relatively new topic. Accurate channel estimates are needed at the BS for carrying out coherent linear processing, e.g. uplink detection and downlink precoding. Due to the large number of antennas in MaMi the instantaneous channel knowledge, which is commonly assumed to be known perfectly in the power control literature, is hard to obtain perfectly. The literature on power control for multi-user MIMO, and even jointly with optimal beamformer design, see for example [7, 55] and the references therein, did not consider the channel estimation error explicitly and the design criterion was based on SE.

We will present power control schemes that optimize the ergodic SE based on only the large-scale fading, which takes into account the channel estimation errors and at the same time simplify the system design since the same power control coefficients are used can be used for all subcarriers and as long as the large-scale fading characteristics are fixed. Since the analysis is based on ergodic SE, an ideal link adaptation is assumed. The power control is formulated as optimization problems for two different objective functions: the weighted minimum SE among the users and the weighted sum SE. A closed-form solution for the optimal length of the pilot sequence is obtained. The optimal power control policy for the former problem is found by solving a simple equation with a single variable. Utilizing the special structure arising from imperfect channel estimation, a convex reformulation is found to solve the latter problem to global optimality in polynomial time. The gain of the optimal joint power control is theoretically justified, and is proved to be large in the low SNR regime. Simulation results also show the advantage of optimizing the power control over both pilot and data power, as compared to the cases of using full power and of only optimizing the data powers as done in previous work.

The questions we want to answer by carrying out this analysis are:

1. Is power control on the pilots needed for MaMi systems? If the answer is yes, how much can we gain from jointly optimizing the pilot power and data power, as compared to

always using equal power allocation or just power control over the data power?

2. In which scenarios can we gain the most from joint optimization?

3. What intuition can be obtained from the optimal power control? This includes the pilot length, and how the pilot and payload power depend on the estimation quality and signal to noise ratio (SNR).

We focus on the main results and algorithms in this sections, while the derivations are available in the MAMMOET publication [11].

3.1.1 System Model

Consider an uplink single-cell MaMi system with M antennas at the BS and K single-antenna users. The K users are assigned K orthogonal pilot sequences of length τ_p for $K \leq \tau_p \leq T$, where T is the number of symbols in the coherence interval in which the channels are assumed to be constant. The channels are modeled to be independent Rayleigh fading. The flat fading channel matrix between the BS and the users is denoted by $\boldsymbol{H} \in \mathbb{C}^{M \times K}$, where the k^{th} column represents the channel response to user k and has the distribution

$$\boldsymbol{h}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{I}), \;\; k = 1, 2, \ldots, K, \tag{3.1}$$

which is a circularly symmetric complex Gaussian random vector. The variance $\beta_k > 0$ represents the large-scale fading including path loss and shadowing, and is normalized by the noise variance at the BS to simplify the notation. The large-scale fading coefficients are assumed to be known at the BS as they are varying slowly (in the scale of thousands of coherence intervals). The power control proposed in this work only depends on the large-scale fading which makes it feasible to optimize the power control online.

In each coherence interval, UE k transmits its orthogonal pilot sequence with power p_p^k to enable channel estimation at the BS. We assume that MMSE channel estimation is carried out at the BS to obtain the small-scale coefficients. This gives an MMSE estimate of the channel vector from UE k as

$$\hat{\boldsymbol{h}}_k = \frac{\sqrt{\tau_p p_p^k \beta_k}}{1 + \tau_p p_p^k \beta_k} \left(\sqrt{\tau_p p_p^k} \boldsymbol{h}_k + \boldsymbol{n}_p^k \right) \tag{3.2}$$

where $\boldsymbol{n}_p^k \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$ accounts for the additive noise during the training interval. During the payload data transmission interval, the BS receives the signal

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{h}_k \sqrt{p_d^k} s_k + \boldsymbol{n} \tag{3.3}$$

where s_k is the zero mean and unit variance Gaussian information symbol from UE k and $\boldsymbol{n} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$ represents the noise during the data transmission. The channel estimates are used for MRC or ZF detection of the payload, which corresponds to multiplying the received signal $\boldsymbol{y}$ with $\hat{\boldsymbol{H}}^H \triangleq [\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_K]^H$ or $(\hat{\boldsymbol{H}}^H \hat{\boldsymbol{H}})^{-1} \hat{\boldsymbol{H}}^H$ to detect the symbols $s_1, \ldots, s_K$. The power control methodologies presented in this subsection can be applied jointly to each subcarrier in an OFDM systems. With the channel hardening effect offered by MaMi, channel variations in different subcarriers can be neglected and the SE in every subcarrier will mainly depend on the large-scale fading. Therefore the whole spectrum can be allocated to every UE and the same power control can be applied to all subcarriers. To make a fair comparison with the scheme

with equal power allocation in which each UE gives the same power to pilot and data, as done in [47] and most other previous work, we impose the following constraint on the total transmit energy over a coherence interval:

$$\tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \ k = 1, \ldots, K \tag{3.4}$$

where E_k is the total energy budget for UE k within one coherence interval. Unlike previous work, we consider the scenario where each UE can choose freely how to allocate its energy budget on the pilots and payload.

3.1.2 Achievable SE With Linear Detection

Since the exact ergodic capacities of the UE channels with channel uncertainty is unknown, lower bounds on the achievable SE are often adopted as the performance metric in MaMi. Here we present lower bounds on the capacity for arbitrary power control.

The capacity of UE k with MRC detection is lower bounded by the achievable ergodic SE

$$R_k = \left(1 - \frac{\tau_p}{T}\right) \log_2(1 + \mathrm{SINR}_k) \tag{3.5}$$

where pilot and payload powers are arbitrary,

$$\mathrm{SINR}_k = \frac{M p_d^k \gamma_k}{1 + \sum_{j=1}^{K} \beta_j p_d^j} \tag{3.6}$$

and $\gamma_k = \frac{\tau_p p_p^k \beta_k^2}{1 + \tau_p p_p^k \beta_k}$.

Similarly, the capacity of UE k with ZF detection is lower bounded by the achievable ergodic SE

$$R_k = \left(1 - \frac{\tau_p}{T}\right) \log_2(1 + \mathrm{SINR}_k) \tag{3.7}$$

where pilot and payload powers are arbitrary,

$$\mathrm{SINR}_k = \frac{(M - K) p_d^k \gamma_k}{1 + \sum_{j=1}^{K} p_d^j (\beta_j - \gamma_j)} \tag{3.8}$$

and $\gamma_j = \frac{\tau_p p_p^j \beta_j^2}{1 + \tau_p p_p^j \beta_j}$. $M > K$ needs to be satisfied for ZF detector to work.

These achievable SEs are the performance metric commonly used in the MaMi literature. Therefore it is used throughout the paper, where τ_p, p_p^k and p_d^k are the variables to be optimized (for $k = 1, \ldots, K$). The optimization can be done at the BS, which can then inform the UEs about the pilot length, the amount of power to be spent on pilots, and the amount of power to be spent on payload data. The aim is to maximize a given utility function $U(R_1, \ldots, R_K)$ where $U(\cdot)$ can be any function that is monotonically increasing in every argument. The utility function characterizes the performance and fairness that we provide to the UEs. Examples of commonly used utility functions are the max-min fairness, sum performance, and proportional fairness. The general problem we address for both MRC and ZF is:

$$\begin{aligned} \underset{\tau_p, \{p_p^k\}, \{p_d^k\}}{\text{maximize}} \quad & U\left(R_1, \ldots, R_K\right) \\ \text{subject to} \quad & \tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \ \forall k, \\ & p_p^k \geq 0, p_d^k \geq 0, \ \forall k, \\ & K \leq \tau_p \leq T. \end{aligned} \tag{3.9}$$

Two important results can be obtained for this problem, when using MRC or ZF detection:

- For any monotonically increasing utility function $U(R_1, \ldots, R_k)$, the energy constraint (3.4) is satisfied with equality for every UE at the optimal solution, i.e.,

$$\tau_p p_p^k + (T - \tau_p) p_d^k = E_k, \ k = 1, \ldots, K.$$

- For any monotonically increasing utility function $U(R_1, \ldots, R_k)$, the problem (3.9) has $\tau_p = K$ at the optimal solution.

Using these properties, we can reduce the number of variables involved in (3.9) and this enables us to find the optimal solutions for certain utility functions in the following sections. We also know that the optimal training period τ_p is equal to the number of UEs being served, and is the same for every UE. Therefore there is no need for assigning pilot sequences of different lengths to different UEs.

3.1.3 Maximize Weighted Minimum SE

In this subsection, we solve the power control problem (3.9) for the class of max-min fairness problem. The max-min fairness problem is selected to provide the same quality-of-service to all users in the cell. The two cases with MRC and ZF will be discussed separately since the SINR expressions are different. With max-min fairness we aim at serving every user with equal weighted SE according to their priorities and make this value as large as possible. We choose $U(\tilde{R}_1, \ldots, \tilde{R}_K) = \min_k \tilde{R}_k$ with $\tilde{R}_k = (1 - \frac{\tau_p}{T}) \log_2(1 + w_k \text{SINR}_k)$ where $w_k > 0$ are weighting factors to prioritize different users. Since $(1 - \frac{\tau_p}{T}) \log_2(1 + w_k \text{SINR}_k)$ is monotonically increasing in $w_k \text{SINR}_k$, it is equivalent to choose the objective as $\min_k w_k \text{SINR}_k$.

Max-Min for MRC

With MRC, the power control problem becomes

$$\begin{aligned} \underset{\{p_p^k\},\ \{p_d^k\}}{\text{maximize}} \quad & \min_k \frac{w_k M p_d^k \gamma_k}{1 + \sum_{j=1}^K \beta_j p_d^j} \\ \text{subject to} \quad & \tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \forall k \\ & p_p^k \geq 0, p_d^k \geq 0, \forall k. \end{aligned} \tag{3.10}$$

Using the epigraph form of (3.10) we have the following equivalent problem formulation:

$$\begin{aligned} \underset{\{p_p^k\},\{p_d^k\},\ \lambda}{\text{maximize}} \quad & \lambda \\ \text{subject to} \quad & w_k M p_d^k \tau_p p_p^k \beta_k^2 \geq \\ & \lambda(1 + \sum_{j=1}^K \beta_j p_d^j + \tau_p p_p^k \beta_k \\ & + \tau_p p_p^k \beta_k \sum_{j=1}^K \beta_j p_d^j), \forall k \\ & \tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \forall k \\ & p_p^k \geq 0, p_d^k \geq 0, \forall k. \end{aligned} \tag{3.11}$$

This problem is non-convex as it is formulated here, however we recognize it as a geometric program (GP). Such programs can be solved efficiently to global optimality with any general-purpose GP solver, for example, [3] with CVX [23]. We have also developed an explicit semi-closed form solution to the problem, which only requires a line search to obtain the global optimum. The exact details can be found in the MAMMOET publication [11].

Max-Min for ZF

Similar to the case of the MRC detector, we can write the problem as max-min weighted SINR as follows:

$$\begin{aligned} \underset{\{p_p^k\},\ \{p_d^k\}}{\text{maximize}} \quad & \min_k \frac{w_k(M-K)p_d^k\gamma_k}{1+\sum_{j=1}^K p_d^j(\beta_j-\gamma_j)} \\ \text{subject to} \quad & \tau_p p_p^k + (T-\tau_p)p_d^k \leq E_k, \forall k \\ & p_p^k \geq 0, p_d^k \geq 0, \forall k. \end{aligned} \tag{3.12}$$

The only difference from (3.10) is the expressions of the SINRs, which is now taken from (3.8) by inserting $\tau_p = K$. Due to the negative terms appearing in the denominator of the SINR expressions, this problem cannot be directly transformed to a GP problem.

Fortunately, we can prove that Problem (3.12) can be reformulated as

$$\begin{aligned} \underset{\{p_p^k\},\ \{p_d^k\}}{\text{maximize}} \quad & \min_k \frac{w_k(M-K)p_d^k\gamma_k}{1+\sum_{j=1}^K p_d^j\beta_j} \\ \text{subject to} \quad & \tau_p p_p^k + (T-\tau_p)p_d^k \leq E_k, \forall k \\ & p_p^k \geq 0, p_d^k \geq 0, \forall k. \end{aligned} \tag{3.13}$$

This implies that solving problem (3.13) gives the same optimal p_d^k, p_p^k as solving problem (3.12), but the objective value is different.

By comparing (3.13) with (3.10), we see that only the difference is that M with MRC is replaced with $M-K$ with ZF. Therefore the power allocation that solves the weighted max-min SE for the MRC also solves the weighted max-min SE for the ZF. The same methods and analytical solutions apply. Practically speaking, this implies that the users do not need to know what kind of detector is used at the BS. While the BS can switch between different detectors according to the data traffic requirements or power consumption restrictions.

3.1.4 Joint Pilot and Data Power Control for Weighted Sum SE

In this subsection, we solve the power control problem (3.9) for the weighted sum SE for MRC and ZF detector. This problem is selected to maximize the total system throughput, and weights are included to provide some fairness between different users. We define the weighted sum SE by choosing $U(R_1,\ldots,R_K) = \sum_{k=1}^K w_k R_k$.

Power control that maximizes sum SE when interference is present is known to be an NP-hard problem in general under perfect channel knowledge [37]. In this part we present a polynomial-time solution to one special case when all sources transmit to the same receiver. When channel estimation errors are present, with the bounding techniques we used for the achievable SE we discover a specific structure that lead to a convex reformulation after a series of transformations. Since optimizing the data power is considered to be a hard problem itself, in the following we first present the case when one only optimizes the data power, then the solution approach is extended to the case of joint optimization of pilot and data power.

Weighted Sum SE for MRC

By utilizing the properties that were presented earlier, (3.9) now becomes the following optimization problem:

$$\begin{aligned} \underset{\{p_d^k\},\ \{p_p^k\}}{\text{maximize}} \quad & \sum_k w_k \log_2 \left(1 + \frac{M p_d^k \gamma_k}{1 + \sum_{j=1}^K \beta_j p_d^j}\right) \\ \text{subject to} \quad & \tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \forall k, \\ & p_d^k \geq 0,\ p_p^k \geq 0, \forall k. \end{aligned} \tag{3.14}$$

Since γ_k depends on p_p^k which is also an optimization variable, the problem is non-convex. However, in [11], we prove that Problem (3.14) can be reformulated into the following form:

$$\begin{aligned} \underset{s,\ \{y_k\}}{\text{maximize}} \quad & \sum_k w_k \log_2 (1 + M y_k) \\ \text{subject to} \quad & \sum_{j=1}^K \beta_j q(y_j, s) \leq 1 - s, \end{aligned} \tag{3.15}$$

where

$$q(y_j, s) = \frac{E_j \beta_j s + (T-K) y_j - \sqrt{E_j^2 \beta_j^2 s^2 - 2(T-K)(E_j \beta_j + 2) y_j s + (T-K)^2 y_j^2}}{2(T-K)\beta_j}. \tag{3.16}$$

The two formulations are equivalent in the sense that they have the same optimal objective values, and the solution to (3.14) can be obtained from solution to (3.15) via $p_d^k = q(y_k, s)/s$. Moreover problem (3.15) is jointly convex in s and y_k. Since we have a convex reformulation (3.15) we can use standard convex solvers to find the optimal solutions efficiently, and the optimal power control parameters can be recovered easily.

Sum SE for ZF

In the case of perfect CSI, maximizing sum SE for ZF is straightforward. This is because the ZF detector completely removes all the interference from other users and creates K parallel channels. However in the case of imperfect CSI, the interference is reduced but still remains, which makes the sum SE problem at least as difficult as with MRC. Fortunately, the techniques we developed for solving the MRC case can be applied here to solve the problem to global optimality. The problem is as follows:

$$\begin{aligned} \underset{t,\ \{p_d^k\},\ \{p_p^k\}}{\text{maximize}} \quad & \sum_k w_k \log_2 \left(1 + \frac{(M-K) p_d^k \gamma_k}{1 + \sum_{j=1}^K (\beta_j - \gamma_j) p_d^j}\right) \\ \text{subject to} \quad & \tau_p p_p^k + (T - \tau_p) p_d^k \leq E_k, \forall k \\ & p_d^k \geq 0,\ p_p^k \geq 0, \forall k. \end{aligned} \tag{3.17}$$

Similar to the MRC case, Problem (3.17) can be reformulated into the following form:

$$\begin{aligned} \underset{s,\ \{y_k\}}{\text{maximize}} \quad & \sum_k w_k \log_2 (1 + (M-K) y_k) \\ \text{subject to} \quad & \sum_{j=1}^K \beta_j q(y_j, s) - \sum_{j=1}^K y_j \leq 1 - s, \end{aligned} \tag{3.18}$$

where $q(y_j, s)$ is given in (3.16) which is the same as in the MRC case. The two formulations are equivalent in the sense that they have the same optimal objective values, and the solution to (3.17) can be obtained from solution to (3.18) via $p_d^k = q(y_k, s)/s$. Moreover problem (3.18) is jointly convex in s and y_k. Hence, it can be solved by general-purpose solvers.

3.1.5 Simulation Results and Discussion

In this subsection, we present simulation results to demonstrate the benefits of our algorithms and compare the performance with the case of no power control (i.e., full equal power) as well as the case of power control on the payload power only (and full power pilots). We consider a scenario with $M = 100$ antennas, $K_0 = 10$ users, and the length of the coherence interval is $T = 200$ (which for example corresponds to a coherence bandwidth of 200 kHz and a coherence time of 1 ms). The users are assumed to be uniformly and randomly distributed in a cell with radius $R = 1000$ m and no user is closer to the BS than 100 m. The path-loss model is chosen as $\beta_k = z_k/r_k^{3.76}$ where r_k is the distance of user k from the BS where z_k represents the independent shadow fading effect. It is chosen to be log-normal distributed with a standard deviation of 8 dB. Due to the long tail behavior of the log-normal distribution there could be some users with very small β_k, therefore in each snapshot the user with the smallest β_k is dropped from service. Therefore the algorithm is run for $K = K_0 - 1 = 9$ users.

The energy budgets $E_k = 10^{-0.5} \times R^{3.76} \times T$ and $E_k = 10^{0.5} \times R^{3.76} \times T$ give a median SNR of -5 dB and 5 dB at the cell edge when using equal power allocation. The weights w_k are set to be equal in all the simulations. The algorithms are run for 1000 Monte-Carlo simulations where in each snapshot the users are dropped randomly in the cell so that the large-scale fading β_k changes.

Max-Min SE Results

We compare 4 schemes:

1. the solution to problem (3.11) (marked as 'Max-min' in the figures);
2. equal power allocation $p_d^k = p_p^k = E_k/T$ (marked as 'Equal Power' in the figures);
3. optimizing only payload power for problem (3.11) by fixing $p_p^k = E_k/T$ (marked as 'Max-min (data)' in the figures);
4. the scheme that maximizes the sum SE is presented as well for reference (marked as 'sum' in the figures).

The same schemes are tested for both MRC and ZF, and low and high SNR scenarios.

In Figure 3.1 (a) and (b) we plot the cumulative distribution function (CDF) of the minimum SE over different snapshots of user locations for MRC at low and high SNR respectively. We observe that without any power control in almost all of the cases the user with the lowest SNR will get less than 0.5 bit/s/Hz in both low and high SNR scenarios. This is not acceptable if we want to provide decent quality of service to every user being served. With max-min power control for both pilot and data we resolve this problem by guaranteeing the users an SE of more than 1 bit/s/Hz with 0.95 probability and 2.75 bit/s/Hz with 0.5 probability. In low SNR scenarios the joint optimization doubles the 0.95 likely point, from 0.5 to 1 bit/s/Hz, which proves the need of joint pilot and data power optimization at low SNR. In this case with data power control the user with the worst channel would have poor channel estimates that

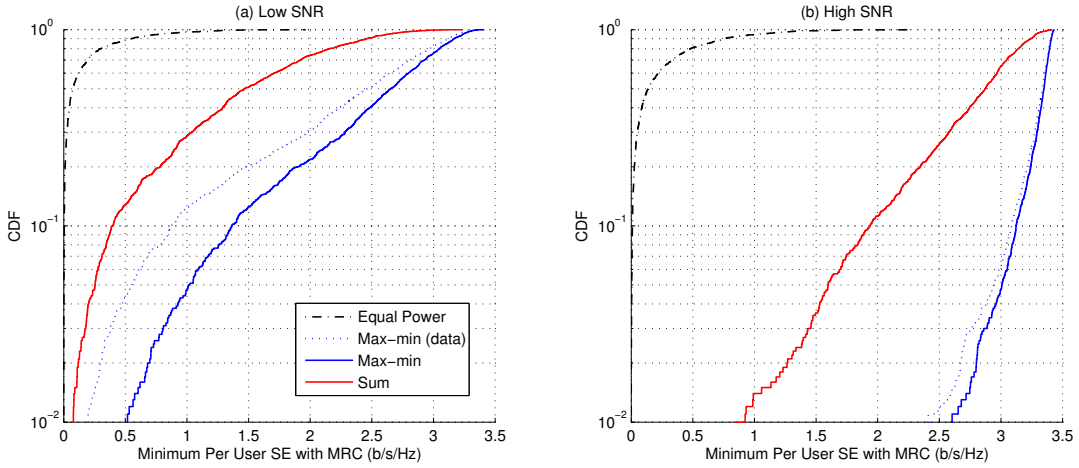


Figure 3.1: CDF of the minimum SE with $M = 100$, $K_0 = 10$, $T = 200$, $R = 1000$ m for MRC. Subplots (a) and (b) correspond to low SNR (-5 dB) and high SNR (5 dB) at the cell edge, respectively.

limits the SE, while with joint power control they borrow power from the data part to enhance channel estimation and thereby increase the SE. However in the high SNR scenarios the gain is marginal by the joint optimization, power control over data is enough. This is because the channel estimates are already good enough for linear detection. The performance of the sum SE formulation is not surprising as it is not designed for improving the minimum SE. It boosts the SE of the users with better channels to increase the sum SE, which in turn scarifies the users with worse channels.

In Figure 3.2 (a) and (b) we plot the CDF of the minimum SE over different snapshots of user locations for ZF at low and high SNR respectively. We observe that all schemes perform similarly and the gains from joint power control with respect to only power control over data are not as large as in the case of MRC. This is because with ZF most interference is removed by the detector, however in low SNR scenarios joint power control is still necessary as it increases the 0.95 likely point from 0.5 to 1 bit/s/Hz compared to power control over data only. The performance of the sum SE formulation is surprisingly good at both low and high SNR and is even better than the max-min scheme with only data power control. This suggests that with ZF detector we can go for the sum SE formulation and push up the total system throughput without sacrificing much of the worse users' performance.

Sum SE Results

We compare 4 schemes:

1. the scheme that maximizes the sum SE (marked as 'Sum' in the figures);
2. equal power allocation $p_d^k = p_p^k = E_k/T$ (marked as 'Equal Power' in the figures);
3. optimizing the data power only for sum SE by fixing $p_p^k = E_k/T$ (marked as 'Sum (data)' in the figures);
4. the max-min scheme is also presented for reference (marked as 'max-min' in the figures).

The same schemes are tested for both MRC and ZF.

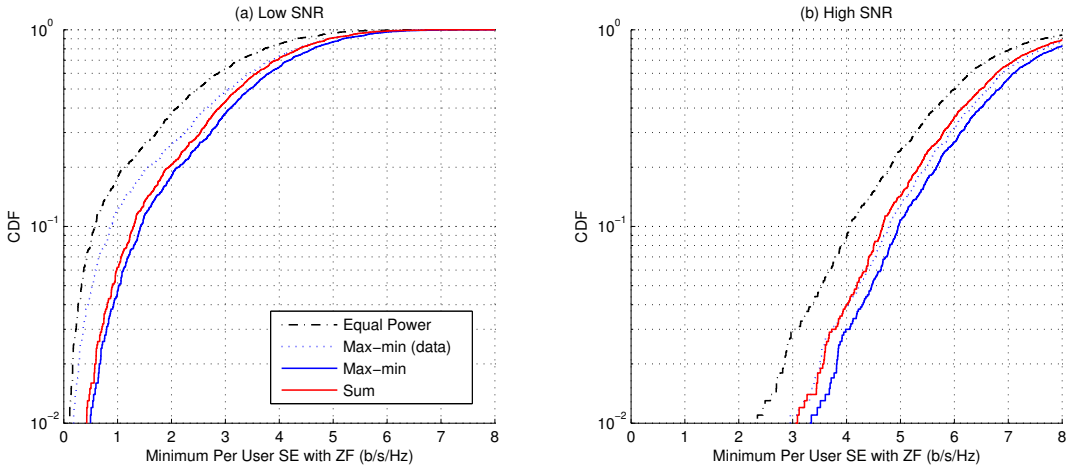


Figure 3.2: CDF of the minimum SE with $M = 100$, $K_0 = 10$, $T = 200$, $R = 1000$ m for ZF. Subplots (a) and (b) correspond to low SNR (-5 dB) and high SNR (5 dB) at the cell edge, respectively.

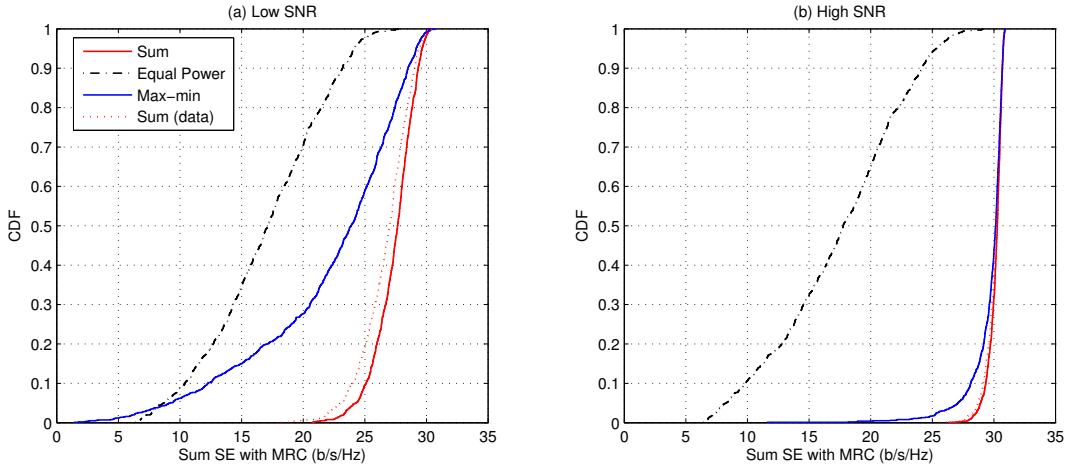


Figure 3.3: CDF of the sum SE with $M = 100$, $K_0 = 10$, $T = 200$, $R = 1000$ m for MRC. Subplots (a) and (b) correspond to low SNR (-5 dB) and high SNR (5 dB) at the cell edge, respectively.

In Figure 3.3 (a) and (b) we plot the CDF of the sum SE for the scenario we described above for MRC at low and high SNR respectively. We observe the optimized power control increases the sum SE significantly. The whole CDF is shifted to the right by almost 15 bit/s/Hz in the low SNR scenario with the proposed power control as compared to equal power allocation. At low SNR the joint power control offers about 10% increase over the case with only data power control. At high SNR the gain is marginal as the SEs of the users have saturated so we are in the log part of the SE already. The max-min scheme performs well at high SNR due to the saturation of SE, but worse at low SNR. This is because enforcing max-min fairness lead to large sacrifices in sum SE at low SNR. The reason is that with high probability there will be some very disadvantaged user, and everyone else has to cut back significantly to avoid causing near-far interference.

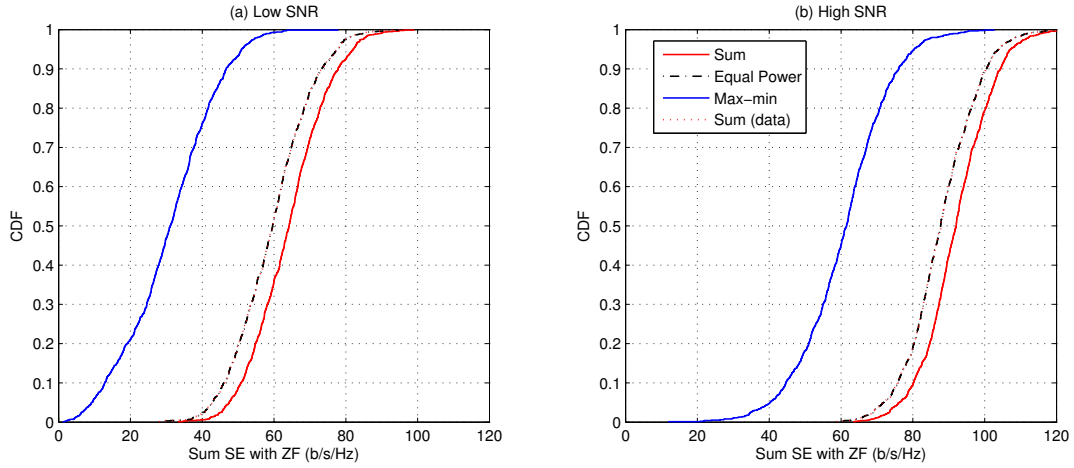


Figure 3.4: CDF of the sum SE with $M = 100$, $K_0 = 10$, $T = 200$, $R = 1000$ m for ZF. Subplots (a) and (b) correspond to low SNR (-5 dB) and high SNR (5 dB) at the cell edge, respectively.

In Figure 3.4 (a) and (b) we plot the CDF of the sum SE for ZF at low and high SNR respectively. We observe that with ZF when we optimize only the data power the optimal scheme is always using full power. The reason for this is that in single cell systems ZF removes most of the interference, the near-far effects are almost removed by the ZF detector thus creating almost parallel channels. Therefore the scheme with equal power allocation is the same as optimizing data power only. The joint power control offers about 10% improvements over the case with only data power control at low SNR and the gain diminishes as the SNR increases. However there will always be a gap between the two schemes, this is because even when the SNR tends to infinity we can always save power on the pilot and use it for data to increase the SE. The max-min scheme performs poorly in both scenarios, this confirms our suggestion that with ZF we should use the sum SE formulation.

Robustness

In this subsection, we present simulation results for the case when the large scale fading parameters are not known perfectly, but obtained through estimation. We assume that the BS collects N processed pilots from each user to perform this estimation. Specifically, denoting each channel realization by $\boldsymbol{h}_k^i$, the processed pilot signals received by the BS for each user can be written as

$$\boldsymbol{y}_k^i = \sqrt{\tau_p p_k}\boldsymbol{h}_k^i + \boldsymbol{w}_k^i, i = 1, \ldots, N, \tag{3.19}$$

where $\boldsymbol{y}_k^i$ is the processed received signal, τ_p is the length of the pilot, p_k is the signal power and $\boldsymbol{w}_k^i$ is additive noise with variance 1. Then we estimate β_k as follows:

$$\hat{\beta}_k = \frac{\sum_{i=1}^{N} ||\boldsymbol{y}_k^i||^2 - MN}{MN\tau_p p_k}. \tag{3.20}$$

This estimate is justified by the fact that

$$\begin{aligned} ||\boldsymbol{y}_k^i||^2 &\approx \tau_p p_k ||\boldsymbol{h}_k^i||^2 + ||\boldsymbol{w}_k^i||^2 \\ &\approx \tau_p p_k \beta_k M + M. \end{aligned} \tag{3.21}$$

Figure 3.5 shows the minimum SE achieved by our max-min scheme with the proposed estimator of the large-scale fading parameters. The number of observations is $N = 10$ and the

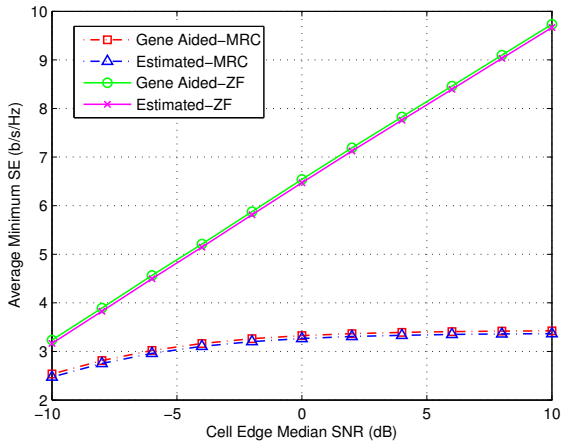


Figure 3.5: Average minimum SE with $M = 100$, $K_0 = 10$, $T = 200$, $N = 10$, $R = 1000$ m for estimated large scale fading parameters.

median SNR at the cell edge ranges from -10 dB to 10 dB; all other simulation parameters are the same as in the previous subsection. The estimated βs are treated as the true βs in the optimization (marked as 'Estimated'). The performance is compared with the case when the βs are known perfectly (marked as 'Genie Aided'). We observe that with the simple, above suboptimal estimator and the small number of training symbols, the performance degradation is almost negligible. We conclude that our scheme shows significant robustness against estimation errors in the large-scale fading parameters.

3.1.6 Conclusion

We considered the optimal joint pilot and data power allocation problems in single cell uplink MaMi systems with MRC or ZF detection. It was first proved that the optimal length of the training interval equals the number of users. Using the SE as performance metric and setting a total energy budget, the power control was formulated as optimization problems for two different objective functions: the weighted minimum SE and the weighted sum SE. The optimal power control policy was found for the case of maximizing the weighted minimum SE. The optimal power control parameters were shown to be the same for MRC and ZF. For maximizing the sum SE a convex reformulation was found and efficient solution algorithms were developed. In [11], we show that these methods can also been extended to handle the case of correlated fading, although a complete treatment of all aspects of that case is left for future work.

Simulation results demonstrated the advantage of joint optimization over both pilot and data power, and how the two objectives behave at low and high cell-edge SNRs. With MRC we have a clear choice to make between max-min and sum SE, which is dependent on the system requirements. With ZF we can maximize sum SE without sacrificing much in min SE. The need of joint pilot and data power control is particularly important at low SNR, while at high SNR optimizing only data power seems to be good enough. Since multi-cell systems are interference-limited, we predict that we will get results similar to the low SNR results, particularly if a large pilot reuse factor is used to get single-cell-like estimation quality. The numerical results were also justified by a theoretical analysis in the low and high SNR regime. This analysis showed that the gain is more substantial when the number of users, K, is small compared to the length of the coherence interval, T.

3.2 Downlink Power Control and Link Adaptation: Throughput-Fairness Trade-Offs

This section deals with power control in downlink single-cell MaMi systems, under practical link adaptation. In a traditional macro-cell some UEs are close to the BS while others are located at the cell-edge, suffering from larger path loss and different large scale fading conditions. Consequently, UEs are subject to different SNRs which in traditional macro-cellular network lead to uneven throughput allocation between inner-cell UEs and cell-edge UEs. To guarantee the highest possible throughput for each UE, link adaptation can be implemented. In LTE, for example, an adaptive modulation and coding scheme (AMC) is adopted [18]. UEs estimate the channel SNR and feed back the corresponding channel quality index (CQI). The BS selects the most efficient modulation and coding scheme, based on the acquired CQI, in order to maximize throughput while reducing retransmissions.

Although link adaptation maximizes the total throughput of the network under predetermined power levels, it does not provide a uniform service to all the UEs. Schedulers, used to allocate frequency resources to the UE, can improve the fairness of a system. However, optimizing schedulers results in a trade-off between network throughput and user fairness [12]. In LTE proper resource allocation over different subcarriers is fundamental in order to take into account small scale fading and distribute frequency resources over users. In contrast, the channel hardening effect in MaMi systems makes the role of the scheduler less crucial because when all the UEs experience a similar, averaged, channel, it is expected that fairness and homogeneity among users is improved without having to consider the frequency dimension. Moreover, the whole spectrum can be simultaneously allocated to all UEs.

Several papers have shown that MaMi can deliver uniformly high throughput to all its users regardless of their positions in the coverage area using proper power control [8,33,60]. However, these papers do not show whether MaMi, similarly to traditional networks, is facing a trade-off between throughput and fairness or can really provide UE fairness at no cost on total network throughput. This trade-off is investigated in this section by comparing several power allocation algorithms. By relying on the channel hardening property of MaMi to compensate for small-scale fading, only slowly-varying large-scale power control is used. Simulations are conducted in both limited and heavy large scale fading scenarios.

3.2.1 System Model

Let us consider the downlink of a multiuser OFDM MaMi system. The BS is equipped with M antennas and serves simultaneously K single antenna users. The received downlink signal $\mathbf{y}_f \in \mathbb{C}^{K\times 1}$ is modeled as

$$\mathbf{y}_f = \alpha \mathbf{L}^{1/2}\mathbf{H}_f\mathbf{W}_f\mathbf{P}^{1/2}\mathbf{s}_f + \mathbf{z}_f, \tag{3.22}$$

where the index f represents the subcarrier. $\mathbf{L}^{1/2}$ is a diagonal matrix with entries $0 \leq \sqrt{l^{(1)}}, \sqrt{l^{(2)}}, \ldots, \sqrt{l^{(K)}} \leq 1$ which represent the inverse path loss (the channel gain) relative to each UE, whose values change very slowly over time. To fairly assess different MaMi scenarios, we ensure that $\mathbb{E}\left[||\alpha\mathbf{L}^{1/2}||^2\right] = K$ introducing the constant $\alpha = \sqrt{K/\mathrm{tr}(L)}$. The channel between the BS and the UEs is $\mathbf{H} \in \mathbb{C}^{K\times M}$. $\mathbf{H}$ is modeled as $\mathcal{CN}(0,\mathbf{I})$ and it is normalized such that $\mathbb{E}\left[||\mathbf{H}||_F^2\right] = KM$. The system is based on time-domain duplexing, including an uplink pilot phase in order to let the BS estimate the channel. Based on this knowledge, the precoder $\mathbf{W} = [\mathbf{w}^{(1)}|\mathbf{w}^{(2)}|\cdots|\mathbf{w}^{(k)}]$, where $\mathbf{w}^{(k)}$ is the M-dimensional beamforming vector, and the diagonal power allocation matrix $\mathbf{P} \in \mathbb{R}^{K\times K}$ are computed. $\mathbf{W}$ and $\mathbf{P}$ are designed under

power constraints. We assume $\text{tr}\left(\mathbf{W}^H\mathbf{W}\right) \leq K$ and $\text{tr}(\mathbf{P}) \leq K$. $\mathbf{W}$ is designed as a Zero Forcing (ZF) precoding which is the pseudo-inverse of $\mathbf{H}$ [48]. $\mathbf{s} \in \mathbb{C}^{K\times 1}$ are the actual stochastic zero-mean data symbols and we further assume $\mathbb{E}[||\mathbf{s}||^2] = 1$. The additive term $\mathbf{z}_f \in \mathbb{C}^{K\times 1}$ is the independent receiver noise generated with distribution $\mathcal{CN}(0, \sigma_z^2 I)$.

In the following, we omit the frequency index without loss of generality since in MaMi channel variations are negligible over the frequency domain [8]. The k-th UE receives the symbol

$$y^{(k)} = \alpha\sqrt{l^{(k)}}\mathbf{h}^{(k)^H}\mathbf{w}^{(k)}\sqrt{p^{(k)}}s^{(k)} + \alpha\sum_{j\neq k}\sqrt{l^{(k)}}\mathbf{h}^{(k)^H}\mathbf{w}^{(j)}\sqrt{p^{(j)}}s^{(j)} + z, \tag{3.23}$$

where the first term represents the desired received precoded symbol at the user k and the second term represents the interfering symbols sent to the others UEs in the system. The ergodic SINR $\gamma^{(k)}$ of the k-th UE is then given by

$$\gamma^{(k)} = \frac{\alpha^2 l^{(k)} p^{(k)} \mathbb{E}\left[|\mathbf{h}^{(k)^H}\mathbf{w}^{(k)}|^2\right]}{\alpha^2 \mathbb{E}\left[\sum_{j\neq k} l^{(k)} p^{(j)} |\mathbf{h}^{(k)^H}\mathbf{w}^{(j)}|^2\right] + \sigma_z^2}, \tag{3.24}$$

which reduce when perfect channel estimation is considered when computing ZF, such that $\mathbf{h}^{(k)^H}\mathbf{w}^{(j)} = 0$, to the following effective SNR:

$$\gamma^{(k)} = \frac{\alpha^2 l^{(k)} p^{(k)} \mathbb{E}\left[|\mathbf{h}^{(k)^H}\mathbf{w}^{(k)}|^2\right]}{\sigma_z^2}. \tag{3.25}$$

Based on $\gamma^{(k)}$ estimation, link adaptation can be used at the BS to select the proper modulation and coding scheme (MCS) and power allocation.

3.2.2 Link Adaptation Procedure

In contrast to the ideal link adaptation assumed in the ergodic rate analysis, carried out in the vast majority of literature on MaMi, we consider practical link adaptation. It makes sure each UE gets the best possible throughput by power and MCS allocation while respecting overall power and possibly fairness constraints. The maximization problem can be stated as:

$$\begin{aligned} \underset{p^{(k)}}{\text{maximize}} \quad & C(\gamma^{(k)}) \\ \text{subject to} \quad & \text{tr}\left(\mathbf{W}^H\mathbf{W}\right) \leq K, \\ & \text{tr}(\mathbf{P}) \leq K, \\ & \text{CWER} \leq \nu. \end{aligned} \tag{3.26}$$

The maximization of the network throughput C is constrained by the total output power and by the desired quality of service (QoS) which is here quantified as codeword error rate (CWER).

Based on the allocated power, the BS estimates the received SNR for each UE. The predicted SNR is used in the link adaptation procedure to choose the MCS which maximizes the throughput of each UE. In practice this is done by selecting the highest MCS which satisfies the CWER requirement. Figure 3.6 shows the CWER as function of SNR, for each transmission mode in a single UE MaMi setting, assuming $M = 100$ antennas. A Rayleigh multi-path channel is assumed. Thanks to the channel hardening effect the small scale fading is averaged

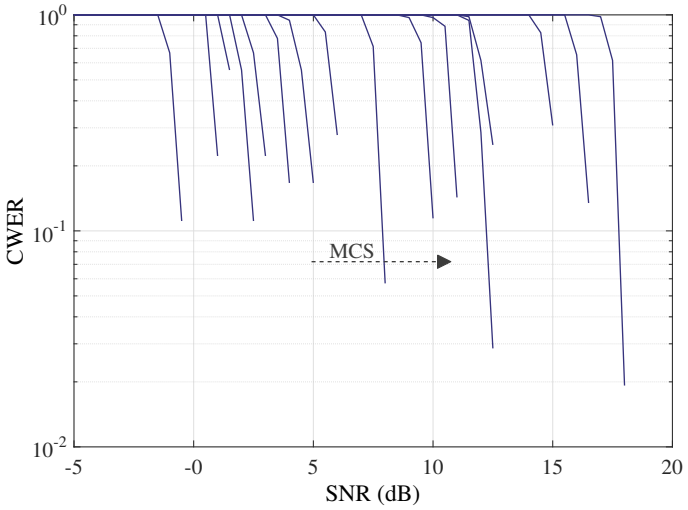


Figure 3.6: CWER as function of received SNR for the different MCS listed in Table 3.1.

out and the same threshold would be obtained in a single antenna AWGN channel given that the MaMi array gain is not included here (the actual SNR can be 20 dB lower than those values with 100 antennas). From Figure 3.6 we can extract the SNR threshold for error-free operation for each MCS. Values are reported on Table 3.1 which shows the modulation order and code rate used, the achieved spectral efficiency and the SNR threshold SNR for each MCS index.

3.2.3 Downlink Power Allocation Schemes

The total transmission power is shared among the UEs based on the selected power allocation scheme (PAS). Thanks to the channel hardening effect, the design of the power allocation matrix $\mathbf{P}$ is based only on the large-scale fading characteristics. Moreover, the same power control is applied over the whole spectrum. Depending on the design of the power allocation matrix $\mathbf{P}$, different performance can be achieved in terms of throughput and fairness but the trade-off is not straightforward. Cell-edge and shadowed UEs are the bottleneck of a communication system as they do not contribute enough to the total throughput of the system. Hence, the most intuitive procedure to optimize the total throughput of the network would be to sacrifice fairness by not equalizing the SNRs over all UEs, but rather dropping the weaker UEs while allocating more power to the best UEs. However, with simple power control, MaMi can ensure the same throughput experience to all the UEs in a cell, given that small-scale fading is removed in MaMi. To study the trade-off between fairness and sum throughput, the different PAS described hereunder are compared.

Inverse Power Allocation

The inverse power allocation strategy aims to compensate the path loss and large scale fading and to guarantee equal received SNR, thus equal quality of service (QoS) to all the UEs present in the cell. UEs located in the cell-edge will observe the main benefit of this technique as more power is allocated to the weaker UEs and less power to the most favorable UEs such that the received SNRs of all the links are equivalent. The power allocation matrix can be expressed as

$$\mathbf{P} = \eta \mathbf{L}^{-1}, \tag{3.27}$$

Table 3.1: Modulation and coding rate mapping.

MCS	Modulation	Code Rate	Spectral eff.	SNR_{th}
1	BPSK	1/2	0.5	0
2	BPSK	2/3	0.66	1.5
3	BPSK	3/4	0.75	2
4	QPSK	1/2	1	2.5
5	QPSK	2/3	1.33	4.5
6	QPSK	3/4	1.50	5.5
7	QPSK	5/6	1.67	6.5
8	16-QAM	1/2	2	8.5
9	16-QAM	2/3	2.67	10.5
10	16-QAM	3/4	3	11.5
12	16-QAM	5/6	3.33	13
13	64-QAM	2/3	4	15.5
14	64-QAM	3/4	4.5	17
15	64-QAM	5/6	5	18

where $\eta = \dfrac{K}{\text{tr}(\mathbf{L}^{-1})}$ is the normalization factor to ensure the total power constraint.

Max-Min Power Allocation

The max-min power allocation has the objective to maximize the minimum achieved rate for each user. Instead of maximizing the total throughput of the system as in (3.26), the max-min algorithm maximizes the throughput of the weakest link to improve fairness:

$$\begin{aligned} \underset{p^{(k)}}{\text{maximize}} \quad & \min(\gamma^{(1)}, \ldots, \gamma^{(k)}) \\ \text{subject to} \quad & \text{tr}\left(\mathbf{W}^H\mathbf{W}\right) \leq K, \\ & \text{tr}\left(\mathbf{P}\right) \leq K, \\ & \text{CWER} \leq \nu. \end{aligned}$$

When the available power is not sufficient to satisfy the minimum requirement for each UE, namely MCS 1, only the strongest UEs are allocated, while one or more UEs with the weakest channels are sacrificed. Max-min and inverse PAS share the objective to equalize the received SNR. When the lower MCS can be guaranteed to each UE, the max-min algorithm converges to the same power allocation in equation (3.27).

Waterfilling

The waterfilling solution has been proven to be the optimal solution to maximize the network throughput by allocating more power on the most favorable channels [61]. The original water-filling solution was derived with the underlying assumption that each spatial stream can support any possible rate which is not true in practical communication system where, usually, a finite set of MCS are available as in Table 3.1. The power allocated to the k-th UE is:

$$p^{(k)} = \left(\mu - \left(\frac{\alpha^2 l^{(k)} \mathbb{E}\left[|\mathbf{h}^{(k)}|^2\right]}{\sigma_z^2}\right)^{-1}\right)^+, \tag{3.28}$$

where μ is the water level chosen to satisfy the power constraint and $(x)^+ \triangleq \max(0, x)$ denotes the positive part of x. The water-filling solution proposed here slightly differs to the traditional solution which is based on singular value decomposition (SVD) of the channel. It allows a simpler algorithm which does not rely on full channel knowledge as in the SVD case but only on the knowledge of the SNR based on the gain of each UE channel, which in MaMi is not affected by small scale fading.

Equal Power Allocation

The available power is equally distributed between the UEs, independently on the channel conditions. The power allocation matrix is simply

$$\mathbf{P} = \mathbf{I}, \tag{3.29}$$

which is the least complex PAS. However, it has been shown to be a suboptimal solution for network throughput optimization in conventional systems [61].

3.2.4 Fairness and Throughput Analysis

In this section, the impact of the different PAS on throughput and fairness in a MaMi system is analyzed using system level simulations. To evaluate the sum rate of the system we define the throughput of the k-th UE as

$$T^{(k)} = (1 - \mathrm{CWER}^{(k)}) N_b^{(k)} r_b^{(k)}, \tag{3.30}$$

where CWER is the codeword error rate, N_b is the modulation order and r_b is the coding rate. To evaluate the fairness of the system we use Jain's fairness index:

$$F = \frac{\left(\sum_{k=1}^{K} T^{(k)}\right)^2}{K \sum_{k=1}^{K} (T^{(k)})^2}, \tag{3.31}$$

which measures the similarity of the achieved throughput over different UEs. When all UEs get the same throughput then the fairness index is 1 which means that the system is 100% fair [26].

Based on LTE we consider a 20 MHz OFDM system with 2048 subcarriers out of which 1200 are actively allocated, the others serving as guard band. The system uses state-of-the-art LDPC coding, derived from the channel coding of the IEEE 802.11ac standard, with codeword length assumed to be 1944 bits. The channel $\mathbf{H}$ is a time-domain Rayleigh with 20 taps of equal expected energy, uncorrelated over both UE and antenna dimensions. This channel is attenuated by user coefficients l_k, describing the large scale fading for each user k. The large scale fading is assumed to be log-normal distributed with $10\log_{10}(l_k) \sim \mathcal{N}(0, \sigma_{ls})$. Results are averaged for 50 channel realizations.

The attenuation due to the large scale fading in a practical scenario strongly depends on the environment and it can assume different values. We evaluate the system behavior in both medium and heavy large scale fading scenarios, represented as $\sigma_{ls} = 5$ or 10 dB, respectively.

Medium Large Scale Fading Scenario

Results demonstrate that in MaMi systems, even though small-scale fading is averaged through channel hardening which eases power allocation and scheduling, fairness is still achieved at the price of throughput reduction, as it is the case for other systems. Figure 3.7 shows Jain's

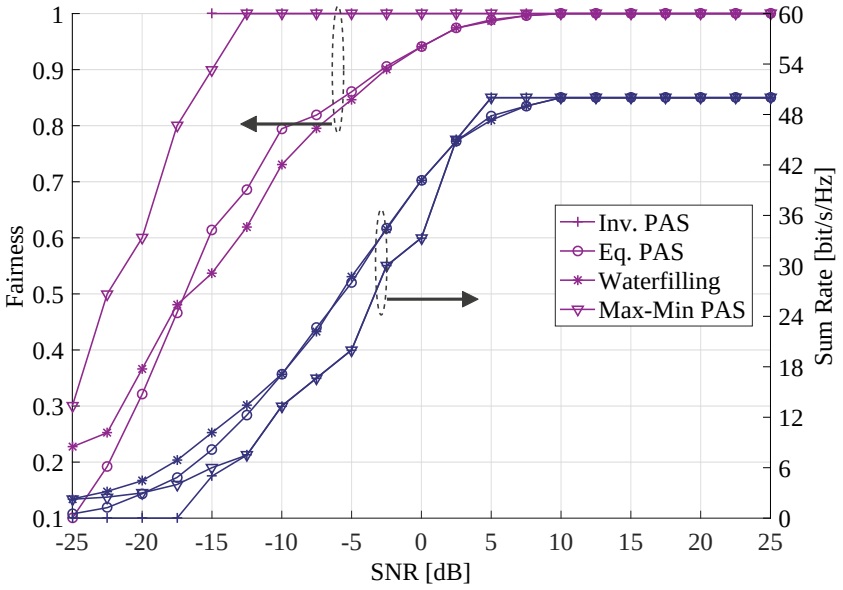


Figure 3.7: Fairness on the left y axis and sum rate on the right y axis in MaMi with $M = 100$, $K = 10$, $\sigma_{ls} = 5$, for different power allocation strategies.

fairness index and total throughput of a MaMi system with $M = 100$ antennas, serving $K = 10$ UEs, for each of the power allocation schemes discussed in Section 3.2.2, assuming $\sigma_{ls} = 5$. The SNR normalization does not include the benefit coming from a $10\log_{10}(M) = 20\,\text{dB}$ MaMi array gain. Waterfilling sets the throughput upper bound, while inverse PAS guarantees full fairness.

At low SNR, inverse power allocation is not able to guarantee the minimum MCS to all the UEs and hence it does not deliver any throughput. On the contrary, the three other algorithms deliver a similar positive throughput but different fairness levels are achieved. Max-min clearly outperforms the other PAS in terms of fairness. For low SNR fairness maximization can be the best strategy as it only creates a negligible throughput reduction. For example at SNR $= -15\,\text{dB}$ max-min provides 90% fairness while waterfilling reduces the fairness to 53%. Even though Jain's index provides some insight into the overall system fairness, it does not help in identifying detailed differences between UEs.

In order to get more insight into individual UEs, Figure 3.8 illustrates the throughput analysis for each UE at a very low SNR $= -15\,\text{dB}$. In the figure, the UEs are sorted in decreasing order based on the received power, that is, UE 1 experiences the best channel condition while UE 10 is subject to the largest fading. It is evident that when enforcing fairness, more UEs are multiplexed and given a similar QoS but the throughput of the most favorable UEs dramatically decreases. The throughput of UE 1 drops from 3 bit/s/Hz using waterfilling to 0.5 bit/s/Hz using inverse PAS, but all the UEs can be served instead of only 7 out of 10. Max-min prefers to sacrifice UE 10 ensuring fairness to the other 9 UEs and providing up to 0.66 bit/s/Hz. Enforcing fairness with max-min PAS also enables UE 6 and UE 7 to experience a slightly improved throughput as compared to waterfilling and equal PAS.

Focusing on intermediate SNRs, waterfilling and equal PAS have similar performance while max-min converges to the inverse PAS. At SNR=0 dB the difference in system fairness decreases. Waterfilling and equal PAS provide indeed fairness higher than 90% while improving the sum rate by some 5 to 7 bit/s/Hz with respect to inverse and max-min PAS. This suggests that throughput or fairness maximization is possible while facing a limited reduction of the other

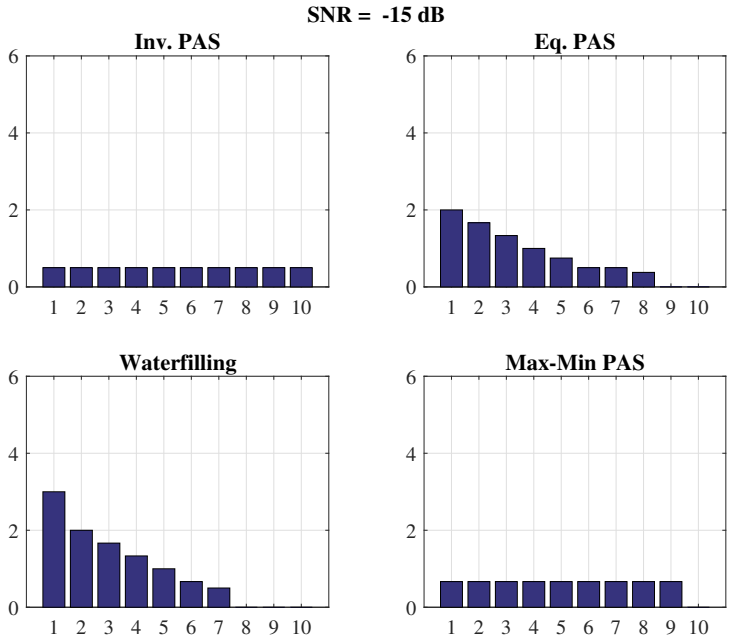


Figure 3.8: Average throughput [bit/s/Hz] achieved for each UE at SNR $= -15$ dB when $\sigma_{ls} = 5$ for the different PAS.

dimension. Figure 3.9 shows the detailed analysis for each UE confirming that both optimization are feasible and high throughput is guaranteed in both cases. At higher SNRs such as 10 dB the system is saturated and all the schemes assign the highest MCS to all the UEs.

Heavy Large Scale Fading Scenario

Assuming a log-normal variance of $\sigma_{ls} = 10$ dB, which is more representative of what can be observed in a real cell, Figure 3.10 shows the sum rate and fairness of the system. Compared to Figure 3.7, the higher attenuation introduced in the system causes a throughput degradation as well as a lower fairness for all the PAS. As we increase the large scale variance, the gap in throughput for low SNR between max-min and equal PAS becomes more evident. Also the throughput losses of max-min and inverse PAS compared to waterfilling and max-min PAS become larger.

For low SNR, both optimization directions are possible. Max-Min has to sacrifice nearly 50% of the throughput with respect waterfilling in order to increase the fairnes from 35% to 70%. Figure 3.11 shows the details for each UE at SNR=-15dB. Inverse PAS is not able to deliver any throughput at that SNR. Max-min can serve 7 UEs instead of 5 selected by equal PAS and waterfilling.

For higher SNR fairness maximization is not the best strategy as the system is subject to a huge throughput loss. For example at SNR=0 dB, max-min and inverse PAS lose nearly 14 bit/s/Hz as compared to waterfilling and equal PAS in order provide full fairness. Figure 3.12 confirms that the losses are too heavy and optimizing for fairness 8 out of the 10 UEs are subject to a throughput reduction.

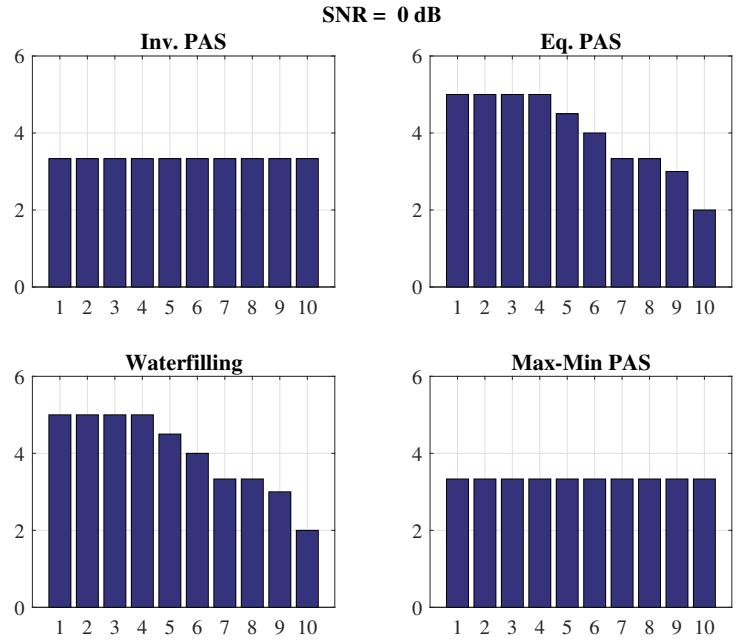


Figure 3.9: Average throughput [bit/s/Hz] achieved for each UE at SNR = 0 dB when $\sigma_{ls} = 5$ for the different PAS.

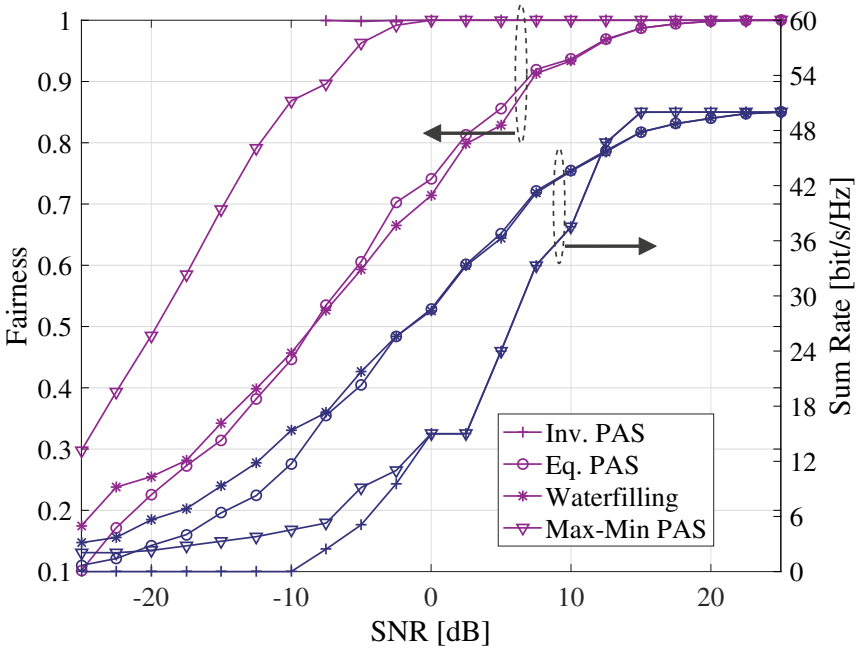


Figure 3.10: Fairness and sum rate in 100x10 MaMi system when $\sigma_{ls} = 10$ for different power allocation strategies.

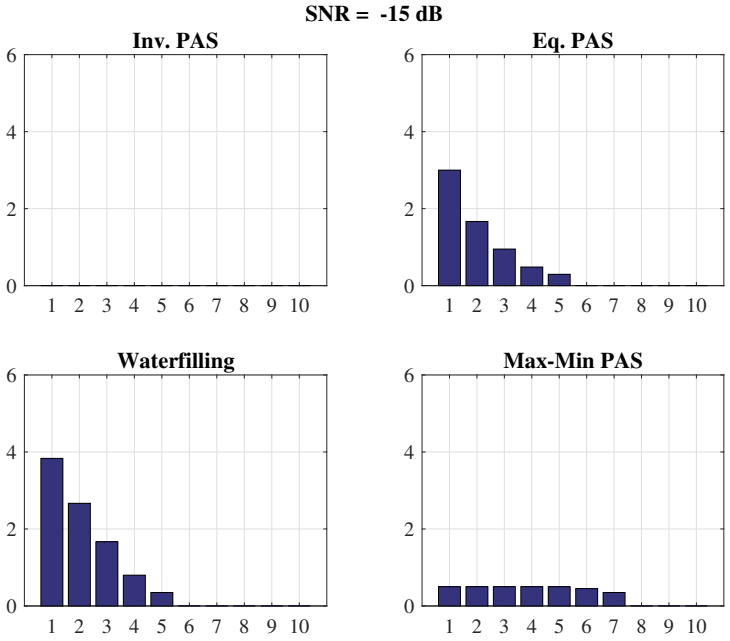


Figure 3.11: Average throughput [bit/s/Hz] achieved for each UE at SNR $= -15$ dB when $\sigma_{ls} = 10$ for the different PAS.

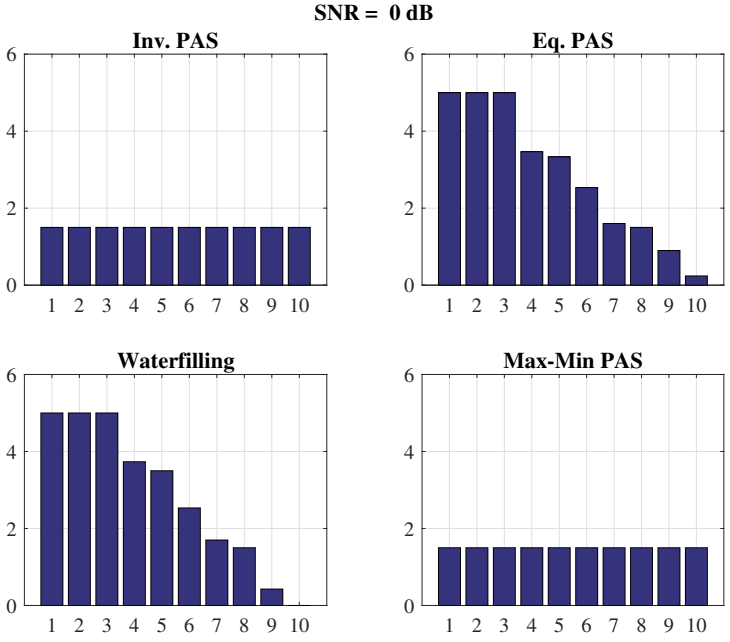


Figure 3.12: Average throughput for each UE at SNR $= 0$ dB when $\sigma_{ls} = 10$.

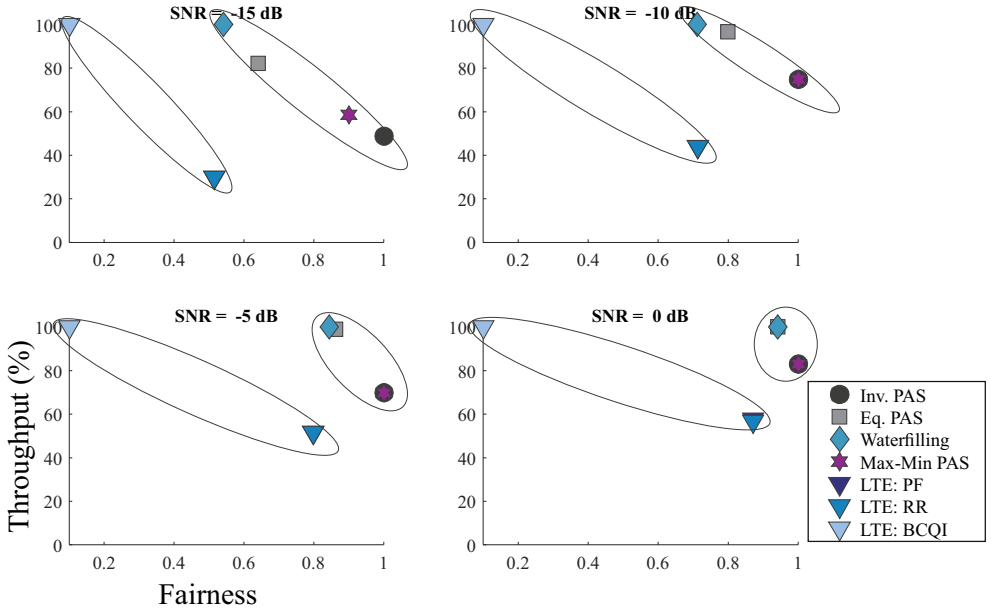


Figure 3.13: Relative throughput losses in percentage required to improve system fairness in LTE and MaMi when $\sigma_{ls} = 5$. Losses are calculated with respect to BCQI in LTE and waetrfilling in MaMi.

Comparison with Traditional Networks

Similarly to traditional networks, MaMi systems are facing the trade-off between throughput and fairness. Let us study the throughput losses of the system and propose a comparison with an LTE cellular network. The comparison is not straightforward as MaMi and LTE operate in a different way. While MaMi eliminates frequency-domain variations, LTE uses schedulers to allocate the UEs over different frequency resources. The scheduler employed determines the fairness of the system. However, it is important to understand whether MaMi, beyond its well known benefits, also simplifies the trade-off between fairness and throughput.

The reference LTE system uses different schedulers. Best CQI (BCQI) aims at throughput optimization. Proportional Fairness aims at fairness optimization and Round Robin assigns the resources to the UEs in a circular fashion without specific throughput and fairness optimization. The reader is refered to [12] and references therein for a more detailed description of the LTE system and schedulers.

Figure 3.13 shows, for $\sigma_{ls} = 5$, the relative throughput losses required to improve the system fairness in LTE and in MaMi with respect to the maximum throughput achieved by each of the two systems. The maximum throughput is provided by the BCQI scheduler and waterfilling PAS respectively in LTE and MaMi. We observe that MaMi has to sacrifice a smaller throughput fraction, independently on the SNR, to maximize the fairness. For example, MaMi is subject to 50% of throughput reduction while LTE has to sacrifice 70% of throughput, when SNR=-15 dB to obtain the best fairness. Moreover, only MaMi is able to provide 100% fairness.

Figure 3.14 shows the heavy fading scenario $\sigma_{ls} = 10$. For higher large scale fading variance MaMi experiences the same benefit as compared to LTE. However, as already pointed out, even MaMi experiences huge losses to improve fairness and more than 50% of throughput reduction is required independently on the SNR, when targeting maximum fairness.

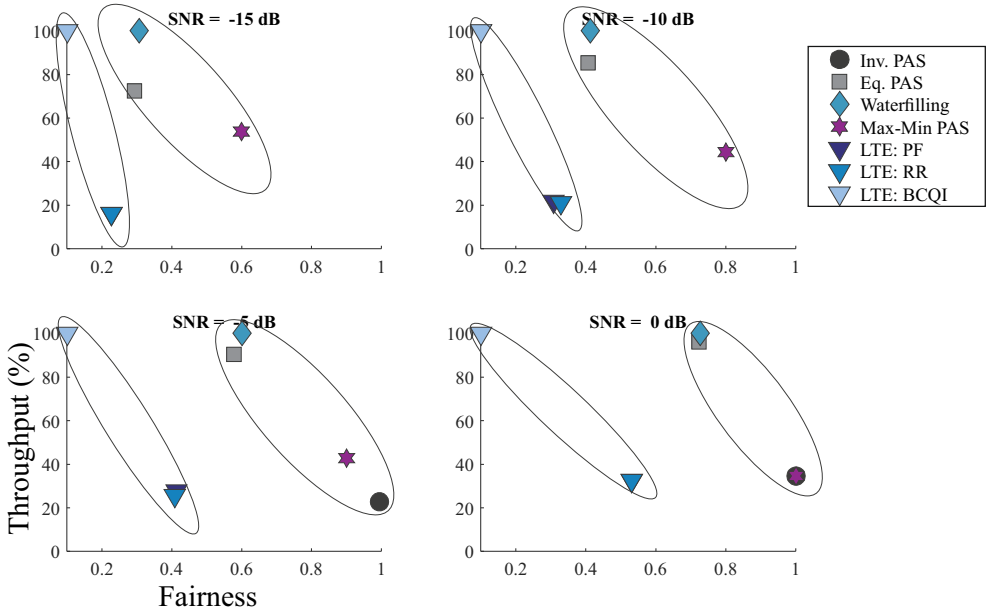


Figure 3.14: Relative throughput losses required to improve system fairness in LTE and MaMi when $\sigma_{ls} = 10$.

3.2.5 Conclusion

This section has investigated the relation between throughput and fairness in MaMi systems when performing long term power control in the downlink, based on large scale fading modeled with a log-normal distribution. Four different power allocation schemes were considered. Two of them, namely inverse power allocation and max-min power allocation, share the objective of fairness maximization. Waterfilling and equal power allocation, on the other hand, aim at maximizing the throughput of the network. Simulations show that, as in traditional networks, the trade-off between fairness and high throughput is still critical in MaMi. System fairness is only achieved at the price of a throughput reduction. For low log-normal variance, simulations suggest that it would be possible to optimize fairness or throughput with minimal impact on the performance of the other dimension. Both optimization can be used depending on the QoS network operator would like to provide. For large log-normal channel fading variations, the throughput cost of enforcing fairness is very large. More than 50% of throughput reduction is observed independently on the SNR.

3.3 Out-of-Band Radiation from MaMi Transmissions

MaMi must be designed with low-cost components, to limit the implementation cost and power consumption when having many RF chains. However, these low-cost components are suffering from imperfections and non-idealities introducing distortion in the transmitted signal. Out-of-band (OOB) radiation is the undesired power of a signal at frequencies outside the allocated frequency band. Such power usually arises from nonlinear circuits and can potentially disturb concurrent transmission in adjacent bands. Therefore, many standards, e.g., LTE [1], limit the amount of out-of-band radiation that is allowed to be emitted. Based on the LTE specifications, OOB radiation should not exceed $-13\,\mathrm{dBm/MHz}$ [2].

Traditionally, OOB radiation has been measured on a per-antenna basis. MaMi uses channel-based MIMO precoding to make the signals add constructively at the intended UE. While in-

band interferences do not recombine constructively [44], and completely saturated PAs only lead 1.5 dB degradation in terms of BER at the intended UEs [14], it is however not clear what the impact of precoding is on the OOB interference at the UE, or more importantly, at any possible location. In a MIMO setting, where many antennas transmit together, measuring the OOB on a per-antenna basis is not necessarily a sensible approach. The radiated power from the transmitting antennas builds up constructively or destructively in the air and the amount of OOB radiation that disturbs transmission in adjacent bands can thus be greater or smaller than what was emitted from any single antenna. Not to disturb other communication, the OOB radiation should therefore instead be limited on the basis of what is actually received by the UEs of adjacent bands.

The phenomenon of OOB radiation in single-antenna systems has been thoroughly studied before, see for example [21]. Methods developed to mitigate OOB radiation, such as digital pre-distortion, are also well known [30]. Many of these methods are, however, impractical in a MaMi system due to the great number of radio chains. In this section, we study the spatial distribution of the OOB radiation in order to gain some fundamental insight into its behavior in multi-antenna systems with nonlinear amplifiers, and to understand how it should be appropriately measured. This will be an important aid for the standardization process of future communication systems. Further details are available in the MAMMOET publication [43].

We show that, in MaMi, OOB does not recombine constructively, in several different scenarios, even using fully saturated PAs. We first give a description of the received power spectrum density (PSD) with the aim to verify whether coherent combination of the signal outside the band of interest occurs. Then we verify our analysis using numerical simulations. In particular, we quantify the interference received by a random UE operating in an adjacent band, exploiting the new concept of MIMO Adjacent Channel Leakage Ratio (ACLR).

3.3.1 System Model

The BS transmits the digital signals $\mathbf{x}[n] \triangleq (x_1[n], \ldots, x_M[n])^{\mathrm{T}}$ on its M antennas by pulse-amplitude modulating them with the pulse $p(\tau)$ into the analog signal

$$\mathbf{x}(t) \triangleq \begin{pmatrix} x_1(t) \\ \vdots \\ x_M(t) \end{pmatrix} = \sum_n \mathbf{x}[n] p(t - nT + \Psi), \tag{3.32}$$

where T is the symbol duration and Ψ is a random variable[1] that is uniformly distributed on the interval $0 \leq \Psi < T$. The bandwidth of the pulse $p(\tau)$ is assumed to be equal to the bandwidth B that is allocated to the BS. The signal $\mathbf{x}(t)$ is amplified to transmit power into $\mathbf{y}(t) \triangleq (y_1(t), \ldots, y_M(t))^{\mathrm{T}}$, where the amplification is modeled as

$$y_m(t) = \sum_{p=1}^{P} \int_{-\infty}^{\infty} b_{mp}(t - \tau) x_m(\tau) |x_m(\tau)|^{2(p-1)} \mathrm{d}\tau, \tag{3.33}$$

where $b_{mp}(\tau)$ is the impulse response of the nonlinear p-th order term of the m-th amplifier [30]. Note that this polynomial model is a special case of the more general Volterra series [53]: all kernels outside the diagonal are set to zero and all dynamic memory is removed. Nonlinear memory effects are typically much weaker (in the order 20 dB) than the direct nonlinearities.

[1] The introduction of Ψ is a way to make pulse-amplitude modulation preserve stationarity [49]; it only appears in this equation.

Their effect is therefore secondary. Further, note that any physical nonlinearity can be approximated by a polynomial—the nonlinearity in (3.33) could therefore also model other hardware architectures.

The received signal $r_\theta(t)$ at a spatial point θ is given by

$$r_\theta(t) = \sqrt{\beta_\theta} \int_{-\infty}^{\infty} \mathbf{h}_\theta^{\mathrm{T}}(\tau)\mathbf{y}(t-\tau)\mathrm{d}\tau, \tag{3.34}$$

where $\mathbf{h}_\theta(\tau)$ is the impulse response of the small-scale fading from the array to the point θ and $\beta_\theta \in \mathbb{R}^+$ a large-scale fading coefficient, which models signal attenuation due to both distance and shadowing.

3.3.2 BS Radiation Pattern

We assume that the BS is serving K single-antenna UEs and that the M transmit signals are produced by linear precoding as

$$\mathbf{x}[n] = \sum_\ell \mathbf{W}[\ell]\mathbf{D}_\xi^{1/2}\mathbf{s}[n-\ell], \tag{3.35}$$

where $\mathbf{s}[n] \triangleq (s_1[n], \ldots, s_K[n])^{\mathrm{T}}$, $s_k[n]$ is the symbol to be transmitted to UE k at symbol time n, $\mathbf{D}_\xi \triangleq \mathrm{diag}(\xi)$ is a diagonal matrix with the relative power allocations $\xi \triangleq (\xi_1, \ldots, \xi_K)^{\mathrm{T}}$, for which $\xi_k \in \mathbb{R}^+$ and $\sum_{k=1}^K \xi_k = 1$, on its diagonal and $\{\mathbf{W}[\ell]\}$ is the impulse response of the precoder.

The discrete-time channel is given by

$$\mathbf{H}[\ell] \triangleq \Big(p(\tau) \star \mathbf{H}(\tau) \star p^*(-\tau)\Big)(\ell T), \tag{3.36}$$

where $\mathbf{H}(\tau) \triangleq (\mathbf{h}_{\theta_1}(\tau), \ldots, \mathbf{h}_{\theta_K}(\tau))^{\mathrm{T}}$ and θ_k is the location of UE k. The simplest linear precoder is the MR precoder, whose impulse response is given by $\mathbf{W}[\ell] = \alpha\mathbf{H}^{\mathsf{H}}[-\ell]$, where α is a real-valued normalization factor that is chosen such that $\sum_\ell \|\mathbf{W}[\ell]\|_{\mathsf{F}}^2 = K$. Another common precoder is zero-forcing, which we use in this section; see e.g. [9, 45] for an exact definition. We assume that the BS knows $\mathbf{H}[\ell]$ perfectly.

Further, we assume that $\mathbf{s}[n]$ is a circularly symmetric i.i.d. stationary process, for which

$$\mathbf{R}_{\mathbf{ss}}[\nu] = \begin{cases} \mathbf{I}_K, & \text{if } \nu = 0, \\ \mathbf{0}_K, & \text{otherwise.} \end{cases} \tag{3.37}$$

Because of the multiuser precoding in (3.35) and of the central limit theorem, the distribution of the discrete-time transmit signals $\mathbf{x}[n]$ is close to circularly symmetric complex Gaussian. Note that this is true independently of whether OFDM or single-carrier transmission is used and independently of the order of the symbol constellation [45]. The autocorrelation function of the unamplified transmit signals $\mathbf{x}[n]$ in a given coherence interval (the expectation is taken with respect to the symbols and is conditioned on the small-scale fading) is

$$\mathbf{R}_{\mathbf{xx}}[\nu] = \mathbb{E}\left[\Big(\sum_\ell \mathbf{W}^*[\ell]\mathbf{D}_\xi^{1/2}\mathbf{s}^*[n-\ell]\Big)\Big(\sum_{\ell'} \mathbf{s}^{\mathrm{T}}[n+\nu-\ell']\mathbf{D}_\xi^{1/2}\mathbf{W}^{\mathrm{T}}[\ell']\Big) \;\Big|\; \{\mathbf{H}[\ell]\}\right] \tag{3.38}$$

$$= \sum_\ell \mathbf{W}^*[\ell]\mathbf{D}_\xi\mathbf{W}^{\mathrm{T}}[\nu+\ell]. \tag{3.39}$$

For example, if maximum-ratio precoding is done, $\mathbf{R}_{\mathbf{xx}}[\nu] = \alpha^2 \sum_\ell \mathbf{H}^{\mathrm{T}}[\ell]\mathbf{D}_\xi \mathbf{H}^*[\ell - \nu]$. The pulse-amplitude modulated $\mathbf{x}(t)$ thus has the autocorrelation function

$$\mathbf{R}_{\mathbf{xx}}(\tau) = \frac{1}{T}\sum_{\nu=-\infty}^{\infty} \mathbf{R}_{\mathbf{xx}}[\nu]\big(p(t) \star p^*(-t)\big)(\tau - \nu T). \tag{3.40}$$

The cross-correlation of the transmit signal is thus

$$\begin{aligned} R_{y_m y_{m'}}(\tau) = \mathbb{E}\Big[\sum_{p=1}^{P}\int_{-\infty}^{\infty} b^*_{mp}(t-\lambda)x^*_m(\lambda)|x_m(\lambda)|^{2(p-1)}\mathrm{d}\lambda \\ \sum_{p'=1}^{P}\int_{-\infty}^{\infty} b_{m'p'}(t+\tau-\lambda')x_{m'}(\lambda')|x_{m'}(\lambda')|^{2(p'-1)}\mathrm{d}\lambda'\Big] \end{aligned} \tag{3.41}$$

$$\begin{aligned} = \sum_{p=1}^{P}\sum_{p'=1}^{P}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b^*_{mp}(t-\lambda)b_{m'p'}(t+\tau-\lambda') \\ \underbrace{\mathbb{E}\Big[x^*_m(\lambda)x_{m'}(\lambda')|x_m(\lambda)|^{2(p-1)}|x_{m'}(\lambda')|^{2(p'-1)}\Big]}_{\triangleq \xi^{(p,p')}_{mm'}(\lambda,\lambda')}\mathrm{d}\lambda\mathrm{d}\lambda'. \end{aligned} \tag{3.42}$$

In the last step, the variable t just translates the integrand. The integral thus does not depend on t and the transmit signals are therefore weak-sense stationary. Because odd moments of Gaussian random variables are zero, we see that $\xi^{(p,p')}_{mm'}(\lambda,\lambda')$ is zero for $m \neq m'$, for all p, p', λ, λ', if the unamplified signals $x_m(t)$ are uncorrelated across the antennas. This means that, when $\mathbf{R}_{\mathbf{xx}}(\tau)$ is diagonal, $\mathbf{R}_{\mathbf{yy}}(\tau)$ is diagonal too.

Using the moment theorem for Gaussian random variables [51], $\xi^{(p,p')}_{mm'}(\lambda,\lambda')$ can be computed for any m, m', p, p', e.g.,

$$\xi^{(1,1)}_{mm'}(\lambda,\lambda') = R_{x_m x_{m'}}(\lambda' - \lambda) \tag{3.43}$$

$$\xi^{(1,2)}_{mm'}(\lambda,\lambda') = 2\sigma^2_{x_m} R_{x_m x_{m'}}(\lambda' - \lambda) \tag{3.44}$$

$$\xi^{(2,2)}_{mm'}(\lambda,\lambda') = 2R_{x_m x_{m'}}(\lambda'-\lambda)\Big(2\sigma^2_{x_m}\sigma^2_{x_{m'}} + \big|R_{x_m x_{m'}}(\lambda'-\lambda)\big|^2\Big), \tag{3.45}$$

where $\sigma^2_{x_m} \triangleq R_{x_m x_m}(0)$. Furthermore, we note that

$$\xi^{(p,p')}_{mm'}(\lambda,\lambda') = \xi^{(p',p)^*}_{m'm}(\lambda',\lambda). \tag{3.46}$$

To study the radiation pattern of the array at different frequencies, we define the frequency response of the channel to the point θ as

$$\tilde{\mathbf{h}}_\theta(f) \triangleq \int_{-\infty}^{\infty} \mathbf{h}_\theta(\tau) e^{-j2\pi\tau f}\mathrm{d}\tau. \tag{3.47}$$

Let $\mathbf{R}_{\mathbf{yy}}(\tau)$ be the matrix, whose (m,m')-th element is $R_{y_m y_{m'}}(\tau)$. The radiation pattern is given by the power spectral density (PSD)

$$\mathbf{S}_{\mathbf{yy}}(f) \triangleq \int_{-\infty}^{\infty} \mathbf{R}_{\mathbf{yy}}(\tau) e^{-j2\pi\tau f}\mathrm{d}\tau \tag{3.48}$$

and the power received at the point θ at frequency f by

$$S_\theta(f) \triangleq \beta_\theta \tilde{\mathbf{h}}_\theta^{\mathsf{H}}(f)\mathbf{S}_{\mathbf{yy}}(f)\tilde{\mathbf{h}}_\theta(f). \tag{3.49}$$

The power radiated by the BS at frequency f is

$$S_{\text{tx}}(f) \triangleq \text{tr}(\mathbf{S}_{\mathbf{yy}}(f)). \tag{3.50}$$

Note that the average received power at a point, where $\tilde{\mathbf{h}}_\theta(f)$ is independent of $\mathbf{S}_{\mathbf{yy}}(f)$ and the fading coefficients are zero-mean and uncorrelated $\mathbb{E}\big[\tilde{\mathbf{h}}_\theta(f)\tilde{\mathbf{h}}_\theta^{\mathsf{H}}(f)\big] = \mathbf{I}_M$, is

$$\mathbb{E}[\, S_\theta(f)\,] = \beta_\theta\, \mathbb{E}[\, S_{\text{tx}}(f)\,]\,. \tag{3.51}$$

The expectation is over all small-scale fading, also over the channels to the UEs, of which the precoding is a function.

3.3.3 Measures of Out-of-Band Radiation

To constrain the amount of OOB radiation of a BS, it is important to be able to easily measure it at the BS. In this subsection, we study the measure conventionally used in single-antenna systems and generalize it to multi-antenna systems. We also propose a framework to analyze how the transmitted signal is beamformed at different frequencies—in-band and out-of-band.

Traditional Single-Antenna Setting

Traditionally, the *transmitted* OOB radiation has been measured at the antenna port in terms of the Adjacent-Channel Leakage Ratio (ACLR). Let $S_{yy}(f)$ be the PSD of the transmit signal in a single-antenna BS. Then the ACLR is defined as [1, 54]:

$$\mathsf{ACLR} \triangleq \frac{\max\{\int_{-3B/2}^{-B/2} S_{yy}(f)\mathrm{d}f, \int_{B/2}^{3B/2} S_{yy}(f)\mathrm{d}f\}}{\int_{-B/2}^{B/2} S_{yy}(f)\mathrm{d}f}. \tag{3.52}$$

This measure compares the amount of power that has leaked over to an immediately adjacent band, which is assumed to have the same width B as the allocated band, to the power in the allocated band. The first term in the numerator of (3.52) is the power in the band just to the left of the allocated band and the second term that in the band to the right. The maximum of the two sideband powers is taken since nonlinear memory effects might create an asymmetric spectrum around the carrier frequency.

We let $\tilde{\mathbf{h}}_\theta(f) = \tilde{h}_\theta(f)$ be the frequency response from the single-antenna BS to the point θ. If the antenna gain is constant over the frequency band $[-3B/2, 3B/2]$, then ACLR can equivalently be measured in a fading environment over the air too as

$$\mathsf{ACLR} = \frac{\max\{\int_{-3B/2}^{-B/2} \mathbb{E}[\, S_\theta(f)\,]\,\mathrm{d}f, \int_{B/2}^{3B/2} \mathbb{E}[\, S_\theta(f)\,]\,\mathrm{d}f\}}{\int_{-B/2}^{B/2} \mathbb{E}[\, S_\theta(f)\,]\,\mathrm{d}f}, \tag{3.53}$$

where averaging is done over the small-scale fading. Note that, because of the averaging, this ratio is the same at every location θ and is equal to ACLR in (3.52). A fading environment can be artificially created in a reverberation chamber, which would lend itself to practical measurements of this kind [24].

Multi-Antenna Setting

The most straightforward way to generalize the ACLR measure to a multi-antenna setting is to define a per-antenna ACLR as

$$\mathsf{ACLR}_m \triangleq \frac{\max\left\{\int_{-3B/2}^{-B/2} \mathbb{E}[\, S_{y_m y_m}(f)\,]\,\mathrm{d}f, \int_{B/2}^{3B/2} \mathbb{E}[\, S_{y_m y_m}(f)\,]\,\mathrm{d}f\right\}}{\int_{-B/2}^{B/2} \mathbb{E}[\, S_{y_m y_m}(f)\,]\,\mathrm{d}f}. \tag{3.54}$$

Since signals from a multi-antenna BS combine in the air however, there is a chance that the received power in an adjacent band is larger than it would be with a single-antenna BS using the same total transmitted power. Therefore it remains to determine what the per-antenna ACLR says about how much a victim, who operates in an adjacent band, really is disturbed.

Based on the observation in (3.53), we define a measure that generalizes the ACLR concept to multi-antenna transmission. We define the MIMO-ACLR as

$$\mathsf{MIMO\text{-}ACLR}(\theta) \triangleq \frac{\max\{\int_{-3B/2}^{-B/2} \mathbb{E}[\, S_{\theta}(f)\,]\,\mathrm{d}f, \int_{B/2}^{3B/2} \mathbb{E}[\, S_{\theta}(f)\,]\,\mathrm{d}f\}}{\int_{-B/2}^{B/2} \mathbb{E}[\, S_{\theta}(f)\,]\,\mathrm{d}f}. \tag{3.55}$$

In this definition, the expectation is taken with respect to the small-scale fading. The small-scale fading $\tilde{\mathbf{h}}_\theta(f)$ is assumed to be independent of that of the UEs $\tilde{\mathbf{h}}_{\theta_k}(f)$, for all k, so that $\tilde{\mathbf{h}}_\theta(f)$ and $\mathbf{S}_{\mathbf{yy}}(f)$ are independent.

We show that the measure MIMO-ACLR has the following properties, if $\mathbb{E}\big[\tilde{\mathbf{h}}_\theta(f)\tilde{\mathbf{h}}_\theta^{\mathsf{H}}(f)\big] = \mathbf{I}_M$, for all θ:

P1 It does not depend on the large-scale fading β_θ and is the same for all θ.

P2 It does not change if the transmitted signal is scaled.

P3 It is equal to the per-antenna ACLR_m and to the ACLR of a single-antenna system with the same radiated power.

The properties P1, P2 and P3 follow from (3.51), which gives

$$\mathsf{MIMO\text{-}ACLR} = \frac{\max\left\{\int_{-3B/2}^{-B/2} \mathbb{E}[\, S_{\mathrm{tx}}(f)\,]\,\mathrm{d}f, \int_{B/2}^{3B/2} \mathbb{E}[\, S_{\mathrm{tx}}(f)\,]\,\mathrm{d}f\right\}}{\int_{-B/2}^{B/2} \mathbb{E}[\, S_{\mathrm{tx}}(f)\,]\,\mathrm{d}f}, \tag{3.56}$$

where the argument θ has been dropped.

Further, we conjecture that the measure MIMO-ACLR has this property:

C1 It depends only weakly on the power allocations $\{\xi_k\}$ and the path losses $\{\beta_{\theta_k}\}$ of the UEs.

The conjectured property C1 remains a conjecture in this study. It is however made plausible by the fact that the optimal transmit direction of each UE k does not depend on its path loss β_{θ_k} in MaMi, see [52].

It is important to note that a MaMi system can radiate less power than a single-antenna system for a given performance requirement by virtue of the high array gain of the precoding. Because the radiated power is reduced, the absolute amount of disturbing power a victim that operates in an adjacent band suffers from is also reduced in the MaMi system, even if the ACLR in the single-antenna system and the MIMO-ACLR in the MaMi system are the same. Property P3 of the MIMO-ACLR measure thus suggests that the MIMO-ACLR for MaMi can be higher than ACLR can be for a single-antenna system without disturbing communication in adjacent bands more—the difference between MIMO-ACLR and ACLR roughly being equal to the array gain of the MaMi system.

Worst-Case Out-of-Band Radiation

If coding can be done over multiple channel coherence intervals, then only the average amount of received OOB radiation is relevant for a victim. However, there are cases, where coding cannot be done over multiple coherence intervals, e.g., because of latency constraints or because the fading is static as in a line-of-sight scenario. In these cases, one has to study whether there are points, to which the OOB radiation is beamformed, in order to protect victims in *every* coherence interval. To study whether there are such points, we study the maximum PSD, which is defined as

$$S_{\max}(f) \triangleq \lambda_{\max}\big(\mathbf{S}_{\mathbf{yy}}(f)\big). \tag{3.57}$$

This corresponds to the highest normalized power received at a given frequency at any point, i.e.

$$\beta_\theta \|\tilde{\mathbf{h}}_\theta(f)\|^2 S_{\max}(f) \geq S_\theta(f), \quad \forall\theta. \tag{3.58}$$

Note that $S_{\max}(f)$ bounds the *maximum* received power at frequency f for all channel vectors $\tilde{\mathbf{h}}_\theta(f)$. There is a possibility, however, that the maximizing channel vector has zero probability to show up in the physical environment. The measure might therefore be a rather loose upper bound, in the sense that the maximum power it indicates is rarely seen by a victim UE.

3.3.4 Simulation of Spatial OOB Distribution

In this section, the spatial distribution of the OOB radiation is studied for some representative scenarios. All continuous-time signals are simulated with $\kappa = 5$-times oversampling. A memoryless, third-order polynomial model is assumed, where $b_{mp}(\tau) = b_{mp}\delta(\tau)$, for $p = 1, 2, \forall m$, and $b_{mp}(\tau) = 0$, for $p > 2$. Then the cross-correlation in (3.42) simplifies into

$$\begin{aligned} R_{y_m y_{m'}}(\tau) = b^*_{m1} b_{m'1} R_{x_m x_{m'}}(\tau) + 2R_{x_m x_{m'}}(\tau) \\ \times \big(b^*_{m1} b_{m'2} \sigma^2_{x_m} + b^*_{m2} b_{m'1} \sigma^2_{x_{m'}} + b^*_{m2} b_{m'2} (2\sigma^2_{x_m} \sigma^2_{x_{m'}} + |R_{x_m x_{m'}}(\tau)|^2)\big). \end{aligned} \tag{3.59}$$

We set $b_{m1} = 1$ and $b_{m2} = -0.03491 + j0.005650$ for all m (obtained through linear regression on measurements on the class AB amplifier that can be run from [32]) and let the amplifier operate at its 1 dB-compression point. As pulse shaping filter, we chose a root-raised cosine with roll-off 0.22, as in LTE [1], which gives the normalized bandwidth $BT = 1.22$.

Two channel scenarios are considered: line-of-sight and independent Rayleigh fading. For simplicity, all UEs are assumed to be at the same distance from the BS and experience the same large-scale fading, i.e. $\beta_{\theta_k} = 1, \forall k$. Equal power allocation is applied, i.e. $\xi_k = 1/K, \forall k$.

In the line-of-sight scenario, there is only one path between each antenna and each UE: the direct non-obscured path. Furthermore, a uniform linear array is considered. Denote the angle to the k-th UE by θ_k. The channel to UE k is then:

$$\mathbf{h}_{\theta_k}(\tau) = e^{j\phi_k} \boldsymbol{\sigma}_k \delta(\tau), \tag{3.60}$$

where ϕ_k is the phase shift due to the propagation delay to the array, and $\boldsymbol{\sigma}_k$ is the steering vector to UE k. The phase shift is assumed to be uniformly distributed over $[0, 2\pi]$. The m-th element of the steering vector, in the case of a linear array with uniform spacing, is given by $[\boldsymbol{\sigma}_k]_m = e^{j2\pi m \Delta \sin(\theta_k)/\lambda}$, where Δ is the distance between the antennas and λ the wavelength of the signal carrier. We study the case of $\Delta = \lambda/2$, which is commonly regarded as the smallest interantenna distance that results in little coupling between antennas.

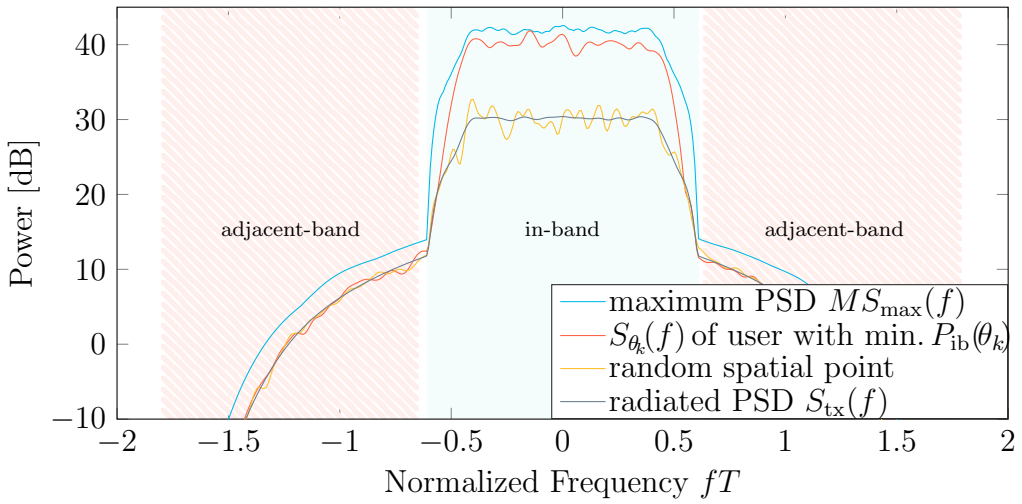


Figure 3.15: Power spectral densities for a system with 10 UEs and 100 antennas in a Rayleigh fading channel.

In the non-line-of-sight scenario, we model the channel as i.i.d. Rayleigh, i.e., each element in the oversampled channel impulse response is i.i.d.

$$\text{LP}\big\{\mathbf{H}(\tau)\big\}(\ell T/\kappa) \sim \mathcal{CN}(0, 1/L), \tag{3.61}$$

where $\text{LP}\{\cdot\}(t)$ is an ideal low-pass filter with cutoff frequency $\frac{\kappa}{2T}$ and where L is the number of non-zero channel taps. We study the case where $L = 15\kappa$, which corresponds to a maximum excess delay of 15 symbol durations.

We define the received in-band power, adjacent-band power and maximum adjacent-band power as

$$P_{\text{ib}}(\theta) \triangleq \int_{-B/2}^{B/2} S_\theta(f)\mathrm{d}f, \tag{3.62}$$

$$P_{\text{ob}}(\theta) \triangleq \max\Big\{\int_{-3B/2}^{-B/2} S_\theta(f)\mathrm{d}f, \int_{B/2}^{3B/2} S_\theta(f)\mathrm{d}f\Big\}, \tag{3.63}$$

$$P_{\text{ob,max}} \triangleq \max\Big\{\int_{-3B/2}^{-B/2} S_{\max}(f)\mathrm{d}f, \int_{B/2}^{3B/2} S_{\max}(f)\mathrm{d}f\Big\}. \tag{3.64}$$

The power spectral densities in Figure 3.15 are from a system with 100 antennas that serves 10 UEs using MR precoding over a realization of a frequency-selective Rayleigh fading channel. Because of channel hardening, generating another channel does not change the general appearance of the curves. By measuring the vertical distance between the transmitted PSD $S_{\text{tx}}(f)$ (black) to the PSD $S_{\theta_k}(f)$ received at the UE with the smallest $P_{\text{ib}}(\theta_k)$ (red), we see that the array gain of the in-band power of even the weakest UE is around 10 dB. Furthermore we see, when the maximum PSD $\mathbb{E}[\|\tilde{\mathbf{h}}(f)\|^2]S_{\max}(f) = MS_{\max}(f)$ (blue) is compared to the transmitted PSD $S_{\text{tx}}(f)$, that the worst-case OOB power has a much smaller array gain, around 2 dB. The received PSD $S_\theta(f)$ at many random points θ were generated, each with an independent Rayleigh fading channel vector. All had the same general appearance as the one that is plotted in yellow. The received power varies around the radiated power level and is well below the maximum PSD.

In Figure 3.16, the adjacent-band power $P_{\text{ob}}(\theta)$ of a line-of-sight system can be OOB seen for different directions around the array. From the peaks, it can be seen that the power OOB is

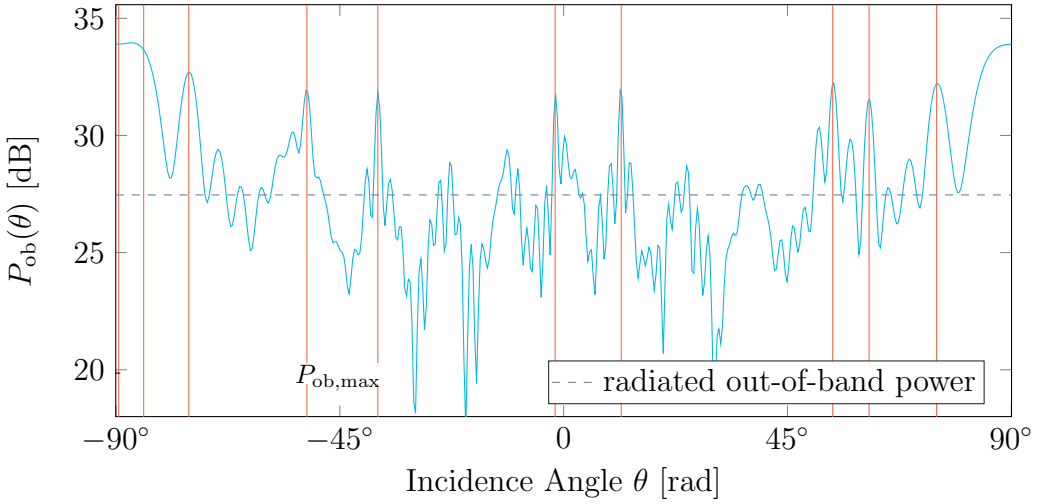


Figure 3.16: The adjacent-band power in different directions in a line-of-sight channel with 100 antennas and 10 UEs. The vertical lines indicate the directions of the UEs.

beamformed in the directions of the served UEs. The highest of these peaks is 4 dB above the transmitted adjacent-band power in this case. This is also how high the maximum adjacent-band power $P_{\text{ob,max}}$ (which upper bounds the adjacent-band power of any victim—not necessarily in line-of-sight) is above the transmitted adjacent-band power. The array gain of the worst-case adjacent-band power is thus slightly higher than in the Rayleigh fading case, but still significantly lower than the array gain seen in-band, which is 10 dB (cannot be seen in the plot). In between the served UEs, we see that the OOB power is approximately equal to or slightly lower than the radiated OOB power $S_{\text{tx}}(f)$.

These observations can also be made by studying the eigenvalue distribution of the correlation matrix $\mathbf{S}_{\mathbf{yy}}(f)$ at different frequencies, see Figure 3.17, where a 100-antenna system that serves both 10 UEs and 1 UE is studied for one realization of a Rayleigh fading channel. We see that, for 10 UEs and frequencies $f < B/2$, 10 out of 100 eigenvalues are 20 dB larger than the rest. These correspond to the directions of the UEs. At OOB frequencies $f \geq B/2$ however, there are no eigenvalues significantly above the average, which is marked by a dot. This means that, even in a worst-case scenario, a victim will not receive significantly more power OOB than on average.

In a single-user MaMi system, the OOB radiation is distributed differently, see the dashed lines in Figure 3.17. The signal OOB is more directive than in the multiuser case and has an array gain of approximately 10 dB in the strongest direction. This should be compared to the signal in-band, which has an array gain of 20 dB. We also see that 20 % of the eigenvalues are 2 dB above the average at $f = \frac{B}{2}$, which means that the probability of an OOB radiation level that is higher than the average is significant.

3.3.5 Simulated PSD and PA Efficiency

We have shown that UEs operating in adjacent channels essentially do not receive any interference enhancement due to the antenna gain. We now verify the validity of this result from a system-level perspective. More precisely, the impact of the system load and precoding design is explored. We further extend our discussion simulating different scenarios to determine how close to saturation and at which efficiency the PA can be operating. The impact of channel correlation over neighboring antennas is also investigated.

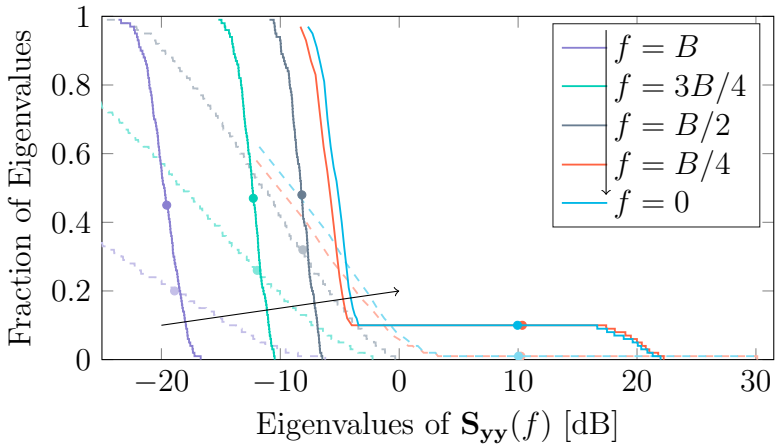


Figure 3.17: The complementary cumulative distribution of the eigenvalues of the correlation matrix $\mathbf{S}_{\mathbf{yy}}(f)$ at different frequencies f for a Rayleigh fading channel with 100 antennas and 10 UEs (solid lines), and 1 UE (dashed lines). The dot on each curve marks the average eigenvalue $S_{\text{tx}}(f)/M$.

Simulated in-band and out-of-band PSD

We assume a 20 MHz OFDM system with 2048 subcarriers of which 1200 are actively allocated, based on LTE. A raised-cosine pulse shaping filter with a roll-off factor of 0.22 is used. The RF power amplifier follows a third-order polynomial model. The PA operating point is generally characterized by the Power Input Backoff P_{IBO}, measuring the input signal margin with respect to $P_{\text{1-dB}}$, the 1-dB compression point where saturation effects become noticeable. In the following an input power backoff $P_{\text{IBO}} = -30\,\text{dB}$ is considered, effectively operating in complete saturation. We consider first a multi-tap independent Rayleigh channel model, i.e. $\mathbf{R} = \mathbf{I}$. The average transmitted power per antenna is normalized to 0 dB. Based on $M = 100$ antennas, the total output power is hence $\gamma = 20\,\text{dB}$.

Figure 3.18 shows the PSD for a system with $M = 100$ transmitting antennas and $K = 1$ UE, using ZF precoding. Observing the received PSD of the target UE, we see that the effect of the array gain applies only in-band, giving a 20 dB gain, while the received power in the adjacent band is equal to the transmitted power. In contrast, the received power of a UE positioned at a random position follows the same behavior as the transmitted power both within and out of the band. The simulation highlights that coherent combination occurs only in band and that a UE randomly located in space is not experiencing this combination gain. MaMi provides in this scenario an ACLR $= -39\,\text{dB}$, despite operating in complete PA saturation.

Similar observations can be extracted from Figure 3.19 where the system has been extended to $K = 10$ UEs. The in-band array gain is still 20 dB, but due to the presence of more UEs sharing the transmitted power, the relative difference between a desired UE and a random UE reduces to 10 dB. Here ACLR $= -29\,\text{dB}$ is obtained. The frequency-domain fluctuations are also smaller with $K = 10$ than with $K = 1$ due to the averaging effects.

Let us assess the impact of different precoding. Figure 3.20 shows the performance of a MaMi with $M = 100$ and $K = 25$ using both ZF and MRT. MaMi provides ACLR $= -26\,\text{dB}$ using MRT while ACLR $= -24\,\text{dB}$ is obtained when ZF is used. The difference between ZF and MRT is hence 2 dB on ACLR for this 100×25 configuration, while for smaller K the characteristics of ZF and MRT tend to be nearly the same from the PSD point of view. This is not the dominant argument for precoder selection, given that MRT simply does not work at a high load such as 100×25, but it is worth noting that the load of the system and the selected

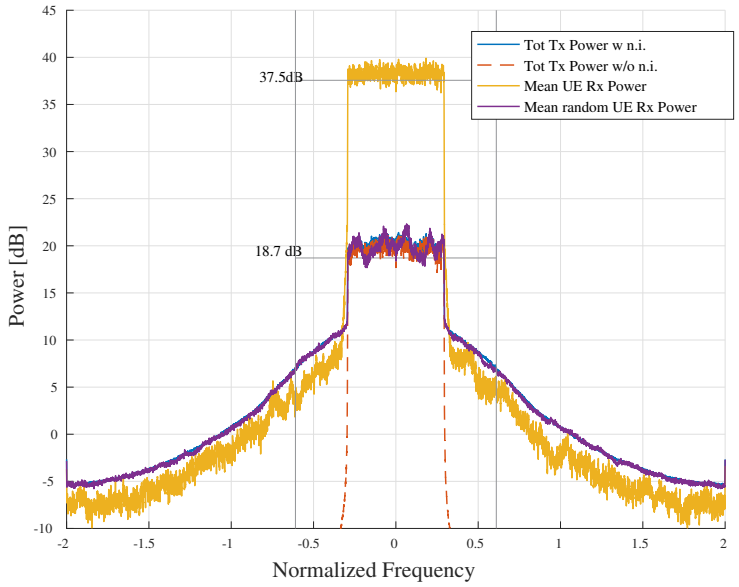


Figure 3.18: MaMi PSD with $M = 100$, $K = 1$, $P_{\text{IBO}} = -30\,\text{dB}$. The desired UE is compared to a random UE at a similar distance. The total BS output PSD is also provided with or without the non-ideal (n.i.) linearity behavior.

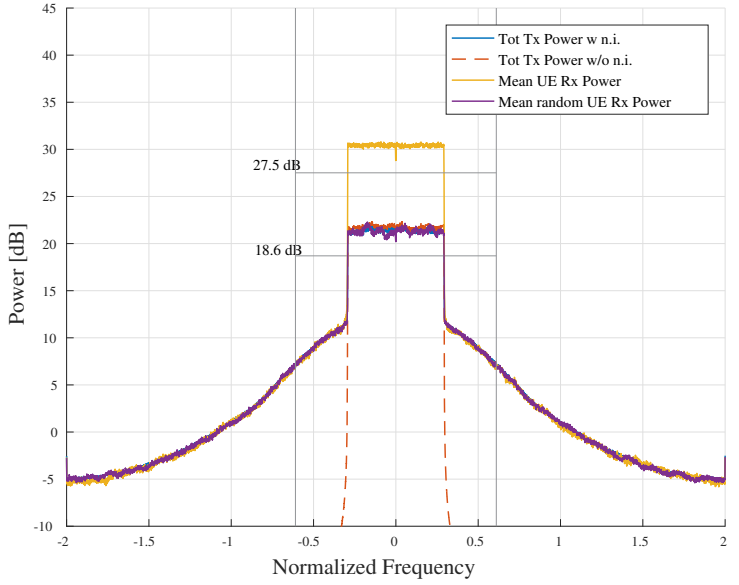


Figure 3.19: MaMi PSD with $M = 100$, $K = 10$, $P_{\text{IBO}} = -30\,\text{dB}$.

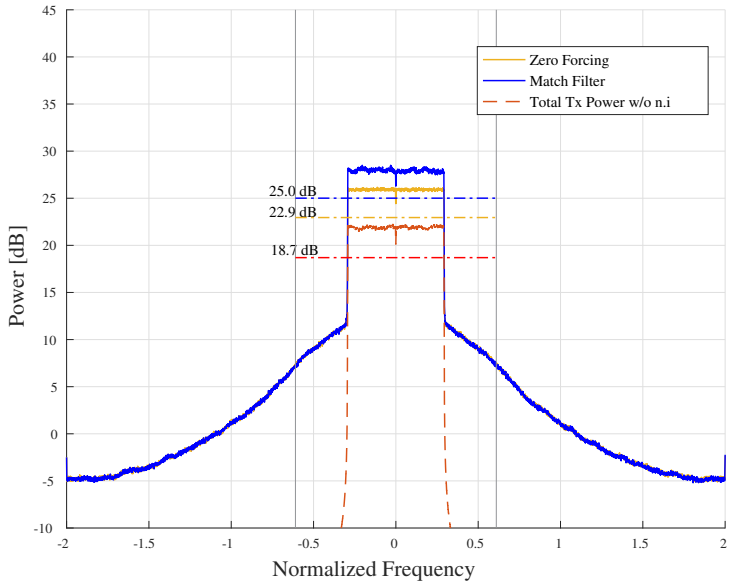


Figure 3.20: MaMi PSD with $M = 100$, $K = 25$, $P_{\text{IBO}} = -30\,\text{dB}$. Comparison between ZF and MRT precoding.

precoder should be jointly designed to meet the OOB requirements.

Power Amplifier Operating Point

In general, the transmitted signal power spectrum is regulated by a spectral mask. Some power input backoff is usually adopted in order to limit spectral regrowth within the mask limit, instead of operating in full saturation. Increasing the backoff reduces the nonlinear distortion, but also reduces PA efficiency. Minimizing power backoff is thus desirable, targeting the highest efficiency while still meeting non-linear distortion constraints. Let us assess the impact of P_{IBO} on the OOB and retrieve the minimum P_{IBO} which meets the LTE spectral mask requirements.

Figure 3.21 shows the PSD for a system with $M = 100$ and $K = 10$ when $\text{P}_{\text{IBO}} = 0\,\text{dB}$. As expected, increasing the power input backoff reduces the nonlinear distortion and ACLR $= -47.5\,\text{dB}$ is obtained, as compared to -29 dB when operating at saturation on Figure 3.19, even if $\text{P}_{\text{IBO}} = 0\,\text{dB}$ is still a very low back-off value in traditional systems.

We define the minimum P_{IBO} which can be selected to satisfy standard-compliant regulations considering as reference the 3GPP requirement [2] which sets the absolute OOB limit to $-13\,\text{dBm/MHz}$. For comparison we convert this absolute specification into an ACLR form. Hence, in the considered 20 MHz bandwidth the total allowed OOB emission is 0 dBm. In SISO-LTE the maximum output power is fixed to $P_{\text{SISO}} = 43\,\text{dBm}$ which leads to a maximum ACLR $= -43\,\text{dBc}$. In MaMi operation, the output power can be lowered, thanks to the array gain:

$$P_{\text{MaMi}} = \frac{K}{M} P_{\text{SISO}}. \tag{3.65}$$

In the considered 100×10 configuration, we can consider a total output power of $P_{\text{MaMi}} = 33\,\text{dBm}$ which in terms of ACLR requirements translates into $\text{ACLR}_{\text{max}} = -33\,\text{dBc}$. Figure 3.22 can be used in order to find the optimal operating point of the PA. We notice that the ACLR requirement is satisfied for $P_{\text{IBO}} = -10\,\text{dB}$. Hence, the PA can work in strong saturation while still satisfying the ACLR requirement. Traditionally the P_{IBO} is a positive value of

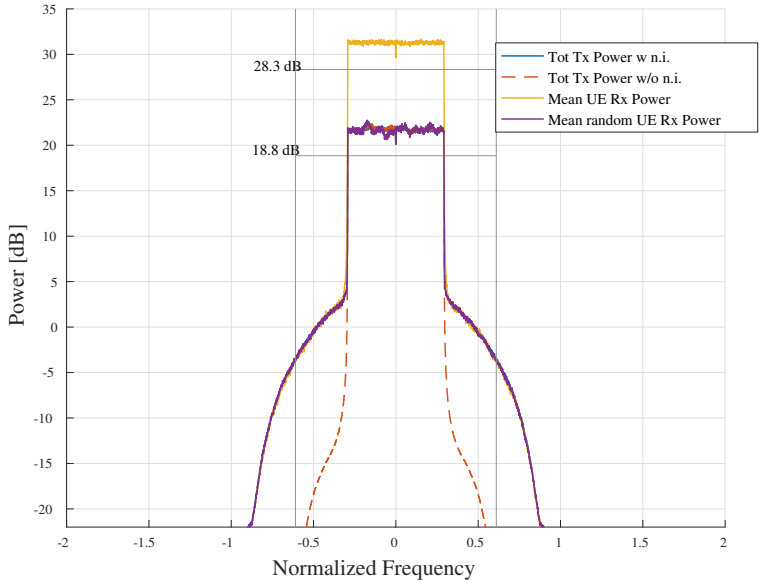


Figure 3.21: MaMi PSD with $M = 100$, $K = 10$, $P_{\text{IBO}} = 0\,\text{dB}$.

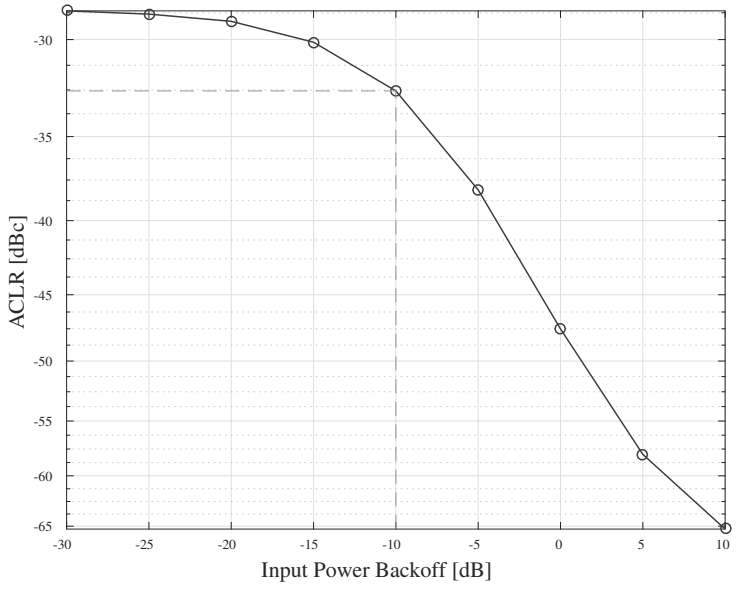


Figure 3.22: Variation of ACLR with P_{IBO} in a MaMi with $M = 100$, $K = 10$. 3GPP requirement is satisfied with $P_{\text{IBO}} = -10\,\text{dB}$.

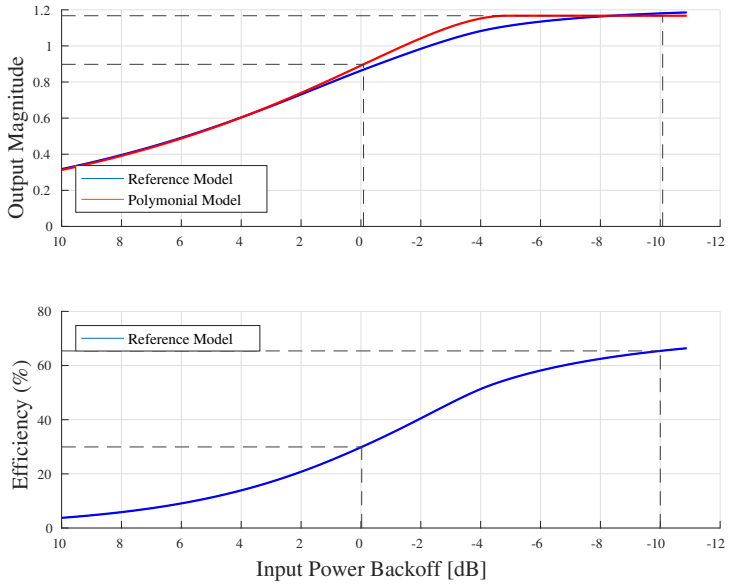


Figure 3.23: Top: characteristic of third-order polynomial model compared to model proposed in [17]. Bottom: efficiency of the reference model over P_{IBO}: $\eta = 65.4\%$ for $P_{\text{IBO}} = -10\,\text{dB}$.

several dB, however in MaMi even a completely saturated PA satisfies the 3GPP requirement, enabling the use of very high efficiency PAs. To give a numerical value of the PA efficiency, we refer to [17] which proposes a mathematical model that jointly models the linearity and efficiency of PAs. We use this model[2] to approximate the third-order polynomial model and to extract the PAs efficiency. Figure 3.23 shows the third-order polynomial model compared to the reference model. The third-order polynomial model is normalized to have $P_{\text{1-dB}} = 0\,\text{dB}$. Applying $P_{\text{IBO}} = -10\,\text{dB}$, Figure 3.23 shows that at the selected operating point the PA exhibits an efficiency of $\eta = 65.4\%$. MaMi systems require less stringent OOB specifications and PAs hence tolerate relaxed linearity requirements with respect to traditional communication systems. Due the high array gain, MaMi radiates less power while providing the same QoS as traditional systems, hence the amount of interference experienced by a UE operating in an adjacent channel is lower in MaMi.

Impact of the Channel Correlation

Let us now extend the analysis by considering a covariance matrix $\mathbf{R}$ which includes the spatial propagation environment and array geometry. We assume a Uniform Linear Array (ULA) with correlated antenna elements. We generate the coefficient of $\mathbf{R}$ following the exponential correlation model:

$$r_{ij} = \begin{cases} r^{j-i}, & i \leq j \\ r_{ji}^{*}, & i > j \end{cases}, \quad |r| \leq 1 \tag{3.66}$$

where the parameter r is the correlation coefficient. The effect of the antenna correlation on the system diversity is comparable to a reduction of the number of antenna elements, thus we expect a reduction of in-band gain and consequently a worst ACLR.

We consider $r = 0.7$, which can approximate the measured correlation in Figure 10 of [50]. In a scenario with $K = 10$ UEs, employing a fully saturated PA ($P_{\text{IBO}} = -30$ dB), Figure

[2]We select the quiescent point $Q = 0.61$ and normalized load resistance $R_L/R_{\text{ref}} = 0.9$, based on the empirical results reported in [17].

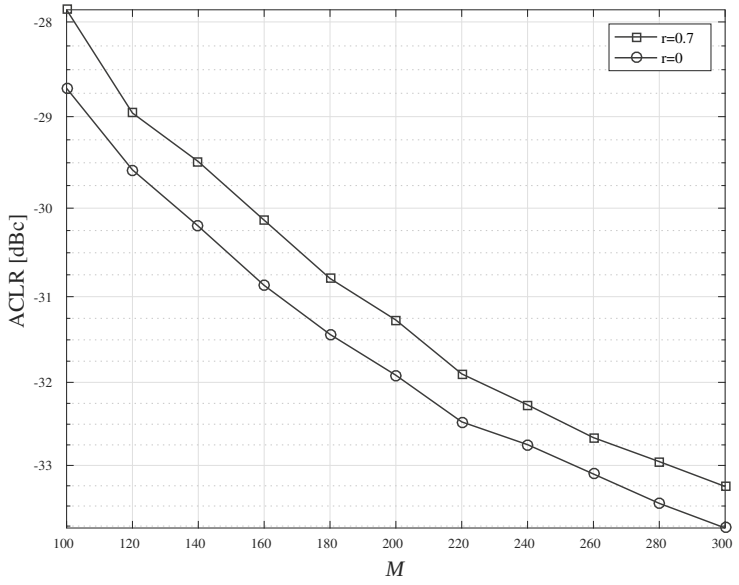


Figure 3.24: Effect of antenna correlation, varying the number of antenna, on the ACLR in a MaMi with $K = 10$, $\mathrm{P_{IBO}} = -30\,dB$.

3.24 shows the variation of the ACLR with the number of antenna elements in the ULA, for $r = 0$ (uncorrelated antenna elements) and $r = 0.7$. We observe that the choice of a different correlation model has a significant impact on the OOB. When using correlated antennas with $r = 0.7$, approximately 20 additional antennas are required to guarantee the same ACLR as the uncorrelated model. However, the reduction in ACLR when not adding antennas is around 1 dB only, thus comparing with the results of the previous section the PAs can still operate in saturation region. In summary, as expected the antenna correlation affects the ACLR metric but MaMi provides enough margin to let the PAs work in high efficiency region.

3.3.6 Conclusions

In this section, we have shown that MaMi systems can operate with lower linearity requirements on the power amplifiers compared to conventional single-antenna systems without increasing the disturbance of communication in adjacent bands. If the amount of radiated power and the linearity constraints are the same, a victim that operates in an adjacent band will receive the same amount of disturbing OOB radiation from a single-antenna system as from a MaMi system. Because of precoding and the large array gain it gives, a MaMi system can lower its radiated power and still serve its UEs with the same quality of service as the single-antenna system. The amount of disturbing OOB power is thus reduced in the MaMi system.

For specific realizations of the channel impulse response however, the small-scale fading of a victim might line up with the signal transmitted OOB and the victim then experiences much higher disturbing OOB radiation compared to the average. Such a worst-case event can be a problem if (i) the fading is time-invariant or (ii) if it occurs often, which can only happen if the small-scale fading of the victim is correlated to the channels of the served UEs. We have seen that the largest amount, by which the OOB radiation received by a victim at a frequency f can increase, is determined by the ratio $MS_{\max}(f)/S_{\mathrm{tx}}(f)$. In multiuser scenarios, this ratio is small; for example, 2–4 dB with 100 antennas and 10 UEs. In a single-user scenario however, this ratio can be much higher—in Rayleigh fading with 100 antennas, it is 10 dB. If coding can

be done over multiple coherence intervals, however, worst-case events are not a problem since data lost during one coherence interval can be recovered.

System simulations confirm the benefit of the non-coherent addition of components coming from the different antennas. We proved that fully saturated PAs in MaMi provide better ACLR compared to traditional communication systems. We find that also assuming channel correlation over neighboring antennas, fully saturated PAs satisfy the 3GPP spectral mask requirements.

Further, we have seen that OOB radiation can be measured over the air in terms of MIMO-ACLR and that MIMO-ACLR is the same as the per-antenna ACLR measured at the BS. To measure and constrain the radiated OOB power at the BS is thus sufficient to limit the average amount of power a victim in an adjacent band is disturbed by.

Usually the power from the different PAs dominates the BS power consumption, due to the large output power. The use of high-efficiency non-linear PAs enables significant power savings, thanks to gains in terms of energy efficiency. The operating region of the PAs is often determined by the OOB radiation. Since MaMi systems require less stringent OOB specifications, PAs can operate at relaxed linearity requirements with respect to traditional communication system. For instance, we find that PAs operating at $P_{\text{IBO}} = -10\,\text{dB}$, satisfy OOB regulation and provide PA efficiency of $\eta = 65\%$.

In conclusion, in MaMi systems, OOB radiation is not a limiting factor, hence enabling the use of highly efficient saturated PAs and reducing the BS power consumption.

Chapter 4

Hardware Implementation of Baseband Processing

This chapter provides an update on the hardware implementation of signal processing algorithms for MaMi. Section 4.1 describes hardware accelerators, implemented in CMOS, that efficiently handles the key baseband processing tasks. The energy consumption related to these processing tasks are evaluated in Section 4.2. Finally, Section 4.3 describes a proposed processing architecture, where different tasks are divided between per-antenna, per-subcarrier, and per-UE processing.

4.1 Hardware Accelerators

This section describes the hardware accelerators implemented within the MAMMOET project, which deal with key processing tasks such as OFDM modulation, downlink precoding, and uplink detection.

4.1.1 Low Latency and Area-efficient FFT/IFFT Processor

Wideband MaMi system typically use OFDM as the modulation scheme. With M antennas at the BS, such a system requires M OFDM modulator (IFFT) and demodulator (FFT) blocks for transmission and reception, respectively. An N-point FFT/IFFT has the complexity of $\mathcal{O}(N \cdot \log_2(N))$. Thus, the computational complexity of either uplink demodulation or downlink modulation of a MaMi system is $\mathcal{O}(M \cdot N \cdot \log_2(N))$. In Deliverable 3.2 [39], we showed that the FFT/IFFT processor has three times higher complexity than the MIMO precoding/decoding blocks. Moreover, the FFT/IFFT processors are key in the processing path, where critical latency constraints have been to guarantee a swift uplink-downlink turnaround time.

To tackle the aforementioned design challenges, we propose a low-latency and area-efficient FFT implementation. The main idea is to use the OFDM guard bands to reduce the operation counts and processing time, resulting in 42% latency reduction compared to single-input pipelined FFT processors reported in the literature. In order to realize this idea, a modified pipelined architecture combined with an efficient data scheduling scheme is proposed. Moreover, using proper resource sharing, the proposed scheme is capable of performing both OFDM modulation and demodulation for two BS antennas, reducing area without sacrificing throughput or latency.

Generally, an N-point FFT has a latency of N clock cycles, given one input per clock, as shown in Figure 4.1(a). To achieve a latency less than the IFFT size, a new approach is

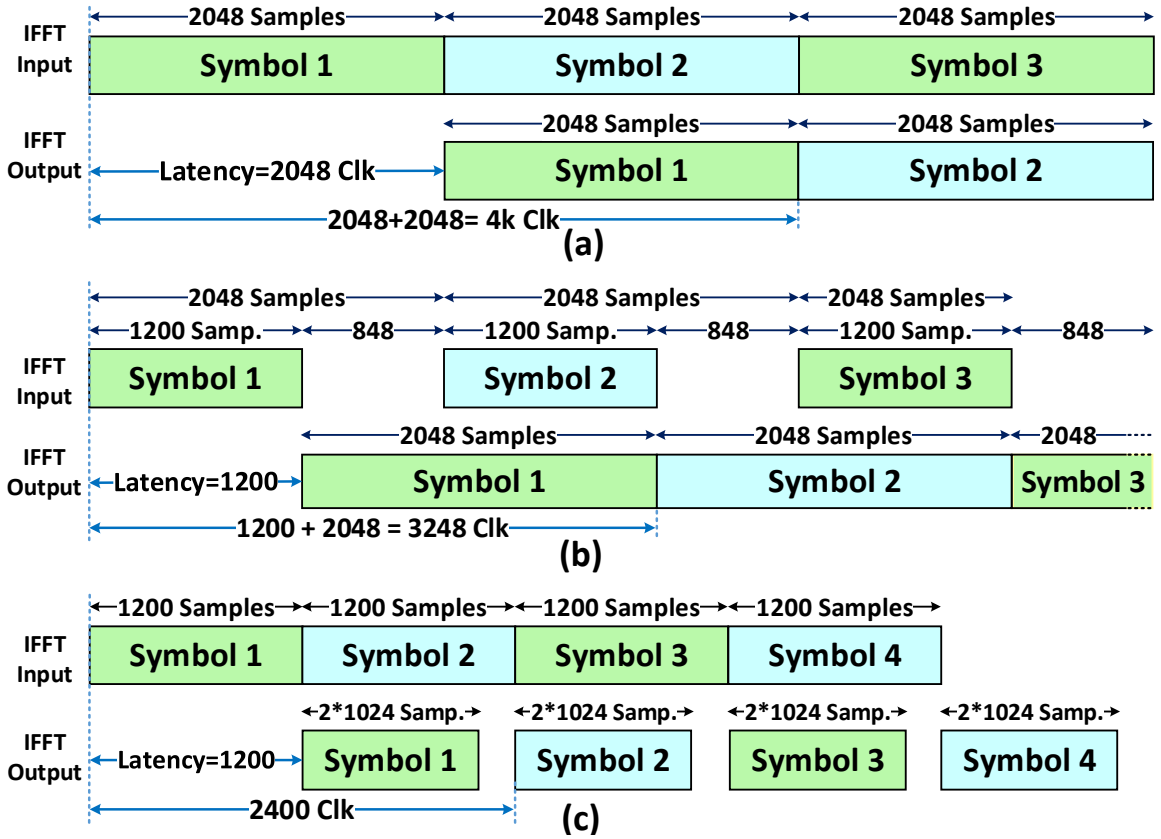


Figure 4.1: Data flow of a single-input pipelined IFFT in a one-antenna scenario for: (a) traditional scheme with continuous input, (b) traditional scheme with non-continuous input, (c) proposed low latency scheme with continuous input. The numbers in these figures are connected to N, P, and Z shown in Figure 4.2.

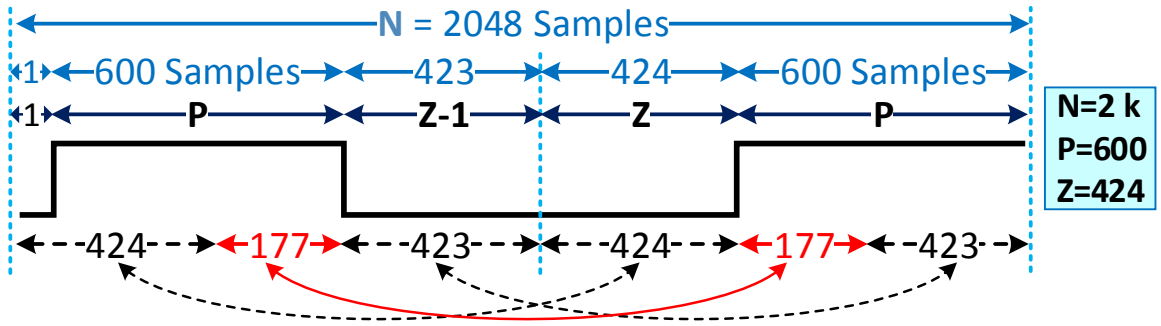


Figure 4.2: Data format of OFDM symbols with $N = 2048$ and 1200 used subcarriers. The proposed scheme can be used for other values of N, P, and Z.

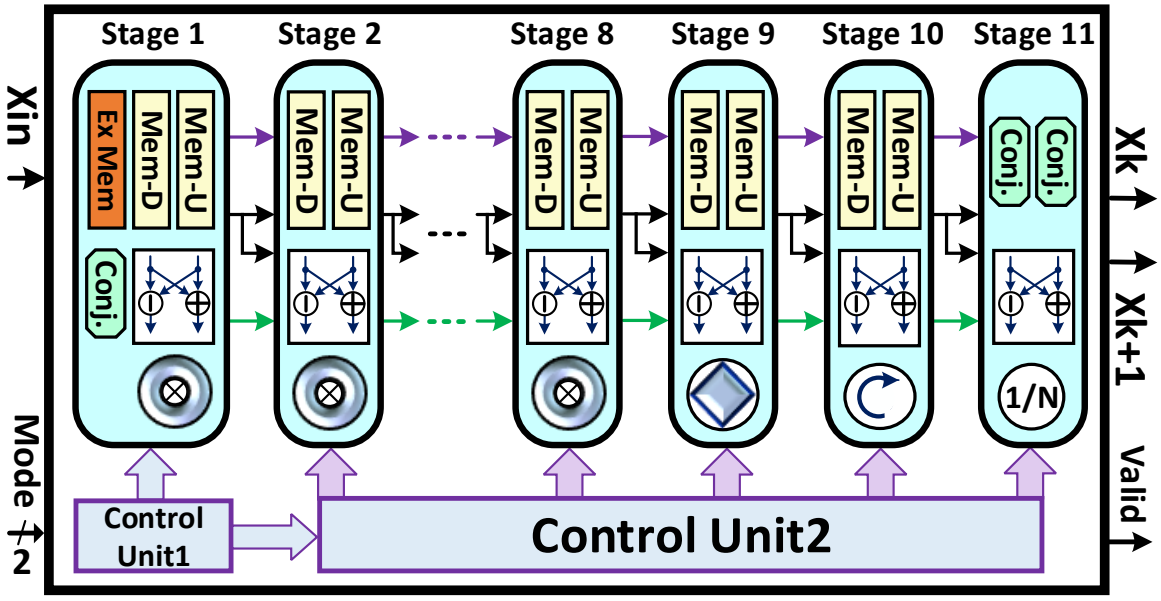


Figure 4.3: Proposed modified pipelined architecture for FFT/IFFT processor.

proposed. As mentioned above, the main concept is to use the OFDM guard bands to modify the IFFT calculation and reduce operation counts, decreasing latency significantly. Figure 4.2 shows the position of guard band zeros in the OFDM symbols. Due to symmetry between zeros and non-zero samples, a radix-2 FFT/IFFT algorithm is chosen, which includes 11 stages (i.e., $\log_2(N)$). In the first stage, the following butterfly operation is done for each pair of the input samples:

$$\begin{cases} y_l = x_l + x_{l+\frac{N}{2}} \\ y_{l+\frac{N}{2}} = x_l - x_{l+\frac{N}{2}} \end{cases} \qquad l = 0, 1, \ldots, \frac{N}{2} - 1. \tag{4.1}$$

There are two types of input pairs in Figure 4.2. In Type I, each pair includes one pre-known zero-sample (dotted arrows in Figure 4.2) and in Type II both samples are non-zero (solid arrow in Figure 4.2). As long as a non-zero sample of Type I enters the IFFT, the result of (4.1) is known without doing the butterfly. This means that, all Type I pairs can skip the butterfly in Stage 1 and go directly to Stage 2, without waiting for the remaining input samples. This significantly reduces the operation count, processing time, and latency. The corresponding data flow of the proposed scheme is shown in Figure 4.1(c), where the OFDM symbols can be entered and processed without any gaps and latency is decreased to 1200 clock cycles. Figure 4.1(c) also confirms that the proposed scheme can be used to perform both FFT and IFFT for two BS antennas, resulting in around 50% area reduction without sacrificing latency.

Table 4.1: FFT/IFFT Implementation Result with ST 28nm CMOS

Latency (Clocks)	**Latency (μs)**	**Gate Count**	**Clock Frequency**
1200	2.4	167 kG	500 MHz
Area	**Power**	**Throughput**	**Energy Efficiency**
0.08 mm^2	8.2 mW	1 GS/s	8.2 pJ/S

In order to realize the idea of latency/area reduction, a new modified pipelined architecture is proposed, as shown in Figure 4.3. Management of the memories and butterfly of Stage 1 is performed by *Control Unit 1*. *Control Unit 2* performs the same task for the remaining stages.

The proposed FFT processor has been synthesized using 28 nm CMOS technology and provides a throughput of 1 GS/s at 500 MHz. Table 4.1 shows the post-synthesis results.

4.1.2 QRD-based ZF Precoder with Approximative Givens Rotation

Several methods can be used to realize low-complexity ZF operation by leveraging the unique feature of MaMi channel matrix. In Deliverable 3.2 [39], we introduced a QR decomposition based matrix inversion, where 50% complexity reduction has been achieved by exploiting the fact that the Gramian matrix in MaMi system is diagonally dominant.

In this deliverable, our focus is on implementation architecture and the corresponding chip measurement results. Systolic arrays consists of homogeneous hardcoded network of nodes or PE, with each PE usually performing the same sequence of tasks. Due to the homogeneity, these architectures are easily scalable and have relatively lower design time. For the proposed downlink precoding, systolic arrays are used for all stages as shown in Figure 4.4 and described as follows:

- Let $\boldsymbol{H}$ denote the downlink channel matrix of size $M \times K$ in a MaMi system with M BS antennas and K single-antenna UEs. The first operation is to perform matrix multiplication to generate the Gram matrix of $\boldsymbol{H}$, i.e., $\boldsymbol{G} = \boldsymbol{H}\boldsymbol{H}^H$. In case of generating Hermitian or Gram matrices, the same systolic array can be used, but with only half the PE in a triangular form. This is achieved due to the symmetrical property of a Hermitian matrix. Both K^2 (2-D) and K (1-D) systolic arrays are used for QRD which have a time complexity of $\mathcal{O}(K)$ and $\mathcal{O}(K^2)$ respectively. Here, we employ a 2-D systolic array with $K^2/2 + K$ PE, hence having a time complexity of K.

- The second operation is to triangularize the Gram matrix, which is performed by a 2-D triangular systolic array. The approximative QRD can be leveraged to either lower the number of multipliers (gate count) or lower clock cycles (latency). In this step, the Gram matrix $\boldsymbol{G}$ is decomposed into a unitary matrix $\boldsymbol{Q}$ and upper triangular matrix $\boldsymbol{R}$.

- After the QRD, the user data vector $\mathbf{s}$ is multiplied with an orthonormal matrix $\mathbf{Q}$ as

$$\mathbf{u} = \mathbf{Q}^H\mathbf{s}. \tag{4.2}$$

  Generating the orthonormal matrix $\mathbf{Q}$ is an expensive procedure in hardware. This would require another systolic array with K^2 nodes. However, to reduce the gate count a 1-D systolic array is proposed which performs the rotation operations on the data vector (implicit). The coefficient and sequence of rotations need to match that of QRD.

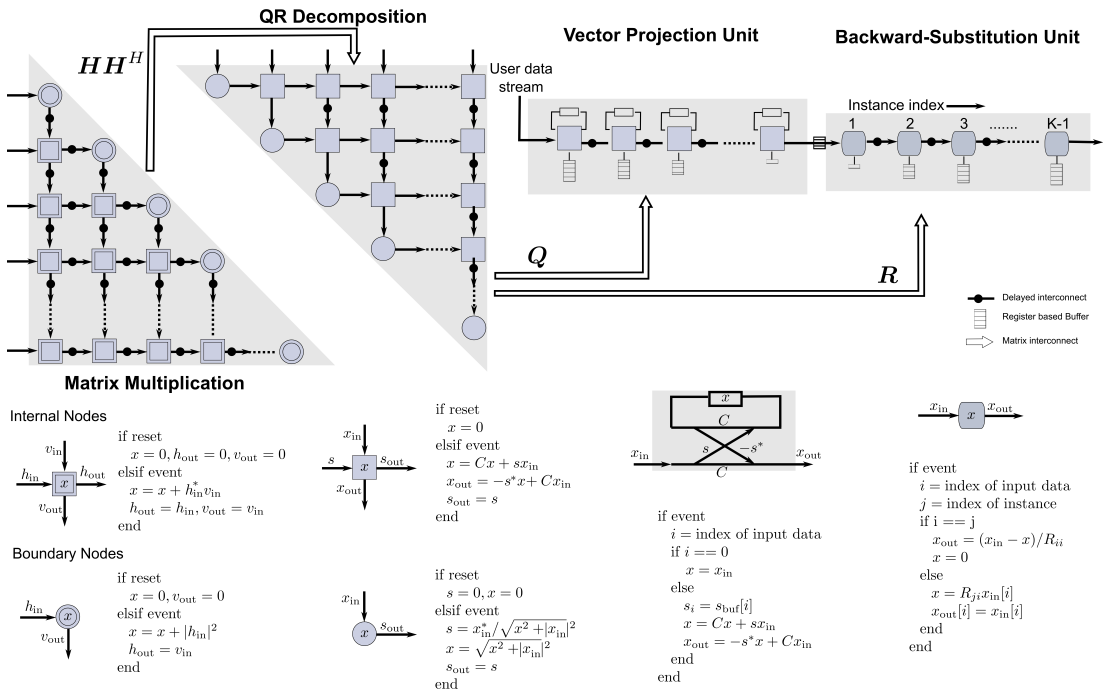


Figure 4.4: Top level description of the systolic downlink precoding system for MaMi. There are four modules and the corresponding PEs are described below the respective modules.

Table 4.2: ZF Precoder Implementation Result with ST 28nm CMOS

Matrix Dimension	**Gate Count**	**Clock Freq.**	**Power**
8×8	138 kG	300 MHz	31 mW
Precoding Rate	**QRD Latency (Cycles)**	**QRD Throughput**	**Energy Efficiency**
300 Mb/s	64	4.7 MQRD/s	103 pJ/b

- After the rotation, the vector $\mathbf{u}$ is multiplied with the triangular matrix as

$$\mathbf{v} = \mathbf{R}^{-1}\mathbf{u}. \tag{4.3}$$

  To avoid explicitly computing the inverse of triangular matrix and then performing the matrix vector multiplication, we employ a backward substitution based linear systolic array.

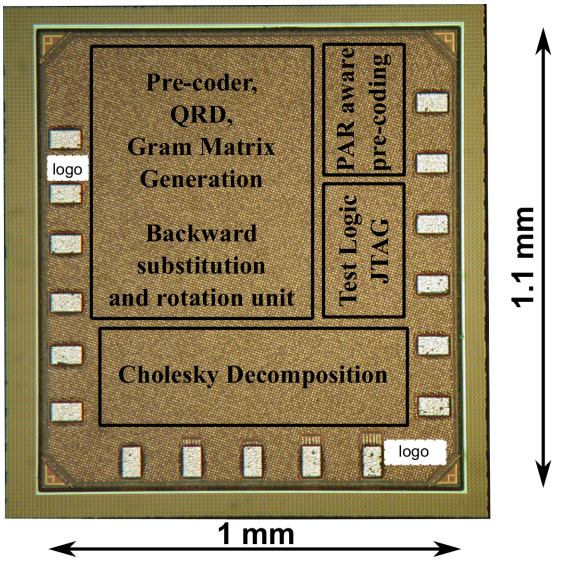
All these operations are envisioned to run in parallel on different subcarriers to fully utilize the hardware.

The proposed architecture has been fabricated using ST 28nm FD-SOI technology. Figure 4.5 shows a chip photo where the QRD-based precoder is located in the upper left corner. Table 4.2 summarizes the chip measurement results.

4.1.3 Uplink Detector using Cholesky Decomposition

Low complexity and near optimal performance makes linear detection an obvious design choice. Consider the received uplink signal $\mathbf{y} = \mathbf{Hs} + \mathbf{w}$. The MRC can be performed as

$$\hat{\mathbf{y}} = \mathbf{Zs} + \hat{\mathbf{w}}\,, \tag{4.4}$$



Technology	28 nm FD-SOI
Supply voltage	0.5 V to 1.0 V
Body-Bias range	-0.3 V to 1.0 V
Chip Gate count	350 K
Test Interface	JTAG
Tested Max. Freq	300 MHz
128x8 Massive MIMO Pre-coder	
Gate Count	138 K
QRD Clock Cycles	64
Algorithm	Approx. GR
Word length	12-bit
128x8 Massive MIMO Detector	
Gate Count	148 K
Cholesky clock cycles	325
Modulation	256-QAM
BSU and VPU cycles	8 cycles/sample

Figure 4.5: Chip photo.

where $\hat{\mathbf{y}} = \mathbf{H}^H\mathbf{y}$, $\mathbf{Z} = \mathbf{H}^H\mathbf{H}$ and $\hat{\mathbf{w}} = \mathbf{H}^H\mathbf{w}$. However, the claims for high performance with MRC assumes spatially uncorrelated channels and high BS antenna to UE ratio. This might not hold true in practical system, e.g., in the case of highly correlated LoS or large K as in stadium scenario. With extra computation, other linear detection, like ZF and MMSE, can improve the performance by performing inteference cancellation. However, to some scenarios, non-linear detection techniques like tree-based algorithms are essential to provide more robust detection performance.

The application of tree-based detection algorithms on (4.4) needs to handle the colored noise $\hat{\mathbf{w}}$. An exhaustive depth-search considering the noise variance will not impact the performance, however it is expensive in hardware. An approach to whiten the noise is to first perform Cholesky decomposition on the Gram matrix $\mathbf{Z} = \mathbf{L}\mathbf{L}^H$, where $\mathbf{L}$ is a lower triangular matrix. Afterwards both sides of (4.4) are multiplied with $\mathbf{L}^{-1}$ as

$$\bar{\mathbf{y}} = \mathbf{L}^H\mathbf{s} + \bar{\mathbf{w}}\,, \tag{4.5}$$

where $\bar{\mathbf{y}} = \mathbf{L}^{-1}\hat{\mathbf{y}}$ and $\bar{\mathbf{w}} = \mathbf{L}^{-1}\hat{\mathbf{w}}$. Computing $\mathbf{L}^{-1}$ explicitly is avoided by employing a forward-substitution module. Performing back-substitution on (4.5) is equivalent of zero-forcing (ZF) linear detection. Furthermore, in (4.5), noise $\bar{\mathbf{w}}$ is now whitened, i.e.,

$$\begin{aligned}
\mathbb{E}(\bar{\mathbf{w}}\bar{\mathbf{w}}^H) &= \mathbb{E}((\mathbf{L}^{-1}\hat{\mathbf{w}})(\mathbf{L}^{-1}\hat{\mathbf{w}})^H) \\
&= \mathbb{E}((\mathbf{L}^{-1}\mathbf{H}^H)\mathbf{w}\mathbf{w}^H(\mathbf{L}^{-1}\mathbf{H}^H)^H) \\
&= \mathbb{E}((\mathbf{L}^{-1})\mathbf{H}^H\mathbf{H}(\mathbf{L}^{-1})^H) \\
&= \mathbb{E}(\mathbf{L}^{-1}\mathbf{L}\mathbf{L}^H(\mathbf{L}^H)^{-1}) = \mathbf{I}_K\,.
\end{aligned}$$

Hence, using Cholesky decomposition for linear detection has an added advantage/ability of switching over to non-linear detection techniques for higher performance.

The proposed framework for adaptive detection is described in Figure 4.6, wherein switching between linear and non-linear detection is accomplished based on performance requirement. The following modes can be envisioned for the architecture

- **Mode 1:** The first detection option in the architecture is MR, which is multiplying the incoming signal with the hermitian of the channel estimate.

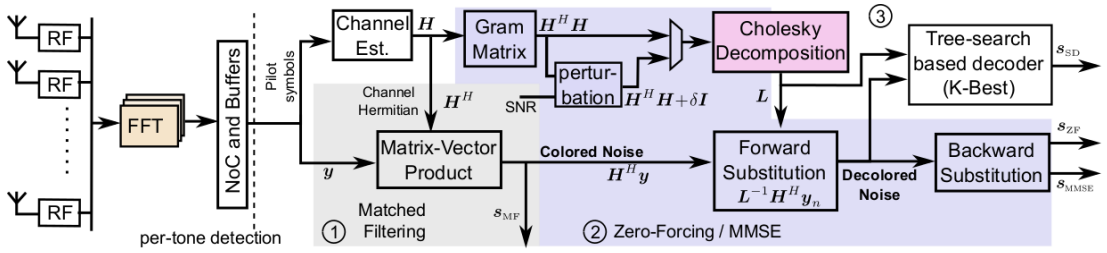


Figure 4.6: Top level architecture of the Cholesky decomposition based adaptive detection.

Alg. 1 Cholesky Decomposition of a $K \times K$ Hermitian positive definite $\mathbf{Z}$ to a lower triangular $\mathbf{L}$.

for $p = 0 \rightarrow K-1$ **do**
  for $q = 0 \rightarrow p$ **do**
    for $a = 0, r = 0 \rightarrow q-1$ **do** // Dot Product
      $a+ = \mathbf{L}[p][r] * (\mathbf{L}^H[q][r])$
    end for
    if $p == q$ **then** // Compute $\mathbf{L}$ values
      $\mathbf{L}[p][q] = \sqrt{\mathbf{Z}[p][q] - a}$
    else
      $\mathbf{L}[p][q] = (\mathbf{Z}[p][q] - a)/\mathbf{L}[q][q]$
    end if
  end for
end for

- **Mode 2:** For ZF or minimum mean square error (MMSE) the output of maximum ratio (MR) is used for further processing. This involves computing the Gram matrix followed by Cholesky Decomposition and then performing forward and backward substitution on (4.4).

- **Mode 3:** In this mode the output after forward substitution *i.e.,* (4.5) is used for non-linear detection schemes. Due to the decolored noise standard tree search implementations, like K-Best or Sphere decoder, can be employed.

The selection of these modes can also be a trade-off between complexity and performance. Also, the higher non-linear detection performance can be leveraged to perform antenna selection and turn-off antennas at base station (BS), with an increased detection processing cost.

The Cholesky decomposition algorithm in Alg. 1 is used for mapping into hardware. It consists of 3 for-loops, outer main loop has K iterations, the inner loops iterate over the index of the previous loop. The inner most loop performs an accumulation and has an $\mathcal{O}(0.5K^3)$. This accumulated value is used to compute elements of $\mathbf{L}$, and requires either a square root or division operation. Different implementations in hardware can be envisioned based on parallelization and pipelining by unrolling the for-loops.

Word-length optimization is a crucial aspect for an efficient hardware implementation. Another important hardware trade-off is between parallelization and cost. In general, reducing the computation time leads to a higher hardware cost. The actual design space has many more parameters, e.g., pipelining factor, targeted frequency, power, area, high speed/low power libraries etc. As a case study, an unrolling factor of 16 and a word-length of 12 bits are used for implementation, corresponding to a latency of 325 clock cycles and Signal-to-Quantization-Noise ratio (SQNR) of around 50 dB. The high accuracy is achieved by employing a bit accurate division and square root units. A standard sequential restoring arithmetic algorithm is used for both square root and division implementation [31]. This approach has a large critical path mainly due to the repetitive subtractions and comparisons. To overcome the speed compared to approximative fast techniques (Newton Raphson), a two stage pipelining is performed as shown in Figure 4.7. The architecture consists of a multiplexer network which feeds data from the

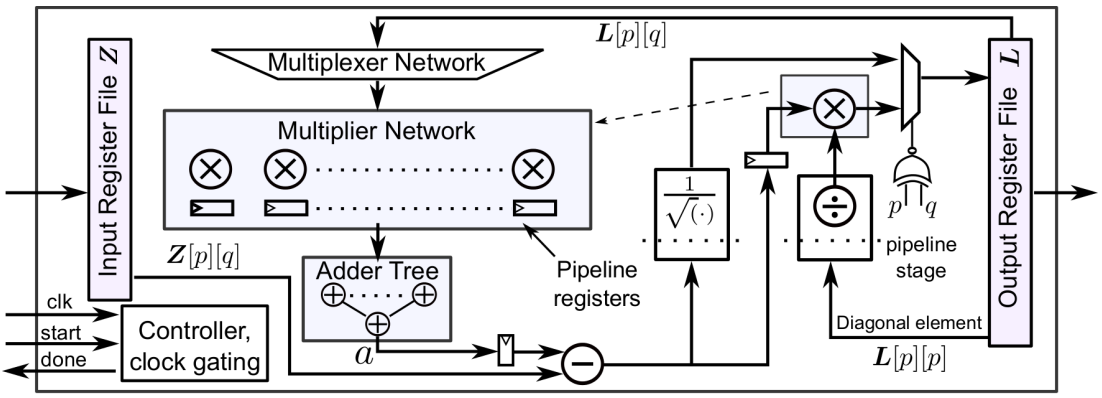


Figure 4.7: Top level architecture for Cholesky Decomposition.

Table 4.3: Uplink Detector Implementation Result with ST 28nm CMOS

MIMO Dimension	**Gate Count**	**Clock Freq.**	**Power**
128×8	148 kGE	300 MHz	18 mW
Detection Rate	**Modulation**	**Area Efficiency**	**Energy Efficiency**
300 Mb/s	256-QAM	2.02 Mb/s/kGE	60 pJ/b

register file to the multipliers. A generic adder tree network performs the vector dot-product to compute the scalar value. This scalar value (a) is used for further computations and updates the output register file based on the element pointers.

To reduce the power consumption, two techniques are employed, namely global clock gating and body-biasing. The implementation supports different clock gating modes, e.g., automatic clock gating based on module activity. In FD-SOI technology the planar back-side of a gate allows for a higher electrostatic control and body biasing efficiency. The implementation exploits body-biasing to either lower power consumption by performing reverse body-biasing (RBB) or improve performance by forward body-biasing (FBB). In the next section, measurement results of a 28 nm FD-SOI ASIC are presented, mainly focusing on energy and latency.

The Cholesky decomposition processor is also part of the fabricated chip shown in Figure 4.5 (lower part). The chip measurement results of the detector is listed in Table 4.3.

4.2 Energy Consumption Profiling

In Section 4.1, we discussed and presented the key signal processing blocks in a MaMi baseband system. They are all implemented using ST 28nm FD-SOI technology. The corresponding energy efficiency (obtained by chip measurement or post-synthesis simulation) for FFT/IFFT processor, precoder, and detector are 8.2 pJ/sample, 103 pJ/bit, and 60 pJ/bit, respectively.

If we consider a MaMi system with 128 BS antennas, 20 MHz bandwidth, 30.72 MS/s sampling rate (out of 2192 samples, 1200 are for data transmission, 144 for CP, and 848 for guard band), and 16-QAM modulation. This system serves 8 signal-antenna UEs simultaneously. The total data rate of such a system setup is 538 Mb/s.

The power consumption of the FFT processor can be calculated by

$$p_{FFT} = 8.2\,pJ/S \cdot 30.72\,MS/s \cdot \frac{N_{FFT}}{2192} \cdot M \cdot S, \tag{4.6}$$

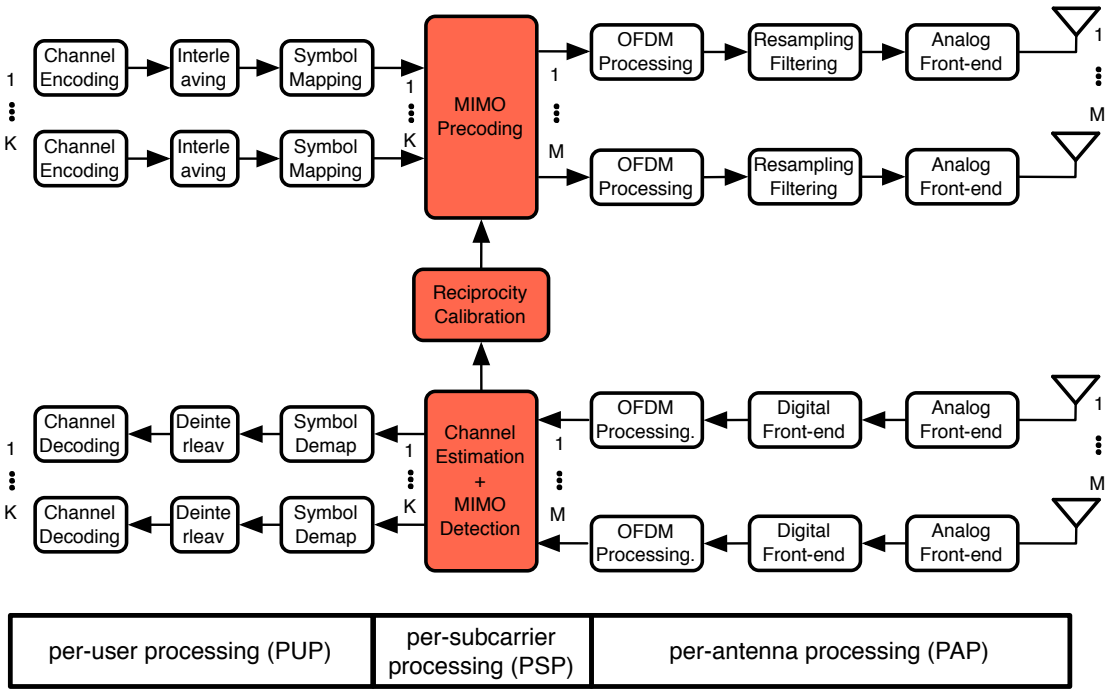


Figure 4.8: Digital Baseband Processing in an OFDM-based MaMi system for M BS antennas and K UE. The highlighted blocks are unique for MaMi and require special treatment due to scaling of complexity. The OFDM processing blocks include cyclic-prefix and guard-band removal on the uplink and cyclic prefix and guard-band addition on the downlink.

where M is the number of BS antennas, N_{FFT} is the FFT size, and S is the scaling factor between post-synthesis power and chip measurement result. Here, we take 1.5 for S based on our experience. In the system described above, the power consumption for FFT processor is 34.5 mW. The corresponding power consumption of precoder and detector are 55.4 mW and 32 mW, respectively.

4.3 Processing Architectures

The different parts of the OFDM based MaMi digital baseband processing system can be grouped into PAP (per-antenna processing), PSP (per-subcarrier processing), and PUP (per-UE processing), as shown in Figure 4.8. We can identify inherent parallelism and observe that processing complexity scales with the number of BS antennas M, the number of UEs K, or both.

- PAP: Scales with M as each antenna requires OFDM processing, digital/analog front-end.
- PSP: Scales with M and K as the channel matrix grows and is required per subcarrier.
- PUP: Scales with K, i.e., the number of UEs.

Note, that MRC/MRT algorithms and channel estimation may be performed on a per-antenna basis. However, to concentrate all reconfigurable processing inside specific blocks, we consider those to be part of the PSP throughout in the rest of this section. This simplifies overall partitioning and system design by providing a constant sharp edge between the different

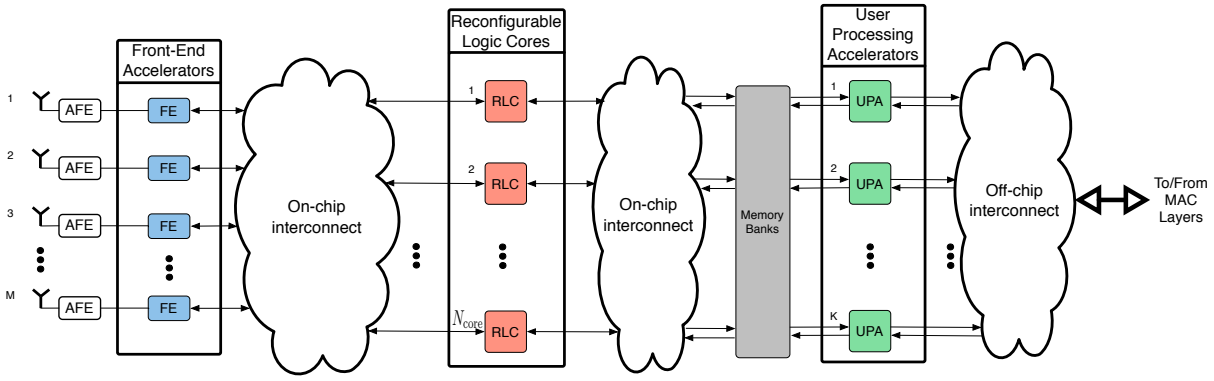


Figure 4.9: Possible High-Level System Architecture of a MaMi system with the different processing blocks; Front-End (FE) for per-antenna processing, Reconfigurable Logic Core (RLC) for per-subcarrier processing, user processing accelerator (UPA) for per-user processing

domains independent of the applied algorithms. As a ground step, the system can be partitioned according to the data flow in PAP, PSP and PUP.

Figure 4.9 shows this generalized partitioning, which takes into account, that the processing characteristics and implementation requirements are different, e.g., processing latency, available parallelism and reconfigurability. Thereby, a heterogeneous architecture is necessary to achieve implementation efficiency.

We also separate two different implementation approaches. First, accelerators which are parameterized hardware dedicated for a certain functionality. Second, reconfigurable hardware defined as hardware designs which allow high reconfigurability and are capable to map arbitrary functionalities within a certain application domain.

The digital part of PAP mainly consists of digital front-end and OFDM processing and is encapsulated in the front-end (FE). The functionality is relatively fixed but to some extent reconfigurable, e.g., variable-length FFT/IFFT processor to support different bandwidths. These reconfigurations are well implementable using accelerators. Given that PAP processing can be performed at each antenna node individually, extensive parallelism can be explored which is proportional to the number of BS antennas M.

PSP performs the channel estimation, detection and precoding including reciprocity compensation and is encapsulated in the Reconfigurable Logic Cores (RLCs). RLC have to be highly reconfigurable in order to adapt to changing operating conditions like current SNR regime, number of connected UEs and correlation among UE channels. As discussed in Sec. 4.1, one may use MR, ZF, MMSE, or even non-linear processing depending on current use cases. To lower processing throughput and latency requirements, we take advantage of the subcarrier independence in OFDM and distribute the overall subcarriers over N_{core} different RLCs. The multi-mode uplink detector we presented can be one example of such reconfigurable hardware.

The user processing accelerators (UPAs) perform PUP on the transmit/receive bits containing symbol demapping, deinterleaving and decoding on the uplink side; and encoding, interleaving and symbol mapping on the downlink side. Modulation order and code rate may change during run-time based on different SNR scenarios but overall functionalities remain quite constant. Furthermore, deinterleaving/interleaving, as well as coding/decoding, in general require large memories and therefore, accelerators are suited best.

Chapter 5

Summary

The signal processing is the tool required to achieve the advanced precoding and detection properties of MaMi, where tens of users can be served at the same time and frequency. It is easy to over-dimension that signal processing capability and requirements of MaMi systems, by presuming that the same processing tasks need to be carried out with the same resolution as in a legacy system—but with tens of more BS antennas and users. Fortunately, the complexity can be greatly reduced by tailoring the algorithms and implementation to the MaMi hardware and system characteristics. In this deliverable, we present and evaluate efficient algorithms for channel estimation, prediction and interpolation in the time-frequency domain. The robustness that different detection algorithms have towards low-resolution quantization and man-made interference has been described. Efficient power control schemes for the uplink and the downlink, which exploit channel hardening to lower complexity, have been further described and analyzed. The changes in characteristics for the OOB radiation when using many antennas have been analyzed. Finally, hardware implementation of key processing tasks have been described and analyzed.

This deliverable serves as a continuation and validation of results disseminated in MAMMOET D3.1 and D3.2. The key new findings and contributions are summarized as follows:

- Channel predication can be used to prolong the time interval of the downlink transmission, by adapting the precoding to predicted channel variations. The predictors are relatively robust to imperfect statistics.
- The frequency-interpolation scheme, developed in D3.2, leads to an acceptable increase in the complexity, making it feasible from a system perspective.
- The use of low-resolution ADCs at the BS is feasible, from both a rate and energy efficiency perspective. As a rule-of-thumb, 3-4 bit per real dimension is sufficient for efficient operation. Fewer bits are also possible, but with a noticeable performance loss.
- The M-MMSE detector, developed in D3.2, is robust to man-made interference and can, generally, reject any type of interference—including pilot contamination. Hence, interference is not a fundamental limiting factor but a design question, where the number of antennas and complexity of the processing scheme determine the interference level.
- Power control algorithms that utilize only large-scale fading characteristics provide an efficient mean to optimize the sum rate or max-min fairness of MaMi systems.
- The OOB radiation is basically the same in MaMi as with legacy systems, using the same total transmit power. This effectively means that a reduction in hardware resolution,

which would result in more OOB radiation, should be combined with a corresponding reduction in total transmit power.

- The basic signal processing tasks of OFDM modulation and ZF detection/precoding have been successfully implemented in CMOS, for a typical MaMi setup. The complexity and energy consumption is highly feasible for practical implementation.

We believe the MAMMOET project efforts collected in this deliverable, as well as D3.1 and D3.2, serve as a solid foundation for hardware-aware MaMi signal processing and validation of the practical feasibility of MaMi implementation. Many of the indications from previous information-theoretic studies, around the potential of reducing the MaMi implementation complexity and hardware resolution, have been thoroughly evaluated and proved. There are still open problems that remain and new advanced algorithms to implement in hardware, which we and others hopefully can solve after the MAMMOET project time span.

Bibliography

[1] 3GPP TS36.141 3rd generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (e-utra); base station (BS) conformance testing (release 10), 2011.

[2] Evolved Universal Terrestrial Radio Access. Base station BS radio transmission and reception, 3GPP TS 36.104. *V10*, 2011.

[3] MOSEK ApS. *MOSEK Optimization Suite Release 8.0.0.42*, 2016.

[4] F. Athley, G. Durisi, and U. Gustavsson. Analysis of massive MIMO with hardware impairments and different channel models. In *Proc. Eur. Conf. Antennas Propag.*, pages 1–5. IEEE, April 2015.

[5] M. Biguesh and A. B. Gershman. Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals. *IEEE Transactions on Signal Processing*, 54(3):884–893, Mar. 2006.

[6] E. Björnson, J. Hoydis, and L. Sanguinetti. Pilot contamination is not a fundamental asymptotic limitation in massive MIMO. In *Proc. IEEE ICC*, 2017. submitted.

[7] E. Björnson and E. Jorswieck. Optimal resource allocation in coordinated multi-cell systems. *Foundations and Trends in Communications and Information Theory*, 9(2-3):113–381, 2013.

[8] E. Björnson, E. G. Larsson, and T. L. Marzetta. Massive MIMO: ten myths and one critical question. *IEEE Communications Magazine*, 54(2):114–123, February 2016.

[9] E. Björnson, M. Bengtsson, and B. Ottersten. Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure. *IEEE Signal Process. Mag.*, 31(4):142–148, July 2014.

[10] 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, November 2014.

[11] H. V. Cheng, E. Björnson, and E. G. Larsson. Optimal pilot and payload power control in single-cell massive mimo systems. *IEEE Trans. Signal Process.*, 2017.

[12] A. Chiumento, S. Pollin, C. Desset, L. Van der Perre, and R. Lauwereins. Analysis of power efficiency of schedulers in LTE. In *2012 19th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, pages 1–4, Nov 2012.

[13] R. H. Clarke and W. L. Khoo. 3-D mobile radio channel statistics. *IEEE Transactions on Vehicular Technology*, 46(3):798–799, May 1997.

[14] C. Desset, S. Blandino, L. Van der Perre, E. Björnson, E. G. Larsson, B. Debaillie, A. Bourdoux, S. Pollin, W. Dehaene, O. Edfors, L. Liu, F. Tufvesson, D. Franz, J. Lorca, E. Karipidis, K. M. Koch, and T. Marzetta. Massive MIMO: the scalable 5G technology. In *European Conference on Networks and Communications 2016: Air Interfaces (PHY, MAC, RRM) (EuCNC2016-AirInt)*, 2016.

[15] C. Desset and B. Debaillie. Massive MIMO for energy-efficient communications. In *EuMC*, London, UK, October 2016.

[16] C. Desset, B. Debaillie, and F. Louagie. Modeling the hardware power consumption of large scale antenna systems. In *2014 IEEE Online Conference on Green Communications (OnlineGreenComm)*, 2014.

[17] H. Enzinger, K. Freiberger, and C. Vogel. A joint linearity-efficiency model of radio frequency power amplifiers. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 281–284, May 2016.

[18] ETSI. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 12.4.0 Release 12)*, 2015.

[19] L. Fan, S. Jin, C.-K. Wen, and H. Zhang. Uplink achievable rate for massive MIMO with low-resolution ADC. *IEEE Commun. Lett.*, 19(12):2186–2189, December 2015.

[20] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson. Massive MIMO in real propagation environments: Do all antennas contribute equally? *IEEE Trans. Commun.*, 63(11):3917–3928, 2015.

[21] K. G. Gard, H. M. Gutierrez, and M. B. Steer. Characterization of spectral regrowth in microwave amplifiers based on the nonlinear transformation of a complex Gaussian process. *IEEE Trans. Microw. Theory Tech.*, 47(7):1059–1069, July 1999.

[22] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer, 1992.

[23] M. Grant and S. Boyd. *CVX: Matlab software for disciplined convex programming, version 2.0 beta*, Sept 2013.

[24] C. L. Holloway, D. Hill, J. M. Ladbury, P. F. Wilson, G. Koepke, and J. Coder. On the use of reverberation chambers to simulate a Rician radio environment for the testing of wireless devices. *IEEE Trans. Antennas Propag.*, 54(11):3167–3177, November 2006.

[25] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer. Massive MIMO with low-resolution ADCs. *ArXiv E-Print*, February 2016. arXiv:1602.01139 [cs.IT].

[26] R. Jain, D.-M. Chiu, and W. R Hawe. *A quantitative measure of fairness and discrimination for resource allocation in shared computer system*, volume 38. Eastern Research Laboratory, Digital Equipment Corporation Hudson, MA, 1984.

[27] W. C. Jakes. *Microwave Mobile Communications*. IEEE Press, 1993.

[28] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath. Pilot contamination and precoding in multi-cell TDD systems. *IEEE Transactions on Wireless Communications*, 10(8):2640–2651, August 2011.

[29] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 1. Prentice Hall, 1993.

[30] J. Kim and K. Konstantinou. Digital predistortion of wideband signals based on power amplifier model with memory. 37(23):1417–1418, November 2001.

[31] I. Koren. *Computer Arithmetic Algorithms, second edition.* Ak Peters Series. Peters, 2002.

[32] P. Landin, S. Gustafsson, C. Fager, and T. Eriksson. Weblab: A web-based setup for PA digital predistortion and characterization. *IEEE Microw. Mag.*, 16(1):138–140, February 2015.

[33] 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.

[34] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian. Analog-to-digital converters. *IEEE Signal Processing Magazine*, 22(6):69–77, Nov. 2005.

[35] Y. Li, C. Tao, L. Liu, A. Mezghani, and A. L. Swindlehurst. How much training is needed in one-bit massive MIMO systems at low SNR? August 2016. arXiv:1608.05468 [cs.IT].

[36] Y. Li, C. Tao, L. Liu, G. Seco-Granados, and A. L. Swindlehurst. Channel estimation and uplink achievable rates in one-bit massive MIMO systems. In *Proc. Sensor Array and Multichannel Signal Process. Workshop*, July 2016.

[37] Z. Q. Luo and S. Zhang. Dynamic spectrum management: Complexity and duality. *IEEE Journal of Selected Topics in Signal Processing*, 2(1):57–73, Feb 2008.

[38] 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.

[39] MAMMOET. MAMMOET Deliverable D3.2, Distributed and centralized baseband processing algorithms, architectures, and platforms. https://mammoet-project.eu/downloads/publications/deliverables/MAMMOET-D3.2-baseband-processing-PU-M24.pdf, Jan. 2016.

[40] T. L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans. Wireless Commun.*, 9(11):3590–3600, 2010.

[41] J. Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, March 1960.

[42] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, Jr. Uplink performance of wideband massive MIMO with one-bit ADCs. February 2016. preliminary version available at: ArXiv E-Print, arXiv:1602.07364 [cs.IT].

[43] C. Mollén, U. Gustavsson, T. Eriksson, and E. G. Larsson. Out-of-band radiation measure for MIMO arrays with beamformed transmission. In *The Proceedings of the IEEE International Conference on Communications*, pages 1–6, May 2016.

[44] C. Mollen, E. G. Larsson, and T. Eriksson. On the impact of pa-induced in-band distortion in massive MIMO. In *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, 2014.

[45] C. Mollén, E. G. Larsson, and T. Eriksson. Waveforms for the massive MIMO downlink: Amplifier efficiency, distortion and performance. *IEEE Trans. Commun.*, April 2016.

[46] B. Murmann. ADC performance survey 1997-2016. Technical report.

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

[48] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta. Massive mu-MIMO downlink TDD systems with linear precoding and downlink pilots. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 293–298, Oct 2013.

[49] A. Papoulis and S. U. Pillai. *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education, 2002.

[50] S. Payami and F. Tufvesson. Channel measurements and analysis for very large array systems at 2.6 ghz. In *Antennas and Propagation (EUCAP), 2012 6th European Conference on*, pages 433–437, March 2012.

[51] I Reed. On a moment theorem for complex Gaussian processes. *IRE Transactions on Information Theory*, 3(8):194–195, April 1962.

[52] L. Sanguinetti, E. Björnson, M. Debbah, and A. L. Moustakas. Optimal linear precoding in multi-user MIMO systems: A large system analysis. In *The Proceedings of the IEEE Global Communications Conference*, pages 3922–3927, December 2014.

[53] M. Schetzen. *The Volterra and Wiener Theories of Nonlinear Systems*. Krieger Publishing Company, 2006.

[54] D. Schreurs, M. O'Droma, A. A. Goacher, and M. Gadringer. *RF power amplifier behavioral modeling*. Cambridge University Press, 2009.

[55] S. Shi, M. Schubert, and H. Boche. Rate optimization for multiuser mimo systems with linear processing. *IEEE Transactions on Signal Processing*, 56(8):4020–4030, Aug 2008.

[56] A. Sripad and D. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5):442–448, Oct. 1977.

[57] S. Stanczak, G. Wunder, and H. Boche. On pilot-based multipath channel estimation for uplink CDMA systems: an overloaded case. *IEEE Transactions on Signal Processing*, 54(2):512–519, Feb 2006.

[58] T. Sundstrom, B. Murmann, and C. Svensson. Power dissipation bounds for high-speed nyquist analog-to-digital converters. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 56(3):509–518, Mar. 2009.

[59] C. Svensson. Towards power centric analog design. *IEEE Circuits and Systems Magazine*, 15(3):44–51, 2015.

[60] H. Yang and T. L. Marzetta. A macro cellular wireless network with uniformly high user throughputs. In *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pages 1–5, Sept 2014.

[61] W. Yu and J. M. Cioffi. On constant power water-filling. In *Communications, 2001. ICC 2001. IEEE International Conference on*, volume 6, pages 1665–1669 vol.6, 2001.

[62] J. Zhang, L. Dai, S. Sun, and Z. Wang. On the spectral efficiency of massive MIMO systems with low-resolution ADCs. *IEEE Commun. Lett.*, 20(5):842–845, May 2016.

List of Abbreviations

ACF autocorrelation function

ACLR adjacent-channel leakage ratio

ADC analog-to-digital converter

AGC automatic gain control

AMC adaptive modulation and coding scheme

AR auto-regressive

ARMA auto-regressive moving average

ASIC application specific integarted circuit

BER bit error rate

BCQI best CQI

BS base station

CDF cumulative distribution function

CMOS complementary metal-oxide semiconductor

CSI channel state information

CQI channel quality index

CWER codeword error rate

DFT discrete fourier transform

DSP digital signal processor

FBB forward body-biasing

FD-SOI fully depleted silicon on insulator

FE front-end

FFT fast Fourier transform

GOPS giga complex arithmetic operations per second

GP geometric program

IBO input-back-off

IFFT inverse fast Fourier transform

LDPC low-density parity-check

LoS line-of-sight

LS least squares

LTE long term evolution

M-MMSE multi-cell MMSE

MaMi massive MIMO

MCS modulation and coding scheme

MIMO multiple-input multiple-output

MMSE minimum mean square error

MR maximum ratio

MRC maximum ratio combining

MRT maximum ratio transmission

MSE mean square error

OOB out-of-band

OFDM orthogonal frequency-division multiplexing

PAS power allocation scheme

PAP per-antenna processing

PE processing element

PQN pseudoquantization noise

PSD power spectral density

PSP per-subcarrier processing

PUP per-user processing

QAM quadrature-amplitude modulation

QoS quality of service

RBB reverse body-biasing

RF radio frequency

RLC reconfigurable logic core

SE spectral efficiency

S-MMSE single-cell MMSE

SINR signal-to-interference-plus-noise ratio

SINQR signal-to-interference-thermal-and-quantization-noise ratio

SISO single-input single-output

SNR signal-to-noise ratio

SQNR signal-to-quantization-noise ratio

SVD singular value decomposition

UE user equipment

ULA uniform linear array

UPA user processing accelerator

ZF zero-forcing