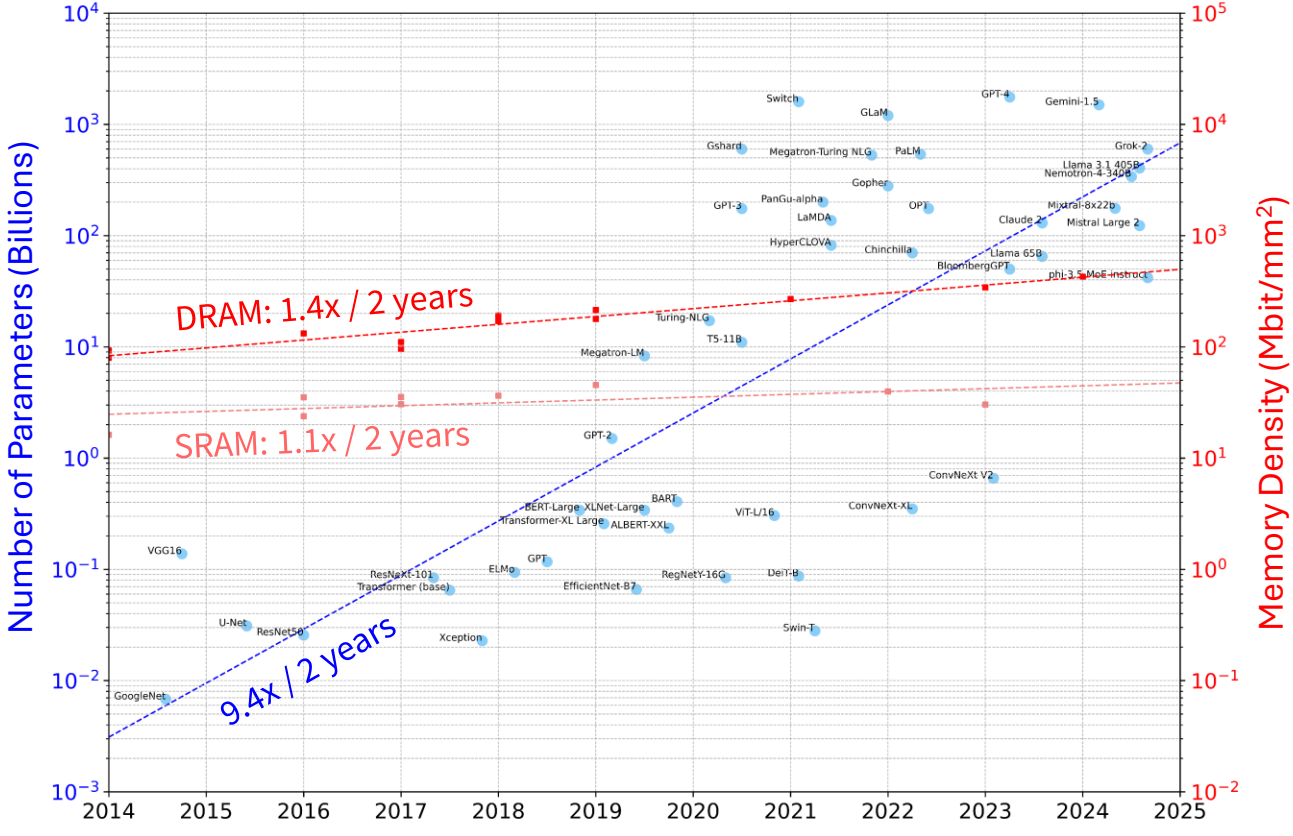
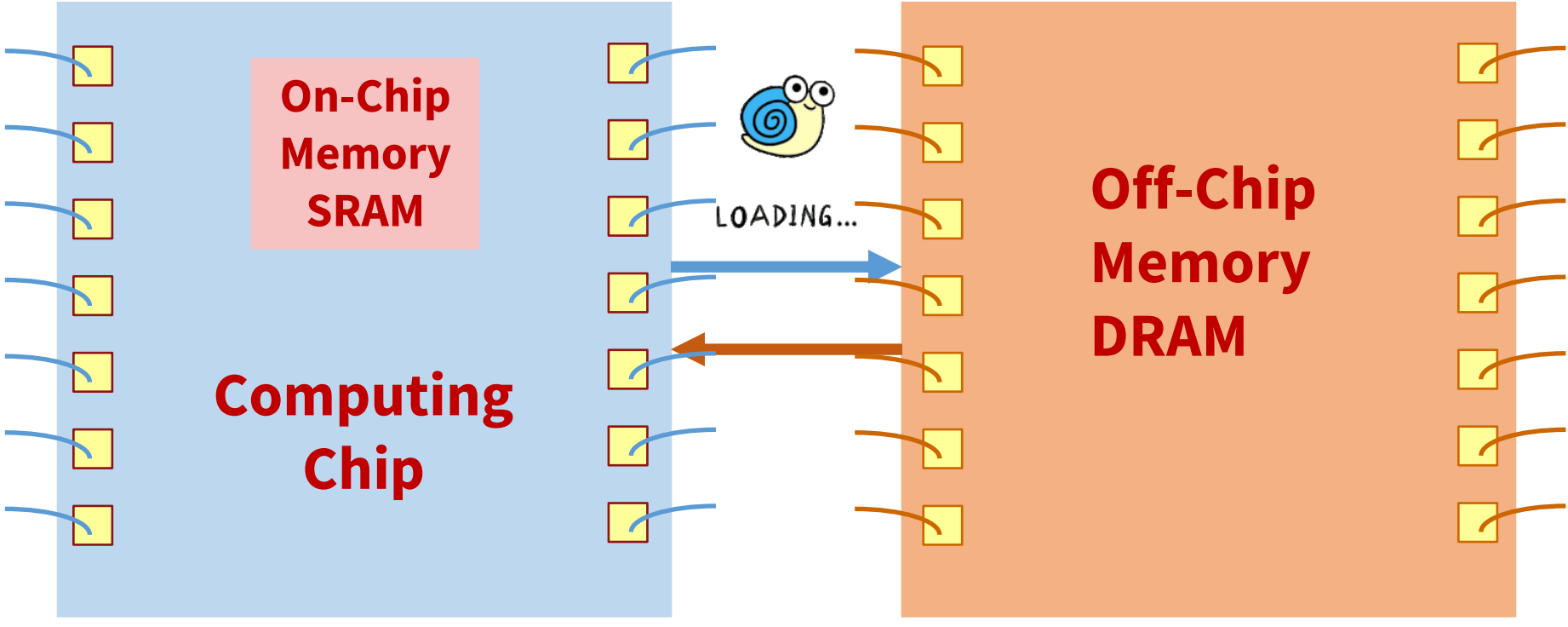
Data >> Memory



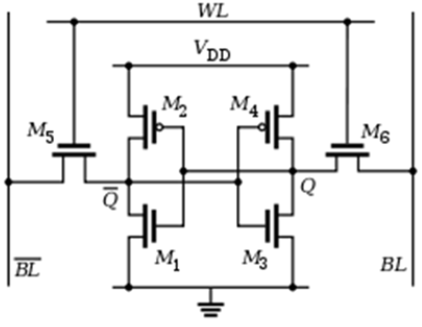
Sources: K. Akarvardar (TSMC) &
https://nano.stanford.edu/downloads/technology-integration-trend

Memory Wall Problem

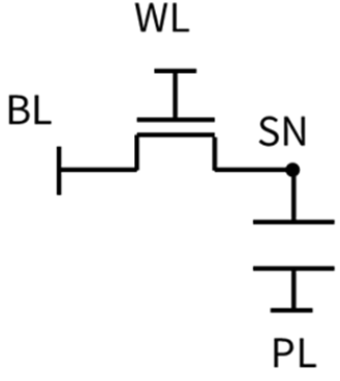


Why DRAM off-chip?

6-Transistor SRAM



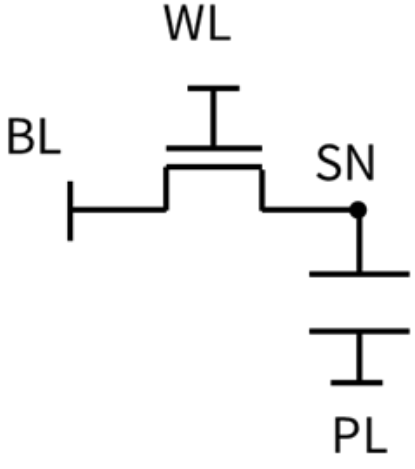
1 Transistor 1 Capacitor DRAM



Capacitor process not compatible

Smaller capacitor?

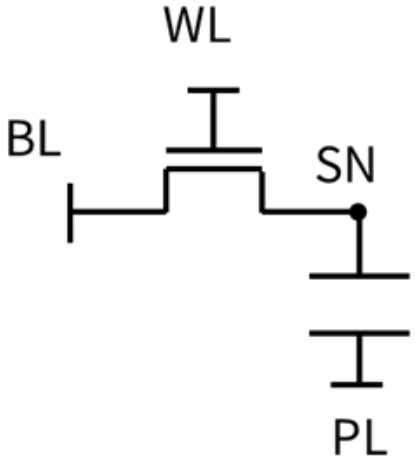
1T1C DRAM



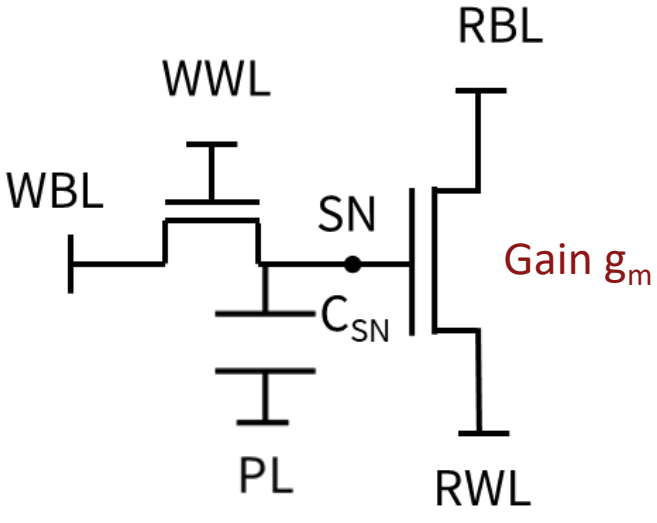
too little charge to read out

Gain: read transistor amplification

1T1C DRAM



2T1C Gain Cell

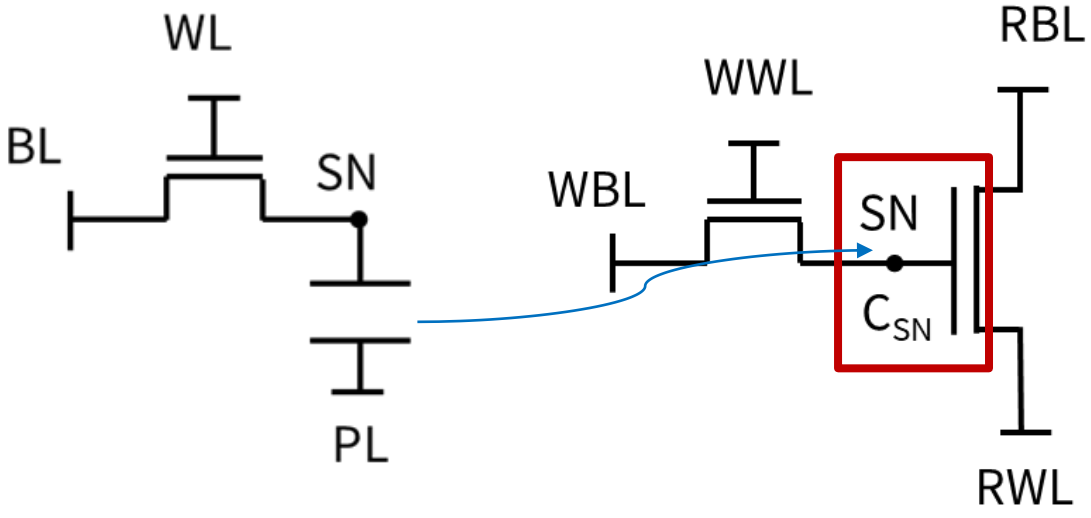


too little charge to read out → Add an "amplifier"

Gain Cell: SRAM-like pseudo-DRAM

1T1C DRAM

2T Gain Cell

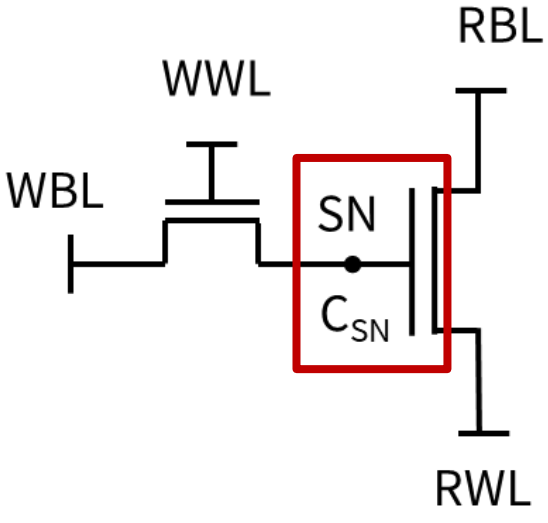


Storage capacitor → Gate capacitance of read transistor

Shukuri, Kure and Nishida, IEDM 1992

Another Problem of Small Capacitance

2T Gain Cell

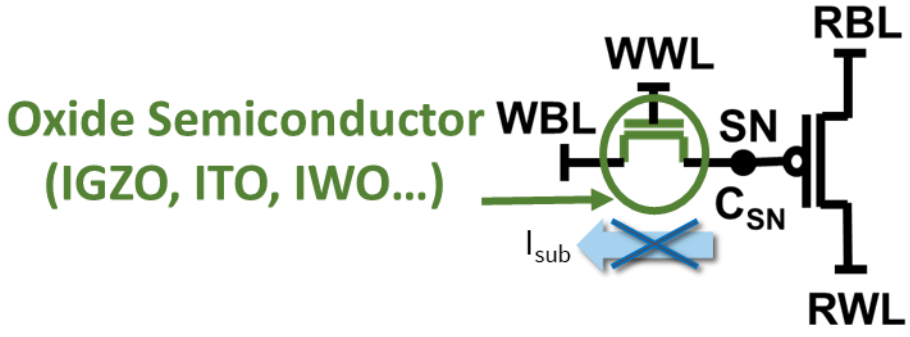


Gate capacitance of read transistor

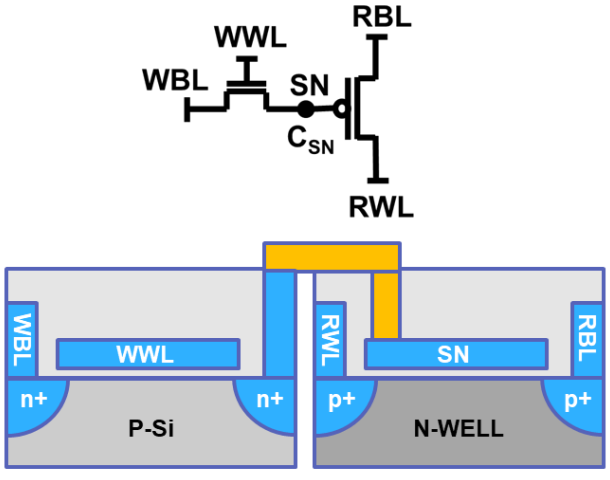
$$t = \frac{Q}{I}$$

Low Retention

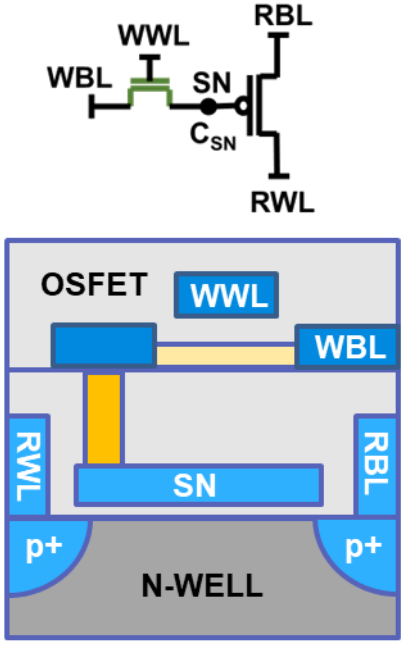
Oxide Semiconductor: Low Leakage



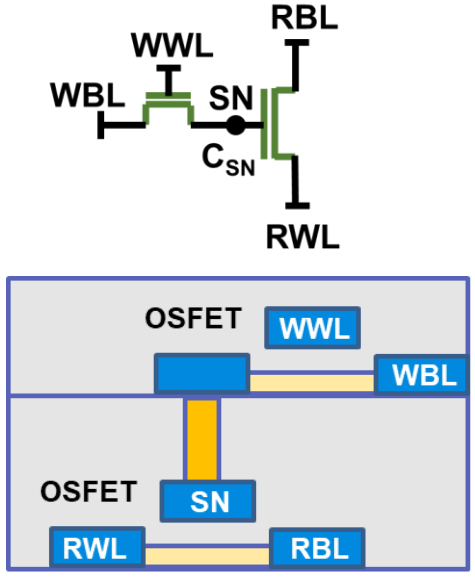
Gain Cell Family



Si-Si gain cell



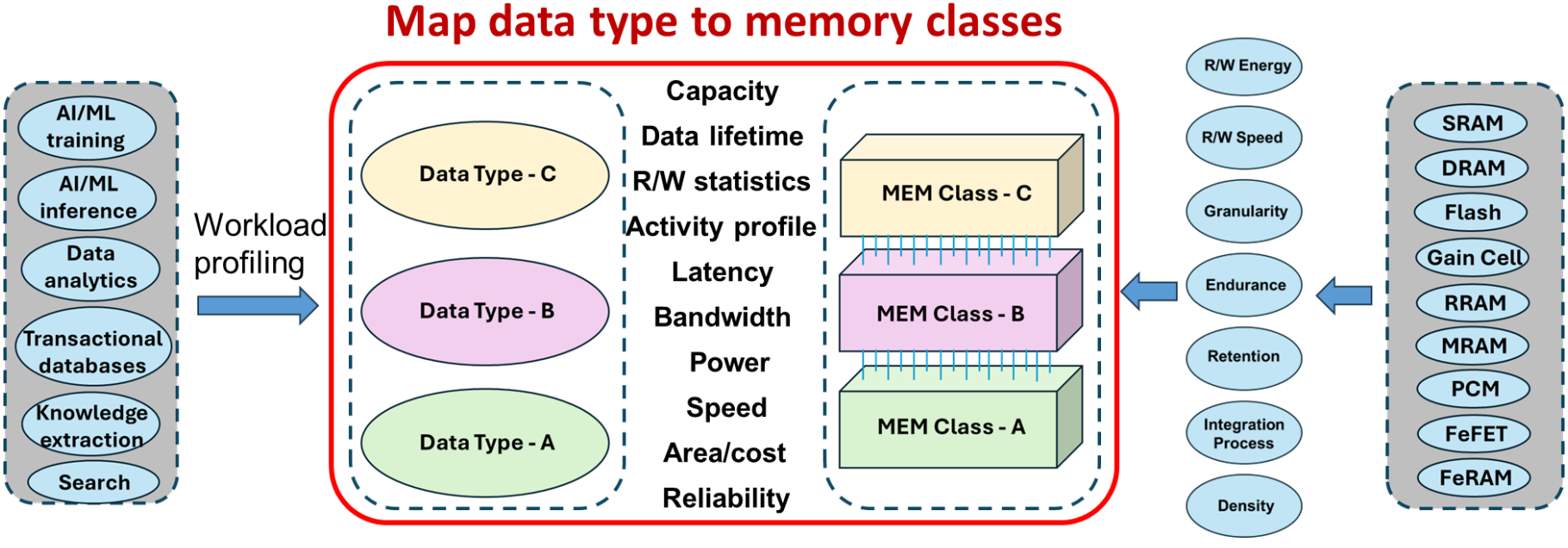
Hybrid (OS-Si) gain cell



OS-OS gain cell

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

Expose Hardware to Software



Shuhan Liu, …, H.-S. Philip Wong, "Future of Memory: Massive, Diverse, Tightly Integrated with Compute – from Device to Software", IEDM 2024

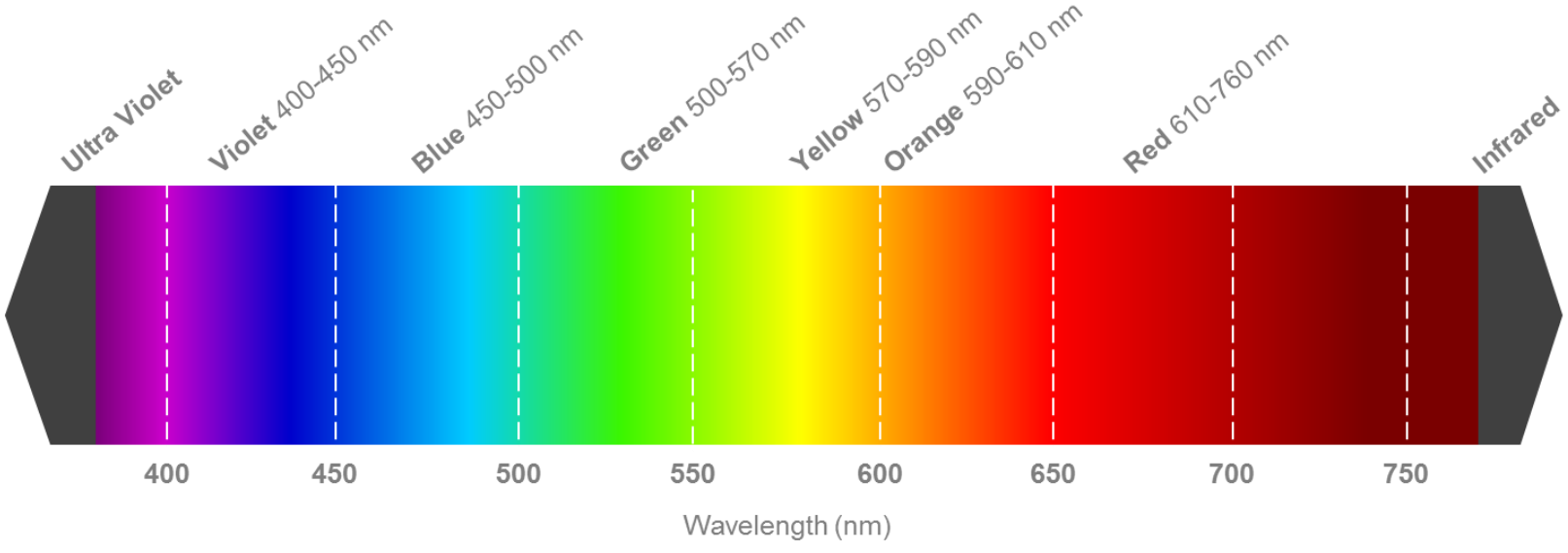
Memory Table?

	Read	Write	Standby including refresh	Standby excluding refresh	Write Endurance	Retention
SRAM	1	1	0.1	0.1	infinite	N/A
DRAM	50	50	0.12	0.05	infinite	64 ms
Si Gain Cell	2	2	0.2	0.05	infinite	10 us
Oxide Gain Cell	10	10	6	5	infinite	10 s
RRAM	20	1000	0	0	10^5	10 y
MRAM	10	1000	0	0	10^{10}	10 y
FeRAM	100	100	0	0	10^{15}	10 y
FeFET	10	100	0	0	10^5	10 y

Can span wide range with trade-off engineering

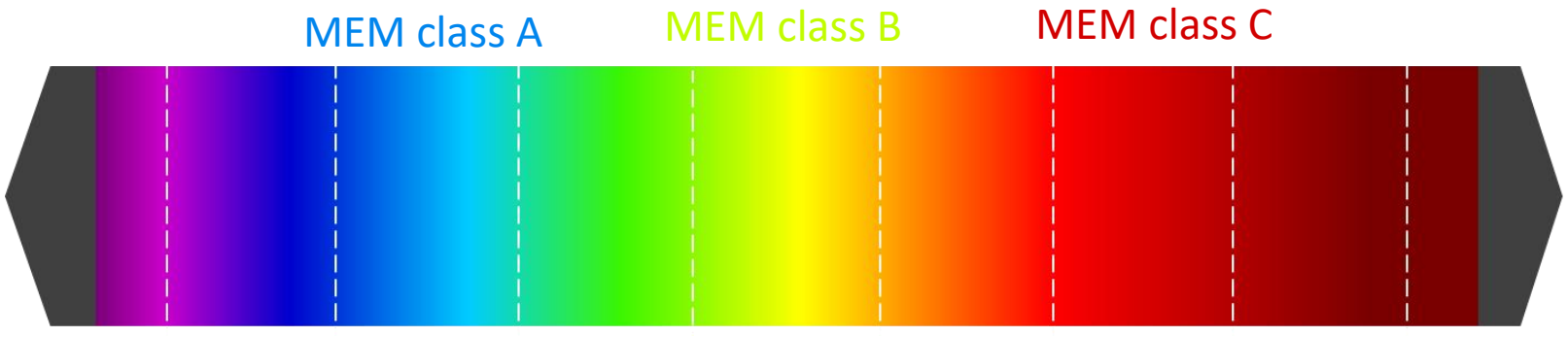
	Read	Write	Standby including refresh	Standby excluding refresh	Write Endurance	Retention
SRAM	1	1	0.1	0.1	infinite	N/A
DRAM	50	50	0.12	0.05	infinite	64 ms
Si Gain Cell	2	2	0.2	0.05	infinite	10 us
Oxide Gain Cell	10	10	0	5	infinite	10 s
RRAM	20	1000	0	0	10^5	10 y
MRAM	10	1000	0	0	10^{10}	10 y
FeRAM	100	100	0	0	10^{15}	10 y
FeFET	10	100	0	0	10^5	10 y

Light Spectrum



- Continuous spectrum
- Indexed by wavelength
- Group into color band

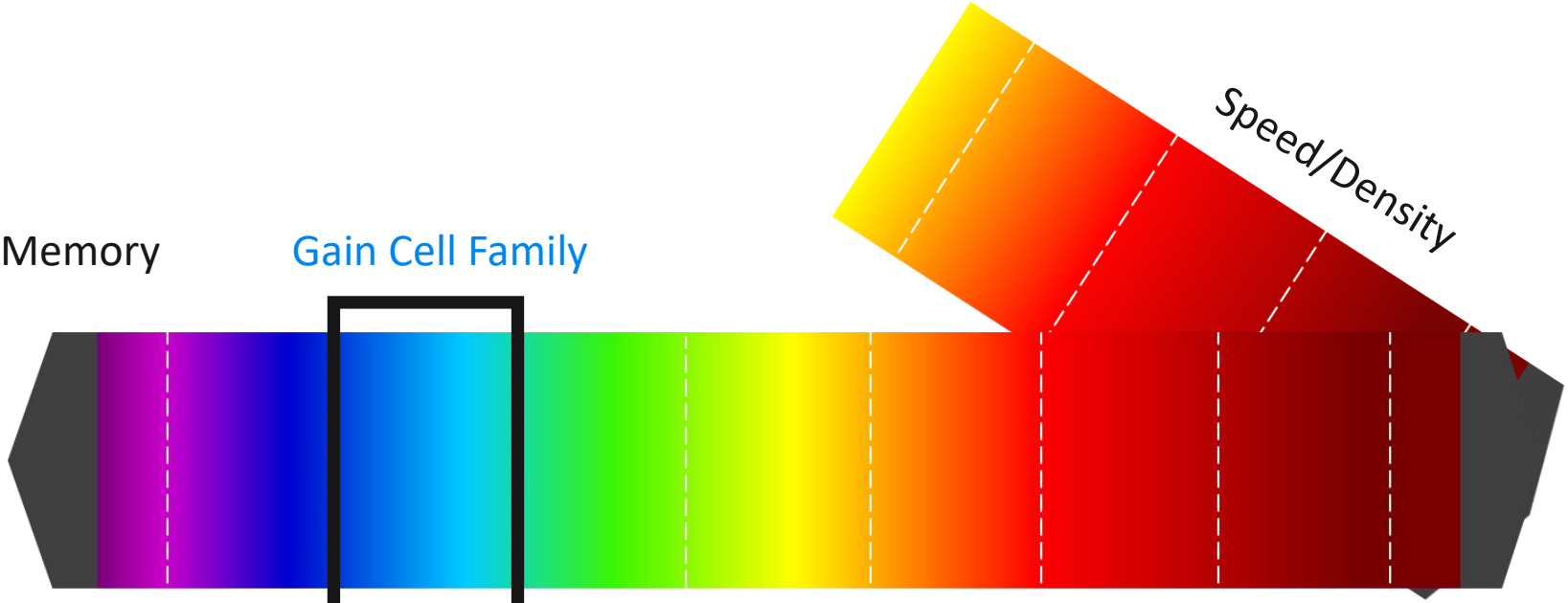
Memory Spectrum



Index: Attribute (retention, endurance, speed etc.)

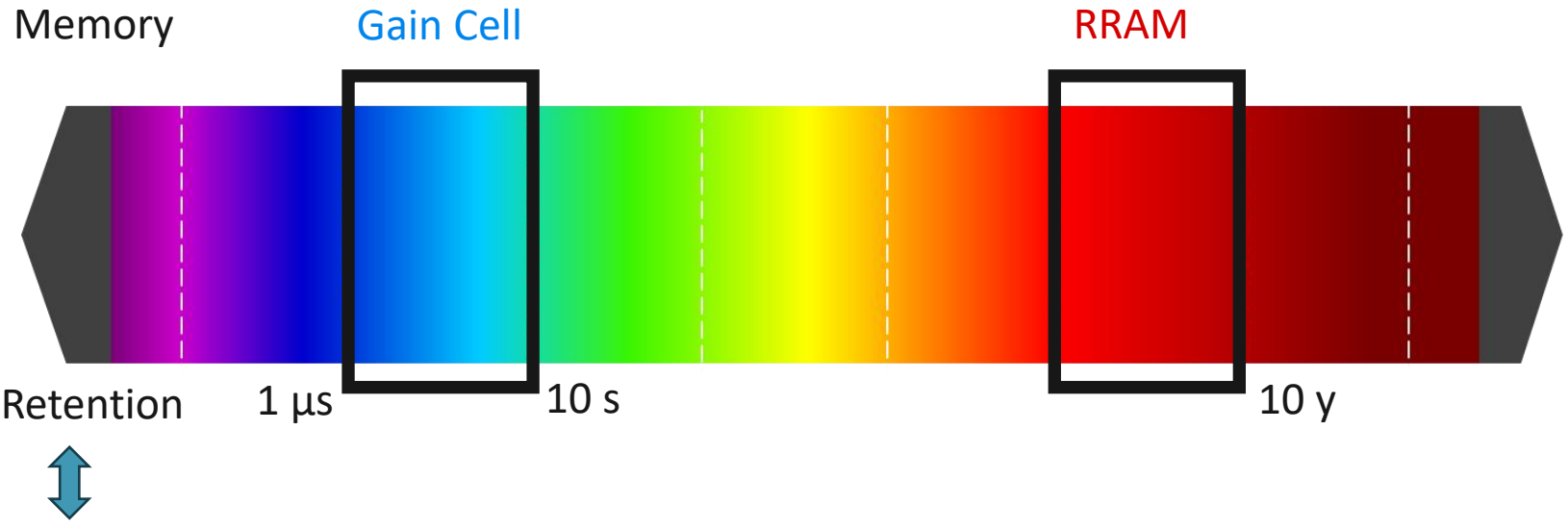
- Continuous spectrum
- Indexed by attribute(s)
- Group into memory class

BRIDGE
Blended Retention-Indexed Diverse Gain cEll



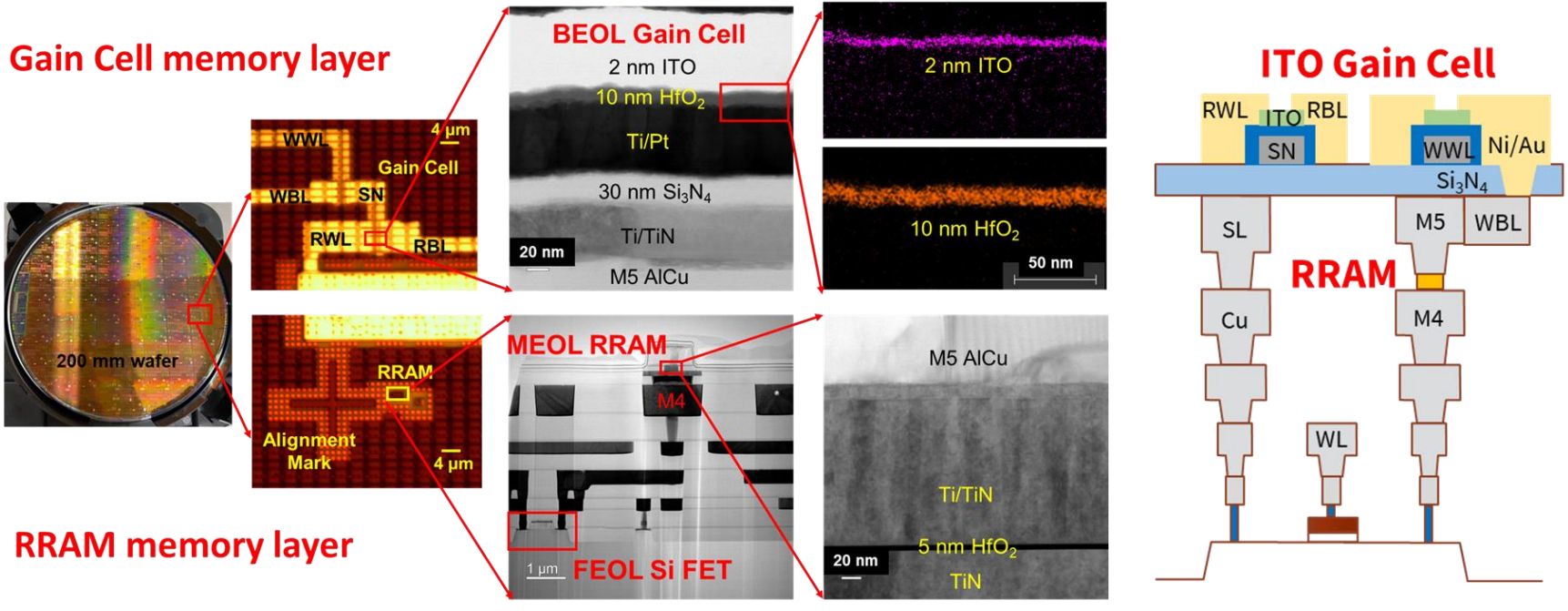
Data lifetime from software workload profiling

Longer Retention? -> Next Band



Data lifetime from software workload profiling

Diverse Memory Integration on CMOS chip

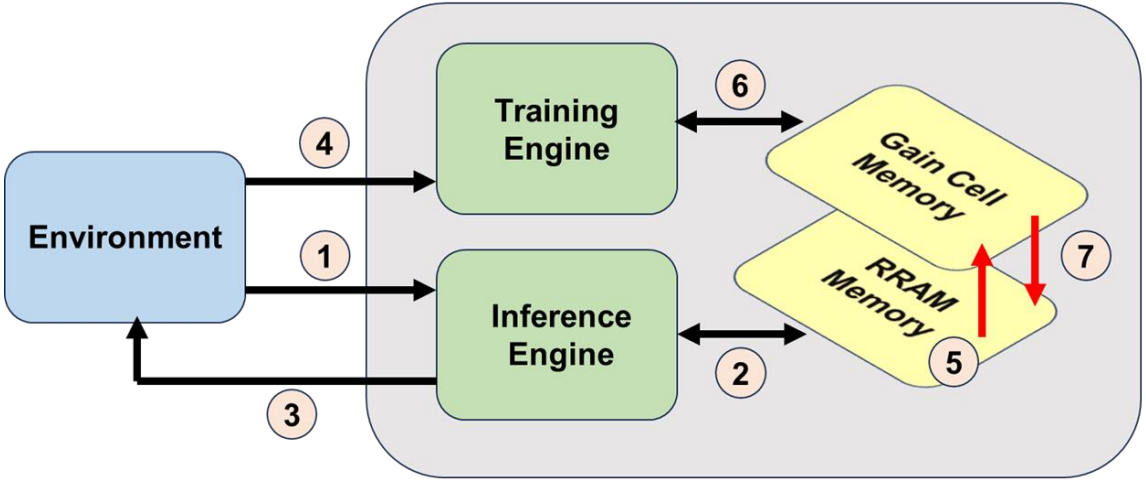


ITO: Indium Tin Oxide

Shuhan Liu, …, H.-S. Philip Wong, IEDM 2024 (best student paper award)

Gain Cell for Training; RRAM for Inference

- ✓ Adaptive to Environment Interaction
- ✓ Split Memory for Training and Inference
- ✓ High-Bandwidth 3D In-Memory-Macro Transfer



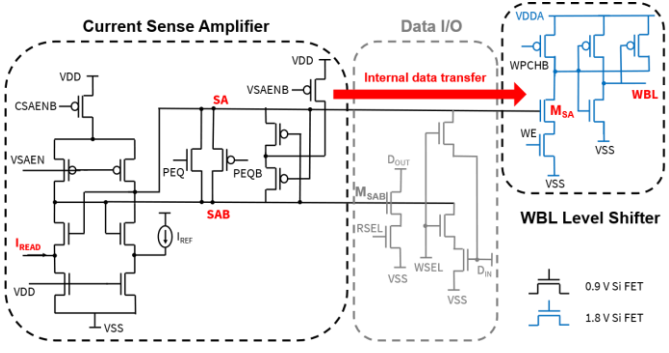
0. Training base model
1. Inference input
2. Inferring with inference memory
3. Inference output
4. Fine-tuning input based on 3
5. Weight transfer
6. Fine-tuning with training memory
7. Weight transfer

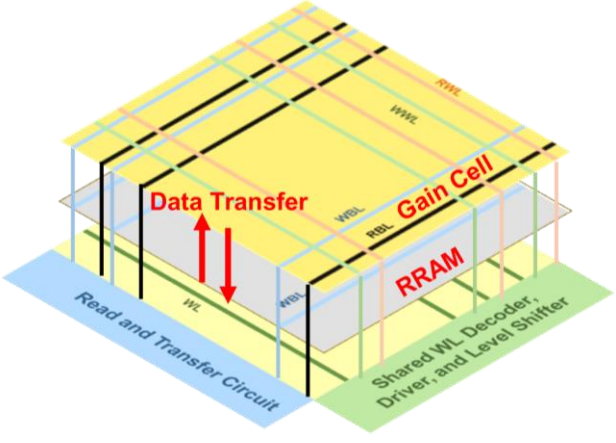
Shuhan Liu, …, H.-S. Philip Wong, IEDM 2024 (best student paper award)

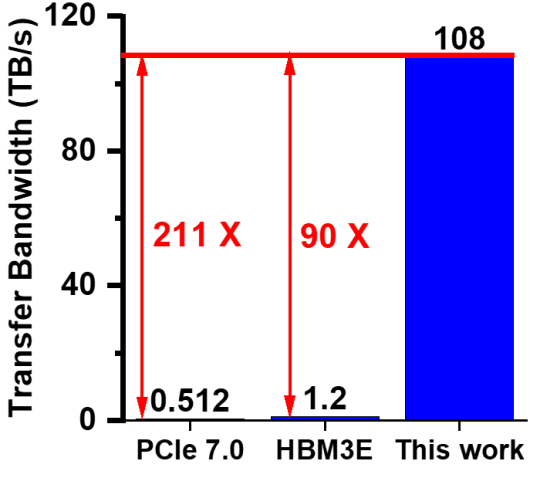
Workload Switch → Data Type Change → Data Transfer

High Bandwidth Data Transfer

- In-memory macro transfer circuit
- Parallel M3D via connection



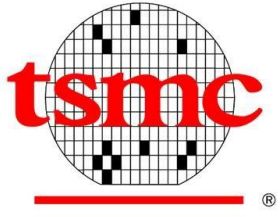




Shuhan Liu, …, H.-S. Philip Wong, IEDM 2024 (best student paper award)

Acknowledgments


Semiconductor Research Corporation



DAM
Stanford Differentiated Access Memories Project

Stanford University

Stanford | NMTRI
NON-VOLATILE MEMORY TECHNOLOGY RESEARCH INITIATIVE