A 1.96 Gb/s Massive MU-MIMO Detector for Next-Generation Cellular Systems

Chi-Chih Wen, Yu-Chi Lee, Yi-Chung Wu, Chen-Chien Kao, Chia-Hsiang Yang
Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan

Abstract

This work presents a massive multi-user MIMO (MU-MIMO) detector for cellular systems that can support 256 base station (BS) antennas and 32 users with modulation of up to 256-QAM. The chip delivers a throughput of 1.96 Gb/s and dissipates 87 mW at 290 MHz. Compared to the state-of-the-art detectors, the chip achieves a 5.7-to-30.5x higher area efficiency with 9.2-to-31.2x lower normalized energy.

Introduction

Massive MU-MIMO has become one of the key technologies for next-generation systems as the demand for system capacity becomes increasingly high. Higher throughput, higher reliability and higher spectral efficiency can be attained compared to small-scale MIMO by employing tens-to-hundreds of antennas at the BS to serve multiple users, as shown in Fig. 1. Despite the advantages brought by massive MU-MIMO, a critical task in such a system is the computational complexity of the data detection due to the increased number of base station antennas and the increased number of users. Previous designs [1][2] rely on explicitly computing the minimum mean square error (MMSE) detection, which involves high computational complexity for matrix inversion. A detector that implements the approximate message-passing algorithm achieves a higher area efficiency [3], but the complexity is still high because of the iterative computations for statistics related parameters.

This paper presents the first dichotomous coordinate descent (DCD) [4] based MU-MIMO detector. Fig. 2 shows the work flow of the detection algorithm. The DCD algorithm is originally used to obtain the optimal solution of least-squares problems without the need for multiplications and divisions. Since the problem formulation for MIMO detection also has the least-squares form, the DCD algorithm is used to obtain the zero-forcing (ZF) solution first. The solution derived from DCD is then used as an initial solution. The error performance can be further improved by leveraging interference cancellation through orthogonal projection [5]. To achieve better performance with error-correcting codes, the proposed detector also generates soft output, which can be implemented by evaluating piecewise mapping for soft information (i.e., log likelihood ratio (LLR)), where S_b^0 and S_b^1 denote sets of constellation symbols with the b-th bit being 0 and 1, respectively.

System Architecture

Fig. 3 shows the architecture of the proposed MIMO detector. We consider an uplink massive MU-MIMO system with U single-antenna users and B BS antennas. The detector mainly comprises an initial solution engine, an interference canceller, and a soft information extractor. The detector is pipelined so as to increase the overall throughput. The data buffer for $\mathbf{G} = \mathbf{H}^H\mathbf{H}$ matrix is folded by leveraging the symmetric characteristic, reducing the buffer size by 50%. Conventionally, DCD sequentially updates each coordinate of the solution vector, resulting in a high latency. In this work, a modified DCD scheme is proposed by updating coordinates in parallel, thereby significantly reducing the latency. In the initial solution engine, criteria for a better solution are checked in parallel for all coordinates and the result is used to accumulate the solution vector. The auxiliary $\mathbf{q}$ vector is updated by summing the columns of $\tilde{\mathbf{G}}$, where $\tilde{\mathbf{G}}$ is the real-valued representation of $\mathbf{G}$. To reduce hardware complexity and increase utilization, additions for half of the coordinates are performed and the same blocks are reused for the remainder, as shown in Fig. 4. However, this also introduces a bottleneck on pipelining for the initial solution engine. Thus, interleaved processing is adopted to reduce the duration for generating one initial solution. The generated initial solution then sent to a parallel-to-serial converter sequentially and enters the interference canceller in a symbol-by-symbol manner.

The interference cancellers consists of two multiply-accumulate (MAC) arrays and two quantizers. A MAC array that consists of 32 MAC units, each corresponding to one user, is deployed to perform matrix-vector multiplications. In order to generate soft information efficiently, piecewise mapping is adopted. Piecewise linear functions for each modulation and bit index exhibit a certain degree of symmetry, which can be utilized to reduce the implementation complexity. Aside from the symmetry, the piecewise linear functions of certain bit index contain multiple neighboring segments whose slopes are close to each other. These segments can be merged into a single slope segment that can still connect to both ends to simplify the evaluation hardware of the piecewise linear functions. Fig. 5 shows the circuits of the original and the approximate implementations. The area of the hardware for $\mathbf{q}$ vector updater is reduced by 48% and the number of segments for soft information is reduced by 79%, as shown in Fig. 6.

Experimental Results and Conclusion

Fig. 7 shows the BER performance of the original DCD scheme and the proposed parallel one. A 99% reduction in the number of iterations (k) can be achieved under the same performance requirement. The number of iterations of the proposed parallel DCD scheme is chosen as 12 due to its sufficient performance. Fig. 8 shows the BER performance of this work for the Rayleigh fading channel and for the spatially correlated channel, where ρ denotes the correlation coefficient between adjacent BS antennas. A *beyond* MMSE (the limit of [1] and [2]) detection performance is achieved with gains of 0.7dB and 1dB for the Rayleigh fading channel and the spatially correlated channels, respectively. Note that only the packet error rate is reported in [3], so a fair comparison is unavailable. To test the effectiveness of the generated soft output, the output of the detector is fed into a serial concatenated turbo code decoder. Fig. 9 shows that an approximately 3 dB SNR gain can be obtained compared to the hard-output design.

Fig. 10 shows the chip micrograph. The chip integrates 1.07M logic gates in an area of 0.73 mm^2 in 40-nm CMOS. The chip dissipates 87 mW and delivers a throughput of 1.96 Gb/s at 290 MHz from a 1.1V supply. It is able to support up to 256 BS antennas and 32 users with modulation order ranging from 4-QAM to 256-QAM. Table I shows the comparison with the state-of-the-art designs [1-3]. This work supports the largest number of users and BS antennas for the correlated channel. It also provides soft output to facilitate error correction. The proposed MU-MIMO detector does not involve complex operations, thus providing high scalability to support even larger MIMO configurations. Compared to prior art, this work achieves a 5.7-to-30.5x higher area efficiency with 9.2-to-31.2x lower normalized energy, as shown in Fig. 11.

Acknowledgments

This work is supported by MOST and Intelligent & Sustainable Medical Electronics Research Fund in NTU. The authors also thank TSRI for technical support on chip design and fabrication.

References

[1] H. Prabhu *et al.*, *IEEE ISSCC*, pp. 60-61, Feb. 2017.
[2] W. Tang *et al.*, *IEEE ISSCC,* pp. 224-225, Feb. 2018.
[3] C. Jeon *et al.*, *IEEE SSCL*, vol. 2, no. 9, pp. 127-130, Sept. 2019.
[4] Y. Zakharov *et al.*, *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3150-3161. July 2008.
[5] M. Mandloi *et al.*, *IEEE Comm. Lett.*, vol. 21, no. 3, pp. 568-571, Mar. 2017.

978-1-7281-9942-9/20/$31.00 ©2020 IEEE

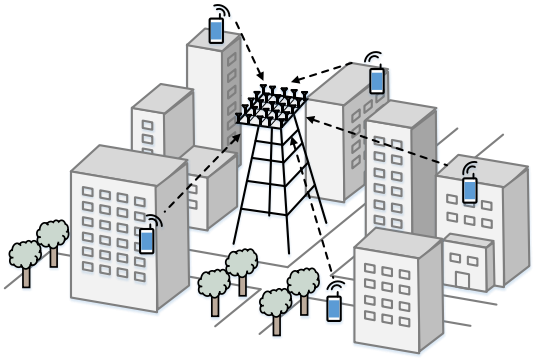

Fig. 1. Uplink massive MU-MIMO system.

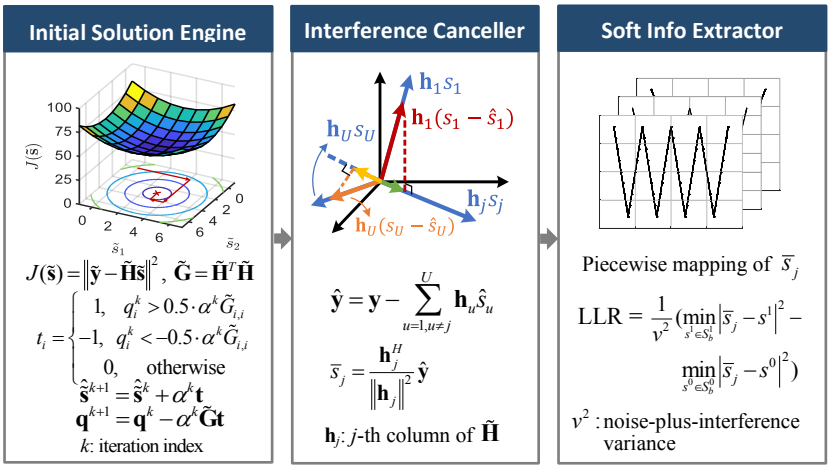

Fig. 2. Work flow of the proposed MIMO detection algorithm.

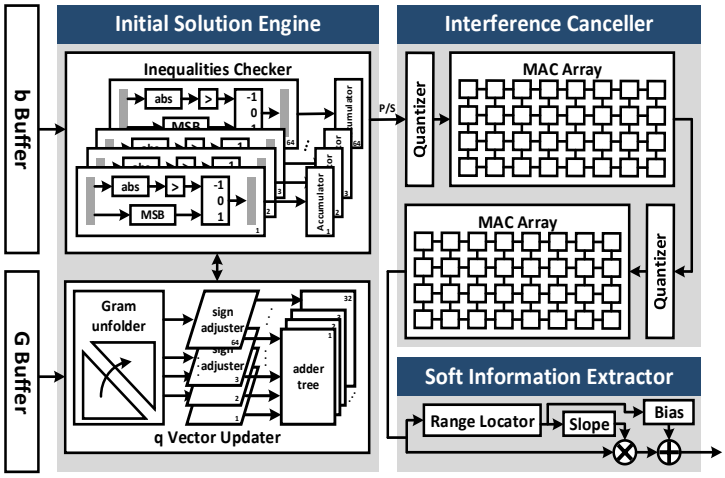

Fig. 3. System architecture of the proposed detector.

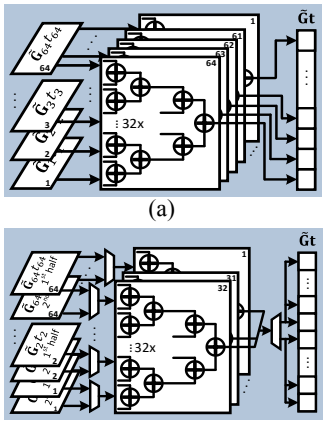

Fig. 4. Architecture optimization of **q** vector updater: (a) direct-mapped and (b) modified.

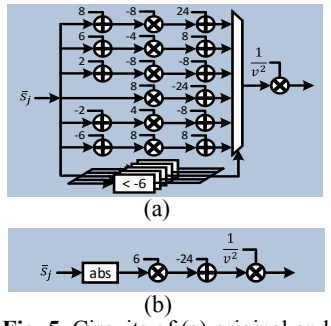

Fig. 5. Circuits of (a) original and (b) approximate implementations of piecewise linear functions (for the 2nd bit in 64-QAM).

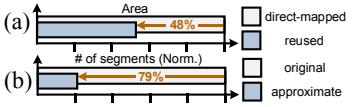

Fig. 6. Complexity reduction.

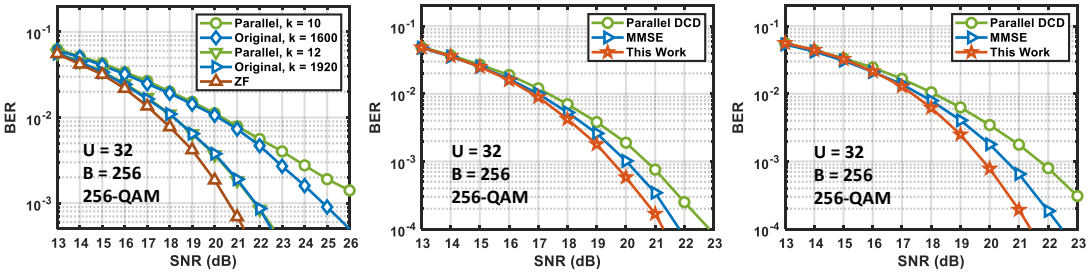

Fig. 7. BER performance of original and parallel DCD schemes.

(a) (b)

Fig. 8. BER performance of this work under (a) Rayleigh fading channel and (b) spatially correlated channel (with $\rho = 0.6$).

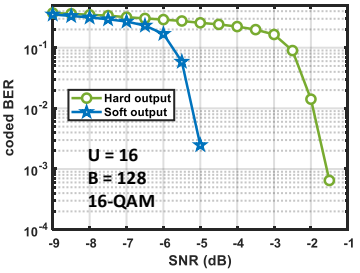

Fig. 9. BER performance of hard output and soft output (using a serial concatenated turbo code).

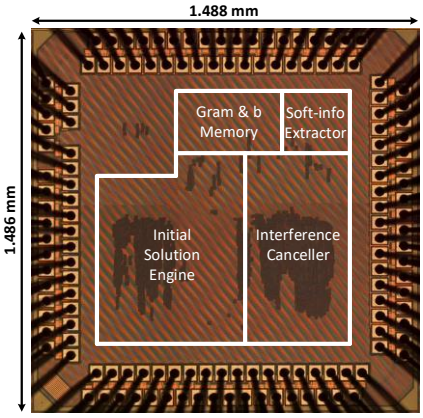

Fig. 10. Chip micrograph.

	ISSCC'17 [1]	ISSCC'18 [2]	SSCL'19 [3]	This work
MIMO (B x U)	128 x 8	128 x 16	256 x 32	256 x 32
Modulation (QAM)	256	4-to-256	256	4-to-256
Soft Output	No	No	Yes	Yes
Correlated Channel	-	Yes	Yes	Yes
Technology [nm]	28	28	28	40
Area [mm^2]	-	2.0	0.37	0.73
Gate Count [kGE]	288	3607	1110	1073
Supply Voltage [V]	0.9	1.0	0.9	1.1
Frequency [MHz]	300	569	400	290
Power [mW]	18	127	151	87
Throughput [Gb/s]	0.3	1.8	0.35	1.96
Norm. Energy [pJ/bit] [a, b]	1371	403	431	44
Area Efficiency [Mb/s/kGE] [b]	0.06	0.12	0.32	1.83

[a] Normalized to 40nm. [b] Normalize by (U/32)^2.

Table I. Comparison with state-of-the-art designs.

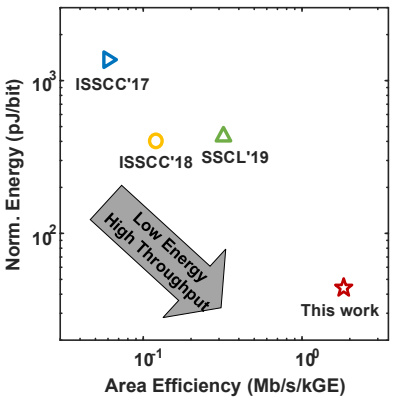

Fig. 11. Performance summary.