From Embedded World to High Performance Computing using STT-MRAM

Lionel Torres, Sophiane Senni

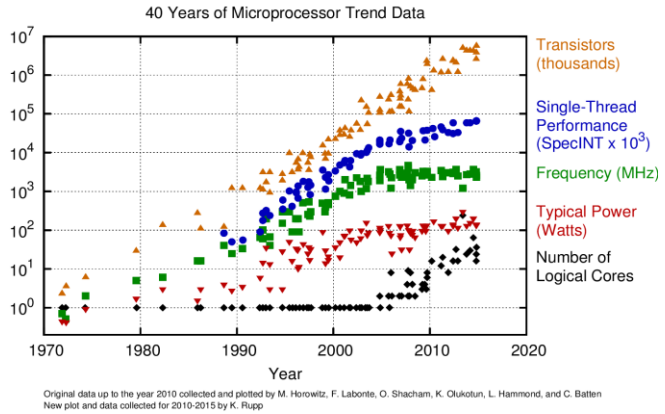
Paris, France
May 29, 2017

OUTLINE

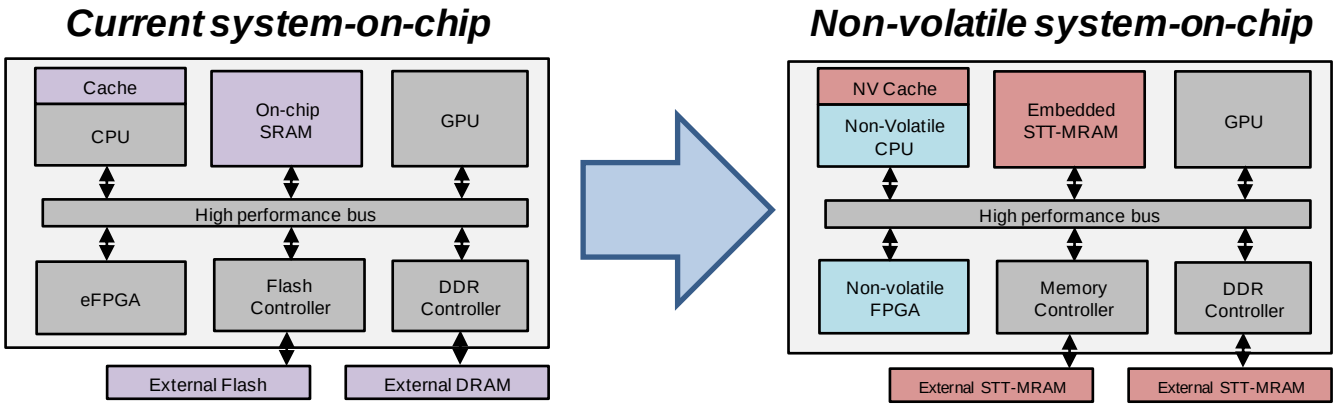
1. Motivation
2. Spintronics
    1. Basics
    2. STT-MRAM technology
3. STT-MRAM exploration at system level
    1. Embbeded systems & High Performance Computing
4. Conclusions and Future Work

Motivation

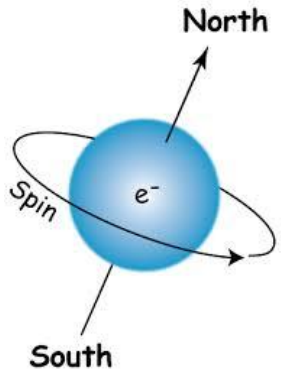
- **CMOS scaling issues are observed…**
  - Heat dissipation
  - Performance saturation
- **Due to..**
  - High leakage current
  - High power density
- Thermal constraints ➔ partially turn off the system
- Turning off the memory part ➔ the execution state is lost



Need to go beyond CMOS



Spintronics

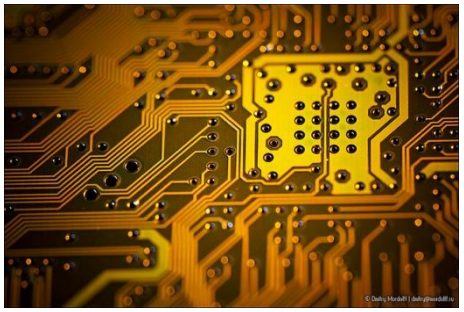


Electron properties

Mass
Electric charge
Spin

Electronics

Electrons are moved (current) by acting on the charge

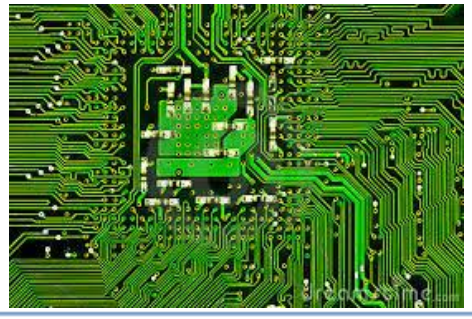


Spintronics

Motion by acting on the spin !

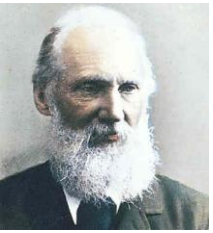
Phenomena related to spin

Magnetoresistance
Spin Transfer torque



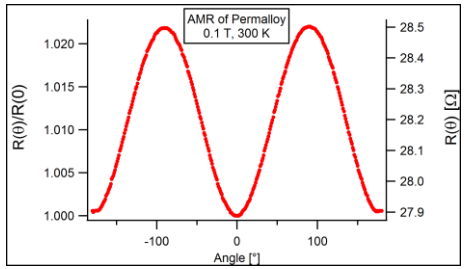
Spintronics

Anisotropic magnetoresistance



William Thomson
1824-1907

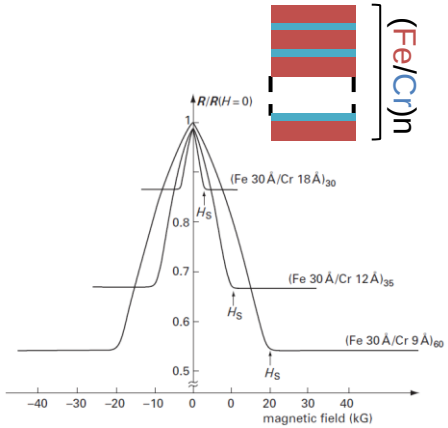
- The electrical resistance of magnetic metal varies with the presence of an external magnetic field



- Resistance variation
  → 2% - 5% at room temperature

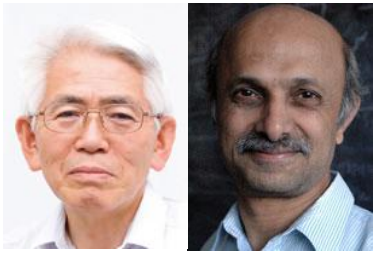
Giant magnetoresistance

Peter Grünberg
Albert Fert
2007 Nobel Prize
(Physics)

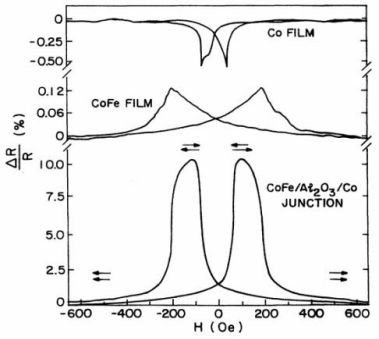


- Large increase of the conductance with structure alternating ferromagnetic / non-magnetic layers

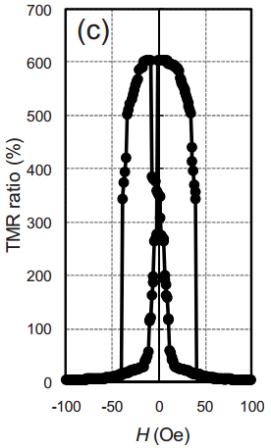
Tunnel magnetoresistance



T. Miyazaki
J. Moodera
(not in the pictures: M. Jullière)



CoFe/Al2O3/Co
J. S. Moodera 1995

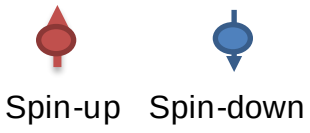


CoFeB/MgO/CoFeB
S. Ikeda 2008

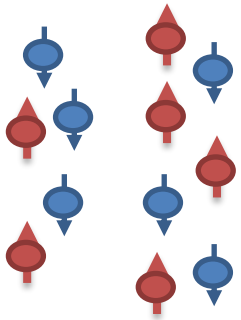
- Unlike GMR, the barrier is an isolant
- With MgO, TMR of 608% reached at room temperature

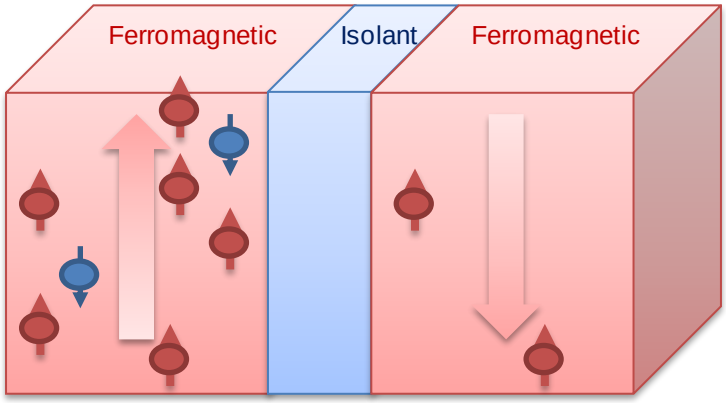
Spintronics

Tunnel magnetoresistance principle

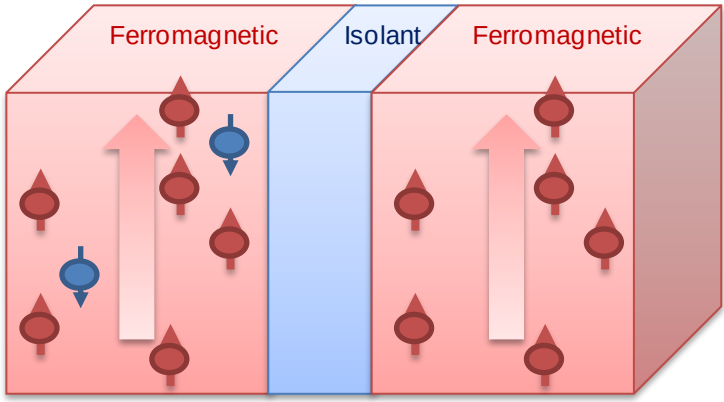


- **The transport of the electrons through the material is spin-dependent**





Antiparallel configuration



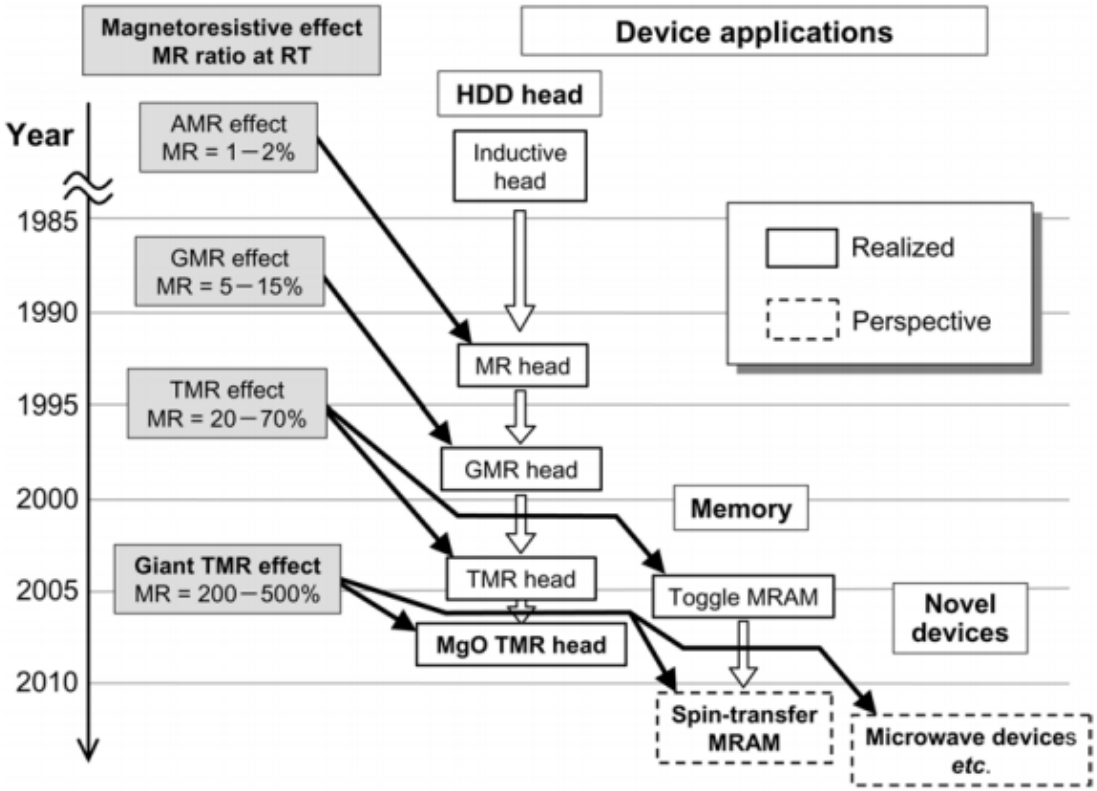
Parallel configuration

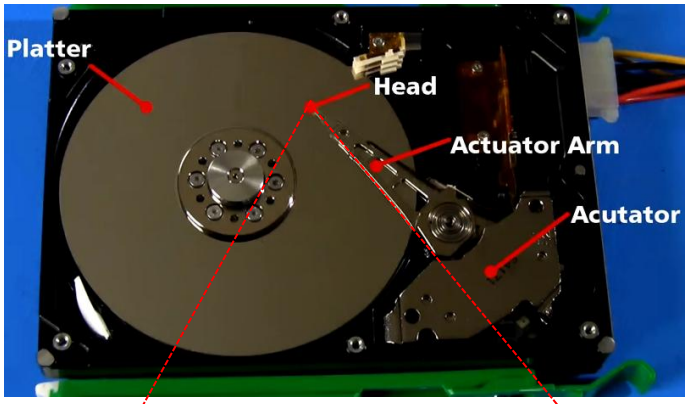
$R_{MAX} \rightarrow$ ‘1’

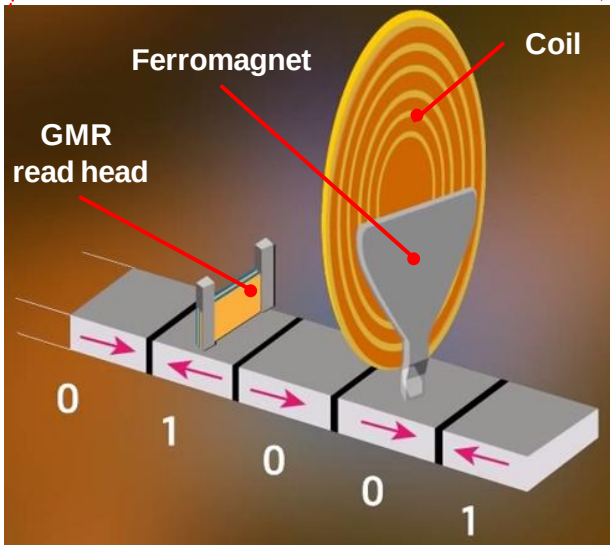
$R_{MIN} \rightarrow$ ‘0’

Spintronics

Applications

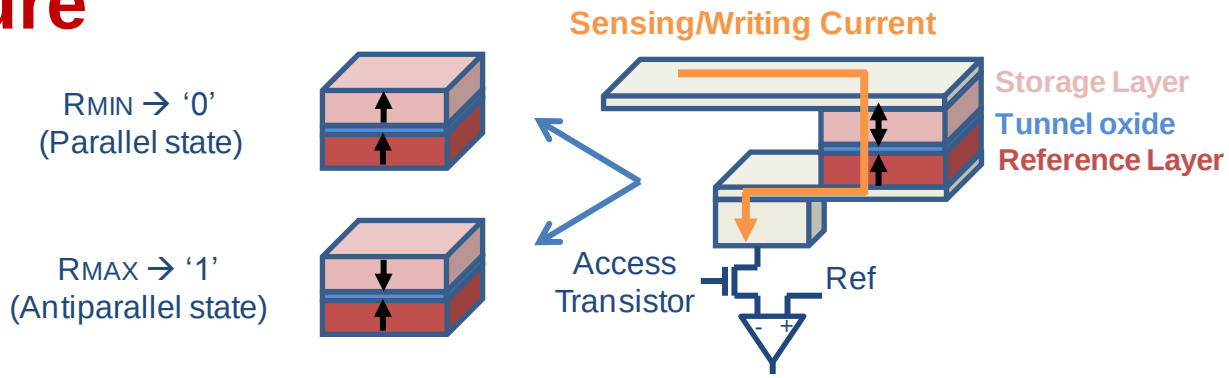






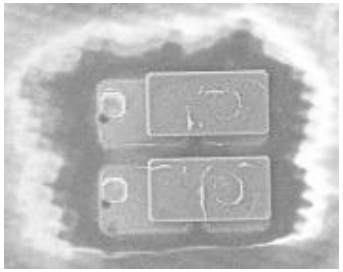
STT-MRAM technology

Bit Cell Structure

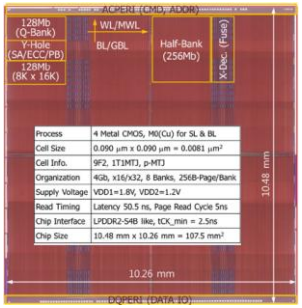


- STT-MRAM can be used to build:
  - Flip-Flops
  - Cache memories
  - Main memories

NVFF STT-MRAM [1]



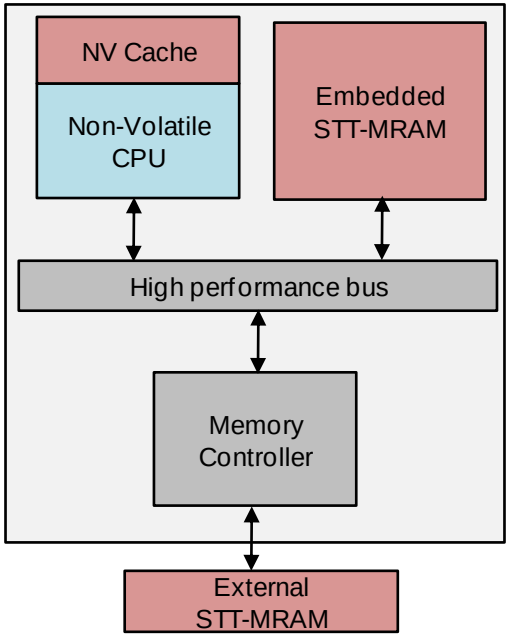
4Gb LPDDR2 STT-MRAM [2]



[1] B. Jovanovic et al., "**A hybrid magnetic/complementary metal oxide semiconductor three-context memory bit cell for non-volatile circuit design**," AIP Journal of Applied Physics, April 2014.

[2] K. Rho et al., "**A 4Gb LPDDR2 STT-MRAM with compact 9F2 1T1MTJ cell and hierarchical bitline architecture**," Solid-State Circuits Conference (ISSCC), February 2017.

STT-MRAM exploration



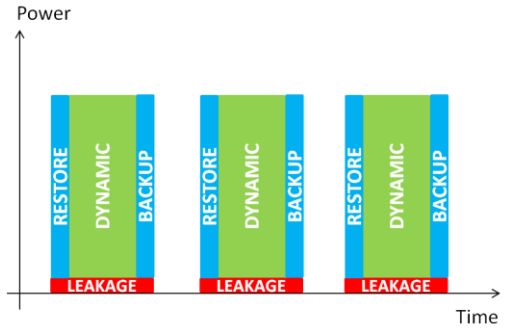
- **The main objectives are…**
  - Evaluate the impact at system level of using STT-MRAM
  - Explore new applications
    - Non-volatile working memories (registers, cache…)
    - In-memory computing

- **This talk focuses on..**
  - Non-volatile processor for embedded applications
  - STT-MRAM exploration framework for High Performance Computing

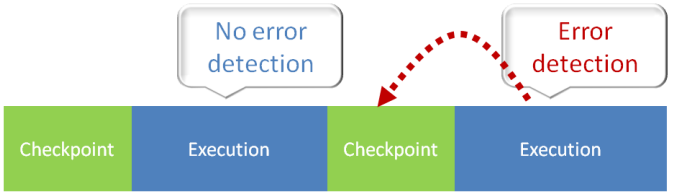
Non-volatile processor based on STT-MRAM

Non-volatile processor based on STT-MRAM

- **Two application under study…**
  - Normally-off Computing
    - The system is normally off
    - The execution state is preserved after a shutdown
    - Fast wakeup, near-zero leakage power in sleep mode



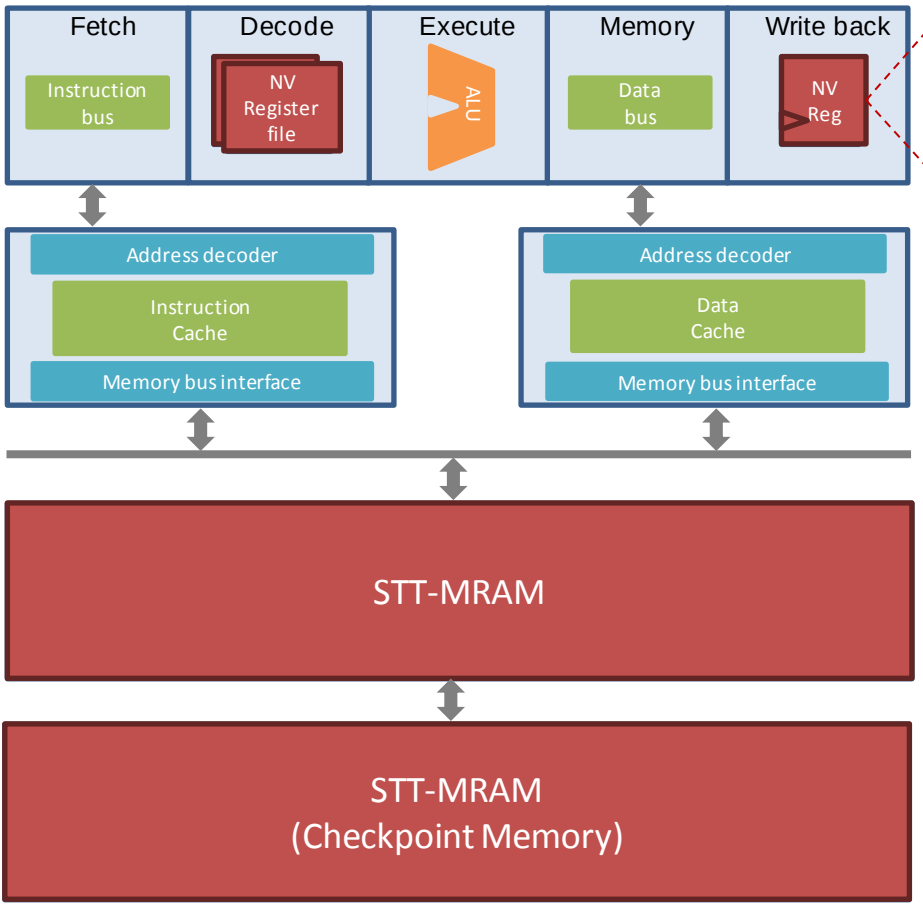
  - Checkpoint/Rollback
    - Restore a safe state of the processor for instance after an execution error or a power failure



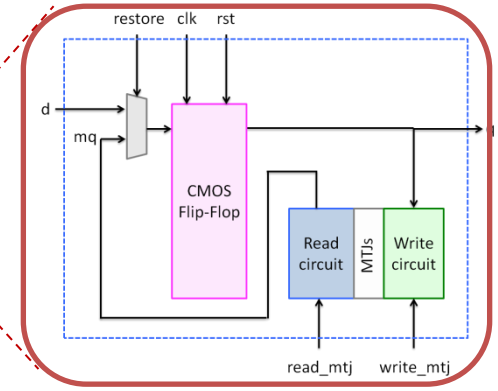
- **Two 32-bit RISC processors considered…**
  - Secretblaze (MIPS like)
  - Amber (ARM like)

Non-volatile processor based on STT-MRAM

Non-volatile Processor Architecture



Hybrid CMOS/STT-MRAM flip-flop



- Speed of CMOS
- Non-volatility of STT-MRAM

STT-MRAM main memory

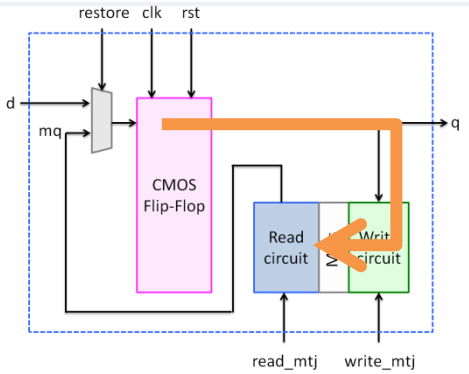
- Data are preserved after a shutdown

Checkpoint memory for the Rollback

- Store a valid state of the system to be tolerant against execution errors and power failures

Non-volatile processor based on STT-MRAM

Normally-off Computing



1 *Back up the register's state*

2 *POWER DOWN*

Main memory based on MRAM

↓

Data preserved

3 *POWER UP*

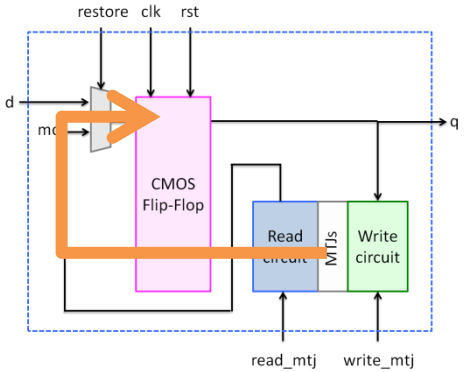
Main memory based on MRAM

↓

Data available

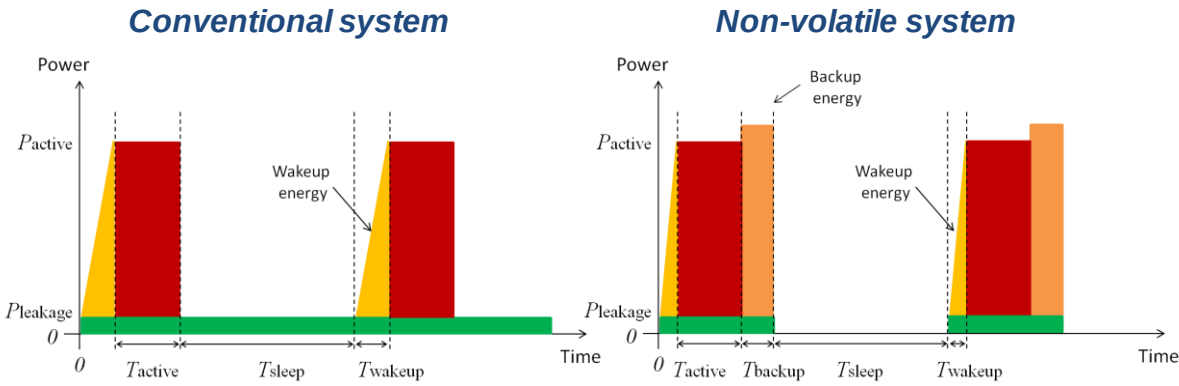
4

Restore the register's state



Non-volatile processor based on STT-MRAM

- Conventional system
  - Leakage power during sleep mode
- Non-volatile system with instant-on/off
  - Near-zero leakage during sleep mode
  - Backup energy



Minimum $Tsleep$ required to be more energy efficient ?

$$Pleakage \times Tbackup + Ebackup < Pleakage \times Tsleep$$

$$\frac{Pleakage \times Tbackup + Ebackup}{Pleakage} < Tsleep$$

Synthesis of the Amber processor

(Industrial 40nm CMOS low-power process)

$$\begin{cases} Pleakage = 973\mu W \\ Ebackup = 1nJ \\ Tbackup = 20ns \end{cases} \Rightarrow Tsleep > 1.05\mu s$$

Synthesis of the Secretblaze processor

(Industrial 40nm CMOS low-power process)

$$\begin{cases} Pleakage = 775\mu W \\ Ebackup = 1nJ \\ Tbackup = 20ns \end{cases} \Rightarrow Tsleep > 1.32\mu s$$

Non-Volatile Flip-Flops Performance

	Latency (ns)		Energy (pJ)	
Technology	Restore	Back-up	Restore	Back-up
STT-MRAM	0.2	4	0.012	0.5

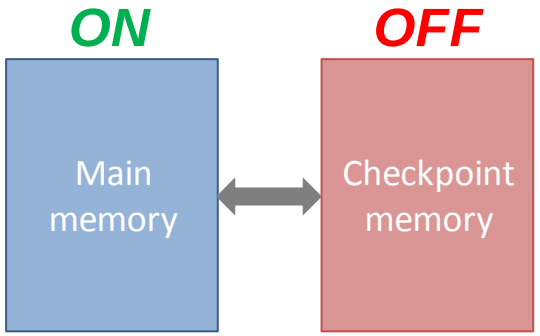
* D. Chabi et al., "**Ultra low power magnetic flip-flop based on checkpointing/power gating and self-enable mechanisms**," IEEE Transaction on Circuits and Systems I, January 2014.

Non-volatile processor based on STT-MRAM

Checkpoint Rollback

NORMAL EXECUTION

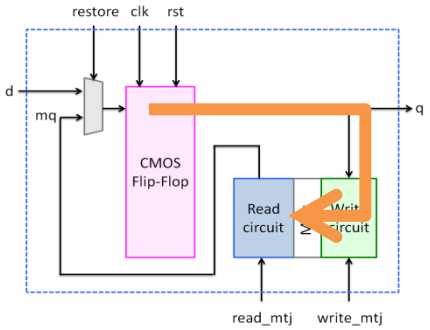
- ***Only the main memory contents are modified***
- ***The checkpoint memory is turned off***



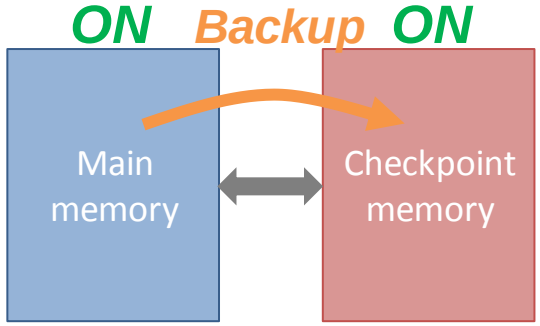
CHECKPOINT

- ***Back up registers***
- ***Back up memory***

=

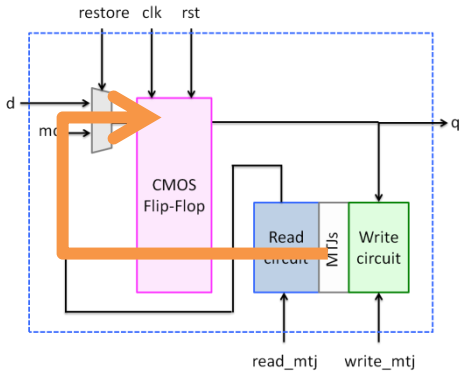


+

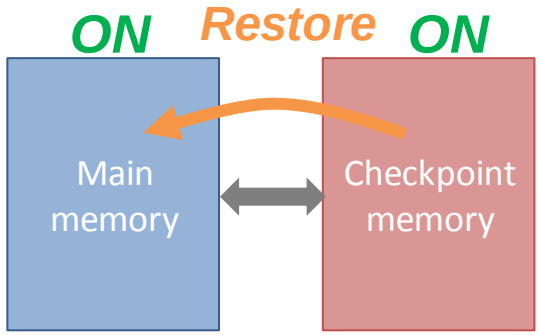


ROLLBACK

1. ***Stall the processor***
2. ***Restore the checkpoint***
3. ***Execution***

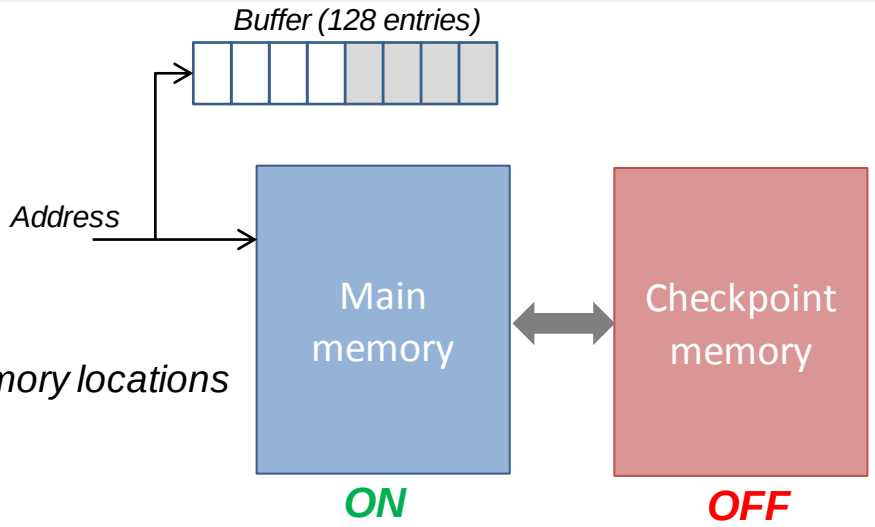


+



Non-volatile processor based on STT-MRAM

Checkpoint/Rollback (Memory part)

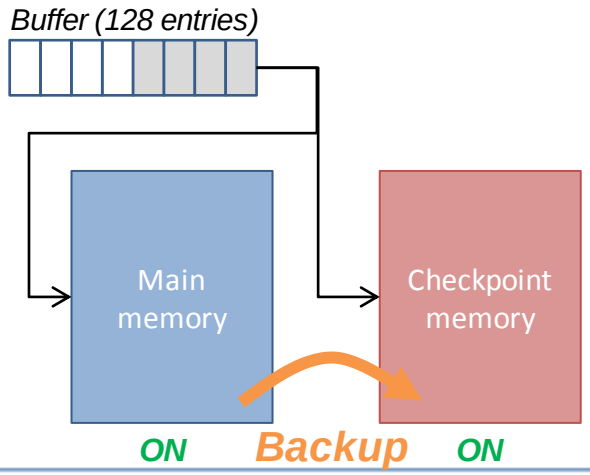


NORMAL EXECUTION

- *Only the main memory contents are modified*
- *Buffer to back up the addresses of the modified memory locations*

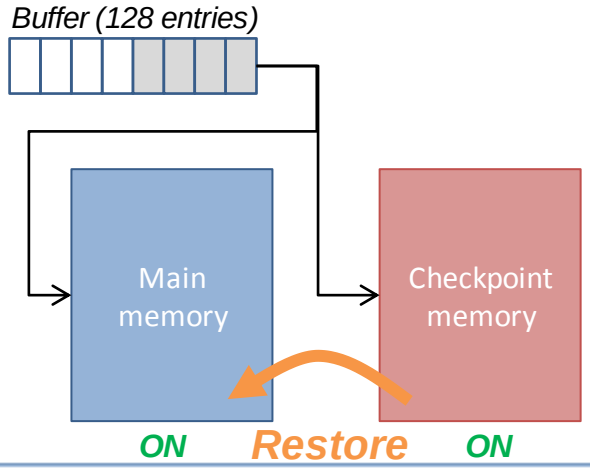
CHECKPOINT

- *Only the modified memory locations are copied*



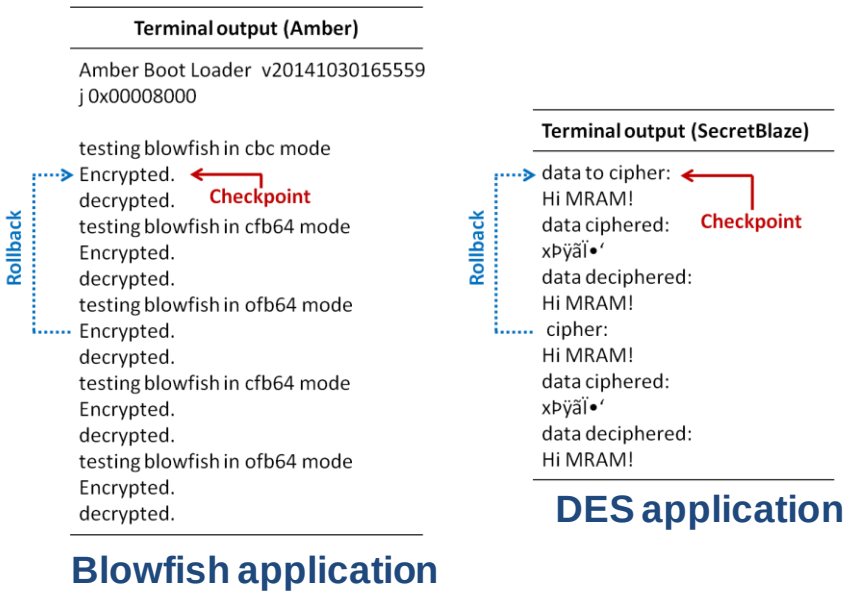
ROLLBACK

- *Only the modified memory locations are restored*



Non-volatile processor based on STT-MRAM

- **Validation of the backup/recovery of the system**

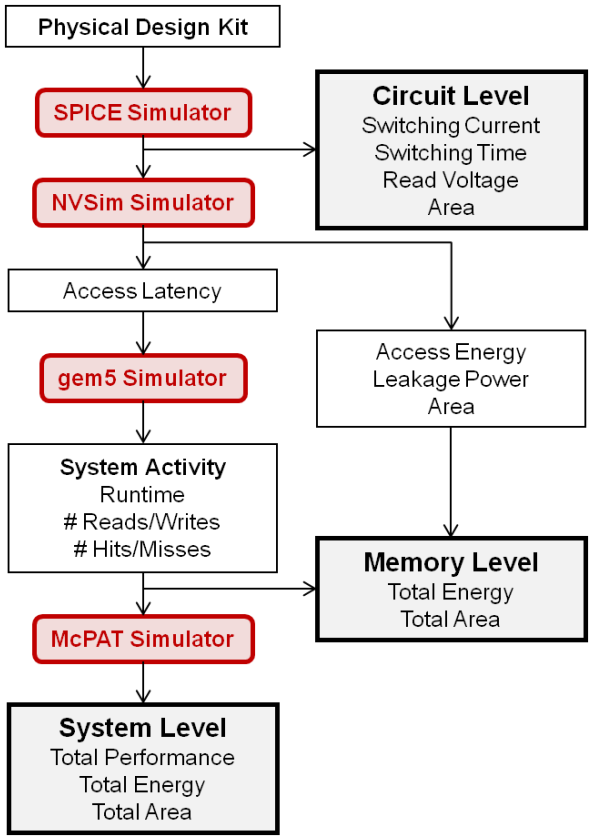


- **Evaluation of the cost**
  - Register level (Data from real flip-flop design)
    - Backup: ≈1nJ (<20ns)
    - Restore: <25pJ (≈1ns)
  - Main memory level (Data from NVSim)
    - 1MB Main memory / 4kB Checkpoint memory
      - Backup: <100nJ (<20µs)
      - Restore: <100nJ (<20µs)

High Performance Computing using STT-MRAM

High Performance Computing using STT-MRAM

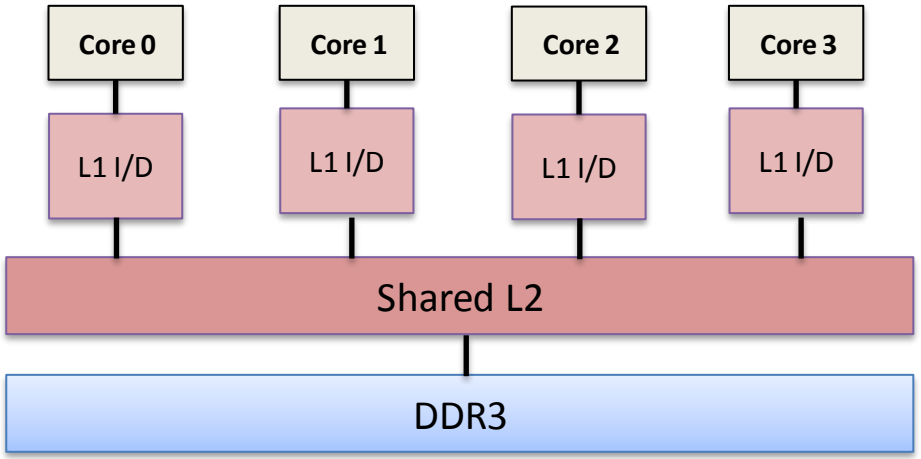
- **A simulation framework has been developed to…**
  - Explore the impact of STT-MRAM at system level
  - Provide essential feedback to enhance the development of STT-MRAM devices
  - Explore different memory technologies



- **A cross-layer investigation is done…**
  - Device level → Physical Design Kit
  - Circuit level → Bit cell
  - Memory level → Cache, main memory…
  - System level → Multi-core architectures

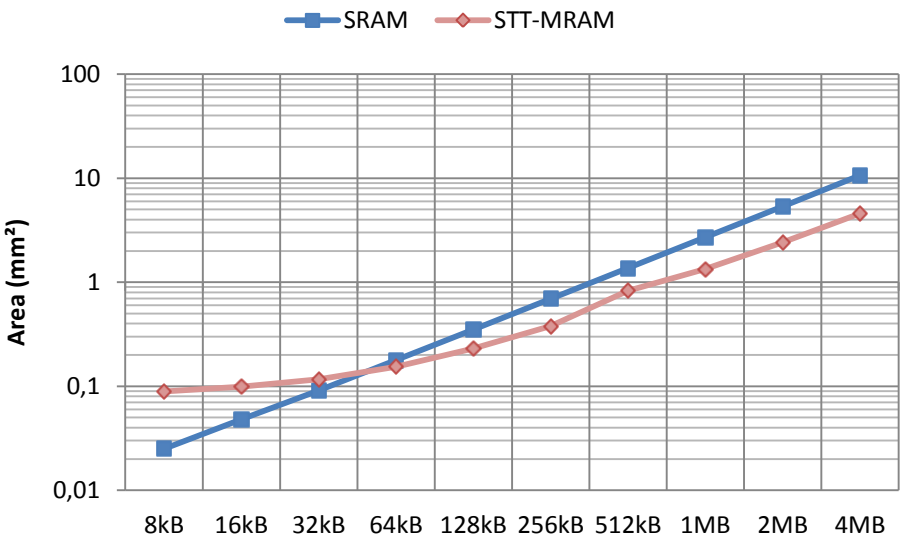
High Performance Computing using STT-MRAM

- **Case study…**
  - **Architecture considered**
    - 4-core out-of-order (ARMv7 ISA)
    - 32kB L1 instruction cache (SRAM)
    - 32kB L1 data cache (SRAM)
    - 1MB shared L2 cache
      - Two scenarios (SRAM / STT-MRAM)
    - 512MB DRAM DDR3 main memory
  - **Benchmarks**
    - PARSEC
    - SPLASH-2



High Performance Computing using STT-MRAM

- Circuit-level analysis…
  - **Area**



Process	Technology	1MB L2 (mm^2)	32kB L1 (mm^2)
45nm	SRAM	2.7	0.091
	STT-MRAM	1.12	0.116

- STT-MRAM is denser for large cache capacity
  - STT-MRAM cell size smaller than that of SRAM
- STT-MRAM needs large transistors for write operations

High Performance Computing using STT-MRAM

- Circuit-level analysis…
  - **1MB cache performances**
    - Based on NVSim

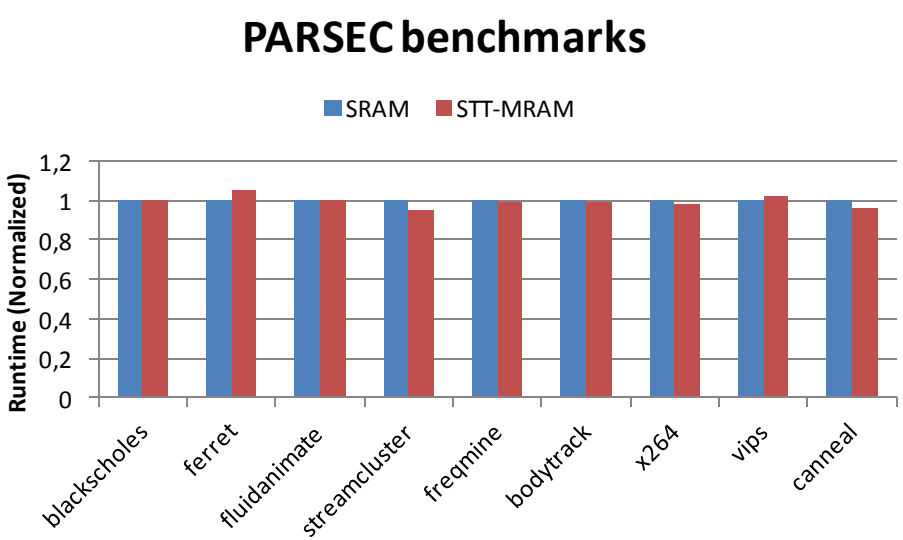
		Read		Write		Standby
Node	**Technology**	**Latency (ns)**	**Energy (nJ)**	**Latency (ns)**	**Energy (nJ)**	**Leakage (mW)**
45nm	SRAM	10.6	0.51	10.6	0.05	630
	STT-MRAM	7.6	0.15	16.7	0.65	24

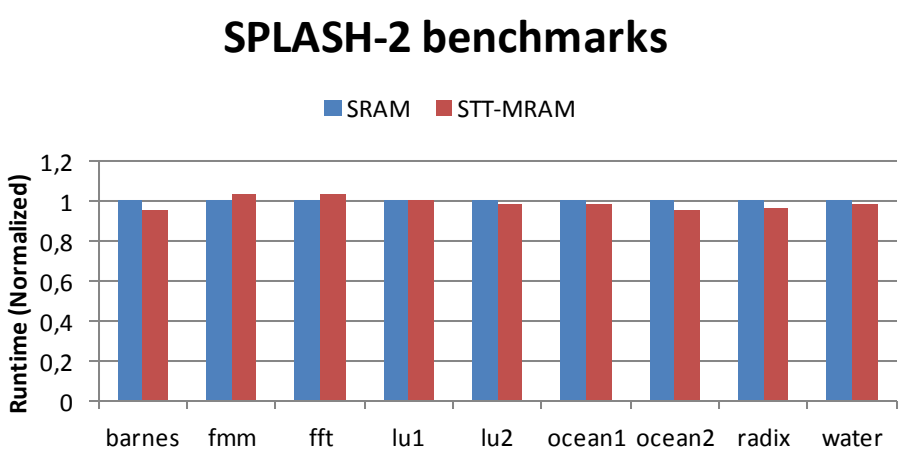
/26

- STT-MRAM < SRAM for reads
  - Small area of STT-MRAM
- STT-MRAM > SRAM for writes
- STT-MRAM << SRAM for static power

High Performance Computing using STT-MRAM

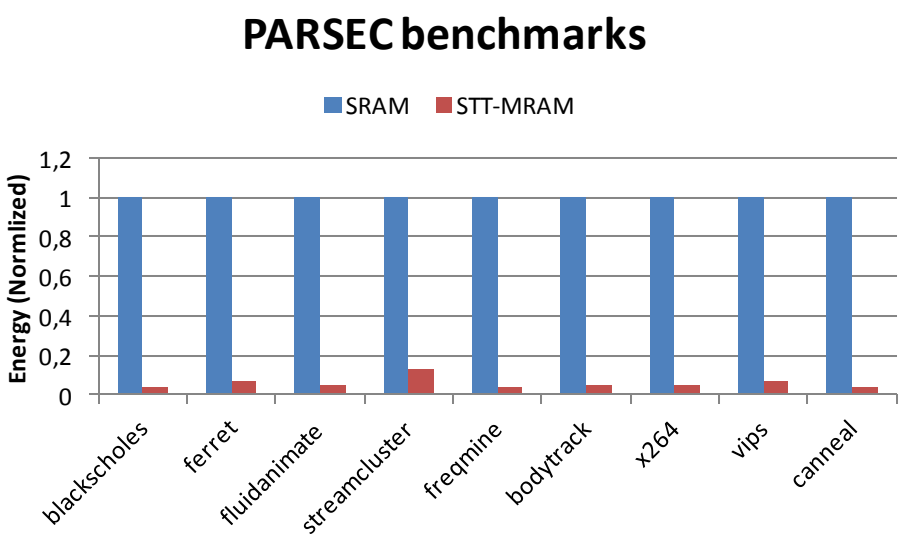
- Set of results…
  - Runtime
    - Similar performance when using STT-MRAM

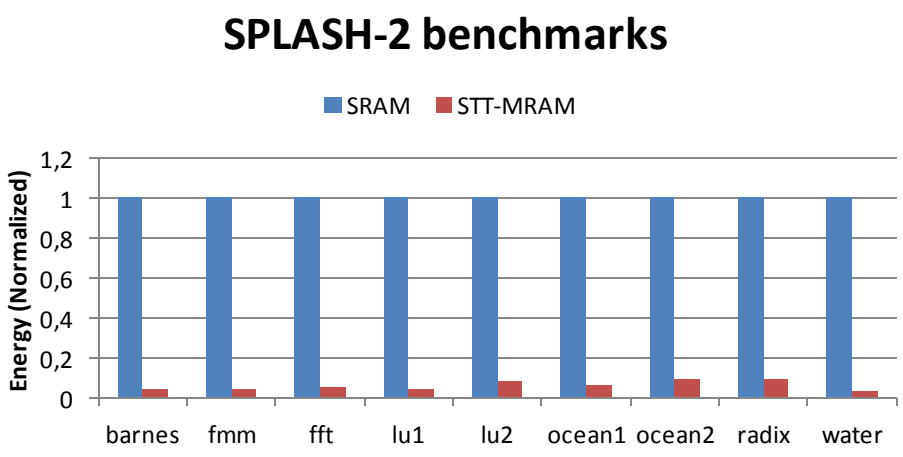




High Performance Computing using STT-MRAM

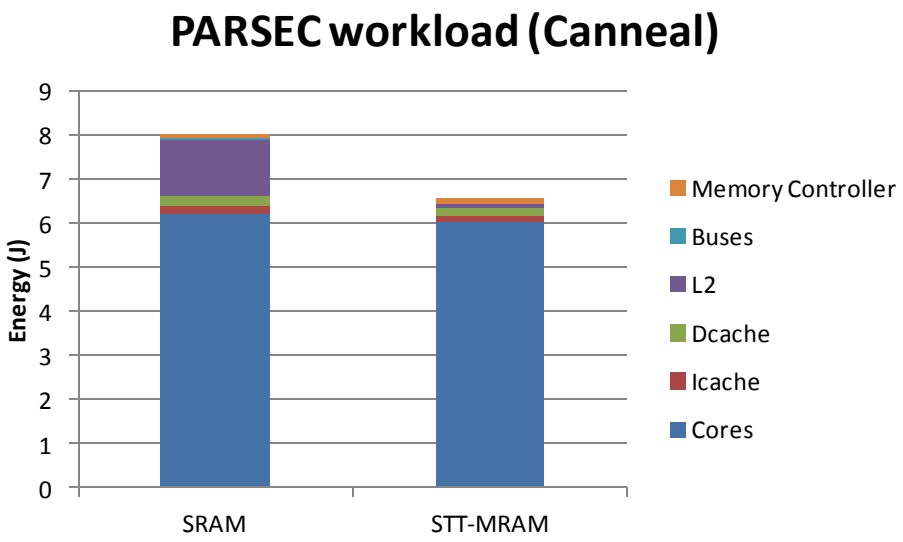
- Set of results…
  - L2 cache energy
    - STT-MRAM based L2 cache consumes >80% less energy than SRAM based L2

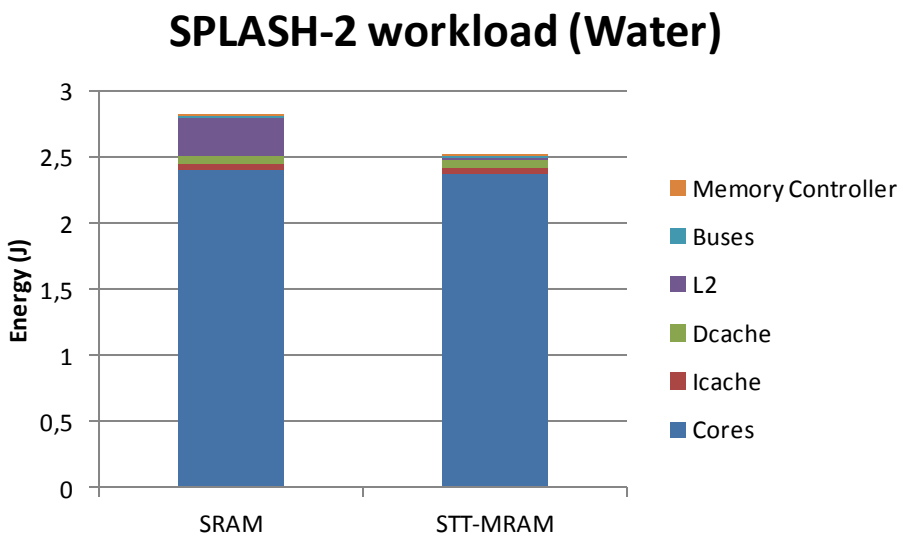




High Performance Computing using STT-MRAM

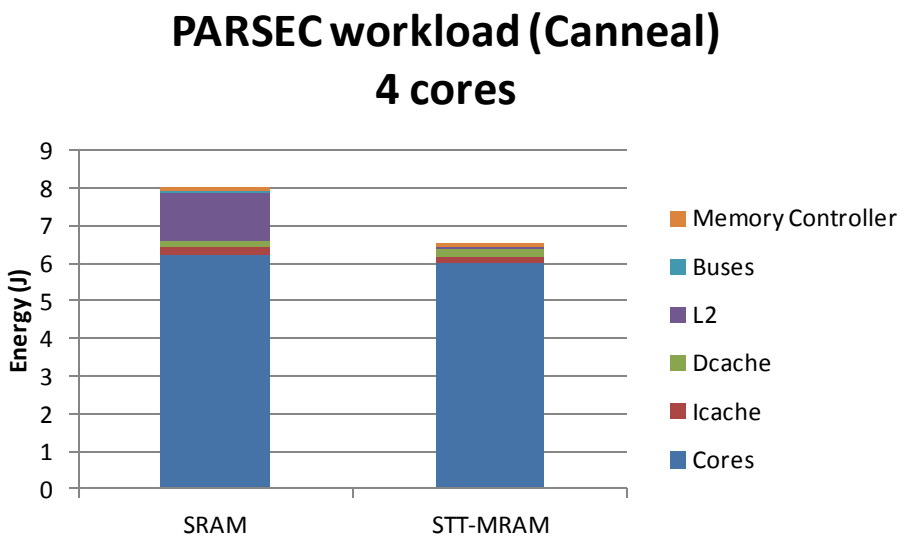
- Set of results…
  - System energy
    - Evaluate the impact of the memory part compared to the rest of the system

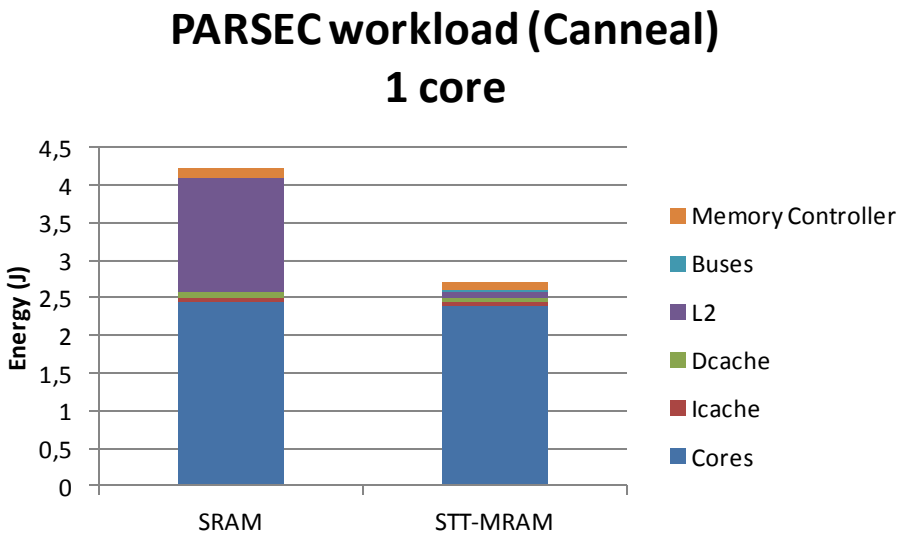


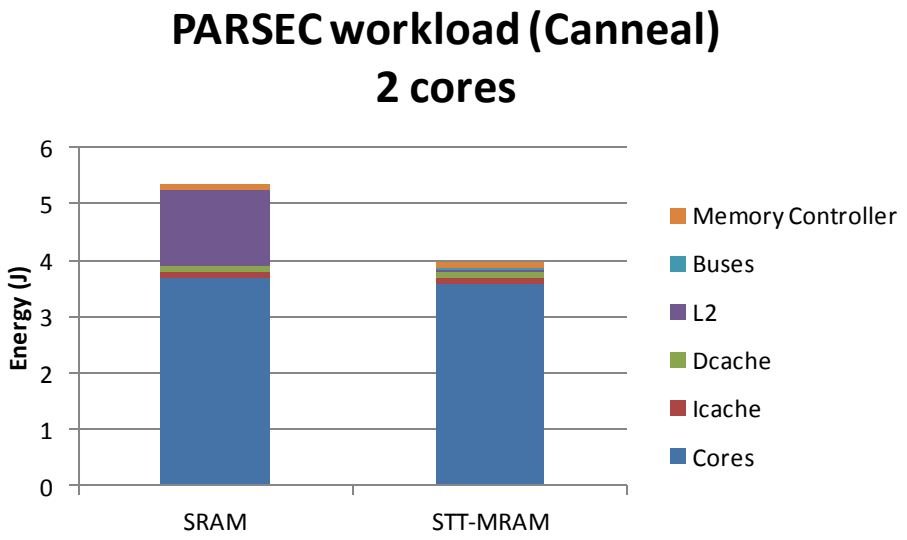


High Performance Computing using STT-MRAM

- Set of results…
  - System energy
    - The impact for different number of cores







Conclusions

- **STT-MRAM is promising for:**
  - Energy-efficient & Reliable embedded systems
    - Normally-off computing
    - Checkpoint / Rollback
  - Caches memories for High Performance Computing

- **A system level simulation framework is developed to enhance the developement of STT-MRAM and other memory technologies**

Future Work

- **Strenghten the results by designing a real system-on-chip based on STT-MRAM**
  - Ongoing work (European Project → GREAT)
- **Explore STT-MRAM at main memory level**
  - Ongoing work
    - Extension of the simulation framework
- **Explore other memory technologies**
  - Spin-Orbit-Torque MRAM
  - Voltage-Controlled MRAM