

CMOS+X: Stacking Persistent Embedded Memories based on Oxide Transistors upon GPGPU Platforms

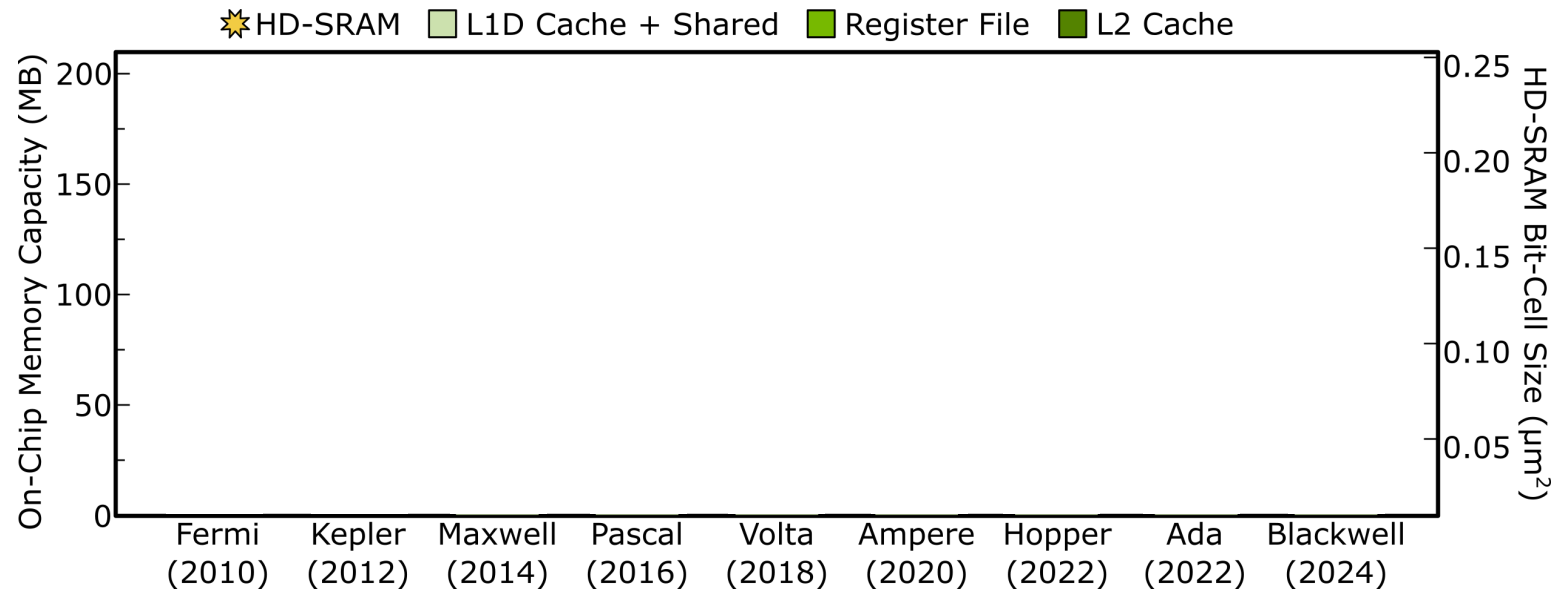
Faaq Waqar, Ming-Yen Lee, Seongwon Yoon,
Seongkwang Lim, Shimeng Yu

School of Electrical & Computer Engineering, Georgia Institute of Technology



Laboratory for Emerging Devices and Circuits

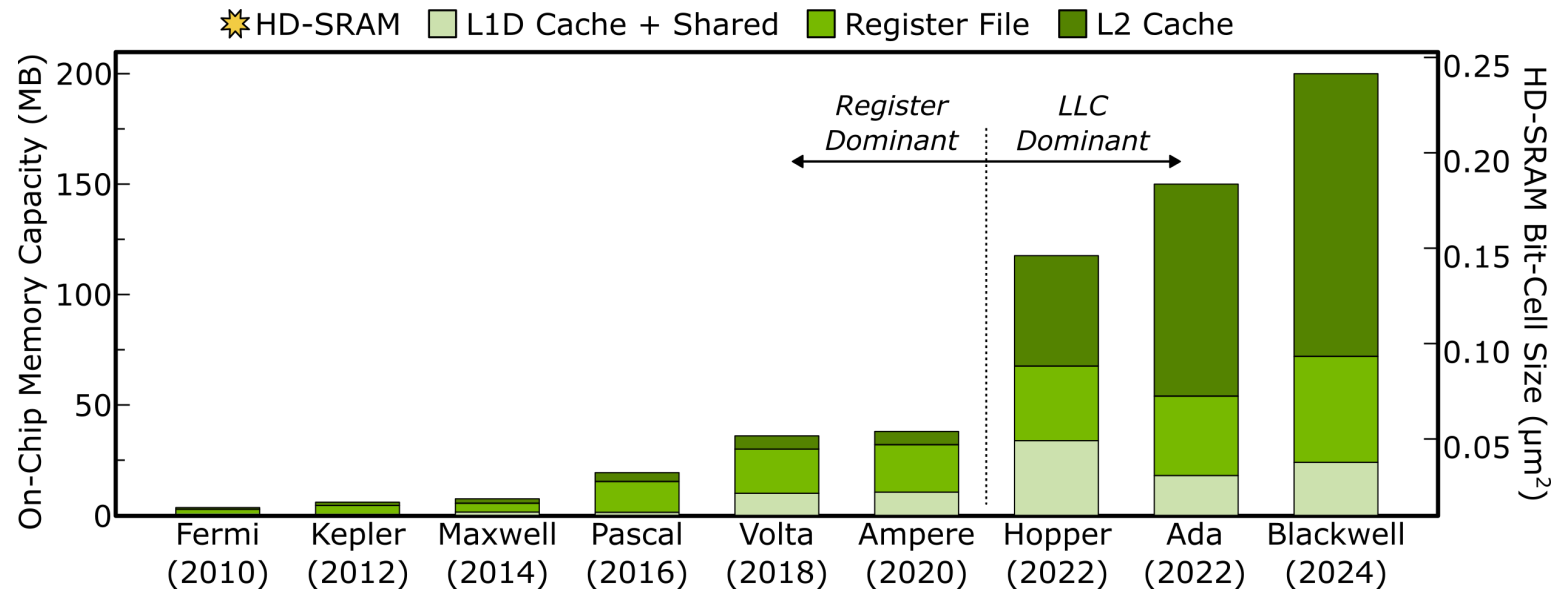
On SRAM Densification and GPU Memories



NVIDIA **GPGPU On-Chip Total Memory Capacity** has grown **>100×** in 15 years



On SRAM Densification and GPU Memories

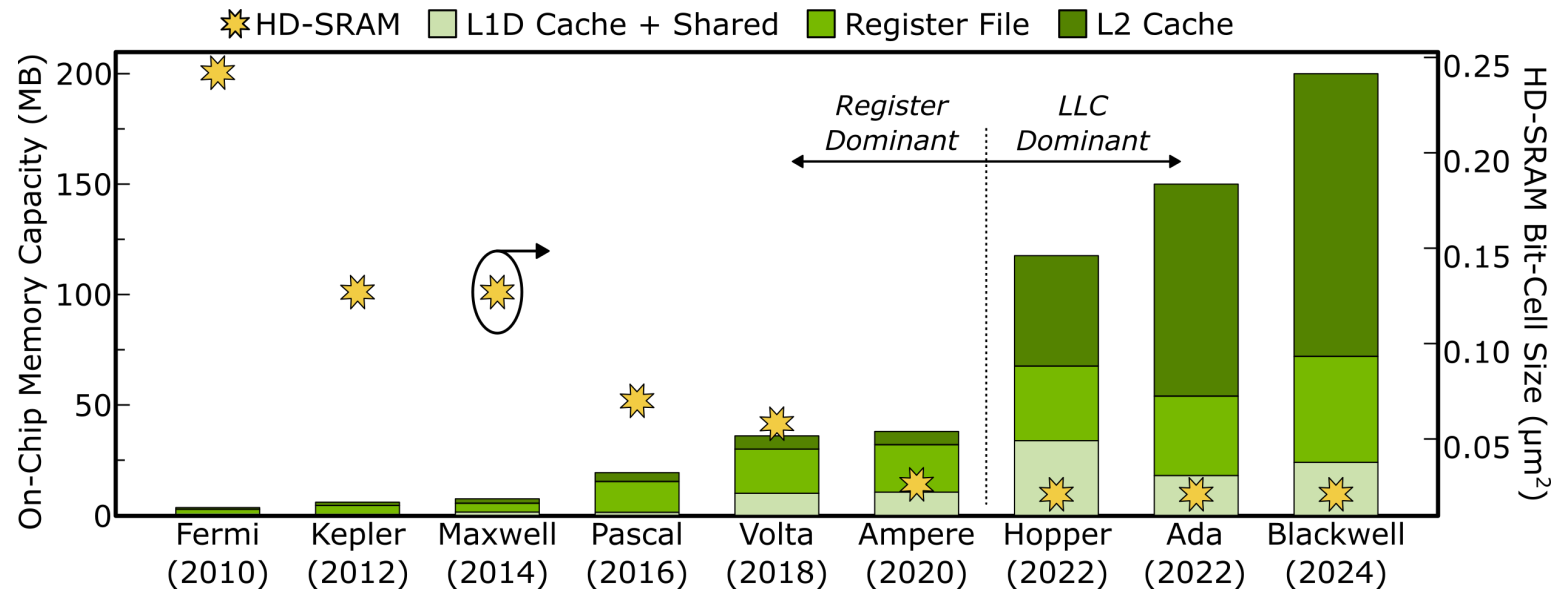


NVIDIA GPGPU Total Memory Capacity has grown **>100×** in 15 years

NVIDIA GPU Memory Scaling in 2 Eras:

- (1) **Register Dominated**, Compute Driven (**Pre-2022**)
- (2) **Last Level Cache Dominated**, Data Driven (**Post-2022**)

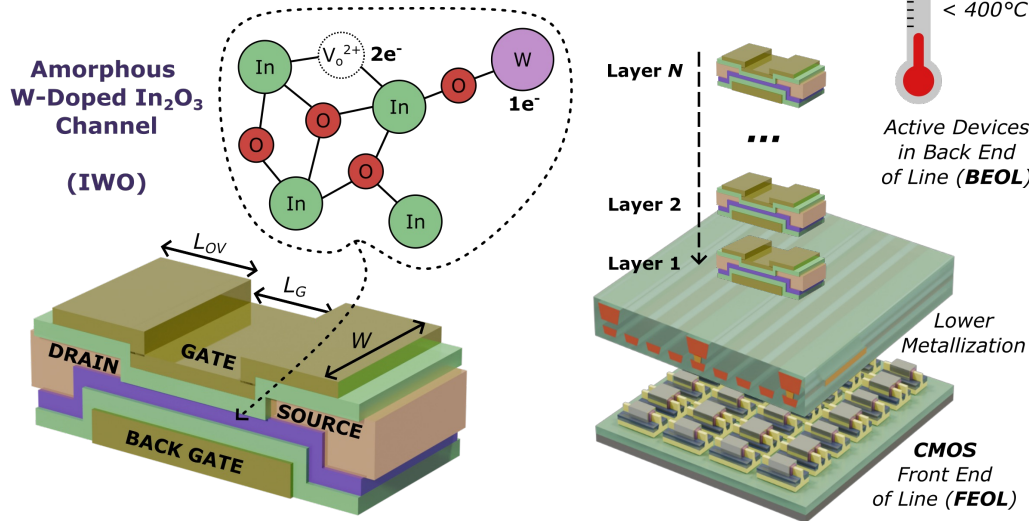
On SRAM Densification and GPU Memories



Meanwhile, (HD-)SRAM bit cell density has only improved by ~10×

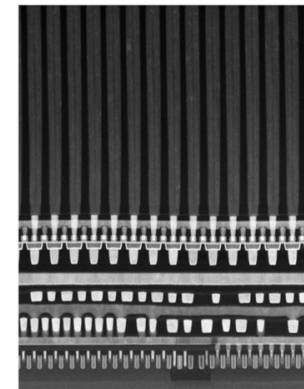
GPU memory capacity growth is driven by architectural demand, not technology scaling

A Path Upwards – Monolithic 3D Integration



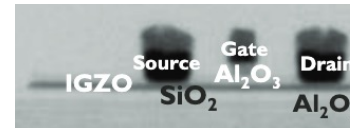
Opportunity: Amorphous Oxide Semiconductor (AOS) Transistors

- (+) Back-End-of-Line (BEOL) **compatible** fabrication temp.
- (+) **Ultra-Low Leakage** ($< 10^{-19} \text{ A}/\mu\text{m}$), Wide Band-Gap
- (+) **Moderate Mobility** ($\sim 20 \frac{\text{cm}^2}{\text{Vs}}$), short write access



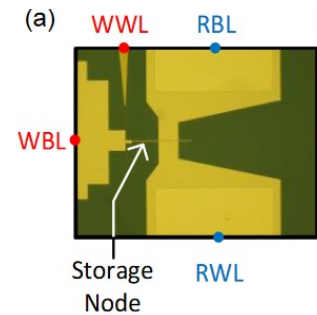
TSMC, Vertical BEOL eDRAM

K. Chiang et al. VLSI 2025



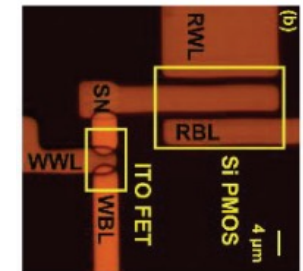
IMEC, Self-Aligned BEOL FET

A. Belmonte et al. IEDM 2021



GT, BEOL Double-Gated FET

H. Ye et al. IEDM 2020



Stanford, BEOL Hybrid Gain-Cell

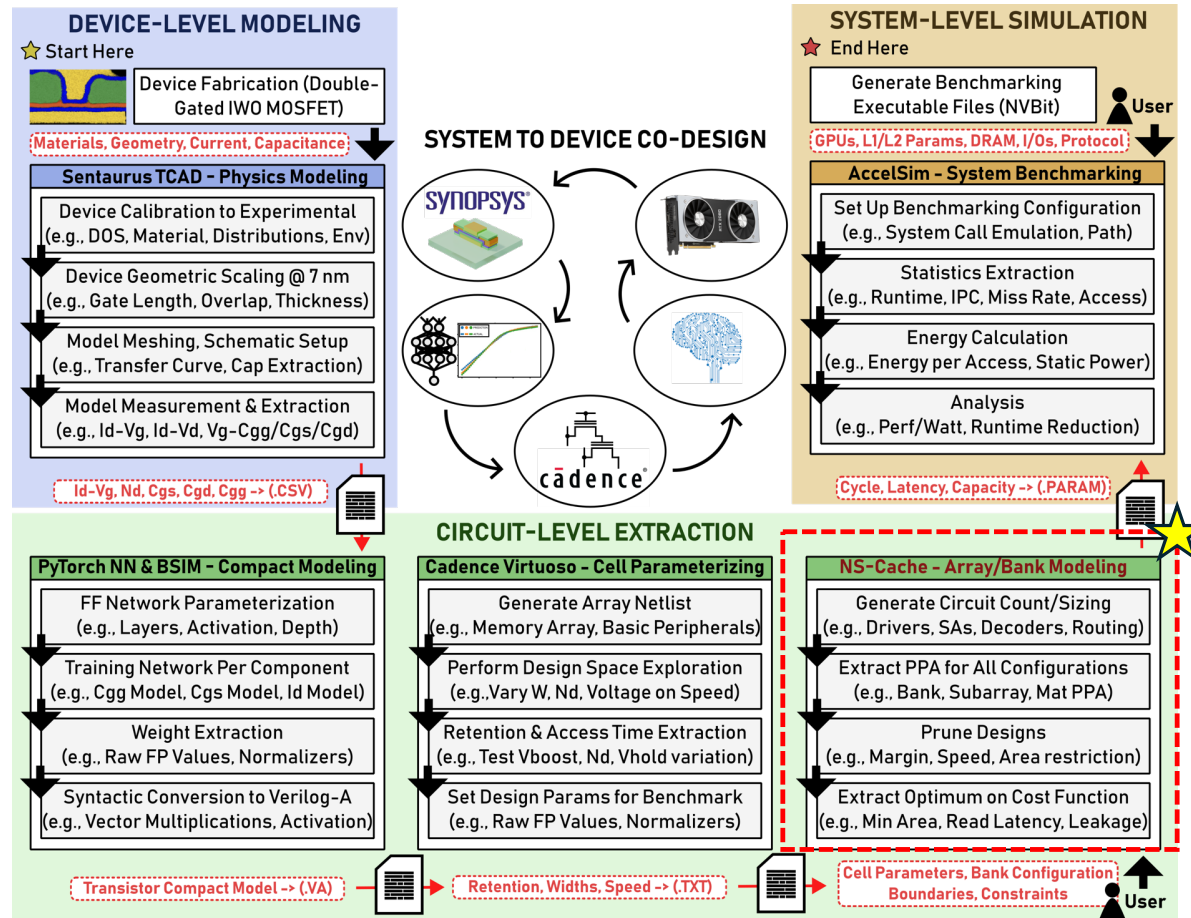
S. Liu et al. VLSI 2024

Questions Given the Technology:

1. In what GPU memory subsystems are AOS candidates (given their tradeoffs) most suitable?
2. What memory-cell topologies present the most significant opportunities?

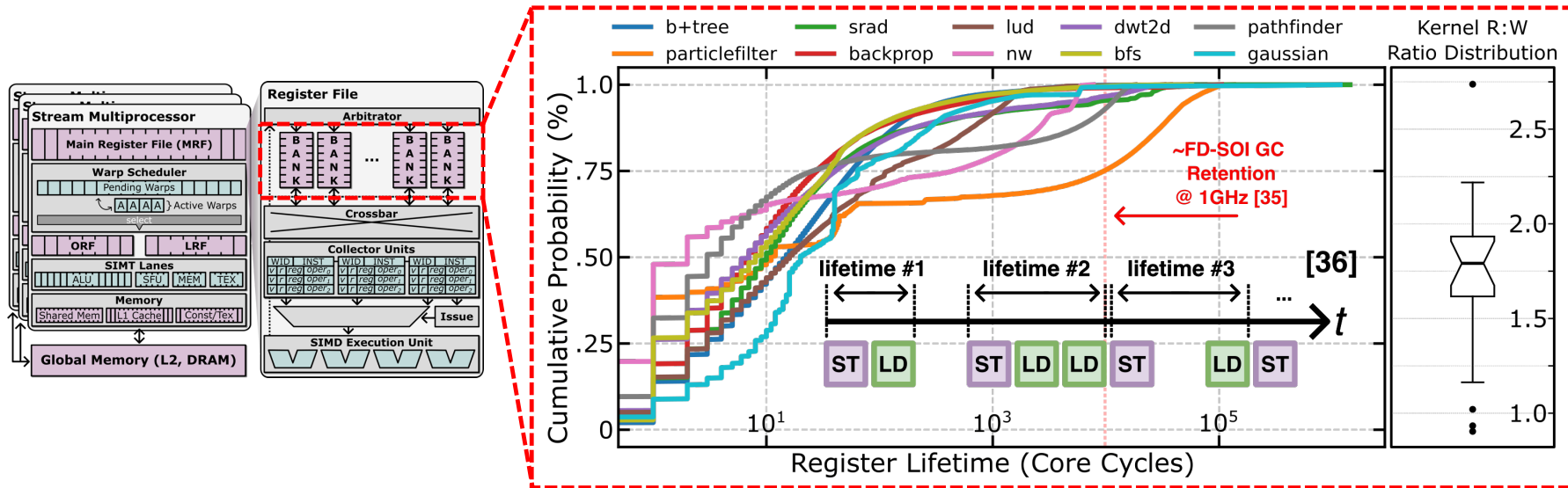


Simulation Methodology



*To be open-sourced –
Currently under review*

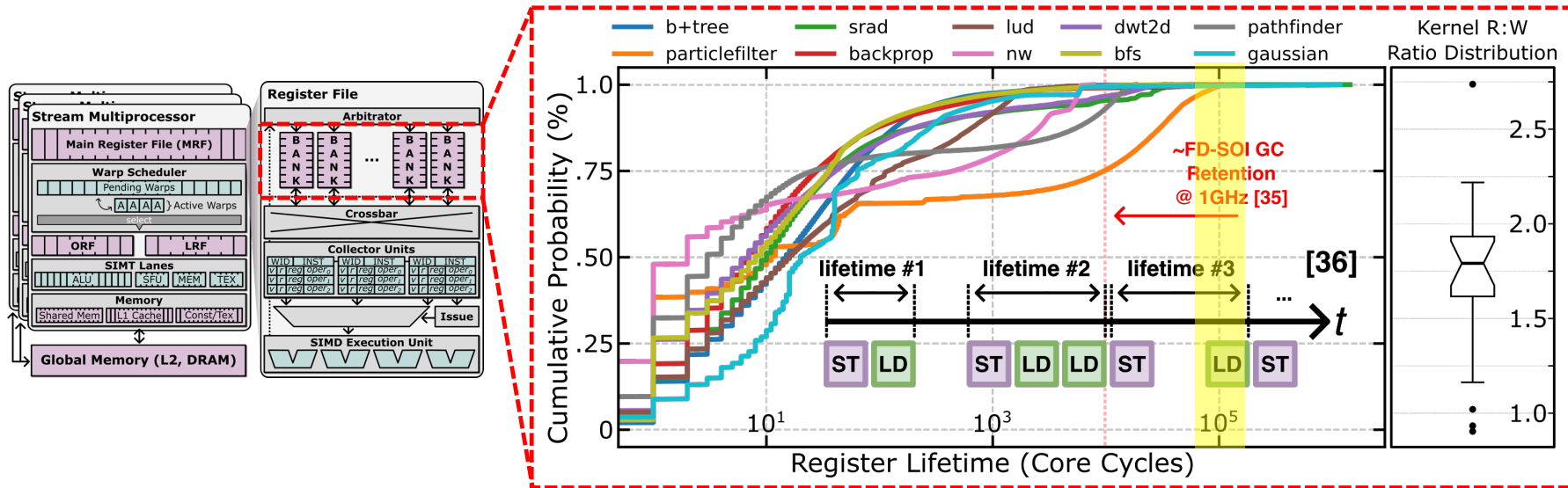
On Register Operands



The GPU Register File (RF) dictates computational capacity:

- Each stream-multiprocessor (SMP) contains **65k registers** to hold operands
- RF capacity dictates the number of resident threads/warps per SMP
 - RF parallelism (i.e. banking) dictates warp size

On Register Operands



Register Operands Don't Have Long Lifetimes,:

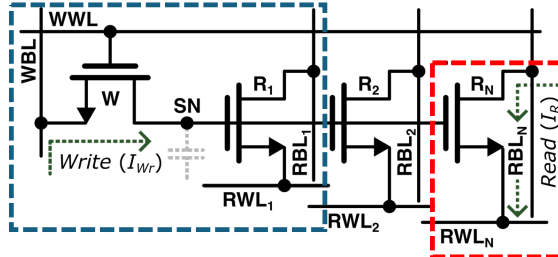
- Measuring on Rodinia, nearly all operands live in a 10^5 -cycle boundary (100 μ s @ 1GHz)
 - Most kernels have read-heavy behavior (indicative of NVIDIA 3R1W ORF)

Takeaway: Retention times in AOS exceed operand lifetimes, multi-read porting desirable

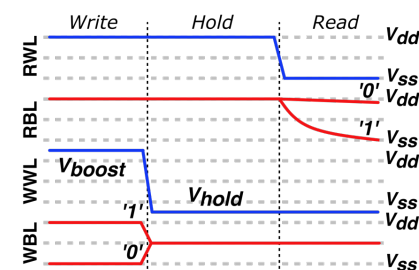
Multi-Ported AOS Gain-Cell

1 Write Transistor +
 N_{RP} Read Transistors
(AOS)

(a) NT0C Multi-Port AOS Gain-Cell [NR1W]

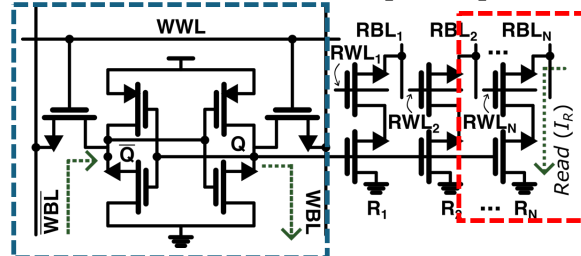


(b) MP Gain Cell Waveform

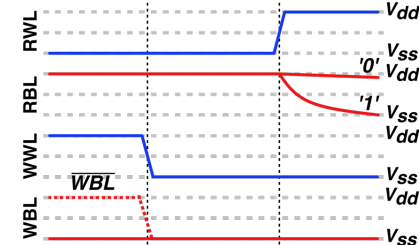


6 Transistors (Base) +
 $2 * N_{RP}$ Read
Transistors (Si)

Multi-Port Si SRAM [NR1W]



MP SRAM Waveform

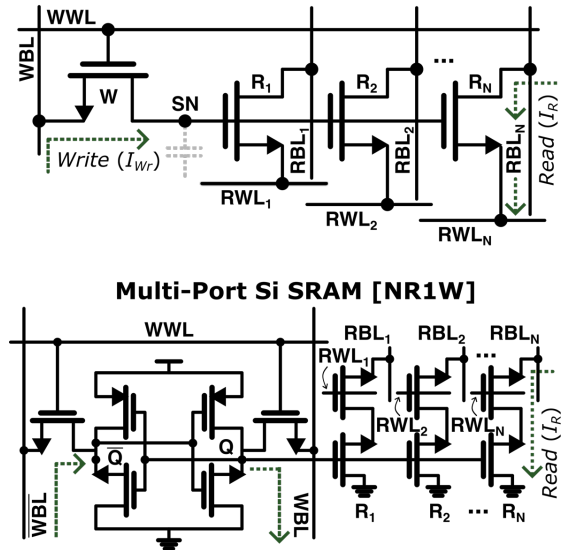


Gain-cell topologies are intrinsically bifurcated:

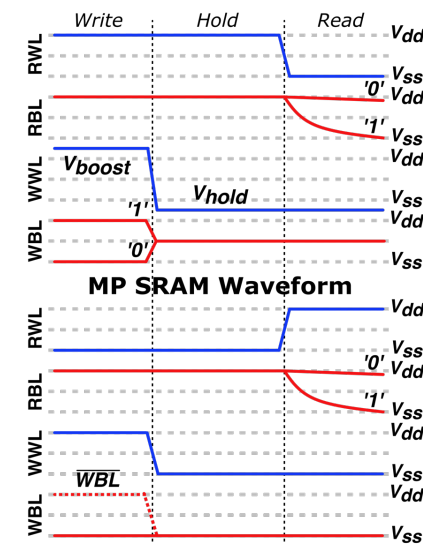
- Write access transistor (W) AOS → long retention despite using parasitic C_{SN}
- SN can be attached to additional read paths → opportunity for single transistor read port (R_i)

Multi-Ported AOS Gain-Cell

(a) NT0C Multi-Port AOS Gain-Cell [NR1W]



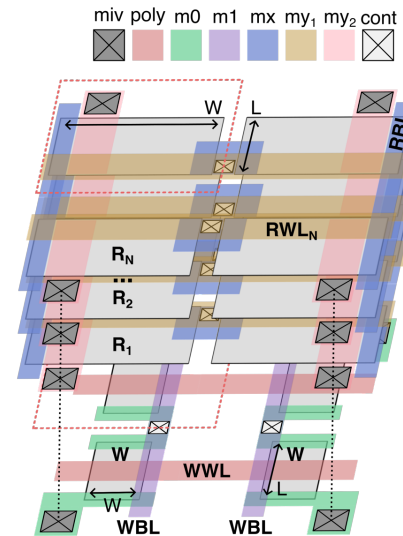
(b) MP Gain Cell Waveform



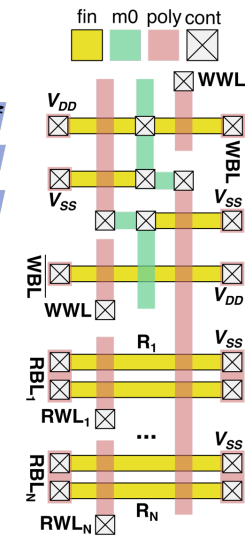
Opportunity for
M3D Stacking

Each Read Port
Takes Large FEOL
Area

(c) MP AOS Gain Cell



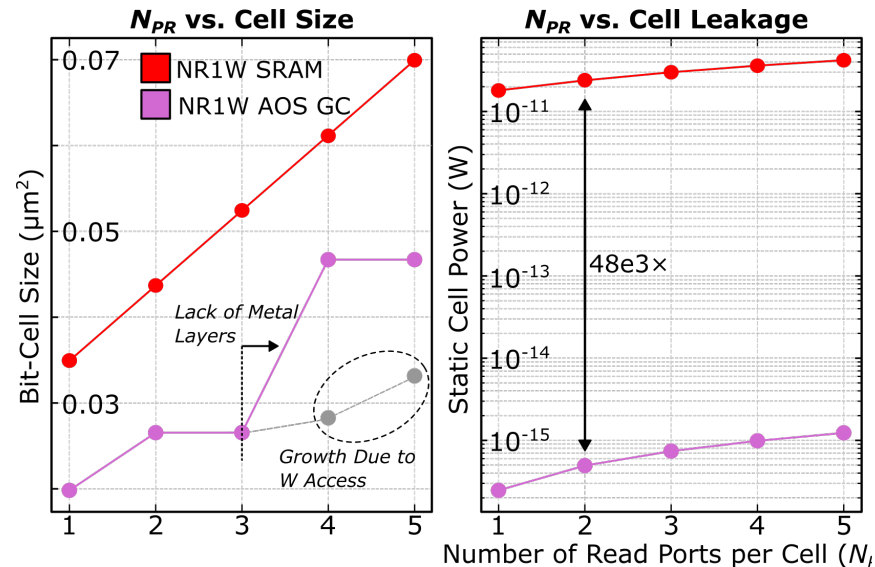
(d) MP SRAM Cell



Gain-Cell topologies are naturally bifurcated:

- Write access transistor (W) AOS → long retention despite using parasitic C_{SN}
- SN can be attached to additional read paths → opportunity for single transistor read port (R_i)

As a Function of N_{RP} (Number of Read Ports)

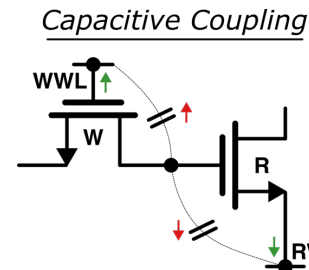
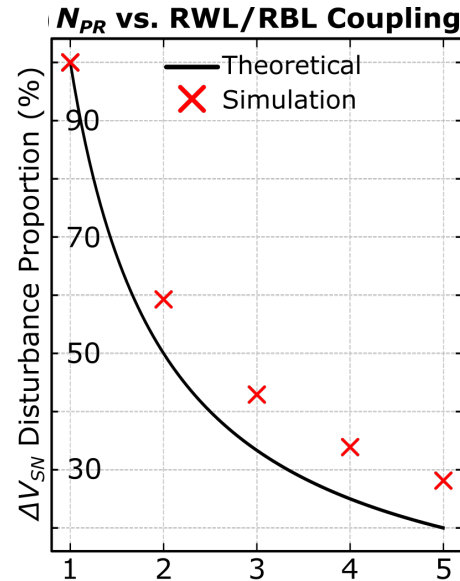


At 3R1W, the MP AOS GC is **76% smaller planar footprint** than SRAM counterpart

- Control over RWL & RBL at S/D of AOS FET + Exclusion of cross-coupled inverters **sharply reduces static power ($>10^4 \times$)**

Takeaway: Beyond M3D Integration, MP AOS GC can offer strong PPA advantages over Si Solution

As a Function of N_{RP} (Number of Read Ports)



$$\Delta V_{SN} \propto \frac{W_{WA}}{W_{WA} + 2W_{RA}N_P} \Delta V_{WWL}$$

$$= \frac{W_{RA}}{W_{WA} + 2W_{RA}N_P} \Delta V_{RWL}$$

Where...

$W_{WA} \equiv$ Width of Write Access

$W_{RA} \equiv$ Width of Read Access

$N_P \equiv$ Number of Ports

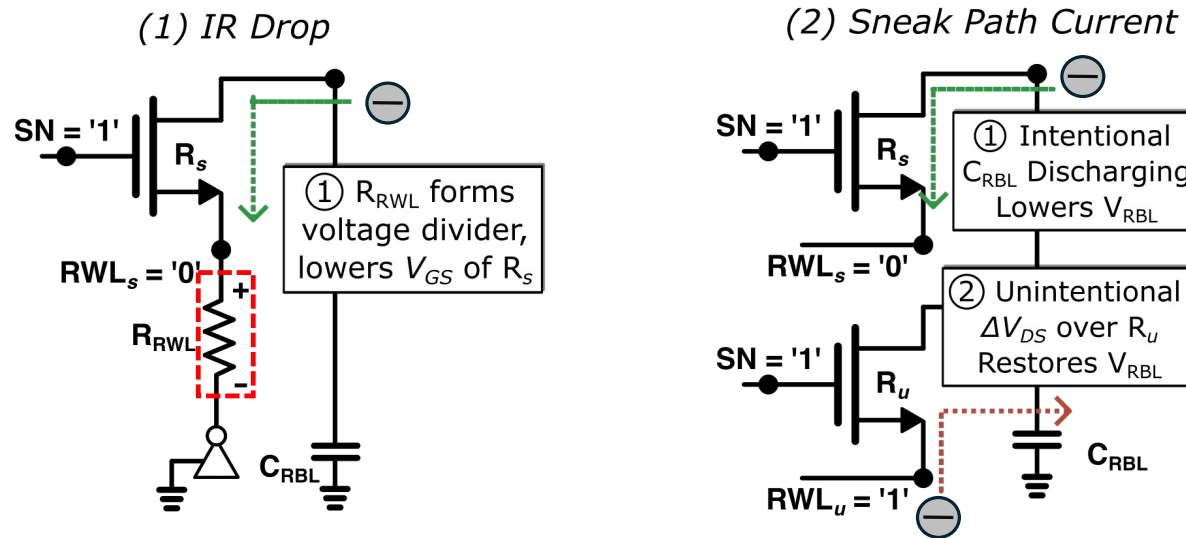
Capacitive coupling during RWL/WWL toggling:

- Proportional to ratio of targeted capacitance & all parasitically coupled capacitances to storage node
 - Increased porting increases number of coupled caps $\rightarrow$ reduced CC

Takeaway: Combined with Geometric Solutions, MP cells provide topological pathway to reduce CC

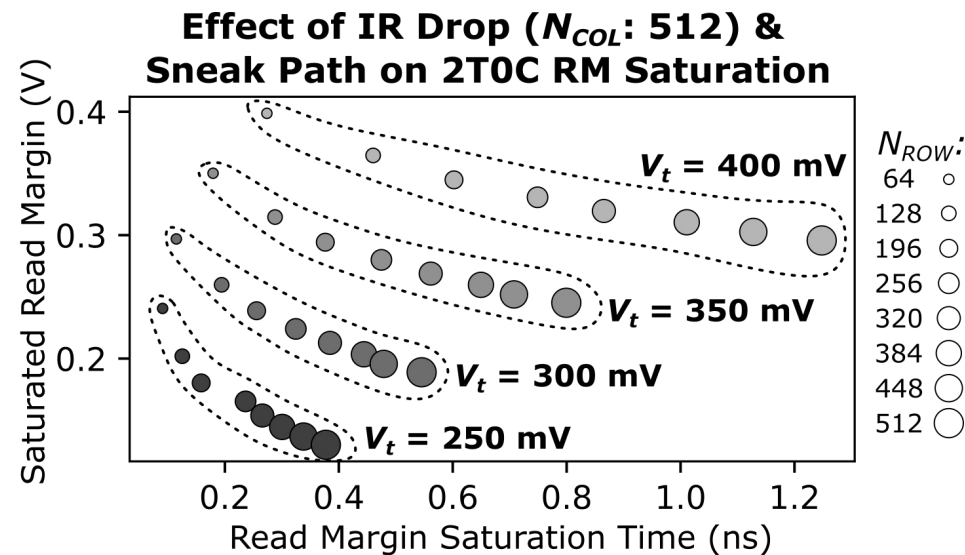
Quantifying Read Margin Degradation

Charge loss in single-transistor read port



Takeaway: Single transistor read port comes with difficulties, how much read margin is lost?

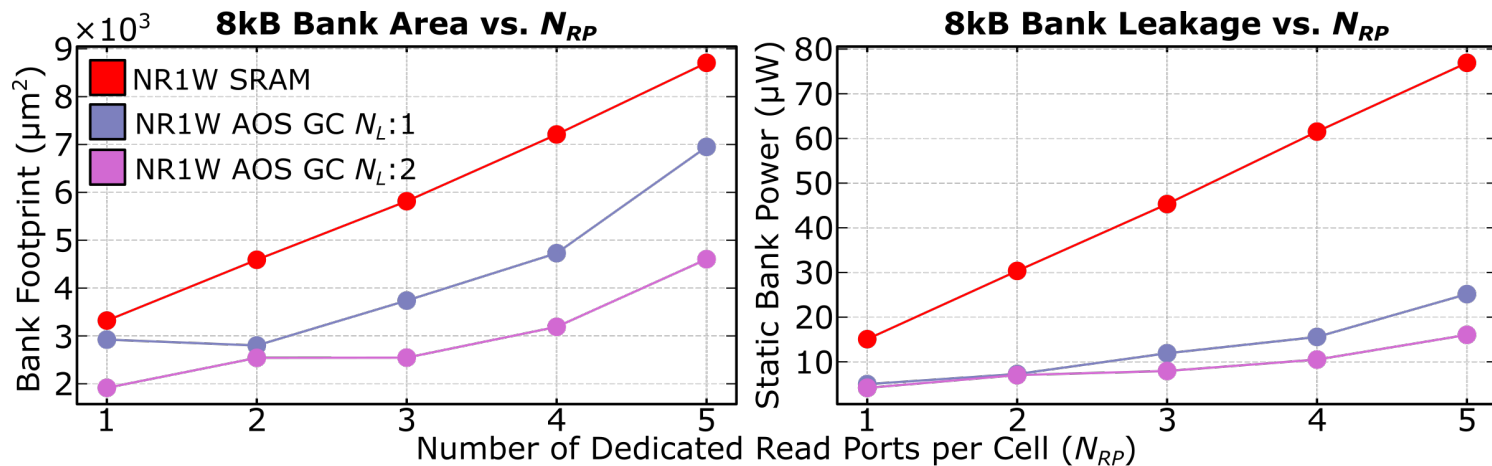
Quantifying Read Margin Degradation



Lower Read V_t Doesn't Always Improve Read Speed:

- Lower $V_t \rightarrow$ Larger Sneak Path currents from unselected '1' cells
- In high row counts, the V_t vs Speed relationship is parabolic; large N_{ROW} may not be readable

As a Function of N_{RP} (Number of Read Ports)

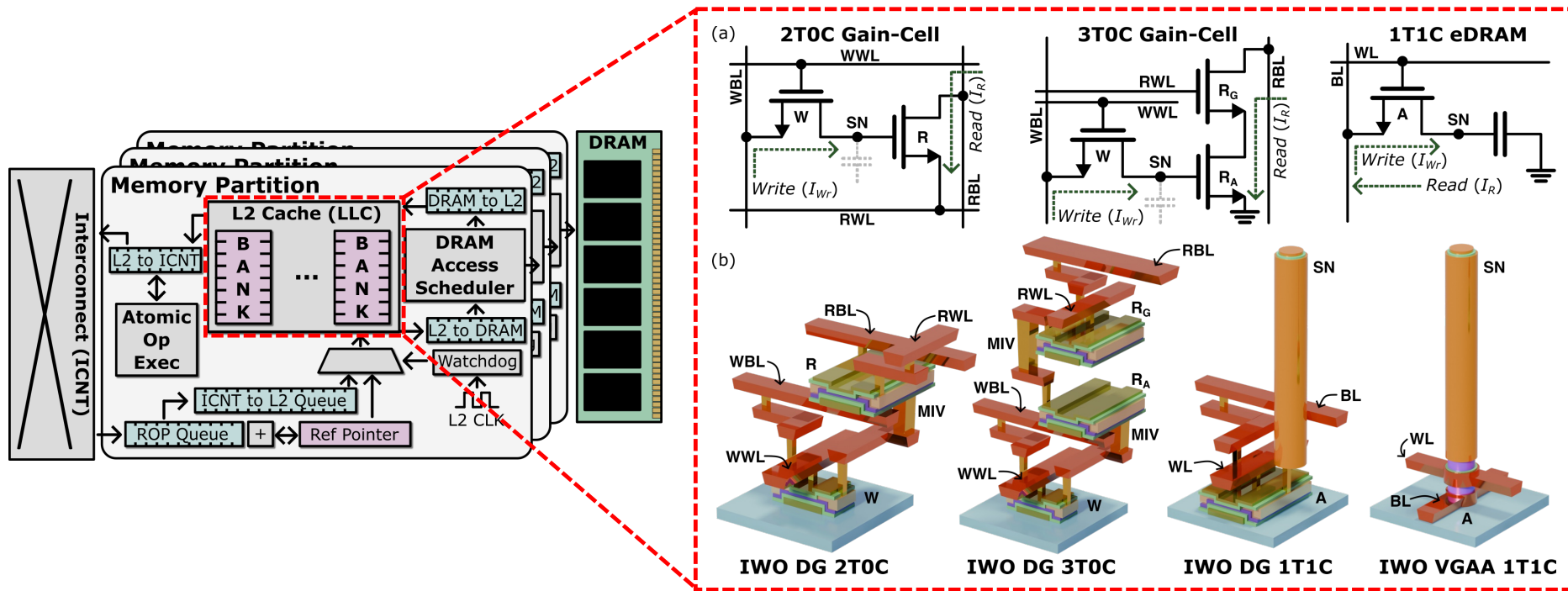


Modeling After NVIDIA Fermi's MRF (8kB Bank, 16B wide, 32b per register) under 750ms subarray RCT constraint

MP can reduce RF area (23-50%), static power (80%).

Static power dominated by peripherals. Multi-tier options improve static power by reducing load (and thus optimal sizing)

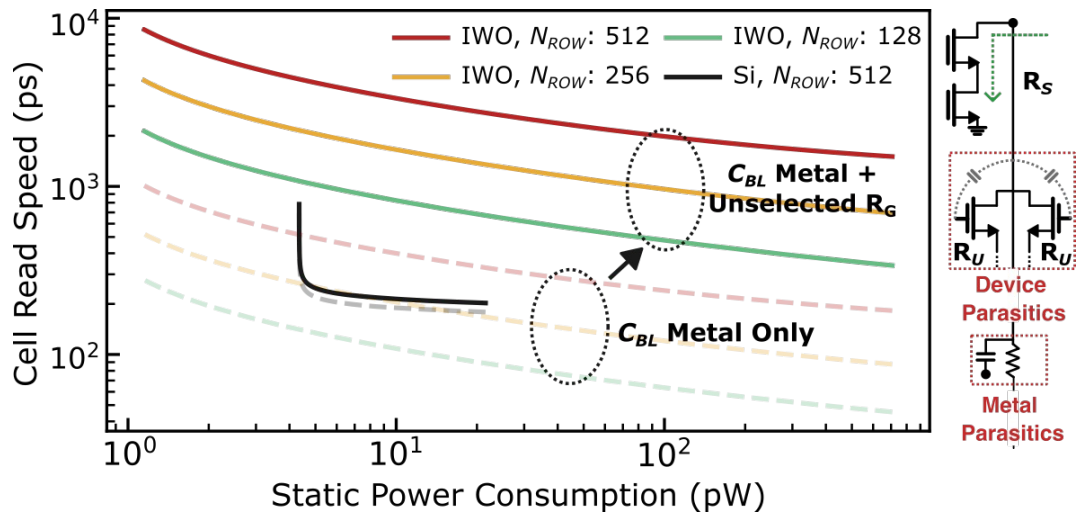
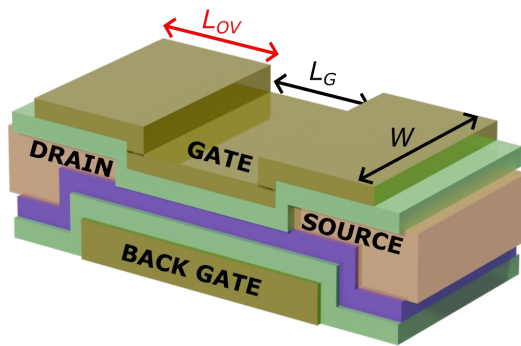
Candidates for Last Level Cache



What about cache? Access-optimized capacitive memories with high density

How do density, speed, and persistency tradeoffs affect their viability?

On 3T0C Leakage and Read Speed



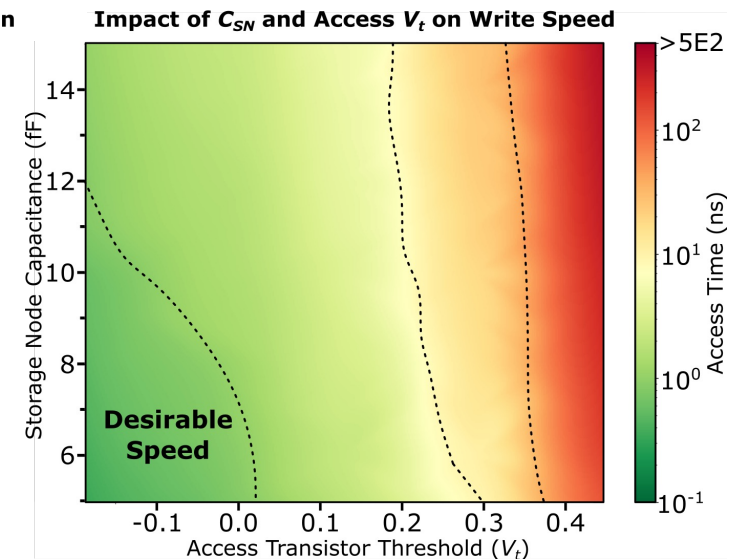
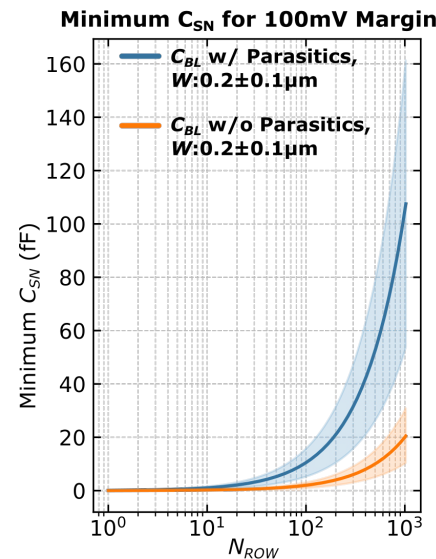
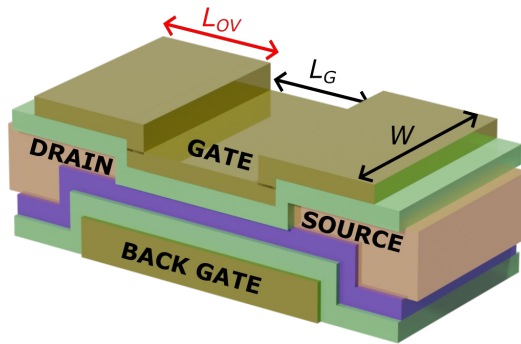
Overlap Contributes to a Speed-Static Power Tradeoff in 3T0C:

- L_{OV} contributes to BL capacitance, requiring greater read current density at iso-speed
- Lower V_t yields larger potential over read port, greater leakage.

Takeaway: the combined higher C_{para} + lower mobility leads to poor read-power tradeoff in 3T0C

On Access Speed and Readout in 1T1C

$$\Delta V_{BL} = \frac{1}{1 + C_{BL}/C_{SN}} \left(\frac{1}{2} V_{DD} - \frac{I_{leak} \cdot t_{ret}}{C_{SN}} \right)$$

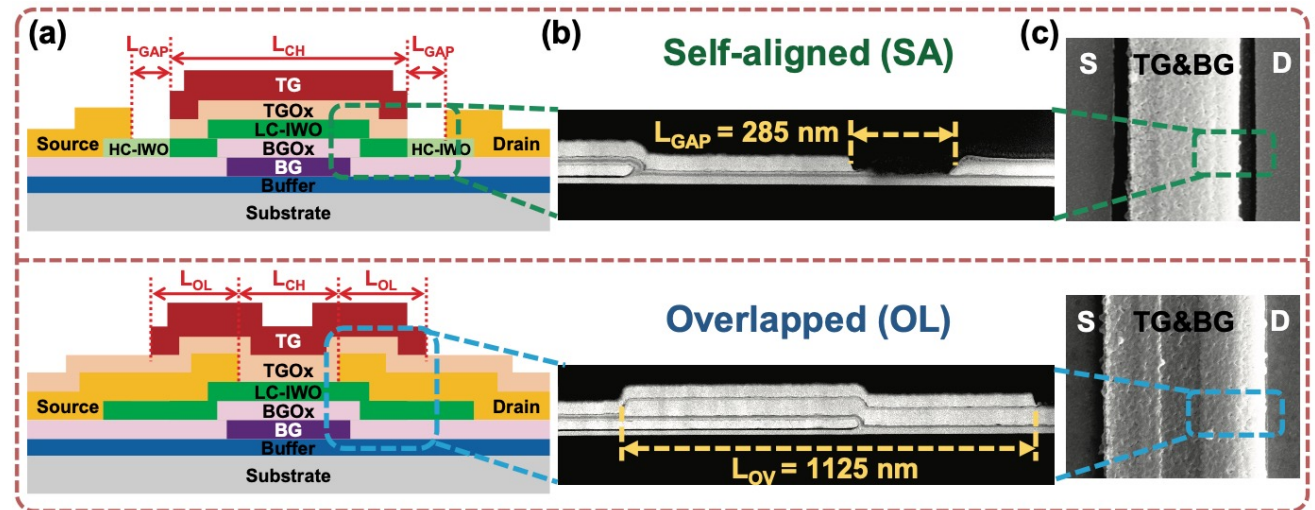
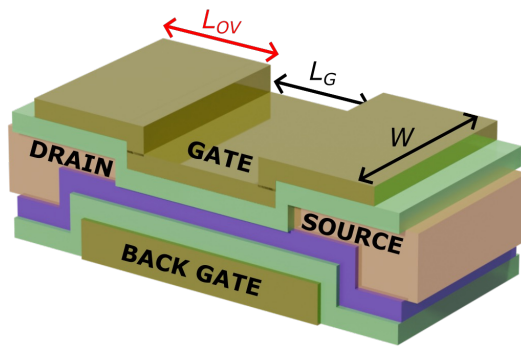


Overlap Contributes to a larger C_{SN} requirement

Scaling to >128 rows requires substantial capacitor due to C_{BL}/C_{SN} ratio, impeding stackability, sizing

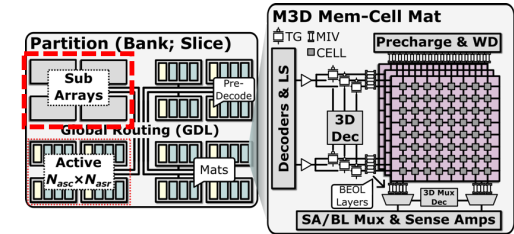
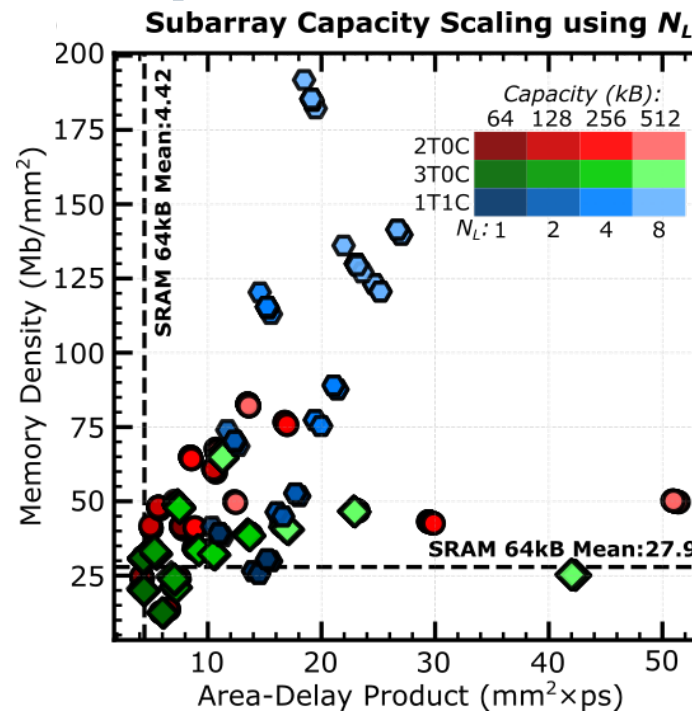
Large C_{SN} comes at cost of access speed to accumulate charge (Q_{SN}), requiring V_t tuning

Emerging Solution - Self-Aligned Structures



A potential solution to parasitic dominance exists in self-aligned structures!
Check us out at IEDM 2025! – S. Deng, J. Sowane, F. Waqar, O. Phadke et al.

Subarray Level Comparison



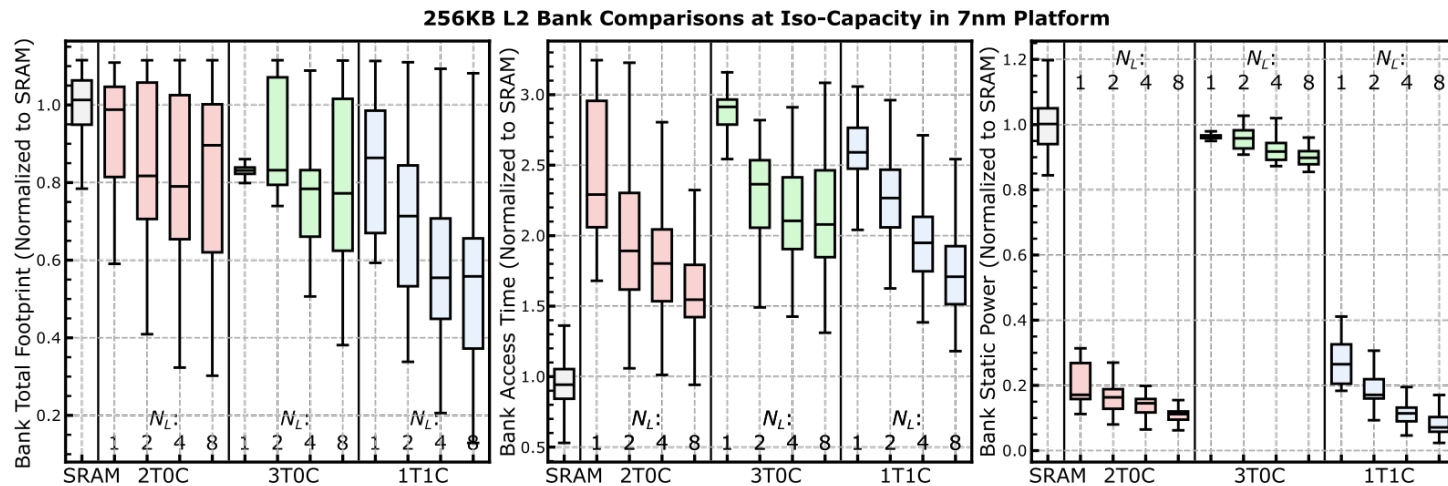
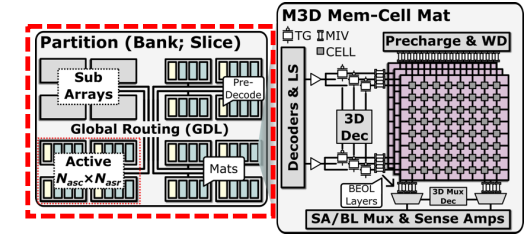
Varying Capacity with Number of Layers (N_L) in Single Subarray under 1.5ns RCT

(1) AOS memories overall **exhibit higher area-delay products than SRAM** counterparts (low mobility)

(2) **Density** benefits of **gain-cell topologies taper off**, while 1T1C holds steady:

higher peripheral overheads, T-gates in 3D-decoding in GC, reduced voltage swing in 1T1C

Bank Level Comparison



Varying Bank N_L under Fixed Capacity at 1.5ns RCT at fixed Footprint

- (1) **3T0C RA to RG Loading increases RC time**, required greater partitioning
- (2) Both access time and density benefits level off in AOS gain-cells, reduced in 1T1C
- (3) **Static power** in 2T0C/1T1C significantly **improved** over SRAM, while speed tradeoff in 3T0C halts improvement

Benchmarking Setup

Table 3: Evaluated Benchmarks and Corresponding Domain

Application	Abbrev.	Domain
Covariance Computation [7]	cov	Pattern Recognition
Particle Filter [7]	pfil	Medical Imaging
Discrete 2D Wavelet Transform [7]	dwt2d	Image/Video Compression
Convolutional Neural Network Training [2]	cnn	Deep Learning
Matrix Transpose and Vector Multiplication [23]	atax	Linear Algebra
Back Propagation [7]	backprop	Pattern Recognition
Matrix Vector Product and Transpose [23]	mvt	Linear Algebra
Pathfinder [7]	pfin	Dynamic Programming
3D Convolution [7]	3dconv	Image Processing
GEMM Kernel Inference [2]	gemm	Deep Learning
Needleman-Wunsch [7]	nw	Bioinformatics
B+ Tree [7]	b+tree	Search
RNN + GRU Training [2]	rnn	Deep Learning
Correlation Computation [7]	corr	Signal Processing
3 Matrix Multiplication [23]	3mm	Linear Algebra

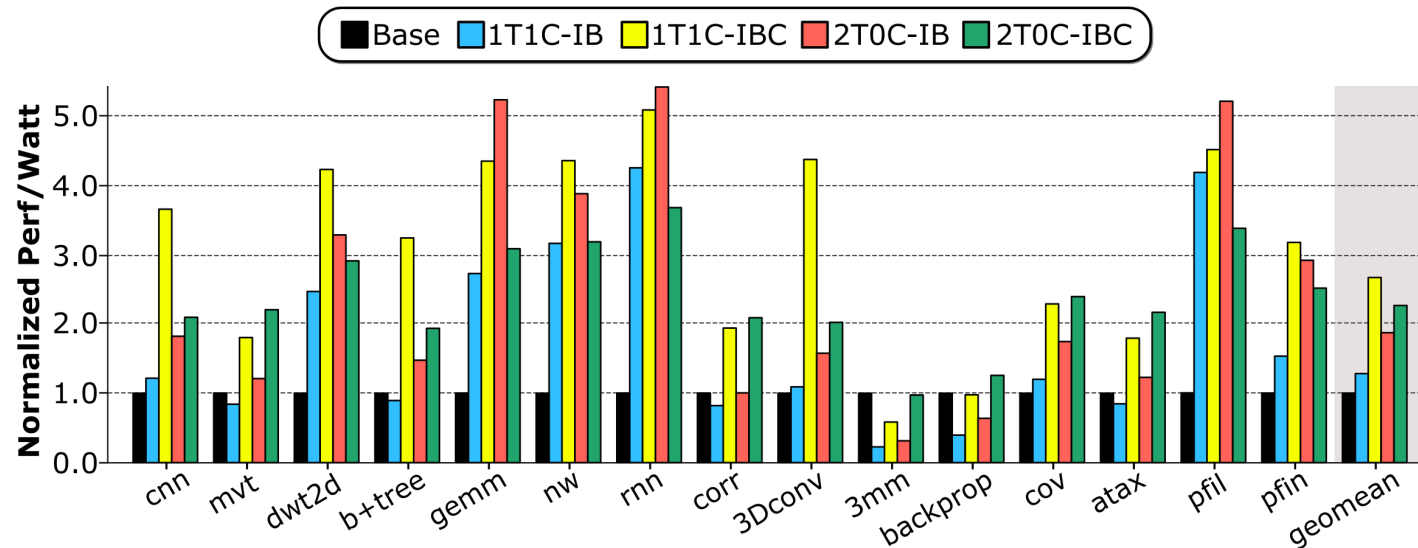
Table 4: Benchmarking L2 Cache Parameters

Config	Memory Type	Num Banks	L2 Capacity	N_L	L2 Clock Domain	L2 Area p. Partition	ROP Latency	Ref Period
Baseline	SRAM	8×2	4 MB	1	1132 MHz	160,328 μm^2	187 cyc.	N/A
2T0C IB	2T0C IWO	8×2	8 MB	2	1067 MHz	130,453 μm^2	184 cyc.	859 μs
2T0C IBC	2T0C IWO	8×8	16 MB	8	1132 MHz	195,044 μm^2	188 cyc.	215 μs
1T1C IB	1T1C IWO	8×2	32 MB	8	724 MHz	175,771 μm^2	190 cyc.	244 μs
1T1C IBC	1T1C IWO	8×16	32 MB	8	724 MHz	166,785 μm^2	190 cyc.	244 μs

Building upon Accel-Sim Modeling of NVIDIA Ampere RTX3070 with 2 Scenarios 2T0C & 1T1C:

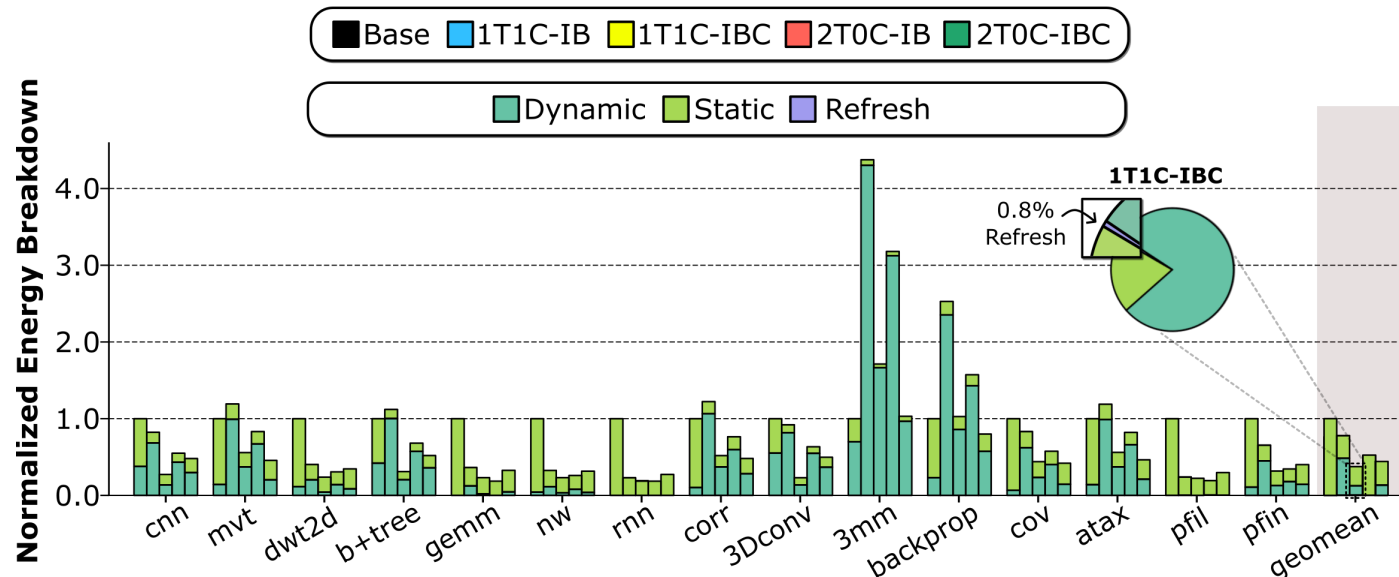
- (1) Iso-Banking (IB): Same number of banks & footprint, maximizing capacity
- (2) Iso-Bank Capacity (IBC): Same bank capacity & footprint, maximizing number of banks

Performance and Runtime



Though gain cell topologies offer superior access times and less restrictive sizing, **reduced peripheral overhead in AOS 1T-1C is more suitable for LLC** (2.8x perf/watt improvement)

Energy and Miss Ratio



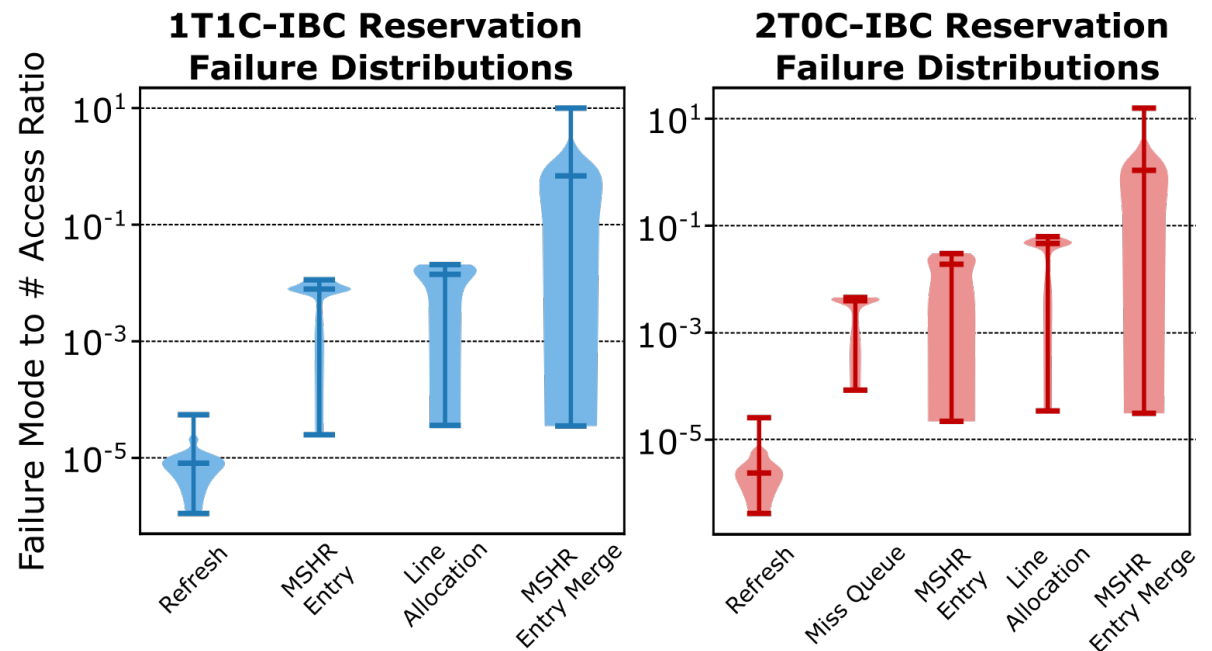
Refresh operations in AOS 1T1C and 2T0C caches add only a minor energy overhead—typically <0.8% of total cache energy—since refreshes are infrequent due to long retention (ms-scale)

Static power drops sharply ($\approx 70\text{--}80\%$) compared to SRAM, while dynamic power becomes the dominant contributor, mainly from **write operations and larger driver capacitances**

Reservation Failure Modes and Refresh

Reservation Failures:

- (i). **Refresh:** Refresh in Progress
- (ii). **MSHR Entry:** MSHR table full
- (iii). **Line Allocation:** All ways reserved
- (iv). **MSHR Entry Merge:** Merge list full
- (v). **Miss Queue:** Miss buffer full



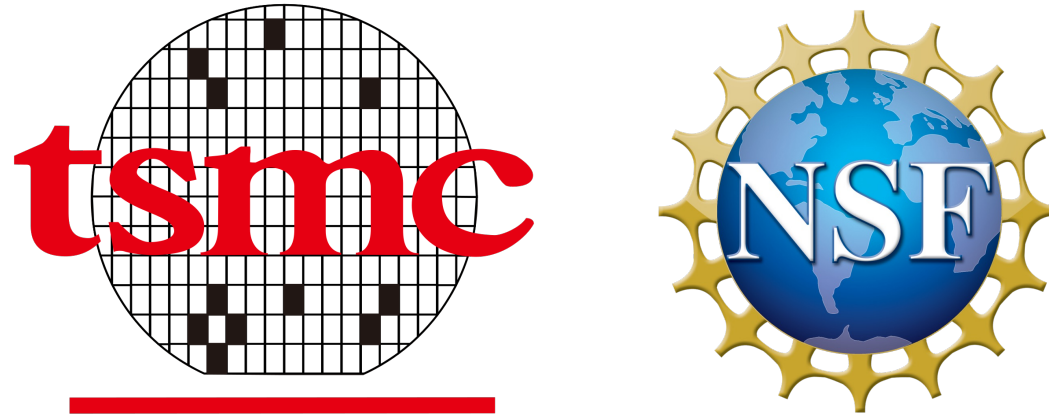
Takeaway: Refresh-induced reservation failures at $\sim 100 \mu s$ periodicity are dwarfed by other failure modes, **modest ms-level retention needed**

Conclusions and Summary

Takeaway: The memory-hungry demands of future GPGPU systems exceed SRAM's capabilities

- The **lifetime of operands is short** ($\sim 10^5$ cycles upper bound) → **opportunity for persistent memories**
 - Intrinsically bifurcated **gain-cells** can be used to **reduce RF area** (23-50%), **static power** (80%), and **improve upon cell-level capacitive coupling** (>70% @ 5-Ports)
 - Though gain cell topologies offer superior access times and less restrictive sizing, **reduced peripheral overhead in AOS 1T-1C is more suitable for LLC** (2.8x perf/watt improvement)
 - **Overlap parasitics play a decisive role** in row limitations and SN requirements of AOS memories
 - **A focus on access time pays off** when exceeding ms-level retention of AOS cells in tail-states
- 

This Work Was Supported By



**Thank you for listening. Special thanks to Yu-Ming Lin
and Huai-Ying Huang of TSMC, Taiwan**

Questions?

Email: faaiq.waqar@gatech.edu, shimeng.yu@ece.gatech.edu



Laboratory for Emerging Devices and Circuits