A 772 Mbit/s 8.81 bit/nJ 90 nm CMOS Soft-Input Soft-Output Sphere Decoder

Filippo Borlenghi*, Ernst Martin Witte*, Gerd Ascheid*, Heinrich Meyr*†, Andreas Burg‡
*Institute for Communication Technologies and Embedded Systems, RWTH Aachen University, 52056 Aachen, Germany
email: {borlenghi,witte,ascheid,meyr}@ice.rwth-aachen.de
†Visiting Professor at the Integrated Systems Laboratory 1, EPFL, 1015 Lausanne, Switzerland
‡Telecommunications Circuits Laboratory, EPFL, 1015 Lausanne, Switzerland
email: andreas.burg@epfl.ch

***Abstract*—Multiple-input multiple-output (MIMO) wireless transmission can approach its full potential in terms of spectral efficiency only with iterative decoding, i.e., by exchanging soft information between the MIMO detector and the channel decoder. Solving the soft-input soft-output (SISO) MIMO detection problem entails a very high complexity, which can typically be reduced only at the cost of a communication-performance penalty. The single tree-search (STS) sphere-decoding (SD) algorithm covers a wide range of this complexity-performance tradeoff. In this paper, we describe the silicon implementation of SISO STS SD. The 90 nm CMOS ASIC operates at a lower signal-to-noise ratio than other MIMO detectors. The maximum throughput is 772 Mbit/s at an energy efficiency of 8.81 bit/nJ.**

I. Introduction

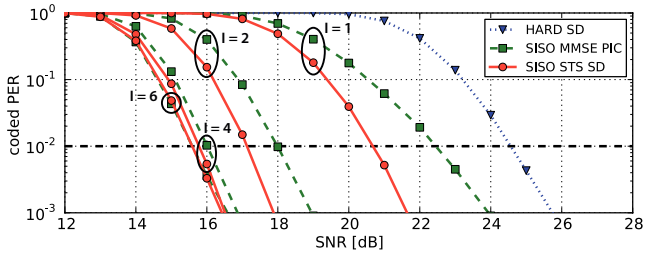
Multiple-input multiple-output (MIMO) transmission can significantly increase the data rate in wireless communication systems by spatial multiplexing, without additional usage of limited resources such as bandwidth and transmit power. Unfortunately, in terms of digital baseband processing in the receiver, MIMO also considerably increases the complexity of the detector. Therefore, most circuit implementations accept a sub-optimal communication performance to reduce complexity. Linear detectors, based on zero forcing or minimum mean square error (MMSE) criteria, and successive interference cancellation exhibit low complexity but also poor error-rate performance. Maximum-likelihood performance is approached by hard-output sphere decoders. A further performance gain over hard-output methods is achieved, with additional complexity, by providing soft information, as log-likelihood ratios (LLRs), to the channel decoder.

Iterative MIMO detection and decoding is the final hurdle towards approaching channel capacity [1], [2]. Introducing a feedback loop enables a soft-input soft-output (SISO) detector to improve its estimates based on extrinsic information computed by the channel decoder. Unfortunately, the resulting performance gain comes at the expense of a much higher detection complexity compared with non-iterative schemes. Only recently, the first silicon implementation of a SISO MIMO detector has been presented in [3], based on SISO MMSE parallel interference cancellation (PIC). This algorithm shows considerable communication performance gains over non-iterative detectors, but, like other (quasi-)linear methods, it fails to exploit the spatial diversity provided by MIMO. This limitation is overcome by SISO single tree-search (STS) sphere decoding (SD) [4], which has max-log maximum *a posteriori* (MAP) performance and the ability to fully exploit spatial diversity. Fig. 1 compares the communication performance, in terms of coded packet error rate (PER), of the non-iterative (iteration number $I = 1$) hard-output SD and the iterative ($I \geq 1$) SISO STS SD and SISO MMSE PIC algorithms for two communication scenarios. For a given number of iterations, STS SD always outperforms the MMSE PIC method. In Fig. 1(a) (fast Rayleigh fading channel), the communication-performance gap between the two algorithms ultimately diminishes for $I = 4$ since the strong code takes advantage of the rapidly changing channel conditions. Unfortunately, this type of diversity is typically not available or cannot be exploited by a weaker code. In this case, shown in Fig. 1(b), with $I = 6$, STS SD still reaches the target 1 % PER at a 3 dB lower signal-to-noise ratio (SNR) than MMSE PIC, showing a significantly better robustness to the operating scenario. Moreover, for a given SNR, STS SD typically achieves the target PER with fewer iterations: for instance, with $I = 2$ STS SD already outperforms MMSE PIC at $I = 6$. In addition to adjusting the number of iterations, the complexity of STS SD can be tuned at run-time and traded off with communication performance, hence scaling the detection effort to the target PER and to the SNR operating point.

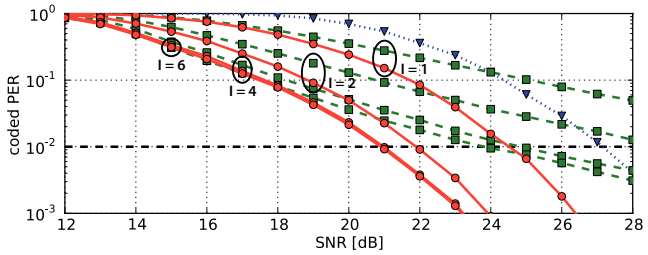
Contributions: We present—to the best of our knowledge—the first silicon implementation of SISO STS SD. Improving the architecture presented in [5], this 90 nm CMOS ASIC demonstrates the scalability of STS SD, achieving at high SNR a maximum throughput of 772 Mbit/s, twice as high as [5] and compatible with recent standards such as IEEE 802.11n, and an energy efficiency of 8.81 bit/nJ. At low SNR this ASIC provides, at a reduced throughput, a communication performance gain and a better robustness to channel conditions than other state-of-the-art detectors.

II. MIMO Detection by SISO STS SD

A spatial-multiplexing MIMO system with M_T transmit and $M_\mathrm{R} \geq M_\mathrm{T}$ receive antennas is assumed [1]. The transmitter sends a symbol vector $\mathbf{s} = [s_1, ..., s_{M_\mathrm{T}}]^T \in \mathcal{O}^{M_\mathrm{T}}$, where each s_i ($i = 1 \,..\, M_\mathrm{T}$) is obtained by mapping Q bits $x_{i,b} \in \{+1, -1\}$ ($b = 1 \,..\, Q$) to an element of the complex-valued constellation $\mathcal{O}$. The received signal is given by the complex symbol vector $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$, where $\mathbf{H} \in \mathbb{C}^{M_\mathrm{R} \times M_\mathrm{T}}$



(a) Fast Rayleigh fading channel (4×4, 64 QAM).



(b) Quasi-static Rayleigh fading channel (4×4, 64 QAM).

Fig. 1. Communication performance of MIMO detection algorithms.

is the channel matrix and $\mathbf{n} \in \mathbb{C}^{M_\mathrm{R}}$ is a white circularly-symmetric Gaussian noise vector with element-wise variance N_0. Tree-search detection is enabled by QR-decomposing (QRD) $\mathbf{H}$ as $\mathbf{H} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{C}^{M_\mathrm{R} \times M_\mathrm{T}}$ with $\mathbf{Q}^H\mathbf{Q} = \mathbf{I}$ and $\mathbf{R} \in \mathbb{C}^{M_\mathrm{T} \times M_\mathrm{T}}$ being upper triangular[1]. Hence, $\mathbf{y}$ is pre-processed as $\tilde{\mathbf{y}} = \mathbf{Q}^H\mathbf{y} = \mathbf{Rs} + \mathbf{Q}^H\mathbf{n}$. SISO MIMO detection can then be performed as an STS within a tree of order 2^Q and height M_T. Each node on tree level i is a partial candidate symbol vector $\mathbf{s}^{(i)} = [s_i, ..., s_{M_\mathrm{T}}]^T$, with a metric $\mathcal{M}_\mathrm{P}(\mathbf{s}^{(i)}) = \sum_{j=i}^{M_\mathrm{T}} \mathcal{M}_\mathrm{P}(s_i)$. Channel- ($\mathcal{M}_\mathrm{C}$) and *a priori*-based ($\mathcal{M}_\mathrm{A}$) contributions, always non-negative, determine the metric increment on level i by $\mathcal{M}_\mathrm{P}(s_i) = \mathcal{M}_\mathrm{C}(s_i) + \mathcal{M}_\mathrm{A}(s_i)$. The STS approach computes in a single tree traversal the MAP solution, with the overall minimum $\mathcal{M}_\mathrm{P}(\mathbf{s}^{(M_\mathrm{T})})$, and the $M_\mathrm{T}Q$ minimum counter-hypothesis metrics, from which the output extrinsic LLRs $\{L^\mathrm{E}_{i,b}\}$ can be easily derived [4].

The search complexity is efficiently reduced by branch and bound strategies relying on *enumeration*, i.e., sorting child nodes based on their metrics. This allows to prune, based on a *pruning criterion*, large parts of the tree that cannot add new information. Each node checked against a pruning criterion is an *examined node*, resulting in the complexity metric *number of examined nodes per detected symbol vector* N_en. Additional techniques can be applied to reduce N_en. In particular, *extrinsic-LLR clipping* [4] can restrict the pruning criterion, with a relevant decrease of N_en at the cost of a performance penalty. Hence, LLR clipping is key for the scalability of STS SD, enabling to trade off complexity and communication performance over a wide range and thus adapt the detection effort to the target PER and operating scenario.

[1] An i.i.d. Rayleigh fading channel, perfect channel knowledge and sorted QRD [6] are assumed. In the fast scenario $\mathbf{H}$ changes independently for each $\mathbf{y}$, in the quasi-static scenario $\mathbf{H}$ is constant for the duration of one packet (assumed equal to a code word). The bit-interleaved coded modulation employs a convolutional channel code (rate $1/2$, generator polynomials $[133_o, 171_o]$, constraint length 7) decoded by a max-log BCJR channel decoder with perfect termination knowledge and a random interleaver with 576 information bits. The SNR is $M_\mathrm{T}E_\mathrm{s}/N_0$, with $E_\mathrm{s} = \mathbb{E}[|s|^2], s \in \mathcal{O}$.

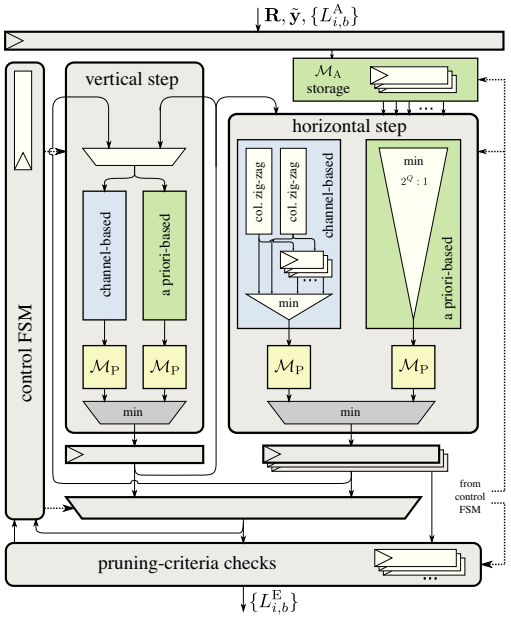


Fig. 2. High-level block diagram of the SISO STS SD architecture.

III. VLSI Architecture

The architecture presented here is designed according to the STS algorithm and the one-(examined-)node-per-cycle (ONPC) principle [5], which requires to check in each cycle one tree node against the pruning criterion. Depending on the check result, the next node to examine is selected among the children, the siblings and the siblings of the parents of the checked node. This ONPC tree-traversal strategy is optimal in terms of complexity since examining multiple nodes concurrently bears the risk of unnecessary computations. The corresponding high-level VLSI architecture is depicted in Fig. 2. It consists of three main computational units [5]: The *vertical-step* unit identifies the first child of the current node with the minimum $\mathcal{M}_\mathrm{P}$. Concurrently, the *horizontal-step* unit computes the next sibling. Based on these outputs and on the current tree-search status, the *pruning-criteria checks* unit determines whether to proceed with the traversal along the vertical or the horizontal direction. In both cases, the next node is available since the vertical- and horizontal-step units run concurrently. The pruning-criteria checks unit also computes the output $\{L^\mathrm{E}_{i,b}\}$ and applies LLR clipping. The architecture also includes a control state machine and input/output registers and supports run-time configurability for the modulation order and the number of antennas. In the following, M_T and Q refer to the run-time setup while $M_\mathrm{T,max}$ and Q_max indicate the maximum configuration supported by the design.

A. Hybrid-Enumeration Architecture

A key issue in the implementation of SISO STS SD is enumeration, since soft inputs prevent the use of simplified methods relying on the geometric properties of $\mathcal{O}$. The straightforward approach of computing and sorting the $\{\mathcal{M}_\mathrm{P}\}$ of all the 2^Q children is very expensive in hardware. A much more efficient solution is to separately determine in each cycle the two best nodes based on $\mathcal{M}_\mathrm{C}$ and $\mathcal{M}_\mathrm{A}$ and then

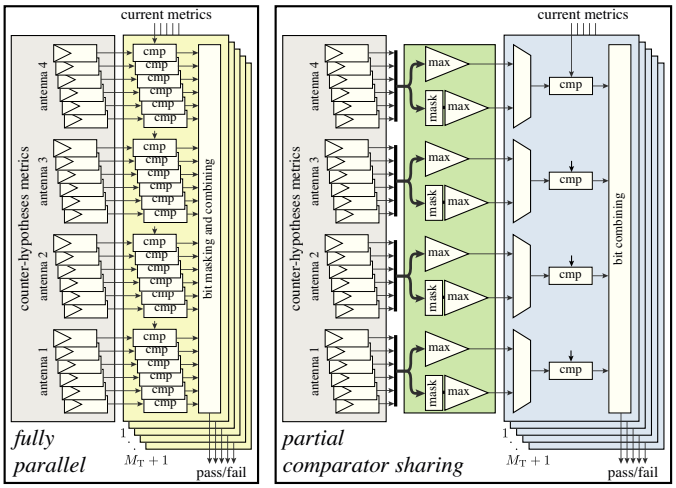


Fig. 3. Pruning-criteria checks architecture with fully parallel comparisons (left) and partial comparator sharing (right) for 4×4 64 QAM.

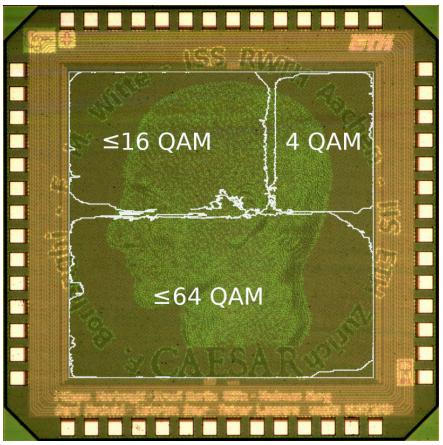


Fig. 4. Chip micrograph with the three SISO STS SD cores highlighted.

select the one with the minimum $\mathcal{M}_\mathrm{P}$ for the next tree-search step (*hybrid enumeration* [7]). The two concurrent $\mathcal{M}_\mathrm{C}$- and $\mathcal{M}_\mathrm{A}$-based enumerations are much less complex than a joint $\mathcal{M}_\mathrm{P}$-based one. In the vertical step, the initialization finds the two minima among the $\{\mathcal{M}_\mathrm{C}\}$ and $\{\mathcal{M}_\mathrm{A}\}$, which can be determined without computing nor sorting metrics [5]. In the subsequent enumeration (horizontal step) the hybrid scheme enables to use the simplified $\mathcal{M}_\mathrm{C}$-based enumeration algorithms developed for non-iterative detectors. Our architecture employs a column-wise decomposition that partitions the constellation points into $2^{Q_\mathrm{max}/2}$ groups with constant real part: the enumeration order within each group is a zig-zag pattern. The selection among the groups is normally based on the comparison of the best candidates, requiring the computation of $\mathcal{M}_\mathrm{C}$ for $2^{Q_\mathrm{max}/2}$ candidates in parallel. As opposed to previous architectures, we notice that not all of these candidates are needed immediately, so that only two, instead of $2^{Q_\mathrm{max}/2}$, $\mathcal{M}_\mathrm{C}$ computation units and a small cache must be instantiated, independently of Q_max. For 64-QAM modulation, this optimization reduces the $\mathcal{M}_\mathrm{C}$-based enumeration area by 30 % compared to a fully parallel implementation [5]. Since this unit lies in the critical path, it is important to notice that this solution comes with no timing penalty. Moreover, several contributions to $\mathcal{M}_\mathrm{C}$ can be precomputed in the vertical-step unit and stored, which ultimately leads to a critical-path reduction of nearly 20 % (gate-level synthesis results).

For $\mathcal{M}_\mathrm{A}$-based enumeration, the horizontal-step unit employs a minimum search over all the 2^Q symbols on the current tree level, based on the $\{\mathcal{M}_\mathrm{A}\}$ computed in parallel to the first $2M_\mathrm{T}$ cycles of the tree search and stored in the $\mathcal{M}_\mathrm{A}$ *storage*. A full compare-select (CS) tree would dominate the critical path, however most of the CS units can be removed by exploiting known relationships among the $\{\mathcal{M}_\mathrm{A}\}$ [5].

B. Pruning-Criteria Checks

Hybrid enumeration requires different pruning criteria for the vertical and the horizontal step. Both criteria involve the comparison of the current metric with one or more (up to $M_\mathrm{T,max}Q_\mathrm{max}$) stored metrics where ultimately only the largest reference metric involved in the comparison determines the outcome. To achieve a short critical path, all comparisons can be carried out in parallel, with a final stage of logic masking the irrelevant ones. However, minimizing the number of cycles for the tree traversal requires $M_\mathrm{T,max}+1$ concurrent pruning-criteria checks, each with the high area costs of $M_\mathrm{T,max}Q_\mathrm{max}$ comparators, as shown on the left of Fig. 3. The architecture in [5] reduces the pruning-criteria checks to two by serializing the search for the next valid node once the forward tree traversal stalls. Such a strategy results in an increased average number of cycles per symbol vector $\mathbb{E}[N_\mathrm{en}]$ and hence a lower average throughput, given by

$$\Theta = \frac{QM_\mathrm{T}}{\mathbb{E}[N_\mathrm{en}]} f_\mathrm{clk} \quad [\mathrm{bit/s}] \quad .$$

A better alternative, implemented in this ASIC, is to rely on concurrent pruning-criteria check units and to reduce the complexity of each unit. To this end, we first select the relevant reference metrics with $2M_\mathrm{T,max}(Q_\mathrm{max}-1)$ CS units that are shared among the $M_\mathrm{T,max}+1$ pruning-criteria checks. Each of these checks then employs only $M_\mathrm{T,max}$ comparators to finalize the decision, as shown on the right of Fig. 3. Compared with [5], this architecture reduces the minimum execution time from $2M_\mathrm{T}+1$ to $M_\mathrm{T}+2$ cycles and, also due to the shorter critical path, achieves a twice as high maximum throughput.

IV. Implementation Results

The proposed SISO STS SD architecture has been implemented in a 90 nm CMOS technology using a standard-performance standard-cell library. The ASIC (Fig. 4) includes three instances of the architecture, all supporting $M_\mathrm{T,max}=4$ but a different Q_max of 2, 4 and 6 respectively. The reference 64-QAM core occupies 0.97 mm^2 (at 69 % area utilization) and reaches a maximum frequency f_max of 193 MHz. The 16-QAM instance has an area of 0.54 mm^2 (at 66 % area utilization) and $f_\mathrm{max}=244$ MHz. Finally, the smallest 4-QAM design requires an area of 0.27 mm^2 (at 67 % area utilization) and achieves $f_\mathrm{max}=330$ MHz. We observe that a 2-bit increase of Q_max doubles the area, meaning that the total hardware complexity grows as $O(2^{Q_\mathrm{max}/2})$. At the same time, f_max degrades only by 20 to 25 % since Q_max only affects tree-structured parts of the critical path, whose depth scales as $O(\log_2 Q_\mathrm{max})$.

TABLE I
IMPLEMENTATION RESULTS AND COMPARISON

	This work	[3]	[8]	[9]
Number of antennas	$\leq 4 \times 4$	$\leq 4 \times 4$	$\leq 8 \times 8$	4×4
Modulation order	≤ 64	≤ 64	≤ 64	64
Iterative MIMO decoding	YES	YES	NO/soft	NO/hard
CMOS technology [nm]	90	90	130	130
Supply voltage [V]	1.0	1.2	1.3	1.3
Area [kGE][a]	212[b]	410	350[b]	114[b]
Max. throughput [Mbit/s]	772	757	624[c]	946[c]
Max. area eff. [Mbit/s/kGE]	3.62	1.85	1.78[c]	8.30[c]
Max. energy eff. [bit/nJ]	8.81	4.00	18.11[c]	12.21[c]

[a] One gate equivalent GE corresponds to a 2-input drive-1 NAND gate.
[b] Required QRD not included because not executed at symbol rate.
[c] General technology scaling [10] to 90 nm and $V_{\text{dd}} = 1.0\,\text{V}$ according to $A \propto 1/S^2$, $f \propto S$, $P \propto 1/U^2$.

TABLE II
SISO DETECTORS COMPARISON (QUASI-STATIC CH., 4×4, 64 QAM)

Scenario	Best comm. performance achievable		Same comm. performance $I_{\text{MMSE}} = 6$		Same comm. performance $I_{\text{MMSE}} = 4$	
	This work	[3]	This work	[3]	This work	[3]
SNR [dB]	**21**[a]	24[a]	24[b]		25[b]	
Iterations	6	6	**2**	6	**1**	4
Throughput [Mbit/s]	5.7	126	96	126	154	189
Area eff. [Mbit/s/kGE]	0.03	0.31	**0.45**	0.31	**0.73**	0.46
Energy eff. [bit/nJ]	0.07	0.67	**1.04**	0.67	**1.76**	1.00

[a] Minimum SNR to have PER $\leq 1\,\%$ with $I = 6$.
[b] Minimum SNR for MMSE PIC to have PER $\leq 1\,\%$ with $I = I_{\text{MMSE}}$; SD run-time constraints and I set to have PER $\leq 1\,\%$ at the same SNR.

A. Peak Performance

At high SNR, the detector reaches its maximum throughput since the number of cycles $\mathbb{E}[N_{\text{en}}]$ approaches the minimum $M_{\text{T}} + 2$. For $M_{\text{T}} = 4$, the 64-QAM reference design achieves 772 Mbit/s while the 16- and 4-QAM instances reach 651 and 440 Mbit/s at nominal supply voltage, respectively. In this regime, the 64-QAM core is twice as area efficient as the only other known SISO detector [3], as shown in Tbl. I. Compared with non-iterative detectors [8], [9], the large SNR-performance gain provided by SISO STS SD, which for instance can exceed 5 dB in the case shown in Fig. 1(a), only entails a rather low degradation in the efficiency. The same observations apply to energy efficiency, which is 8.81 bit/nJ for the 64-QAM core. The 4- and 16-QAM instances achieve 8.29 and 7.82 bit/nJ respectively.

A comparison of the three cores running in the same configuration and frequency shows the costs of modulation flexibility. The detection of a 16-QAM signal on the 64-QAM core requires up to 24 % more energy compared to a circuit that is limited to 16-QAM. Similarly, 4-QAM detection consumes up to 119 % more energy on the 64-QAM instance than on the specialized 4-QAM core. The energy overhead for supporting a higher Q_{max} is however much lower than the related area costs, which double with a 2-bit increase in Q_{max}.

B. Average Performance

Due to the variable STS SD execution time, the measured area and energy efficiency varies with the SNR and with the target communication performance. Tbl. II compares this work with the SISO MMSE PIC ASIC [3] in different scenarios, all with a quasi-static channel and a 4×4 64-QAM configuration, which corresponds to the mode with the highest STS SD complexity. The SISO STS SD ASIC reaches a 3 dB lower (better) operating SNR at $I = 6$, though with a very high complexity and hence a reduced area and energy efficiency (Tbl. II, col. 1). The complexity of STS SD, however, can effectively be reduced by introducing run-time constraints so that the performance matches the target PER at the given SNR. This adaptivity positively affects area and energy efficiency: in fact, when compared with MMSE PIC at the same SNR and target PER (Tbl. II, cols. 2 and 3), the STS SD ASIC often achieves a better efficiency. Moreover, STS SD often reaches the target PER with fewer iterations, thus reducing the channel decoding effort and providing additional energy savings.

V. CONCLUSION

In this paper we have shown how key SISO SD implementation issues can be solved with efficient tree-search and enumeration strategies, improving the architecture presented in [5] to double the maximum throughput and support run-time configurability. Measurements show that our ASIC achieves a competitive area and energy efficiency and, at the same time, a better communication performance and robustness to the channel conditions than other known MIMO detectors.

ACKNOWLEDGEMENTS

This work is supported by the UMIC Research Centre, RWTH Aachen University. We thank F. Gürkaynak and B. Muheim (Microelectronics Design Center, ETH Zürich) for the invaluable support in implementing and measuring the ASIC.

REFERENCES

[1] B. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389–399, Mar. 2003.
[2] L. Schmitt, H. Meyr, and D. Zhang, "Systematic design of iterative ML receivers for flat fading channels," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 1897–1901, Jul. 2010.
[3] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using parallel interference cancellation," in *IEEE Journal of Solid-State Circuits, to appear*, Jul. 2011.
[4] C. Studer and H. Bölcskei, "Soft-input soft-output single tree-search sphere decoding," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4827 –4842, Oct. 2010.
[5] E. M. Witte *et al.*, "A scalable VLSI architecture for soft-input soft-output single tree-search sphere decoding," *IEEE Trans. Circuits Syst. II*, vol. 57, no. 9, pp. 706–710, Sep. 2010.
[6] D. Wübben *et al.*, "Efficient algorithm for decoding layered space-time codes," *Electronics Letters*, vol. 37, no. 22, pp. 1348–1350, Oct. 2001.
[7] C.-H. Liao *et al.*, "Combining orthogonalized partial metrics: Efficient enumeration for soft-input sphere decoder," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2009, pp. 1287 –1291.
[8] C.-H. Liao, T.-P. Wang, and T.-D. Chiueh, "A 74.8 mW soft-output detector IC for 8×8 spatial-multiplexing MIMO communications," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 2, pp. 411–421, Feb. 2010.
[9] M. Shabany and P. G. Gulak, "A 0.13 µm CMOS 655 Mbit/s 4×4 64-QAM k-best MIMO detector," in *Dig. Techn. Papers, IEEE ISSCC*, Feb. 2009, pp. 256–257,257a.
[10] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits - A Design Perspective*, 2nd ed. Prentice Hall, Dec. 2002.