



SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA





SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA



Energy Efficient Memory

Sushil Sakhare (PhD)

(ssak@veevx.com)



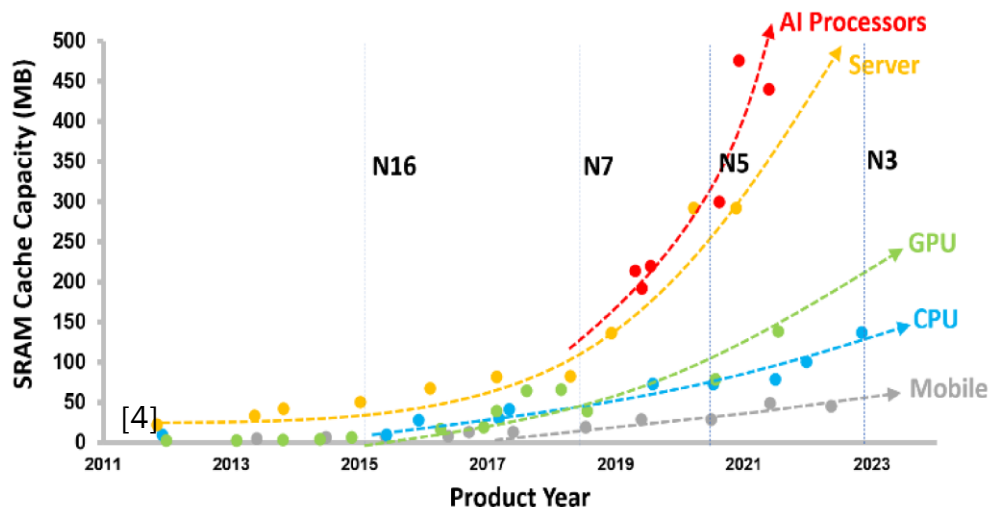
Agenda

- Motivation:
 - Rocketing cache demand Vs SRAM scaling
 - Future demands energy efficiency
 - Build large energy-efficient caches
 - MRAM recovers Avg. power of IOT systems
- Solution: iRAM
 - Energy-efficient chiplet memory
 - Overcomes standard MRAM deficiencies
- iRAM design
 - Chiplet Architecture
 - Fastest read access time [3ns]
 - Write repulsing & self-timed auto-tracking
- Conclusion

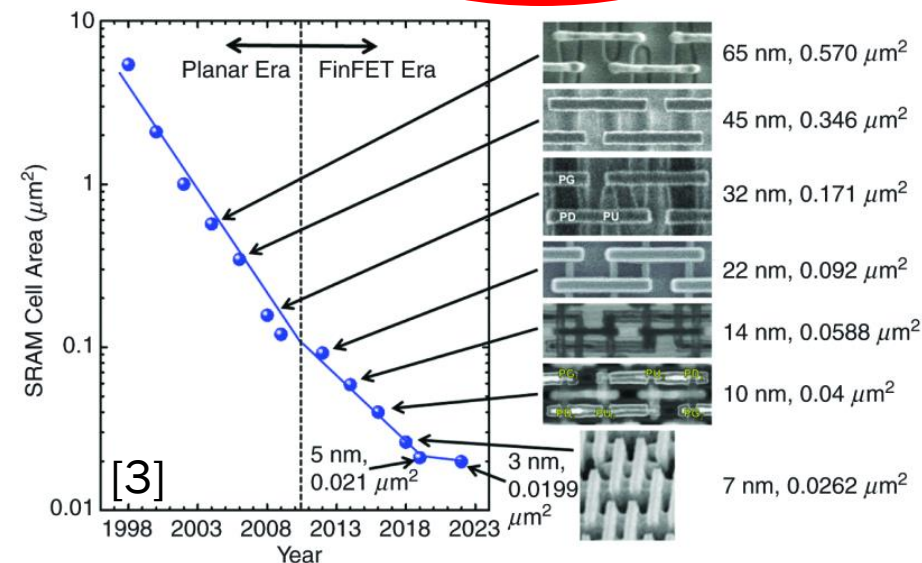
Agenda

- Motivation:
 - Rocketing cache demand Vs SRAM scaling
 - Future demands energy efficiency
 - Build large energy-efficient caches
 - MRAM recovers Avg. power of IOT systems
- Solution: iRAM
 - Energy-efficient chiplet memory
 - Overcomes standard MRAM deficiencies
- iRAM design
 - Chiplet Architecture
 - Fastest read access time [3ns]
 - Write repulsing & self-timed auto-tracking
- Conclusion

Rocketing cache demand Vs ~~SRAM scaling~~



Rocketing cache demand



Dead SRAM scaling

- SRAM is widely used as the on-chip cache for CPU/GPU and AI/ML SoCs.
 - Growing demand for SRAM cache memory in modern computing chips [4].
- AMD' 3D V-Cache, enables the 96MB giant LLC for HPC
 - 2nd generation of 3D V-Cache uses a less advanced 7-nm node for SRAM dies while the processor cores are on a more advanced 5-nm node[3].

- SRAM scaling 5-nm to 3-nm node 5% area reduction is achieved [1-2].
- Challenging to realize[4]
 - Cell & Macro area.
 - Minimum Operating voltage(V_{min})
 - Macro speed
 - Yield at High densities.
 - Leakage power

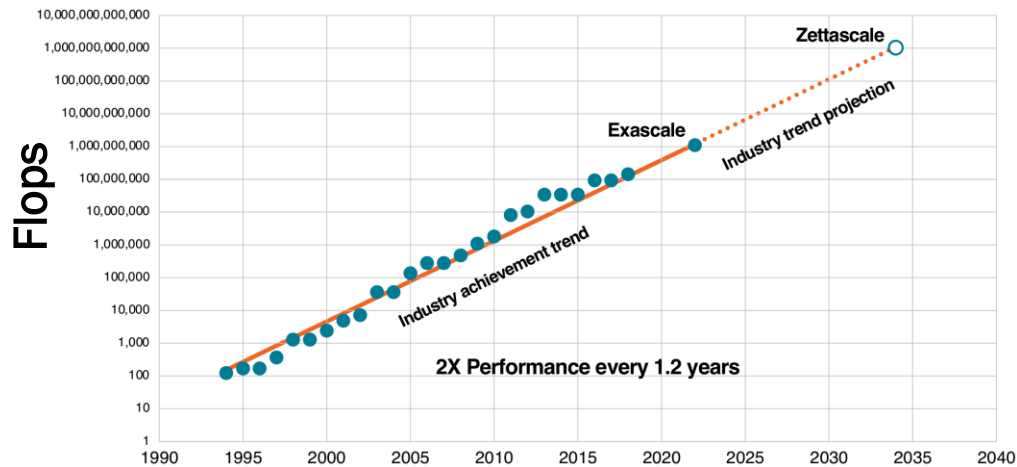
[1] S. Yu and T.-H. Kim, "Semiconductor Memory Technologies: State-of-the-Art and Future Trends," in *Computer*, vol. 57, no. 4, pp. 150-154, April 2024, doi: 10.1109/MC.2024.3363269

[2] C.-H. Chang et al., "Critical process features enabling aggressive contacted gate pitch scaling for 3nm CMOS technology and beyond", *Proc. Int. Electron Devices Meeting (IEDM)*, pp. 27.1.1-27.1.4, 2022.

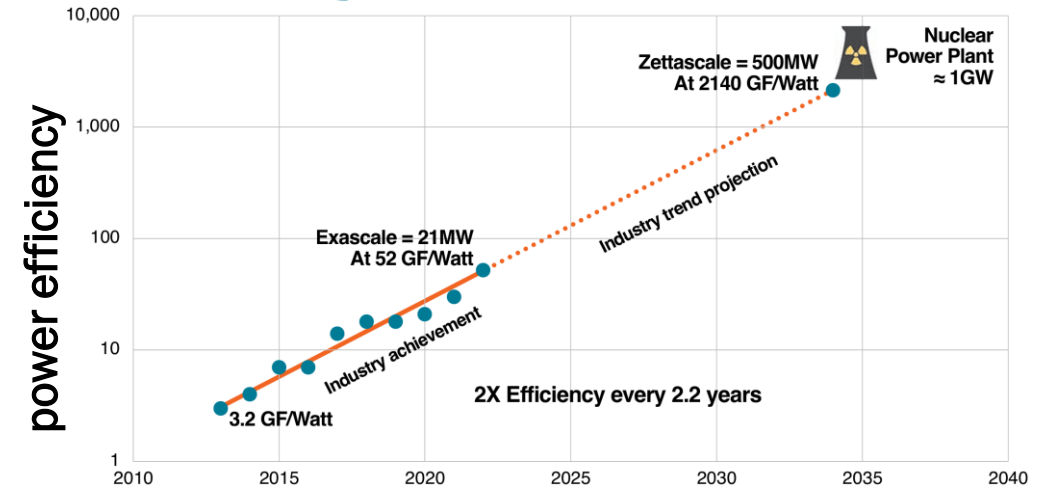
[3] J. Wu et al., "3D V-Cache™: The implementation of a hybrid-bonded 64MB stacked cache for a 7nm x86-64 CPU", *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 428-429, 2022.

[4] Y. Wang et al., "High-Speed Embedded Memory for AI and High-Performance Compute," *2023 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2023, pp. 1-4, doi: 10.1109/IEDM45741.2023.10413818.

Future demands energy efficiency



The trend in High Performance Linpack Flops of top performing super-computers on top500.org list [1]



The trend in the power efficiency of top-performing super-computers on top500.org list[1]

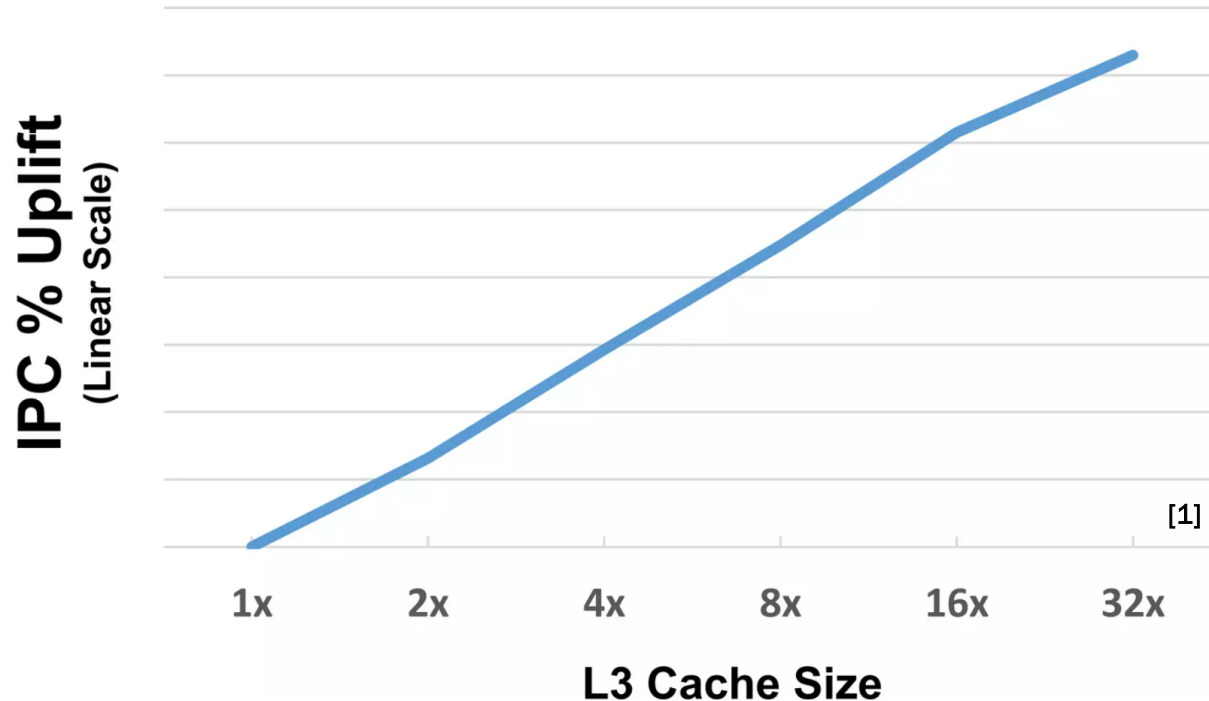
- The rate of improvement in energy efficiency (Gigaflops per watt) is less than half that of the rate of performance increase over the past two decades [1].
- Industry Goal: Zettaflop of computing in ~10 years
 - However, it would require 0.5GW of electricity equivalent to half a nuclear power plant (not a viable solution)[1].
 - Thus, the Industry must embark on
 - New processor architectures
 - Reduce communications overhead by innovating in system design and more efficient on-package and off-package communication
 - Techniques to bring **giant energy-efficient memory closer to compute** [1-3].

[1] L. Su and S. Naffziger, "1.1 Innovation For the Next Decade of Compute Efficiency," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 8-12, doi: 10.1109/ISSCC42615.2023.10067810.

[2] <https://www.slideshare.net/AMD/3d-vcache>

[3] J. Wu et al., "3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU," 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2022, pp. 428-429.

Build a large, energy-efficient caches.



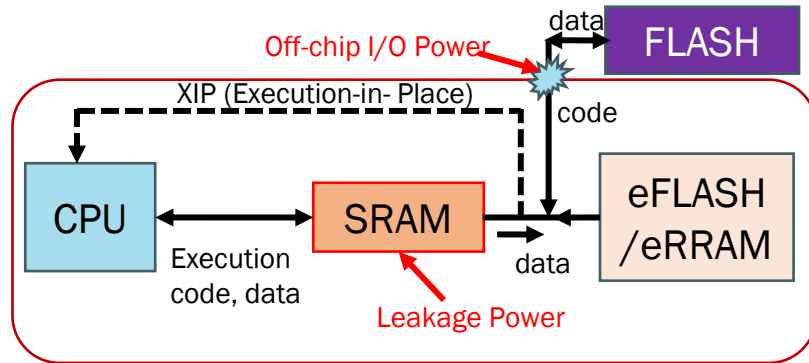
- Large L3 cache uplifts the Instruction Per Clock Cycle (IPC) linearly.
- Cache consumes a significant portion of system power [3].
- Thus, build a large, energy-efficient L3 cache.

[1] <https://www.slideshare.net/AMD/3d-vcache>

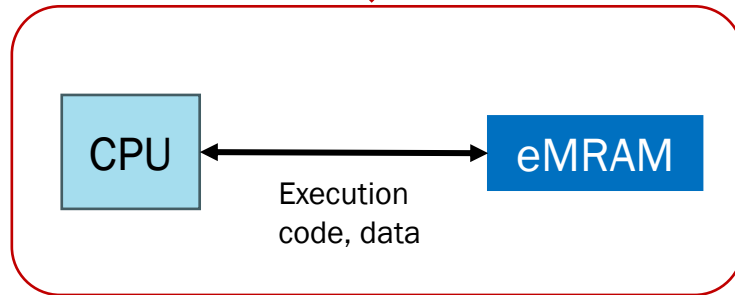
[2] J. Wu et al., "3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU," 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2022, pp. 428-429.

[3] S. Thoziyoor, J. H. Ahn, M. Monchiero, J. B. Brockman and N. P. Jouppi, "A Comprehensive Memory Modeling Tool and Its Application to the Design and Analysis of Future Memory Hierarchies," 2008 International Symposium on Computer Architecture, 2008, pp. 51-62

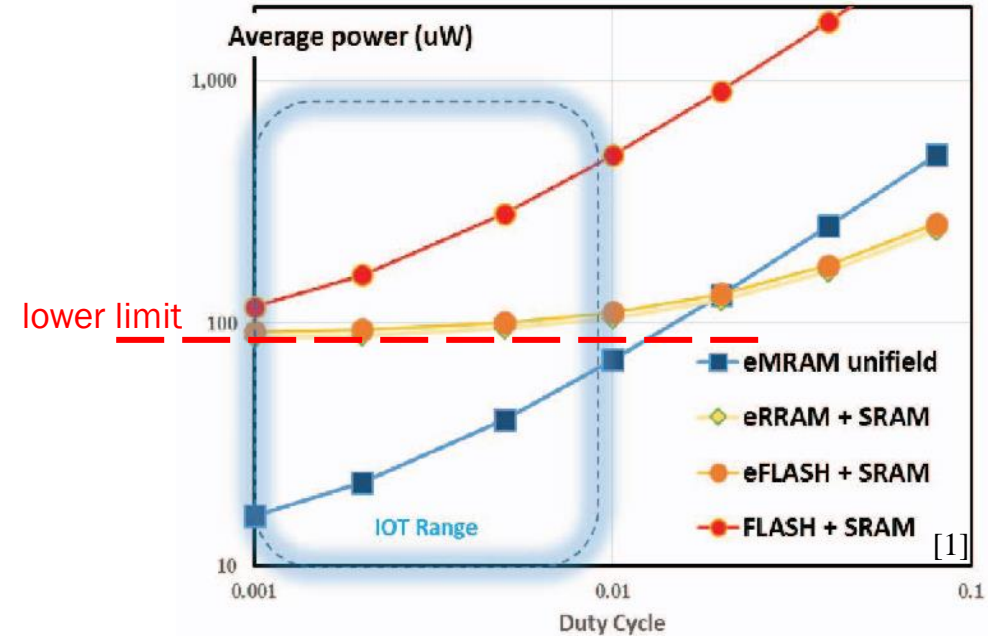
MRAM recovers Avg. power of IoT Systems



conventional memory subsystem in an IoT device [1]



A unified MRAM in an IoT device by Qualcomm[1]



• MRAM in IOT:

- Qualcomm in [1] showed ultra-low average power using MRAM.
 - The leakage from an SRAM macro has set the lower limit of average power as shown in the graph.
- Vega [2]: Always-on IoT SoC achieved
 - Up to 32.2GOPS (@49.4mW) peak performance on near-sensor analytics algorithms (NSAAs), including mobile DNN inference.
 - exploiting 1.6MB of state-retentive SRAM and 4MB of non-volatile MRAM.

[1] Y. Lu et al., "Fully functional perpendicular STT-MRAM macro embedded in 40 nm logic for energy-efficient IoT applications," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, 2015, pp. 26.1.1-26.1.4, doi: 10.1109/IEDM.2015.7409770.

[2] D. Rossi et al., "4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2021, pp. 60-62, doi: 10.1109/ISSCC42613.2021.9365939

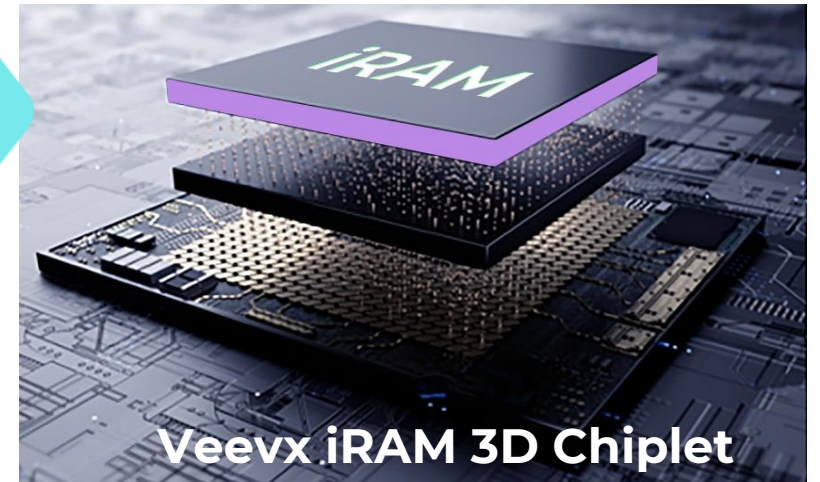
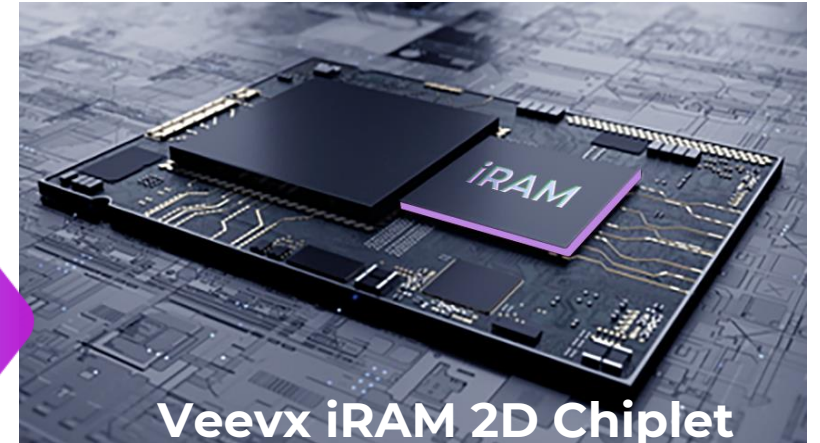
Agenda

- Motivation:
 - Rocketing cache demand Vs SRAM scaling
 - Future demands energy efficiency
 - Build large energy-efficient caches
 - MRAM recovers Avg. power of IOT systems
- Solution: iRAM
 - Energy-efficient chiplet memory
 - Overcomes standard MRAM deficiencies
- iRAM design
 - Chiplet Architecture
 - Fastest read access time [3ns]
 - Write repulsing & self-timed auto-tracking
- Conclusion

iRAM – Energy-efficient Chiplet Memory

Veevx's 16nm node Chiplet (@TSMC)

High speed operation	3ns read & 25ns write
Ultra-high density	16Mb/mm ²
deep sleep leakage	15nA for 32Mb (1/50,000 th of SRAM)
Standby Leakage	1uA/Mb (1/25 th of SRAM)
Endurance	>10 ¹²
Retention	10 year @ 125 °C
deep sleep wake-up time	<200ns (quickest recovery)
One Time Programmable (OTP) bits	256/Mb
CPU	Arm Cortex-M0+
Selectable & Programmable Bus Interfaces	
User-defined start-up code written in MRAM for CPU	Can be a boot code or any program code
Chip ID realized in MRAM	PUF
Supports Periodic Data Integrity Checking	In field repair



iRAM- overcomes standard MRAM deficiencies

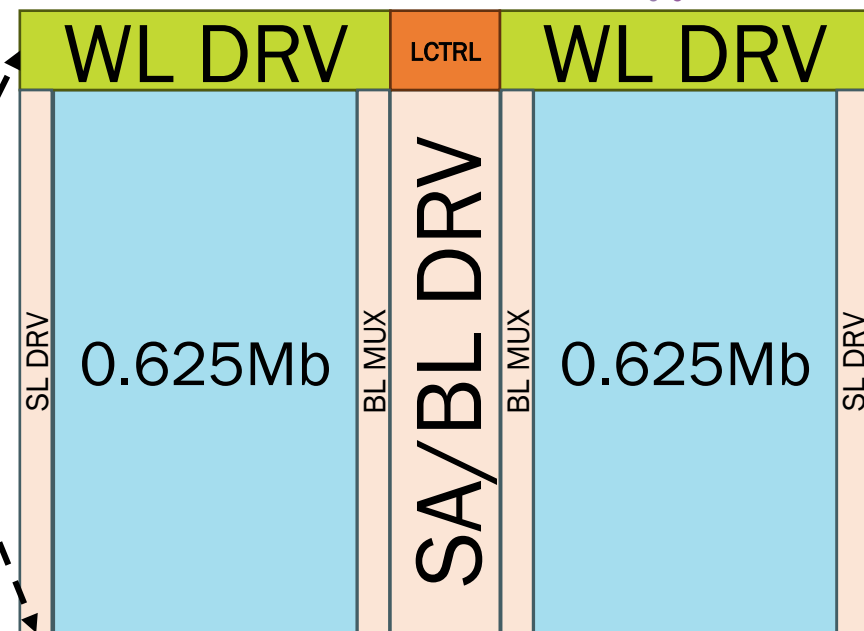
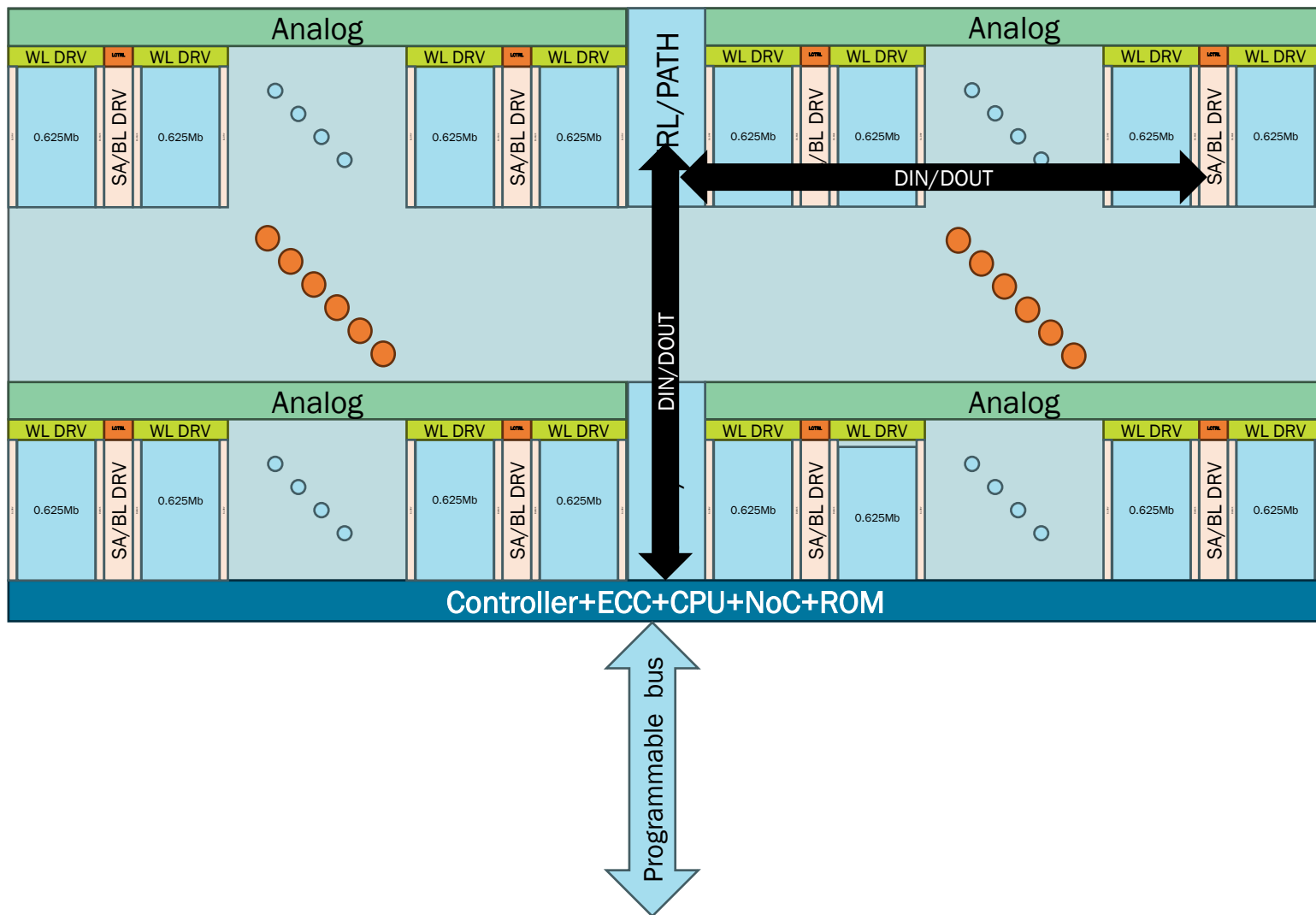


- Fastest read performance in the market
 - Innovative reference generation
 - High performance & accurate sense amplifier.
 - Self-timing tracking schemes
 - Innovation in design
 - Improved write performance & reliability
 - Fastest deep-sleep wakeup and tiny standby power.
 - Configurable Chiplet interfaces
- Results:
 - IPC uplift
 - Energy-efficient giant L3/L4 caches

Agenda

- Motivation:
 - Rocketing cache demand Vs SRAM scaling
 - Future demands energy efficiency
 - Build large energy-efficient caches
 - MRAM recovers Avg. power of IOT systems
- Solution: iRAM
 - Energy-efficient chiplet memory
 - Overcomes standard MRAM deficiencies
- iRAM design
 - Chiplet Architecture
 - Fastest read access time [3ns]
 - Write repulsing & self-timed auto-tracking
- Conclusion

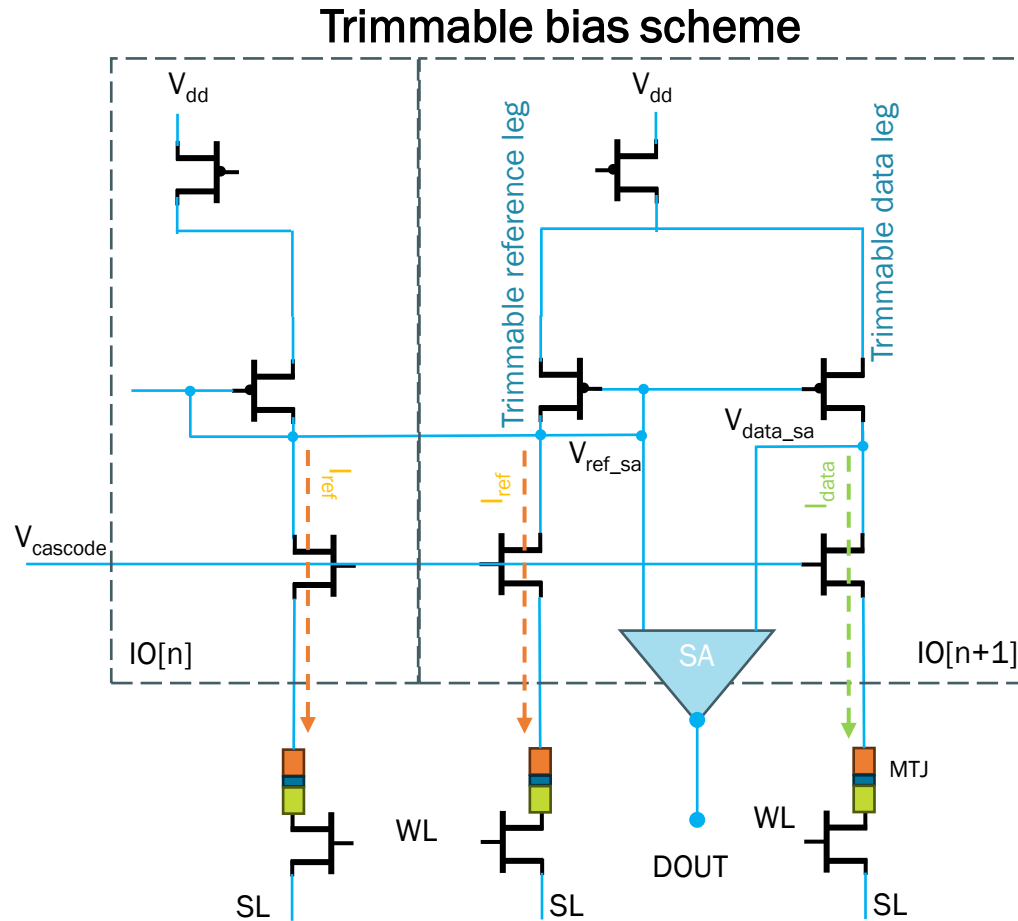
iRAM Chiplet architecture



• Array Optimization:

- Structural optimization of word-line (WL) and bit-line (BL) routing to mitigate parasitic.
- Source line is shared in an IO.
- Common single-sided word-line driver for read & write.
- Each block has two planes of 0.625Mb.
- BL mux is placed about the common-sense amplifier & BL driver.

Fastest read access [3ns]



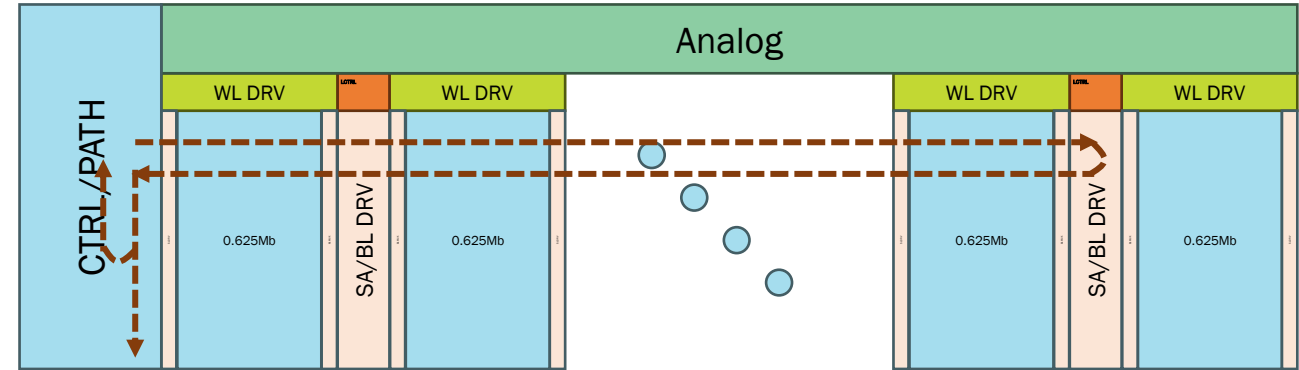
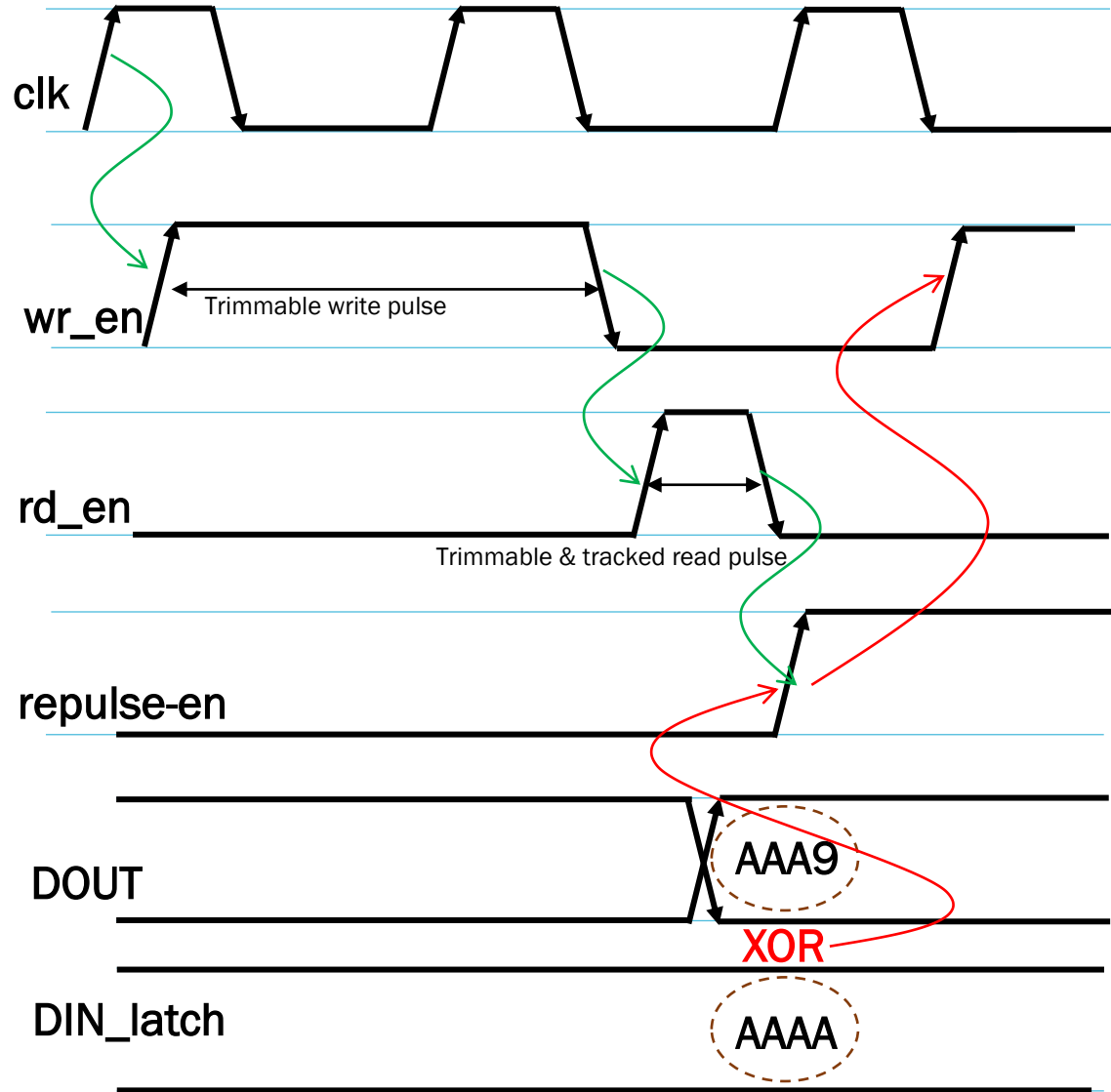
- Trimmable bias scheme

- A simple current mirror is used to reflect the I_{ref} from the reference leg to the data leg,
- The reference legs between neighboring IO cells are shorted to generate the common V_{ref_sa} voltage.
- Current through data leg sets the V_{data_sa} voltage based on the resistance offered by the MTJ
- Analog voltage $V_{cascode}$ limits the voltage across the MTJ.
- Trimming is used to mitigate the input offset of SA. The trim algorithm runs within 60ns of wakeup.

- High-speed sense amplifier (SA).

- A High-speed latch-type sense amplifier is used to read logic 0/1.
- Inputs are equalized before building differential at the input of a sense amplifier.

Write repulsing & self-timed auto-tracking



- The read cycle follows the write cycle during the write mode.
- The DOUT is XORed with latched DIN
 - If it does not match, memory starts writing the DIN again by asserting wr_en.
 - 4-bit repulse_cnt can set up to 16 repulse cycles.
 - Repulsing is self-timed to ensure that all DOUTs reach the data path and are compared to generate the repulse-en.
 - The write voltage is set by tracking the ambient temperature.
 - During the entire write operation MRAM generates a 'Busy' signal for handshake.

Agenda

- Motivation:
 - Rocketing cache demand Vs SRAM scaling
 - Future demands energy efficiency
 - Build large energy-efficient caches
 - MRAM recovers Avg. power of IOT systems
- Solution: iRAM
 - Energy-efficient chiplet memory
 - Overcomes standard MRAM deficiencies
- iRAM design
 - Chiplet Architecture
 - Fastest read access time [3ns]
 - Write repulsing & self-timed auto-tracking
- Conclusion

Conclusion: iRAM

- Rocketing cache demand & SRAM scaling is dead
- Future demands energy efficiency
- Large L3 caches provide IPC uplift
- Thus, build a large energy-efficient L3/L4 cache.
- iRAM is energy & density efficient solution.
 - Reverses the trend: **Improving performance while lowering power**

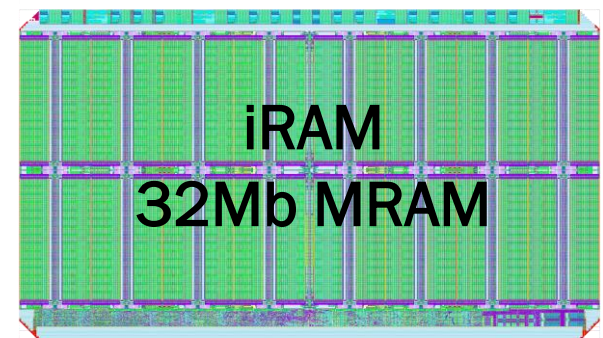


veevx

For further information contact:

Dana McCarty

dmccarty@veevx.com





**THE CHIPS
TO SYSTEMS
CONFERENCE**

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA



Thanks

For further information contact:

Dana McCarty

dmccarty@veevx.com

