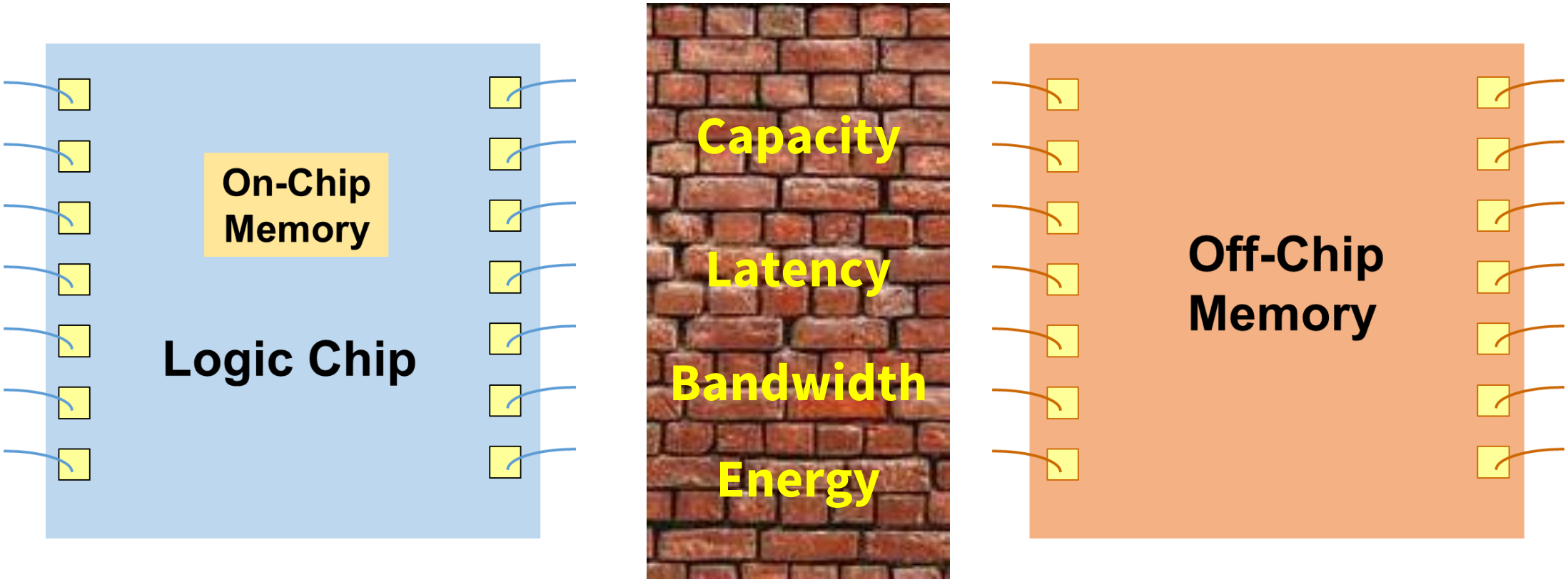
OpenGCRAM: An Open-Source Gain Cell Compiler Enabling Design-Space Exploration for AI Workloads

Xinxin Wang

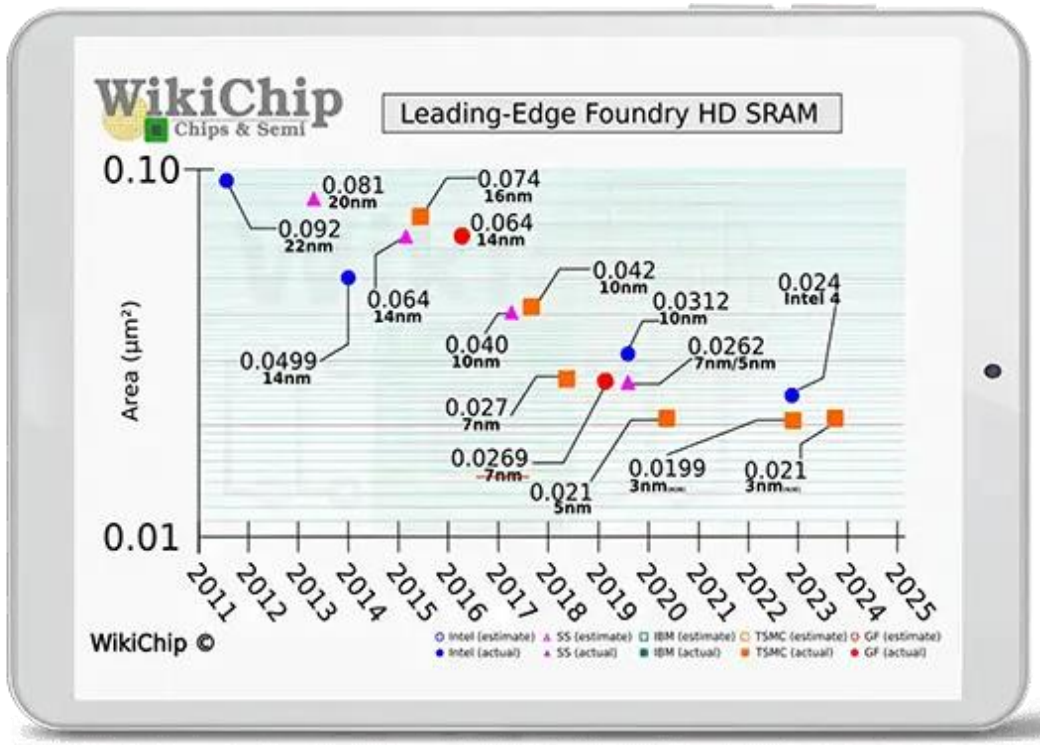
04/28/2025

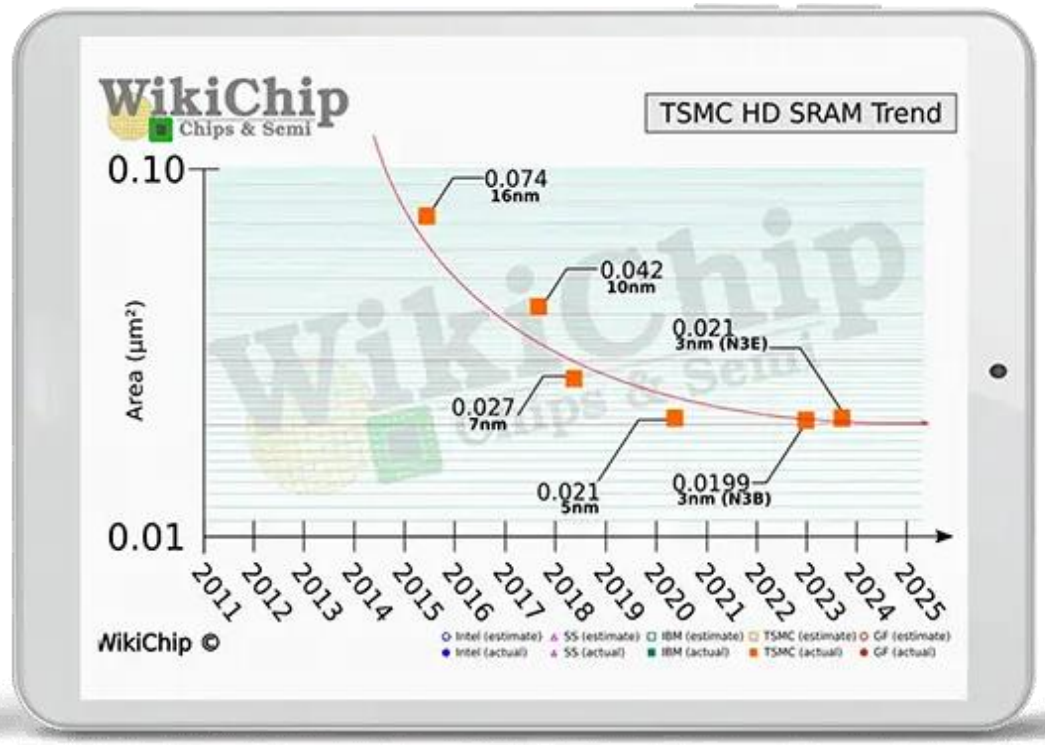
Stanford University

Memory Wall


On-Chip Memory
Logic Chip
Capacity
Latency
Bandwidth
Energy
Off-Chip Memory

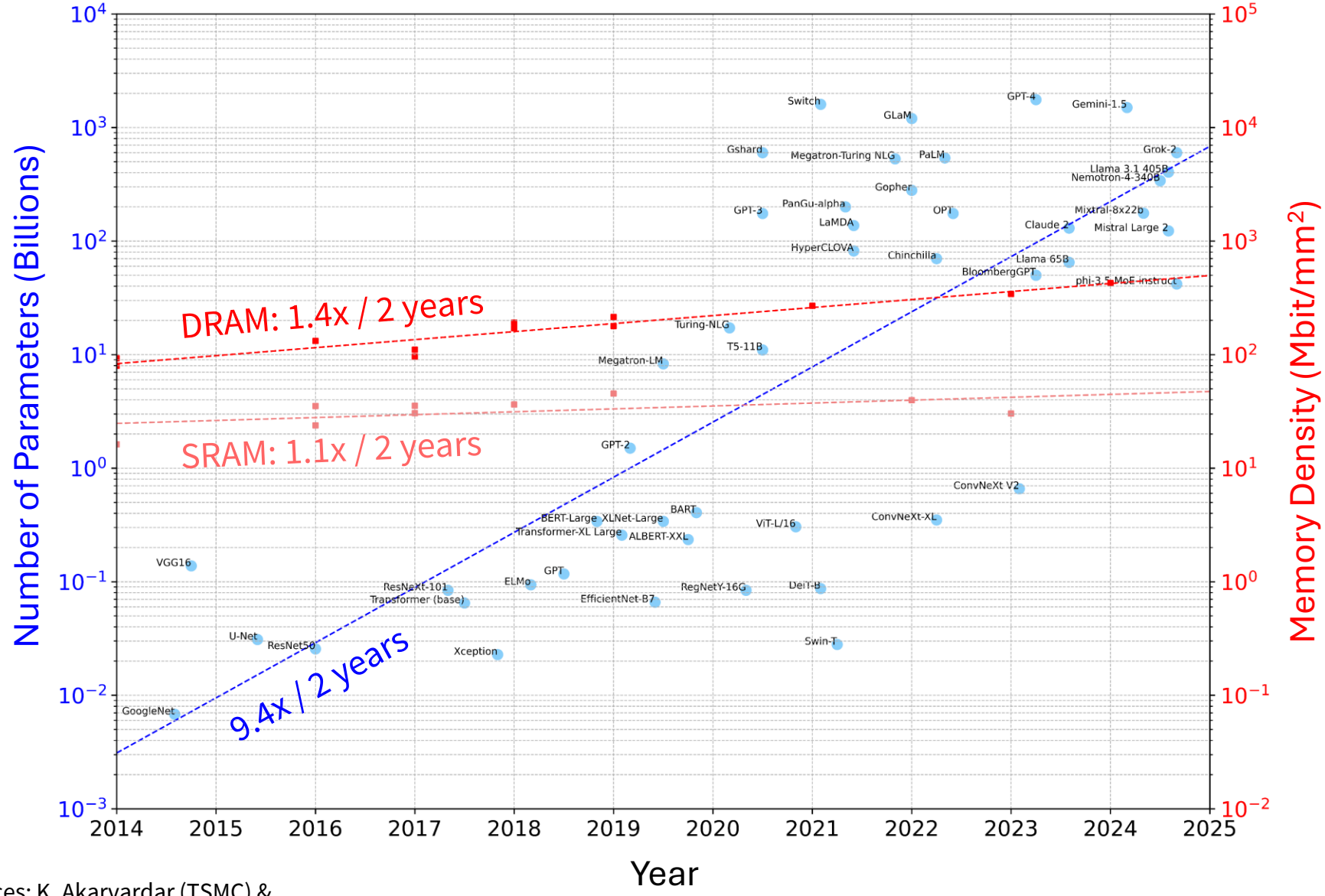
SRAM scaling miniaturization





Source: https://raaam-tech.com/technology/

Memory Needs Outpace Memory Advances

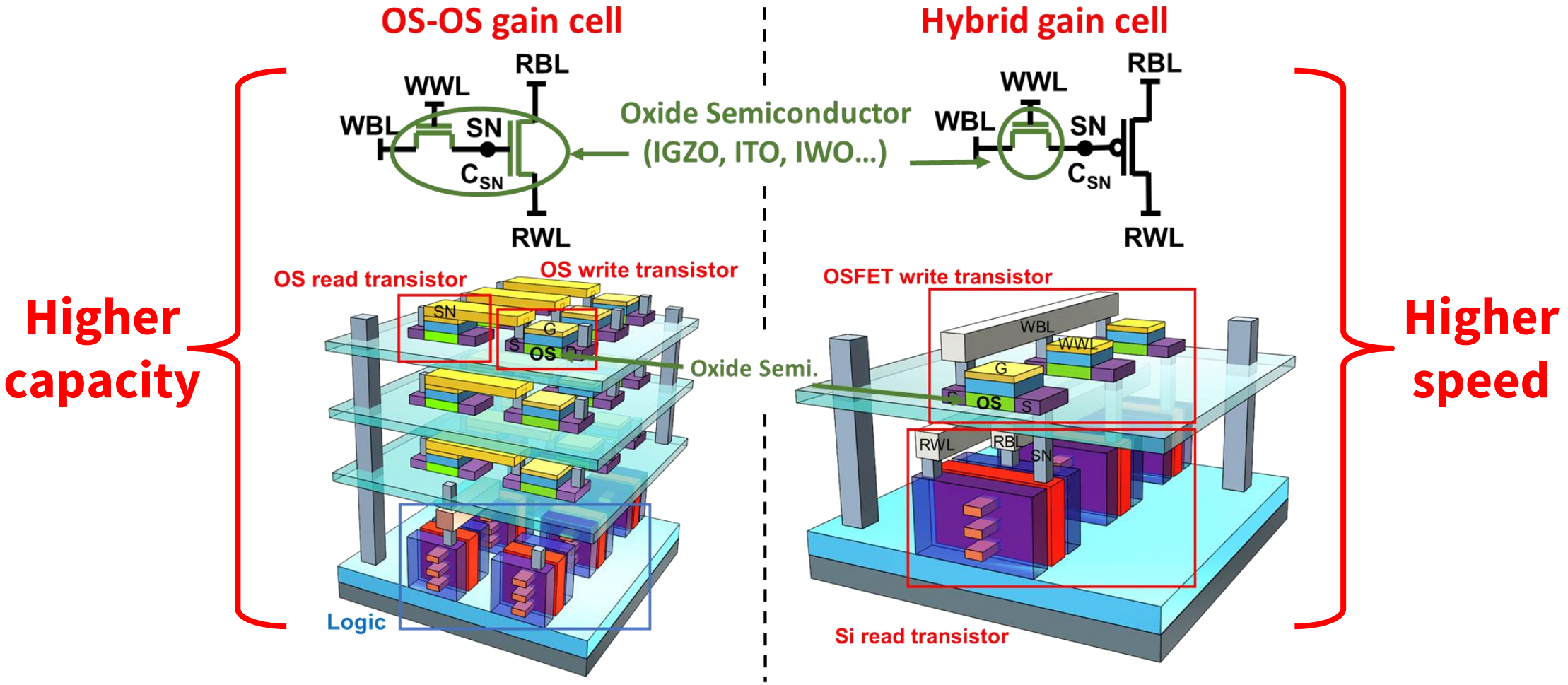


We need innovations for high-density on-chip memory

Sources: K. Akarvardar (TSMC) &
https://nano.stanford.edu/downloads/technology-integration-trend

Slide by Shuhan Liu

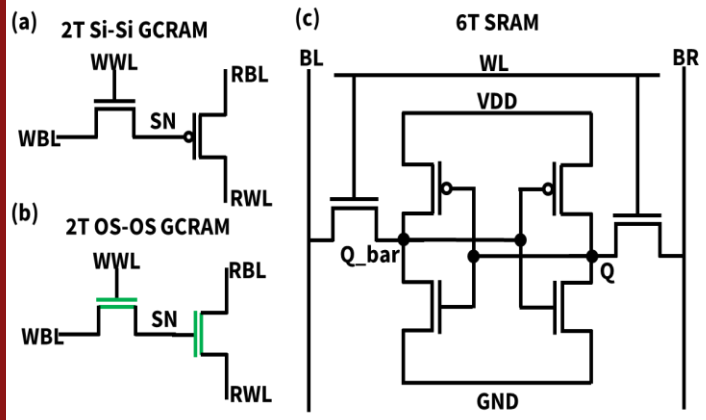
Gain Cell memory: higher density than SRAM

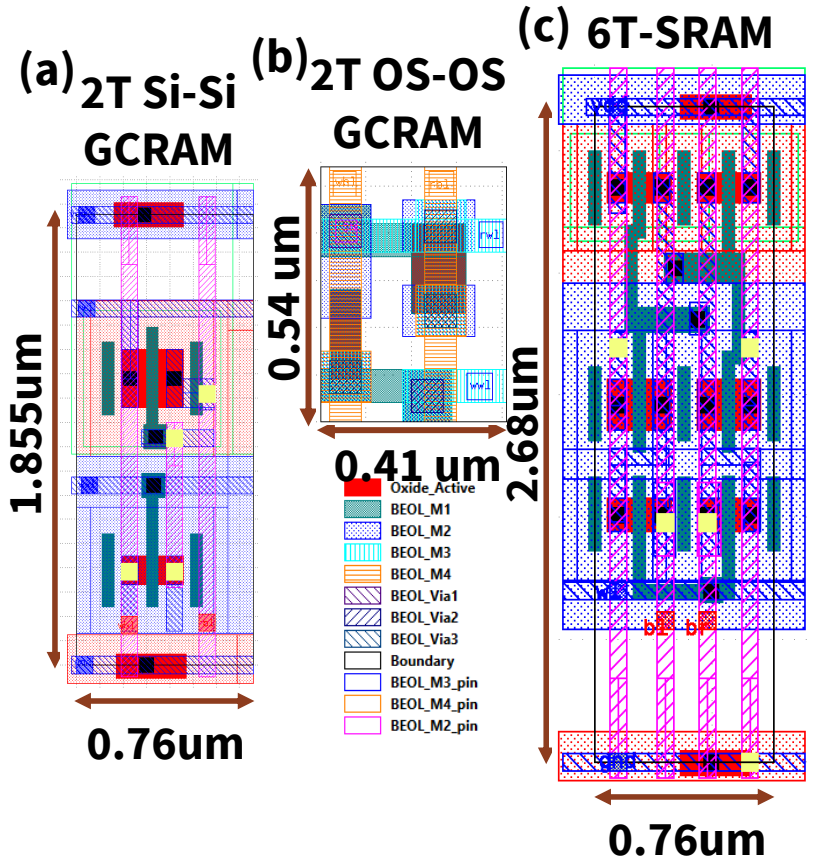


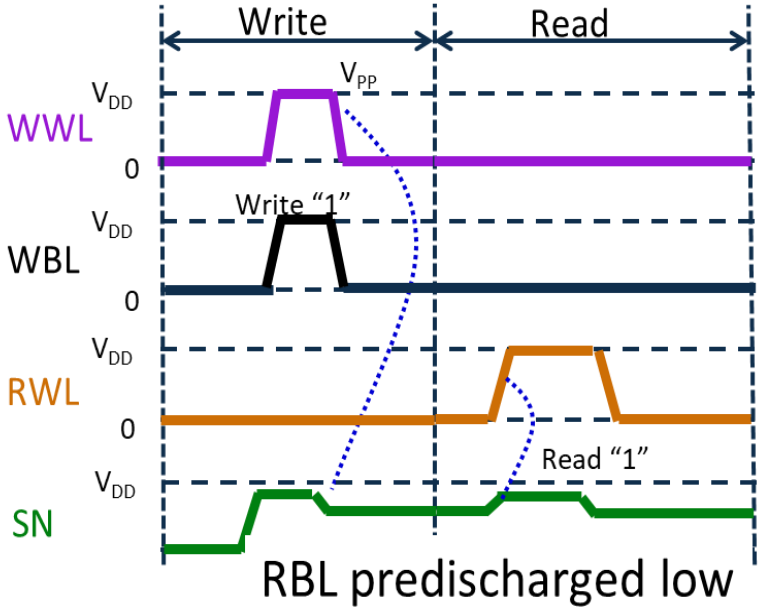
Shuhan Liu, …, H.-S. Philip Wong, IEDM 2023, T-ED 2024, VLSI 2024

Slide by Shuhan Liu

Gain Cell memory: higher density than SRAM







Koustav Jana,… H.-S. P. Wong, *ITC*, poster # 05_1003, 2024, S. Liu …et al., *EDL*, 2024.

To enable fast, accurate, customizable, and optimized Gain Cell bank generation and performance simulation targeting for AI workloads:

We need a Gain Cell compiler

M. R. Guthaus, et al., “OpenRAM: An Open-Source Memory Compiler”, ICCAD, 2016

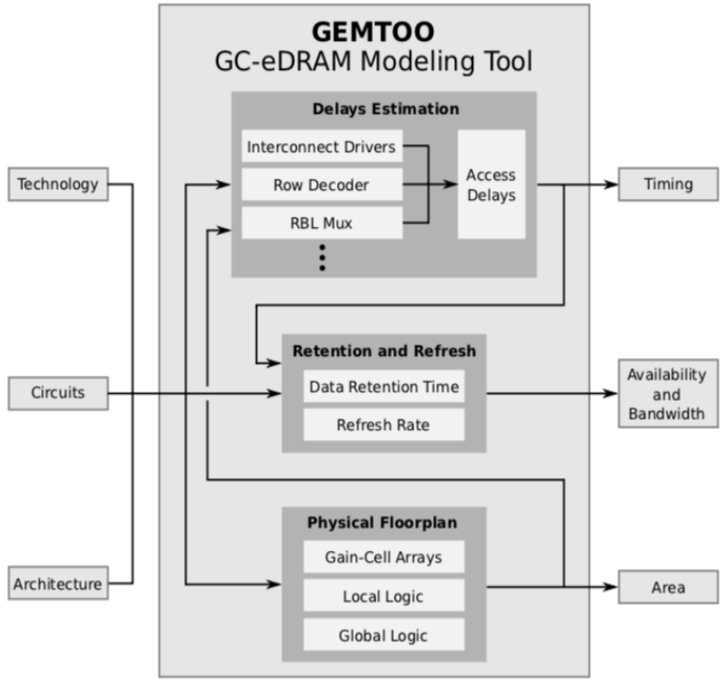
Related work

RAAAM GCRAM

- Up-to **2X Higher density** vs. SRAM
- Up-to **50% area reduction** vs. SRAM
- Standard **SRAM interface**
- **Extended interface options** vs. SRAM
- **Single/two** ported
- Up to **2Mbit** instances
- **Standard CMOS** process
- **Single cycle** operation
- **Customizable**

Limitations:
- 3T GCRAM
- commercialized
- not open-access
- **Compiler in development**

GEMTOO Simulator



Limitations:
- No netlist and layout
- No power evaluation
- Very rough delay estimation

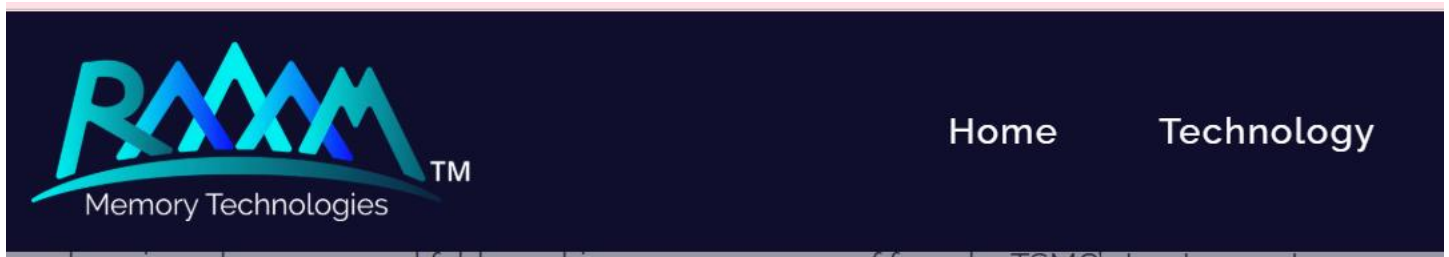
OpenRAM Compiler

- Open-source SRAM compiler
- Support open-source PDKs
- SRAM bank netlist & layout generation
- Performance simulation

Limitations:
- No Gain Cell support
- No commercial PDK support

Related work

- RAAAM GCRAM



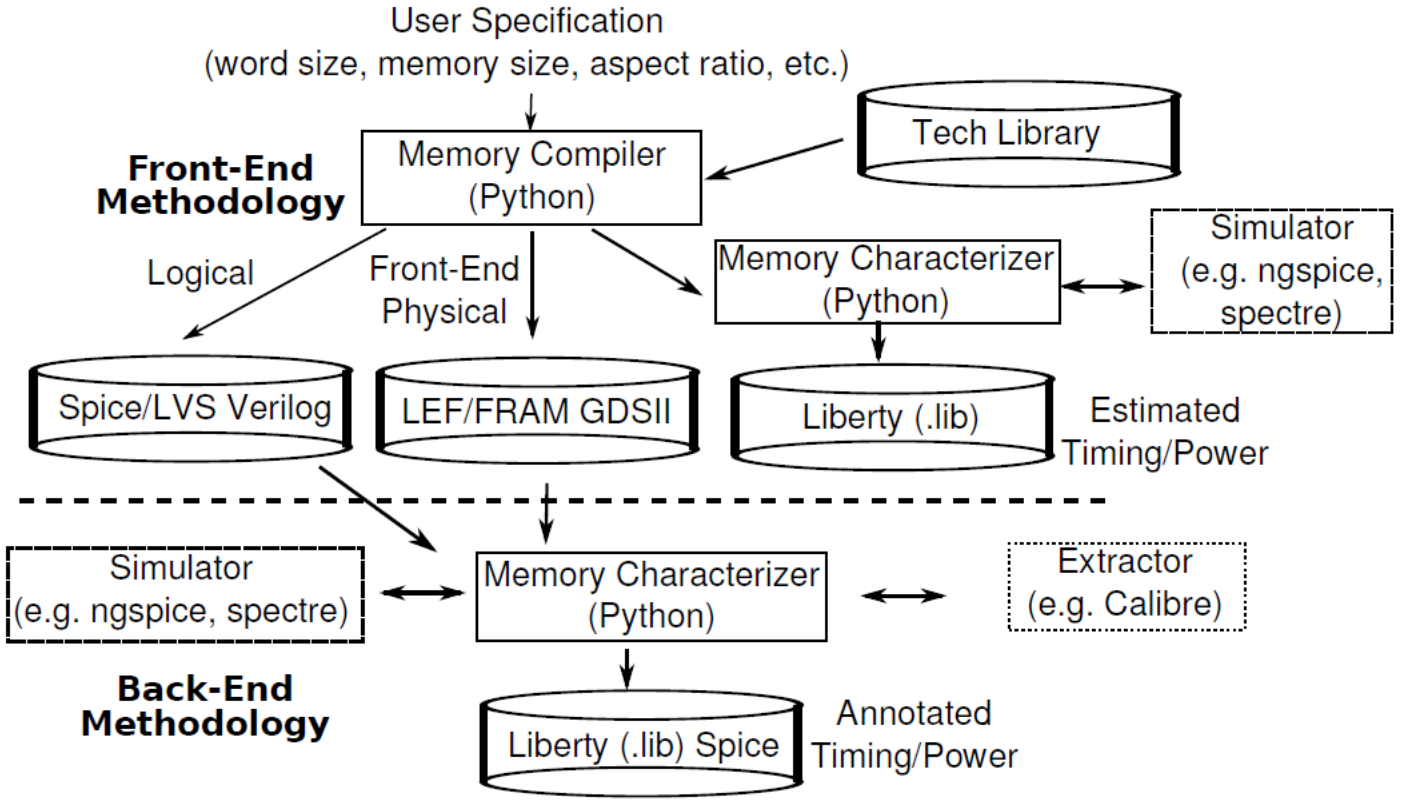
has signed an unnamed fabless chip company – one of foundry TSMC's top ten customers – as a licensee and investor. RAAAM raised $4 million in a seed funding round and in July 2024 announced it had been awarded €5.25 million (about $5.8 million) from the European Innovation Council. This comprises a €2.5 million grant and a follow-on equity investment of €2.75 million.

Having already designed test chips on a 16nm FinFET process, RAAAM is working with the unnamed licensee and TSMC on a test chip tape-out aiming for first 5nm GCRAM silicon in the first quarter of 2025. Giterman said that if the test chip performs as expected it could become a hard macro available for licensing. The company is also writing a compiler so that custom IP blocks can be generated.

Peter Clarke is a veteran reporter and analyst covering the global electronics industry.
This article was first published by The Ojo-Yoshida Report on September 17, 2024.
See www.ojoyoshidareport.com for more of such reports.

OpenRAM SRAM compiler

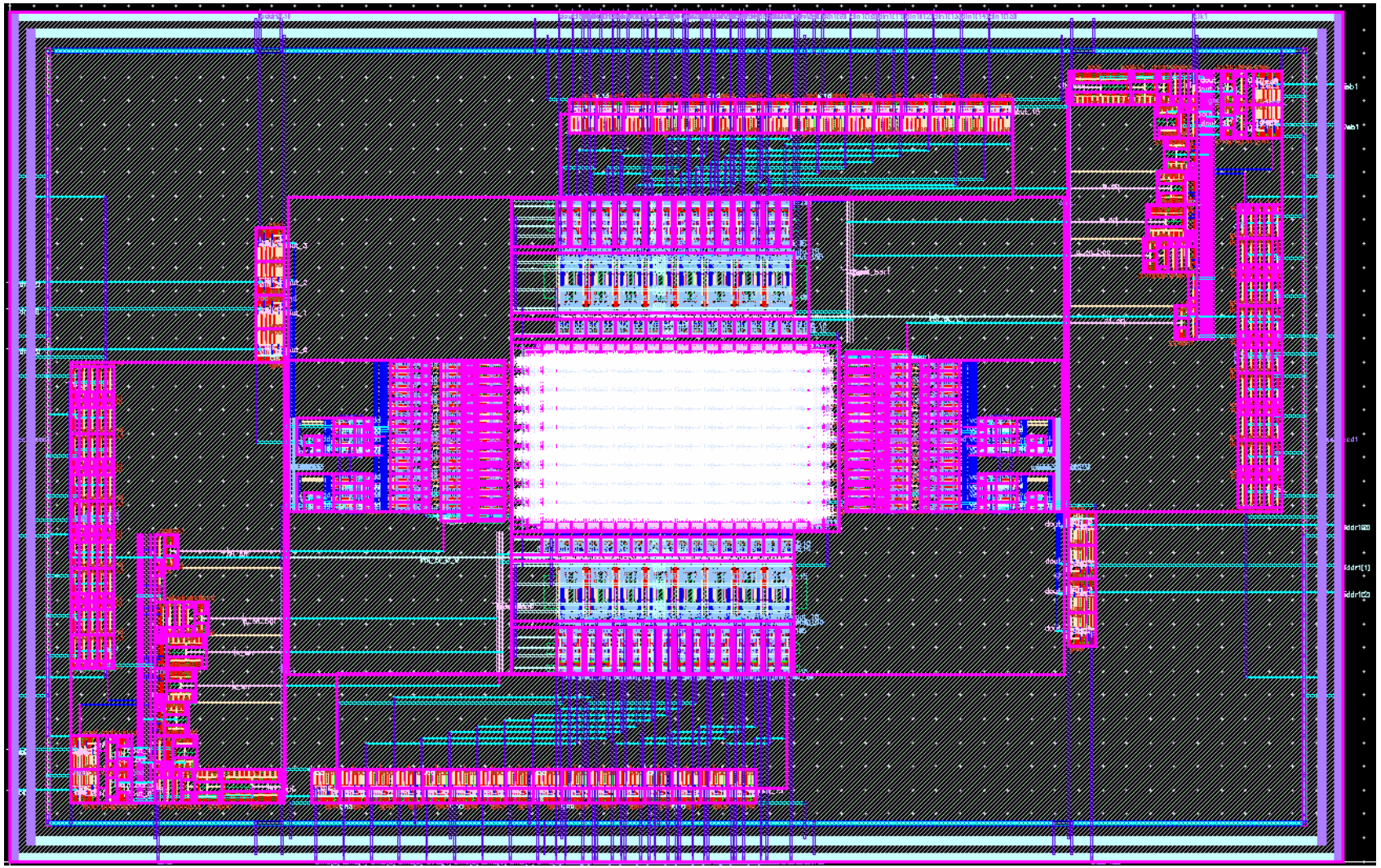
- Front-end: Create Netlist and Layout, DRC & LVS check;
- Back-end: Perform simulations on these generated files



M. R. Guthaus, et al., "OpenRAM: An Open-Source Memory Compiler", ICCAD, 2016

OpenRAM SRAM compiler

- Example: 16x16 SRAM macro (file:///E:/OneDrive%20-%20Stanford/CNT-ITO/OpenRAM/sram_16x2.html)



M. R. Guthaus, et al., “OpenRAM: An Open-Source Memory Compiler”, ICCAD, 2016

OpenRAM SRAM compiler

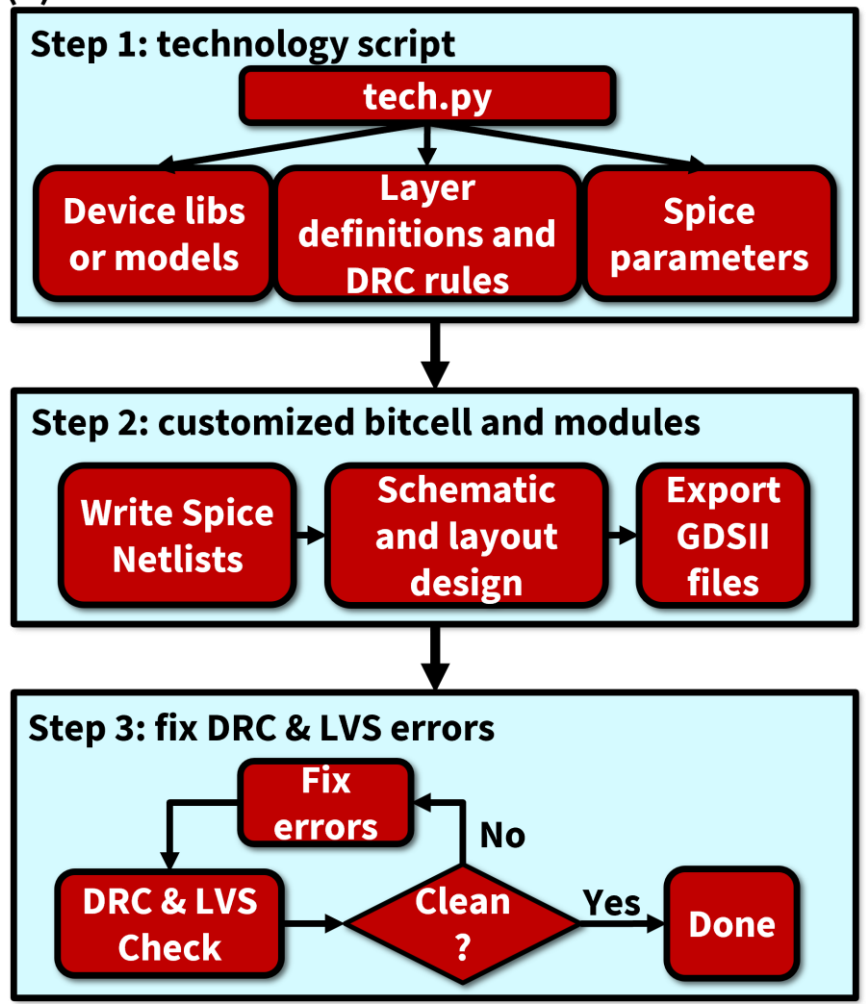
- Limitations:
  - Support NCSU FreePDK 45nm, MOSIS 0.35um (SCN4M_SUBM), Skywater 130nm (sky130), no advanced tech nodes
  - Only SRAM is supported
  - Only one macro architecture is supported

M. R. Guthaus, et al., “OpenRAM: An Open-Source Memory Compiler”, ICCAD, 2016

Methodology to extend OpenRAM functionality

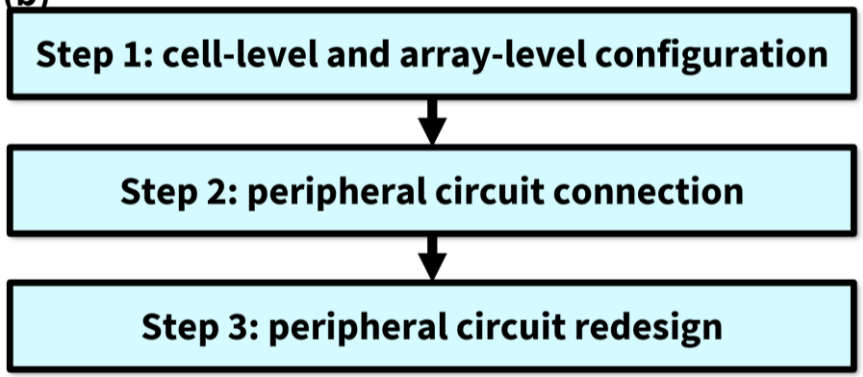
- Step 1: Porting to new PDKs

(a)

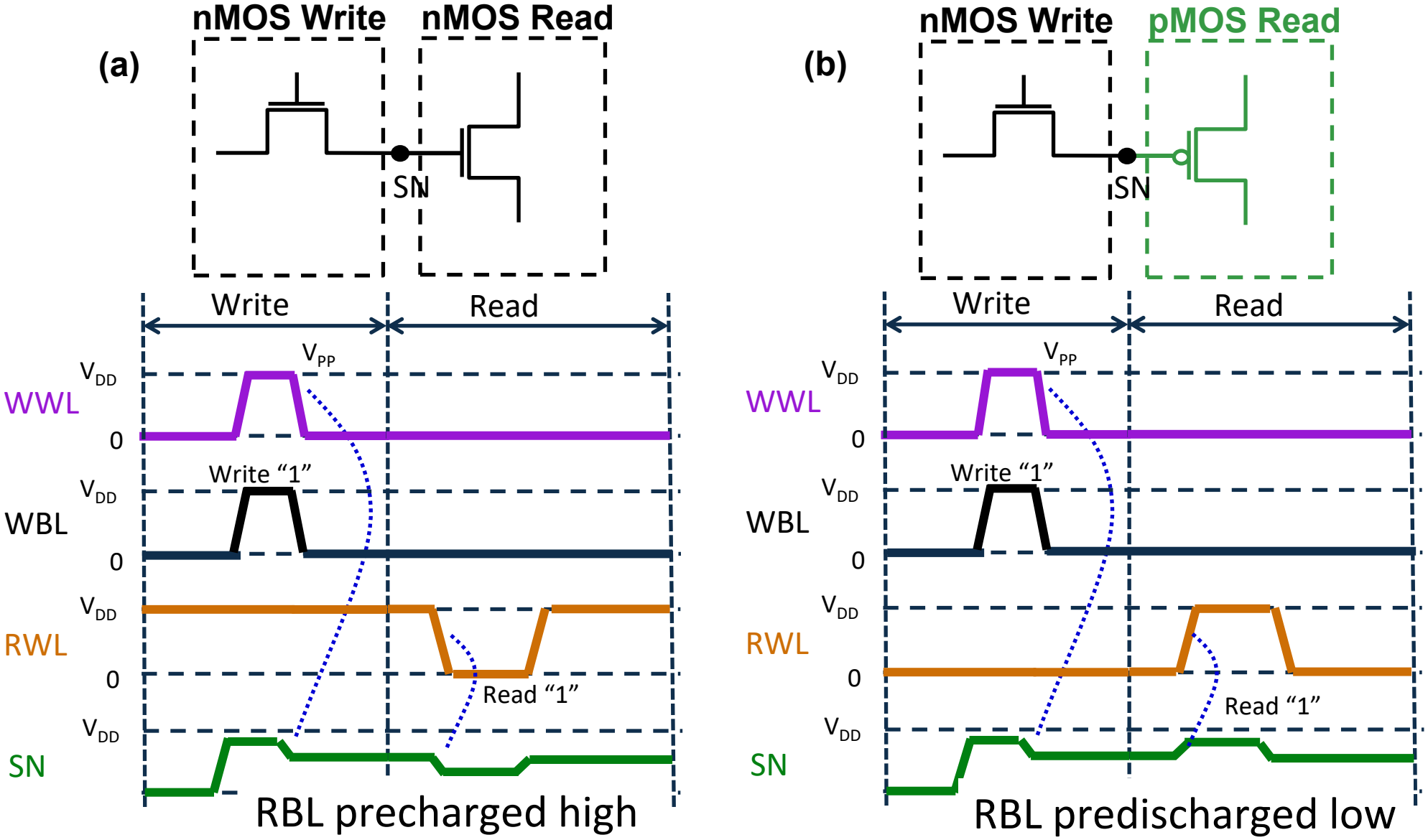


- Step 2: Adding new memory technologies

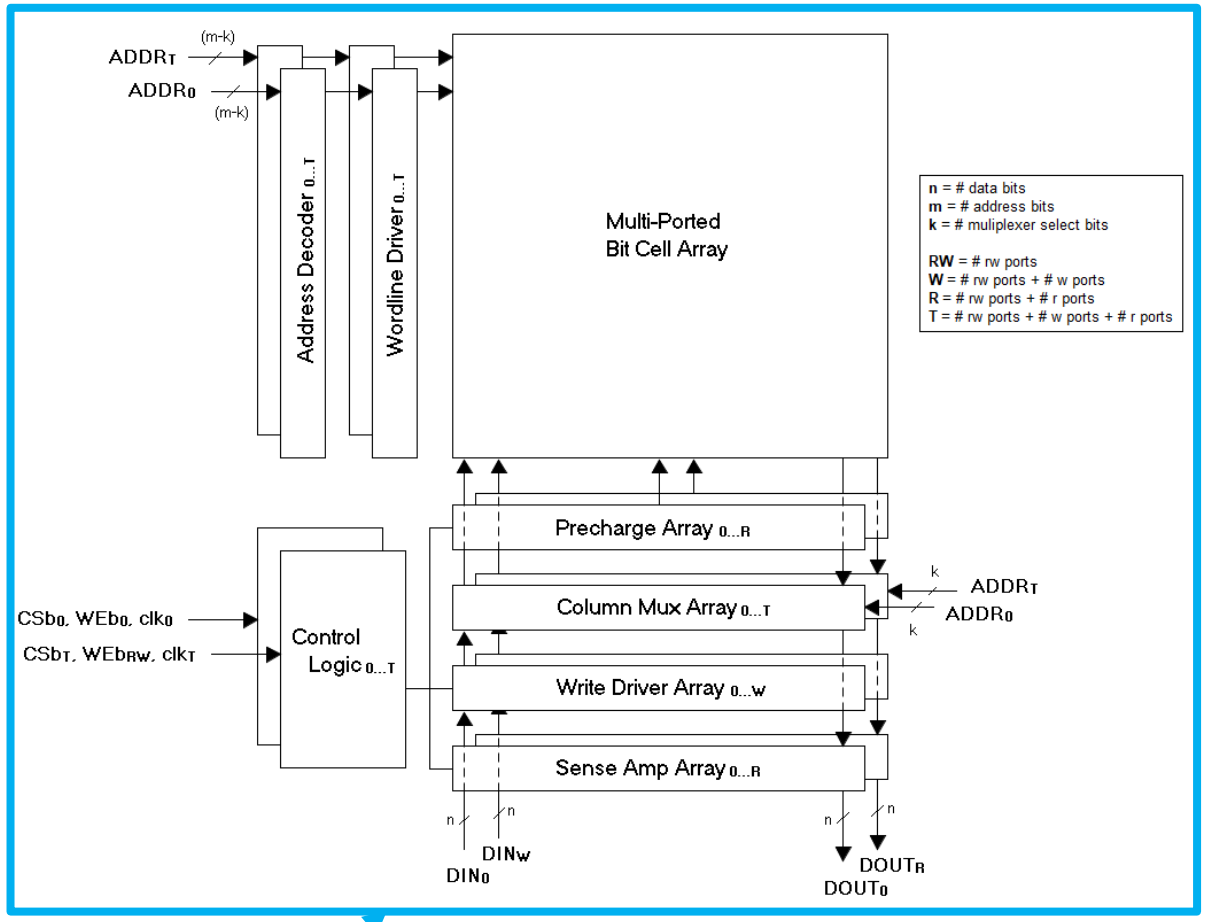
(b)

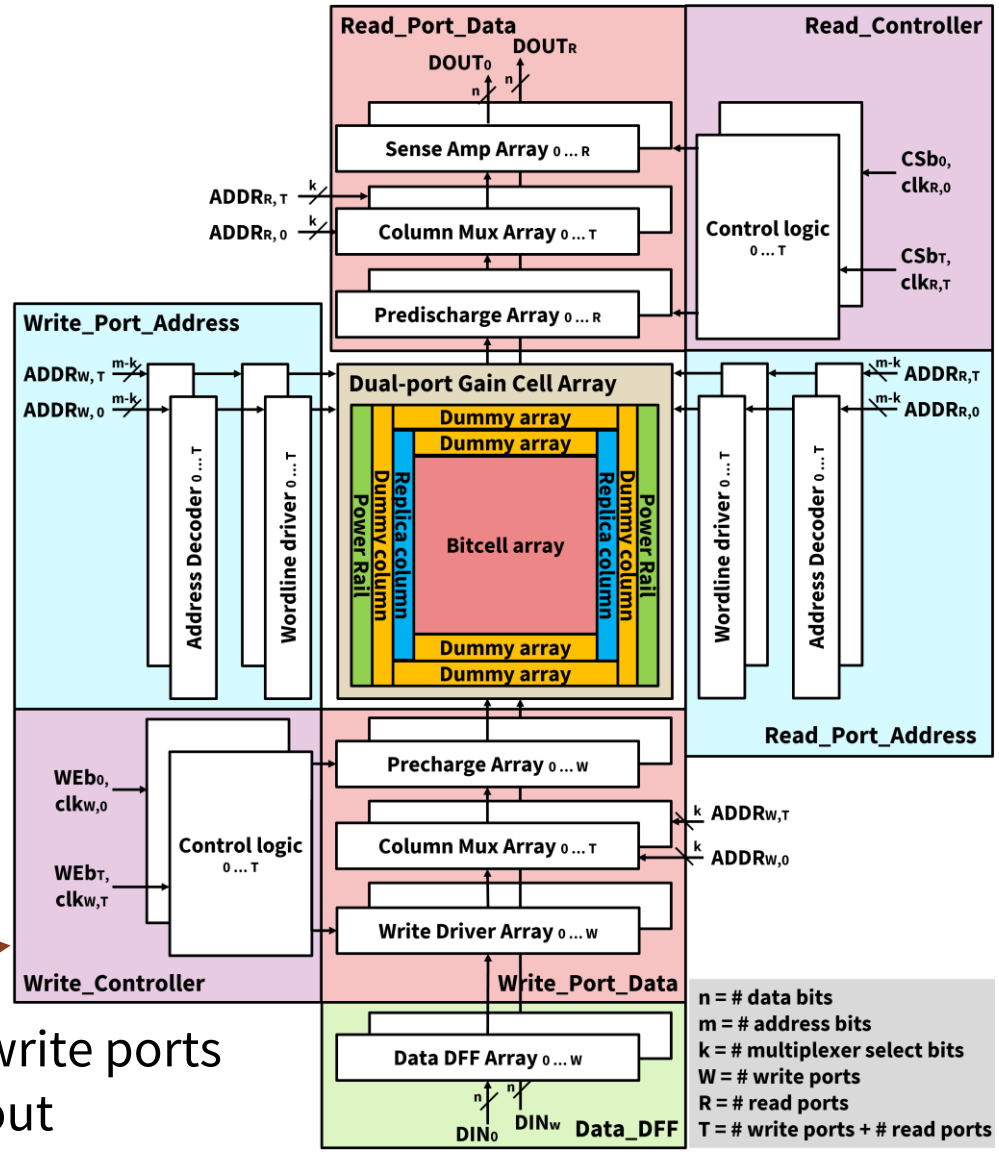


Gain Cell memory operations



OpenGCRAM: bank architecture



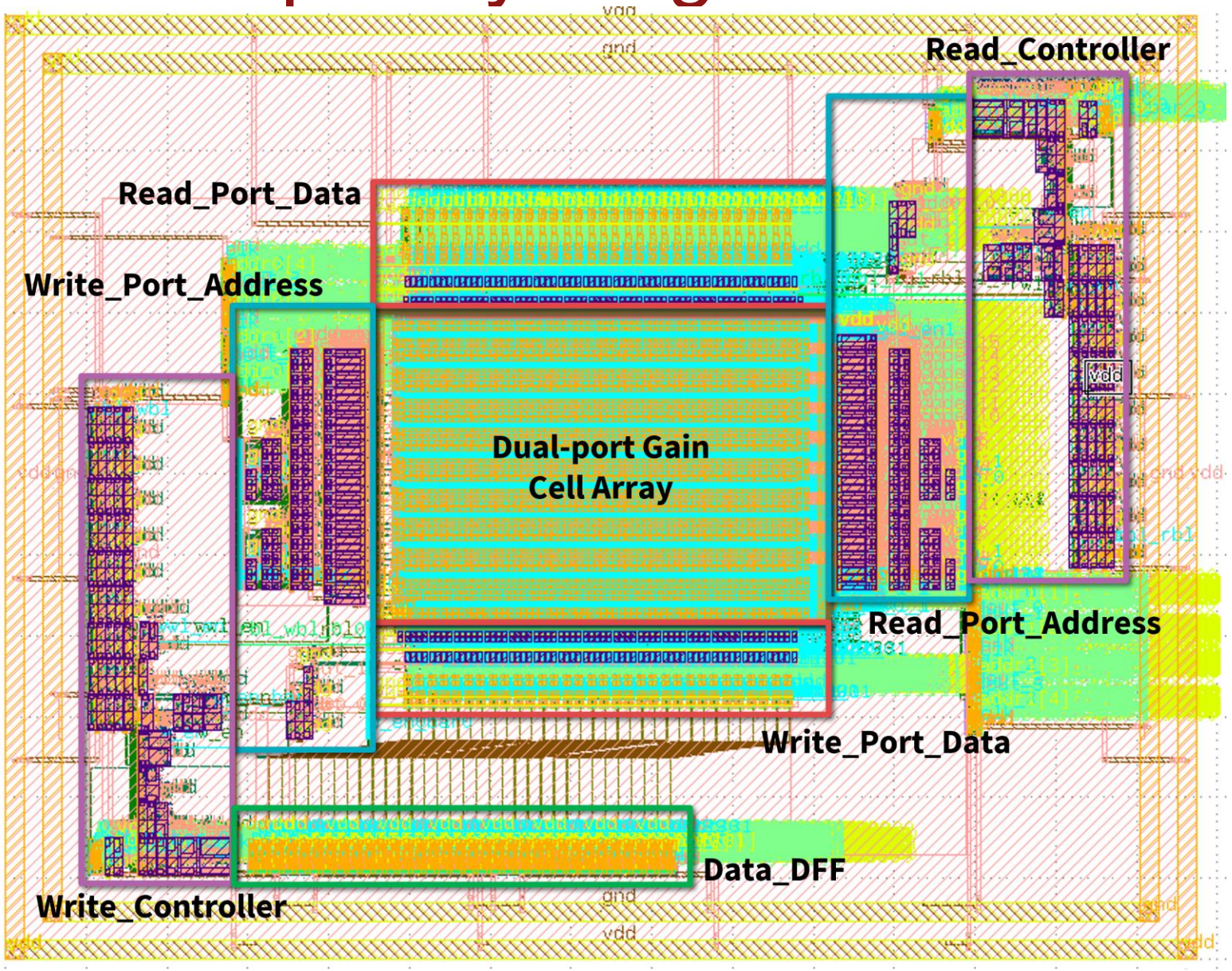


SRAM bank → **Gain Cell bank**

- separate read and write ports
- single-ended read out
- predischarge array for read port

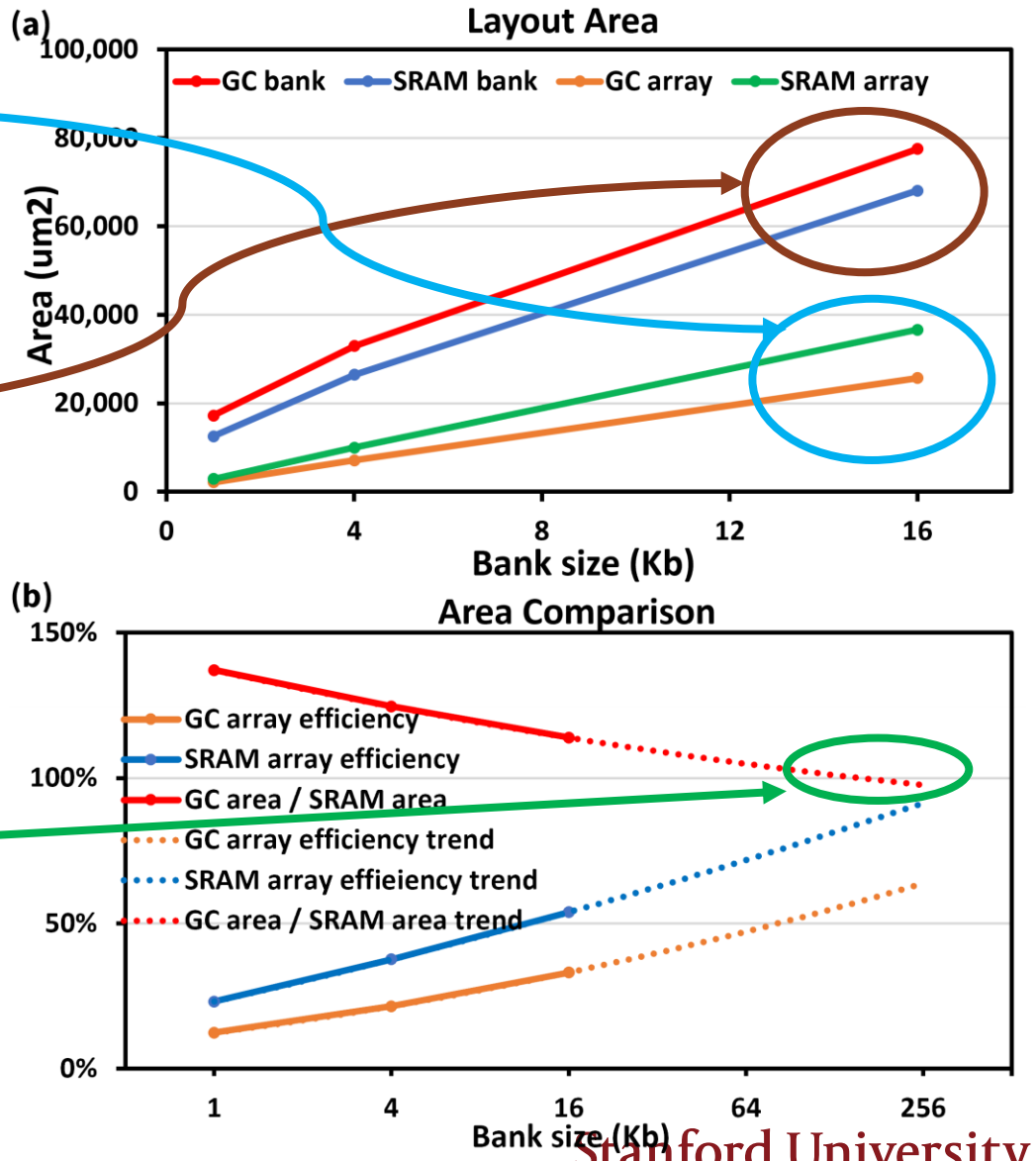
OpenGCRAM: Example layout generation

A 32x32 Gain Cell bank generated by OpenGC



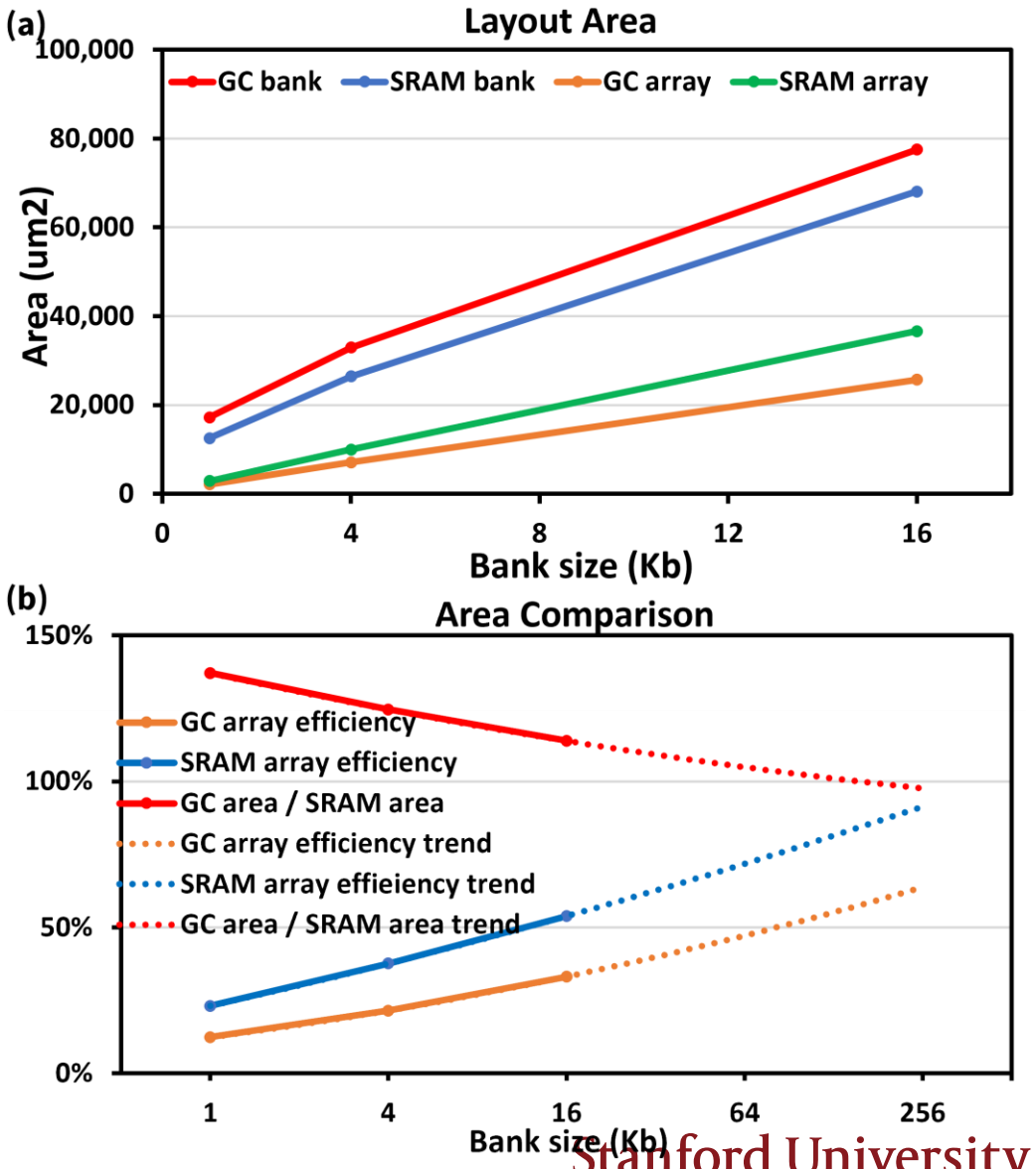
Gain Cell vs. SRAM: area

- GC array area < SRAM array area
  → GC cell size < SRAM cell size
- GC bank area > SRAM bank area for small bank size
  → GC has separate peripheral circuits for different ports
- GC bank area < SRAM bank area for large bank size
  → Peripheral circuits are amortized



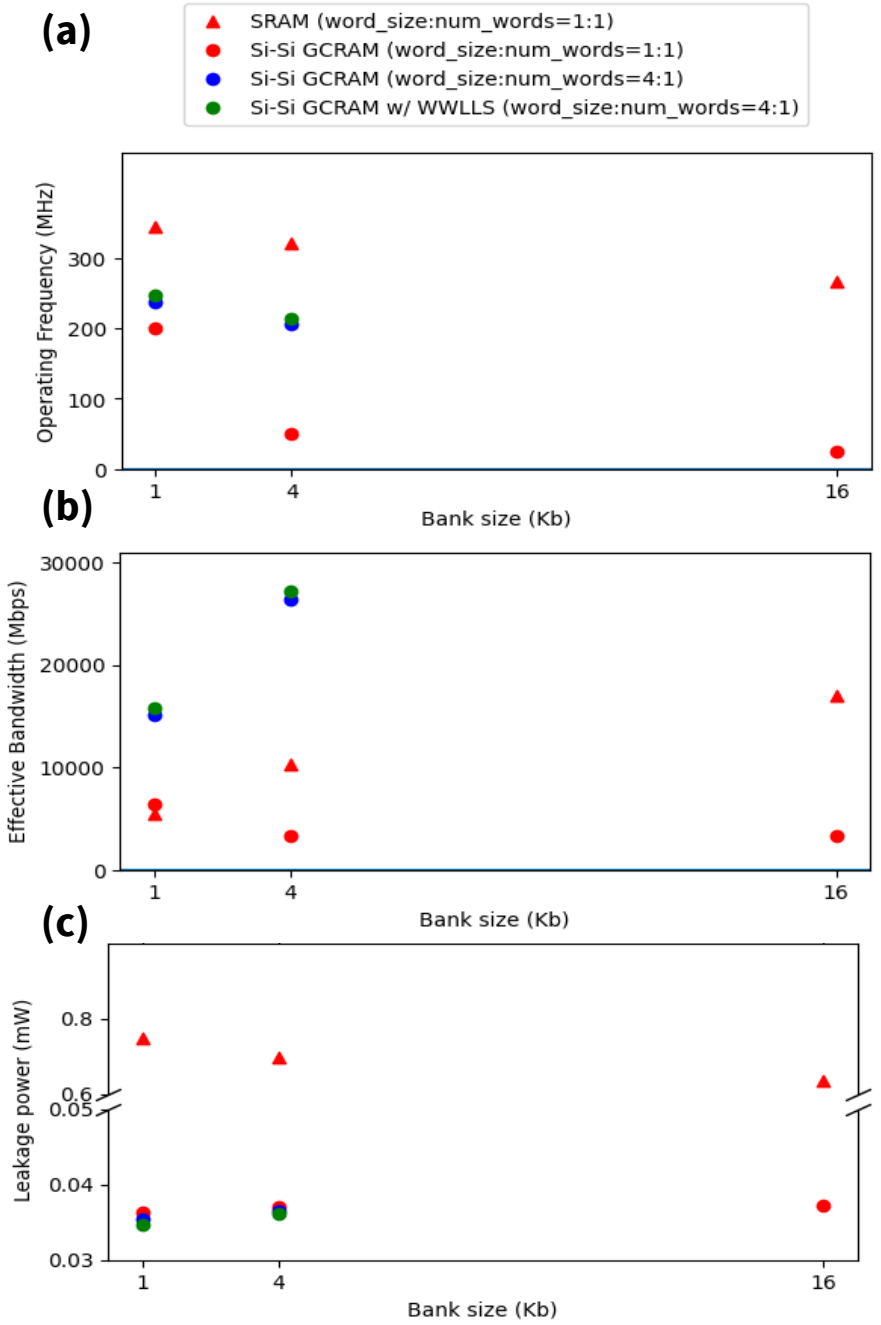
Gain Cell vs. SRAM: area

- Dual-port peripheral of GC bank allows simultaneous read and write operations, resulting in high bandwidth
- Dual-port SRAM bank area is ~2x of single-port SRAM bank area, which is larger than GC bank.



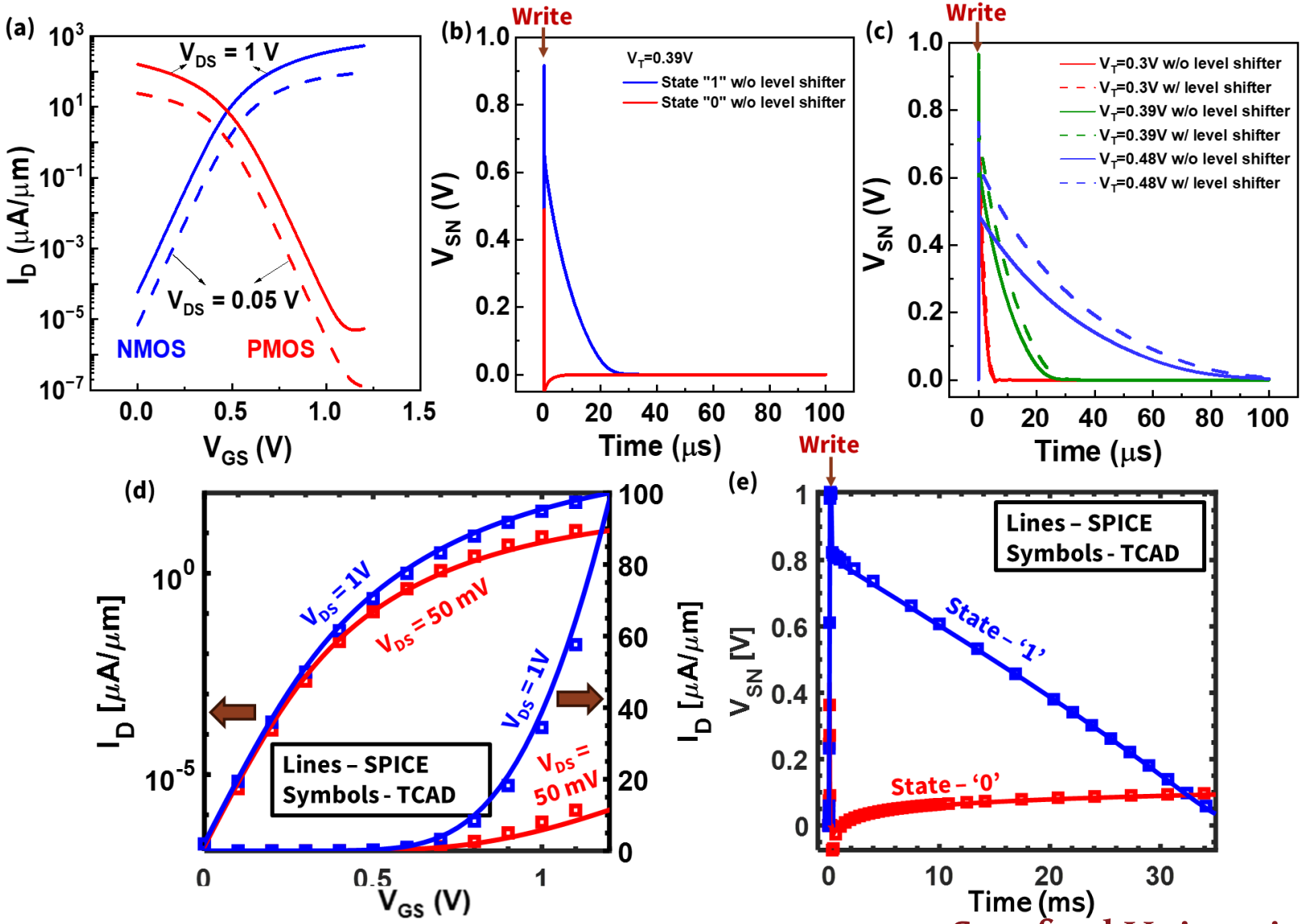
Gain Cell vs. SRAM: delay and power

- GC frequency < SRAM frequency
  - → GC uses single-ended readout
  - → GC storage node voltage is Vdd-Vth when writing "1"
  - →GC frequency can be boosted with design space exploration
- GC leakage power < SRAM leakage power
  - → There is no direct path from Vdd to GND in GC
- An abrupt change of GC freq and power from 1 Kb to 4 Kb
  - → Increase in delay chain stages



Gain Cell retention modulation

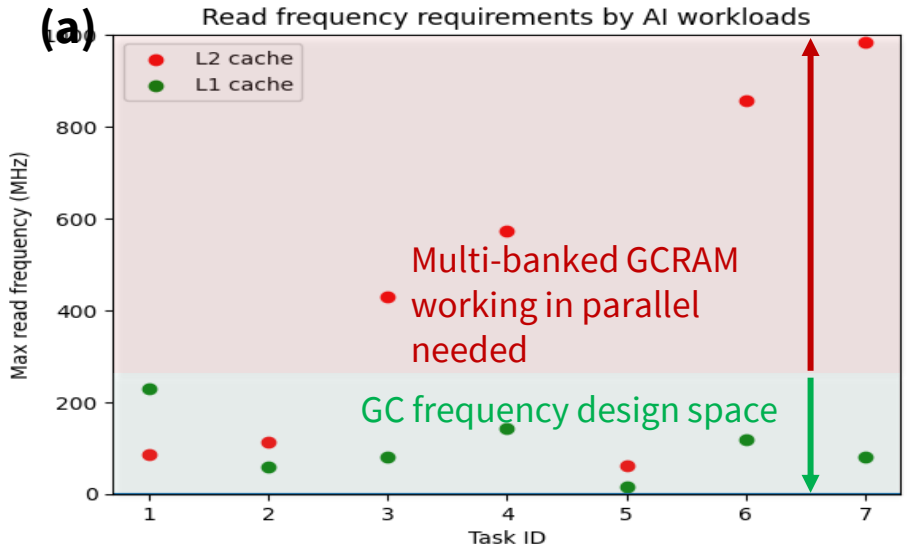
By adjusting the transistor design (like Vth and channel material), the retention can be tuned to accommodate for different applications (such as activation caches and weight memory in AI inference).

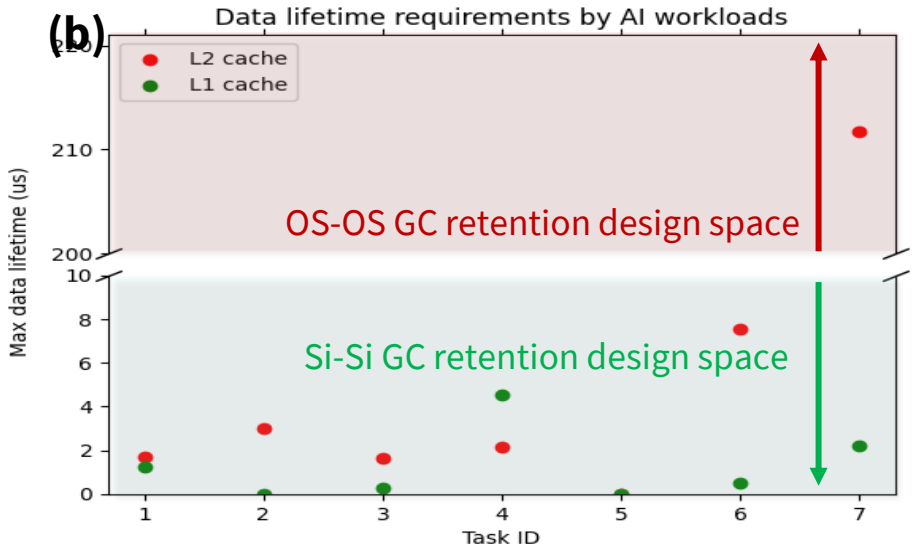


Design space exploration for AI workloads

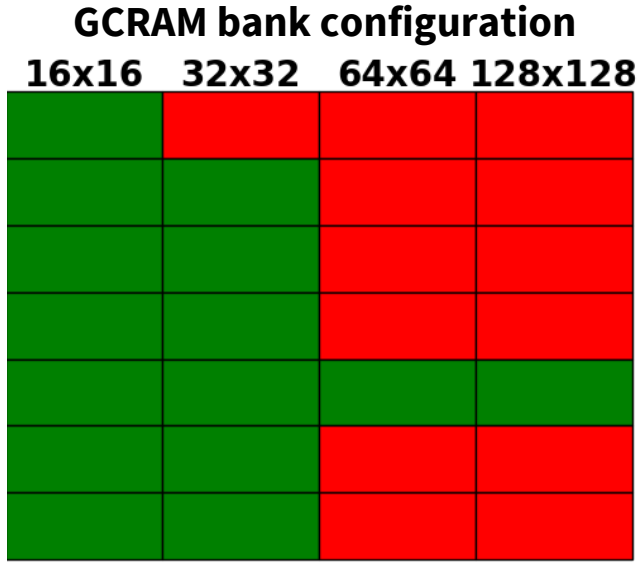
Task	L1 lifetime Max (us)	L2 lifetime Max (us)	L1 read freq (MHz)	L2 read freq (MHz)
polybench-2DConvolution	1.25	1.68	230.43	84.93
llama-3.2-11b-vision	4.50	2.14	142.82	574.17
bert-base-uncased	0.48	7.53	117.11	856.99
llama-3.2-1b	0.28	1.63	79.78	428.50
stable-diffusion	2.17	211.76	79.33	984.24
polybench-3DConvolution	0	3.01	57.66	113.18
resnet-18	0	0	16.31	62.47

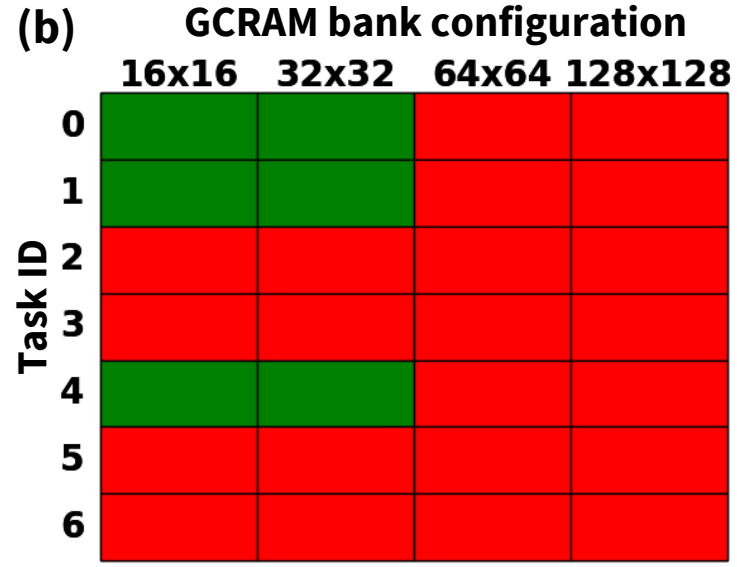
Design space exploration for AI workloads





Task ID	1	2	3	4	5	6	7
Task name	Polybench-2DConvolution	Llama-3.2-11b-vision	Bert-base-uncased	Llama-3.2-1b	Stable-diffusion	Polybench-2DConvolution	Resnet-18





Summary

- We developed OpenGC to enable fast, accurate, customizable, and optimized Gain Cell bank design as high-density on-chip memory
- We introduced a standard methodology for porting OpenRAM compiler to new PDKs and memory technologies
- OpenGC supports Gain Cell bank design generation with TSMC N40 PDK and precise Spice simulations for performance evaluation
- By following the proposed methodology, OpenGC can be extended to support additional types of Gain Cell memory and other PDKs
- Design space exploration is enabled to accommodate for different AI workloads