University of Pittsburgh

The Applications of Spintronic Memory in Microprocessors

Yiran Chen (yic52@pitt.edu)

Swanson School of Engineering
Department of Electrical and Computer Engineering

Outline

- Introduction
- STT-RAM as On-Chip Cache
  - On-Chip Cache for Multicore Architecture
  - Retention Time Relaxation
  - Multi-Level Cell Cache
- Conclusion

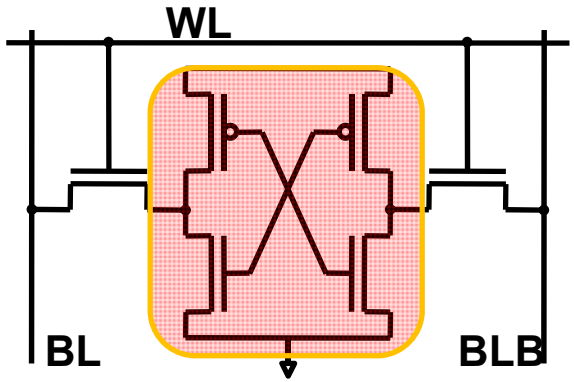
Outline

- **Introduction**
- STT-RAM as On-Chip Cache
  - On-Chip Cache for Multicore Architecture
  - Retention Time Relaxation
  - Multi-Level Cell Cache
- Conclusion

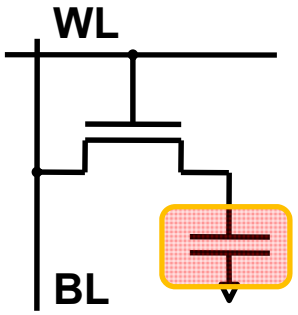
ITRS Projection

	SRAM	DRAM	NOR	NAND	MRAM
Data Retention	N	N	Y	Y	Y
Memory Cell Factor (F^2)	50-120	6-10	10	2-5	16-40
Read Time (ns)	1	30	10	50	3-20
Write /Erase Time (ns)	1	50	10^5-10^7	10^5-10^6	3-20
Endurance	10^{16}	10^{16}	10^5	10^5	10^{15}
Power Consumption – Read/Write	Low	Low	High	High	Med/ High
Power Consumption – Other than R/W	Leakage Current	Refresh Power	None	None	None
Embedded/SoC Friendly	Y	N (Thermal)	N (Thermal)	N (Thermal)	Y

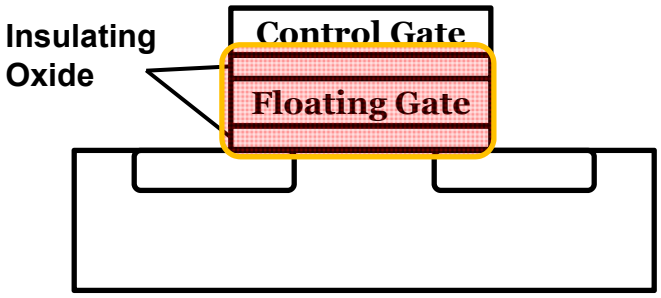
Magnetic RAM (MRAM)



6T-SRAM

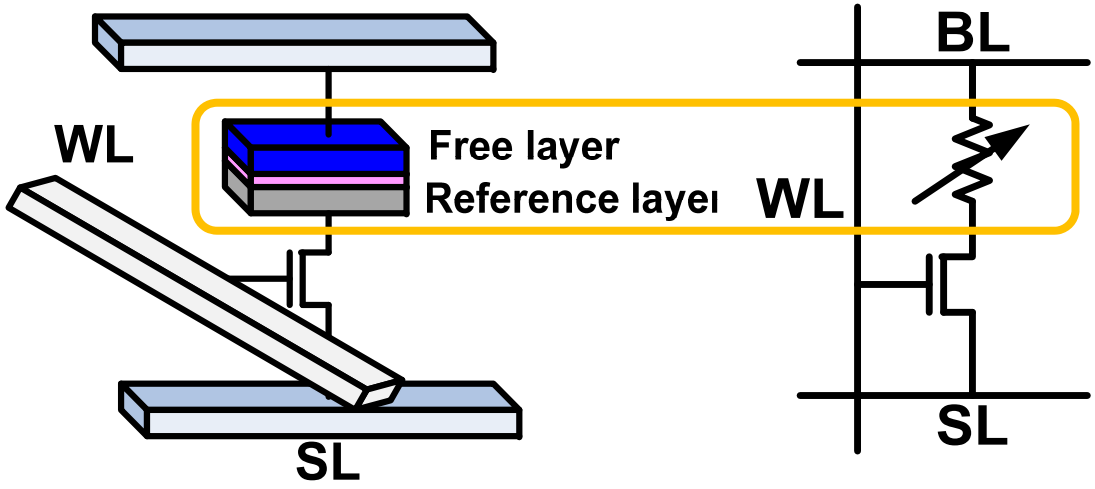


DRAM

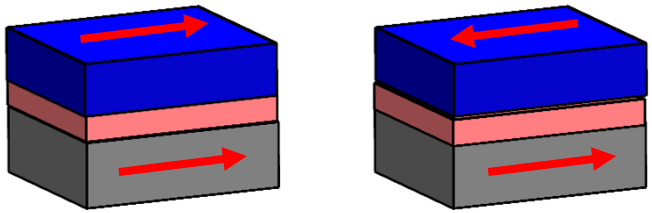


Flash

Electronic Charge



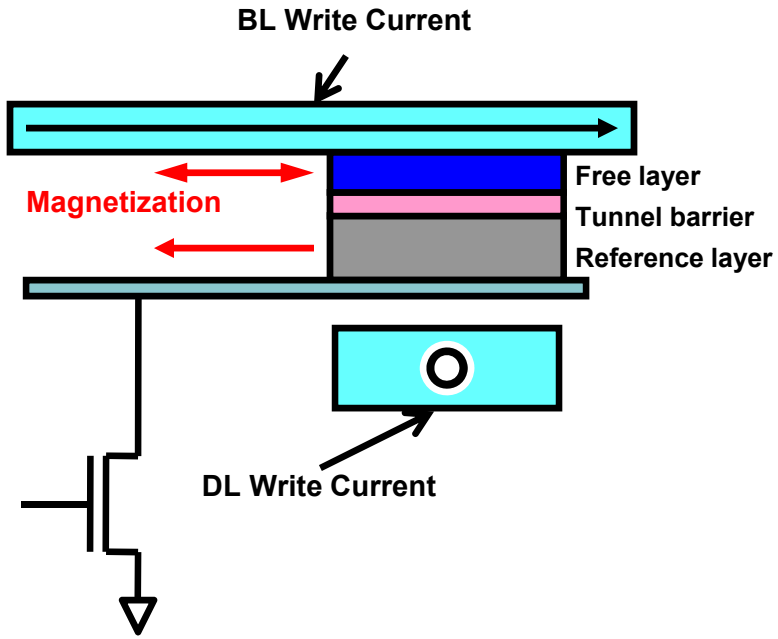
MTJ – Magnetic Tunneling Junction



"0" – Parallel "1" – Anti-parallel

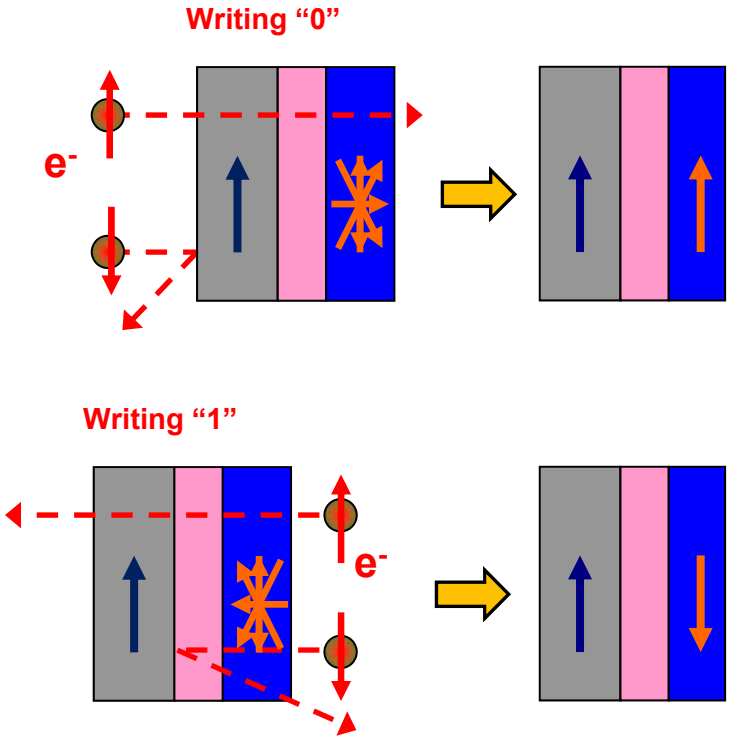
Traditional MRAM vs. STT-RAM

Traditional MRAM



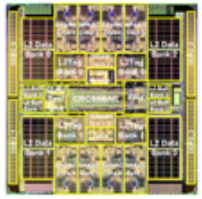
- Current induced magnetic field
- 7-8 year ago
- When scaling down, write current ↑

Spin-Transfer Torque

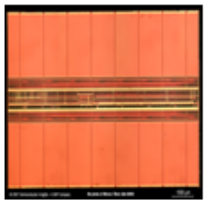


- Spin-Transfer Torque
- Recently
- When scaling down, write current ↓

Memory Hierarchy Review



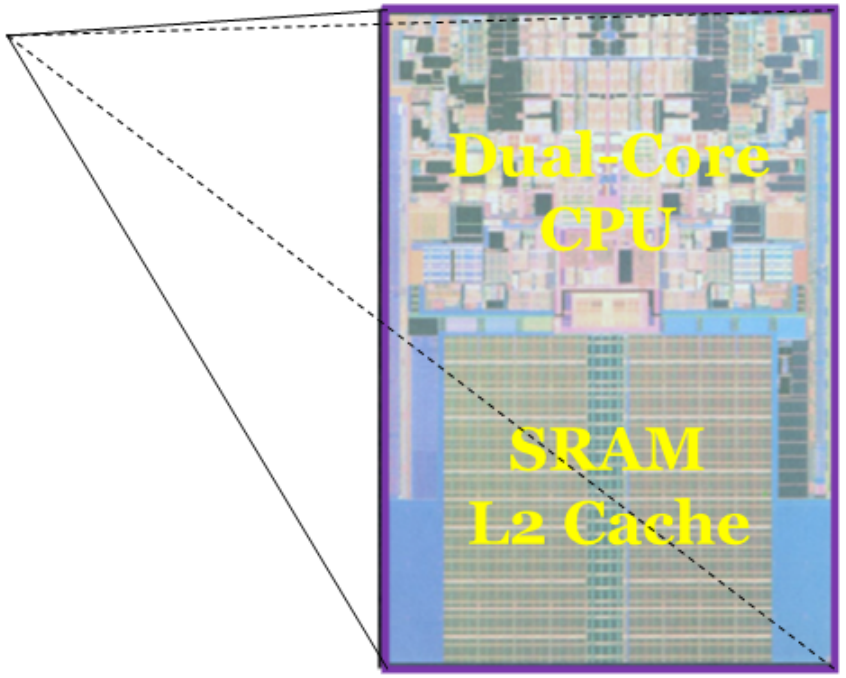
On-chip memory
(SRAM)
1~30 cycles



Off-chip memory
(DRAM)
100~300 cycles

Solid State Disk
(Flash)
25K~2M cycles

Secondary Storage
(HDD)
> 5M cycles



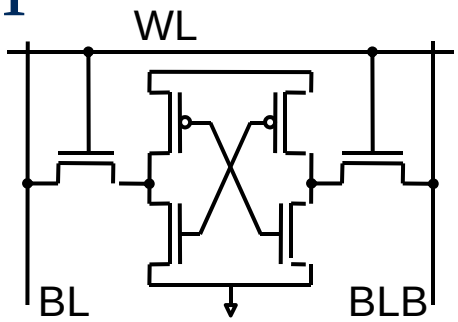
Courtesy: Al Fazio (Intel)

Outline

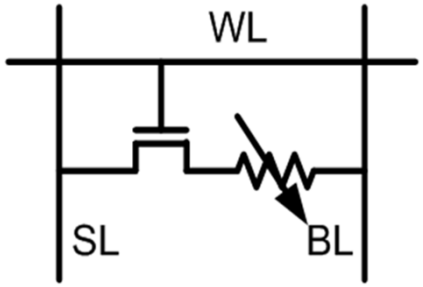
- Introduction
- **STT-RAM as On-Chip Cache**
  - **On-Chip Cache for Multicore Architecture**
  - **Retention Time Relaxation**
  - **Multi-Level Cell Cache**
- Conclusion

SRAM? STT-RAM?

- SRAM is widely used as cache
- SRAM challenges when scaling down
  - Leakage, reliability, device mismatch, variation…
- Replace SRAM w/ STT-RAM?
- STT-RAM has the similar electrical interface as SRAM



6T-SRAM



STT-RAM

SRAM vs. MRAM (STT-RAM)

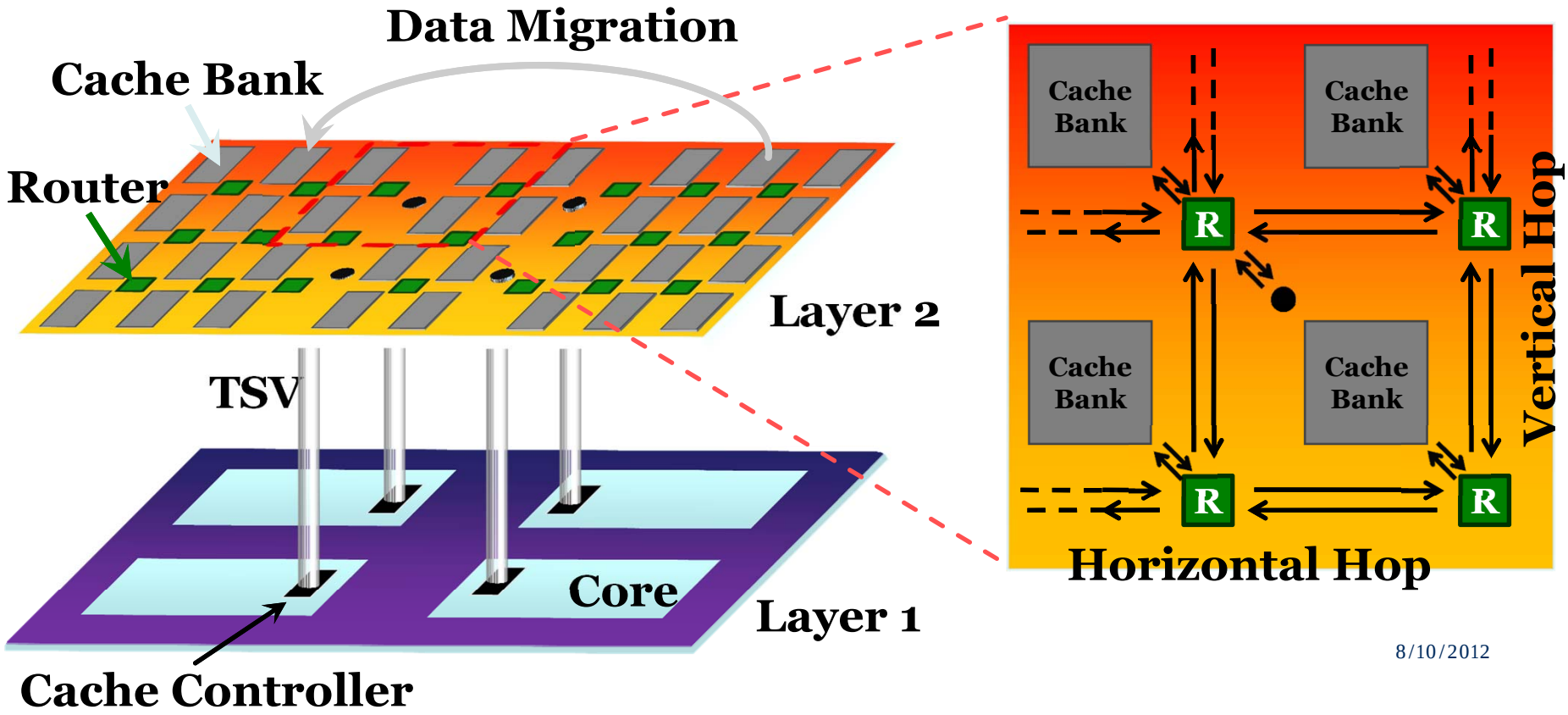
Area (65nm)	3.66mm² SRAM	3.30mm² MRAM
Capacity/Bank	128KB	512KB
Read latency	2.25ns	2.32ns
Write latency	2.26ns	11.02ns
Read energy	0.90nJ	0.86nJ
Write energy	0.80nJ	5.00nJ

Cache configurations	Leakage power
2MB (16x128KB) SRAM cache	2.09W
8MB (16x512KB) MRAM cache	0.26W

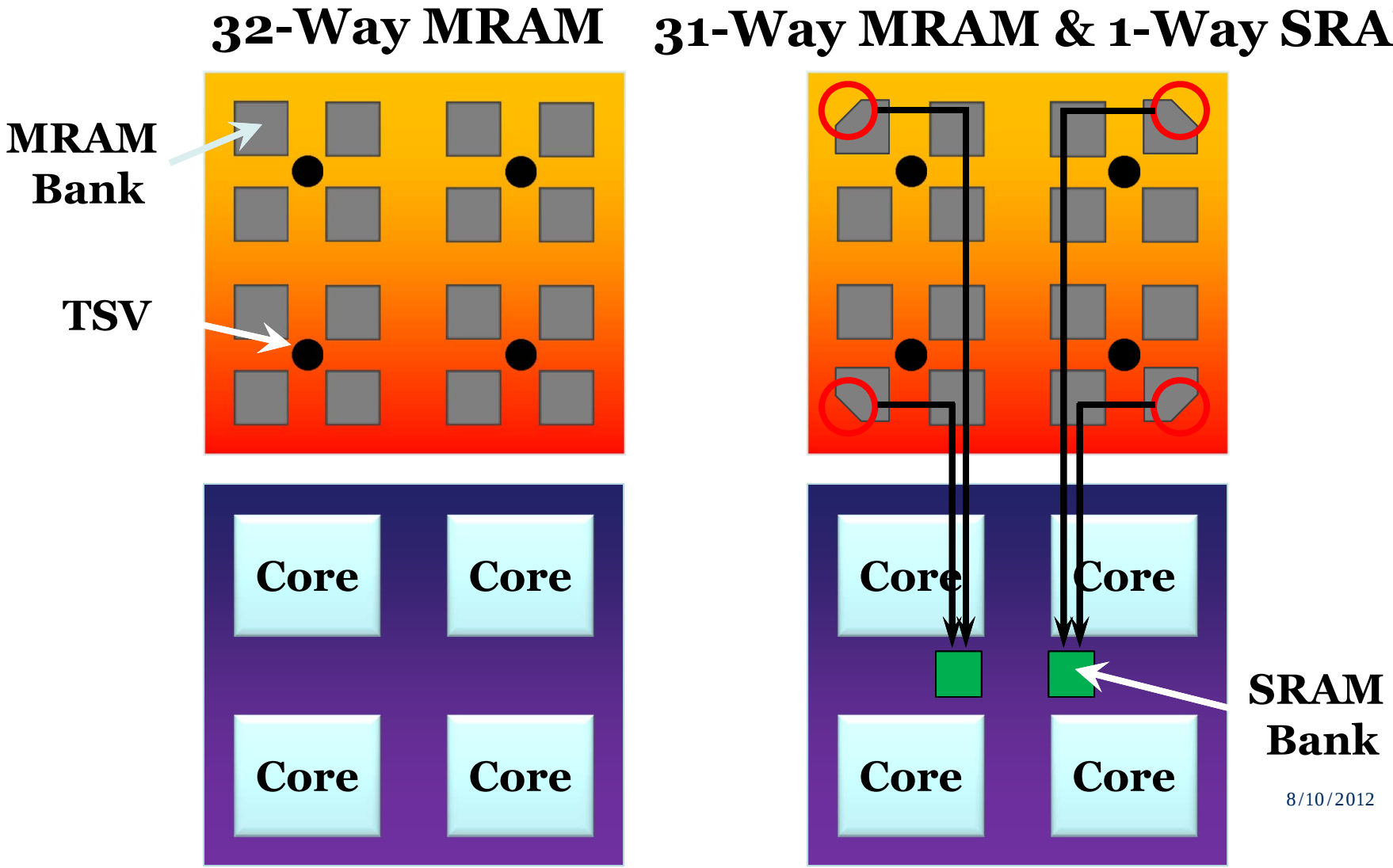
- **Pros:** Low leakage power, high density.
- **Cons*:** Long write latency and large write power

Baseline 3D Architecture

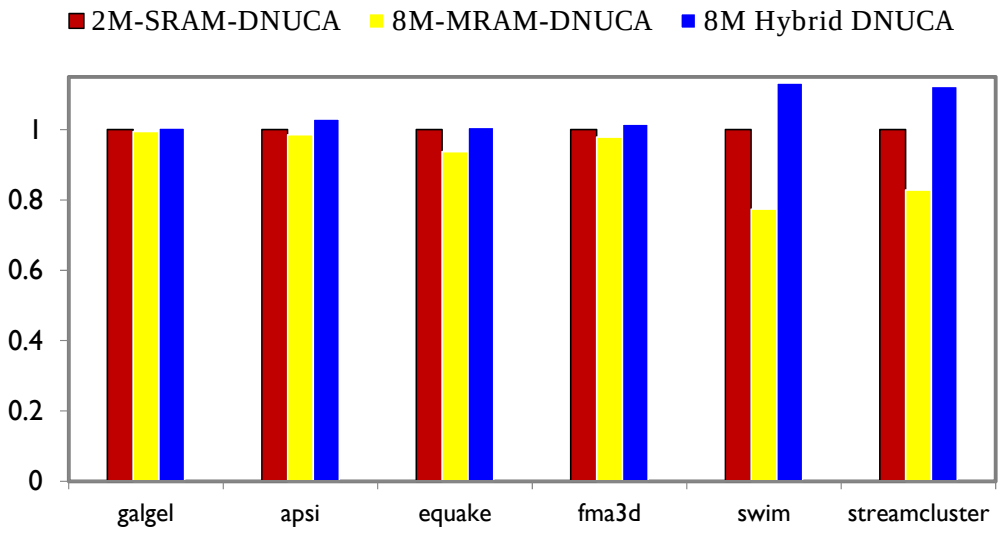
- Core Layer + Cache Layers.
- NUCA caches with NOC connections.



SRAM-MRAM Hybrid L2 Cache

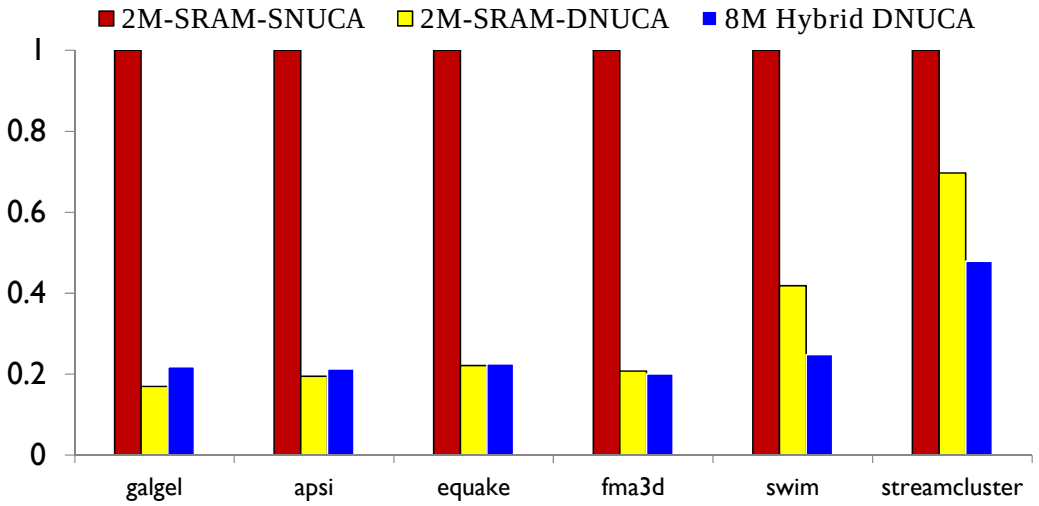


IPC (Instruction Per Cycle) Result

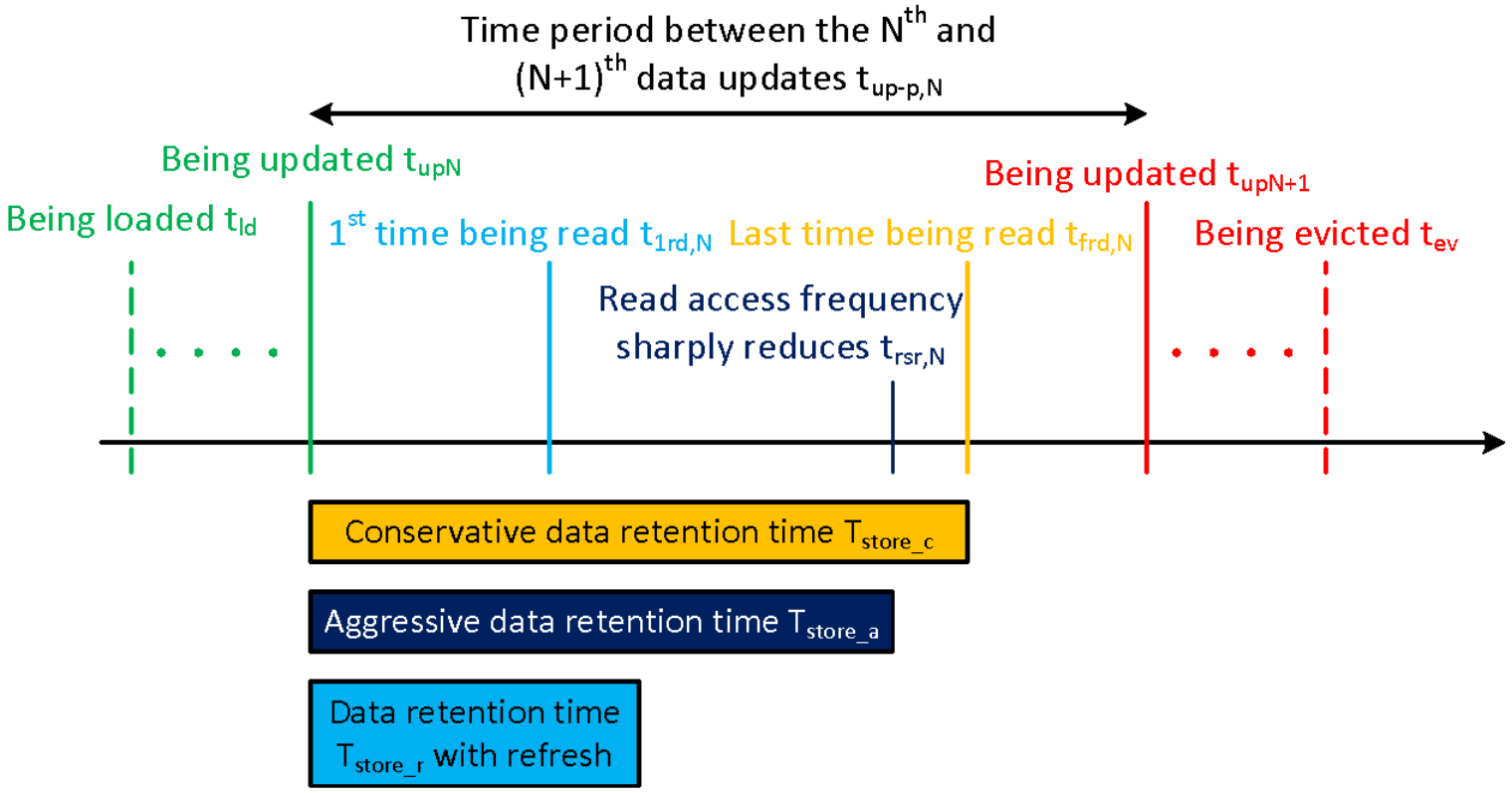


The average IPC (Instruction Per Cycle) is increased by 15%.

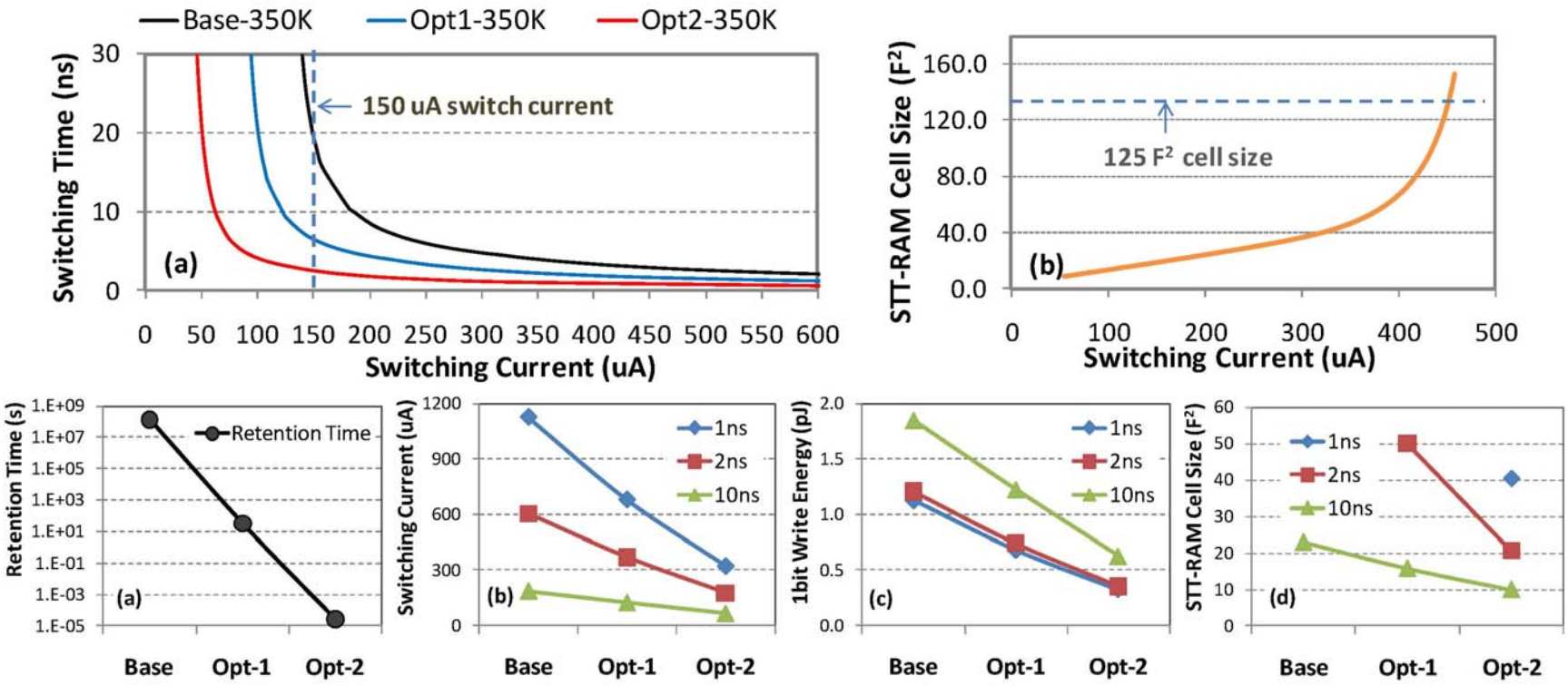
The average total power is also reduced substantially.



Data retention time requirement

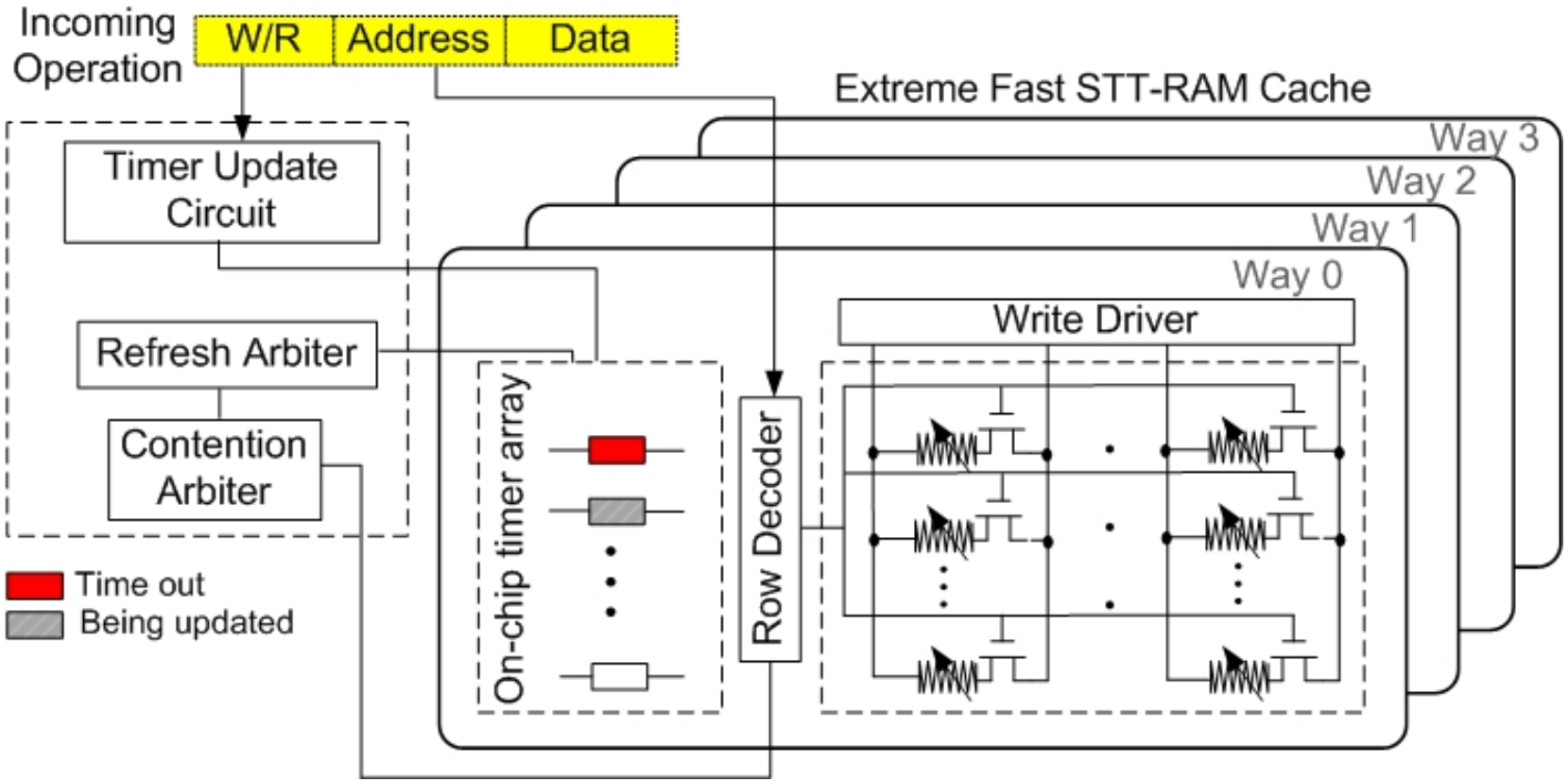


- No data will be kept in the memory hierarchy forever.
- Data becomes quiet even before it retires or is replaced.
- The actual retention time may be much shorter.

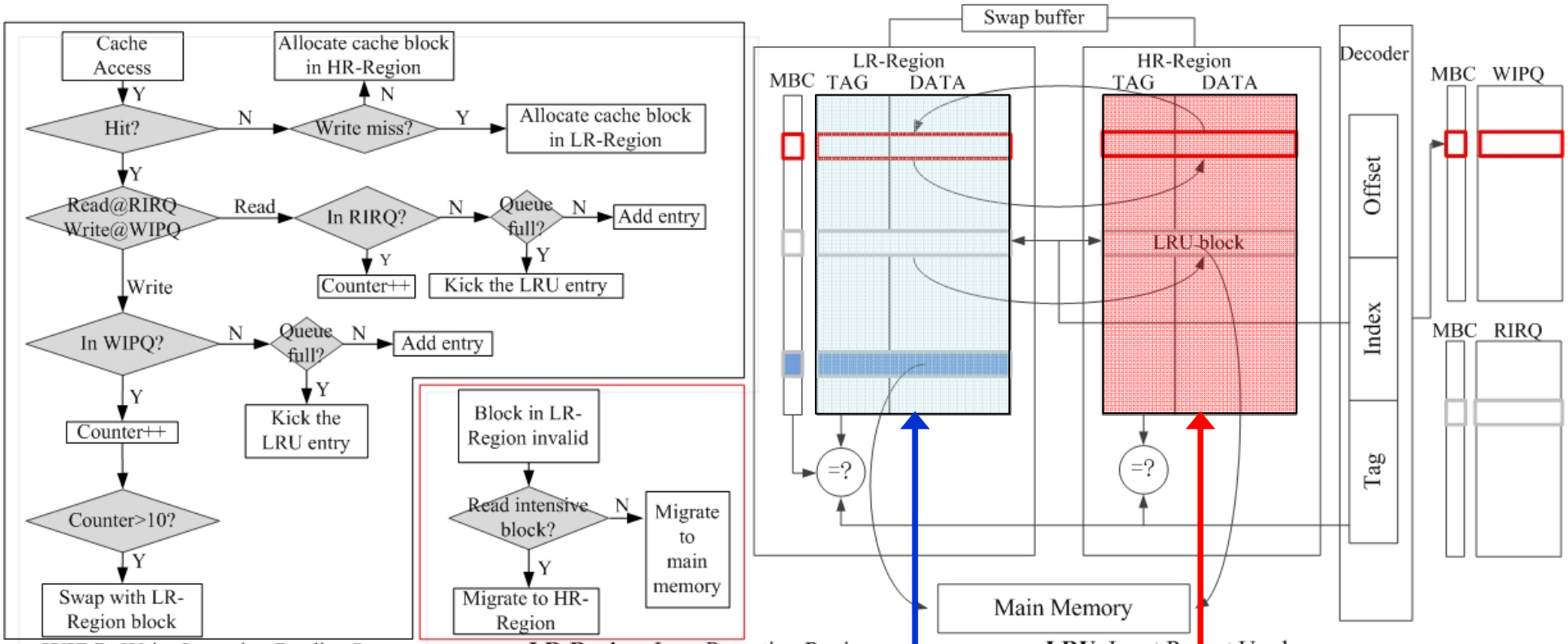


- MTJ data retention time is scaled from 4.27 years (“Base”) to 26.5us (“Opt2”).
- The required MTJ switching current decreases from 185.14uA (“Base”) to 62.5uA (“Opt2”) for a 10ns switching time at 350K.
- At a MTJ switching current of 150uA, the corresponding switching times of all three MTJ designs varied from 20ns to 2.5ns.

STT L1 w/ Data Monitor & Refreshing



STT L2 w/ Multi-Retentions



WIPQ: Write Intensive Predict Queue
RIRQ: Read Intensive Record Queue

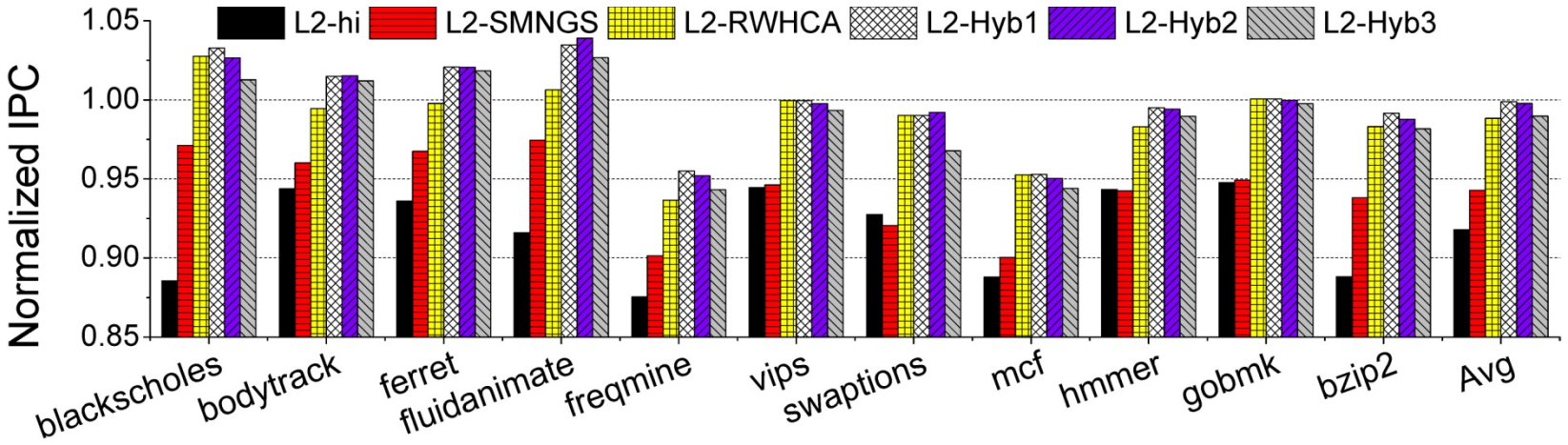
LR-Region: Low-Retention Region
HR-Region: High-Retention Region

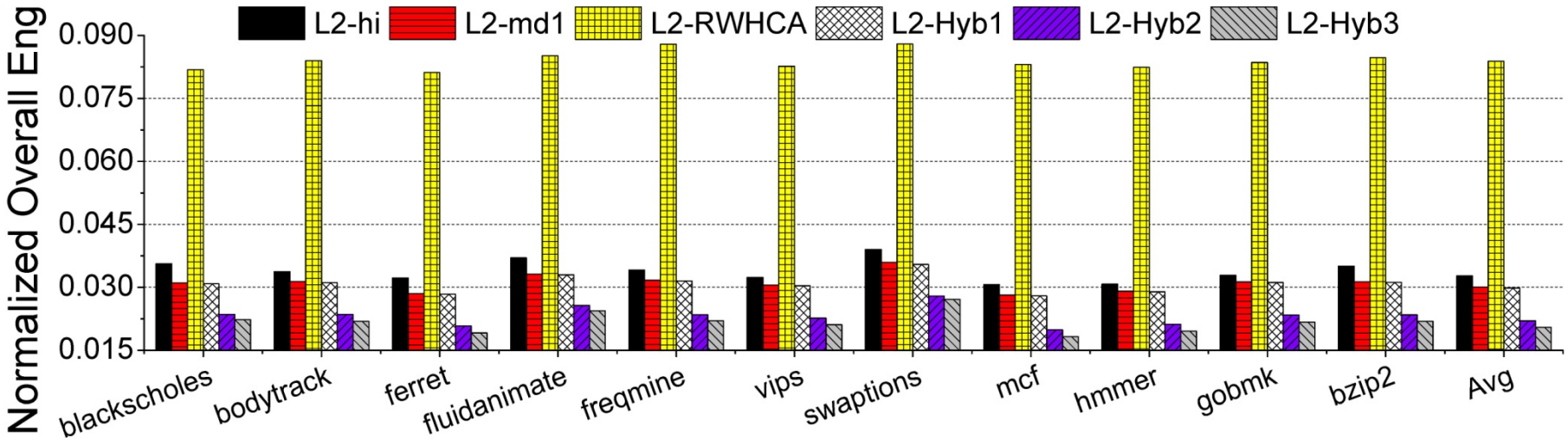
LRU: Least Recent Used
MBC: Memristor Bit Counter

Low retention region: Way 1

High retention region: Way 2~16

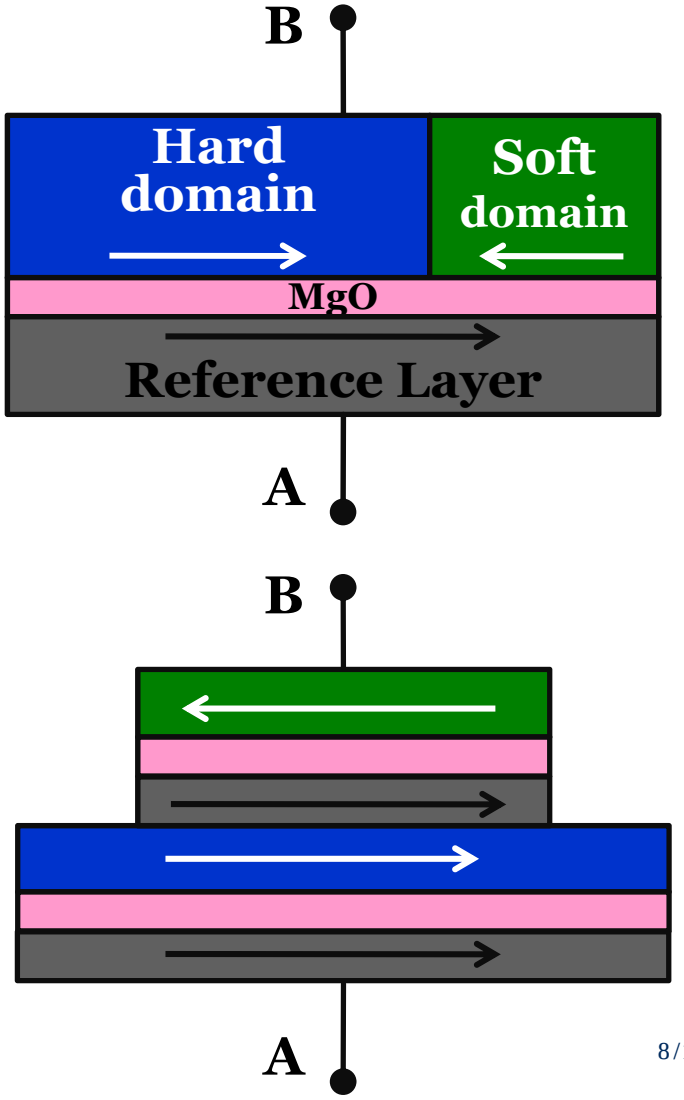
Dramatic Power and Energy Improvement



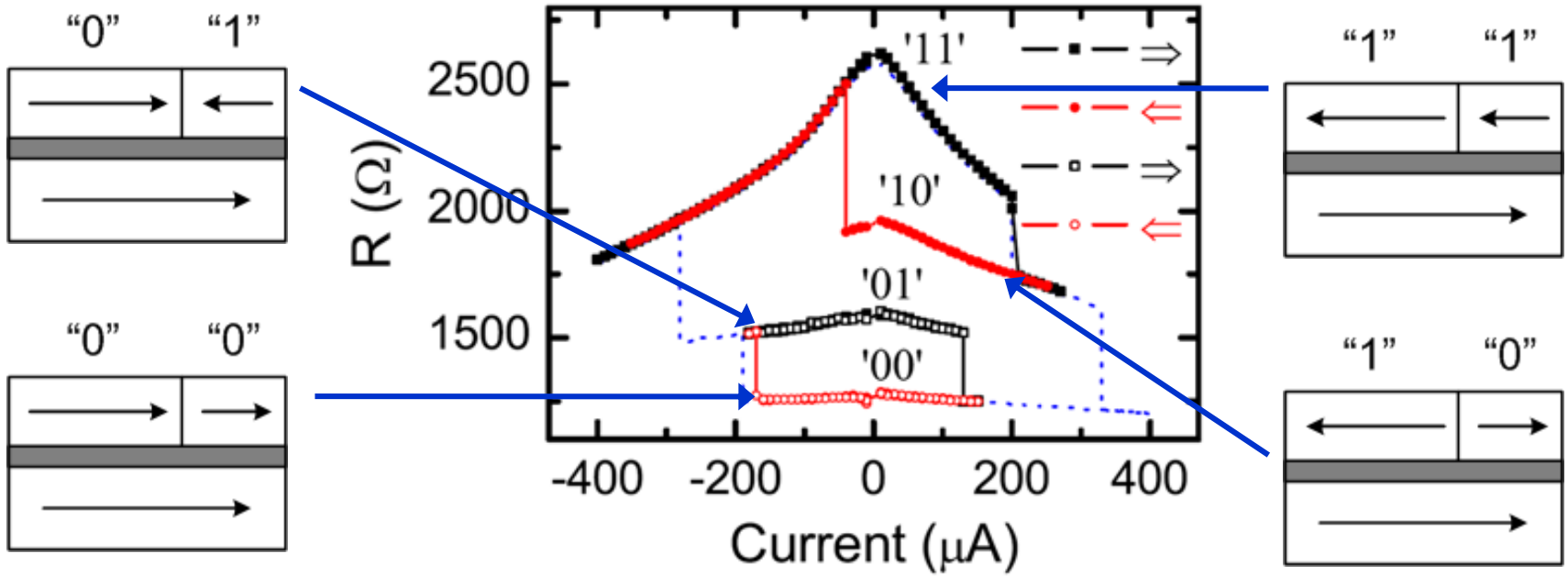


Multi-Level Cell (MLC) MTJ

- Two magnetic domains/devices
  - Hard & soft
- Four resistance states: 00-11
- Transitions are realized by passing the spin-polarized currents with different amplitudes and/or directions.
  - Soft: by a small current.
  - Hard: by only a large current.



MLC MTJ R-V Sweep Curve

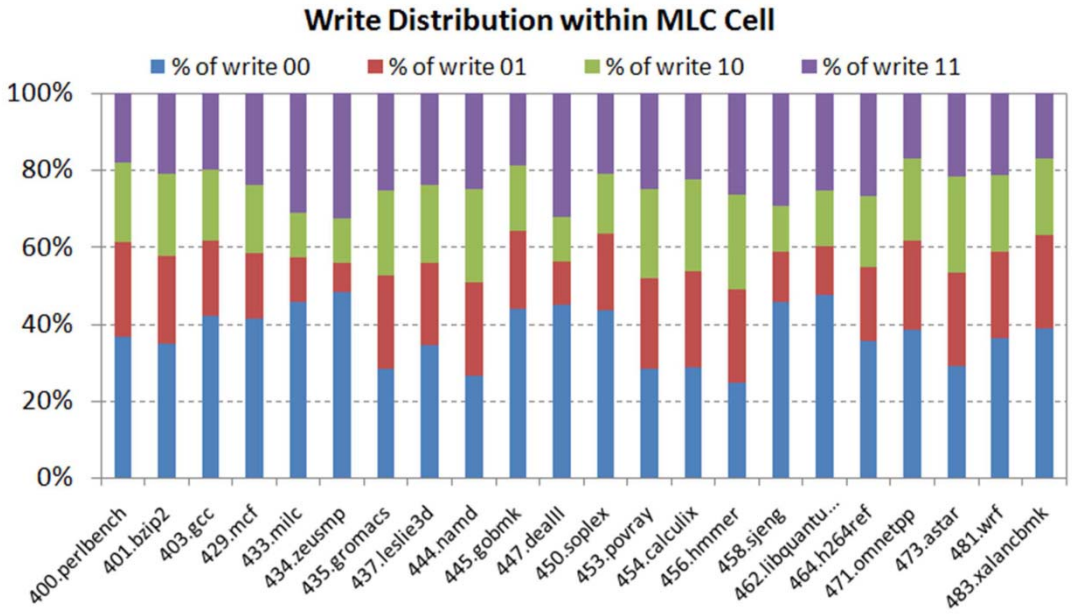


- X. Lou, et al, "Demonstration of Multilevel Cell Spin Transfer Switching in MgO Magnetic Tunnel Junctions," APL. 93, 242502 (2008)

MLC STT-RAM Cell Specifications

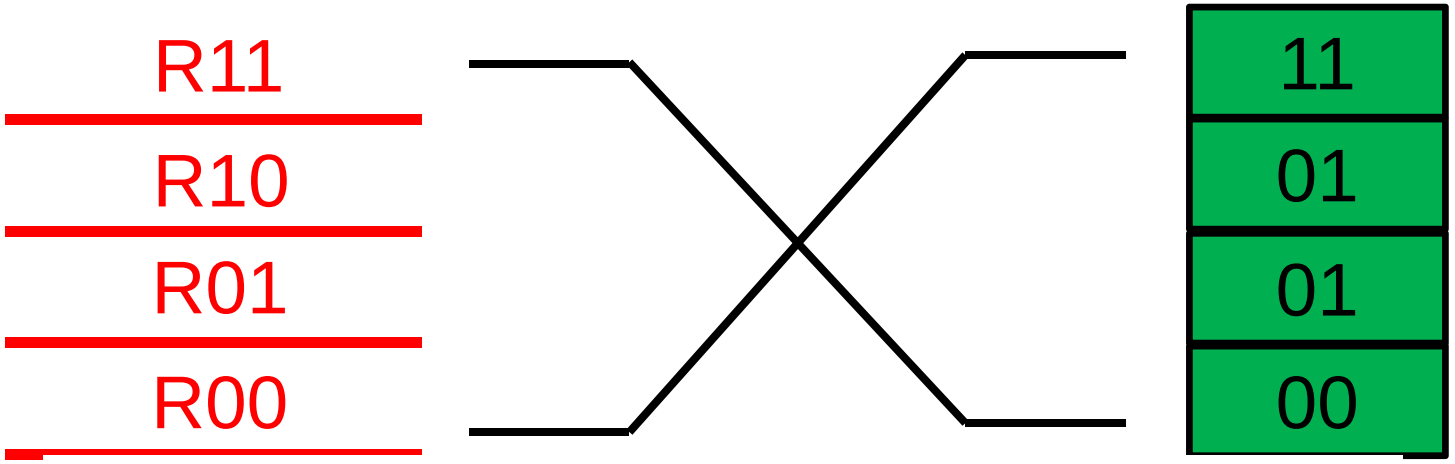
Switching Currents of MLC MTJ (μA) at 45nm

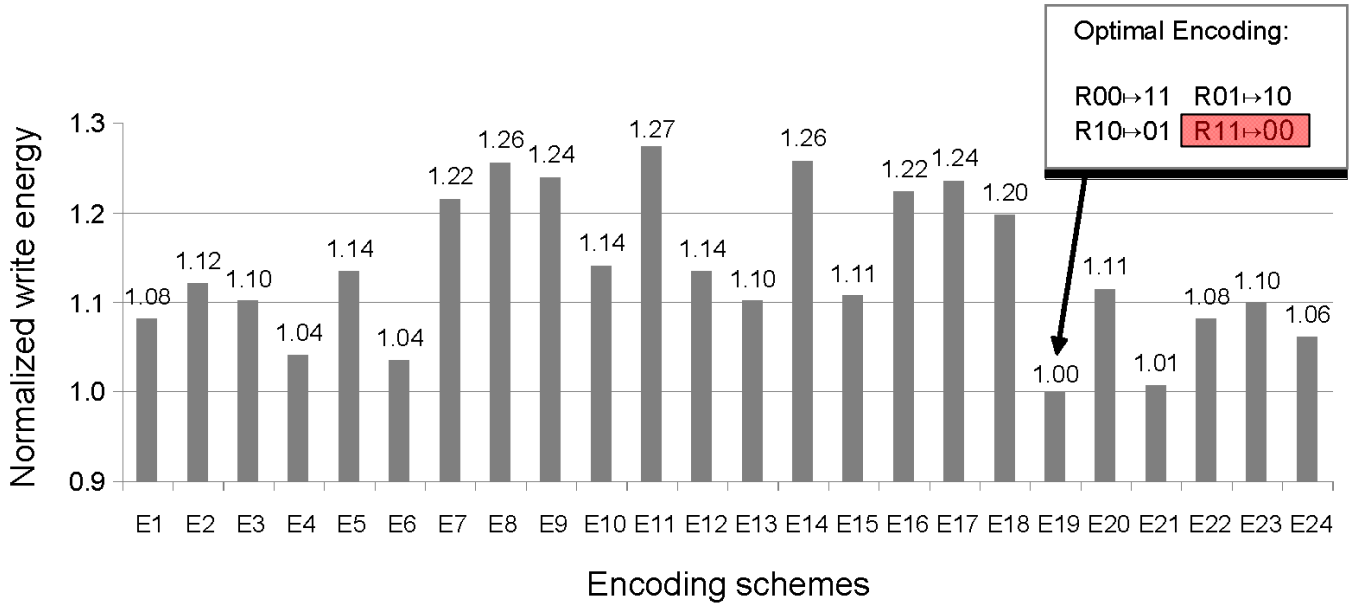
	R00	R01	R10	R11
R00	0	-38.3	X	-56.7
R01	26.3	0	X	-56.7
R10	66.4	X	0	-9.1
R11	66.4	X	39.7	0



Normal V_{DD} = 1.0V, V_{WL} = 1.0V	Min. NMOS width (nm)	174.4	715.0
	Memory cell area (F^2)	14.6	50.7
Overdriving V_{DD} = 1.0V, V_{WL} = 1.2V	Min. NMOS width (nm)	88.1	148.6
	Memory cell area (F^2)	9.0	12.9

Resistance-Logic State Encoding





Outline

- Introduction
- STT-RAM as On-Chip Cache
  - On-Chip Cache for Multicore Architecture
  - Retention Time Relaxation
  - Multi-Level Cell Cache
- **Conclusion**

Conclusion

- **STT-RAM demonstrates great potentials in modern microprocessors, in terms of power, performance, and area.**
- **Novel architecture designs have been deployed to accommodate the unique properties of STT-RAM.**
- **STT-RAM technology also needs to be further polished for on-chip applications.**

THANK YOU!
Q&A