Mantissa-masking for energy-efficient floating-point LTE uplink MIMO baseband processing

D. Guenther[1], T. Henriksson[2], R. Leupers[1], G. Ascheid[1]
{guenther, leupers, ascheid}@ice.rwth-aachen.de, tomas.henriksson@huawei.com
[1]RWTH Aachen, Germany [2]Huawei Technologies Sweden AB

***Abstract*—The increasingly diverse wireless communication ecosystem has given rise to flexible, programmable platforms for wireless baseband processing. This industry case study presents advance development results of a fully programmable, flexible floating-point DSP architecture for uplink (UL) multiple-input, multiple-output (MIMO) baseband processing with runtime-adaptive precision. By tuning the floating-point precision to the application needs, energy consumption can be reduced by up to 23 % per task.**

I. Introduction

The growing amount of wireless communication standards and their inherent complexity has motivated the concept of software defined radio (SDR). SDR platforms do not need dedicated circuits for each communication standard. Instead, they contain flexible circuits that implement the communication standards in software. SDR platforms are commonly equipped with fixed-point processor cores. Fixed-point cores generally exhibit a lower energy consumption per instruction than similar floating-point architectures. Recent work [1] shows, though, that floating-point processing allows for programs with lower execution time due to a decreased need for numerical stabilization. Thus, floating-point systems can achieve superior energy efficiency on a per-task level. Precision requirements in MIMO baseband processing vary among different use cases (e.g., antenna setup) but also within sections of the baseband algorithm itself [1]. Runtime-adaptive precision by means of mantissa-masking (setting a number of least significant bits to zero) exploits this fact to reduce switching activity and thereby energy consumption. This case study applies the findings of [1] to the domain of LTE uplink MIMO detection in an industrial setting. We present a suitable advance development baseband processor with adaptive mantissa precision and explore how this flexibility can be used under various conditions to improve energy efficiency.

II. Fundamentals of UL MIMO Detection

For a communication system with N_t transmit and N_r receive antennas, the transmission of a vector $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ over a frequency-flat fading channel with additive white Gaussian noise is modeled by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{u} \tag{1}$$

with receive vector $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$, channel matrix $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$, and noise-plus-interference vector $\mathbf{u} \in \mathbb{C}^{N_r \times 1}$, where interference is caused by neighboring cells for example. Linear minimum mean square error (MMSE) MIMO detection derives an estimate $\hat{\mathbf{x}}$ for the originally transmitted vector x using the equalizer matrix

$$\mathbf{W}^H = \mathbf{A}^{-1}\mathbf{H}^H\mathbf{R}_{uu}^{-1} \tag{2}$$

$$\mathbf{A} = \mathbf{H}^H\mathbf{R}_{uu}^{-1}\mathbf{H} + \mathbf{V}_s^{-1} \tag{3}$$

with noise-plus-interference covariance matrix $\mathbf{R}_{uu} = \mathbb{E}\{\mathbf{u}\mathbf{u}^H\}$. For iterative detection where an estimate $\mathbf{s}$ of the most likely transmitted symbol vector is available, diagonal matrix $\mathbf{V}_s$ contains the element-wise variances of $\mathbf{s}$. Otherwise, it is set to identity. In the first iteration (*it-0*) without a priori knowledge, we use

$$\hat{\mathbf{x}} = \mathbf{W}^H\mathbf{y}, \tag{4}$$

while in subsequent iterations (*it-1..it-n*), we first cancel out the impact of the most likely transmitted vector, then apply MMSE filtering to the residual error $\mathbf{y} - \mathbf{H}\mathbf{s}$ to estimate the error of the estimate $\mathbf{s}$, and finally correct $\mathbf{s}$ by that error.

$$\hat{\mathbf{x}} = \mathbf{s} + \mathbf{W}^H(\mathbf{y} - \mathbf{H}\mathbf{s}) \tag{5}$$

The matrix inversion in (2) has a high numerical precision requirement due to the high dynamic range. The inversion can be stabilized by means of matrix factorization, resulting in a reduced precision requirement [2].

LDLh factorization rewrites a Hermitian matrix $\mathbf{A}$ as the product $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^H$ with lower triangular matrix $\mathbf{L}$ and diagonal matrix $\mathbf{D}$. Thus, $\mathbf{A}^{-1}$ can be expressed as

$$\mathbf{A}^{-1} = \mathbf{L}^{-H}\mathbf{D}^{-1}\mathbf{L}^{-1}, \tag{6}$$

which contains only triangular/diagonal matrix inversions.

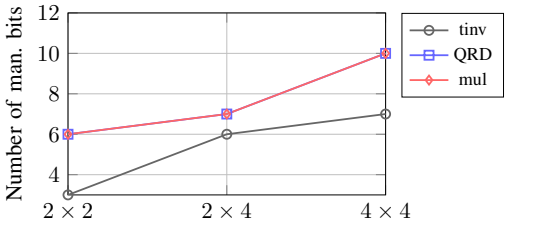
QR decomposition (QRD) of a matrix $\mathbf{A}$ into a unitary matrix $\mathbf{Q}$ and a triangular matrix $\mathbf{R}$ allows to simplify matrix inversions by applying a base transformation via $\mathbf{Q}$ first. Thereafter, only triangular matrix $\mathbf{R}$ has to be inverted. We use Givens rotation for QRD due to its low precision requirements and dynamic range [2].

Direct matrix inversion (DMI) calculates the inverse of matrix $\mathbf{A}$ directly without prior factorization. It also allows divide-and-conquer approaches to invert bigger matrices based on a subdivision of the original problem (e.g., breaking down a 4×4 matrix into 2×2 matrices) [3].

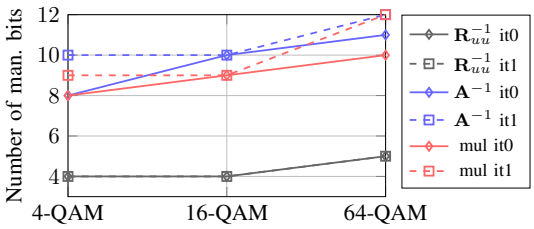
III. DSP with Mantissa-Masking

To facilitate the ease of programming and reduced execution time, our processor core supports a floating-point number format with 12 mantissa bits. Mantissa-masking is implemented by applying a configurable bitmask to the mantissa of each input of each arithmetic unit. Since MIMO baseband processing is largely based on complex-valued vector arithmetic, the core supports single-instruction multiple-data (SIMD) instructions on complex scalars. It is layed out as a very long instruction word (VLIW) processor where the SIMD instructions make out one instruction slot. The programming language is ANSI C with intrinsics to access specialized functionality. The core was synthesized for TSMC $28\,\mathrm{nm}$ CMOS technology, where the hardware overhead of mantissa-masking showed to be negligible compared to the overall core. The same results were obtained for the software overhead, since precision

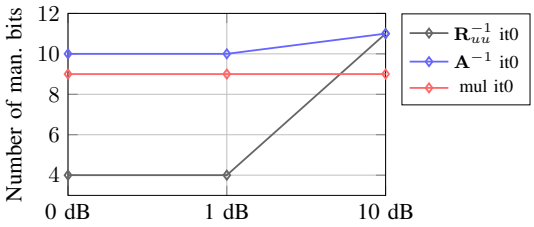
978-3-9815370-6-2/DATE16/



(a) Antenna setup (it-0, 16-QAM, QRD, IoT 0 dB)



(b) Constellation (2x4 MIMO, LDLh, IoT 0 dB)



(c) Interference (2x4 MIMO, it-0, 16-QAM, LDLh)

Fig. 1: Impact of use case parameters on numerical precision requirements

configuration instructions can be executed in parallel to vector arithmetic operations in free VLIW instruction slots.

IV. Precision Evaluation

Multiple characteristics of each use case influence the precision requirements for MIMO detection. We present the impact of relevant parameters in an otherwise constant scenario. Precision requirements are evaluated by Monte Carlo simulations, where mantissa precision is reduced as much as possible without impairing the achieved frame error rate. For LDLh, requirements are listed for the inversion of $\mathbf{R}_{uu}$, inner matrix $\mathbf{A}$ (see (3)), and the purely multiplicative part (mul). Precision results for DMI coincide with the LDLh variant. QRD is divided into the triangular matrix inversion of $\mathbf{R}$ (tinv), the actual QRD, and the multiplicative part. Figure 1 gives an overview of our findings.

Antenna Setup: The precision requirements to reliably separate the transmit signal at the receiver increases with the number of antennas as illustrated in Figure 1a for a 16-QAM transmission and detection using QRD.

Symbol Constellation: The reliable detection of a symbol from a denser constellation requires higher numerical precision due to the narrower margin of error, as shown in Figure 1b for LDLh inversion. Per use case, matrix inversion exhibits the highest precision requirements, followed by the multiplicative part and finally the inversion of $\mathbf{R}_{uu}$. The latter needs a comparably small precision, but note that numbers were obtained for 0 dB interference level.

Alg.	Constel.	P_{std} [mW]	P_{mem} [mW]	P_{sum} [mW]	E_{sum} [nJ]	Saving [%]
LDLh	Full	55.4	15.4	82.1	11.5	
	4-QAM	43.0		69.7	9.8	15.1
	16-QAM	46.6		73.3	10.3	10.7
	64-QAM	49.7		76.4	10.7	6.9
QRD	Full	56.2	13.5	81.0	21.6	
	4-QAM	37.4		62.1	16.5	23.3
	16-QAM	40.3		65.1	17.3	19.7
	64-QAM	40.3		65.1	17.3	19.7

TABLE I: Power/energy benchmark and mantissa-masking power saving (2x4 MIMO, it-0, IoT 0 dB)

Interference: Figure 1c shows the impact of the interference over thermal noise ratio

$$IoT = 10 \log \frac{I+T}{T} \tag{7}$$

on the precision requirements, with interference power I and thermal noise power T. Interestingly, this parameter mainly influences the precision for calculating $\mathbf{R}_{uu}^{-1}$. For low interference levels (1 dB), the impact of interference is non-measurable. For 10 dB the precision rises to the same level as for $\mathbf{A}^{-1}$. The underlying cause is that for high interference levels, $\mathbf{R}_{uu}$ turns from a scaled identity matrix to a seemingly random Hermitian matrix.

V. Energy Benchmarking

The previous precision evaluation is used to reduce the energy consumption of the baseband processor by means of mantissa-masking. Exemplary results are given in Table I, listing power consumption of memories P_{mem} and standard cells P_{std}, sum power P_{sum} and energy E_{sum}, and relative energy savings achieved by mantissa-masking. Savings increase for sparser constellations with lower precision requirements. The numerically stable QRD variant has the highest potential for energy reduction by mantissa-masking (23%). While having the overall lowest power consumption, LDLh is preferable from an energy perspective due to reduced execution time (as opposed to QRD). The power share of the memory subsystem, which is not impacted by mantissa-masking, ranges from 15.8 to 17.2%.

VI. Conclusion

In this paper, we showed that mantissa-masking is an efficient approach to reduce the energy consumption of embedded processing systems and presented UL MIMO detection for LTE as a case study. In this study, mantissa-masking had no notable negative impact on execution time or hardware complexity, while the energy savings are significant enough to make mantissa-masking an attractive feature for future production-line DSPs.

References

[1] D. Guenther, R. Leupers, and G. Ascheid, "Efficiency enablers of lightweight SDR for MIMO baseband processing," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 2015.
[2] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.
[3] F. Zhang, Ed., *The Schur Complement and Its Applications*, 1st ed. Springer US, 2005.