



SHAPING THE NEXT GENERATION OF ELECTRONICS

**JUNE 23-27, 2024**

MOSCONE WEST CENTER  
SAN FRANCISCO, CA, USA



**JUNE 23-27, 2024**

MOSCONE WEST CENTER  
SAN FRANCISCO, CA, USA

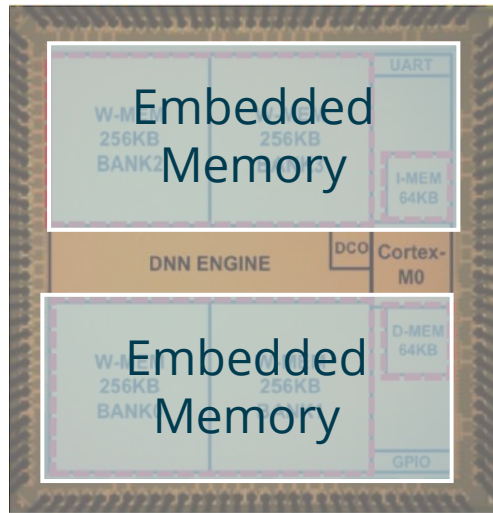
# GCRAM: The Highest Density Embedded Memory in Standard CMOS

Andreas Burg (Head of Technology), Eli Leizerowitz (CBO),  
Robert Giterman (CEO)

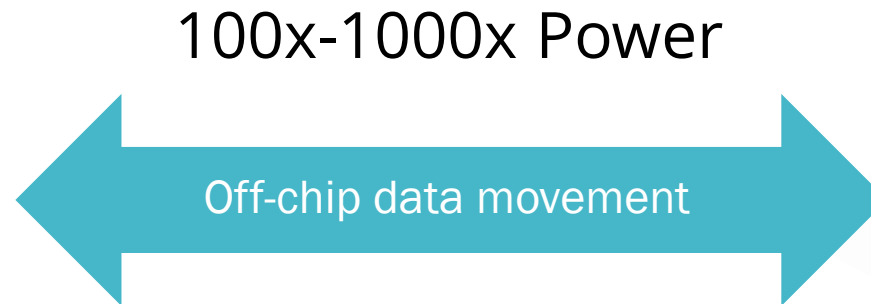
RAAM Memory Technologies  
(visit us at booth #1460)



# Memory is a Bottleneck



SoC



100x-1000x Lower Bandwidth

Significantly higher BOM and  
3rd party dependencies



External DRAM

**External memory should be avoided at all costs and if needed,  
access should be minimized**

# Embedded Memory Dominates Cost of Silicon



Cost of silicon is **proportional to area**, especially in high volume

**Memory dominates** the **die area** of almost all applications

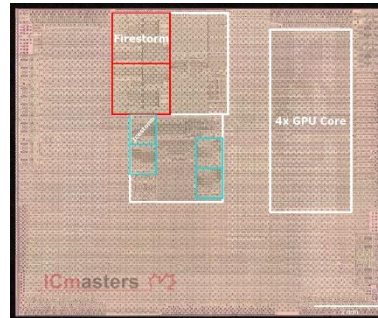


Automotive



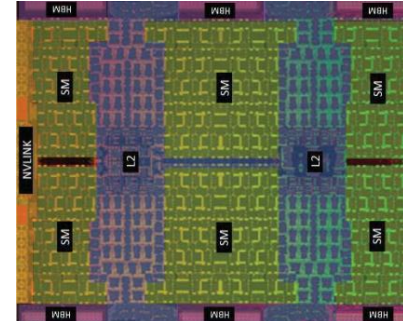
Tesla FSD, 14nm

Mobile



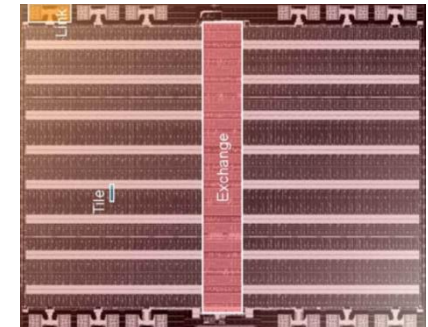
Apple A14, 5nm

Datacenter



Nvidia A100, 7nm

ML/AI & Server



Graphcore IPU, 7nm

Memory  
Area [%]

>35%

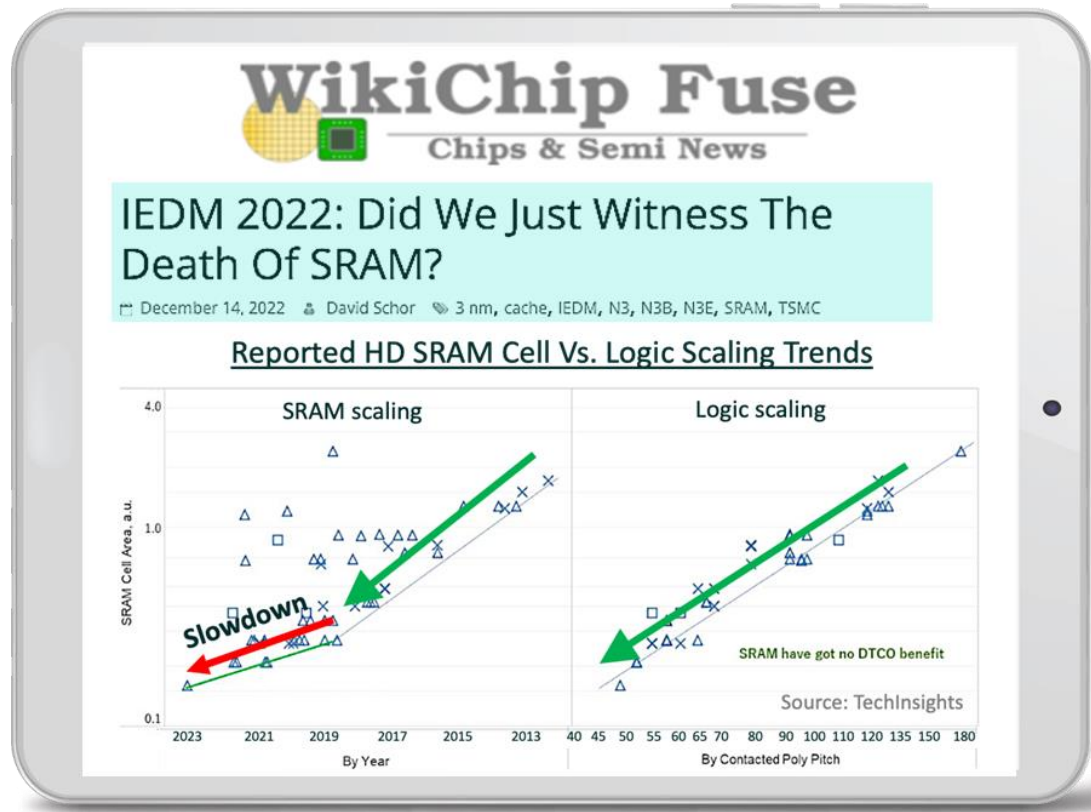
>45%

>60%

>70%

**System Cost of any Digital System is Significantly Reduced Through Memory Density Improvement**

# SRAM Scaling is Coming to an End

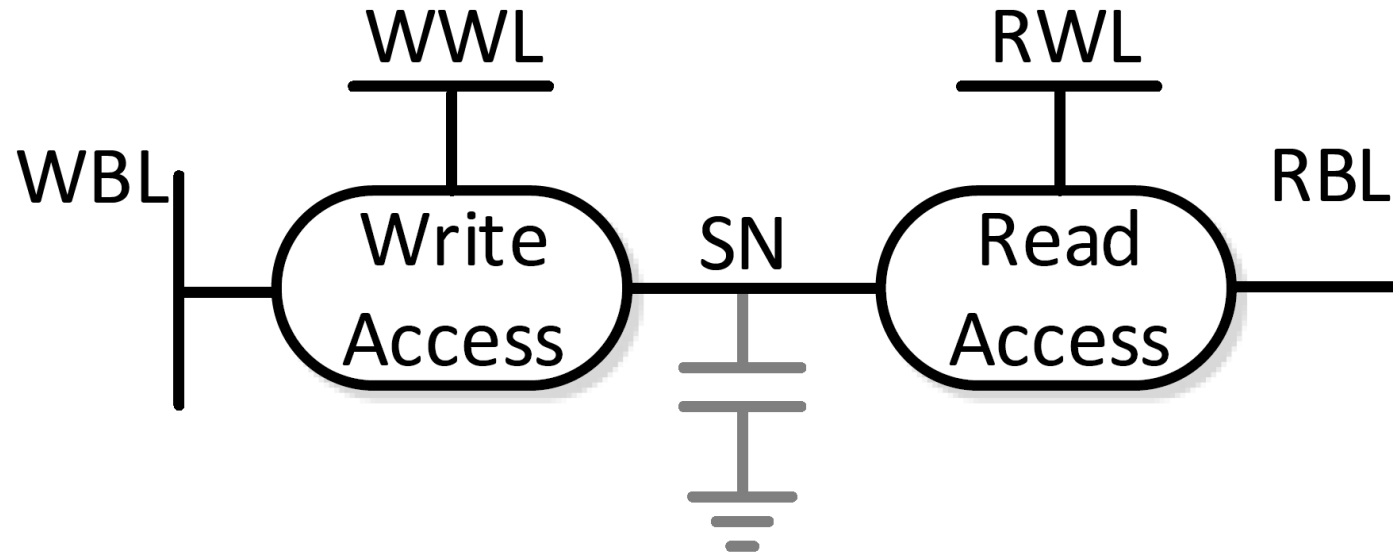


- Diminishing returns from decades of design-technology co-optimization with already pushed design rules that are less amenable to further scaling
- **SRAM Bitcell** is fundamentally sensitive to process variations and **requires complex assist techniques** and periphery, **degrading array efficiency**

“TSMC’s 3nm Node: No SRAM Scaling Implies More Expensive CPUs and GPUs”

**New solution is needed to address the memory bottleneck and the end of SRAM Scaling**

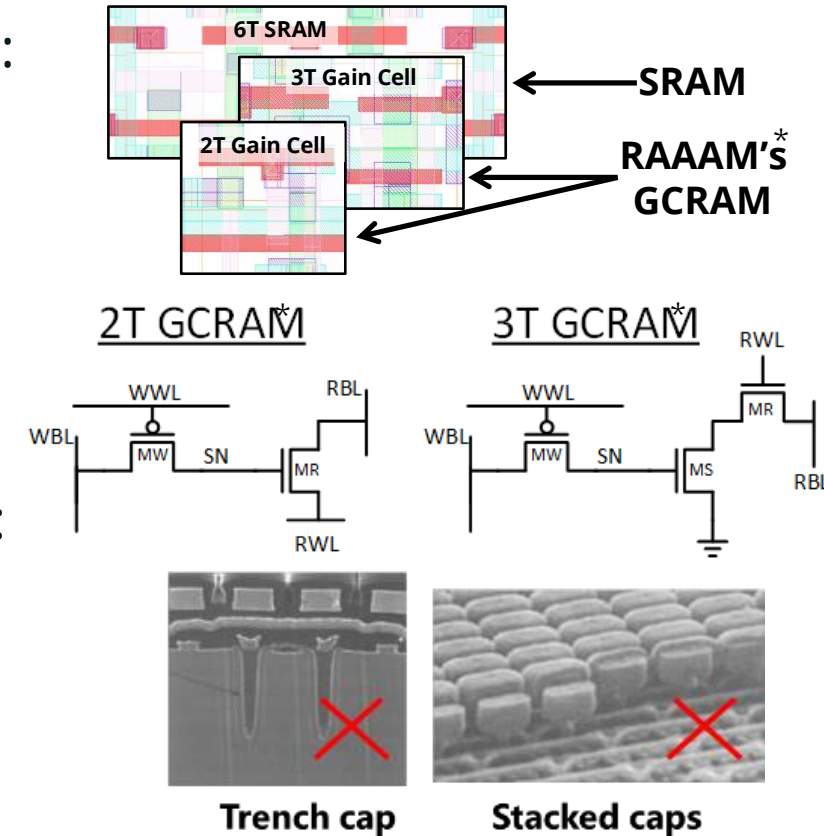
# Basic Concept of RAAAM's Gain Cell RAM (GCRAM)



Data is stored as charge with read- and write-access networks with 1-2 transistors each

# RAAAM's Gain-Cell RAM Technology

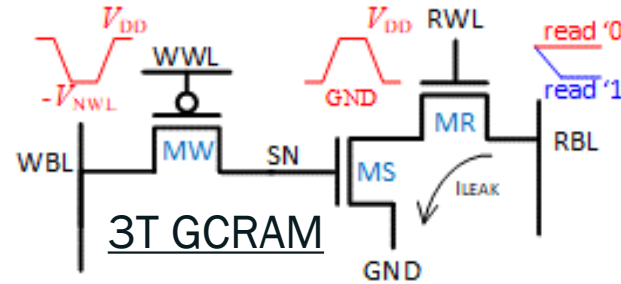
- GCRAM has significant advantages over SRAM and 1T-1C eDRAM:
  - **Up-to 50% smaller cell size** than SRAM
  - **Fully logic-compatible**, no extra cost
  - **Active read port – no charge sharing** as in 1T-1C eDRAM
  - Naturally supports **two ported operation**
- RAAAM's patented designs mitigate memory refresh by applying:
  - Circuit-level techniques for **bitcell leakage reduction**
  - Architecture-level refresh strategies for **100% memory availability**



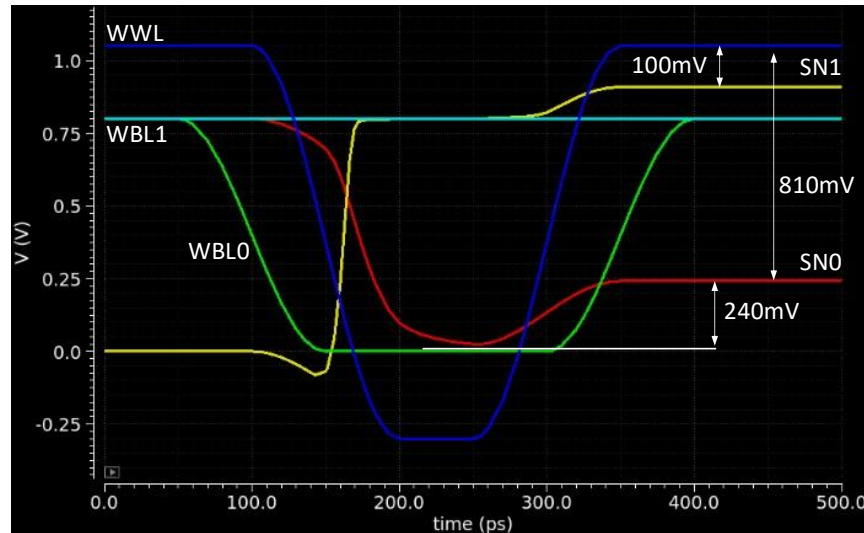
**GCRAM offers the advantages of 1T-1C eDRAM in a standard CMOS Process**

\* Schematics and layout for illustration purposes only. Final implementations may use different device flavors and signals.

# GCRAM Operating Principle

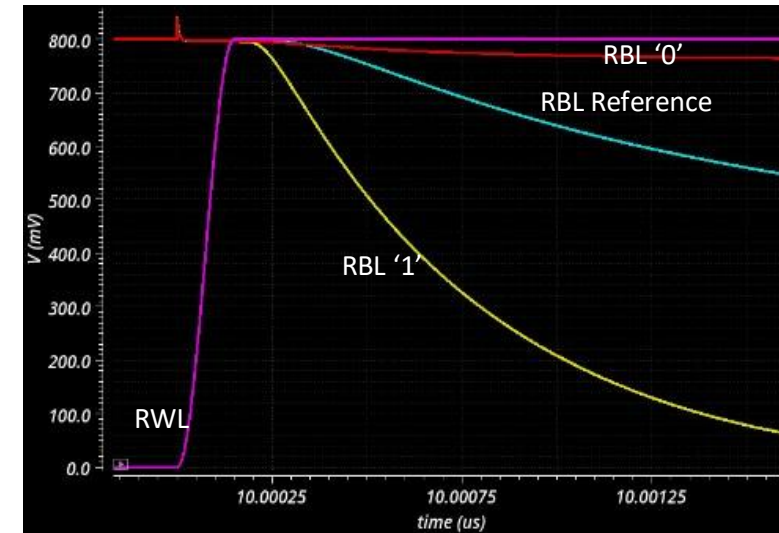


Write Mode



- Negative WWL pulse to overcome PMOS threshold drop for write '0'
- Boosted WWL during retention for reduced leakage

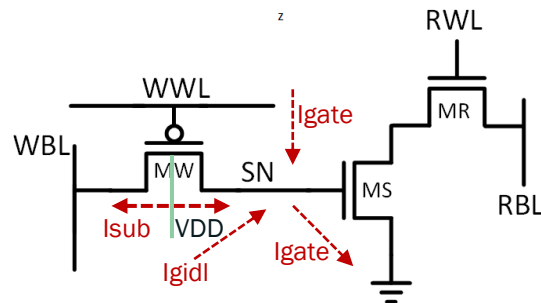
Read Mode



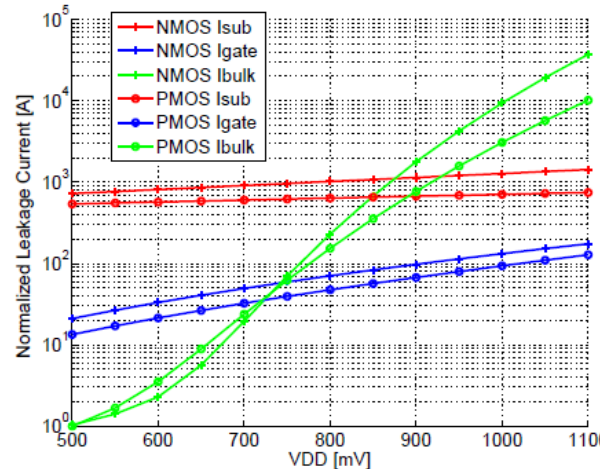
- RBL pre-charge to VDD prior to RWL assertion
- RWL pulse enables conditional RBL discharge
- Reference generated to enable differential readout

# Transistor Leakage in FinFET Nodes

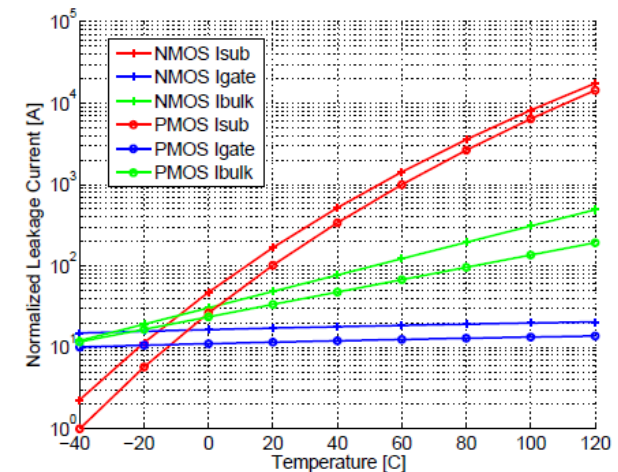
- GCRAM Data retention time is limited by the SN leakage distribution:
  - **Sub-threshold leakage via WBL – high temperature dependency + high variation**
  - **Gate Induced Drain Current (GIDL) – high VGD dependency + low variation:**  
Reduces with time over a retention period due to SN deterioration
  - **Gate tunneling via WWL and MS – low temperature dependency + low variation**



SN leakage components



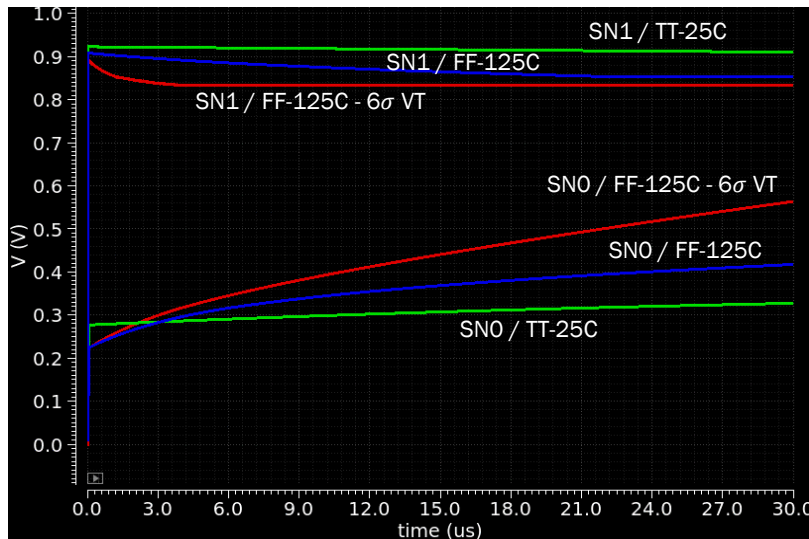
Leakage vs. VDD  
[Source: Shalom'ICECS18]



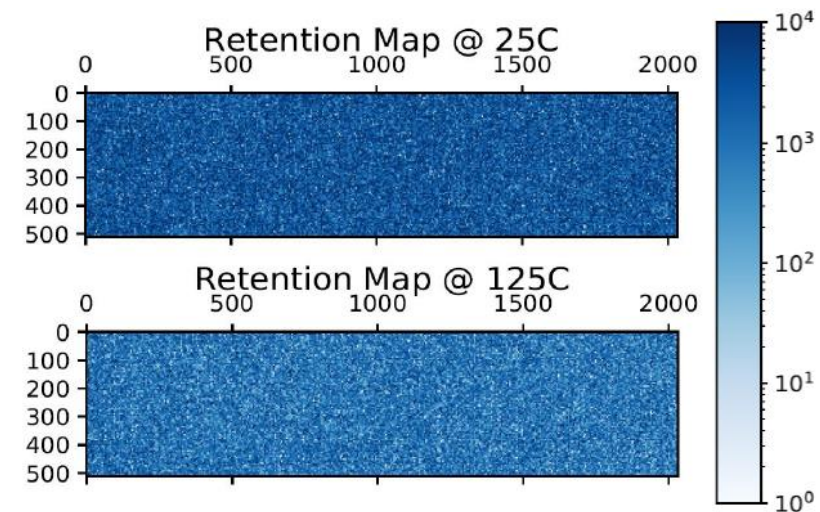
Leakage vs. Temperature  
[Source: Shalom'ICECS18]

# Leakage Impact on Data Retention Time

- **Data retention time follows a lognormal distribution**
  - Sub-threshold leakage variations is dominant
- **High temperatures** lead to **~10X data retention time reduction**
  - Exponential increase in sub-threshold leakage and GIDL



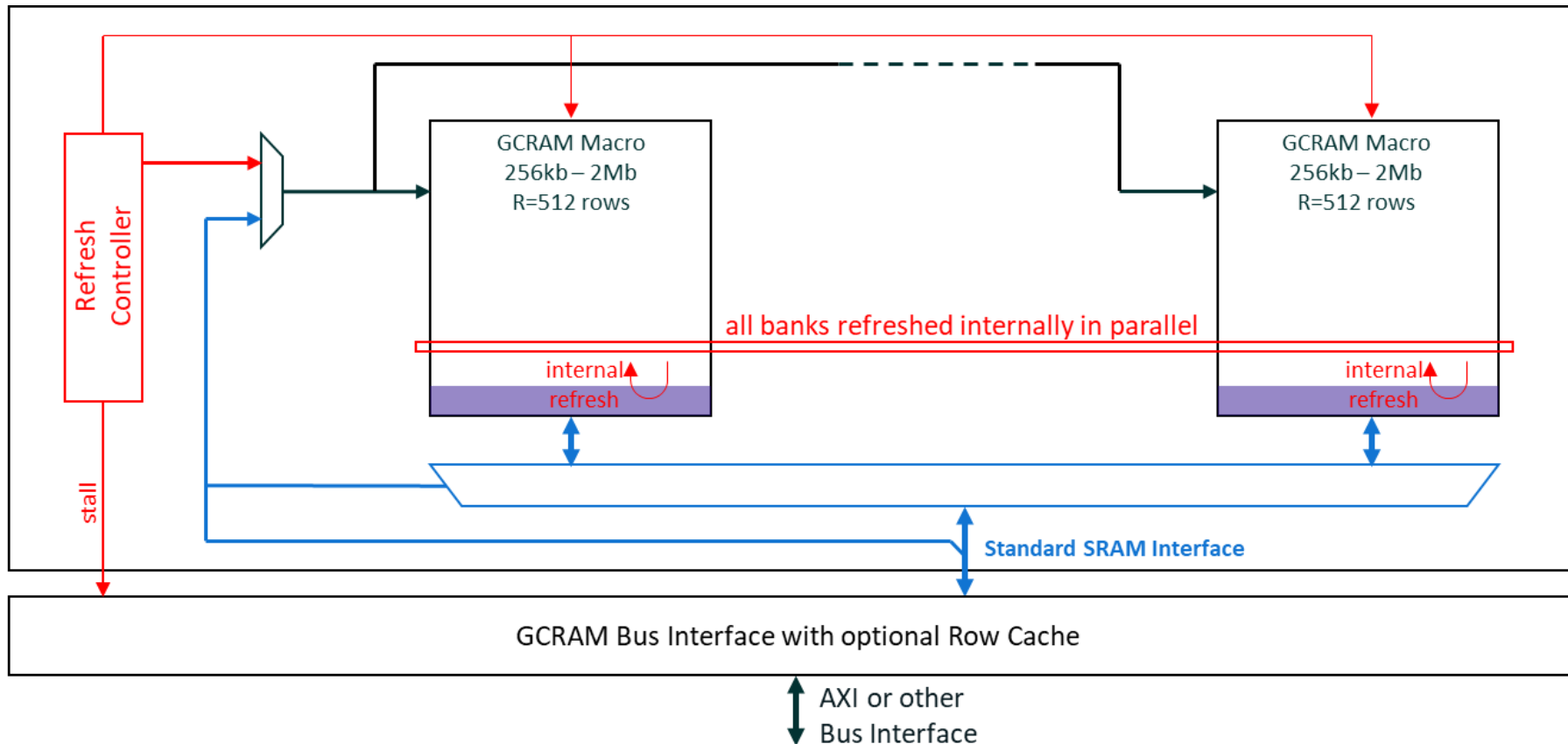
SN deterioration following a write operation for TT-25C, FF-125C, and FF-125C - 6 $\sigma$  VT shift on the write transistor for worst-case leakage



16nm Retention Time Distributions  
[Source: Giterman'ESSCRIC21]

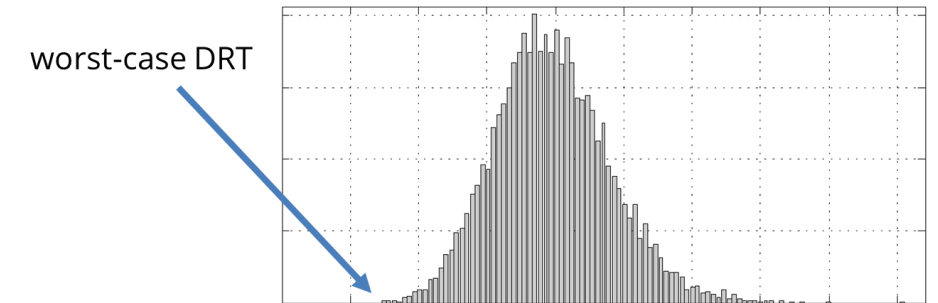
# GCRAM System Integration and Refresh

- Multiple GCRAM macros managed by a shared memory/refresh controller
- **Memory is refreshed row-by-row, but in parallel across multiple macros,** reducing refresh overhead



# GCRAM Refresh Overhead is Negligible

- Data retention time (DRT) varies significantly across PVT variations
- The **refresh period** is **set** according **to** the **worst-case** possible **DRT**
  - Can be obtained through high-sigma analysis or
  - On-chip memory BIST

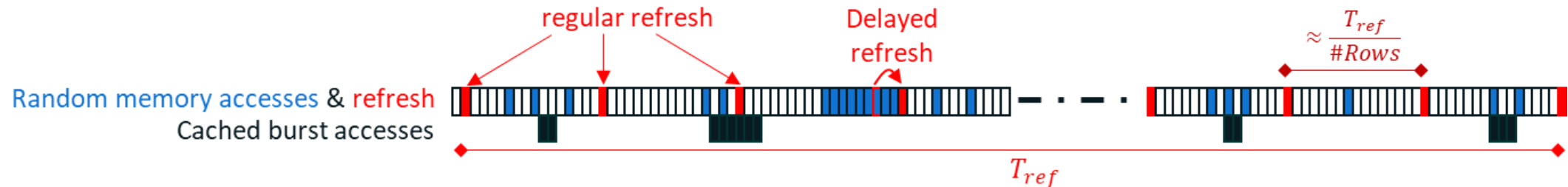


- The **memory availability** is defined as  $1 - \frac{T_{clk}}{T_{ret}} N_r$ 
  - Typical DRT:  $T_{ret} = 100\mu s$
  - Typical access / refresh cycle time:  $T_{clk} = 2ns$
  - Typical number of array rows:  $N_r = 512$

**Typical array  
availability: ~99%**

# GCRAM Refresh Optimization

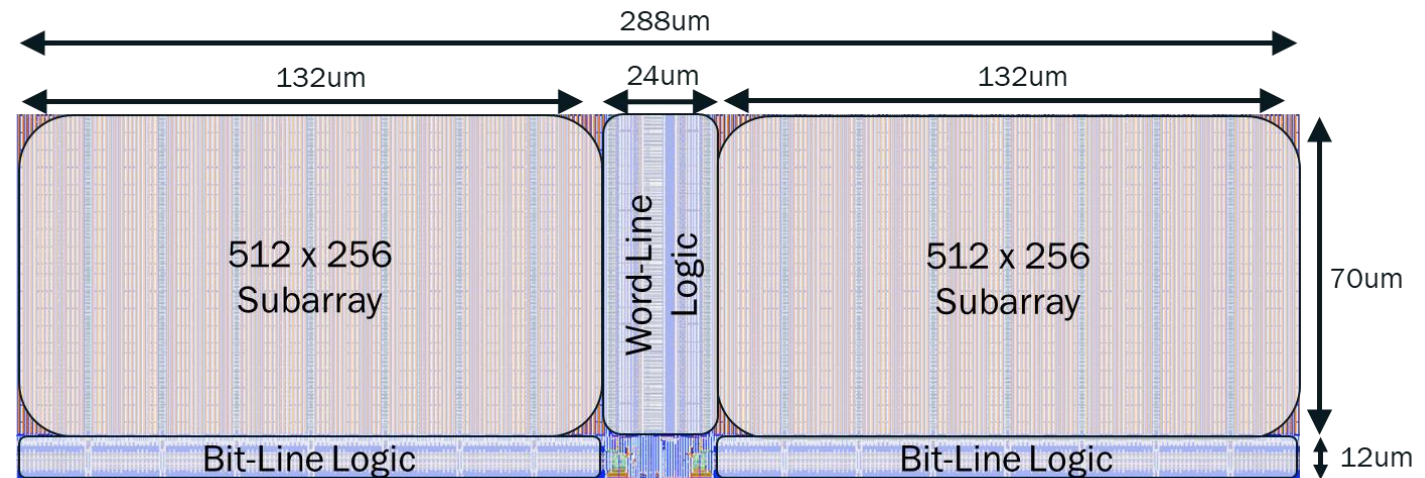
- Smart memory controllers provide potential for further optimization
  - On chip PVT sensors, adapted to GCRAM set refresh period based on PVT operating conditions
  - Tolerance to irregularities in refresh schedule and multiple macros allow to avoid collisions with most access cycles



# 16nm GCRAM Memory Macro Block Diagram

- **512x256 subarrays (x2)** featuring 512x256 bitcell arrays, local RWL pull-down drivers, replica column, and reference rows
- **Word-Line Logic** featuring word-line (WWL / RWL) post decoder, latches, WWL drivers (including write / retention assists), and RWL drivers
- **Bit-Line Logic** featuring RBL pre-charge, sense amplifiers, output buffers, WBL drivers, I/O latches
- **Global Logic** feature pre-decoder, timing generation, I/O sampling, charge pump
- **Single-supply macro design**

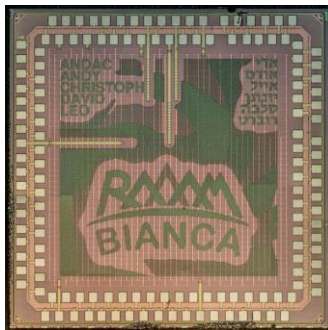
>78%  
Bitcell Utilization



# Recent Successful 16nm FinFET Tapeouts

Technology	Academic GCRAM 3T	RAAAM's GCRAM 2T	HD SRAM 6T	RAAAM's GCRAM 3T	HD SRAM 6T
Macro Capacity	256kbit	256kbit	256kbit	512kbit	512kbit
No. of Supply Voltages	4	1	1	1	1
Area Per Bit [um <sup>2</sup> ] (including periphery)	0.125 (logic rules)	0.09 (logic rules) 0.053 (SRAM rules)	0.109 (SRAM rules)	0.083 (logic rules) 0.058 (SRAM rules)	0.102 (SRAM rules)
Retention Time [us]	< 1us @ 125C	6us @ 105C	NA	35us @ 125C	NA
Cycle Time @ 25C [ns]	2.5ns	< 2ns (refresh )	< 1ns	< 1.5ns (refresh)	< 1ns
Leakage Current [uA]	43.75 / 256Kb @ 25C	244 / 256Kb @ 25C	> 5 / 256Kb @ 25C	6.33 / 512Kb @ 25C	> 10 / 512Kb @ 25C
Read Current / bit [uA / MHz]	0.067 @ 25C	0.027 @ 25C	> 0.07 @ 25C	0.023 @ 25C	> 0.07 @ 25C
Write Current / bit [uA / MHz]	0.064 @ 25C	0.022 @ 25C	> 0.07 @ 25C	0.021 @ 25C	> 0.12 @ 25C
Total Current* [uA] / 10% 512bit read @ 500 MHz	8003 @ 25C	2269 @ 25C	> 3500 @ 25C	1184 @ 25C	> 3500 @ 25C

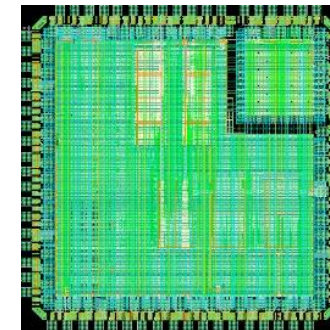
\* Total Current defined as:  $I_{total} = I_{Leakage} + I_{refresh} + \alpha_{read} \times f_{clk} \times I_{read/row}$



RAAAM's 2T 16nm Test Vehicle (2022)



Measurement Setup



RAAAM's 3T 16nm Test Vehicle (2023)

# GCRAM Scales to 5nm FinFET

Technology	GCRAM	HDSRAM	GCRAM	HDSRAM
Memory Capacity	512kbit		512kbit	
Number of Bitcell Ports	2-Port	1-Port	2-Port	1-Port
Corner	TT-0.75V-25C		FF-0.825V-125C	
Bitcell Size (normalized)	0.53	1	0.53	1
Memory Size (normalized)	0.57	1	0.57	1
Minimum Cycle Time (ns)	0.81	0.72	0.66	0.57
Data Retention Time [us]	100	-	30	-
Read Current / bit (uA/MHz normalized)	0.44	1	0.46	1
Write Current / bit* (uA/MHz normalized)	0.9	1	0.91	1
Total Current** – 10% read @ 700MHz (uA normalized)	0.76	1	0.7	1

# Conclusions

- Memory is an essential component for all digital ICs, often occupying >50% of the area (for ML/AI often even more) and dominating power.
- Scaling of SRAM has slowed down significantly below 16nm and has almost stopped completely below 5nm
- GCRAM offers a significant (50%) area advantage and much better power/energy then SRAM
  - The **key is a new bit-cell** that combines advantages of dynamic and static storage
  - **No additional process steps** allow integration with any standard process
  - The **refresh overhead is negligible** and can be fully mitigated even with a very basic controller