EES2 Circuits & Architectures WG

Energy Efficient CMOS Memories

Azeez Bhavnagarwala
Metis Microsystems

April 27th 2023

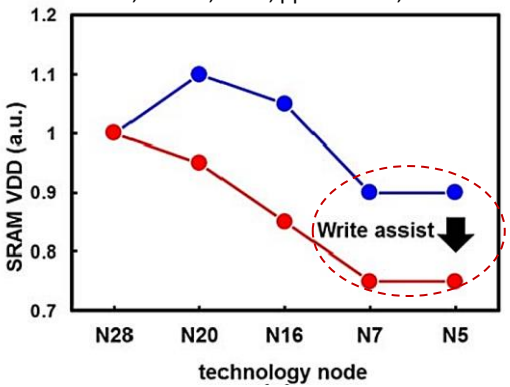
CMOS Memories

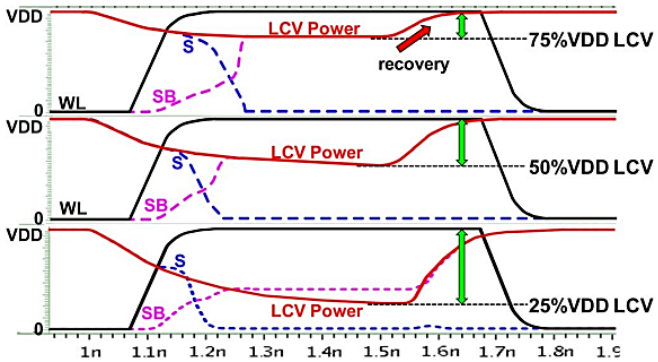
- Built on a *standard CMOS Logic process* – do not require additional masks, process integration steps
- Use *CMOS Logic supply voltages* & comparable device threshold voltages in bit cell transistors
- Delivers *highest access/cycle time performance (<200ps), lowest operating voltages (<0.5V)* of any semiconductor memory technology - highest energy efficiency across all semiconductor memory technologies
- Considered '*Foundational IP*' – arrays present in practically every chip manufactured
- *Variants*: 6T SRAM, 2-port 8T Register File, Dual-port 8T SRAM, CIM, TCAM arrays
- Industry-typical peripheral circuit architectures used around CMOS memories are *inefficient* and *irrelevant* to addressing limitations seen today
- Energy-efficiency scaling *requires breakthroughs in CMOS memory circuits* – voltage scaling knob increasingly difficult and inadequate

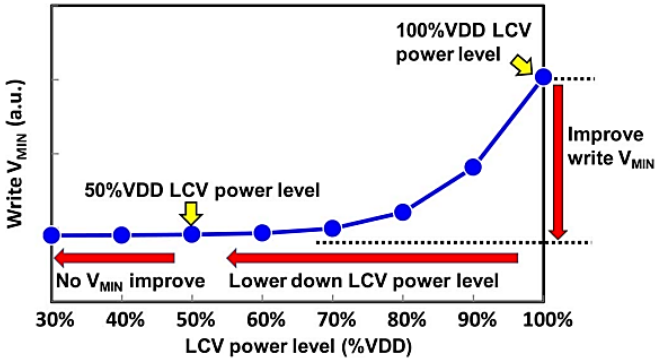
Energy Inefficiencies - Write

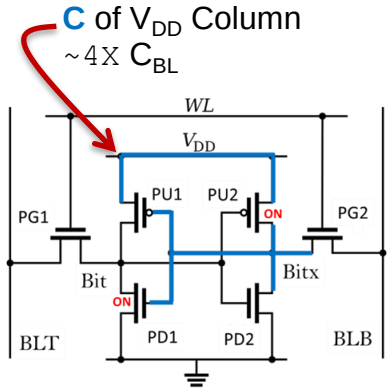
- Voltage scaling increasingly difficult – even with Assist Circuit techniques, given constraints imposed by performance, yield & leakage
- Write Assist schemes LCV, NBL add energy, area and performance overheads that *limit the energy efficiency improvements* from Voltage scaling
- Restoring VDD column to VDD following a 25% drop during WA consumes as much energy as Writing to a BL - need to scale VDD by 30% just to break even

T-Y Chang et al, IEEE Journal of Solid-State Circuits, vol. 56, no. 1, pp. 179-187, Jan. 2021





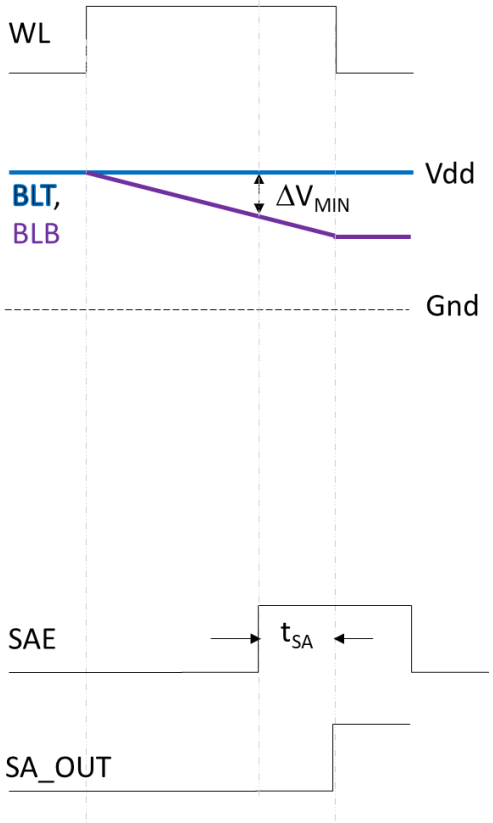


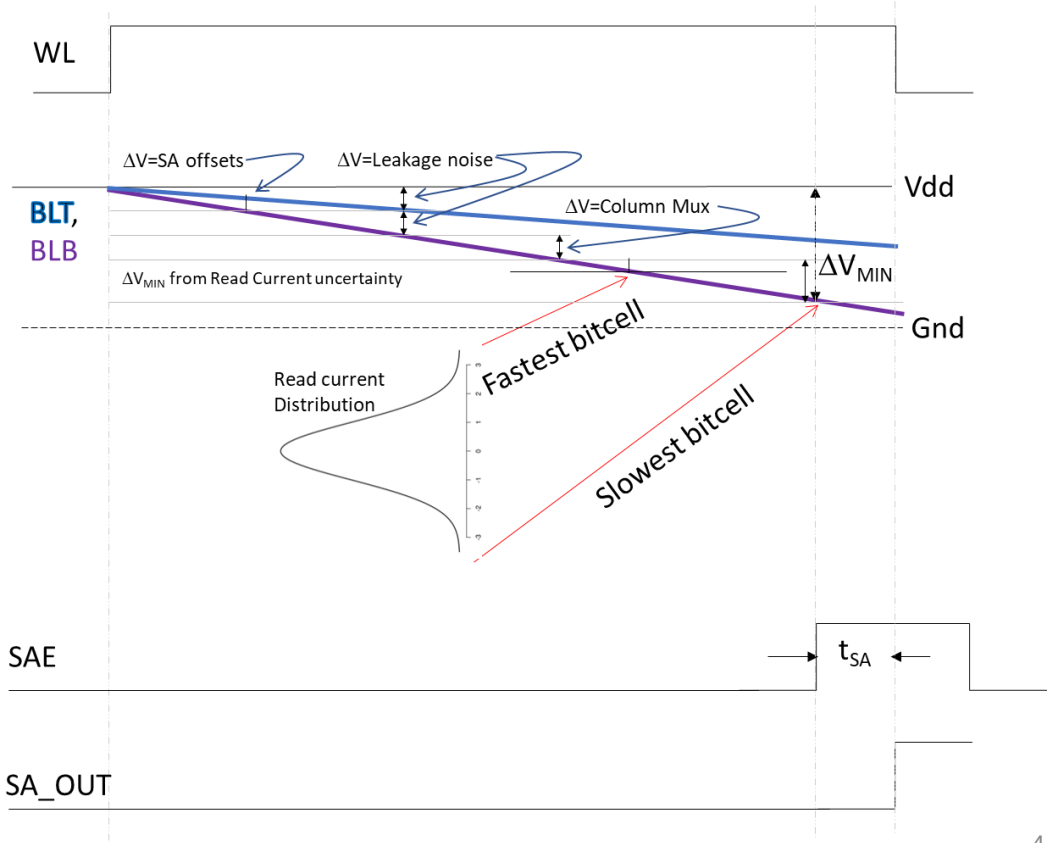


Y-H Chen et al, IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 170-177, Jan. 2015

Capacitance of net driven when restoring VDD to the Supply terminal during Write Assist

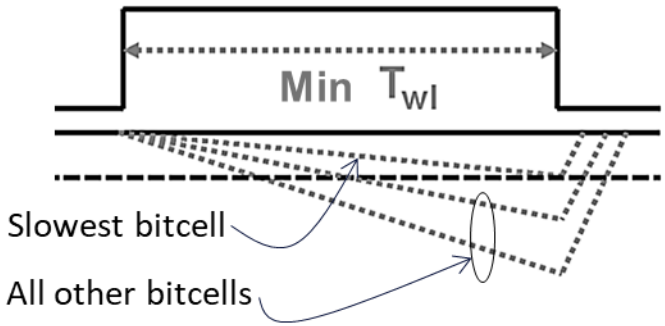
Speed Degradation with Differential Sensing

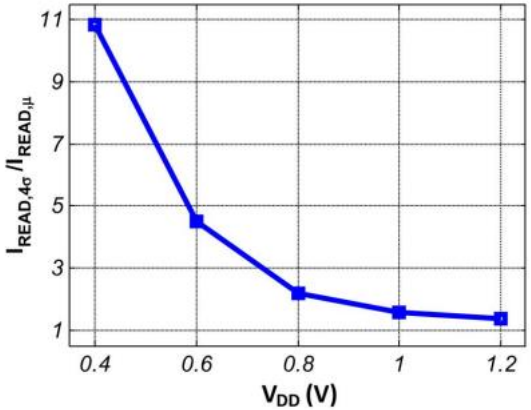




Energy Inefficiencies from Differential Sensing

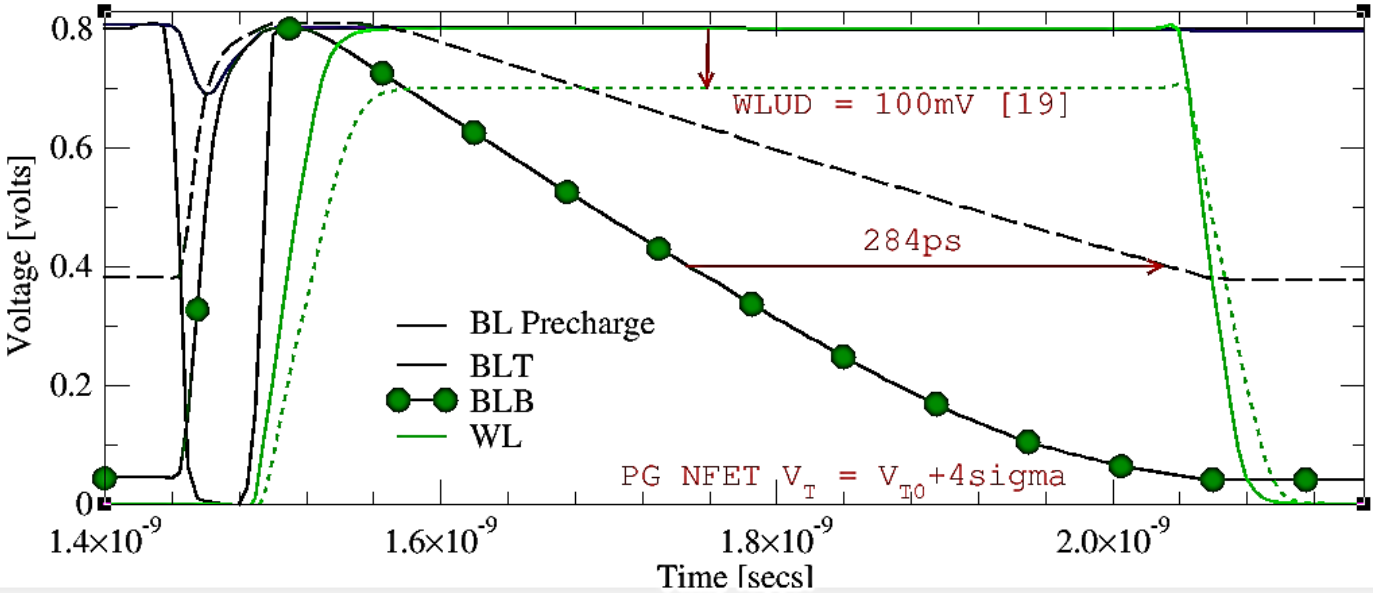
- Variability along bitpath: I_{READ}, *BL* I_{Leak}, *SA offsets, timing uncertainty b/w WL and SAE edges* - limit min BL signal development time in slowest bit cell
- Min BL signal development time in slowest bit cell is sufficient to drain most of precharge from other active BLs - substantially increasing energy overhead of variability in bitpath
- Neither of the expected benefits of differential sensing – fast action or small voltage swing to resolve data are realized





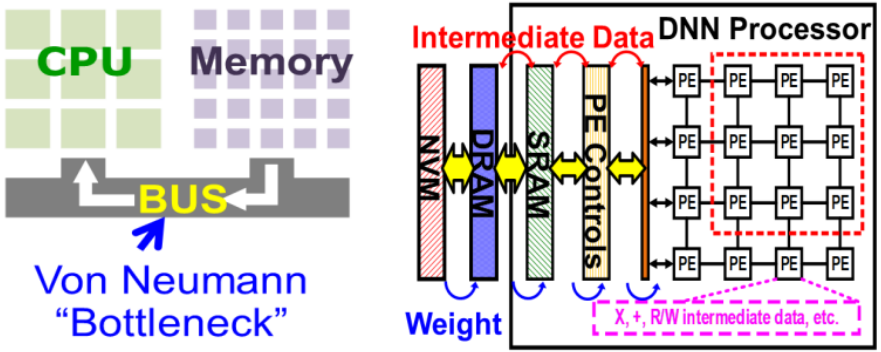
Read Assist degrades energy efficiency further

- WL Under Drive makes the slowest bit cell much slower
- Gives more time to all other bit cells to discharge more of the Precharge
- WLUD for half-select cells during Write, degrades WM improvements from WA
- WLUD ΔV of -100mV on WL increases WL →BL delay by over 2X w PG NFET $V_T = V_{TO} + 4\sigma$



Where is the Energy going?

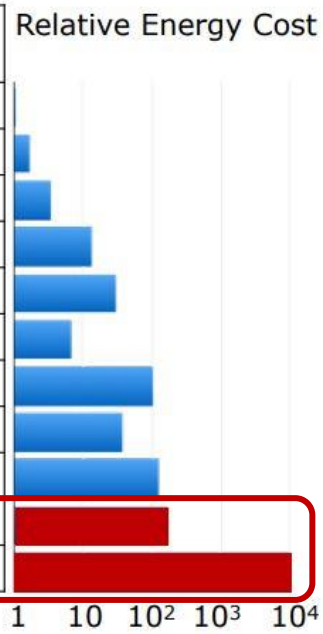
- *Energy & latency costs of data movement* across and within Memory layers



M-F Chang, Tutorial VLSI Ckts Symposium, 2022

- *Inefficiencies* from using *conventional peripheral circuit architectures* in CMOS Memories

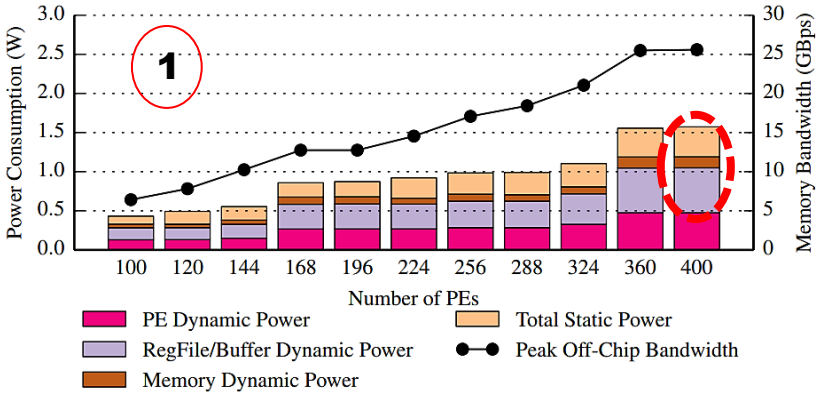
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

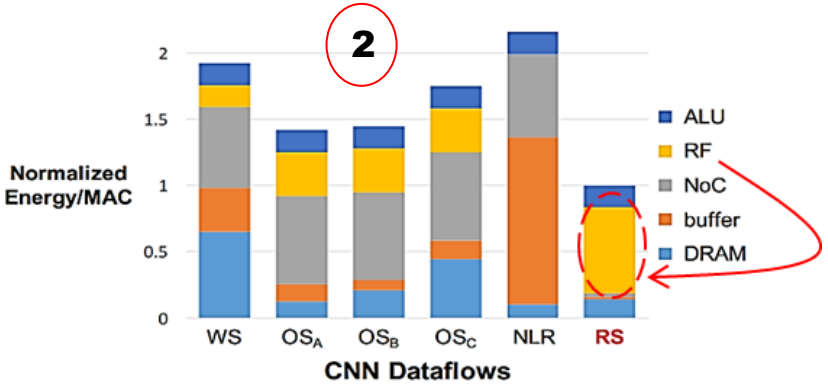


M Horowitz, "Computing's Energy Problem (and what we can do about it)" Keynote at 2014 ISSCC, Feb 2014

CMOS Memories in Accelerators, GPUs

1. Over 2/3 of ASIC accelerator Energy (switching & leakage) consumed by SRAM buffers and Register File (RF) arrays
2. Almost 70% of MAC energy in a GPU consumed by RF SRAM arrays





M Gao et al, "TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory", ASPLOS 2017, pg 751, April 2017

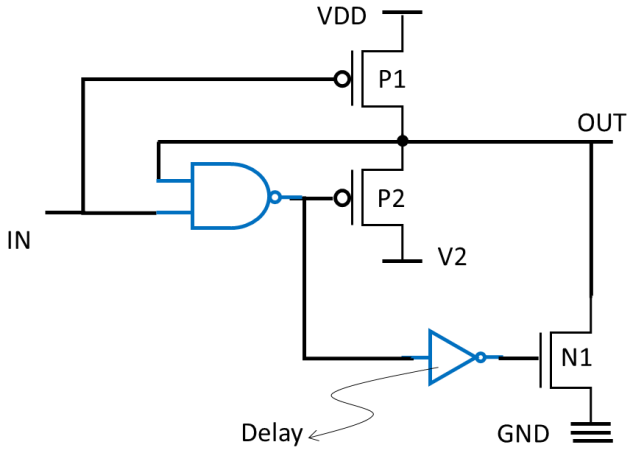
V Sze et al, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey" Proceedings of the IEEE, Vol 105, No. 12, Dec 2017

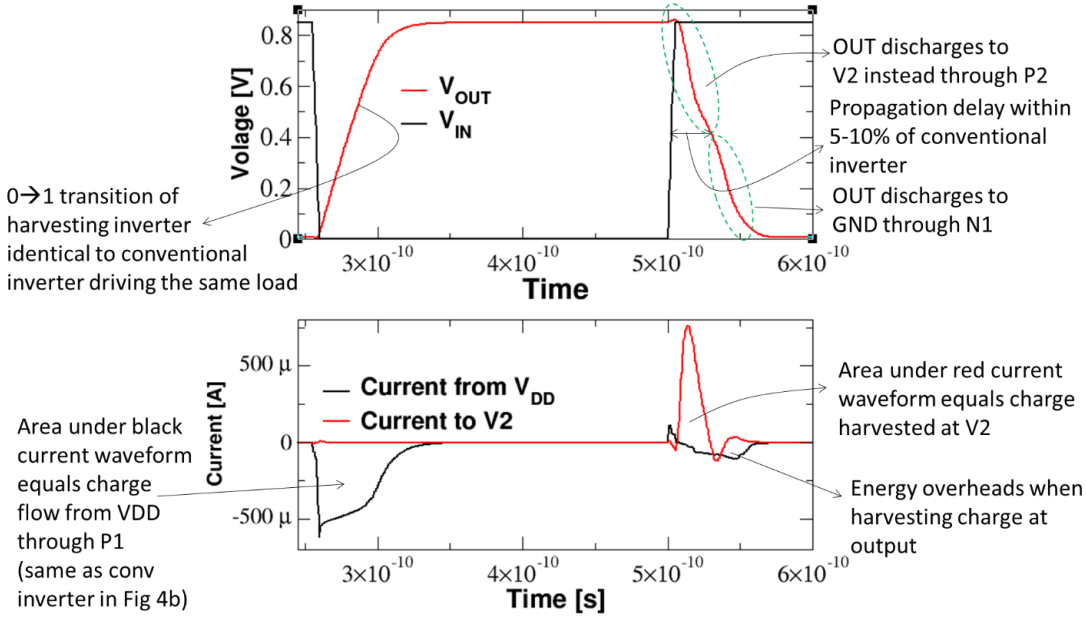
Memory Circuit Solutions Proposed

- Circuit solutions to *holistically improve all metrics* – circuit speed, energy dissipation, R/W margins, reliability (without/with marginal) area overheads and substantial reductions in leakage – using new circuits that
  - *Harvest the information token of data* on chip: electric charge when overwriting, moving or storing data
  - *Self-limiting, self-disabling & self-regulating circuits* to eliminate most of the inefficiencies seen with industry-typical circuit architectures
  - 5X reduction in active energy without lowering VDD, 2X improvement in access/cycle time without raising VDD
  - Above improvements at component level without requiring changes to the bit cell or the CMOS process

Simple circuits to harvest/use harvested charge

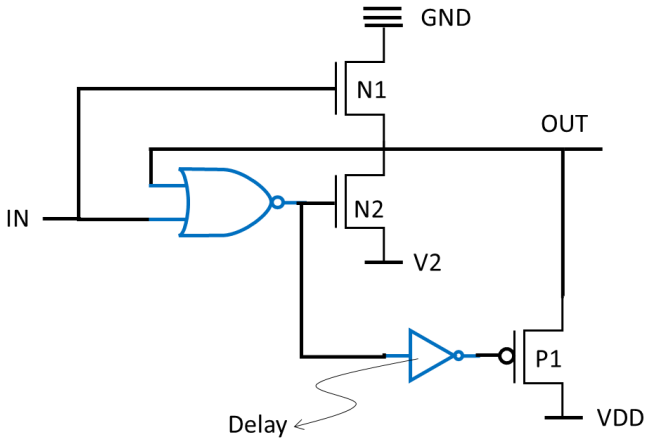
- Full-swing, no reduction in propagation delay, area overheads, sub CV^2 switching energy!
- High fan-out circuits, large loads – recover 30+ % of energy

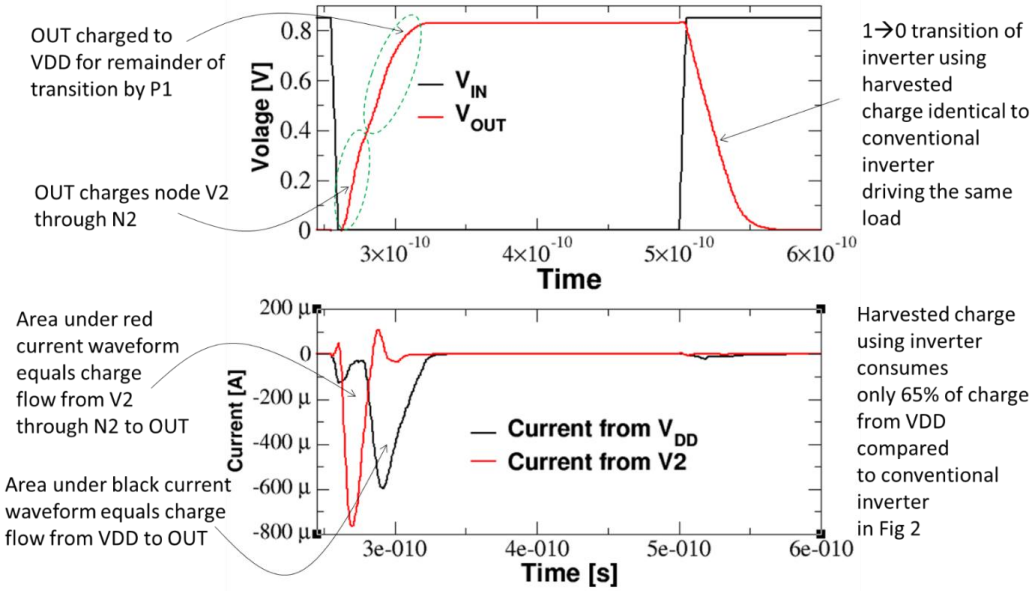




Simple circuits to harvest/use harvested charge

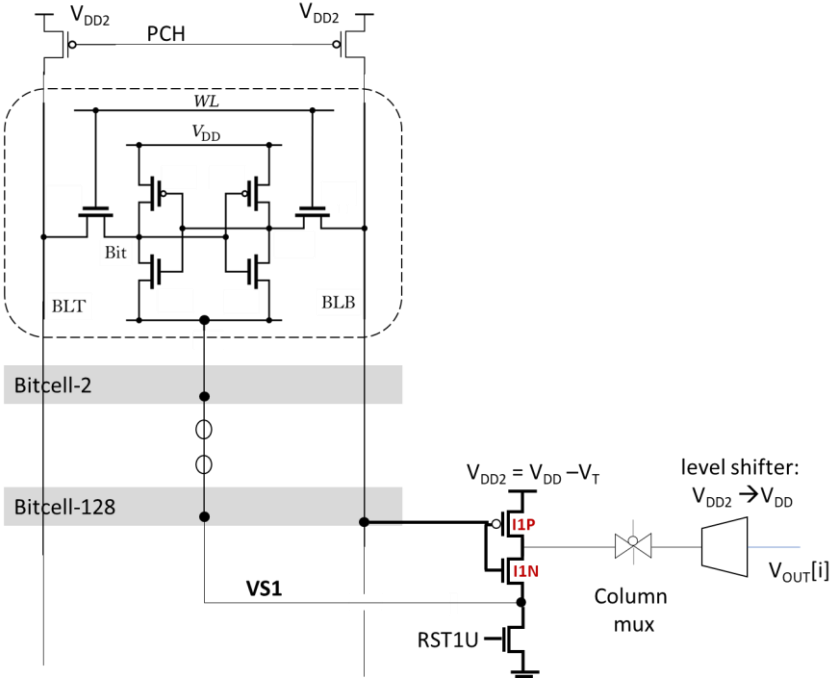
- Logic dual of harvesting circuits – use harvested charge to partially drive 0→1 transitions

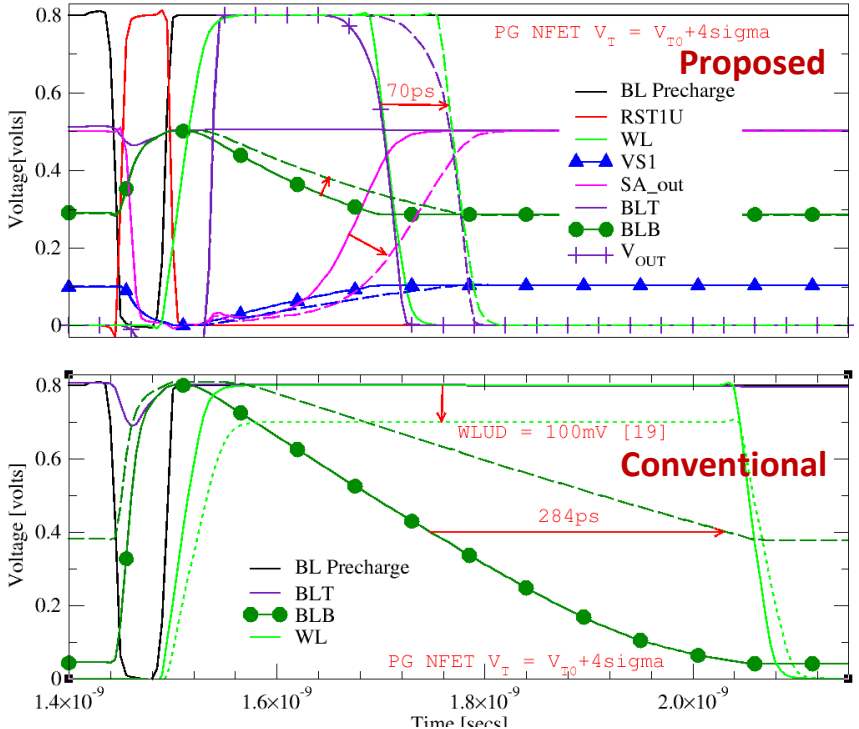




CMOS 6T SRAM Arrays

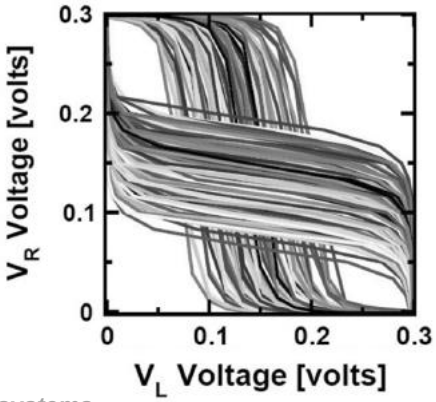
- Variability tolerant, higher cell stability, 2X+ faster access times, 4-5X lower Read power, 30+ % lower Write power

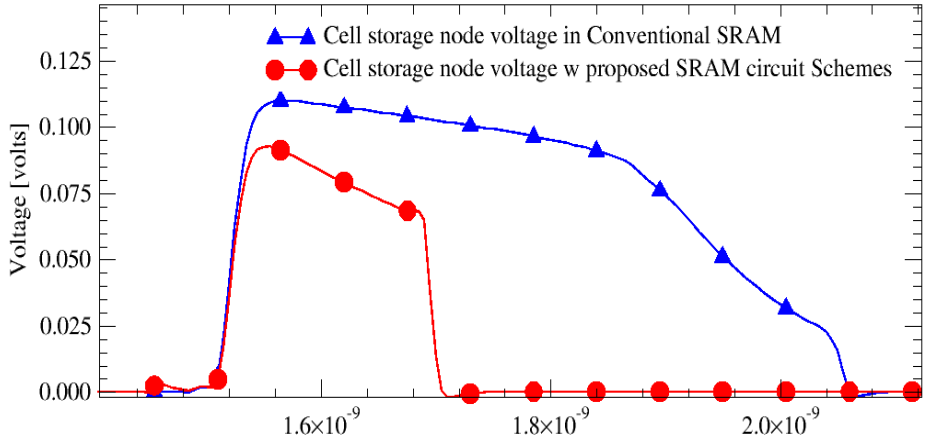


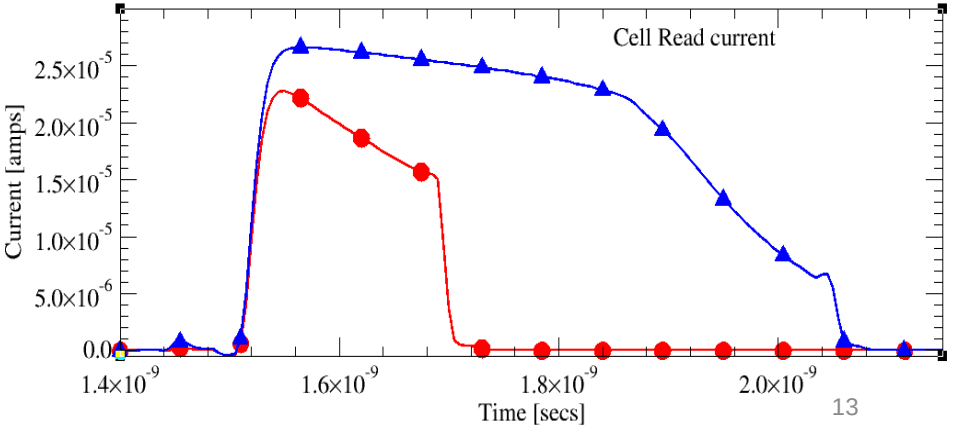


CMOS SRAM Cell Stability

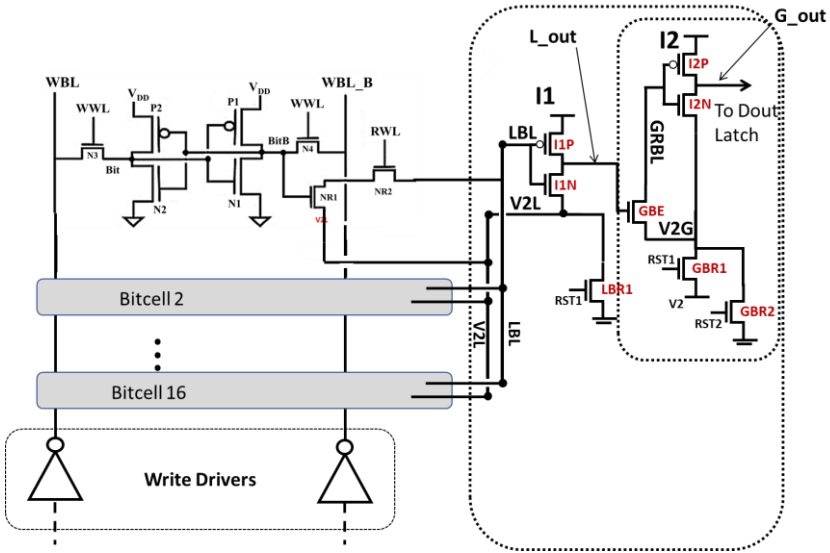
- With less Read current, proposed sensing scheme is 2X faster, 4-5X lower active power from self-limiting, self-disabling discharge of BLs
- Self-limiting & self-disabling action increases read stability during WL active period with retention stability when Read current disabled with WL active

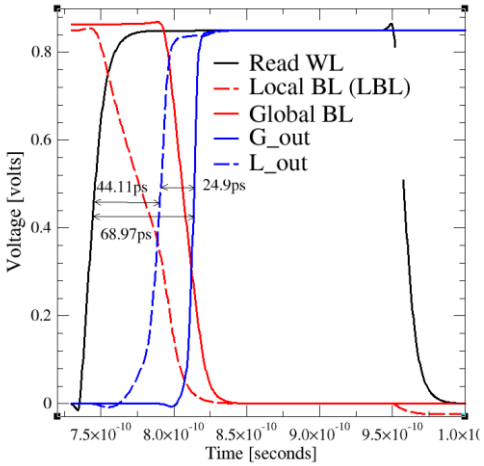


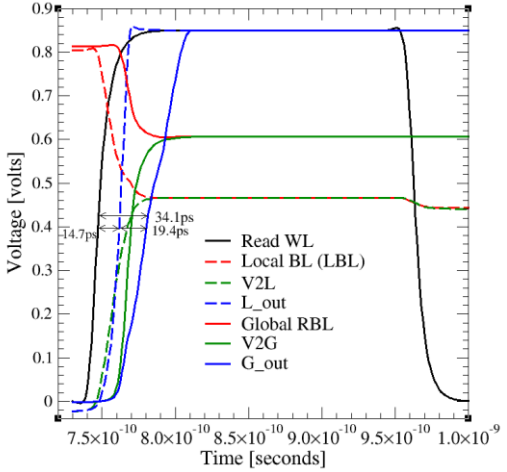




CMOS 8T Register File Arrays



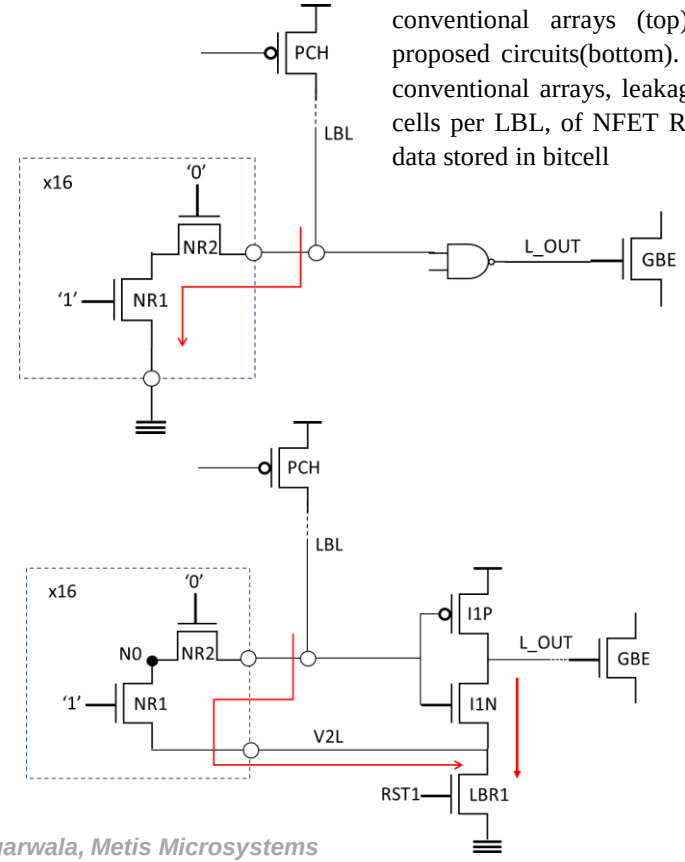


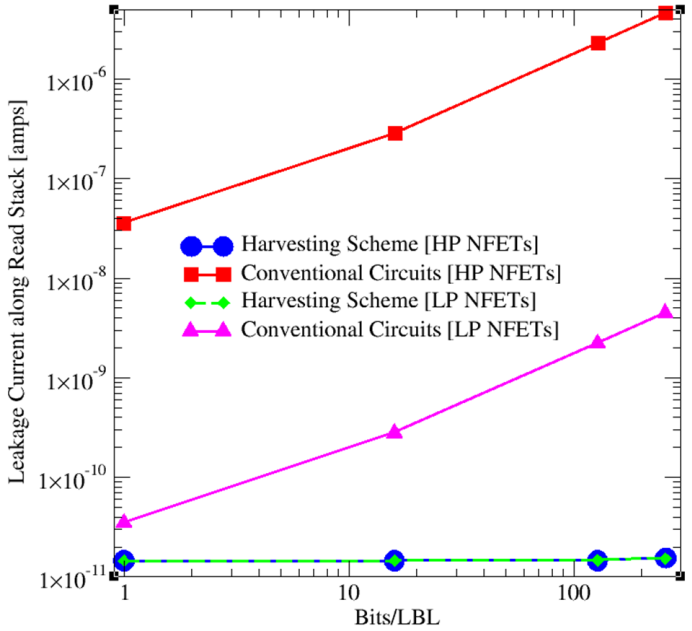


Comparison of RF Array Metrics	WL → Data_out		Read Bitpath Energy		Write Bitpath Energy	
	Delay	% improvement	Energy	% improvement	Energy	% improvement
RF Array with Conventional Circuits	68.97 ps	-	19.1 fJ		13.63 fJ	
RF Array with Proposed Circuits	43.90 ps	36.3%	3.74 fJ	80.4%	9.69 fJ	28.9%
RF Array with Proposed Circuits using LVT NFETs in Read Stack of Cell	32.5 ps	52.8%	3.74 fJ	80.4%	9.69 fJ	28.9%

CMOS 8T Register File Arrays

Leakage paths along decoupled Read stack in conventional arrays (top) and in the arrays with proposed circuits(bottom). In proposed scheme, unlike conventional arrays, leakage is independent of # of bit cells per LBL, of NFET Read Stack device VT and of data stored in bitcell

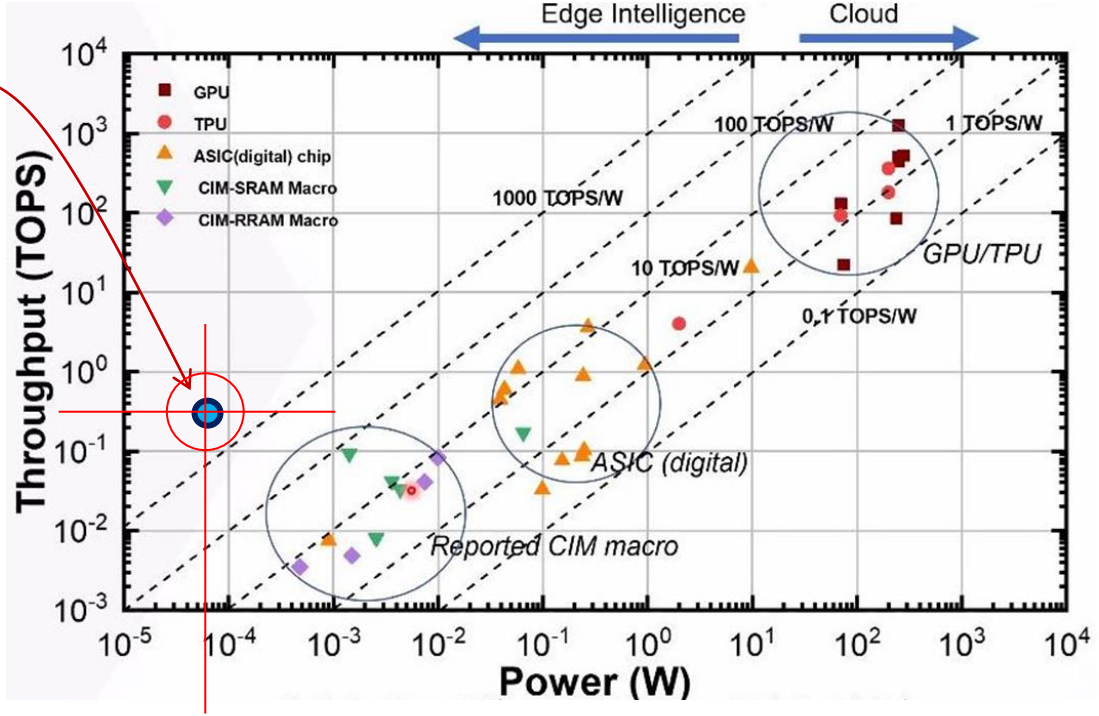




Comparisons with Reported CIM macros & AI Hardware

Metis 8T SRAM based Digital CIM array

CIM Arrays using 8T SRAM	2020 ISSCC	This work (Simulations)
CMOS Technology	7nm	16nm
Cycle Time [ns]	4.5	0.20
CIM	Analog	Digital
Voltage (V)	0.8	0.8
Array Size	4Kb	
Bit Precision	4b/4b	4b/4b
Peak MAC Throughput (GOPS)	372	160
Peak MAC Energy Efficiency (TOPS/W)	611	>5000
MAC Comp. Density (TOPS/mm^2)	116	>116



Rest of figure from: Memory Centric Computer Workshop at the 2021 DARPA ERI Summit, Oct 19th 2021

Summary

- New Circuit architectures enabling fast, energy efficient data movement in CMOS memories are proposed
- Proposed circuits with self-limiting, self-disabling and self-regulating behavior also lower substantial overheads in latency and energy consumption seen in conventional/Industry-typical circuit architectures for CMOS Memories
- Simulations with 16FF parameter decks demonstrate 5X reduction in active energy and 2X improvements in latency across all CMOS memories using proposed circuits – an order of magnitude improvement in the Energy Delay Product
- Impact to energy and performance bottlenecks of accelerators for AI workloads are significant – 70% of the active power dissipations from a MAC operation impacted with above improvements
- With no changes required to the bit cell, CMOS process or (nominal) operating voltages, demonstration of proposed circuits in hardware prototypes is feasible within 9-12 months