Stanford Differentiated Access Memories Project

Future of Memory: Massive, Diverse, Tightly Integrated with Compute – from Device to Software

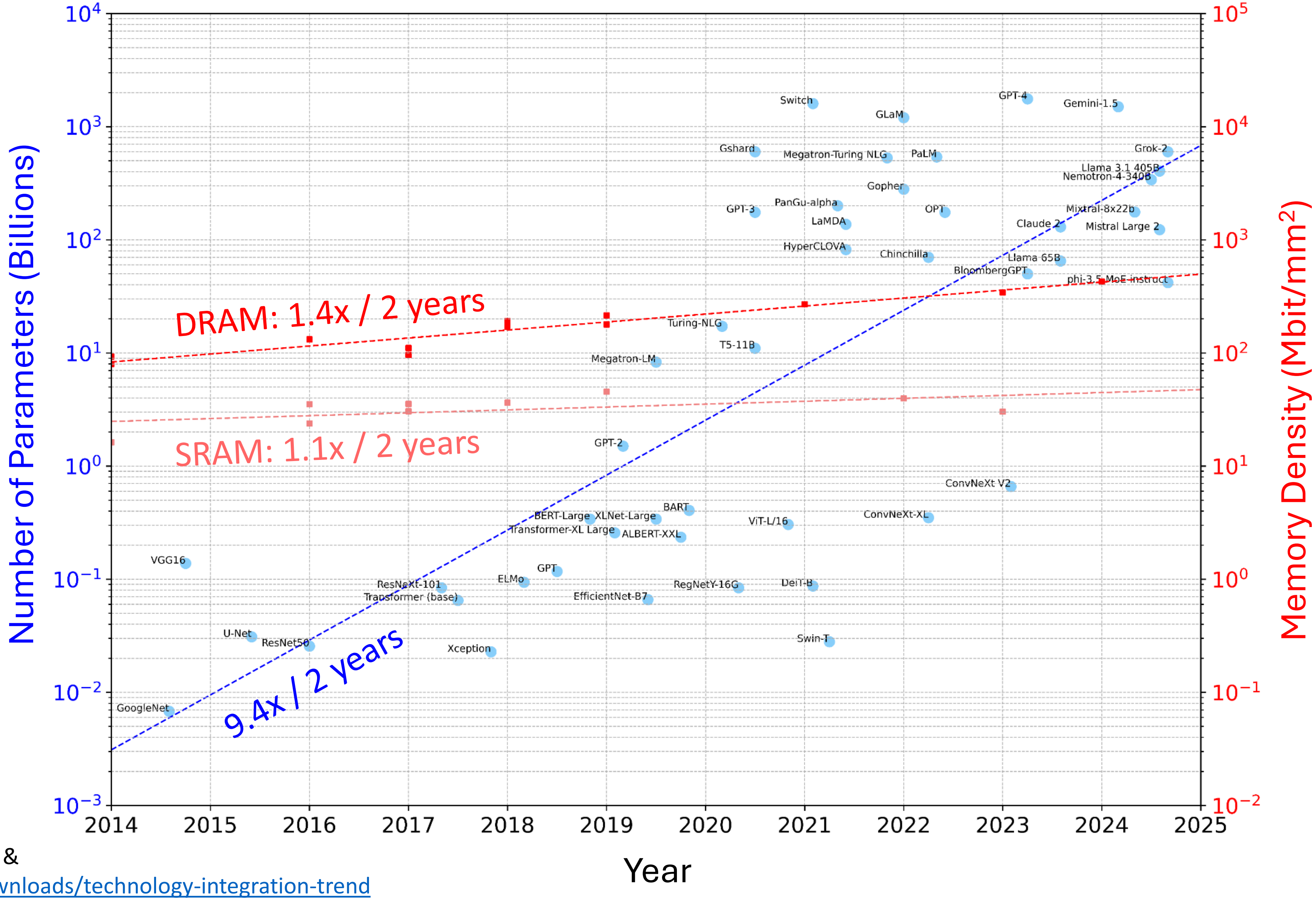
Shuhan Liu[1]*, Robert M. Radway[1], Xinxin Wang[1], Jimin Kwon[1],

Caroline Trippel[2], Philip Levis[2], Subhasish Mitra[1,2], H.-S. Philip Wong[1]*

[1]Department of EE, [2]Department of CS, Stanford University, CA, USA.

(*E-mail: shliu98@stanford.edu, hspwong@stanford.edu)

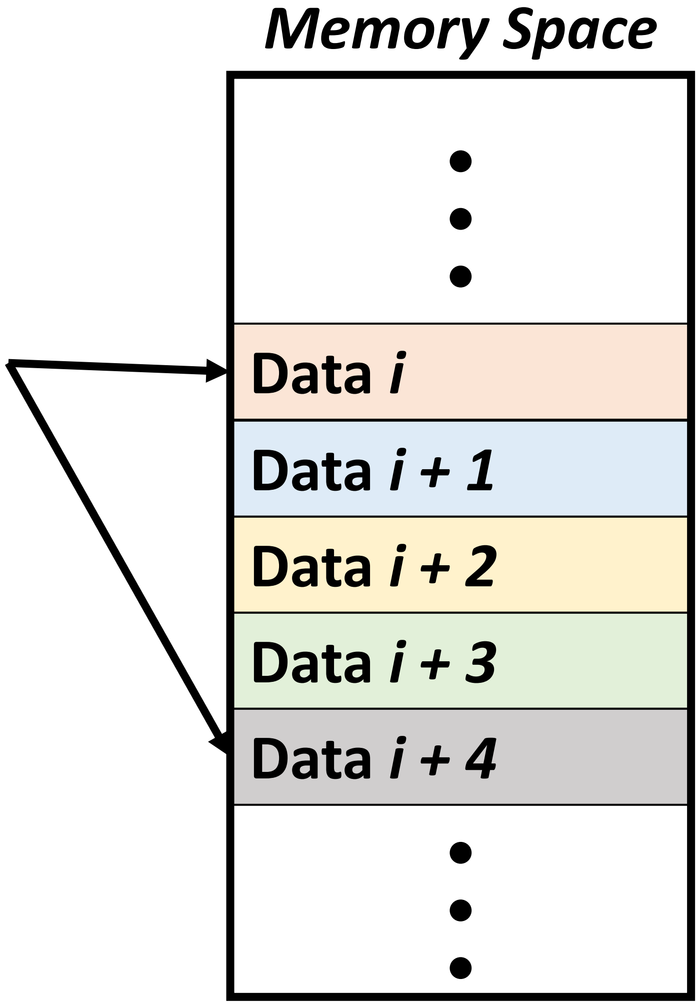
Memory Needs Outpace Memory Advances



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

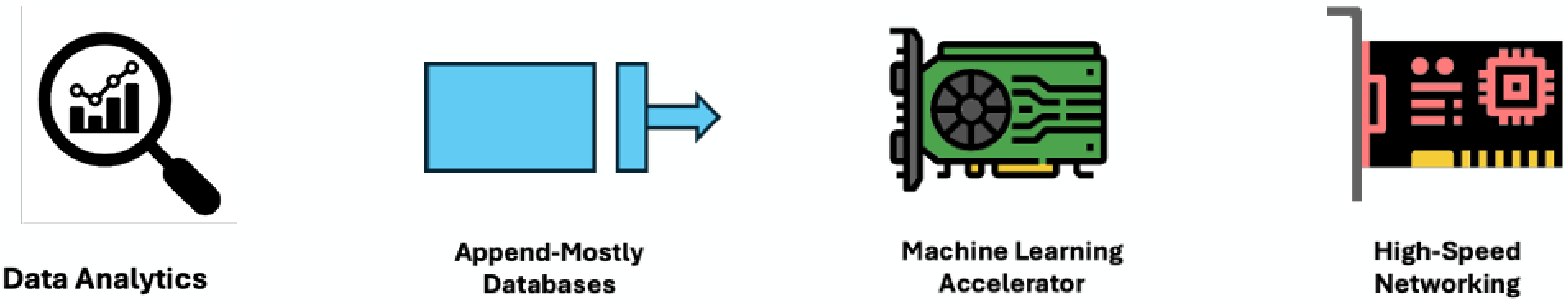
Software Assumes Uniform Memory

Access assumed uniform



A word-addressable random-access uniform memory address space

Software Use of Memory: Very Diverse



Streams of data

- Write-once, read-once
- Filters (scans)
- Joins (random access)

Read >> Write

- Write once
- Mostly append
- Read many times
- Scans
- Random access

Blocked operations

- Blocked operations
- Sparse accesses
- Read multiple times
- Write many times

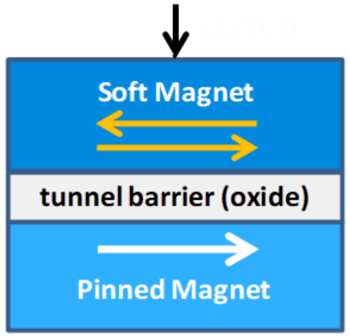
Packet-oriented

- Ultra-low latency
- Header processing
- Packet-oriented
- Read once
- Write once

Philip Levis, Differentiated Memory (DAM) Project white paper, https://dam.stanford.edu/

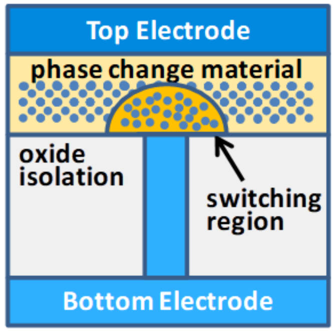
Diverse Memories

> ***Various** parts of this memory have to perform functions which differ somewhat in their nature and considerably in their purpose … —* **J. von Neumann 1946**



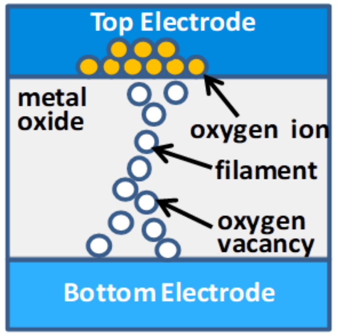
STT-MRAM

Spin transfer **t**orque **m**agnetic **r**andom **a**ccess **m**emory



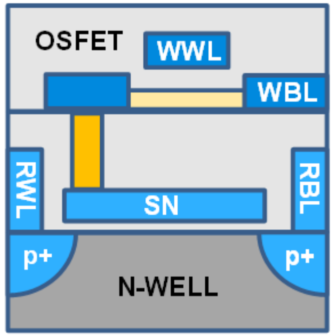
PCM

Phase **c**hange **m**emory



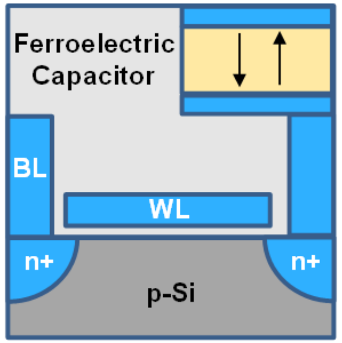
RRAM

Resistive switching **r**andom **a**ccess **m**emory



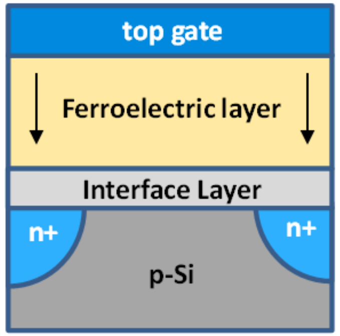
Gain Cell

Gain cell memory (quasi-non-volatile)



FeRAM

Ferro-electric **1T1C** memory (destructive read)



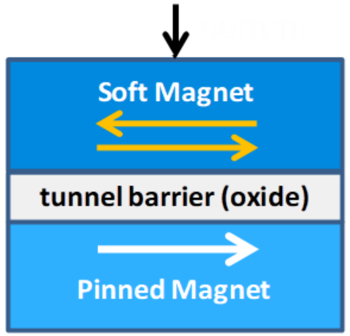
FeFET

Ferro-electric **f**ield **e**ffect **t**ransistor

Updated from: H.-S. P. Wong, S. Salahuddin, *Nature Nanotech.*, 2015.

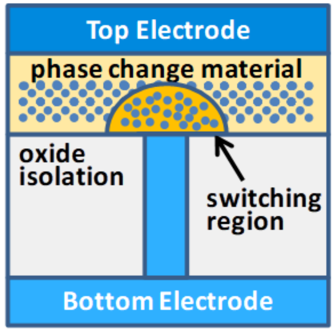
Diverse Memories

Focus on: integration of memory with new capabilities as a tool in our toolbox



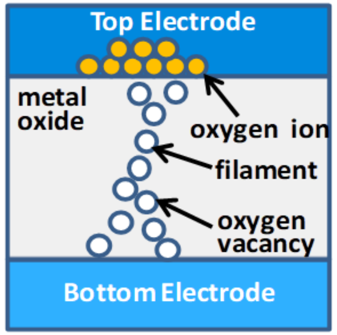
STT-MRAM

Spin transfer **t**orque **m**agnetic **r**andom **a**ccess **m**emory



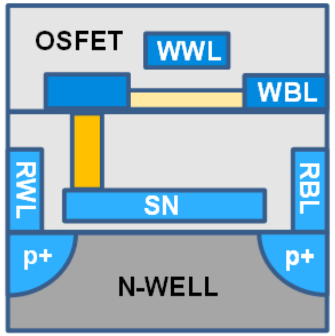
PCM

Phase **c**hange **m**emory



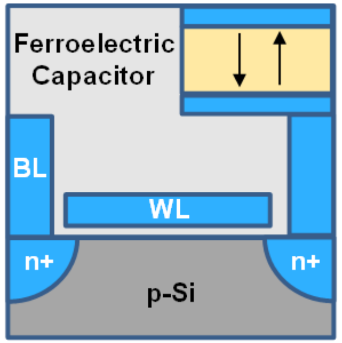
RRAM

Resistive switching **r**andom **a**ccess **m**emory



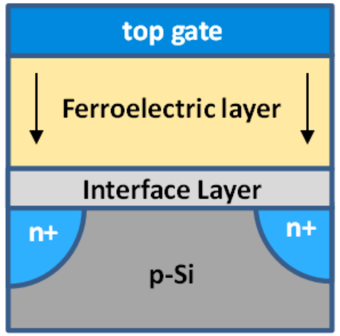
Gain Cell

Gain **cell** memory (quasi-non-volatile)



FeRAM

Ferro-electric **1T1C** memory (destructive read)

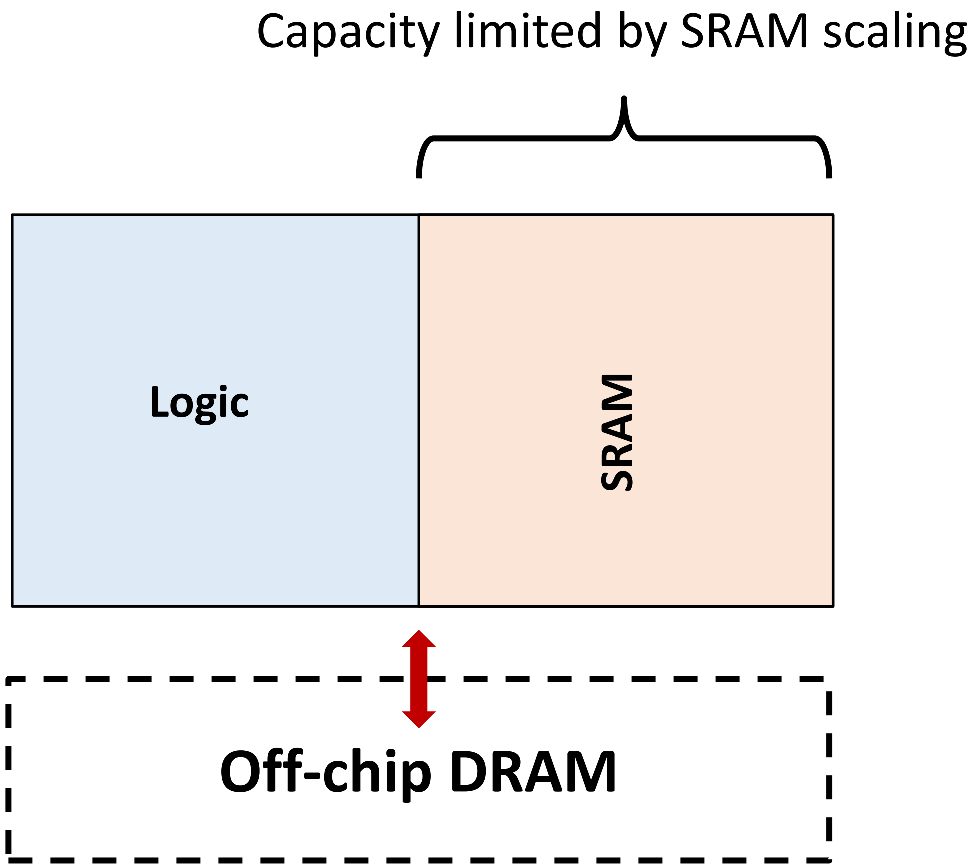


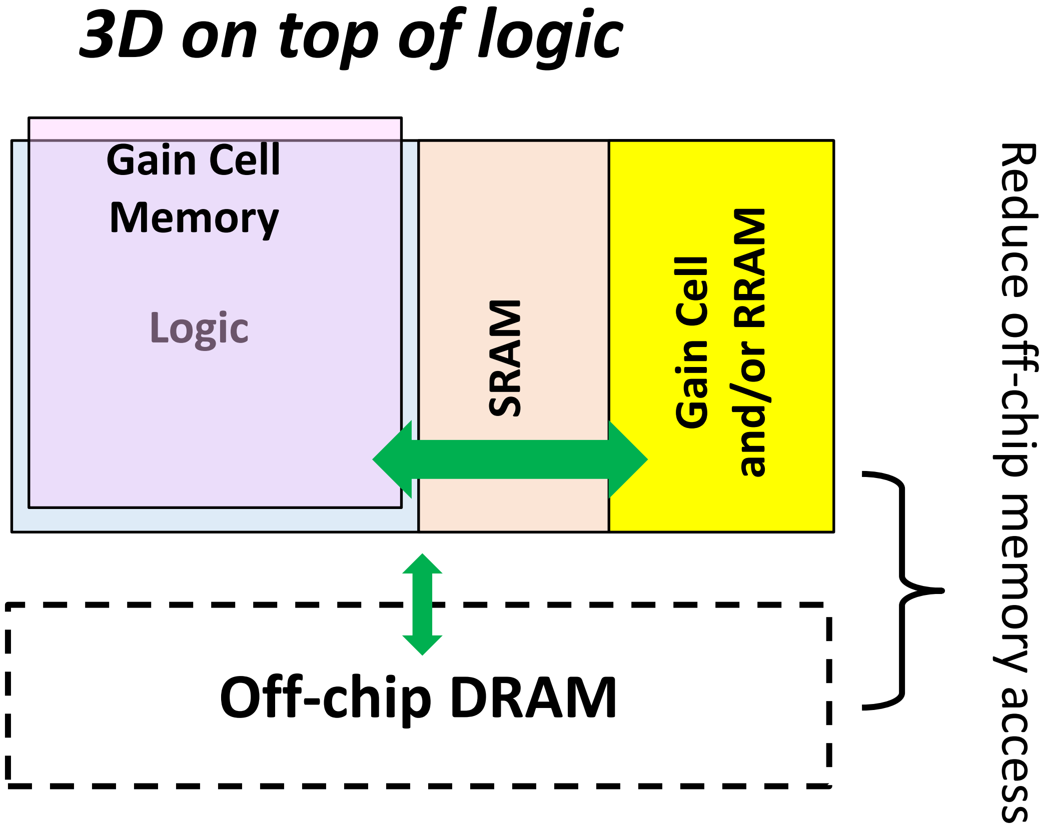
FeFET

Ferro-electric **f**ield **e**ffect **t**ransistor

Updated from: H.-S. P. Wong, S. Salahuddin, *Nature Nanotech.*, 2015.

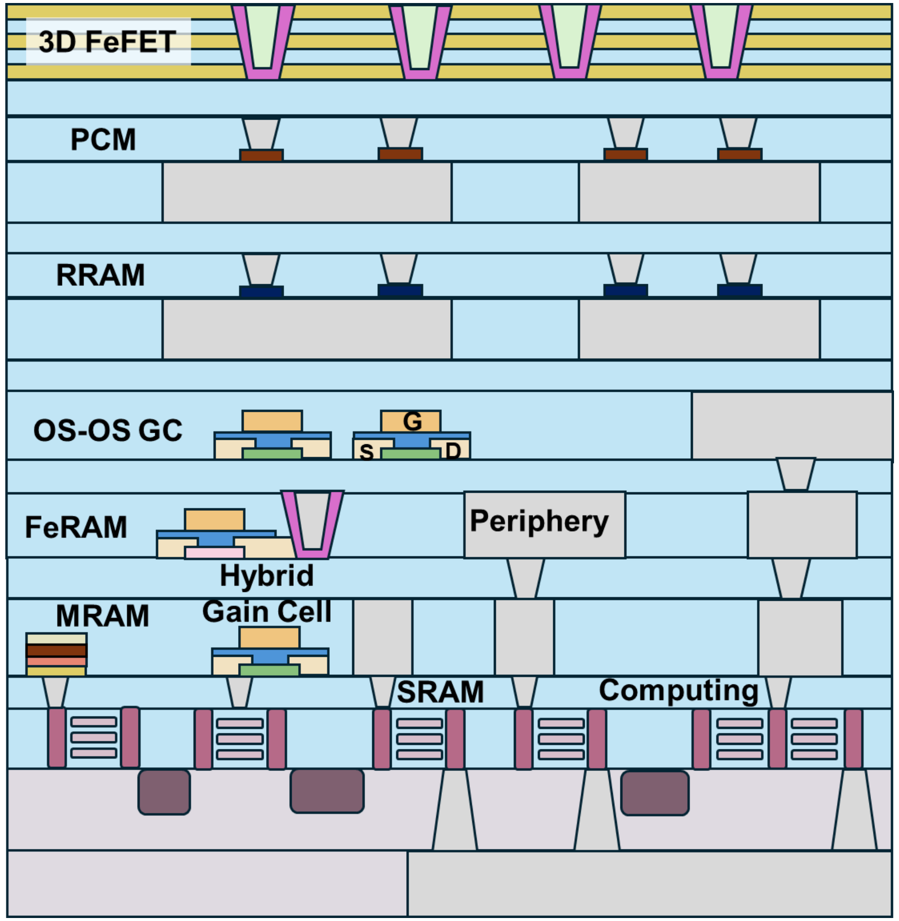
Massive Memory On-Chip





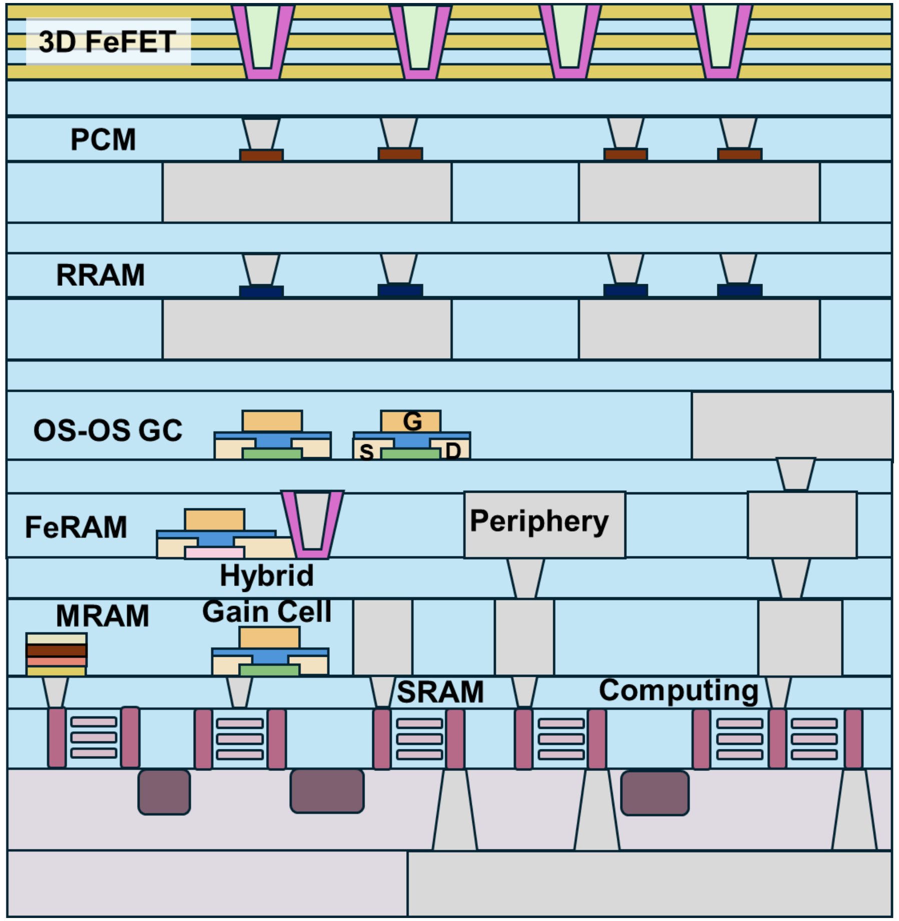
H. Li, W. Wan, and H.-S. P. Wong,, "In- and Near-Memory Computing Using 2D/3D Resistive Memories," *EDTM Tutorial*, 2021.

Future of Memory:
Massive, Diverse, Tightly Integrated with Compute

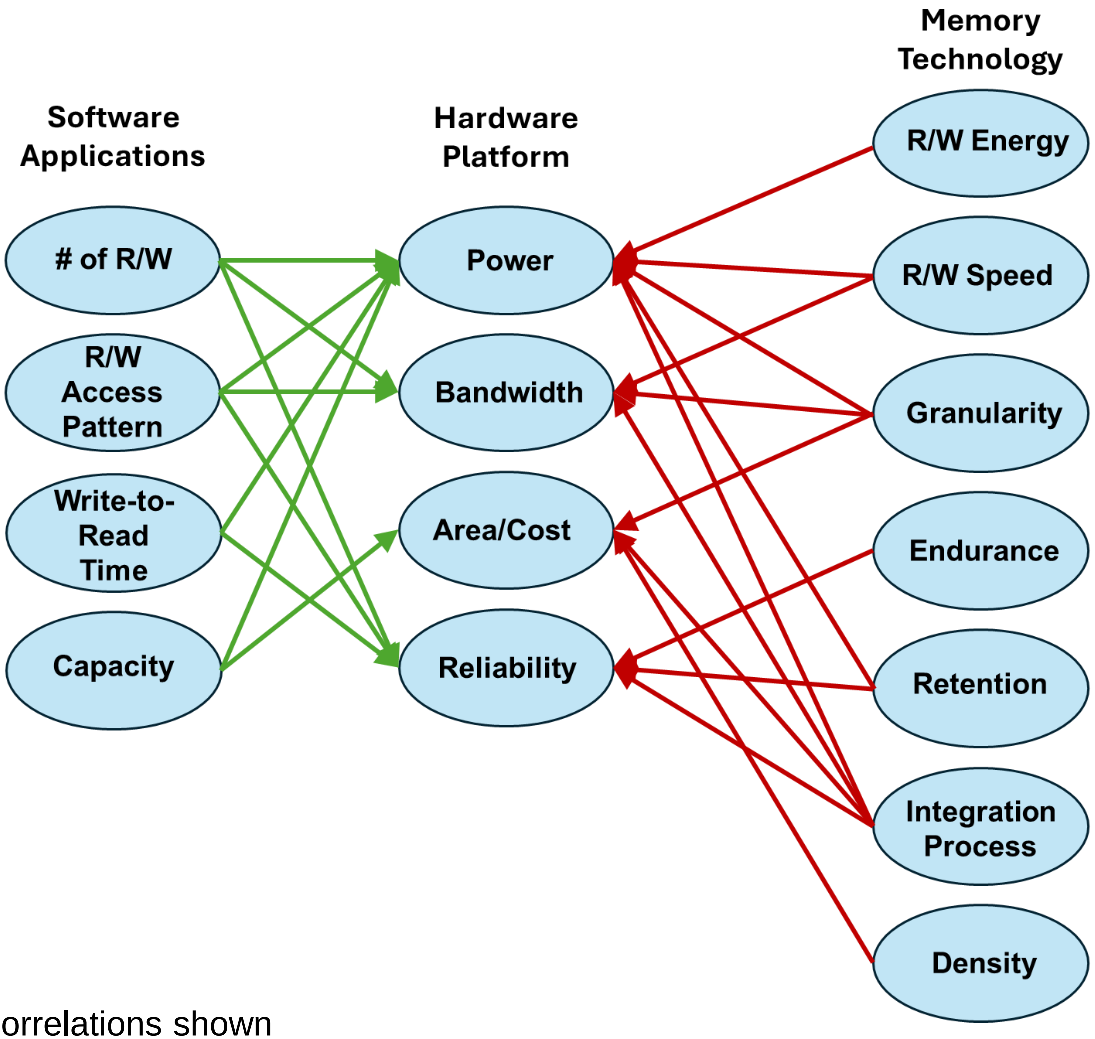


Diverse Memory:

- How to choose?
- How to use?
- What attributes are important?

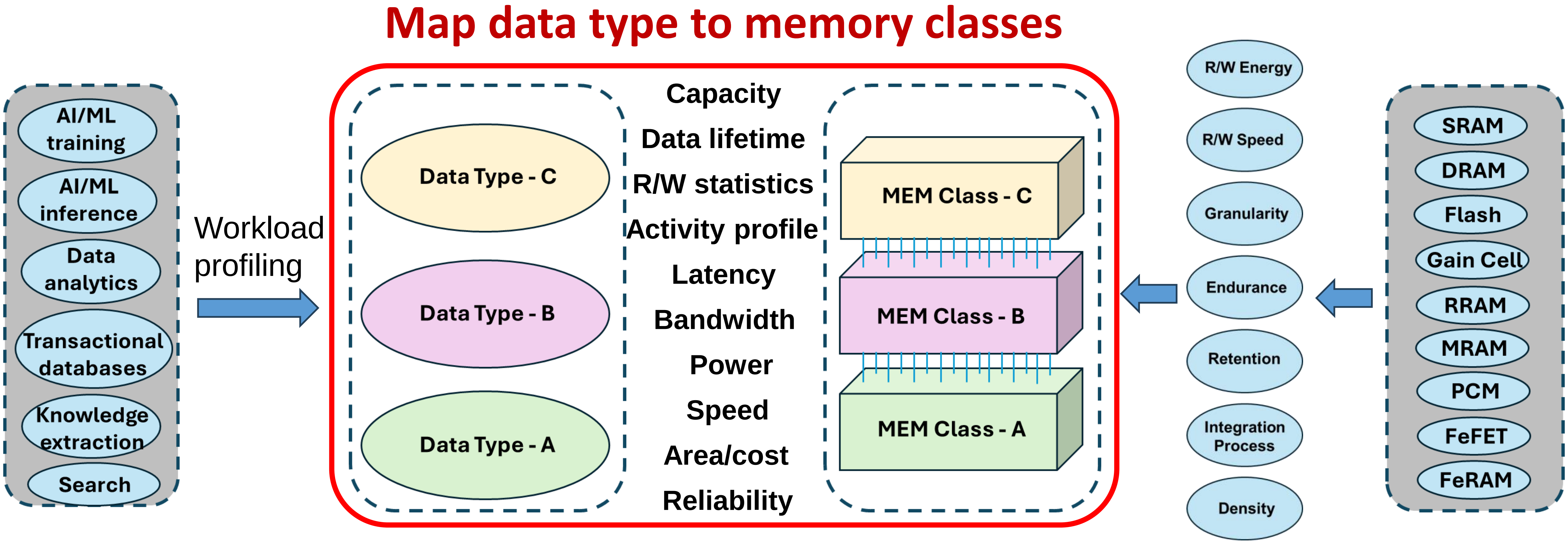


Exposing Hardware to Software



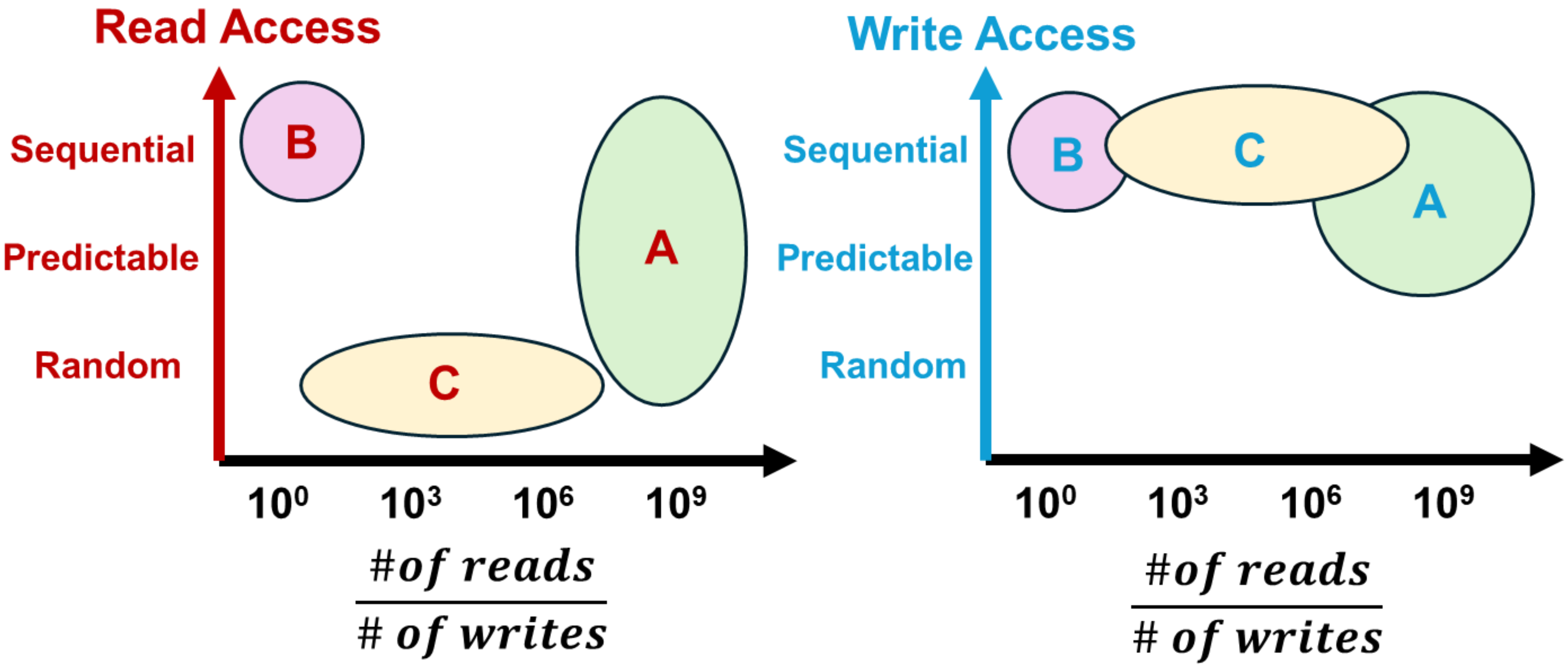
Only major direct correlations shown

Abstraction Layer Needed



Software Data Types

Type A "mostly read" – e.g. AI/ML inference weight memory and processor instruction caches
Type B "streaming data" – e.g. streaming I/O, AI/ML activations, and data analytics
Type C "frequent write" – e.g. buffers for a file system, AI/ML training memory



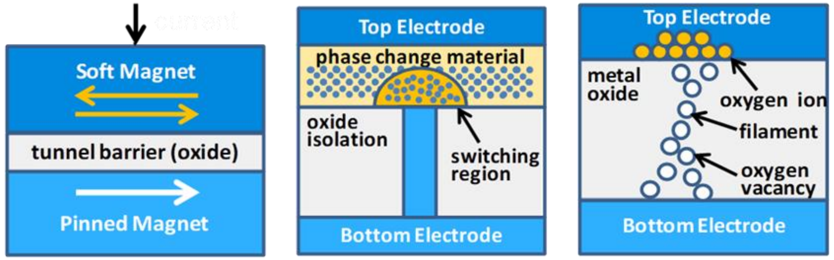
Other attributes:

- Capacity
- Data lifetime
- Access granularity
- Latency…

Type A “mostly read” – Frequent Reads, Infrequent Writes, Predictable Accesses

Trade-off write costs for better read

Data type	Example	Read Energy (pJ/bit)	Read Latency (ns)	Write Energy (pJ/bit)	Write Latency (ns)	Endurance (cycles)	Retention (s)	Capacity	Access granularity	Memory Today	Future Memory
A	Instruction cache	< 0.5	< 1	< 500	< 1,000	$> 1\times 10^8$	> 1	8KB-1MB	Word (8-16B)	SRAM	MRAM, RRAM



STT-MRAM

Spin transfer torque magnetic random access memory

PCM

Phase change memory

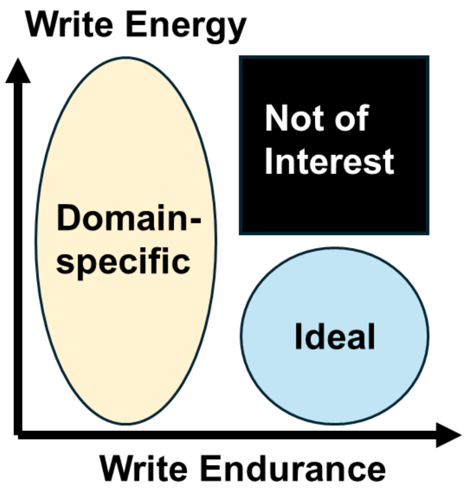
RRAM

Resistive switching random access memory

COMBINATION of Attributes Matters

Trade-off write costs for better read, but write also matters

Data type	Example	Read Energy (pJ/bit)	Read Latency (ns)	Write Energy (pJ/bit)	Write Latency (ns)	Endurance (cycles)	Retention (s)	Capacity	Access granularity	Memory Today	Future Memory
A	Instruction cache	< 0.5	< 1	< 500	< 1,000	$> 1\times 10^8$	> 1	8KB-1MB	Word (8-16B)	SRAM	MRAM, RRAM

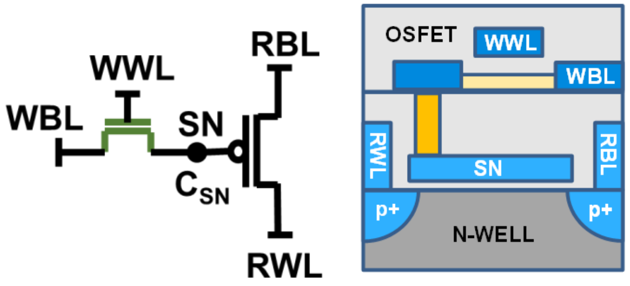


Example RRAM/MRAM :
Write energy & endurance should be optimized together

Type B “streaming data” – Frequent Writes, Few Reads per Write, Short Data Lifetime

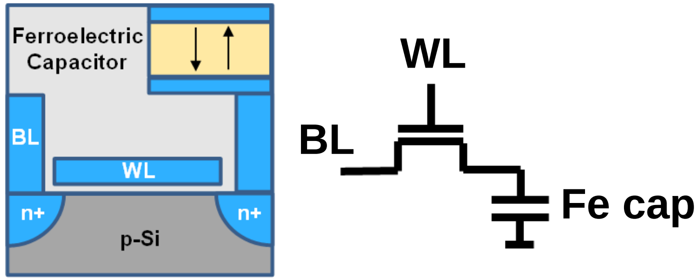
Trade-off retention for speed/density/energy

Data type	Example	Read Energy (pJ/bit)	Read Latency (ns)	Write Energy (pJ/bit)	Write Latency (ns)	Endurance (cycles)	Retention (s)	Capacity	Access granularity	Memory Today	Future Memory
B	Video streaming	< 200	$< 1,000$	< 200	$< 1{,}000$	$> 1\times 10^9$	0.1 - 10	1KB-10MB	Page (KB)	DRAM	FeRAM, Gain Cell



Gain Cell

Gain cell memory (quasi-non-volatile)



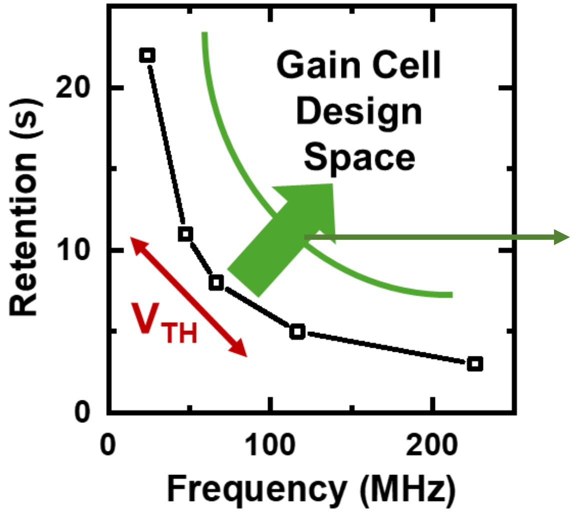
FeRAM

Ferro-electric 1T1C memory (destructive read)

Trade-off Design Knob Matters

Trade-off retention for speed/density/energy

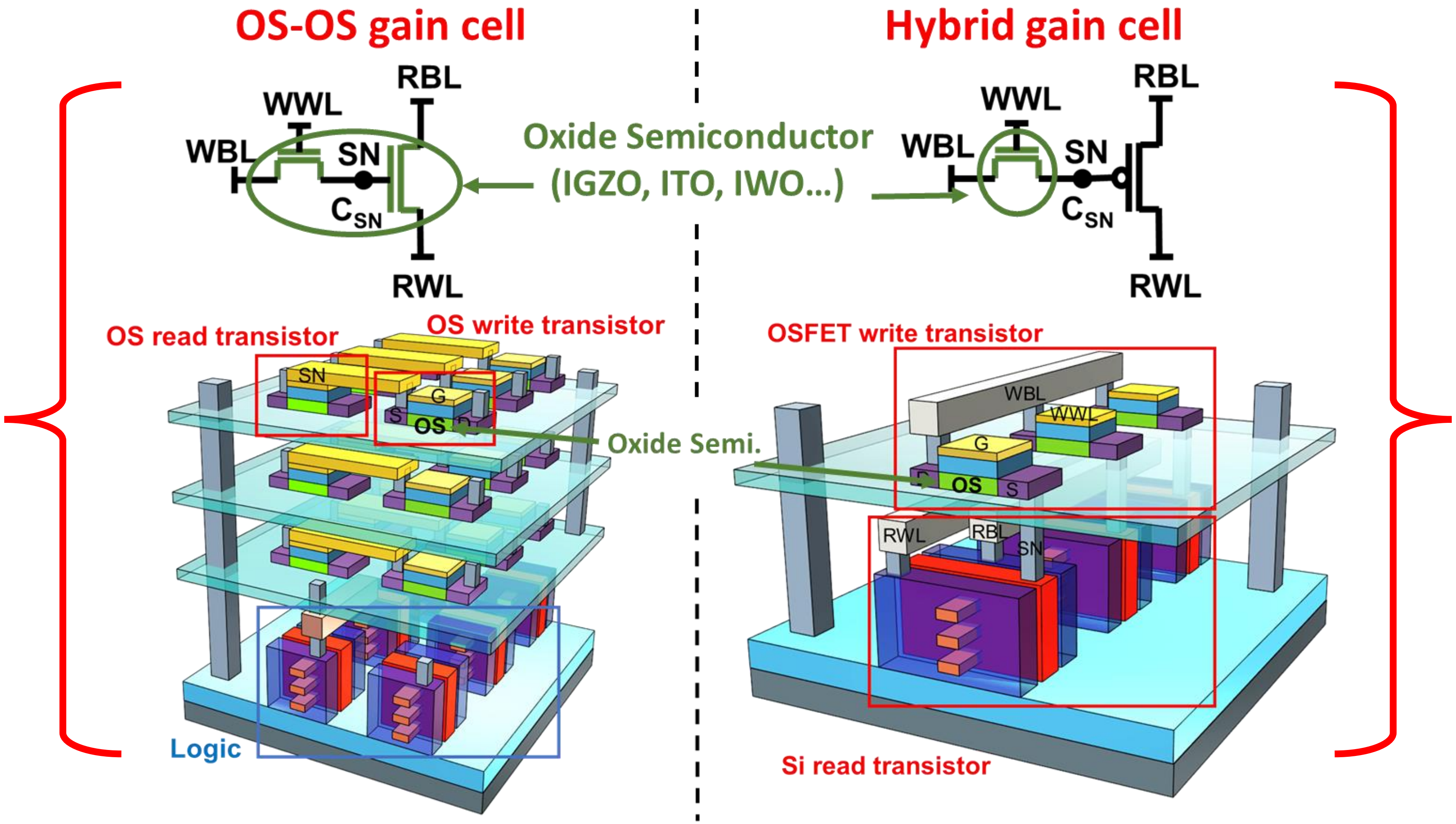
Data type	Example	Read Energy (pJ/bit)	Read Latency (ns)	Write Energy (pJ/bit)	Write Latency (ns)	Endurance (cycles)	Retention (s)	Capacity	Access granularity	Memory Today	Future Memory
B	Video streaming	< 200	< 1, 000	< 200	< 1,000	$> 1\times 10^9$	0.1 - 10	1KB-10MB	Page (KB)	DRAM	FeRAM, Gain Cell



Device: SS, transition region
Circuit: voltage, hybrid gain cell

Oxide Semiconductor Gain Cell

Higher capacity


OS-OS gain cell

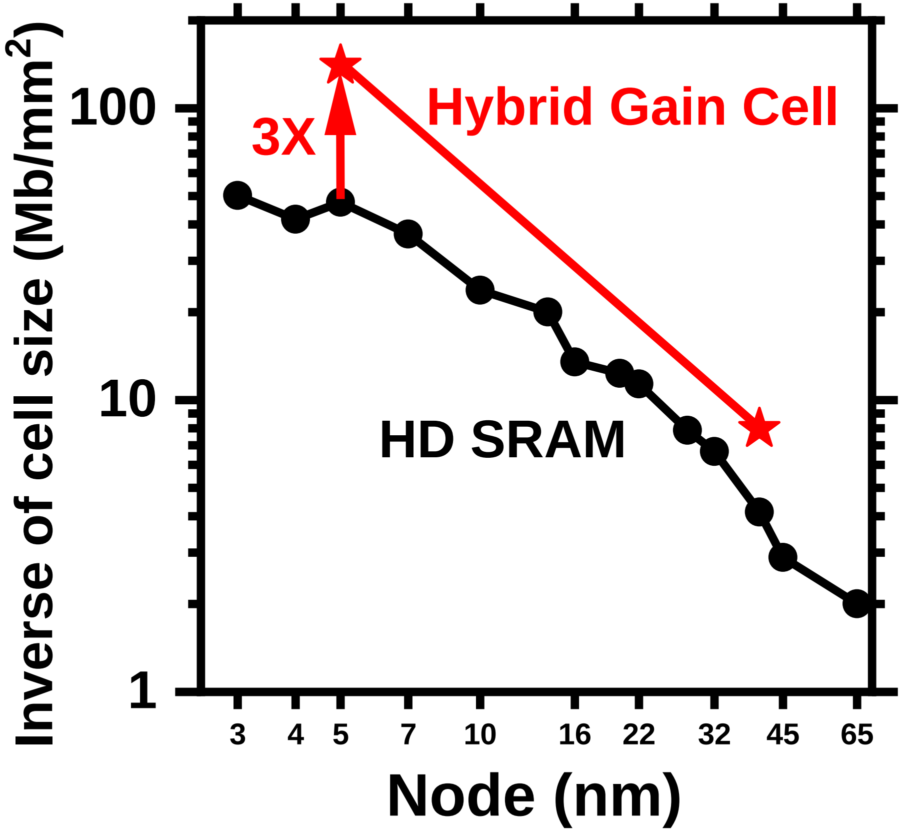
Hybrid gain cell

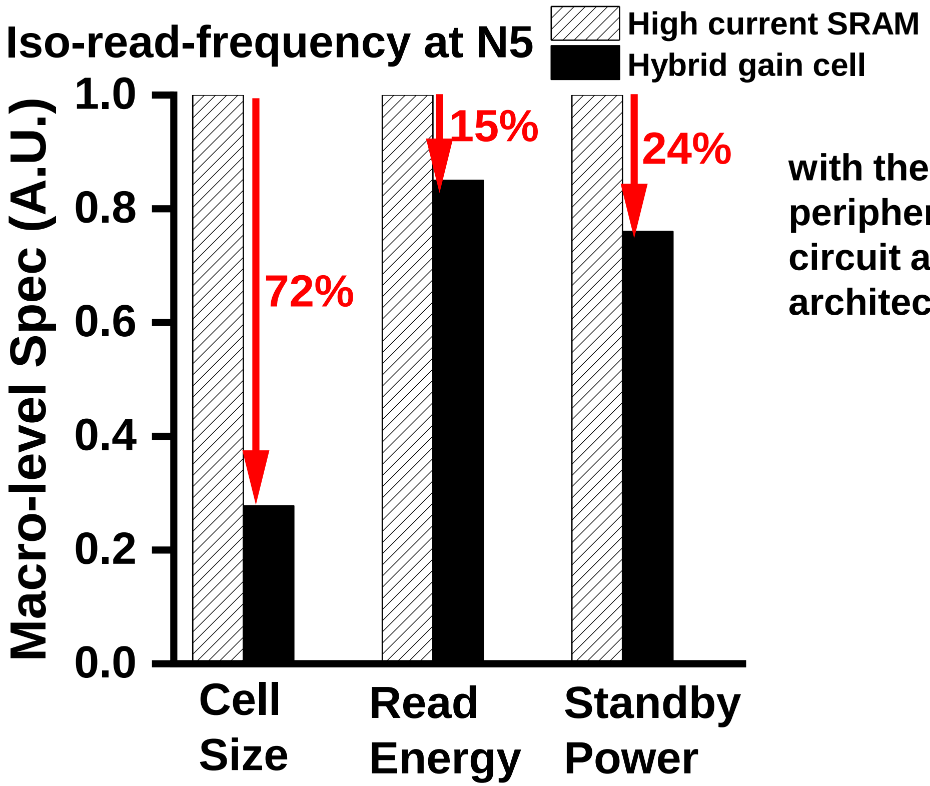
Oxide Semiconductor (IGZO, ITO, IWO...)

Higher speed

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

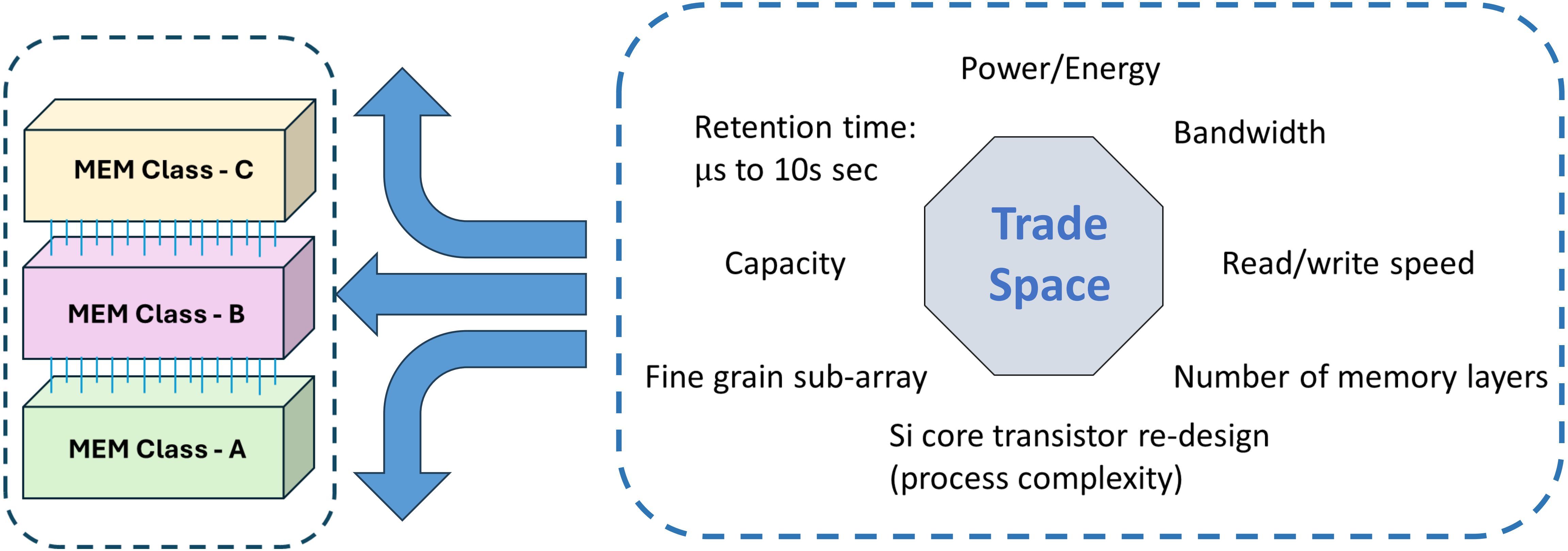
Hybrid Gain Cell –
High-density Scalable to N5





with the same peripheral circuit and array architecture

Optimize Tradeoff Guided by Software Use



MEM Class - C

MEM Class - B

MEM Class - A

Trade Space

Power/Energy

Retention time:
µs to 10s sec

Bandwidth

Capacity

Read/write speed

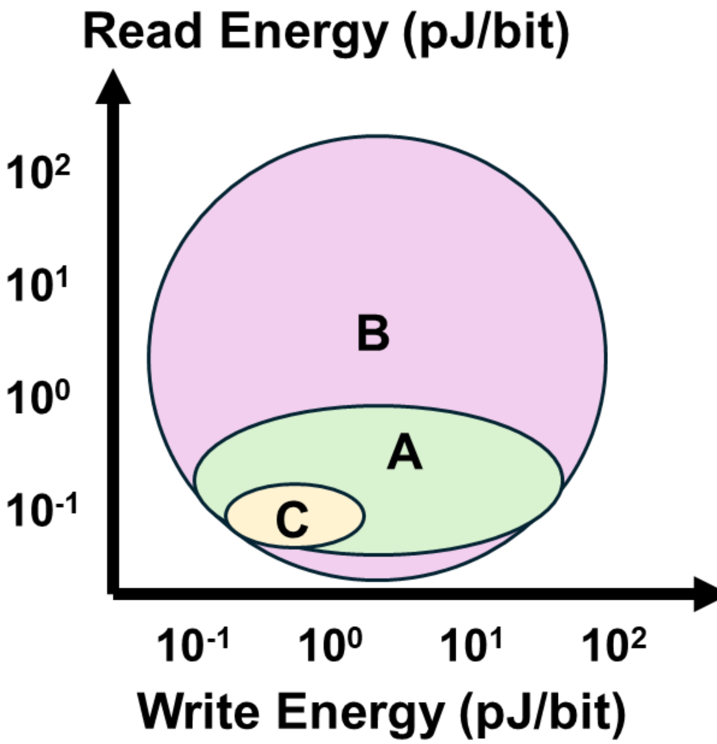
Fine grain sub-array

Number of memory layers

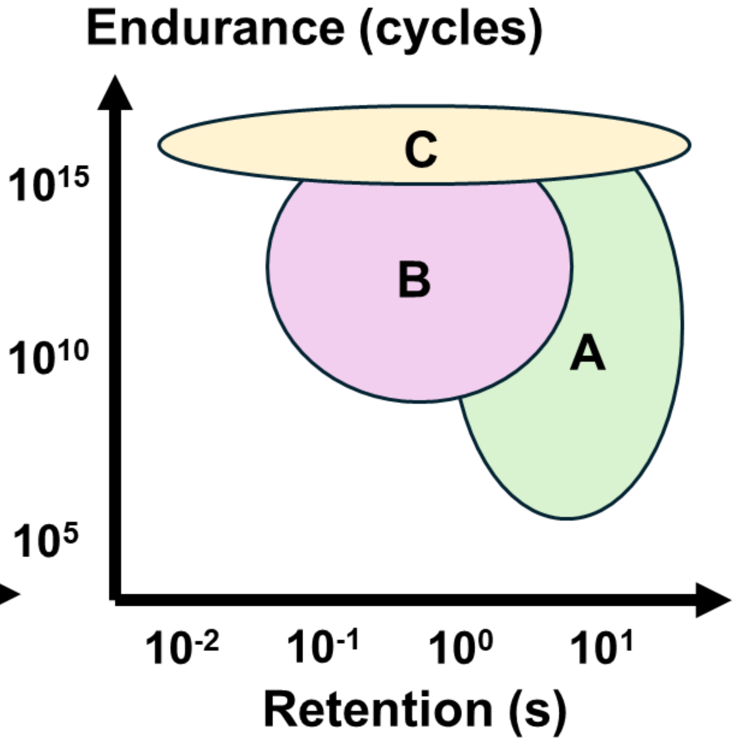
Si core transistor re-design
(process complexity)

Diverse Hardware Specs for Software Data Types A, B, C

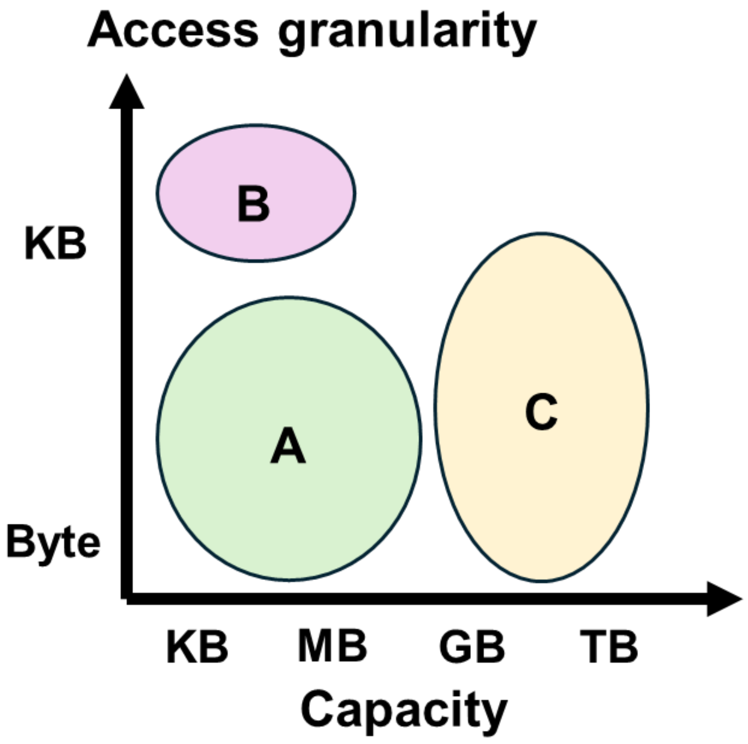
Subset



Intersection



Disjoint



Typical Memory Comparison

- Attributes in isolation
- Not application-correlated

	SRAM	DRAM	RRAM	MRAM
Energy	Low	Medium	High	High
Speed	High	Medium	Low	Low
Density	Low	Medium	High	High
Endurance	High	High	Low	Medium

We may be working too hard for no good reason !

Memory Comparison w/ Improvement Target

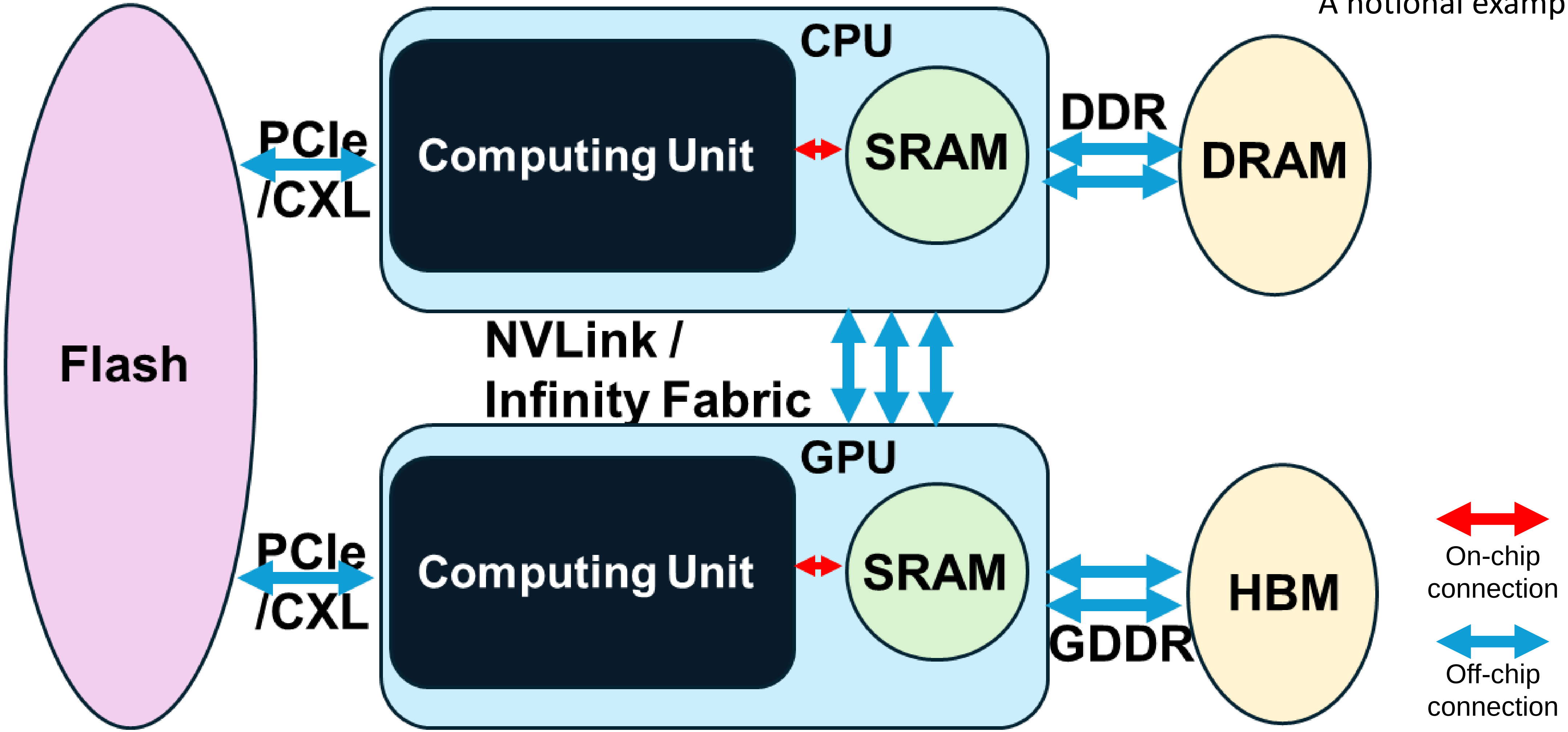
Improvements needed for each memory technology to be used in the software use cases, based on state-of-the-art macro demonstrations.

Data Type	SRAM	3D V-Cache	DRAM	OS-OS Gain Cell	Hybrid Gain Cell	RRAM	MRAM	PCM	FeRAM
B	Density	Standby power	Retention	Capacity	Capacity	Endurance & write energy	Write energy	Endurance & write energy	Read energy

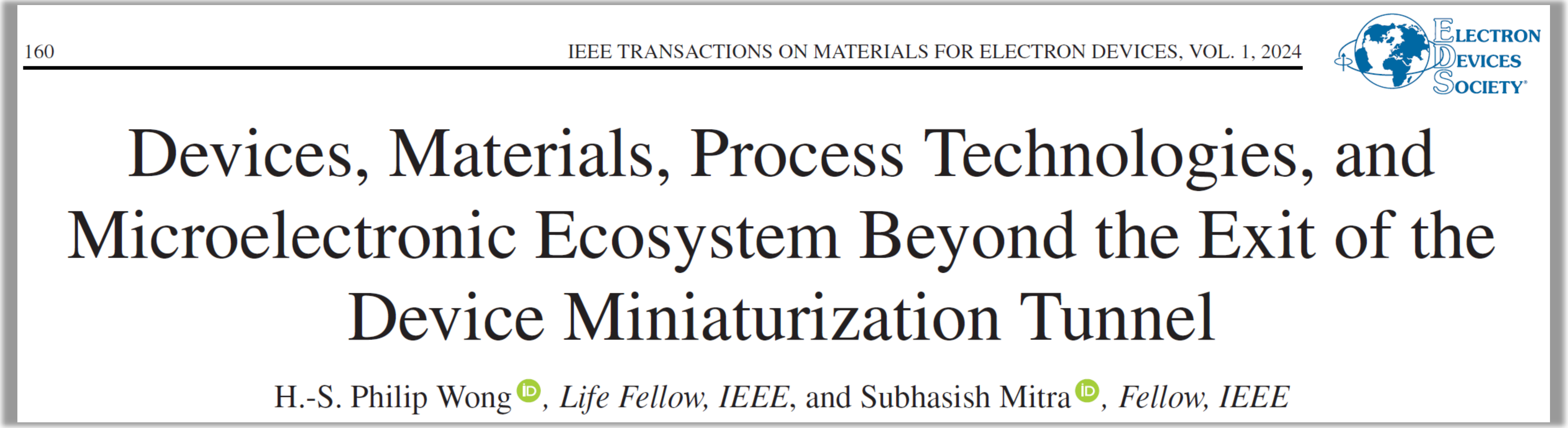
Type B “streaming data” – e.g. streaming I/O, AI/ML activations, and data analytics

Physical Layers with Interface Protocol (Today)

A notional example



The KEY is INTEGRATION



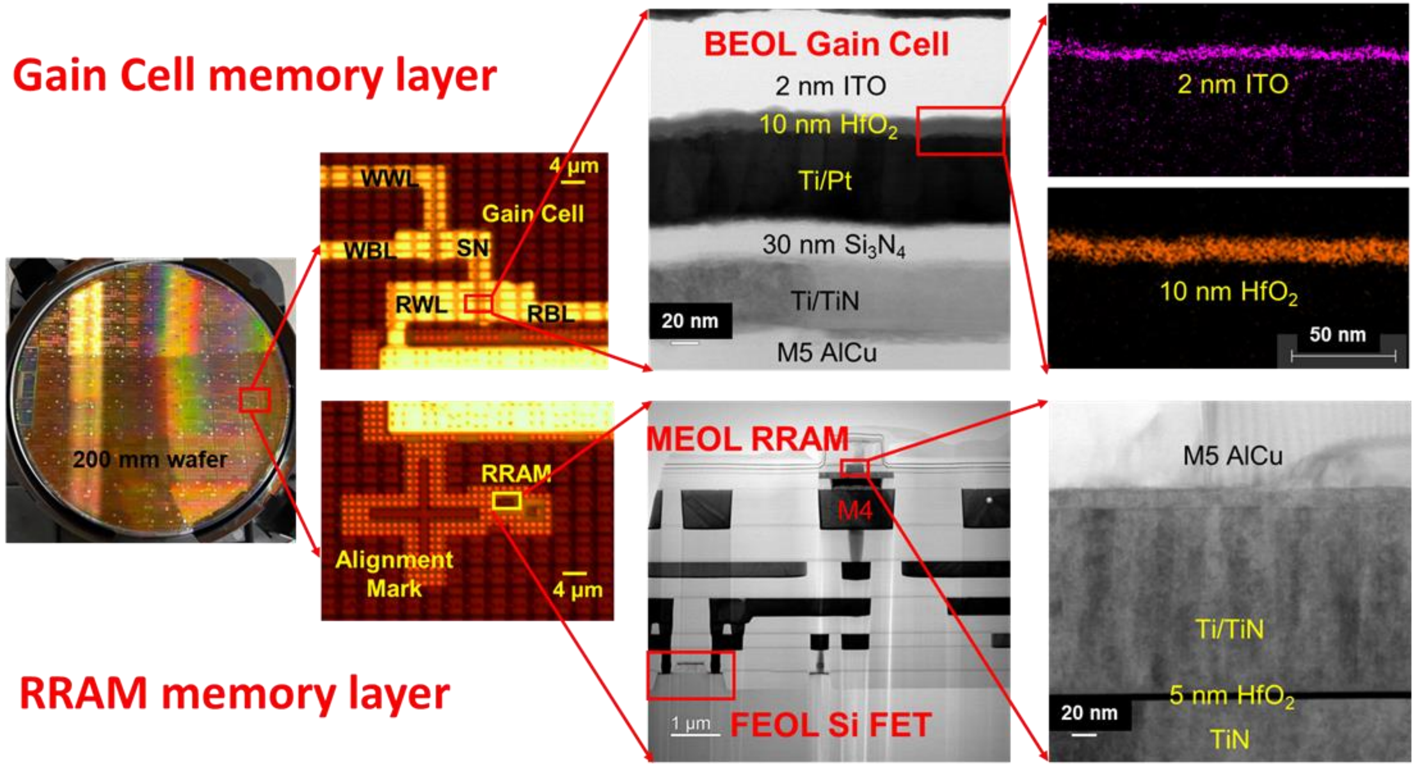
160 IEEE TRANSACTIONS ON MATERIALS FOR ELECTRON DEVICES, VOL. 1, 2024

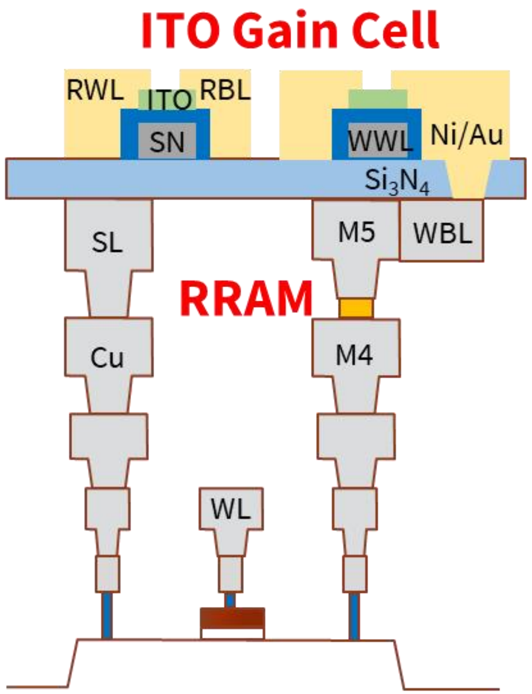
Devices, Materials, Process Technologies, and Microelectronic Ecosystem Beyond the Exit of the Device Miniaturization Tunnel

H.-S. Philip Wong, *Life Fellow, IEEE*, and Subhasish Mitra, *Fellow, IEEE*

H.-S. P. Wong and S. Mitra, *IEEE Trans. Materials for Electron Devices* (T-MAT), 2024.

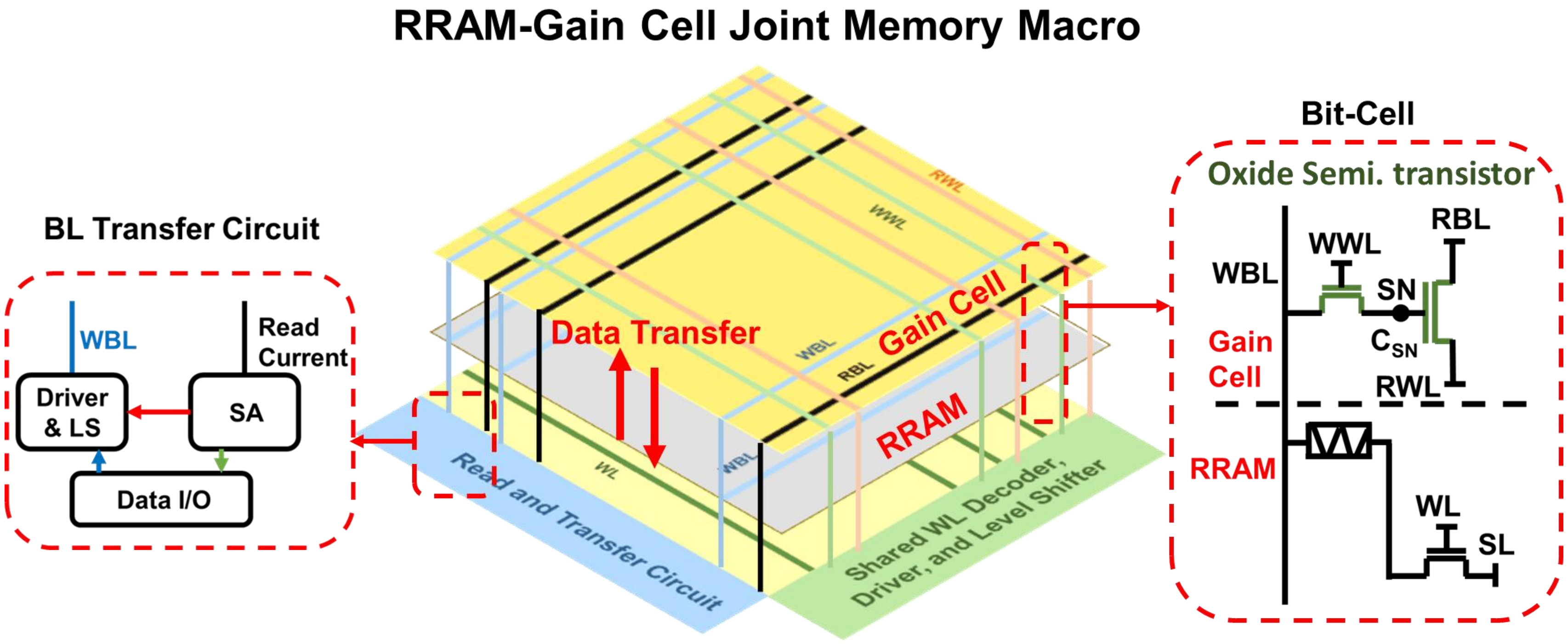
RRAM & Gain Cell Integration on Si CMOS: On-Chip Physical Integration

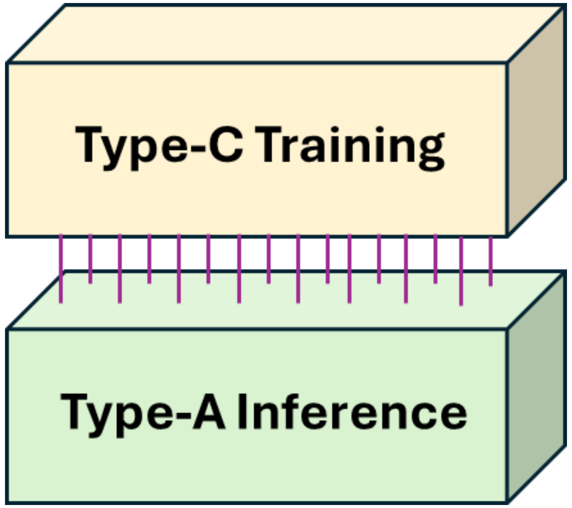




Shuhan Liu, …, H.-S. Philip Wong, IEDM 2024, paper 15-3

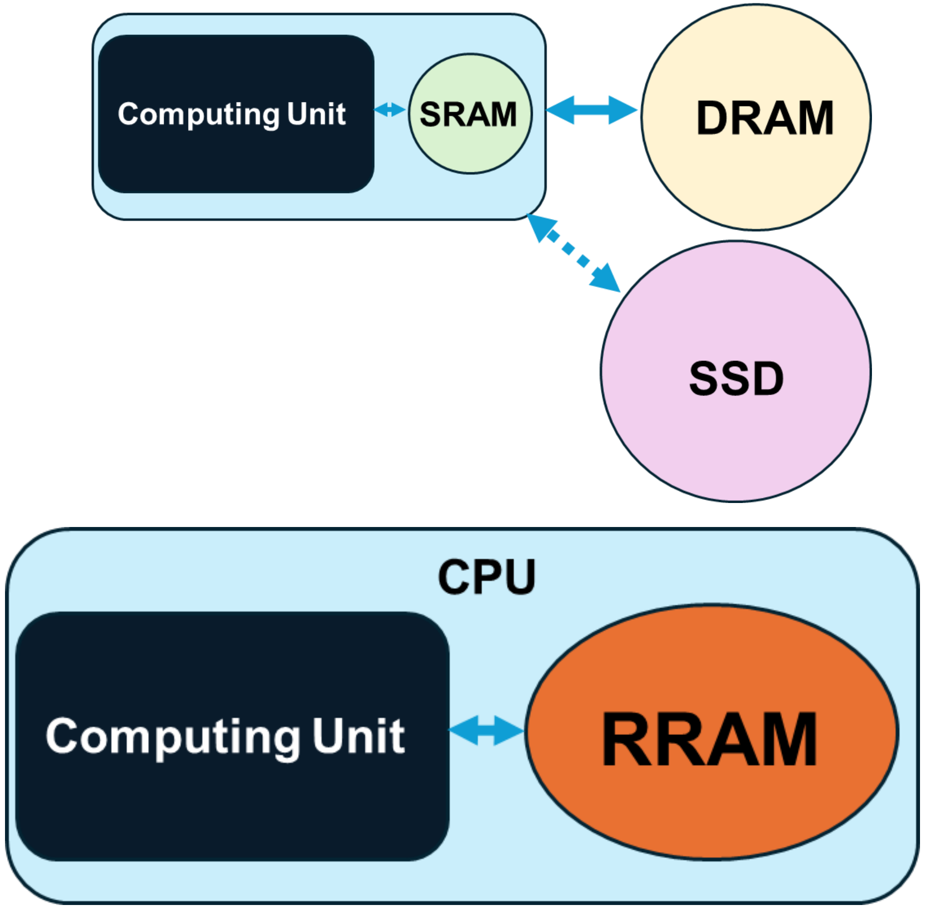
RRAM & Gain Cell Integration on Si CMOS: On-Chip Architectural Integration

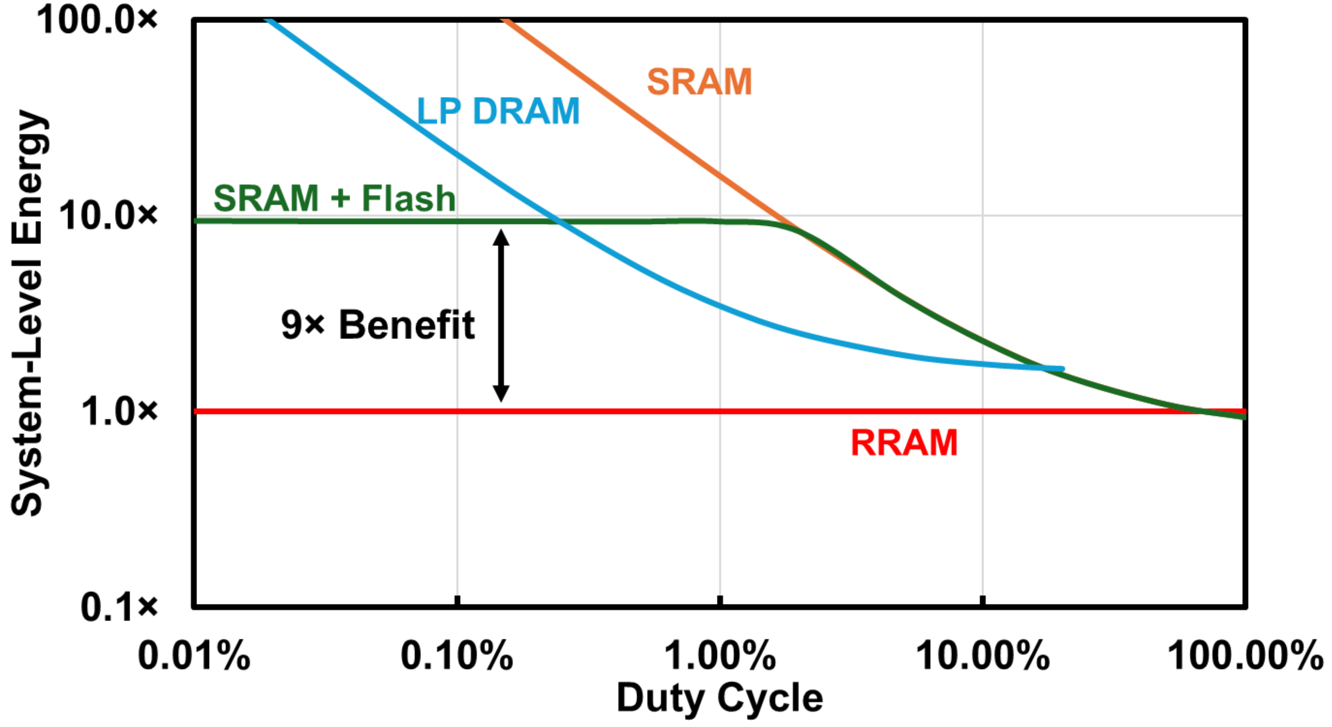




Shuhan Liu, …, H.-S. Philip Wong, IEDM 2024, paper 15-3

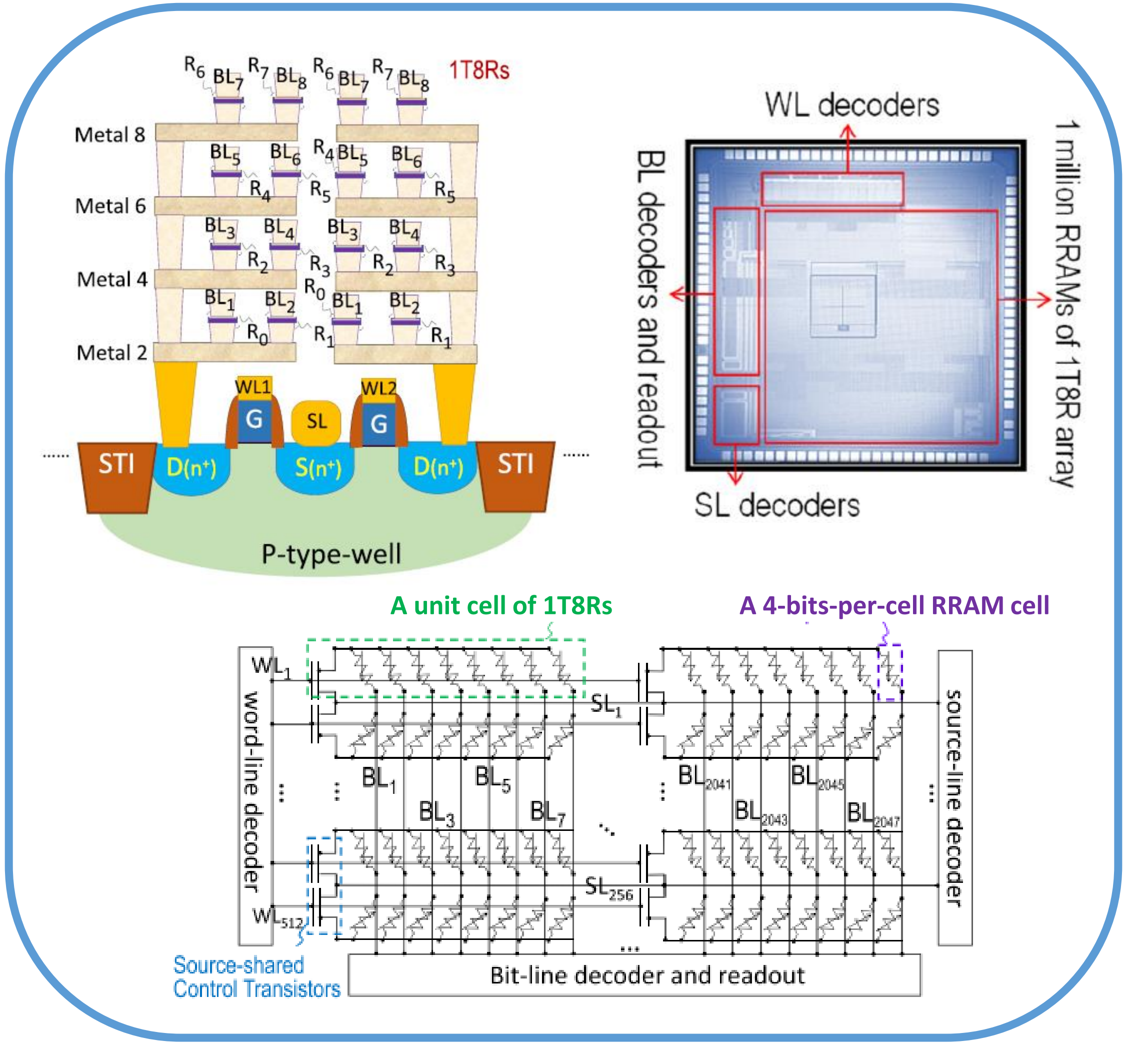
RRAM non-volatility provides 9× System energy benefits



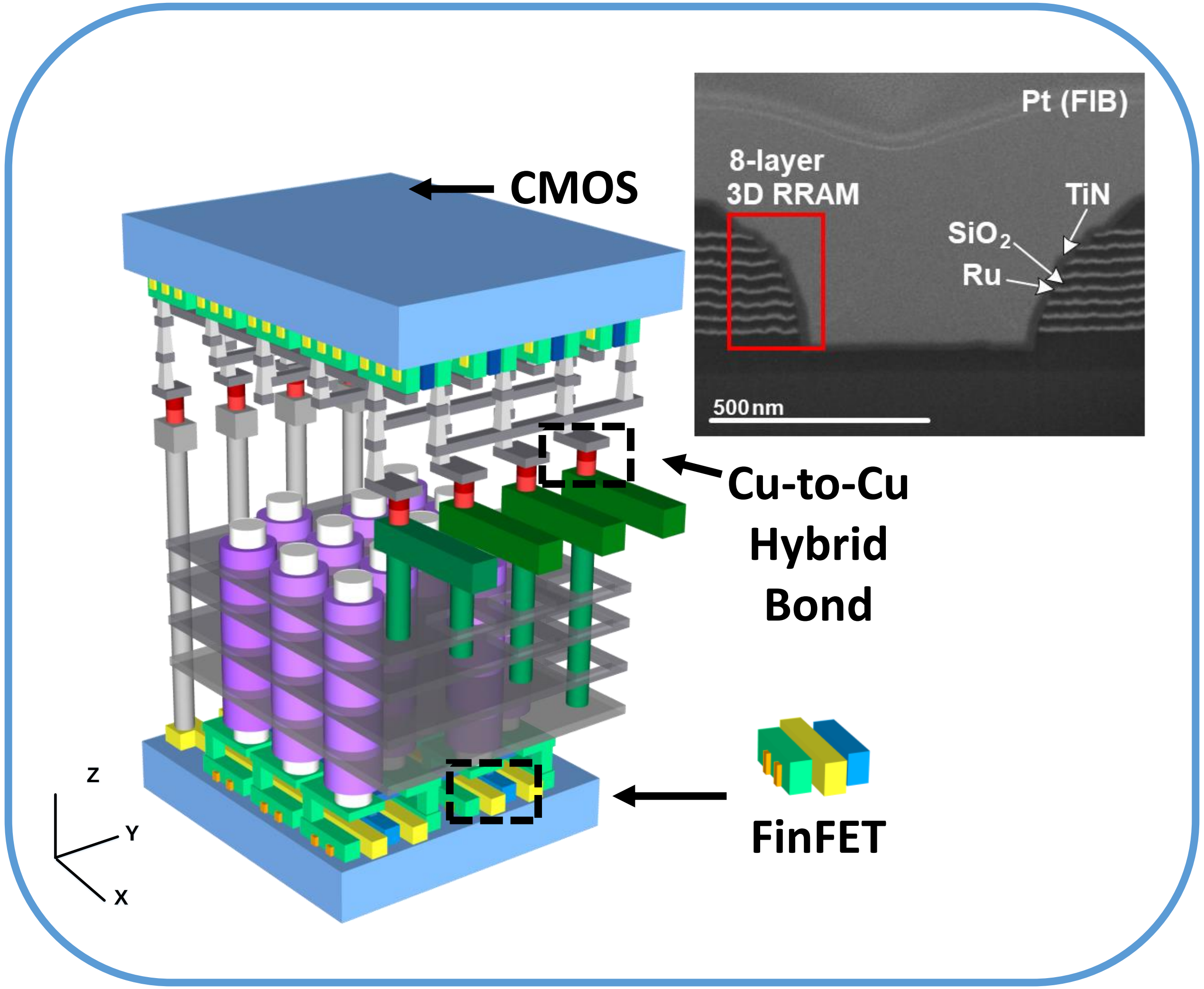


Based on accelerator system in K. Prabhu*, R.M. Radway*, … P. Raina, JSSC 2021.

High-Capacity RRAM: 1T8R, 3D RRAM

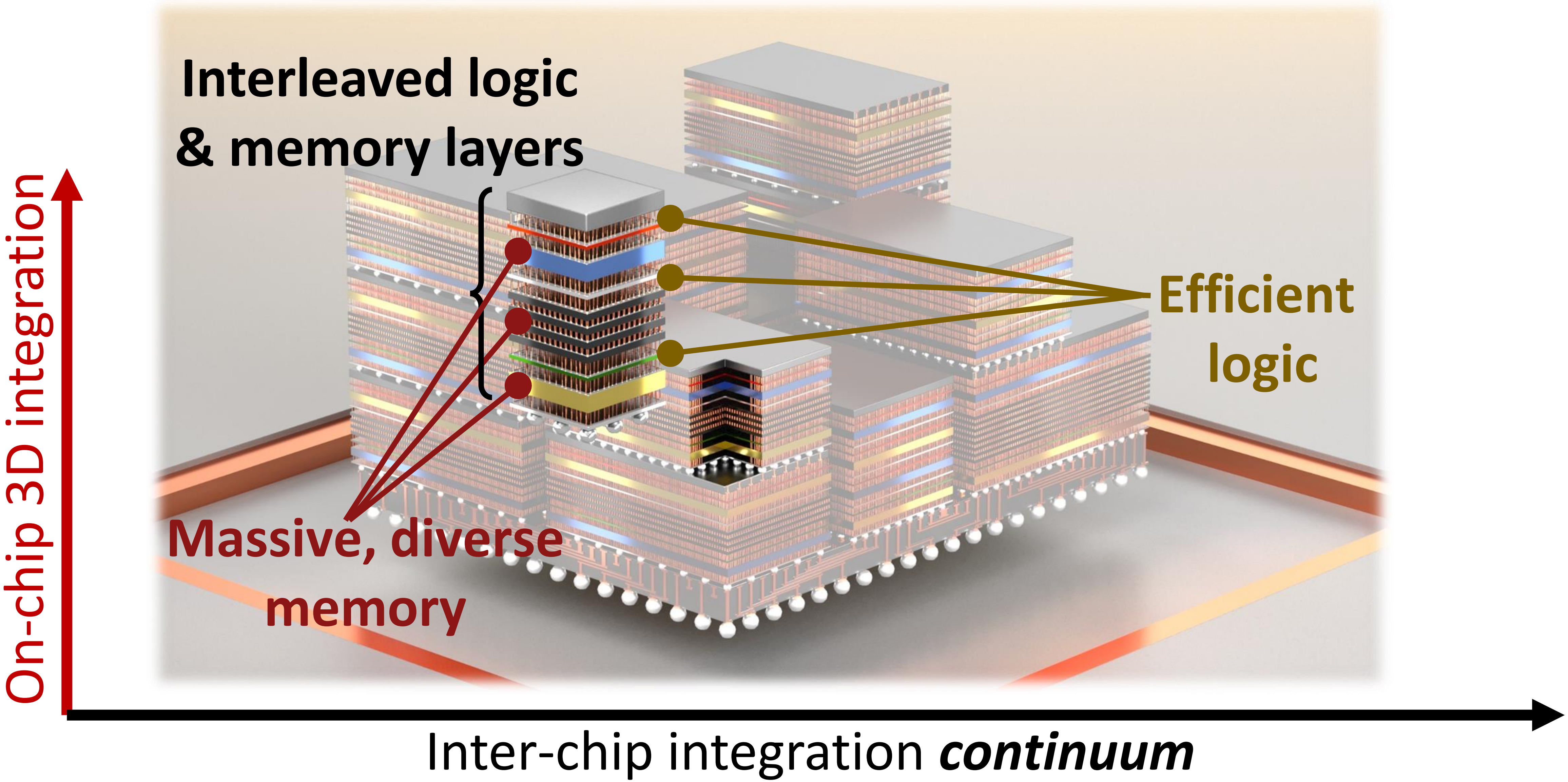


E. R. Hsieh, …, S. Mitra, S. Wong, 2021 EDL.



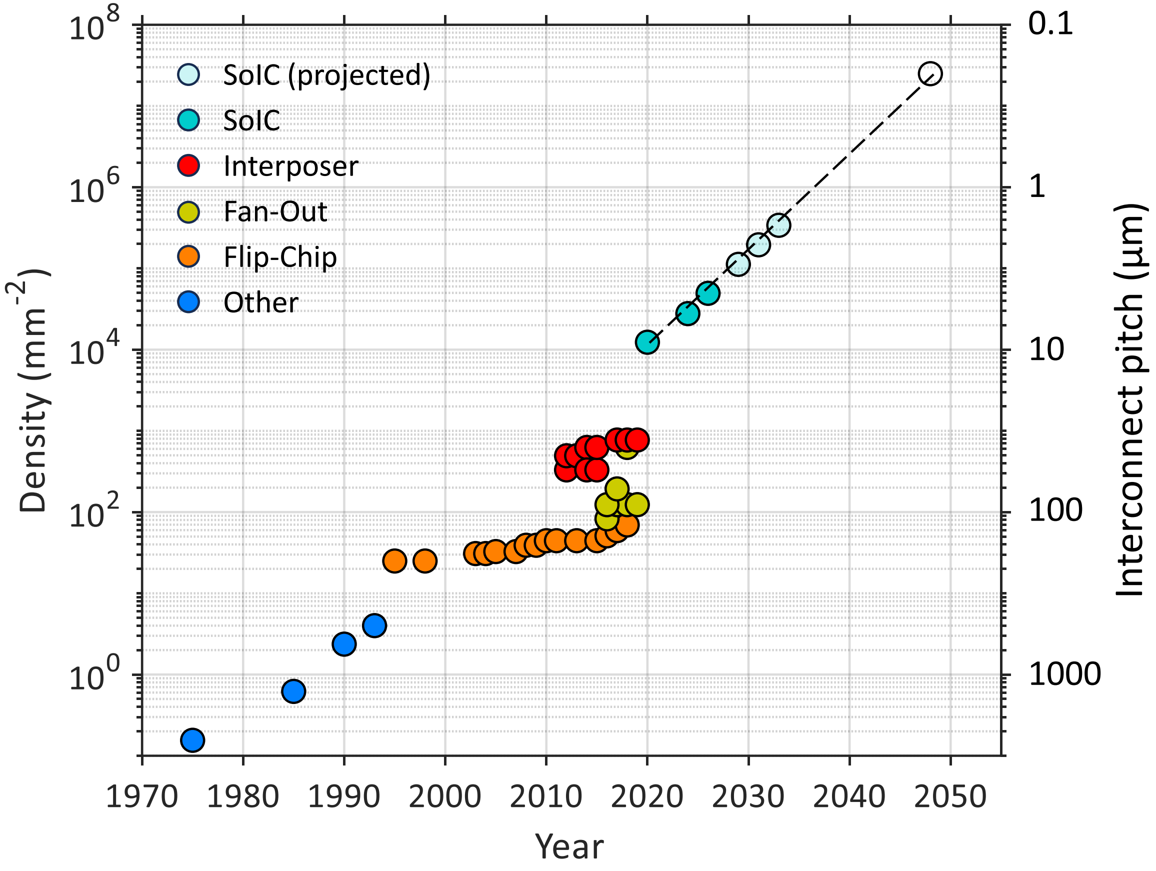
S. Qin, …, H.-S. P. Wong, VLSI 2022, paper T04-3.

Continuum of Interconnection Density



Interleaved logic & memory layers
Efficient logic
Massive, diverse memory
On-chip 3D integration
Inter-chip integration *continuum*

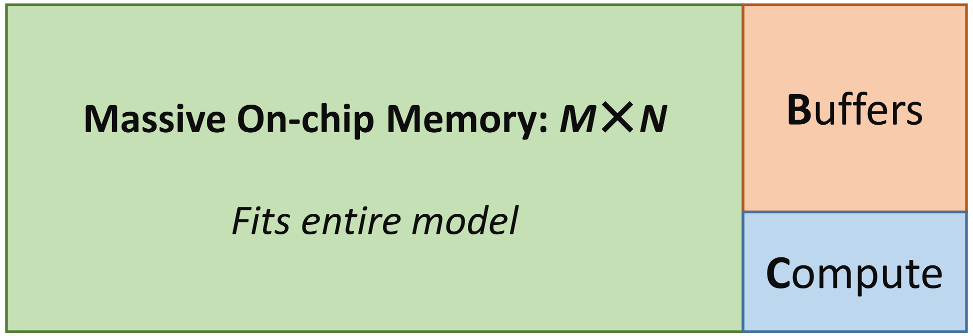
Interconnect Density –
Inter-Chip Physical Integration



H.-S. P. Wong and S. Mitra, *IEEE Trans. Materials for Electron Devices* (T-MAT), 2024.

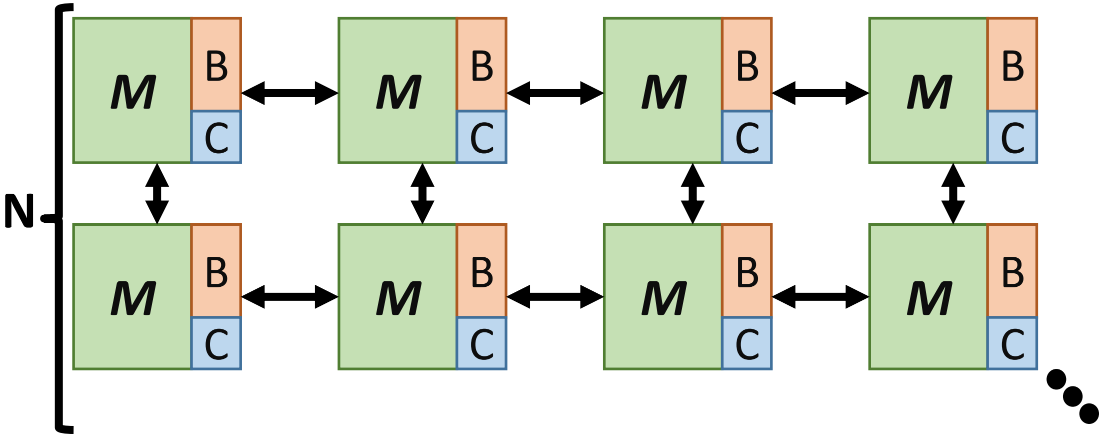
Illusion System –
Inter-Chip Architectural Integration

Dream Chip:



≈

Illusion System:



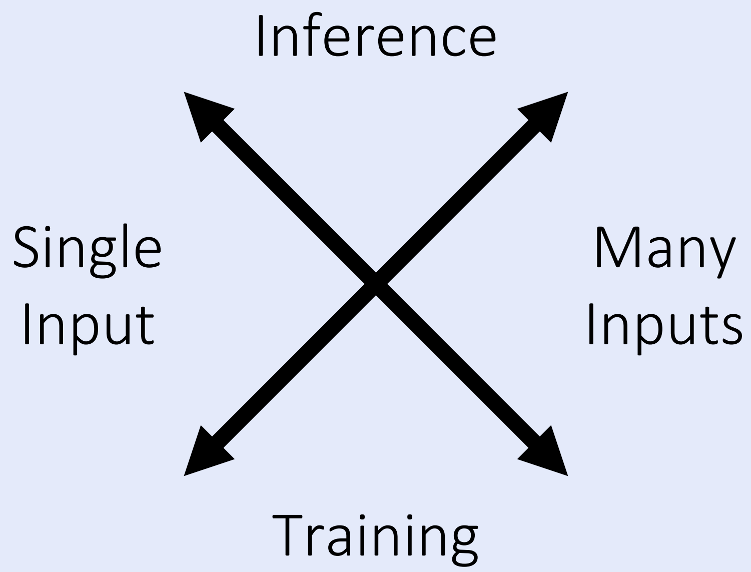
N: chips
M: memory
B: buffer
C: compute

Three Key Ideas: *Enough* on-chip memory + ***Quick*** chip ON/OFF + ***Special*** mapping

R.M. Radway, … Subhasish Mitra, IEDM 2021, paper 25.4
and Nature Electronics 2021.

Illusion within 1.15 × Dream EDP

Edge AI/ML Applications



DNNs, CNNs, LSTMs, D2NNs, Transformers, …

Illusion ≈ Dream

***1.15×* Dream EDP**

Illusion Energy

≤ 1.1×

Dream Energy

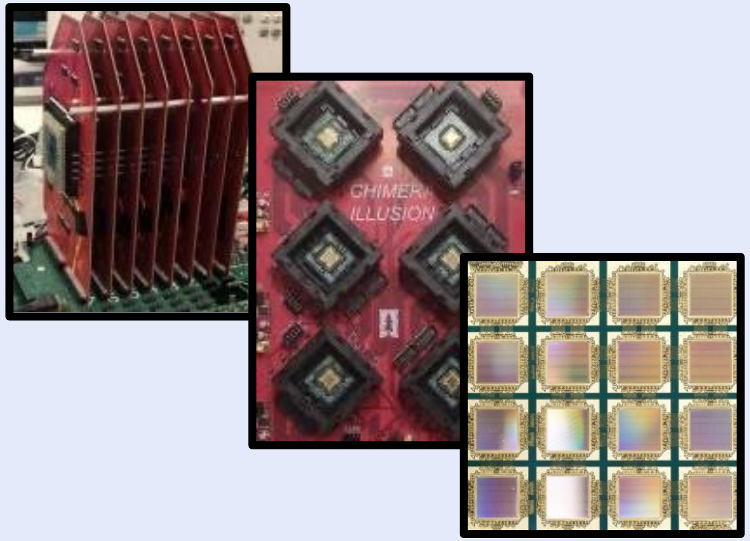
Illusion Exec. Time

≤ 1.05×

Dream Exec. Time

(measured for AI inference)

Hardware-proven *backed by theory*



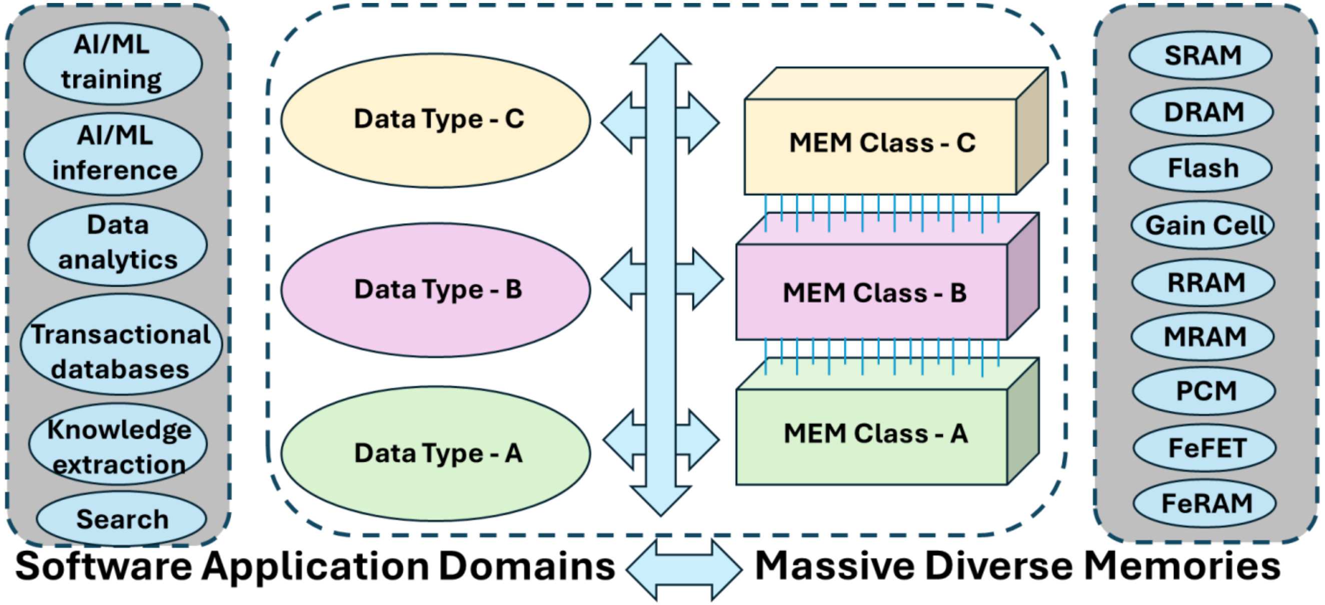
6 to 8 Chip Illusions

32 KB to 96 MB Systems

R.M. Radway, … Subhasish Mitra, IEDM 2021, paper 25.4 and Nature Electronics 2021; K. Prabhu*, R.M. Radway*, … Priyanka Raina, JSSC 2022.

Future of Memory: Massive, Diverse, Tightly Integrated with Compute – from Device to Software

- Massive High-Density On-Chip Memory
- Diverse Memories Exposed to Software
- Tight Integration with Compute Physically and Architecturally



Acknowledgments

Semiconductor Research Corporation

MICROELECTRONICS COMMONS
BRIDGING THE GAP FROM LAB TO FAB

Northwest AI HUB

PRISM

CHIMES
Center for Heterogeneous Integration of Micro Electronic Systems

NSF

DARPA
ERI – 3DSoC

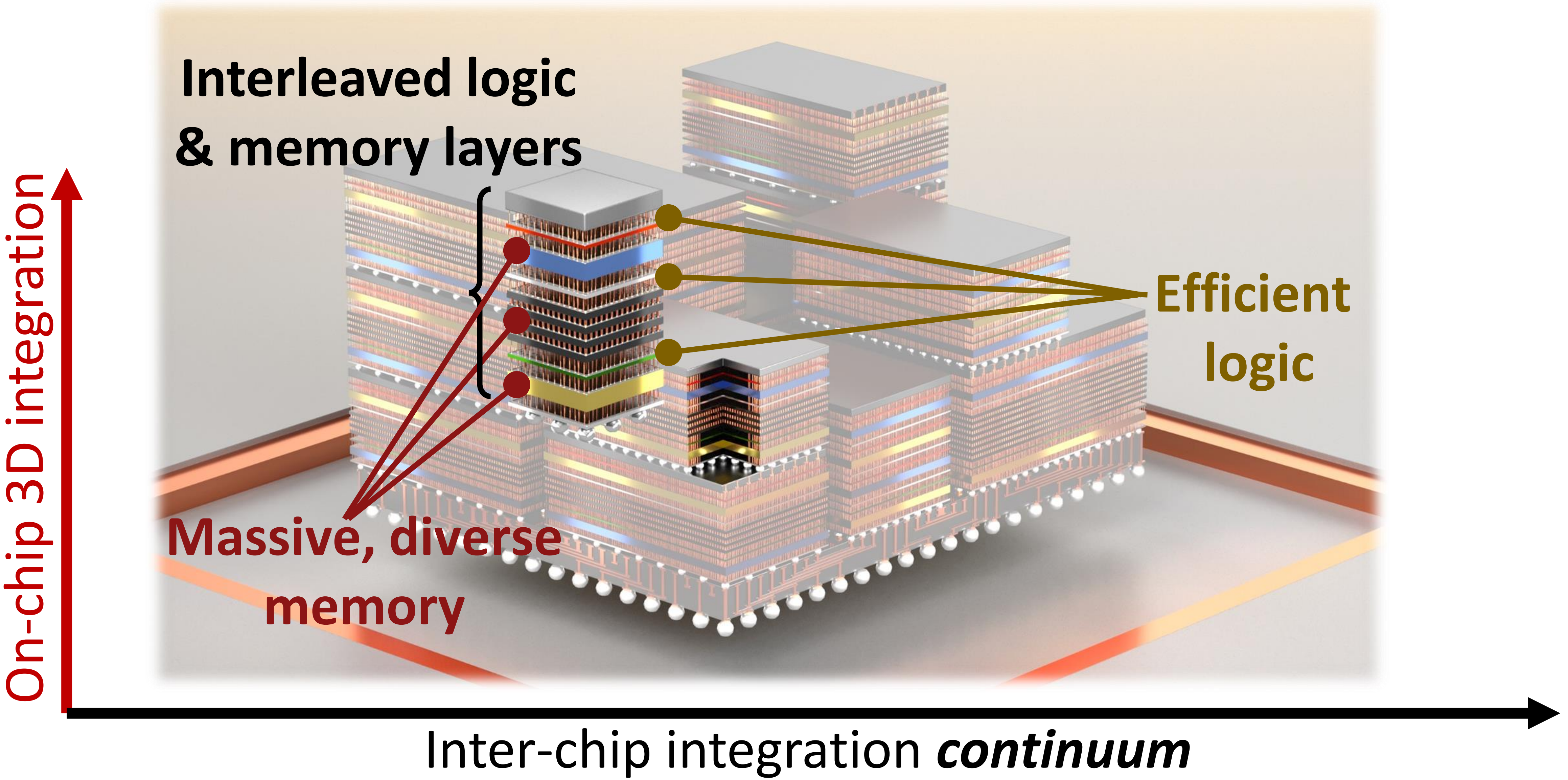
DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA

Latency Capacity Bandwidth Durability
DAM
Stanford Differentiated Access Memories Project

SystemX ALLIANCE
Stanford University

Stanford | NMTRI
NON-VOLATILE MEMORY TECHNOLOGY RESEARCH INITIATIVE

Continuum of Interconnection Density


Interleaved logic
& memory layers
Efficient
logic
Massive, diverse
memory
On-chip 3D integration
Inter-chip integration *continuum*