An 8pJ/bit Charge-Transfer Analog Compute-In-Memory Processor for Massive MIMO

Ying-Tuan Hsu, Michael Laun, Wei Tang, Luke Wormald, Zhengya Zhang, and Michael P. Flynn
Electrical Engineering and Computer Science, University of Michigan
Ann Arbor, Michigan, USA
Email: ythsuf@umich.edu

***Abstract*—A 28nm CMOS SRAM analog computing-in-memory (CIM) detector targeting 64-antenna x 8-user massive multiple-input multiple-output (MIMO) uplink system has a measured energy efficiency of 8pJ/bit, including ADCs. The proposed charge-transfer architecture directly receives analog inputs and supports signed weights for complex MAC, and reduces the ADC count from 128 to 16. The charge-transfer architecture, assisted by ring amplifiers, mitigates the signal attenuation, nonlinearity, and drive limitations of the conventional charge-redistribution architecture while maintaining high energy efficiency. A dual-bit multiplication mechanism improves computation throughput and area efficiency. In measurements, the prototype achieves a BER = 10^{-3} with an SNR loss of only 0.36dB compared to a reference 28-bit MMSE digital detector with ideal ADCs. The area efficiency is 10.9Gb/s/mm^2. The measured energy efficiency is 4.5x better than the reference digital MIMO detector.**

Keywords—Computing-in-memory (CIM), SRAM, Multiple-input multiple-output (MIMO), Charge transfer

I. Introduction

MIMO increases wireless channel capacity but requires extensive processing. Conventional MIMO processing is dominated by analog-to-digital conversion and multiply-and-accumulate (MAC) operations. MIMO processing requires two high-resolution ADCs for each antenna element, which is expensive in terms of die area and power consumption (Fig. 1). Furthermore, digital MAC of the high-resolution ADC outputs is very energy intensive. This work demonstrates the potential of analog computing-in-memory (CIM) to reduce the MAC power consumption as well as the number ADCs, their power consumption and area.

Recent analog CIM designs for ML applications ([1], [2], and [3]) utilize charge redistribution MAC. However, as shown in Fig. 2 the approach is unsuitable for analog-input MIMO processing.

1) Due to the characteristics of the MIMO detection, the output product is concentrated over a small range of values. However, charge redistribution CIM provides no gain from passive MAC. This reduced output swing increases the required resolution and noise-performance of the downstream ADCs, resulting in higher energy consumption.

2) MIMO has stringent linearity and noise requirements. Parasitic capacitance on the charge redistribution summation line reduces CIM gain and linearity. The large routing parasitic capacitance reduces the output swing. The non-linear parasitic capacitance of switches and the input to follow-on circuitry degrades MAC linearity.

3) Bit-serial computation reduces the capacitor area needed for the multi-bit weight multiplication [3] through the reuse of a single capacitor, but the additional cycles required greatly reduce throughput.

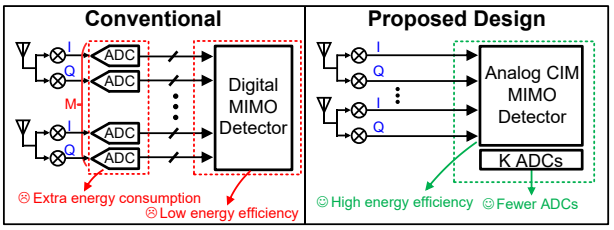


Fig. 1. Proposed system concept.

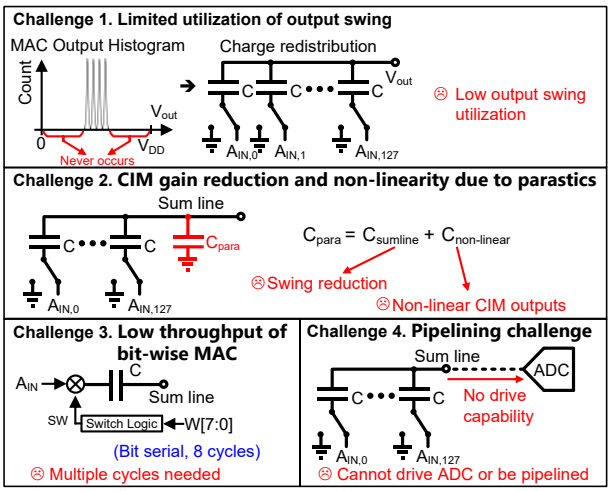


Fig. 2. Challenges of charge-redistribution CIM for MIMO detection.

4) Passive charge redistribution lacks the drive capability required to pipeline the MAC operation and analog-to-digital conversion, thereby limiting the overall throughput.

We introduce a charge-transfer CIM for efficient high-linearity MIMO processing. The prototype MIMO CIM directly processes the analog I/Q antenna signals, reducing the number of ADCs from 128 (2x antennas) to 16 (2x user streams). The area and energy efficiencies are 10.9Gb/s/mm^2 and 8pJ/bit, respectively, including ADCs.

II. Charge Transfer Architecture

We overcome the challenges of prior CIM designs with a new charge-transfer CIM architecture that meets the needs of MIMO processing (Fig. 3):

Charge transfer architecture: Our MIMO CIM processor introduces charge transfer instead of passive charge redistribution to provide higher gain, better linearity and enabling pipelining. In the charge transfer technique, the charge sampled by the input capacitor (C_{1x}) fully transfers to the feedback capacitor (C_f) without being attenuated by all the capacitors connected to the sum line.

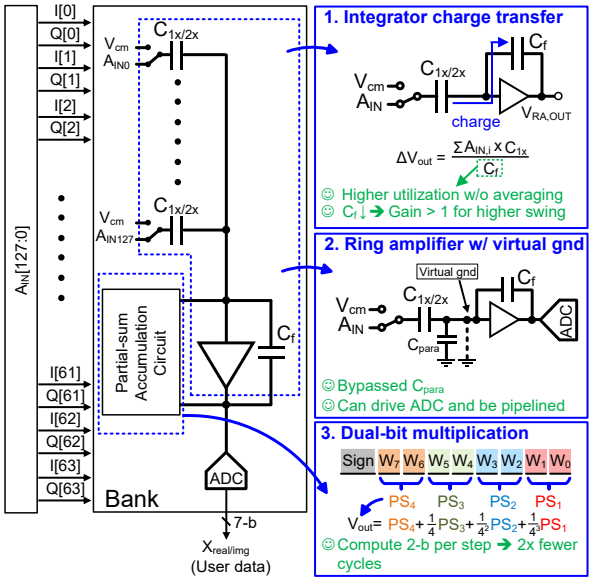


Fig. 3. Proposed features.

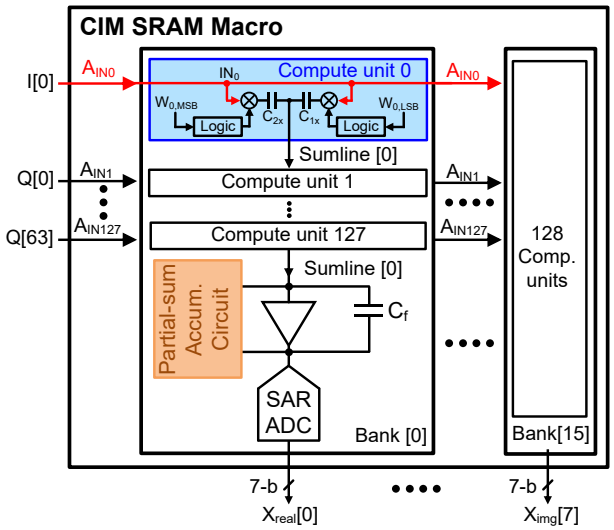


Fig. 4. CIM Macro.

Integrator charge transfer provides important benefits:

1) The integrator provides a high gain (i.e., 10), relaxing the follow-on ADC noise requirements. The ratio of the input capacitance to the feedback capacitance sets the gain enabling gains much larger than 1.

2) A virtual ground terminates the sum line. This low impedance virtual ground of the integrator improves multiplication linearity by minimizing the effect of non-linear parasitics.

3) The integrator circuit scales and sums partial multiplication results.

The ring amplifier in the integrator makes the charge transfer architecture and pipelining practical and efficient. Ring amplifiers excel at slewing capacitive loads and provide a near rail-to-rail output swing.

Partial-sum accumulation and **Dual-bit multiplication** enable two bits per cycle multiplication and pipelining. This halves the number of cycles for signed 9-bit multiplication. Pipelining allows the output ADCs to run concurrently.

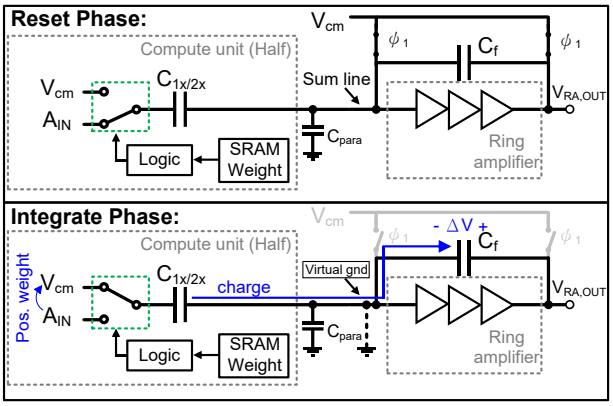


Fig. 5. Operation of charge transfer technique.

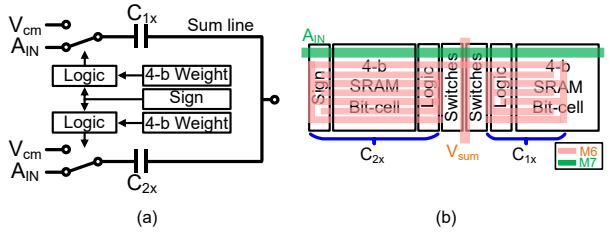


Fig. 6. (a) Schematic and (b)Layout of the compute unit.

Fig. 4 shows the overall architecture of the proposed analog CIM circuit supporting 64 antennas and 8 users. 128 I/Q antenna-element analog inputs (A_{IN}) directly feed 16 banks of the CIM macro. Compute units in each bank perform multiplication. A ring-amplifier-based integrator terminates each bank with a virtual ground and accurately sums the compute unit outputs. Each ring amplifier drives an efficient 7-bit SAR ADC. The 16 SAR ADCs provide I/Q digital outputs for 8 users.

III. CIM Architecture

A. Compute Unit and Charge Transfer

Fig. 5 demonstrates the charge transfer technique with a single input capacitor and no integrator scaling. Each CIM cycle has two phases: Reset and Integrate. The Reset phase clears the charge on the feedback capacitor (C_f) and samples either A_{IN} or the V_{cm} onto the input capacitor, depending on the sign. During the Integrate phase, the input switch toggles to the alternate input (i.e., V_{cm} or A_{IN}). (If the weight value is 0, then the input switch does not change.) Meanwhile, the sampled charge is transferred to the feedback capacitor. The feedback capacitor sums the charge from all the compute units.

B. Patial-sum Accumulation for Dual-bit Multiplication

Fig. 6 shows the implementation of the compute unit. We introduce dual-bit multiplication to take advantage of the area efficiency of bit-serial operation but provide twice the throughput (Fig. 7). Dual-bit multiplication replaces the single input capacitor with a binary-weighted pair (C_{1x}, C_{2x}). C_{2x} (the double-sized capacitor) is responsible for the more significant bit in each cycle. Four dual-bit compute cycles implement the signed 9-bit multiplication. Each cycle produces a partial sum for two of the eight-bit weights. Fig. 8 shows the partial-sum accumulation circuit, which uses the integrator to combine and scale the partial results. Residue

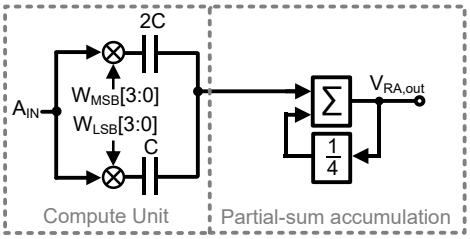


Fig. 7. Concept of dual-bit multiplication and partial-sum accumulation.

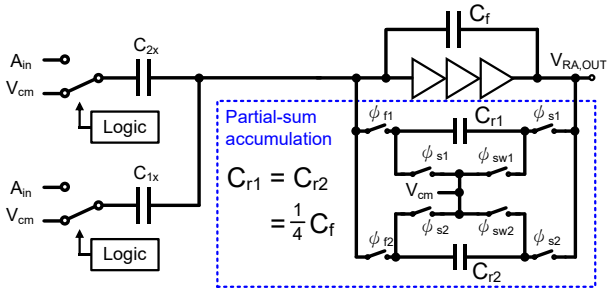


Fig. 8. Schematic of the partial-sum accumulation circuit.

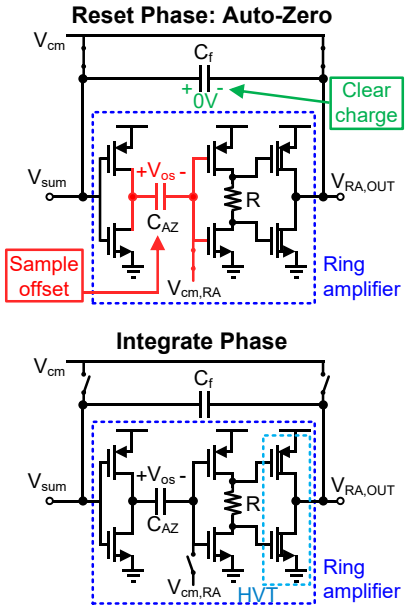


Fig. 9. Operation of the ring amplifier in different phases.

capacitors C_{r1} and C_{r2} attenuate the integrated output by 4x in each cycle to appropriately weight the partial results.

C. Ring Amplifier

The ring amplifier (Fig. 9) is a cascade of three inverters. An offset resistor, R, in the second stage and high-V_{TH} devices in the last stage improve stabilization [7]. The ring amplifier operates with the compute unit in two phases (Reset and Integrate). The Reset phase shorts the input and output of the ring amplifier to V_{cm} and stores the input offset voltage (V_{os}) on an auto-zero capacitor (C_{AZ}). This autozero eliminates the input offset of the ring amplifier and minimizes mismatches between the ring amplifiers in different banks.

Ring amplifiers are smaller and more energy efficient than conventional amplifiers. The dynamic switching provides high-speed slewing and fast settling. Furthermore, the output stage supports a near rail-to-rail swing. These

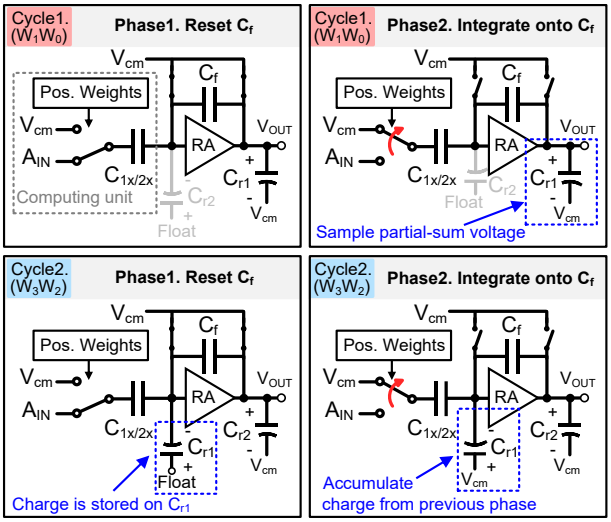


Fig. 10. System workflow for the first two of four cycles (weights W_0-W_3).

benefits make the ring amplifier ideal for analog computing-in-memory.

D. Compute Flow

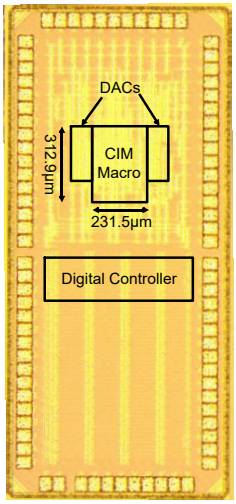
As shown in Fig. 10, four compute cycles implement the signed 9-bit weight multiplication, beginning with the 2 LSBs. The ring amplifier, along with feedback capacitor C_f and residue capacitors C_{r1} and C_{r2}, scale and integrate the charges during the four cycles. Each cycle has two phases. The first phase resets C_f and one of the residue capacitors while the input capacitors C_{1x}, C_{2x} connect to A_{IN} (or V_{cm}, depending on the sign). In the second phase, C_{1x}, C_{2x} connect to V_{cm} (or A_{IN}), and the charge transfers to the feedback capacitor, C_f. Also, one of the residue capacitors, C_{r1} (C_{r2}), samples V_{OUT}. In phase 2 of the next cycle, C_{r1} (C_{r2}) connects to the input of the integrator while C_{r2} (C_{r1}) samples the output. C_{r1} and C_{r2} are one quarter the size of C_f, providing a one fourth scaling each cycle.

IV. Measurement

The prototype is fabricated in 28nm CMOS, and has a core area of 0.082mm^2 including SAR ADCs. Fig. 11 shows a die micrograph. 128 on-chip R-2R DACs emulate analog inputs to facilitate testing. The area and energy efficiencies are 10.9Gb/s/mm^2 and 8pJ/bit, respectively, including ADCs.

Thanks to charge transfer, the measured compute linearity, determined by sweeping the input DAC code, is excellent with an INL_{max} < 1 LSB (Fig. 12(a)). Fig. 12(b) compares the measured and ideal 64x8 MIMO user output for a single MIMO user output over 100 16-QAM MIMO detection cycles, and the measured results closely align with the ideal values. Fig. 12(c) shows the measured BER vs. SNR for an i.i.d. channel. The prototype achieves a 10^{-3} BER with only a 0.36 dB SNR loss compared to a reference 28-bit digital detector with ideal ADCs. Fig. 13 shows the BER at SNR=10dB for multiple chips and the energy breakdown.

Table 1 compares the prototype design with the state-of-the-art digital MIMO detectors. Fig. 11 compares with a 28nm CMOS reference MMSE digital design. Energy efficiency is 4.5x/9.4x better than the reference digital design without/with ADCs.



	Digital Synthesis	This Work
Technology	28nm	**28nm**
MIMO [M x K]	64x8	**64x8**
Modulation	16 QAM	**16 QAM**
Access time (ns)	128	**36**
Power (mW)	8.93	**7.1**
System Throughput (Gb/s)	0.25	**0.89**
Energy Efficiency (pJ/b)	35.7 (no ADCs)	**8 (w/ ADCs)**
Energy Efficiency (TOPS/W)	3.6 (no ADCs)	**16 (w/ ADCs)**
Energy Efficiency (pJ/b) (ADCs included)[1]	75	**8**
Energy Efficiency (TOPS/W) (ADCs included)[1]	1.7	**16**

[1]The energy consumption of the 8-b input ADCs is estimated by Schreier FoM=160dB with Fs=30MHz.

Fig. 11. (a) Die micrograph of prototype and (b) comparison with a reference digital MMSE MIMO detector in the same 28nm technolgy and using same modulation.

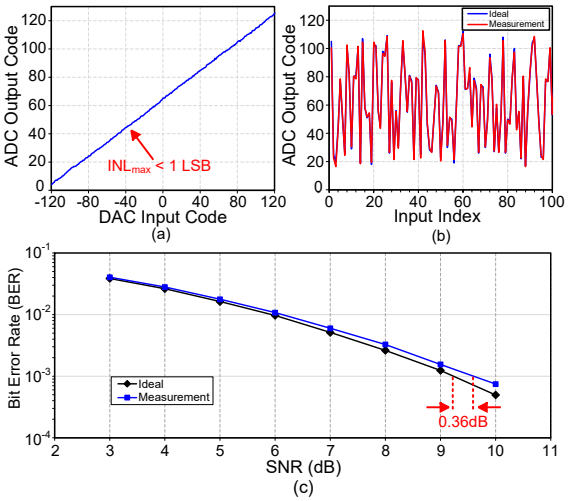


Fig. 12. (a) Measured linearity, (b) sample ADC output, and (c) measured and ideal BER for i.i.d. channel versus SNR shows small SNR loss.

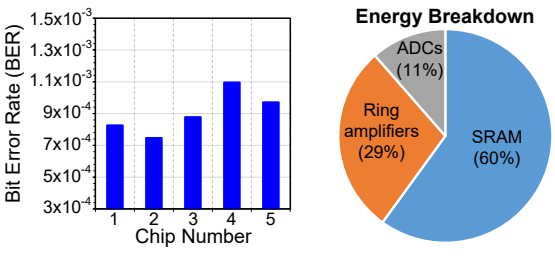


Fig. 13. Measured BER for 5 die and SNR=10dB and energy breakdown.

V. Conclusion

This work introduces a new charge-transfer CIM architecture, which is especially suited for MIMO processing. In addition to reducing the MAC energy, the CIM processor replaces 128 high-resolution front-end ADCs with 16 7-bit SAR ADCs. Charge-transfer CIM provides the linearity needed for MIMO. Instead of a conventional CIM gain < 1, the high gain greatly relaxes the specifications of the output SAR ADCs. Combining dual-bit operation with the pipelining of MAC and ADC increases MIMO detection throughput by approximately 4x. We introduce ring amplifiers in CIM for high energy efficiency, high speed, and high output swing. The auto-zero technique also helps minimize mismatch, resulting in a more compact area.

The fabricated 28nm CMOS prototype achieves area and energy efficiencies of 10.9Gb/s/mm^2 and 8pJ per detected bit, representing a 3x and 2.8x improvement, respectively, over state-of-the-art digital MIMO detectors.

TABLE I. Comparison with State-of-the-Art Digital MIMO Detector

	ISSCC'17 [4]	ISSCC'18 [5]	ISSCC'24 [6]	This Work
Technology	28nm	28nm	40nm	**28nm**
Architecture	Digital	Digital	Digital	**Analog CIM**
Core area (mm^2)	N.A.	2	4.36	**0.082**
MIMO [M x K]	128x8	128x16	8x8	**64x8**
Modulation	256 QAM	4~256 QAM	16 QAM	**16 QAM**
Input ADCs Included?	No	No	No	**Not required**
Output ADCs Included?	No	No	No	**Yes**
Throughput (Gb/s)	0.3	1.8	9.8	**0.89**
Energy Efficiency (pJ/b)[a]	60	17.5	47.5	**8**
Area Efficiency (Gb/s/mm^2)[b]	N.A.	3.6	2.25	**10.9**
Energy Efficiency (pJ/b) (ADC included)[c]	69.8	22.4	49.9	**8**

[a] Energy Efficiency is scaled by (8/K)^2.
[b] Area efficiency is scaled by (K/8)^2.
[c] The energy consumption of the 8-bit input ADCs is estimated using a Schreier FoM=160dB and Fs=30MHz.

References

[1] S. -E. Hsieh et al., "7.6 A 70.85-86.27TOPS/W PVT-Insensitive 8b Word-Wise ACIM with Post-Processing Relaxation," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 136-138.

[2] H. Wang, R. Liu, R. Dorrance, D. Dasalukunte, N. Gowda and B. Carlton, "A PVT Robust 8-Bit Signed Analog Compute-In-Memory Accelerator with Integrated Activation Functions for AI Applications," 2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2024, pp. 1-2.

[3] X. Chen et al., "14.6 A 28nm 64kb Bit-Rotated Hybrid-CIM Macro with an Embedded Sign-Bit-rocessing Array and a Multi-Bit-Fusion Dual-Granularity Cooperative Quantizer," 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 260-262.

[4] H. Prabhu, J. N. Rodrigues, L. Liu and O. Edfors, "3.6 A 60pJ/b 300Mb/s 128×8 Massive MIMO precoder-detector in 28nm FD-SOI," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2017, pp. 60-61.

[5] W. Tang, H. Prabhu, L. Liu, V. Öwall and Z. Zhang, "A 1.8Gb/s 70.6pJ/b 128×16 link-adaptive near-optimal massive MIMO detector in 28nm UTBB-FDSOI," 2018 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2018, pp. 224-226.

[6] Y. Zhang et al., "2.7 BayesBB: A 9.6Gbps 1.61ms Configurable All-MessagePassing Baseband-Accelerator for B5G/6G Cell-Free Massive-MIMO in 40nm CMOS," 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 48-50.

[7] Y. Lim and M. P. Flynn, "A 100 MS/s, 10.5 Bit, 2.46 mW Comparator-Less Pipeline ADC Using Self-Biased Ring Amplifiers," in IEEE Journal of Solid-State Circuits, vol. 50, no. 10, pp. 2331-2341, Oct. 2015.