Can SOT-MRAM replace SRAM in modern HPC CPUs?

-- A case study utilizing gem5 and STREAM

Tamao Takahashi
RIKEN R-CCS, University of Tsukuba
s2310594@u.tsukuba.ac.jp

Jens Domke, Shigeki Tomishima
RIKEN R-CCS

Motivation

Overcoming the Memory Wall with SOT-MRAM

- The Roofline model generally illustrates the system efficiency, where application performance is characterized by being either “Memory Bound” or “Compute Bound”, and it highlights the current problem which HPC and AI workloads are facing. The true “Memory Wall”, the left-hand side of the roofline, refers to the system performance limitation caused by memory access bandwidth and capacity constraints, stemming from the limitations of existing memory devices, such as SRAM, DRAM, HBM, and Flash Memory.
- SRAM has been widely used for cache memory for several decades. While it has fast access speeds, its large cell size makes it difficult to increase capacity.
- Therefore, researchers have proposed using SOT-MRAM for cache. Since SOT-MRAM is non-volatile, it has near-zero leakage power, which is critical for reducing static energy, especially in large-scale caches. For this reason, we aim with our architecture and workloads simulations to demonstrate that SOT-MRAM is a potential memory candidate for HPC.

Why SOT-MRAM?

- High Density: Small cell size allows for larger cache capacity within the same chip area compared to SRAM.[2]
- Non-Volatility: Zero leakage power enables energy-efficient "normally-off" computing.[1][2][3]
- High Speed: Offers faster switching speed than STT-MRAM, making it suitable for L2/L3 caches.

Feature	SRAM	DRAM	SOT-MRAM
Non-Volatility	No	No	**Yes**
Cell Size F^2	Large (~146)	Small (~6)	**Medium (~46–60)**
Leakage Power	High	High (Refresh)	**Near Zero**
Read Latency	Very Fast (<1 ns)	Slow (~30 ns)	**Fast (~3 ns)**
Write Latency	Very Fast (<1 ns)	Slow (~30 ns)	**Fast (~5 ns)**
Endurance	Unlimited (10^{16})	Unlimited	**High (>10^{12})**

Comparison between SRAM, DRAM and SOT-MRAM[1]

Method

Simulation Environment

- gem5: Cycle-accurate full-system simulator for architecture modeling.
- STREAM: Benchmark for measuring sustained memory bandwidth.
- In this study, we simulated a DerivO3CPU architecture including SOT-MRAM for L2 and L3 caches using gem5 and compared the outputs against a baseline.

- Higher Density: The compact nature of SOT-MRAM's 3-terminal structure, compared to the conventional 6-transistor (6T) SRAM.[3]
- →We simulated the SOT-MRAM configuration with double the capacity of the SRAM baseline.

Level	SRAM Baseline Configuration	SOT-MRAM Configuration
L1	32KB I/D(SRAM)	32KB I/D (SRAM)
L2	512 KB (SRAM)	1 MB (2x SOT-MRAM)
L3	32 MB (SRAM)	64 MB (2x SOT-MRAM)

Cache size configuration

Level & Tech	Tag Latency	Data Latency	Response Latency	Total Hit Latency	
L1 (SRAM)	2	2	2	4	Parallel Access(check tag and read the data at the samw time)
L2 (SRAM)	20	20	2	41	
L2 (SOT-MRAM)	20	24	2	46	
L3 (SRAM)	30	30	2	62	
L3 (SOT-MRAM)	30	34	2	66	

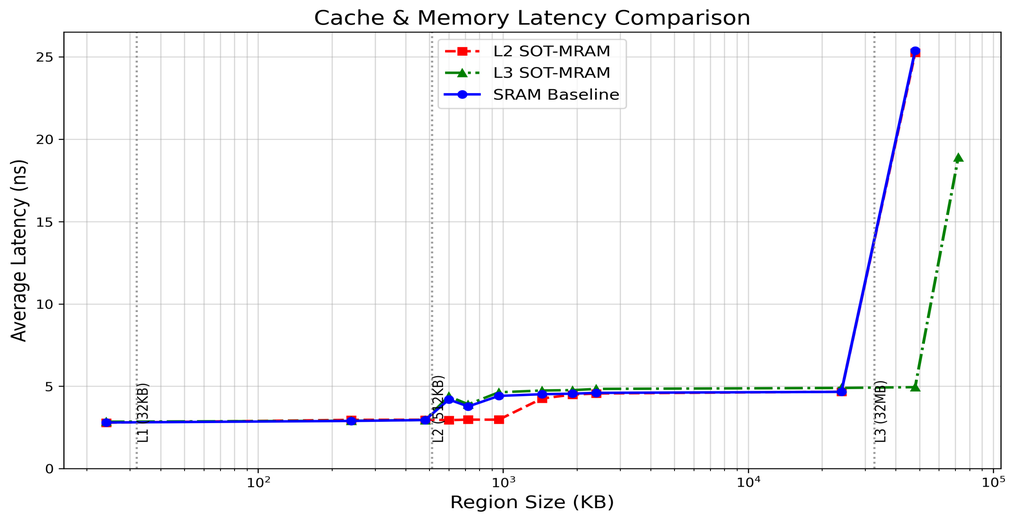
Latency configuration

- Latency: We modeled SOT-MRAM with slightly higher latency than SRAM to account for the physical time required for magnetization switching, which is slower than SRAM’s electrical operation.

Result

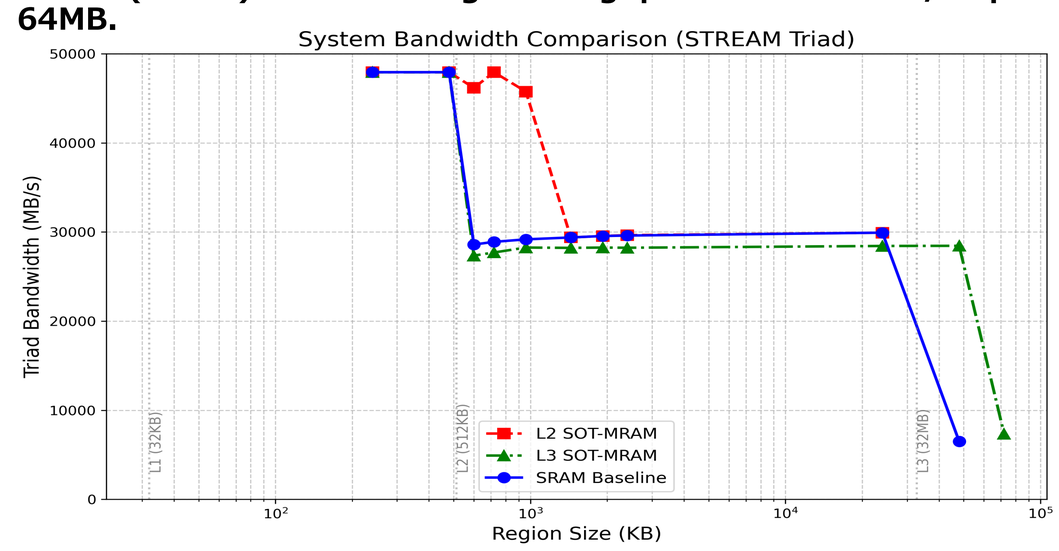
Latency

- Key Findings: Extending the Low-Latency Region
- L2 Benefit (~1MB): While the SRAM baseline degrades at 1MB, the L2 SOT-MRAM (Red) maintains its low latency of 3.0 ns, absorbing the workload within the L2 cache.
- L3 Benefit (>32MB): Crucially, beyond 32MB where other configurations spike to off-chip latency exceeding 25 ns, the L3 SOT-MRAM (Green) sustains performance around 4.9 ns up to 64MB.



Bandwidth

- Key Findings: Sustaining High Bandwidth
- L2 Benefit (~1MB): Unlike the baseline which drops to L3 speed, the L2 SOT-MRAM (Red) sustains peak bandwidth of ~48 GB/s at 1MB, doubling the L2 coverage.
- L3 Benefit (>32MB): Crucially, beyond 32MB where the baseline bandwidth collapses to DRAM levels exceeding 10 GB/s, the L3 SOT-MRAM (Green) maintains high throughput around 28 GB/s up to 64MB.



Conclusion

- the SOT-MRAM L3 cache, with 2x capacity under iso-area constraints, prevented performance degradation in the 32MB–64MB working set region
- Performance Trade-off: Although SOT-MRAM exhibits slightly higher write latency than SRAM, the system-level benefit of reducing off-chip DRAM accesses outweighs this device-level disadvantage.
- Future work will focus on detailed write energy evaluations using NVSim to further optimize the power-efficiency of the proposed hybrid architecture.

Reference

- [1]Oboril,F., Bishnoi,R., Ebrahimi, M, Tahoori, M. B. Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,34(3),367-380, 2015.
- [2]Prenat,G.,Jabeur,K.,Vanhauwaert,P.,DiPendina,G.,Oboril,F.,Bishnoi ,R., ... Gaudin, G. Ultra-fast and high-reliability SOT-MRAM: From cache replacement to normally-off computing. IEEE Transactions on Multi-Scale Computing Systems, 2(1), 49-60, 2015.
- [3] Gupta, M., et al. "High-density SOT-MRAM technology and design specifications for the embedded domain at 5nm node." 2020 IEEE international electron devices meeting (IEDM). IEEE, 513 - 516, 2020.