Master's Thesis Engineering Technology

2024-2025

Hybrid IGZO-Si 3T0C DRAM for Energy-Efficient Near-Memory MAC and Bitwise computing

Mohammed Murad Khalil Albayyouk

Technology Master of Electronics and ICT Engineering

Background and Motivation

The ever-widening "memory wall"—the disparity between processor speeds and DRAM access times (see Figure 1)—now dominates both performance and energy in data-intensive workloads. In-Memory Computing (IMC) directly embeds logic into the memory array, sidestepping costly data transfers. Traditional CMOS-based DRAM achieves high speed but suffers from substantial refresh overhead, whereas emerging oxide TFTs (e.g., IGZO) offer ultra-low leakage at the expense of mobility (see Figure 2).

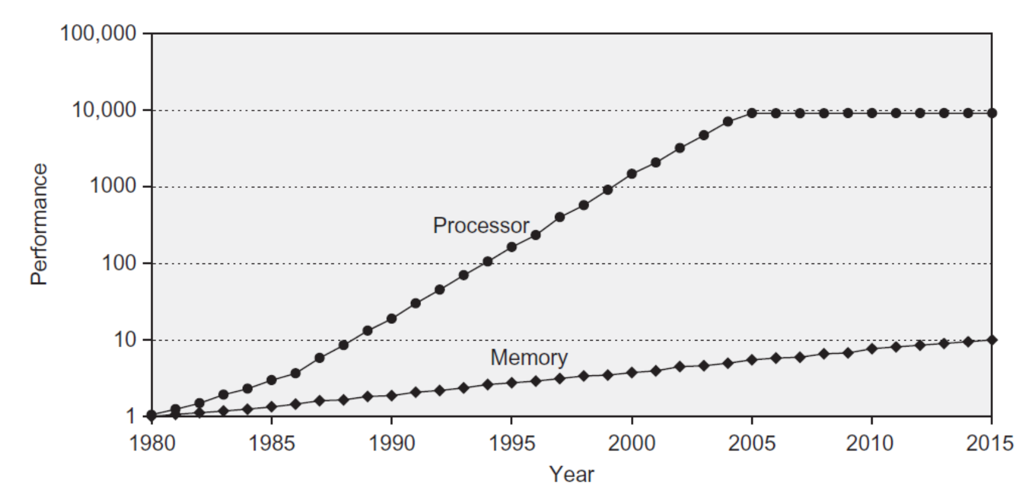


Figure 1: Processor vs. memory relative performance over the years [1, p. 80].

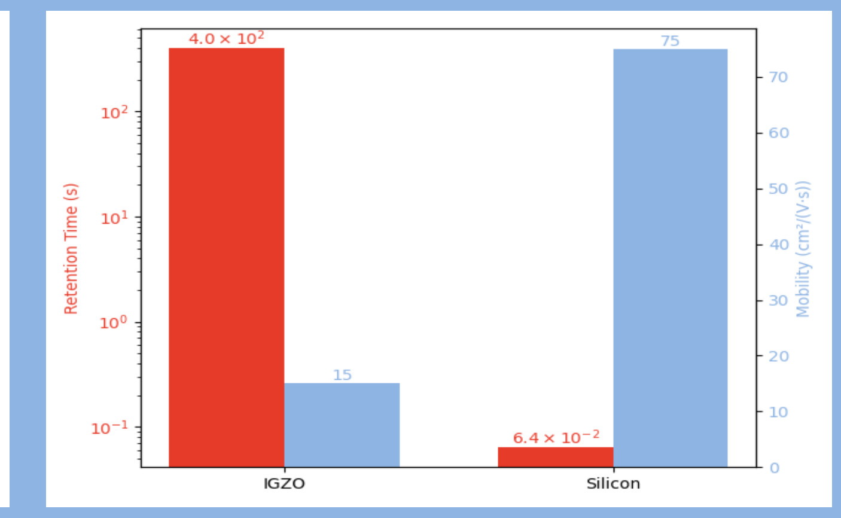


Figure 2: IGZO vs. Si: Mobility & Retention

To reconcile these trade-offs, a hybrid IGZO–Si 3-transistor/0-capacitor (3T0C) cell is proposed, marrying IGZO's retention with silicon's high-speed switching. This work quantifies the potential of such a hybrid for IMC.

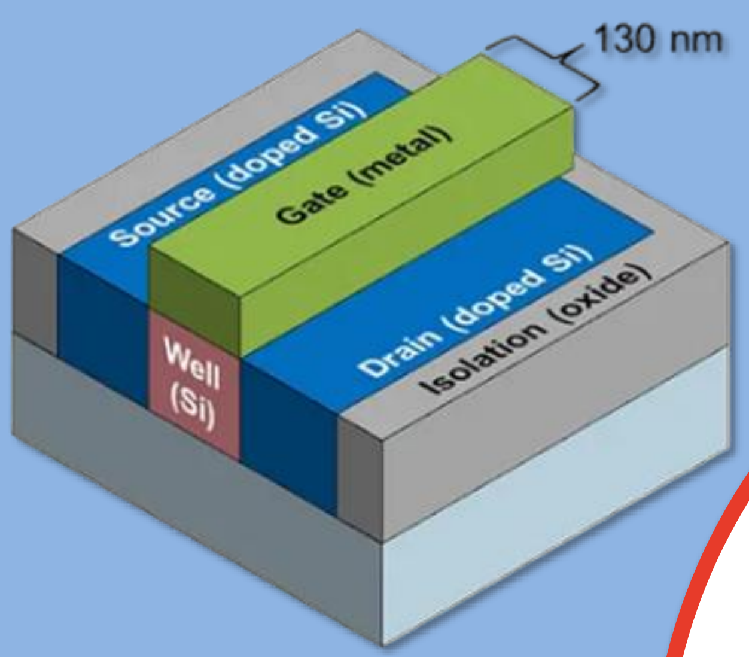
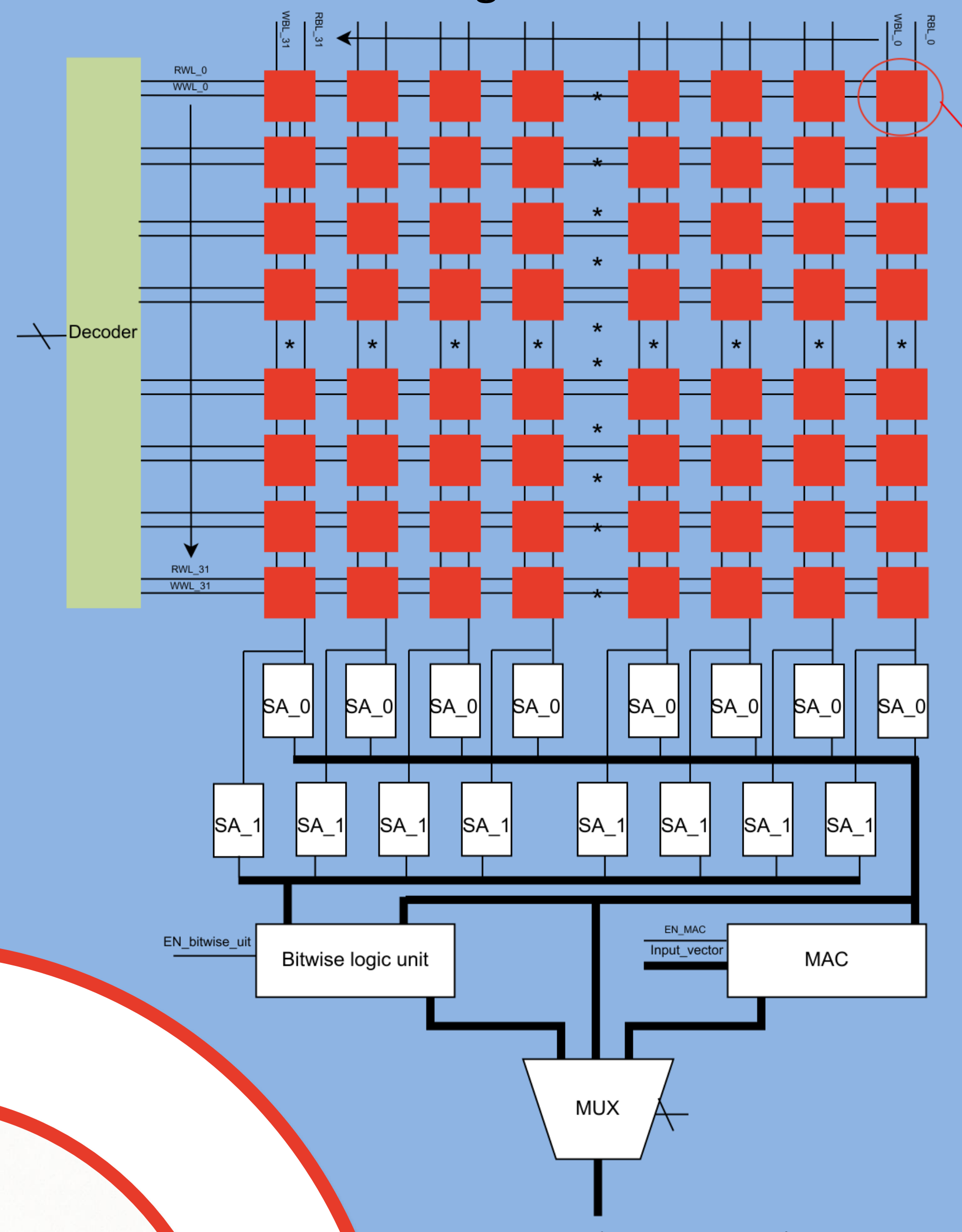


Figure 3: 3D image of a transistor [2].

3T0C Cell and Array Architecture

Each 3T0C cell integrates two IGZO transistors (M1, M2) for low-leakage storage/access and a silicon cascode transistor (M3) for read-path isolation and for fast discharge. IGZO's off-current preserves data; silicon's mobility accelerates sensing.



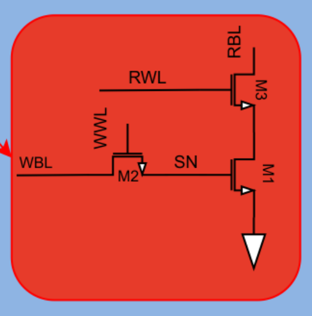


Figure 4: Proposed 32x32 array architecture.

Array and Sensing: A 32×32 DRAM array (Figure 4) uses separate read/write word and bit lines feeding two sense-amplifier banks for fast row activation.

Logic Units: The bitwise unit loads two rows into sense amplifiers, applies a 32-bit Boolean operation in place. The MAC unit reads one row as eight 4-bit weights, multiplies by an input vector, and accumulates the dot product.

IMC

Results, Conclusion and Outlook

Hybrid IGZO–Si 3T0C DRAM achieves over 400 s retention, ~6 450× silicon-only, while delivering 50 ns/55 ns read/write at 116 pJ/131 pJ (table 1). In-memory bitwise (85 ns, 232 pJ) and MAC (55 ns, 144 pJ) operations enable low-energy, high-throughput compute near memory efficiently.

Table 1: Comparison to the state of Art.

	Giterman et al.	Ryu et al.	This Work
Write Access Time	1.3 ns	9.1 µs	55 ns
Write Energy	N/A	N/A	131 pJ
Read Access Time	25 ns	9.1 µs	50 ns
Read Energy	N/A	N/A	116 pJ
Retention	< 0.8 ms	> 1000 s	> 400 s

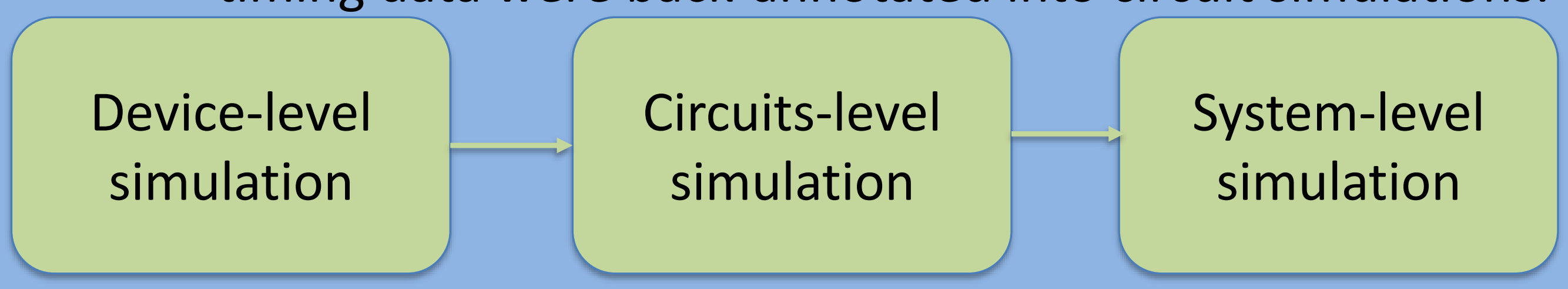
Outlook

Scaling to larger arrays requires analyzing interconnect parasitics, IR drop, and exploring monolithic 3D stacking of IGZO on silicon logic to reduce routing overhead. Advanced sense amplifiers with adaptive biasing and column-parallel activation may lower read latency below 50 ns. Per-bit logic cells or analog compute primitives could enable single-cycle, array-wide operations. End-to-end benchmarks on sparse and bitwise neural workloads and compiler-driven ISA extensions will quantify performance and energy benefits.

Simulation Methodology

Device-level:

All simulations were performed in Cadence. Device-level SPICE models for IGZO transistors were provided by Pragmatic; the silicon cascode model by X-FAB. The logic units and the controller were implemented in VHDL, simulated in Vivado, synthesized with the Genus tool using X-FAB standard cells, and their functional and timing data were back-annotated into circuit simulations.



Device-level simulation → Circuits-level simulation → System-level simulation

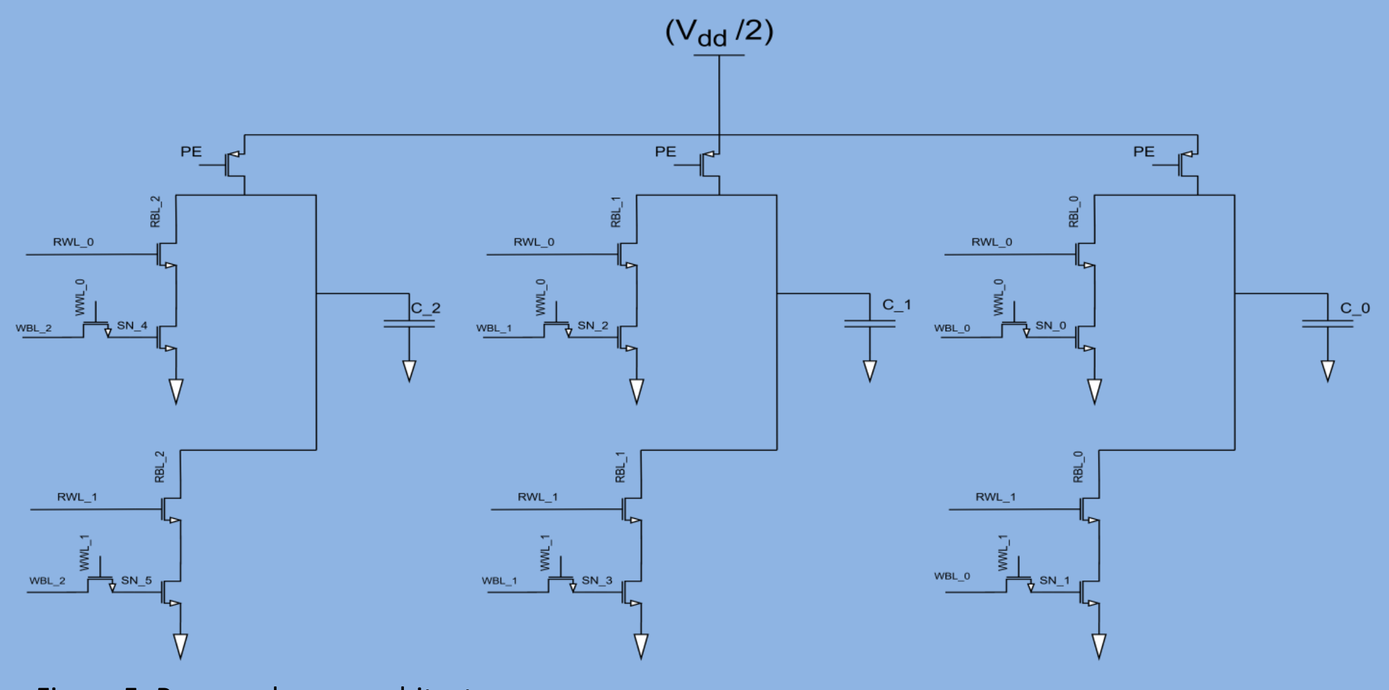


Figure 5: Proposed array architecture.

Circuit- and system-level:

Monte Carlo SPICE on a 2×3 hybrid-cell array under process variation determines read latency, sense-amplifier offset, and retention. System-level behavioral modeling of the full 32×32 array (with controller, sense amp, MAC, and bitwise block) estimates refresh energy and IMC throughput.

Supervisors / Co-supervisors / Advisors: Prof. Dr. Ing. Kris Myny
Ing. Djihad Narcereddine
Ing. Yari Nowicki
Ing. Sven Baerten
Ing. Jelle Biesmans.

[1] J. L. Hennessy, D. A. Patterson, and A. C. Arpaci-Dusseau, *Computer Architecture: A Quantitative Approach*. San Francisco, CA: Morgan Kaufmann, 6 ed., 2019.
[2] M. Traverso, "A Node by Any Other Name: Transistor Size & Moore's Law,". [online]. Available: https://medium.com/predict/a-node-by-any-other-name-transistor-size-moores-law-b770a16242e5. [accessed May 26, 2024].

De opleiding industrieel ingenieur is een gezamenlijke
opleiding van UHasselt en KU Leuven