A 0.58mm^2 2.76Gb/s 79.8pJ/b 256-QAM Massive MIMO Message-Passing Detector

Wei Tang[1], Chia-Hsiang Chen[1,2], Zhengya Zhang[1]
[1]University of Michigan, Ann Arbor, MI, [2]Intel Labs, Santa Clara, CA

Abstract

A 0.58mm^2 40nm CMOS message-passing detector (MPD) is designed for a 256-QAM massive MIMO system supporting 32 concurrent mobile users in each time-frequency resource. Leveraging channel hardening in massive MIMO, a symbol hardening technique is proposed to reduce MPD's complexity by more than 60% with minimal SNR loss. The MPD is implemented in a 4-layer 2-way interleaved architecture to enable a 2.76Gb/s throughput (average 4.9 iterations at 27dB SNR with early termination) using 76% smaller area than a fully parallel architecture. With dynamic precision control and clock gating to exploit algorithmic properties, the energy is reduced to 79.8pJ/b (or 2.49pJ/b per TX antenna).

Introduction

Massive MIMO has been identified as a key disruptive technology for the upcoming fifth generation (5G) wireless communication systems [1]. Massive MIMO is a multi-user MIMO technique that relies on a large number, e.g., hundreds, of base station antennas to serve a multiplicity of, e.g., tens, of autonomous single-antenna users in each time-frequency resource [2]. The large number of antennas provide a high spatial multiplexing gain for an increased capacity; and the radiated energy can be focused to the intended receivers for an energy-efficient downlink transmission. However, massive MIMO requires complex and power-hungry signal processing for uplink processing in base station. Recently iterative message-passing detectors (MPD) have been proposed [3]. An MPD exploits channel hardening in a massive MIMO system, and it approximates the sum of interference using a Gaussian distribution. An MPD has a comparable complexity as an MMSE detector, and yet in a massive MIMO system, it outperforms an MMSE detector by a large margin.

Iterative Message-Passing Detection

In an $N_r \times N_t$ M-QAM massive MIMO uplink system (Fig. 1), N_t users transmit at the same time and frequency, and N_r antennas at the base station pick up not only the intended transmissions but also interference plus noise. A MIMO detector attempts to retrieve the intended transmissions by canceling the interference. An MPD uses a set of N_t interference cancellation PEs (IPE) and a set of N_t constellation matching PEs (CPE) to iteratively estimate the N_t user symbols. An IPE is connected to N_t CPEs (Fig. 3(a)), and it computes a Gaussian approximation of a user symbol after canceling the interference from the other $N_t - 1$ users. A CPE is connected to one IPE (Fig. 3(a)), and it estimates a symbol by considering its likely locations at M constellation points. In every iteration, an IPE requires $8(N_t - 1)$ real-valued multiply-accumulates (MAC) to compute the mean and variance of the Gaussian approximation; and a CPE requires $2\sqrt{M}$ Gaussian evaluations and $4\sqrt{M}$ MACs to compute the mean and variance of a symbol estimate. The complexity of an MPD grows with the number of users N_t and the order of modulation M, presenting an implementation challenge for a high-order massive MIMO system.

Complexity Reduction Leveraging Channel Hardening

In this work, we design an MPD for a 128×32 (N_r = 128, N_t = 32) 256-QAM system. With channel hardening in a massive MIMO system, the variance of symbol estimate converges at a fast pace. The fast convergence permits the use of a small fixed variance to eliminate the variance calculations, saving nearly 4K MACs in 32 IPEs and 1K MACs in 32 CPEs. The use of a small variance further reduces the CPE processing to making one hard symbol decision based on its distribution. The symbol hardening technique eliminates 1K MACs and 1K Gaussian evaluations in 32 CPEs, sacrificing 0.25dB of SNR at 10^{-4} BER, but the optimized MPD still outperforms an MMSE detector by 1dB (Fig. 2).

Layered and Interleaved Architecture

The MPD can be fully parallelized with 32 IPEs and 32 CPEs (Fig. 3(a)), requiring nearly 4K MACs and 10K interconnects. Despite the high throughput, the fully parallel architecture is dominated by global wiring, resulting in a large silicon area, low clock frequency, and high power. Therefore we opt for a more compact design (Fig. 3(b)) that divides 32 users into 4 layers with 8 users per layer for processing, using 1/4 as many MACs in each IPE. In each layer, 32 IPEs compute the sum of interference contributed by 8 users and update the symbol estimates. The updated estimates are then forwarded to the next layer. The intra-iteration forwarding speeds up convergence by nearly 2× compared to the flooding schedule used in the fully parallel architecture. Based on trial designs, the 4-layer architecture reduces area and power by 66% and 61%, respectively, but with faster convergence, its throughput is only 28% lower (Fig. 5(a,b)).

The layered architecture imposes data dependency between layers, requiring one layer to be completed in one clock cycle. To relax the data dependency and reduce area, we halve the number of IPEs to 16 and time-multiplex their use between 2 groups in a 2-stage pipeline (Fig. 3(c)). In each cycle the symbol estimates are computed for either group 1 or group 2 users, which are interleaved to avoid pipeline stalling. Based on trial designs, the 4-layer 2-way interleaved architecture (Fig. 4) reduces area and power by 76% and 65% respectively over the baseline (Fig. 5(a,c)).

Dynamic Precision Control and Clock Gating

The datapath power is dominated by the 512 MACs. To save dynamic power, we adapt the multiplier precision dynamically to exploit the MPD's convergence behavior. In early iterations, the MPD makes coarse symbol estimates using 6b×2b low-precision multiplications; but in late iterations, the MPD fine tunes symbol estimates using 12b×4b full-precision multiplications. Each full-precision multiplier is designed to support the low-precision mode with LSBs disabled (Fig. 6(a)), saving 75% of the switching activity and the associated dynamic power.

Registers are used as data memory to support the wide data access required by the architecture. The memory access is regular (Fig. 6(b)), e.g., the 3Kb interference memory (P MEM) is updated once every 8 cycles. Therefore, we implement clock gating to turn off the clock input when the memory is not updated to save dynamic power.

Test Chip Measurement Results

A massive MIMO MPD test chip (Fig. 8) is fabricated in TSMC 40nm CMOS technology. The chip includes a 0.58mm^2 MPD core, a PLL to generate clock, a test memory to store test vectors, and scan chains for input and output. The chip is measured to run at a maximum frequency of 425MHz at the nominal supply voltage of 0.9V in room temperature, dissipating 221mW. Incorporating the proposed architecture techniques along with dynamic precision control and clock gating, the MPD's power dissipation is reduced by 70% and energy per bit reduced by 52% over the baseline (Fig. 5(a,d)). With early termination enabled on chip, detection converges in 5.7, 5.2 and 4.9 iterations on average at 23dB, 25dB and 27dB SNR, allowing a throughput up to 2.76Gb/s or 86Mb/s per mobile user (Fig. 7). A higher throughput for massive MIMO can be achieved by deploying multiple MPD modules and applying interleaving.

The results are compared with state-of-the-art MIMO detector chips in Table I. Note that all the previous designs, including SD [4]-[6] and MMSE [7], will incur much higher implementation costs in a massive MIMO system. To the best of our knowledge, this chip is the first silicon demonstration that supports large-scale multi-user MIMO detection.

Acknowledgements

The work was supported in part by NSF and Intel.

References

[1] F. Boccardi, et al., *Commun. Mag.*, 2014.
[2] E. G. Larsson, et al., *Commun. Mag.*, 2014.
[3] T. L. Narasimhan, A. Chockalingam, *J. Sel. Topics Signal Process.,* 2014.
[4] F. Borlenghi, et al., *Proc. ESSCIRC*, 2012.
[5] B. Noethen, et al., *ISSCC Dig. Tech. Papers*, 2014.
[6] M. Winter, et al., *ISSCC Dig. Tech. Papers*, 2012.
[7] C.-H. Chen, et al., *ISSCC Dig. Tech. Papers*, 2015.

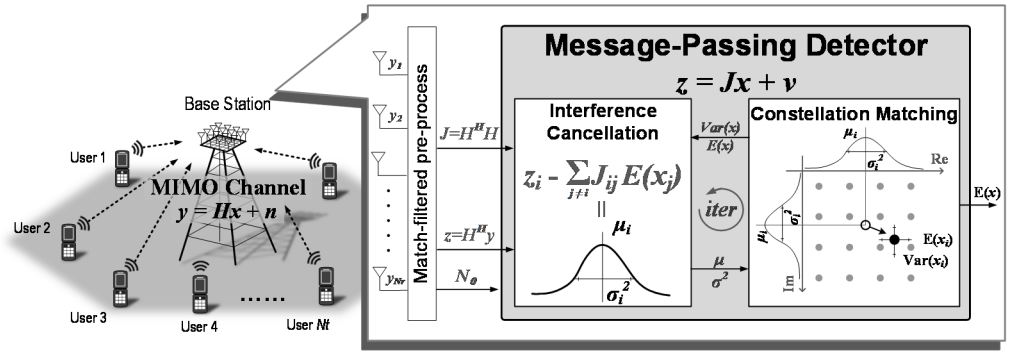

Fig. 1. Illustration of an uplink large-scale MIMO system of N_t single-antenna users and N_r antennas at base station; and a top-level block diagram of a message-passing detector (MPD).

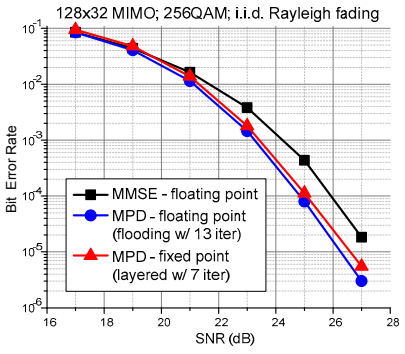

Fig. 2. Bit error rate (BER) performance of 128x32 uplink MMSE detection and MPDs.

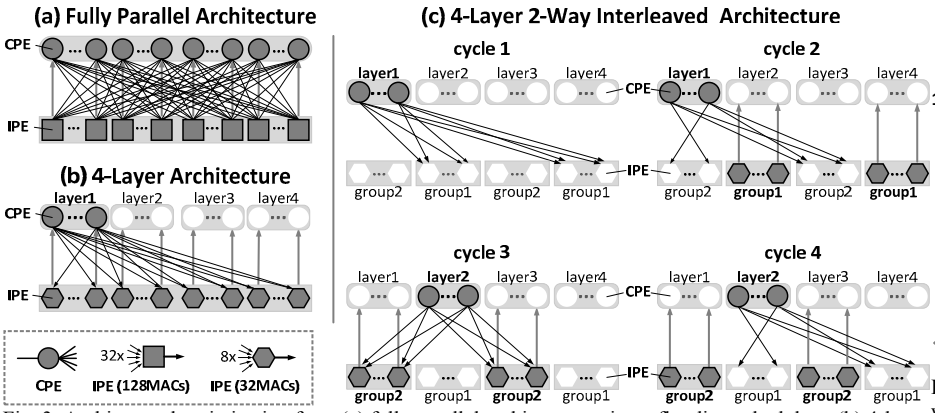

Fig. 3. Architectural optimization from (a) fully parallel architecture using a flooding schedule to (b) 4-layer architecture, and (c) 4-layer 2-way interleaved architecture (the first 4 pipeline cycles are shown).

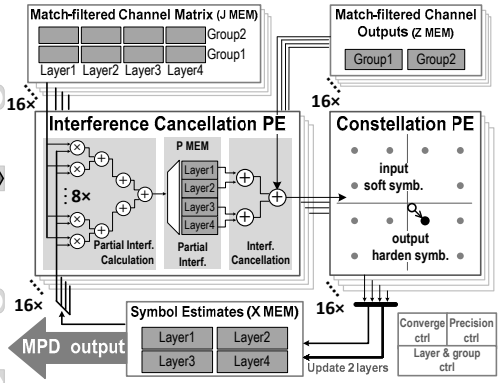

Fig. 4. Detailed block diagram of the proposed 4-layer 2-way interleaved MPD using 16 Interference cancellation PEs (IPEs) and 16 Constellation matching PEs (CPE).

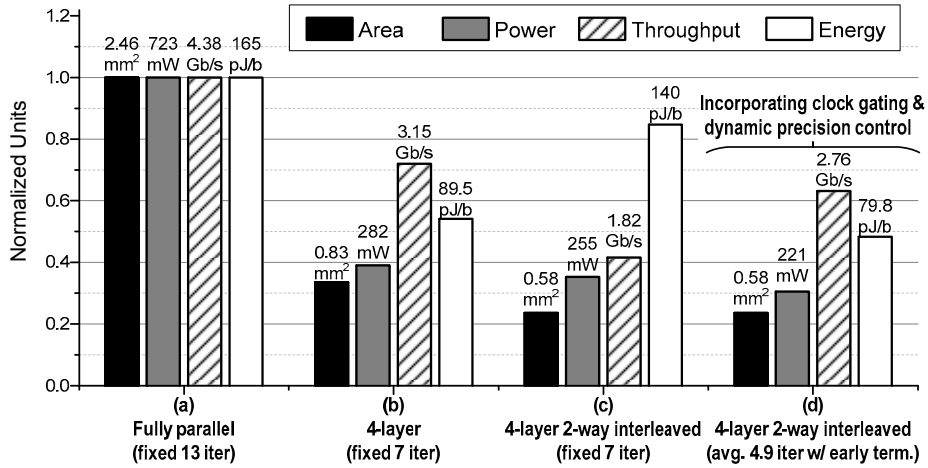

Fig. 5. Area, power, throughput and energy improvement from (a) fully parallel architecture using flooding schedule to (d) 4-layer 2-way interleaved architecture with early termination, incorporating dynamic precision control and clock gating.

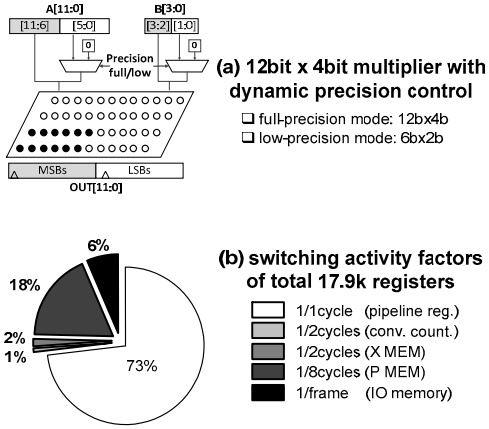

Fig. 6. Low power techniques: (a) dynamic precision multiplier, and (b) clock gating: breakdown of register usage and the corresponding update activity.

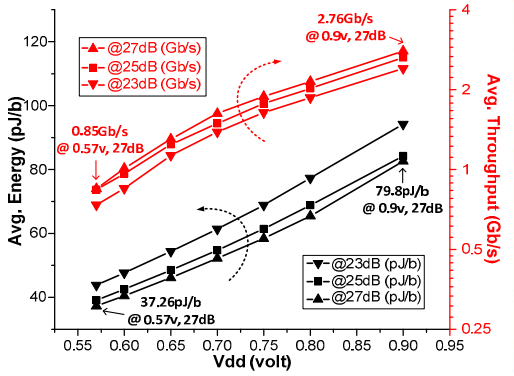

Fig. 7. Measured average throughput (red) and average energy (black) with voltage scaling at different SNRs.

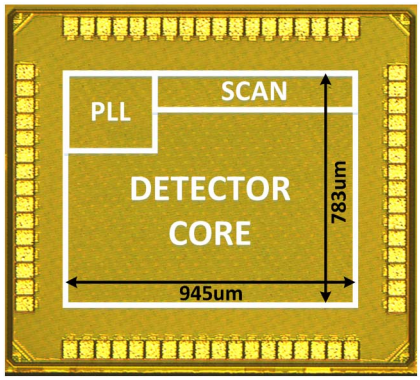

Fig. 8. Chip microphotograph.

TABLE I. COMPARISON TABLE OF STATE-OF-THE-ART MIMO DETECTOR DESIGNS

Detector	Borlenghi [4]	Noethen [5]	Winter [6]	Chen [7]	This Work
Algorithm	SD [a]	SD [a]	SD [a]	MMSE	**MPD [b]**
MIMO size (Nr × Nt)	MIMO ≤ 4×4	MIMO ≤ 4×4	MIMO ≤ 4×4	MIMO 4×4	**Massive MIMO 128×32**
Modulation	≤ 64	≤ 64	≤ 64	256	**256**
Technology [nm]	65	65	65	65	**40**
Core area [mm²]	2.78	-	0.31	0.7	**0.58**
Frequency [MHz]	135	445	333	517	**425**
Power [mW]	-	87	38	26.5	**220.6**
Throughput [Gb/s]	0.194	0.396	0.296-0.807	1.379	**2.76 [c]**
Area efficiency [Gb/s/mm²]	0.07	-	0.96-2.6	1.97	**4.76**
Energy [pJ/b]	920	220	48	19.2	**79.8**
Energy efficiency [pJ/b/TX antenna]	230	55	12	4.8	**2.49**

(a): sphere decoding. (b): message-passing detection.
(c): early termination with average 4.92 iterations at SNR=27dB.