FROM MEMORY TECHNOLOGY AND ARCHITECTURE TO COMPUTING WITH NON-VOLATILE MEMORY

Pr. Lionel TORRES, lionel.torres@umontpellier.fr

Thanks to : S. Senni, G. Patrigeon, F. Ouattara, G. Sassatelli, A. Gamatié, J.Y Peneau, M. Robert, P. Benoit

ARCHI 2017 - Nancy

Summary

1 – Context and objectives of the lecture

2 – Classical technologies and memory architecture overview (SRAM, DRAM, FLASH)

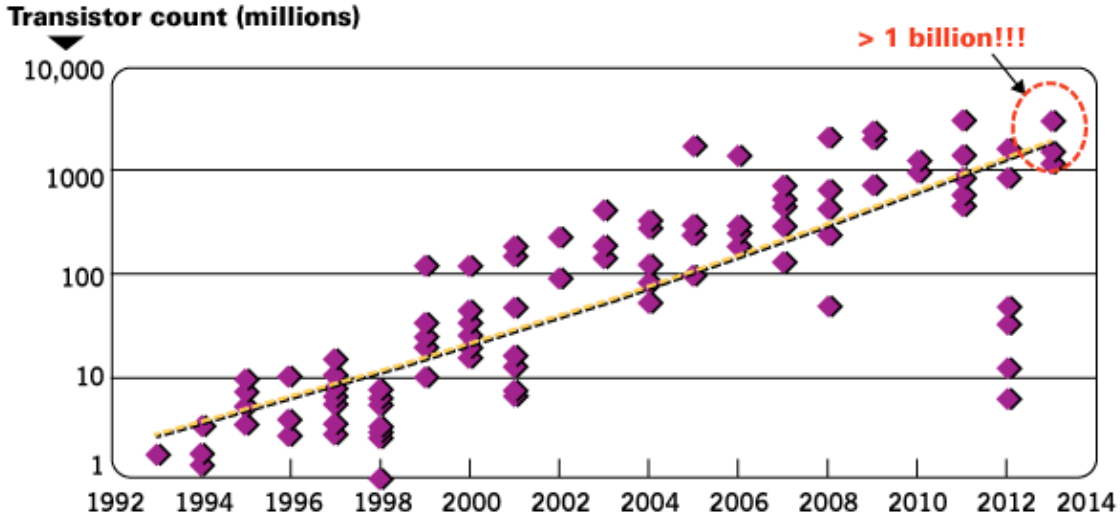
3 – Emerging memory technologies

4 – Computing with Non-Volatile memory technologies

- For high performance computing applications
- For Embedded applications (Non-volatile processor)
- For secure applications

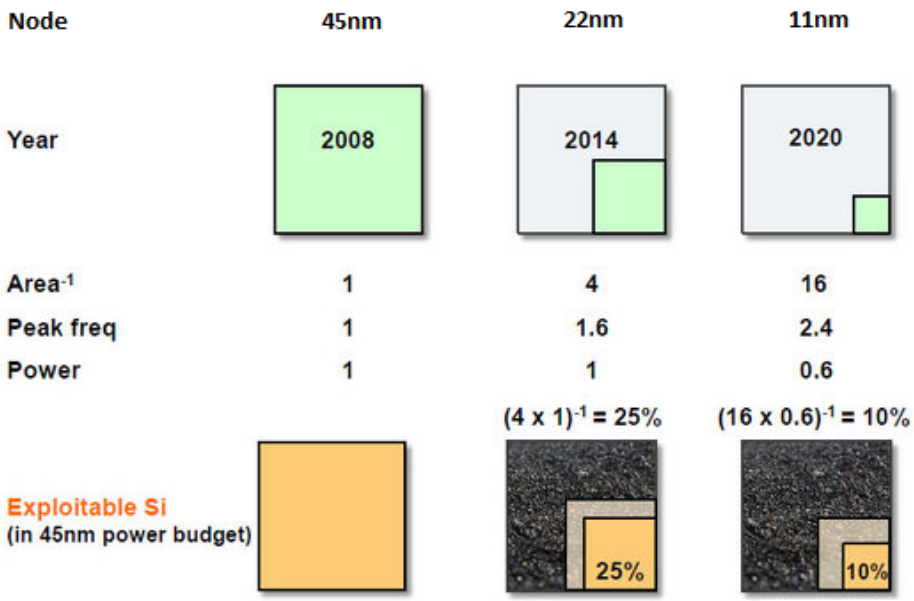
5 - Conclusions

1- Context



*Source : **ISSCC 2013 High-performance digital trends***

Increase of chip complexity



- **Power limits the active silicon area (Dark silicon)**
- **Heat dissipation wall**
- **Power modes**
- **Memory a key component**

*Source : **« ARM CTO warns of dark silicon », eetimes, 2010***

1- Context

- **Observation**
  - Decreasing size of devices
    - ➔ power consumption and heat issues
    - ➔ stagnation of performance
  - Why ?
    - Leakage current of CMOS devices
    - Volatility
- **Memory: a key component**

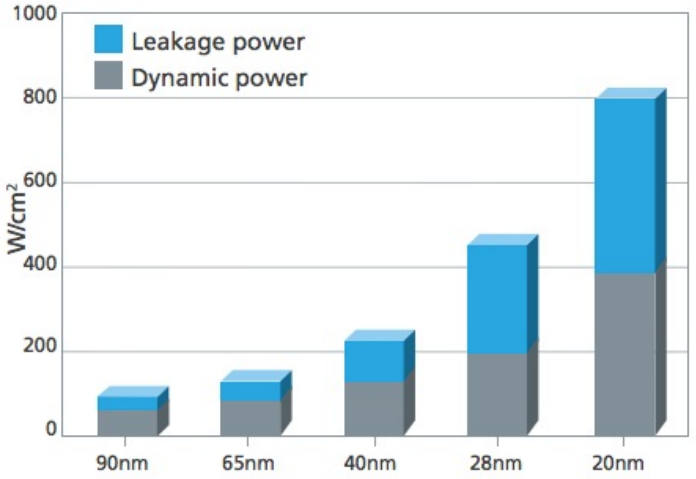
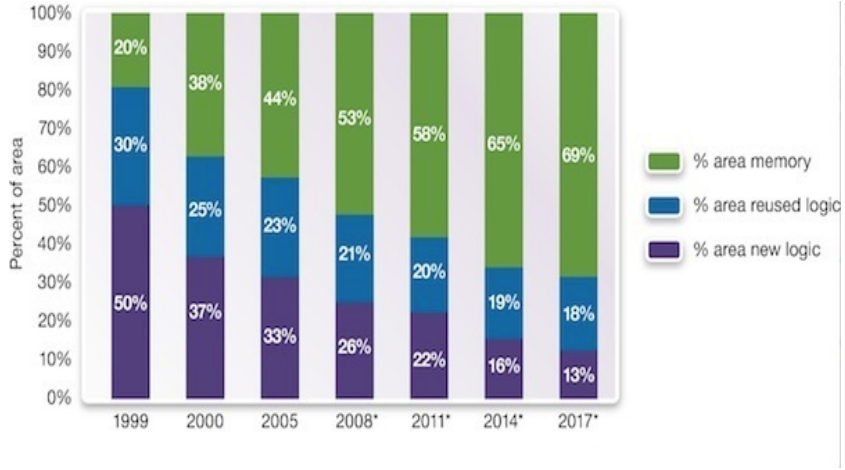
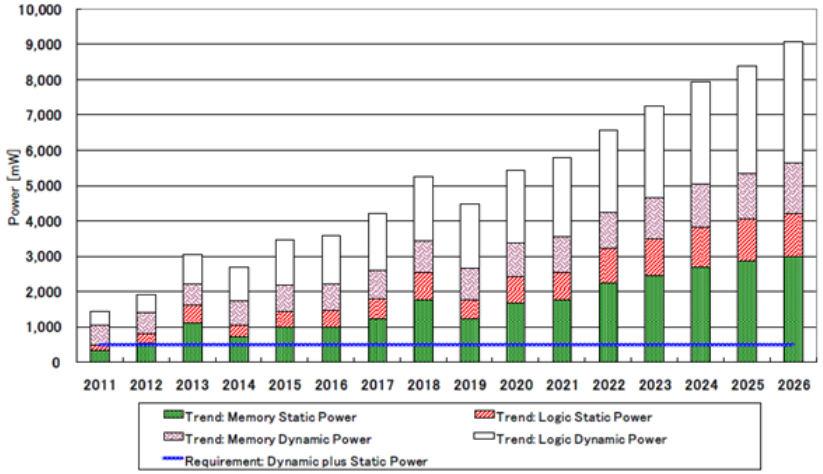


Figure 1: Leakage power becomes a growing problem as demands for more performance and functionality drive chipmakers to nanometer-scale process nodes (Source: IBS).



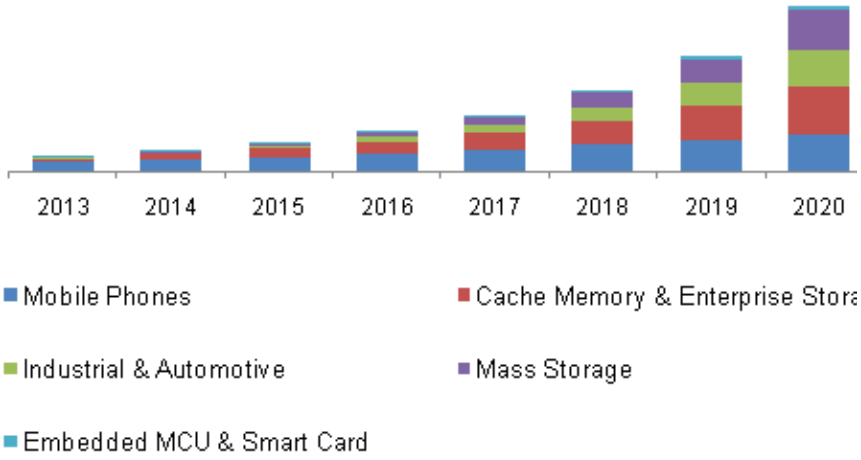
*Source : **Semico Research Corporation***



*Source : **ITRS***

1- Context

Memory market trends



DRM market trends

Flash Market trends

1- Lecture Objectives

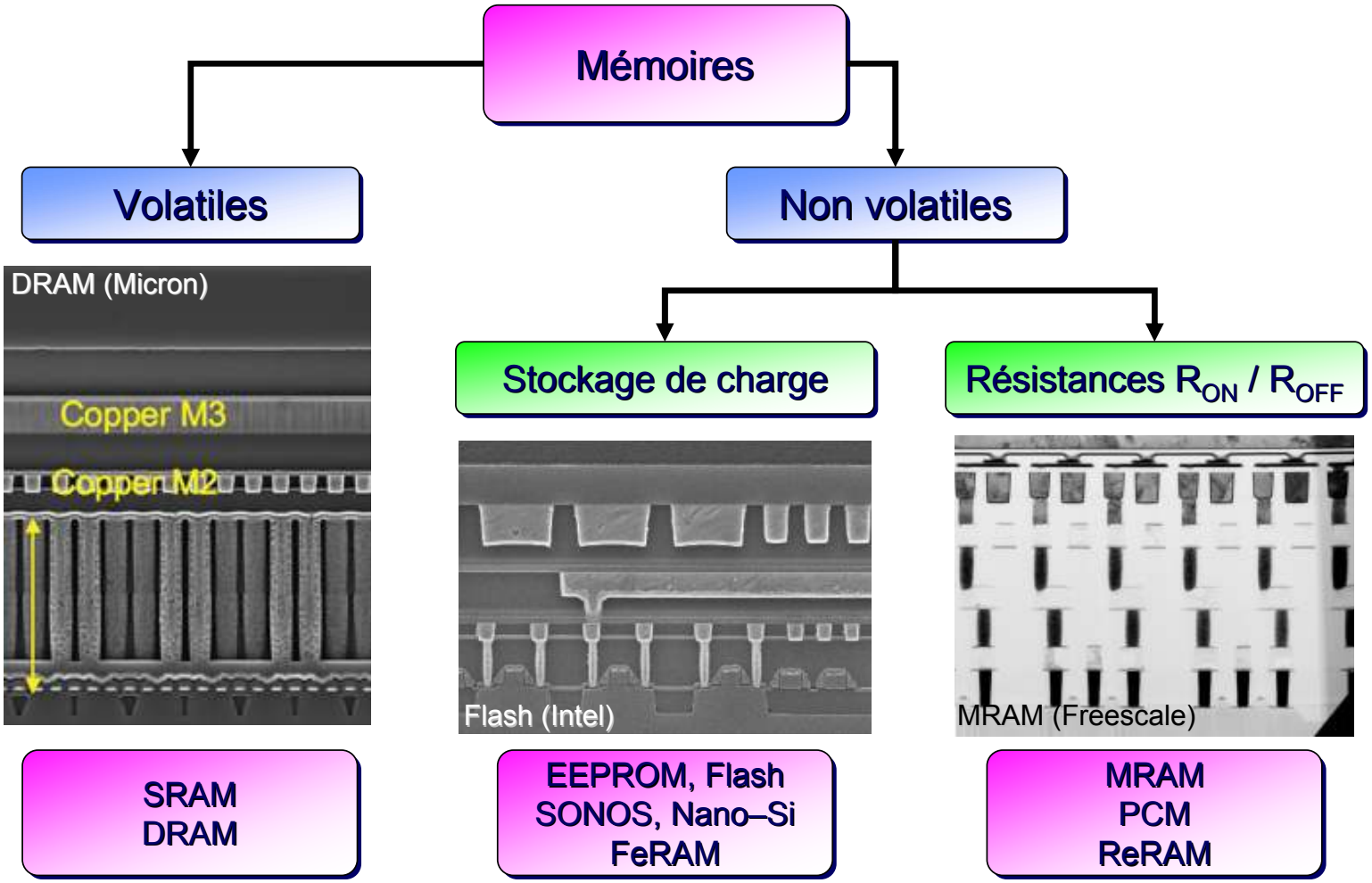
1/ Giving a memory technology architecture overview

2/ Discussing on promising memory technologies

3/ Understanding which type of memory technologies related to applications

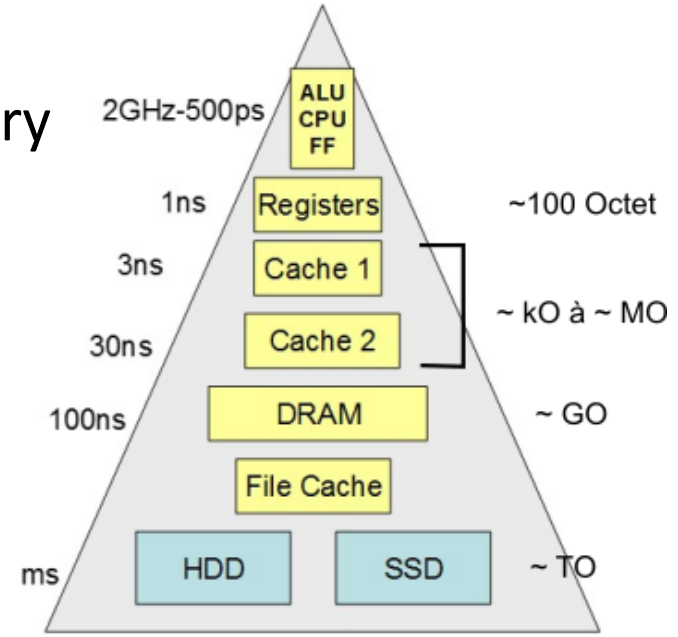
4/ Illustrating some case study to demonstrate that logic in memory could help to reach ultra low power consumption applications

2 - Technology and memory architecture overview



Memory Hierarchy

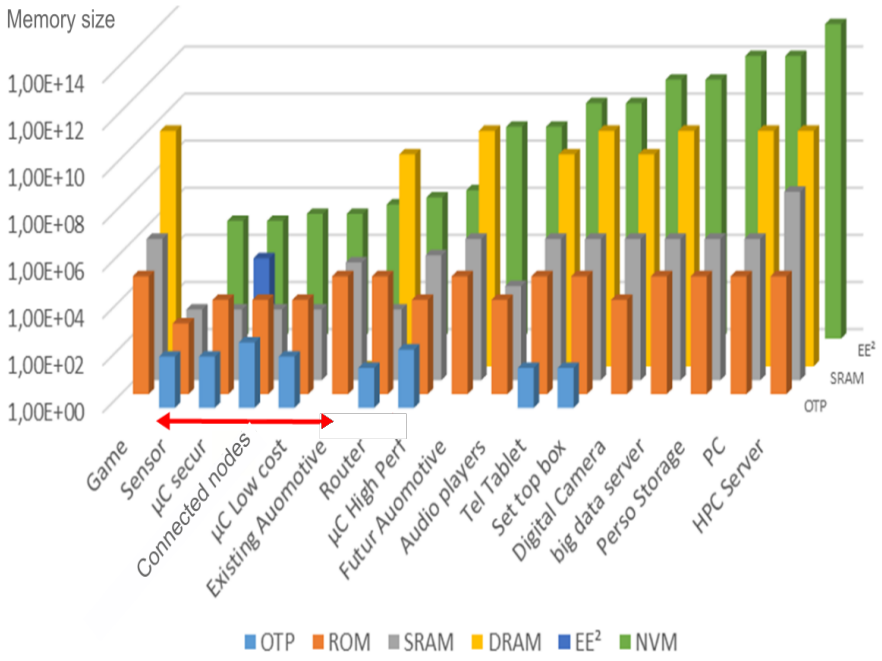
- Closer the memory is to computing/calcul, the faster it must be
- Processor Registers are part of the memory hierarchy
- SRAM Cache memory connected to the processor
- Main memory in general is DRAM
- Data storage, slow but very dense, Non-volatile memory
- What is important : the cell regularity !



T.Kawahara, IEEE Design and test of computers, 52, Janv/Feb 2011)

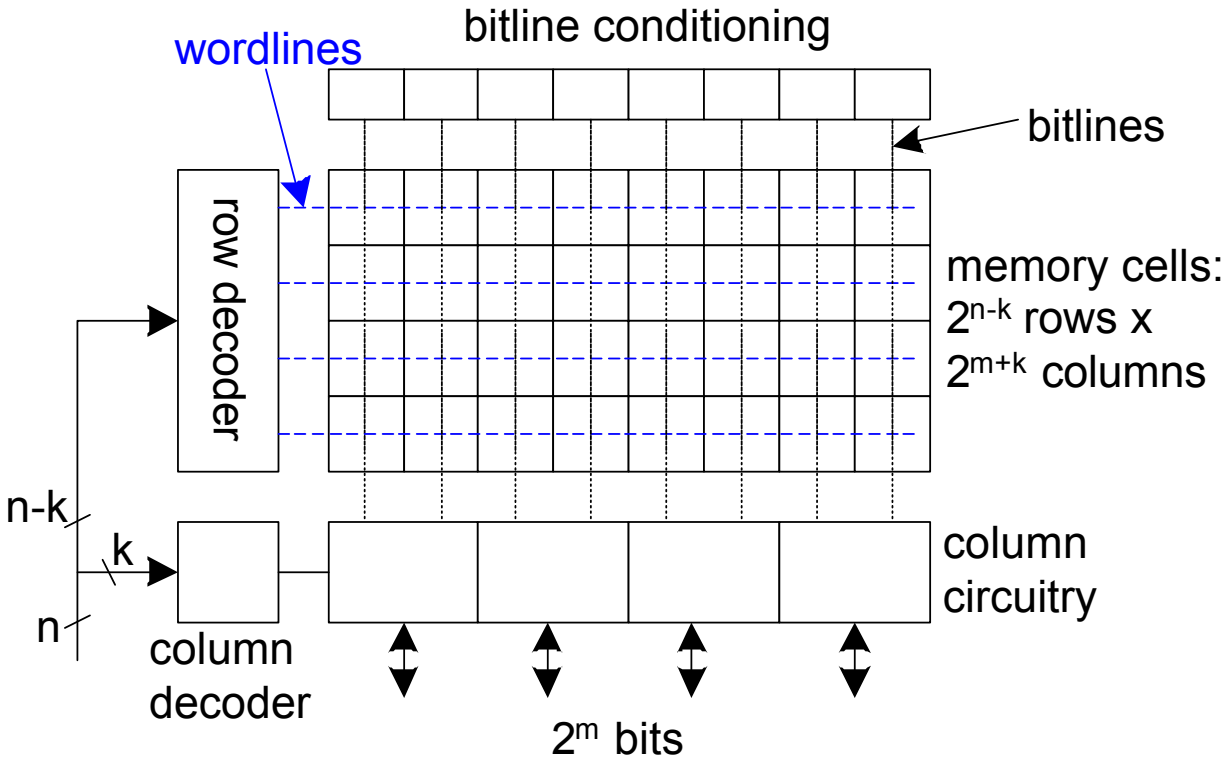
Memory & applications requirements

- **Main metrics**
  - Cost
  - Performances
  - Data retention
  - Security
  - Physical behaviour
  - Power consumption
  - Celle size
  - Scalability



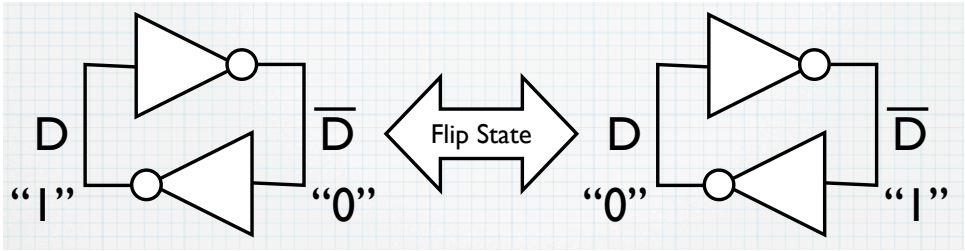
Memory array architecture

- 2^n words of 2^m bits each
- If n >> m fold by 2^k into fewer rows of more columns
- What is important : the cell regularity !

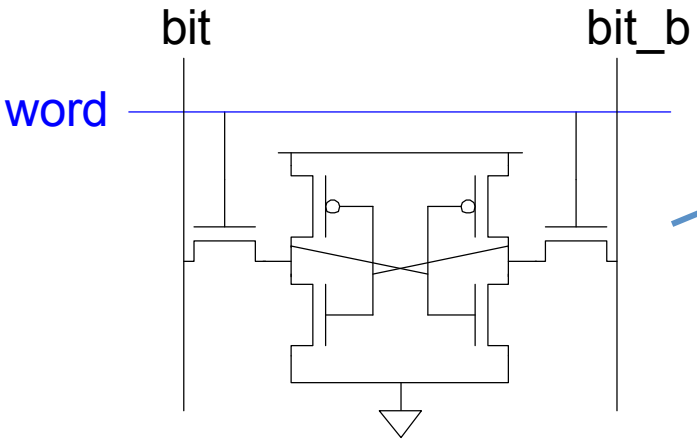


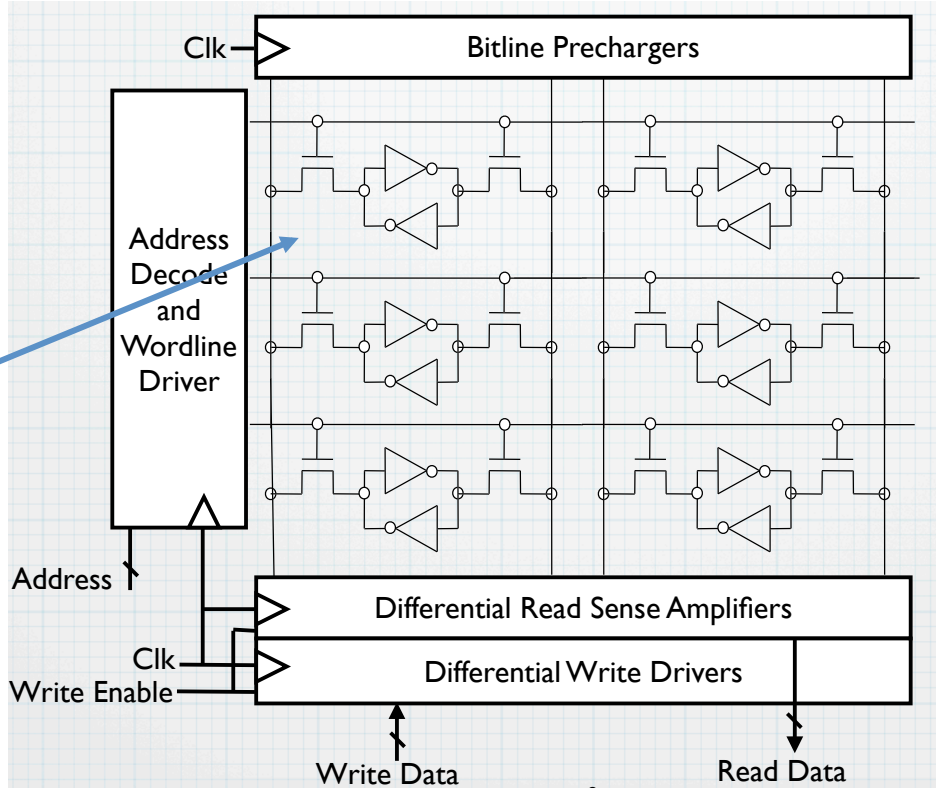
2 - Technology and memory architecture overview

SRAM Technology



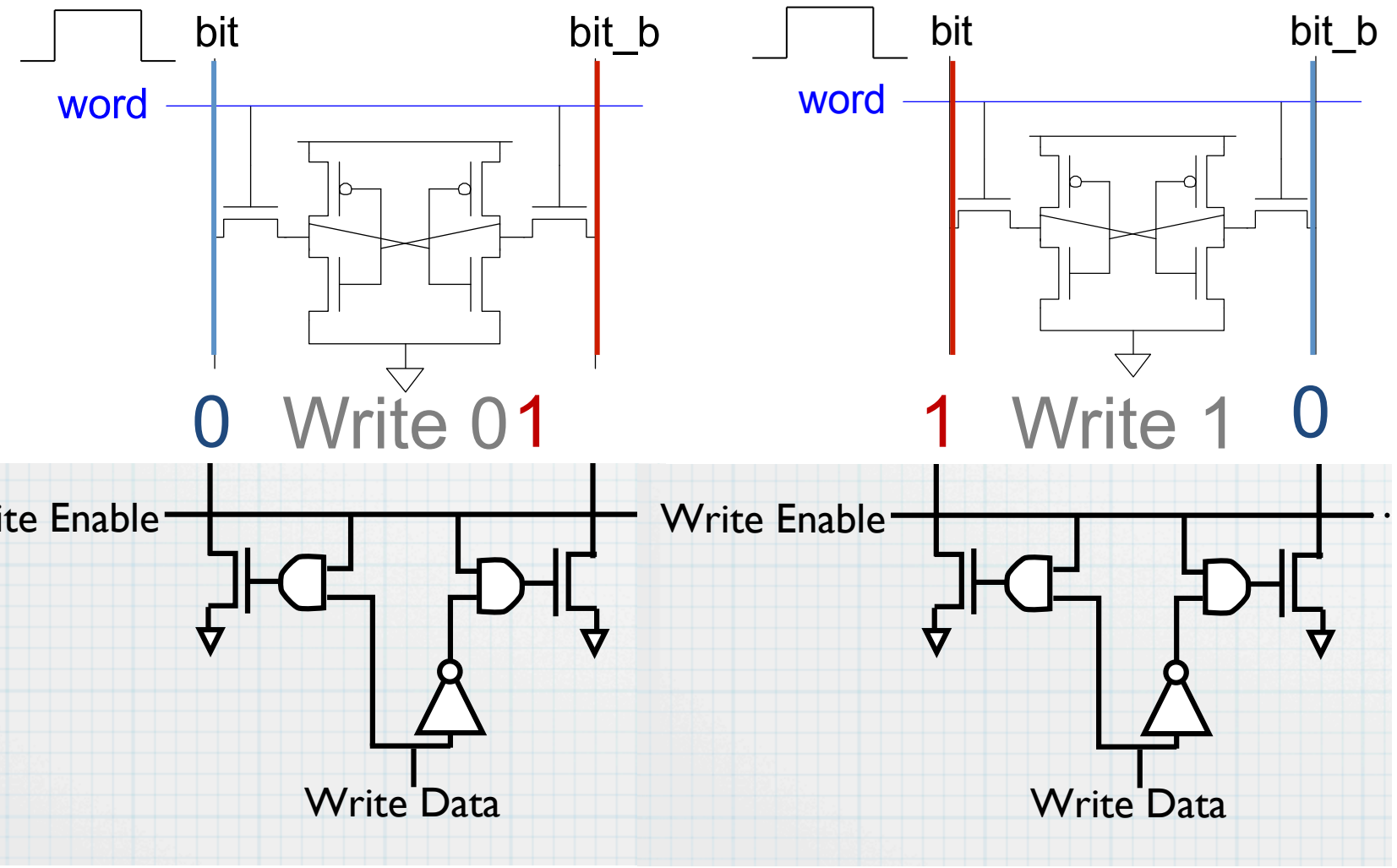
SRAM cell (6T)





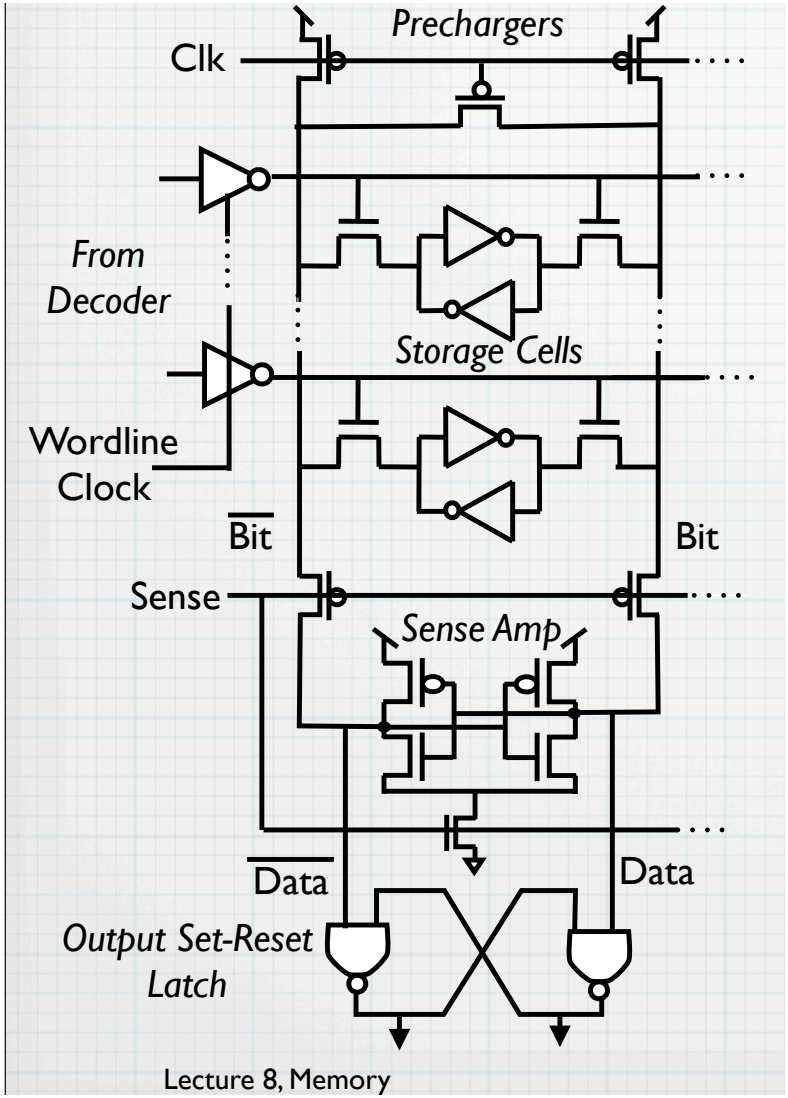
2 - Technology and memory architecture overview

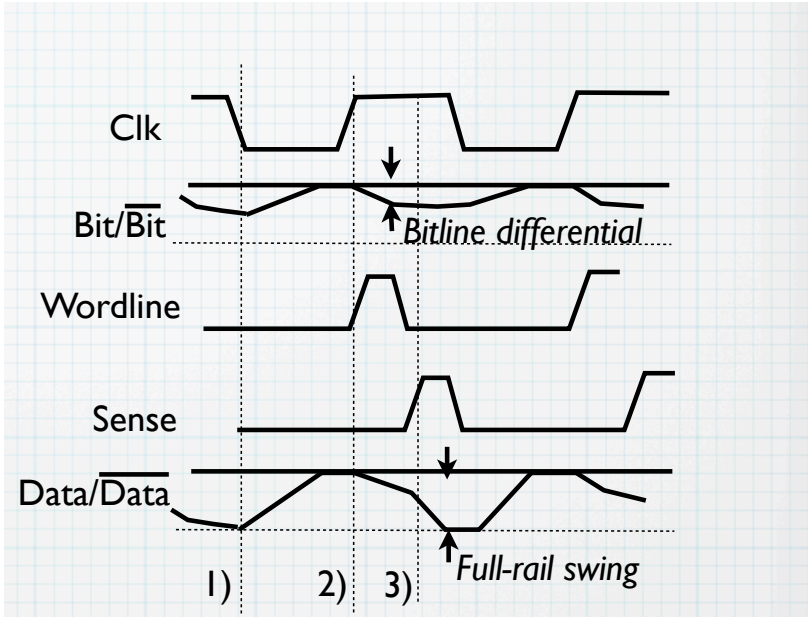
SRAM Write



2 - Technology and memory architecture overview

SRAM Read





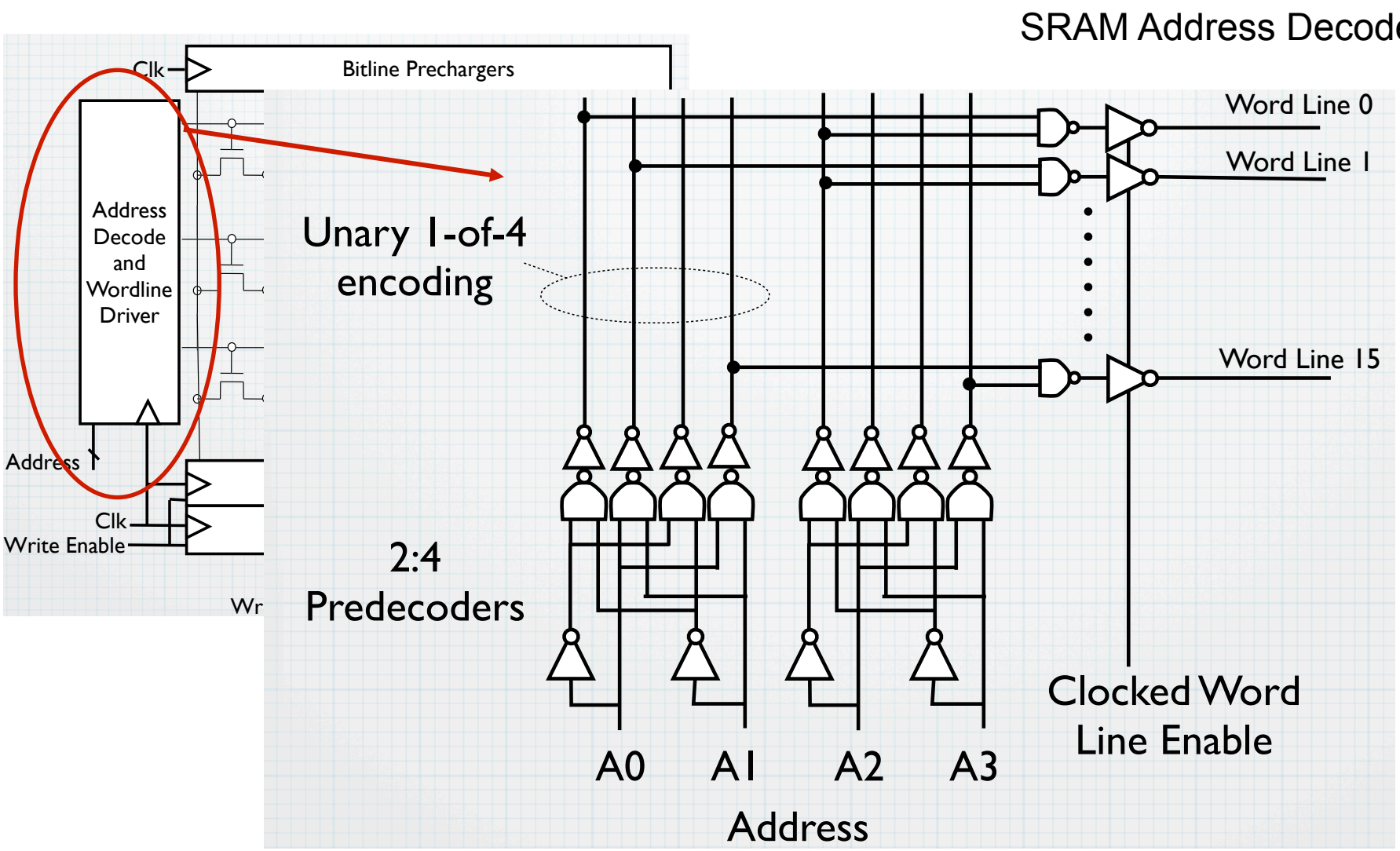
1) Precharge bitlines and senseamp

2) Pulse wordlines, develop bitline differential voltage

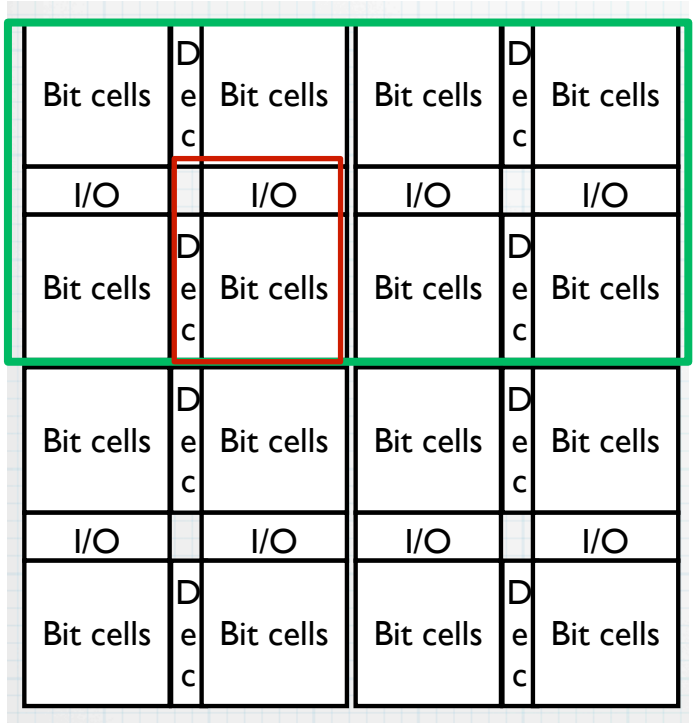
3) Disconnect bitlines from senseamp, activate sense pulldown, develop full-rail data signals

Pulses generated by internal self-timed signals, often using "replica" circuits representing critical paths

2 - Technology and memory architecture overview



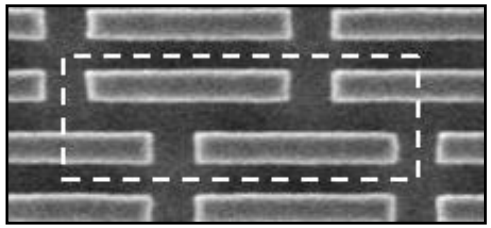
2 - Technology and memory architecture overview



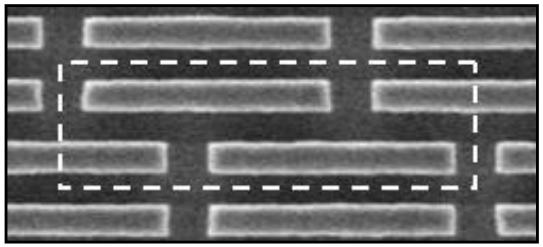
- Using Banks and sub-banks to construct larger array
- Due to RC delays 128-256 bits in row/column (sub-banks)
- For energy efficiency only one Bank (sub-bank) activated at same time
- Delay and energy dominated by I/O wiring

2 - Technology and memory architecture overview

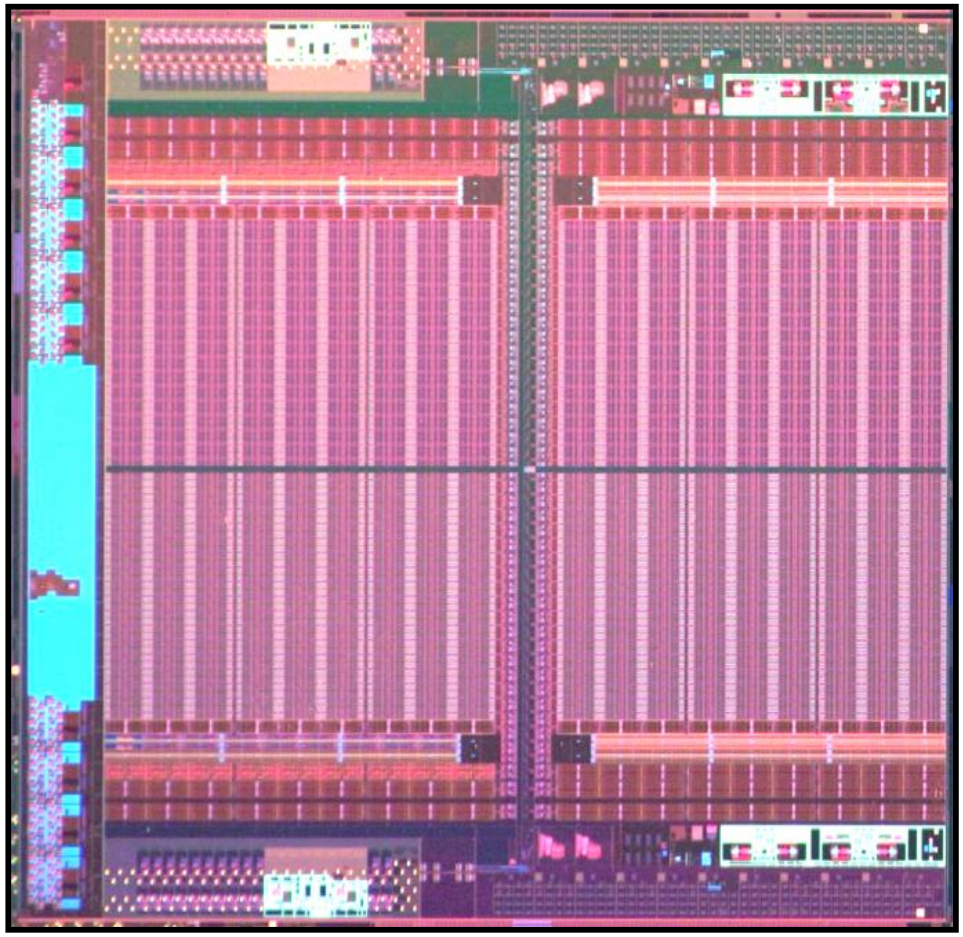
SRAM Layout memory



0.092 um^2 SRAM cell
for high density applications



0.108 um^2 SRAM cell
for low voltage applications

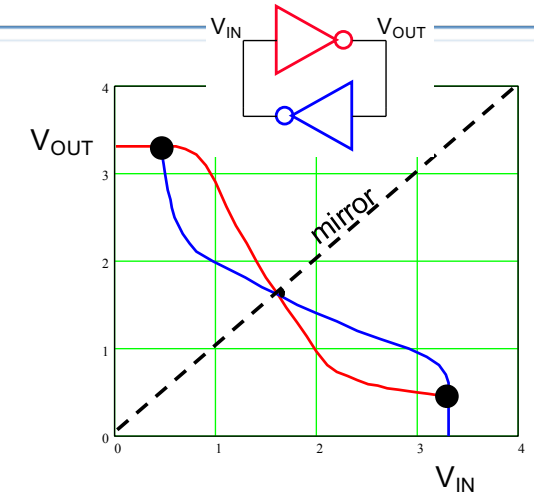


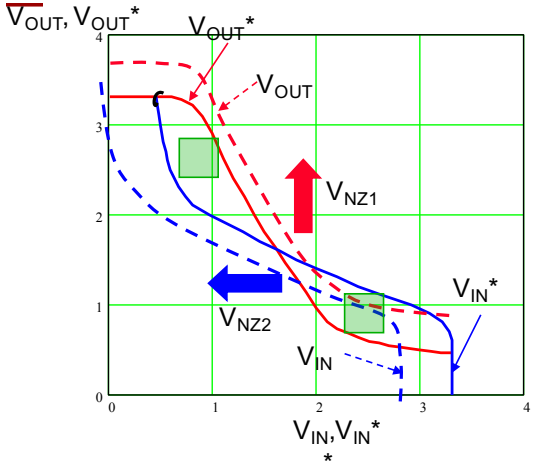
Intel 22 nm SRAM

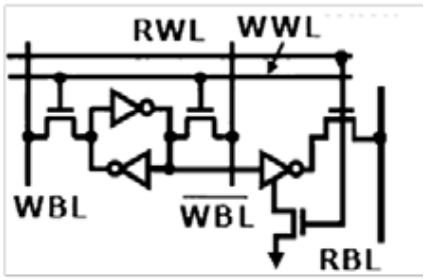
2 - Technology and memory architecture overview

SRAM limitations

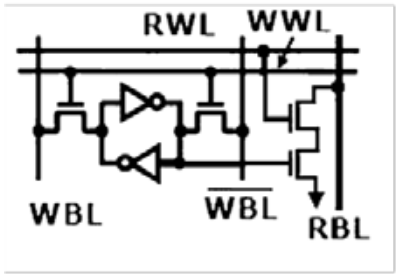
- Static Noise margins
- Very high sensibility to process variability
- High sensibility to temperature
- High Leakage for advance node technology
- To overcome these drawbacks number of Tr per cell increase !





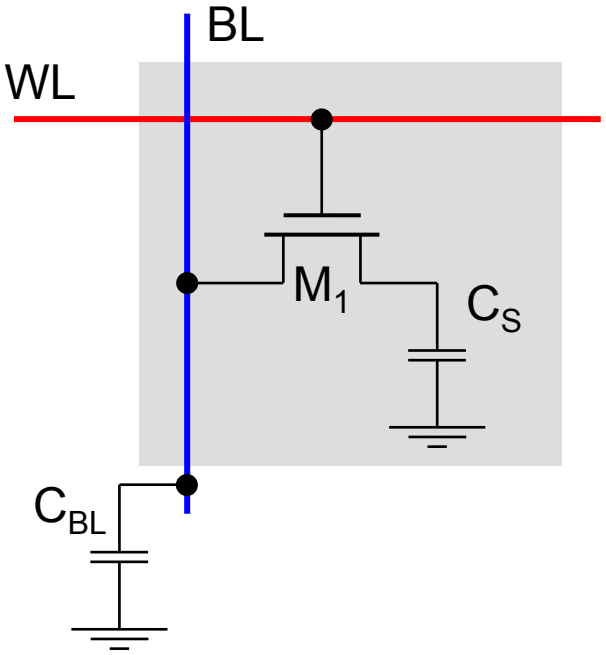


10T SRAM cell [13]



8T SRAM cell[11]

DRAM Technology

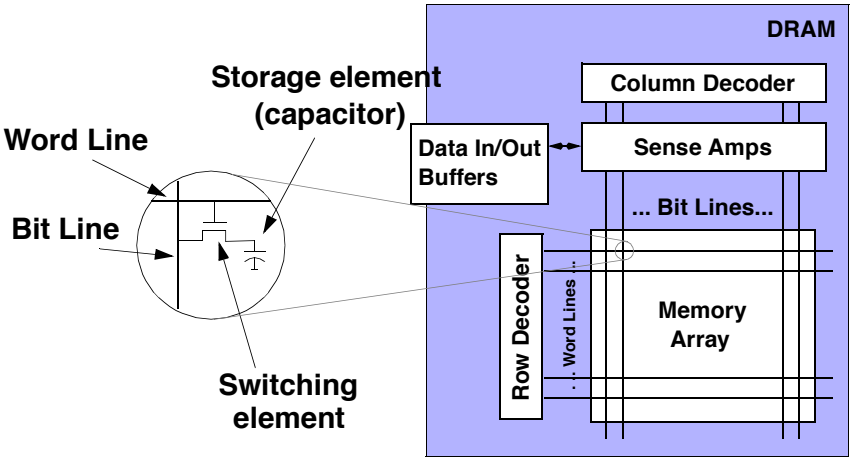


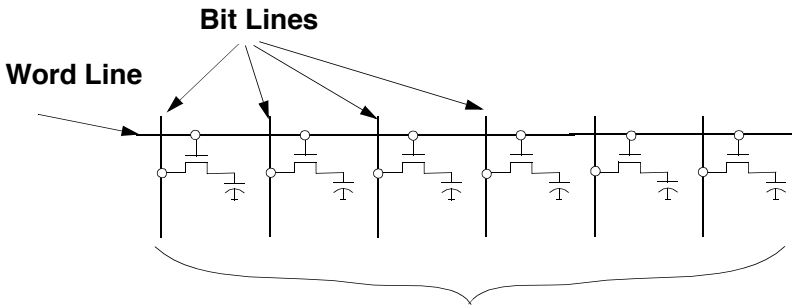
- To write – $W_L = V_{DD}$, B_L is "0" or "1" depending of the value to store
- To read – VB_L precharged at V_{PRE}
  Then activate W_L
  - If "1" – V_{BL} ↗ - 1 is detected
  - If "0" – V_{BL} ↘ - 0 is detected

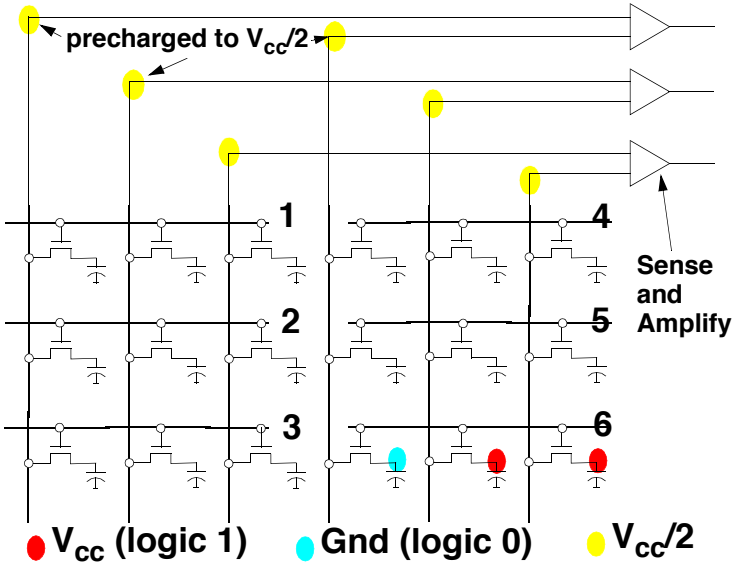
1 – Necessary to refresh the cell (~ms)

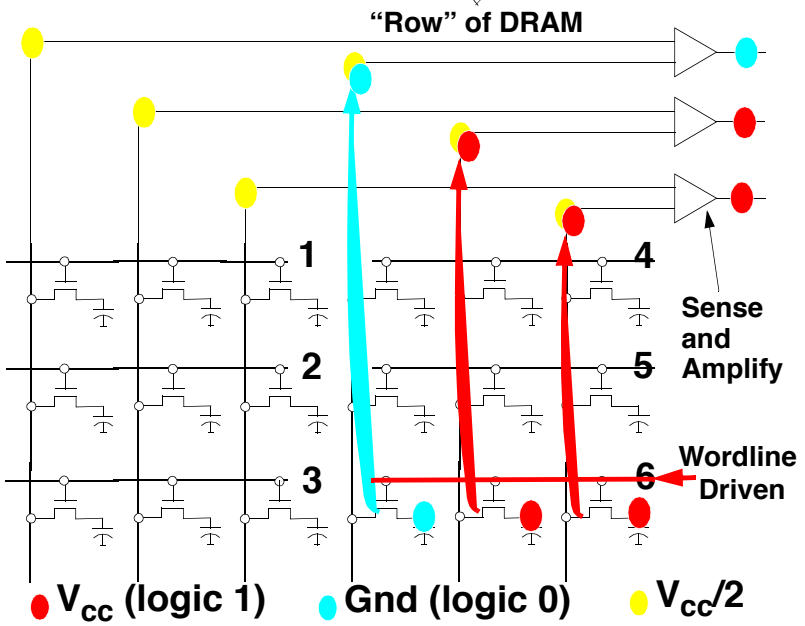
2 – When a read occurs (destructive read), it is necessary to re-write the cell

DRAM Technology



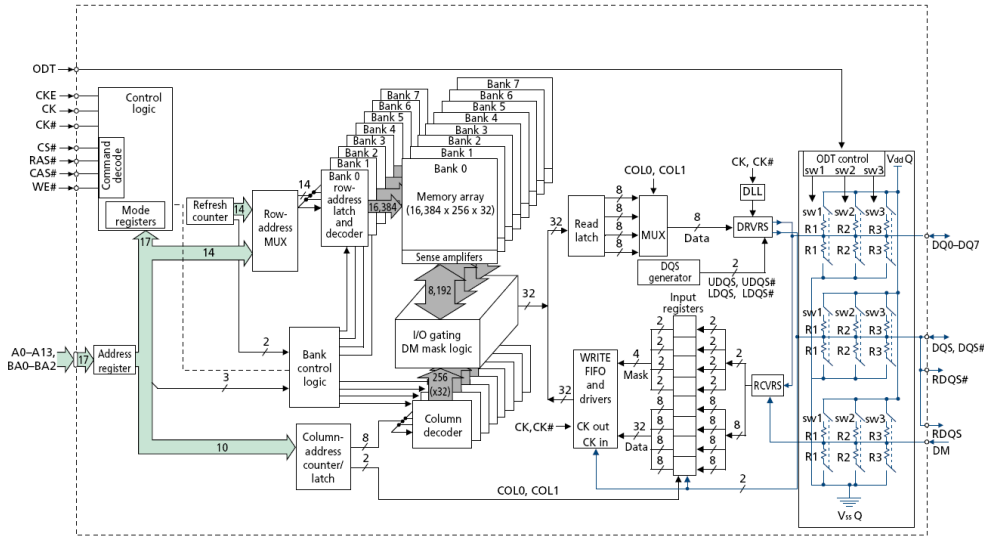


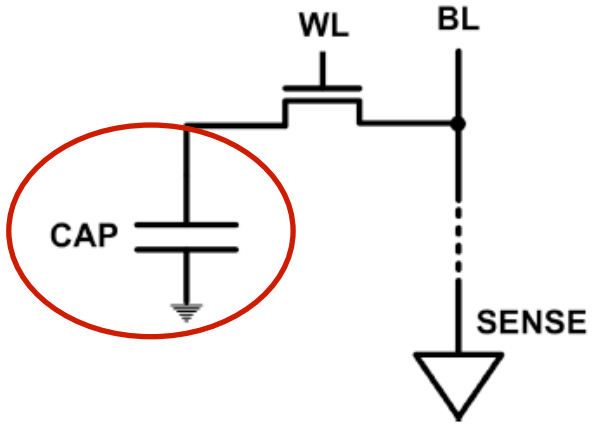




DRAM limitations

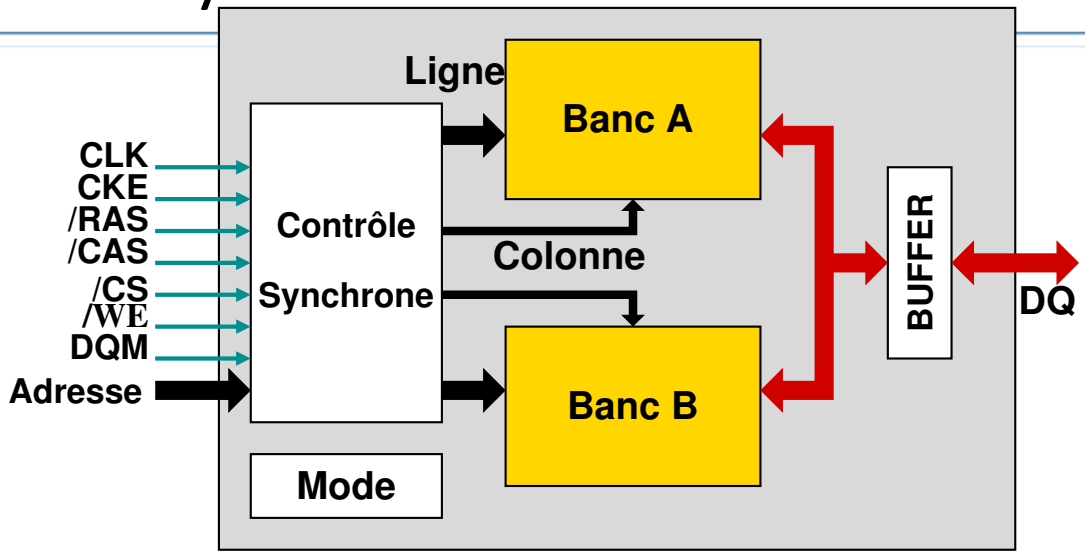
- Capacitor integration (must be large enough)
- Refresh cost is depending of the capacitor
- Access transistor large to avoid static leakage
- DRAM Hard to scale in advance node
- Not easy to embed DRAM (Specific Techno)

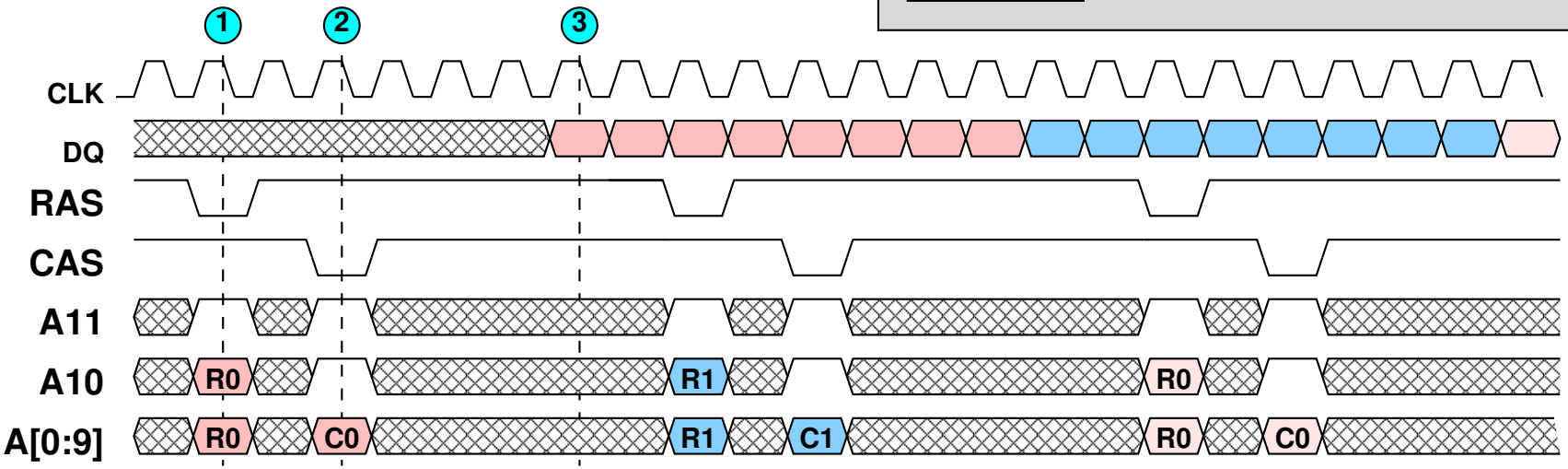




SDRAM

- interleaved (2 banks)– one is refreshing and the other can be accessed
- synchronized to clock and burst mode (without CAS)





Lecture banc A, ligne R0, colonne C0

Lecture banc B, ligne R1, colonne C1

Lecture banc A, ligne R2, colonne C2

Burst = 8 mots

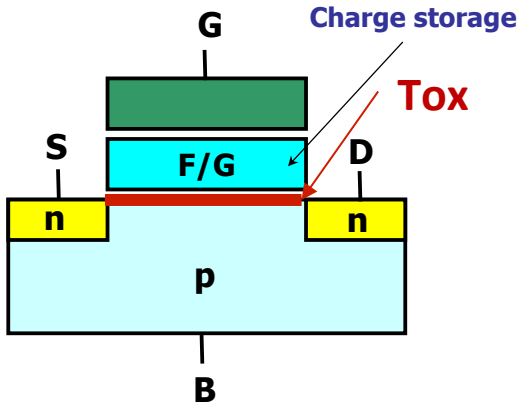
1 Entrée de la ligne R0 Précharge du banc A

2 Entrée de la ligne C0

3 Lecture de la donnée

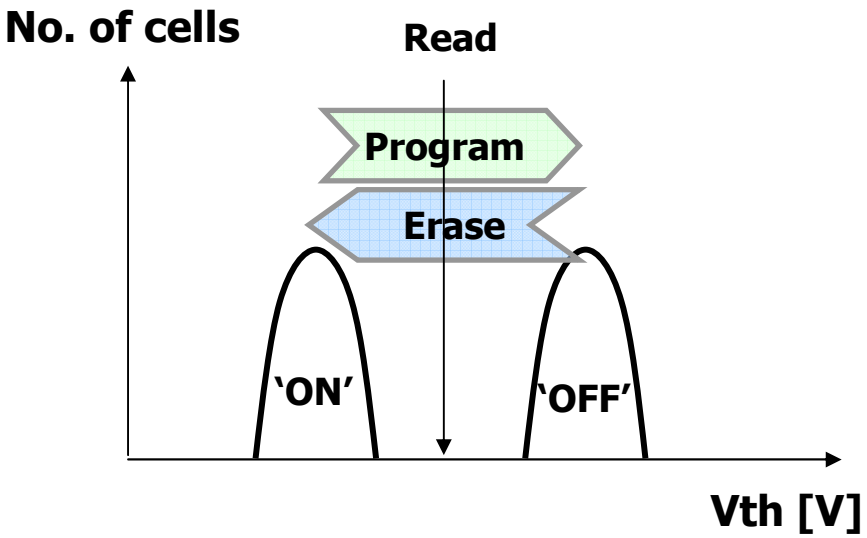
2 - Technology and memory architecture overview

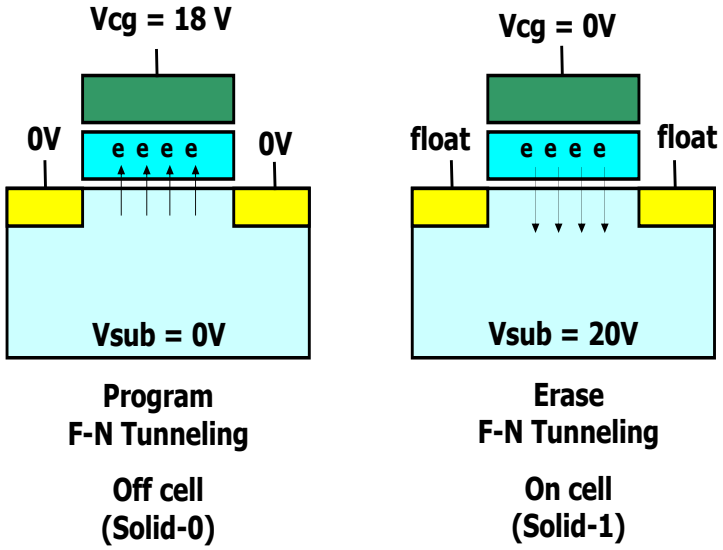
FLASH Technology



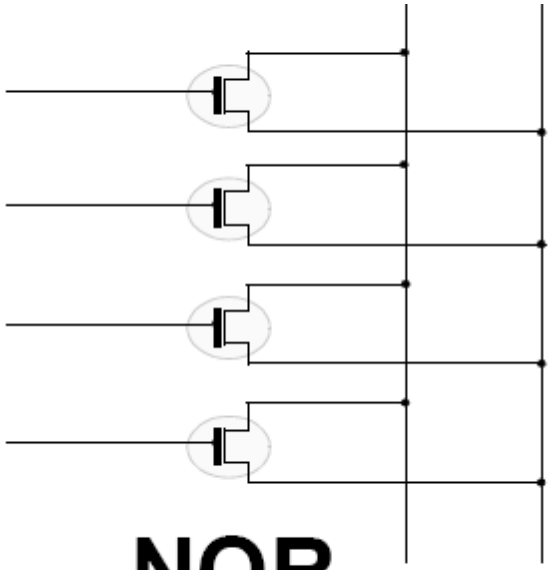
Flash cell

- Double gate where charge storage can be changed – control of the V_{th} of the cell
- Cell Vth changes depending of the amount of F/G charge
- Electrons injected (ejected) into (out of) the F/G through **Tox** with electric field across **Tox**

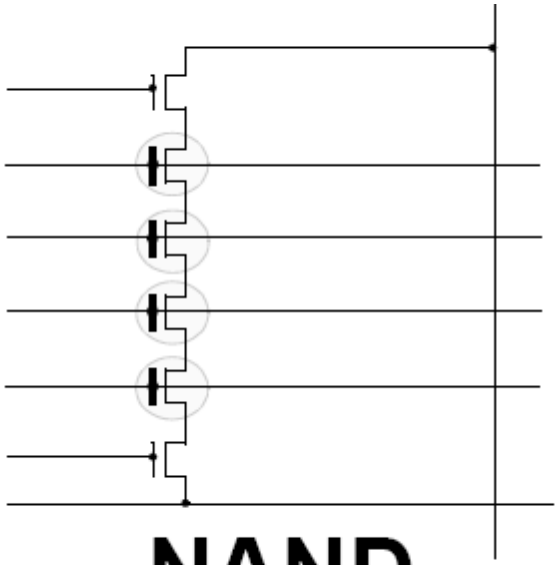




FLASH Technology



NOR

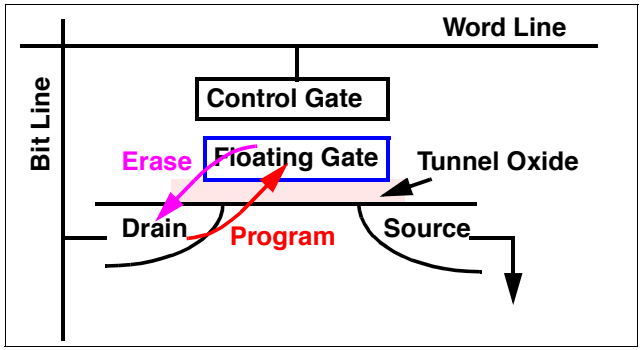


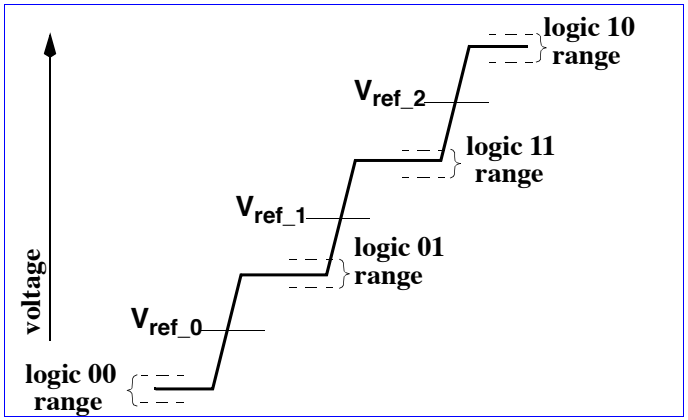
NAND

NOR	NAND
10x better endurance	Smaller cell size
Fast read (~100 ns)	Slow read (~1 µs)
Slow write (~10 µs)	Faster write (~1 µs)
Used for Code	Used for Data

Flash limitations

- Limited number of write/erase (endurance)
- Necessary to generate high voltage (charge pump)
- Access time
- Integration (> 10 masks)
- Scalability, charge retention lose on advanced nodes
- MLC Capabilities appreciated (but complex)





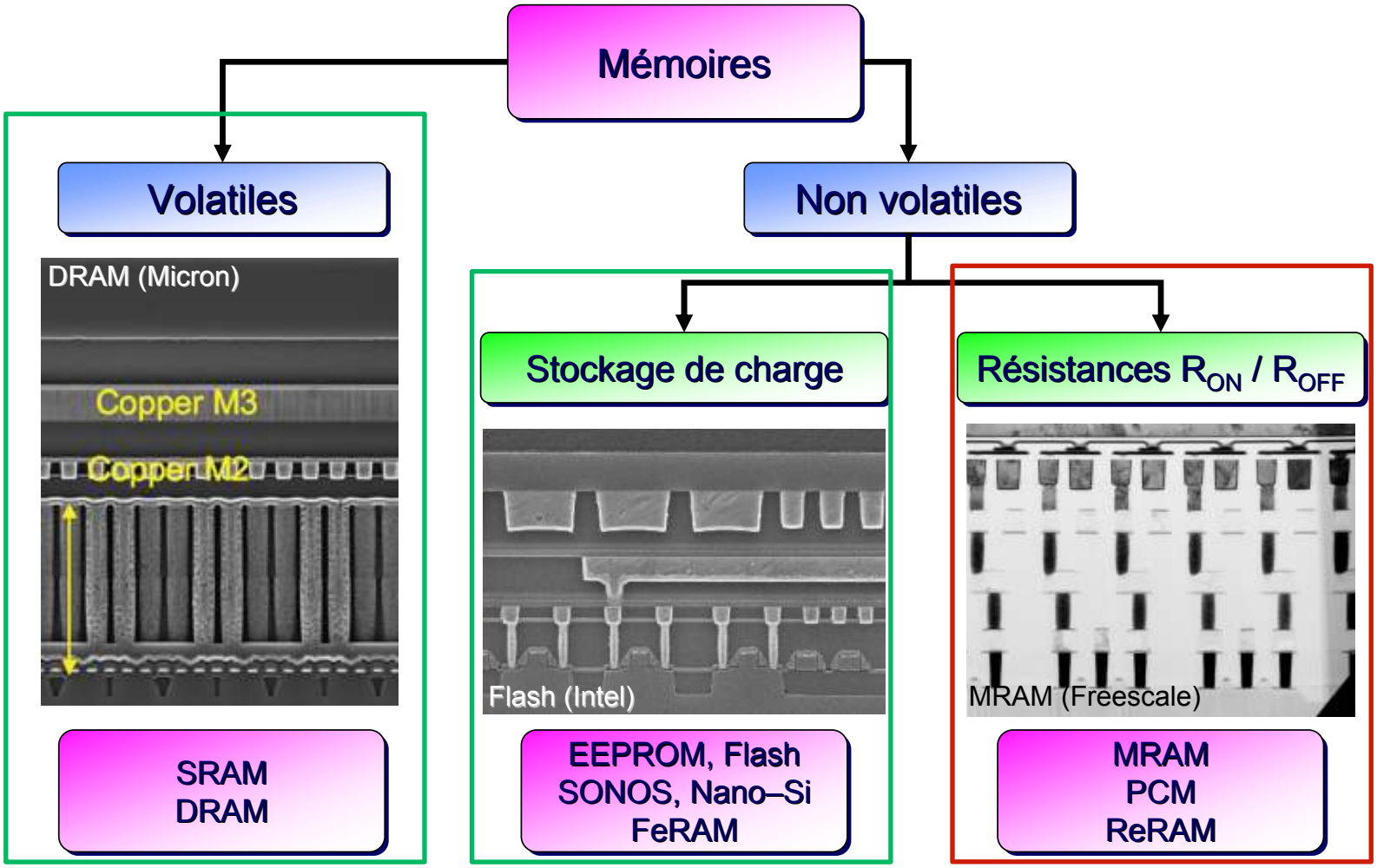
Overall considerations

- Bigger is slower
  - Kbyte – Mbyte → SRAM - fast access time (~ns) – low density (6T/cell) – CMOS compatible – easy to embed - volatile
  - Gbyte → DRAM – reasonable access time (~ 30 ns) – High density (2T/cell) – Specific manufacture process – not easy to embed - volatile
  - > 10 Gbyte → FlASH – Slow access time (~ us) – Very high density (1T/cell) – Specific manufacture process – could be embed (> 10 masks) – non-volatile

- Faster is more expensive
  - SRAM – few $ per Mbyte
  - DRAM - <1$ per Mbyte
  - FLASH - < 1$ per Gbyte

Other technologies have their place as well

3 – Emerging memory technologies



3 – Emerging memory technologies

- **Currently used memories:**
  - SRAM for fast working memories
  - Flash (data storage)
  - FeRAM, smart cards
  - …

- **« Universal memory »candidates**
  - Magnetic tunneling junctions
  - Phase change memory cells
  - Programmable metallization cells
  - OxRRAM
  - …

Universal memory:
"Non volatile RAM"

- Performance of SRAM
- Cell size of DRAM/Flash
- Non volatility of Flash
- Scalability

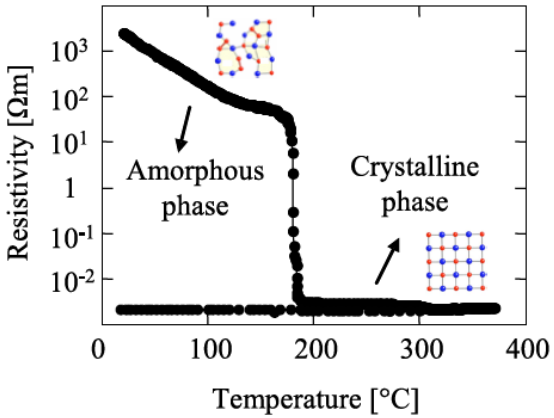
Resistance Switching Memory

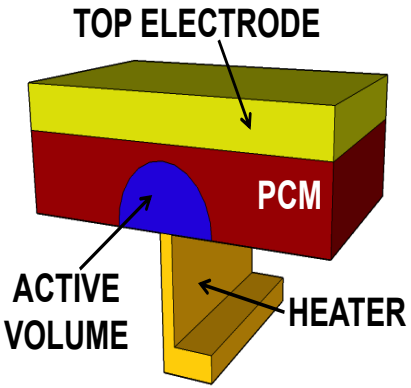
PCRAM Technology

- **Principe de fonctionnement : changement de phase du volume actif de l'élément de stockage de l'information**

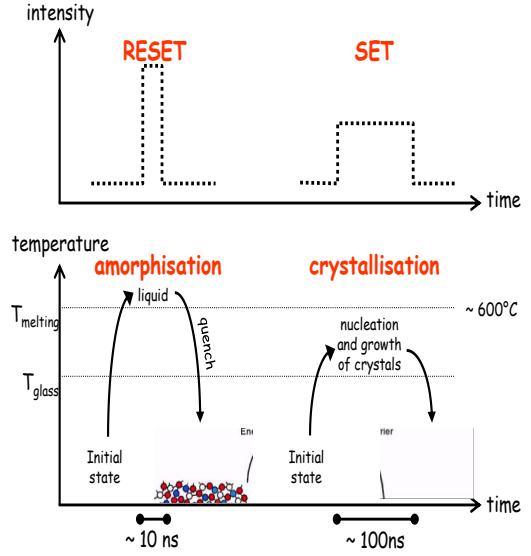
ETAT RESET *Phase amorphe* **ETAT SET** *Phase cristalline*

- **Lecture: contraste de résistance électrique entre la phase amorphe et la phase cristalline.**



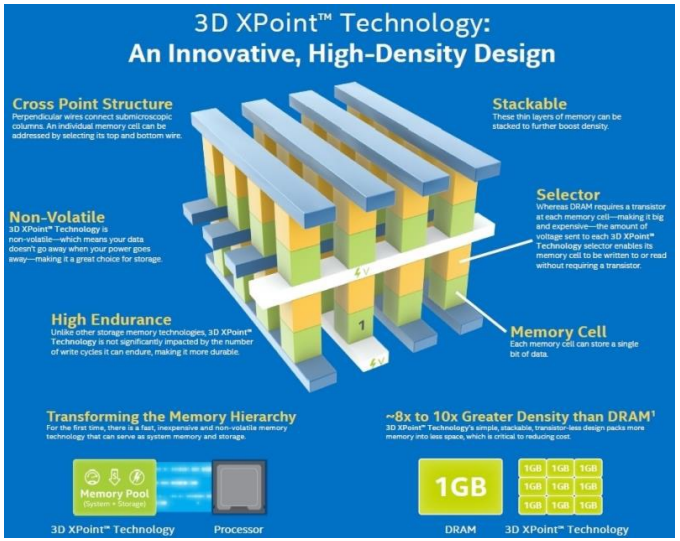


- **Ecriture: changement de phase induit par effet Joule sous l'application d'une impulsion électrique**



PCRAM Technology

- Mémoire non-volatile (NVM)
- Temps d'accès court (12ns)
- Ecriture et effacement rapide (100ns)
- Tension de fonctionnement basse (3V)
- Ratio Roff/Ron important (x1000)
- Haute endurance (10^{12} cycles)
- Rétention prouvée à 10 ans à 150°C y compris à température de soudage !
- Intégration 2D démontrée dans une architecture crossbar pour applications de stockage (cf. 3D XPoint)
- Technologie la plus mature parmi les candidates au remplacement de la flash (technologie industrielle)



From Intel / Micron websites

Défis

- Dérive de résistance et courant de reset relativement haut affectent la mise à l'échelle de la cellule mémoire (élément de stockage et transistor de sélection)

ReRAM Technology

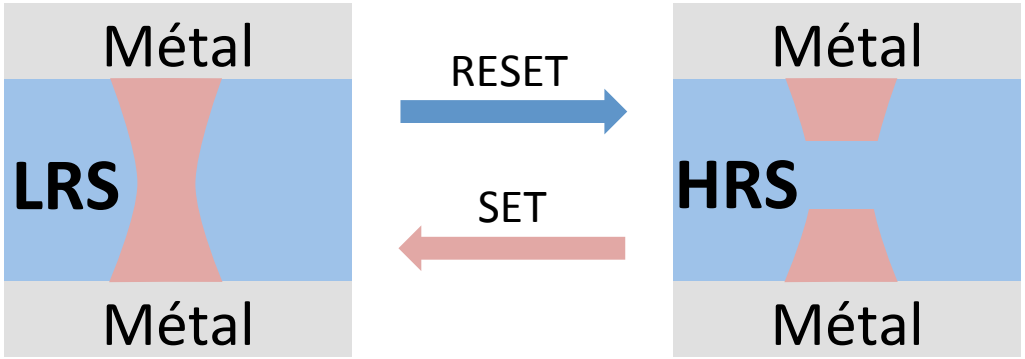
Variation résistance d'une structure Métal\Isolant\Métal contrôlée électriquement

2 Etats

- **HRS → Haute résistance (Etat « 0 »)**
- **LRS → Basse résistance (Etat « 1 »)**

2 Opérations

- **Set → passage de l'état HRS à LRS**
- **Reset → passage de l'état LRS à HRS**



Oxide-base RAM (OxRAM)

- **Filament conducteur = chaine de lacunes d'oxygène (migration)**

Conductive Bridge RAM (CBRAM)

- **Filament conducteur = chaine d'atomes métalliques (Ag, Cu)**

Switching filamentaire unidimensionnel dans les 2 cas

ReRAM Technology

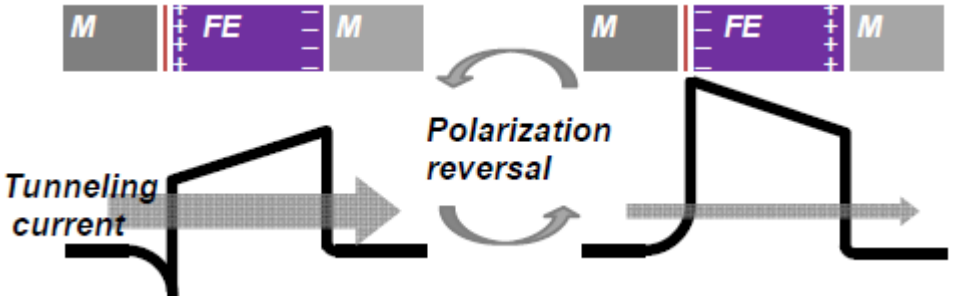
AVANTAGES

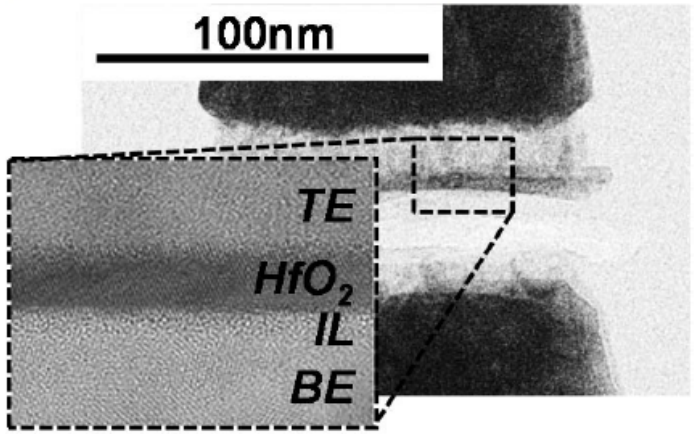
- Structure simple, facile à fabriquer
- Fort potentiel pour la réduction de taille (mécanisme filamentaire)
- Faible tension (1V – 3V)
- Courants de programmation 10-100µA
- Vitesse de commutation < 100ns
- Coût énergétique ~ 100pJ/bit
- Endurance > 10^8 cycles (OxRAM)
- Rétention de l'information > 10 ans à 70°C
- Intégration en BEOL à faible température mais aussi compatible FEOL (OxRAM)
- Des produits déjà démontrés

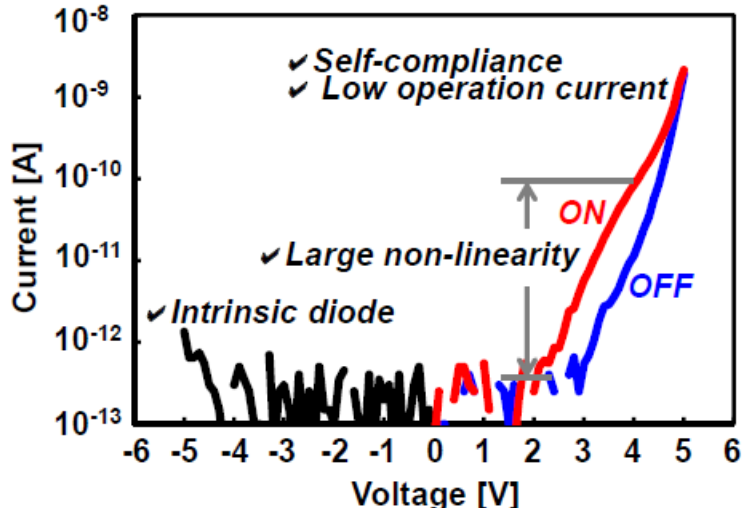
INCONVENIENTS

- Etape initiale de forming à plus haute tension pour créer le filament
- Variabilité importante de l'état HRS

Emerging Technologies : Fe-RAM



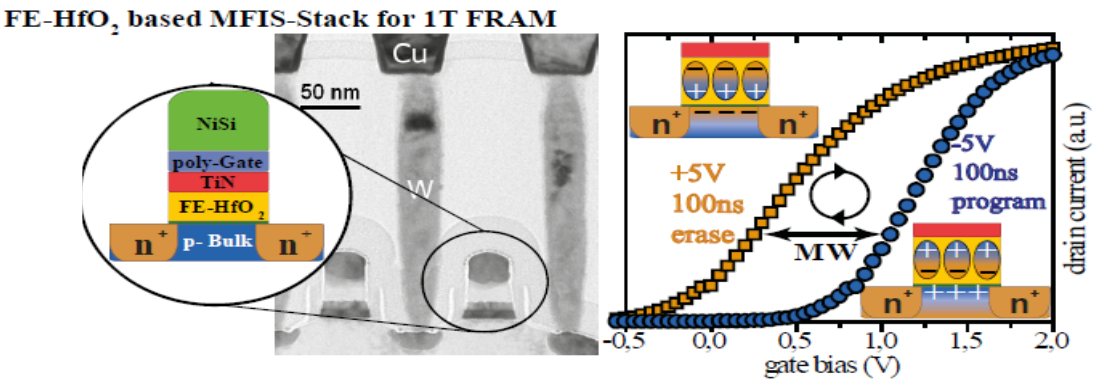




	HfO_2 FTJ
	This work
Operation current	~100nA
½ bias non-linearity	~300
Self-compliance	Yes
Intrinsic diode	Yes

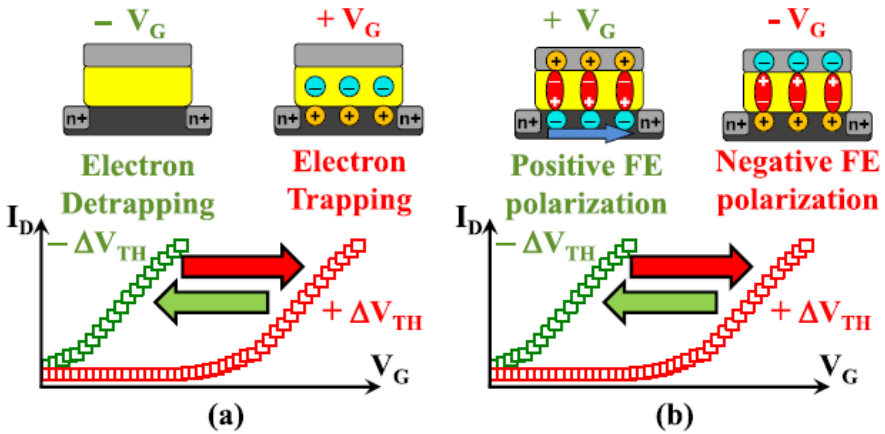
Shosuke F. et al., First demonstration and performance improvement of ferroelectric HfO2-based resistive switch with low operation current and intrinsic diode property, VLSI 2016

Emerging Technologies : Fe-RAM



Müller J et al. Ferroelectric hafnium oxide: a CMOS-compatible and highly scalable approach to future ferroelectric memories. IEDM 2013

Pas un effet de chargement/ déchargement dans l'oxyde de grille (variation contraire de la tension de seuil)



Yurchuk et al., Charge-Trapping Phenomena in HfO2-Based FeFET-Type Nonvolatile Memories, IEEE Transaction on Electron Devices Vol. 63, No. 9, sept. 2016

++: consommation, rapidité

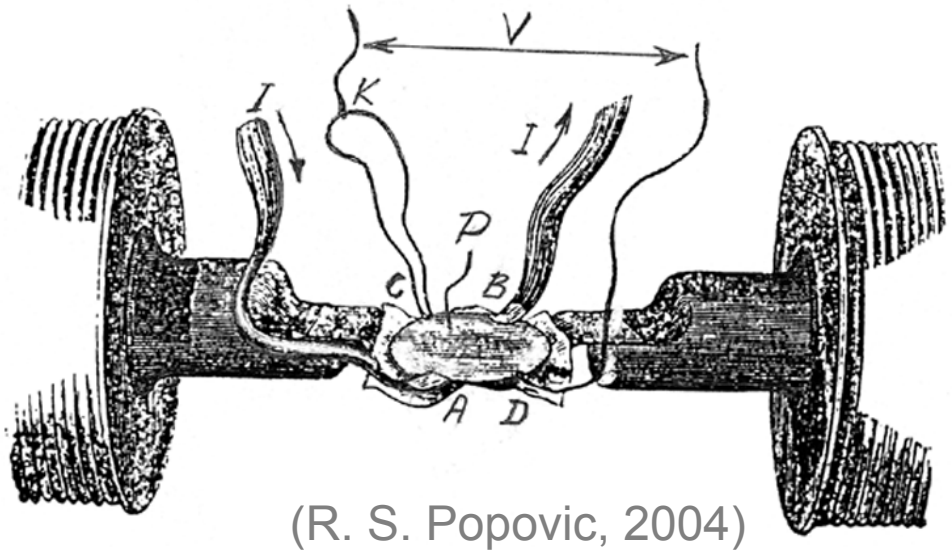
--: cyclage, rétention

- **Concept de mémoire non-volatile très récent, manque de maturité, démonstration d'intégration au nœud 28nm HKMG**
- **HfO2 ferroélectrique: utilisable aussi pour des transistors MOS à faible pente sous le seuil (concept de capacité négative de grille)**

3 – Emerging memory technologies : MRAM

William Thomson
1824-1907

- Conductance of magnetic metal plates is larger in the presence of a magnetic field perpendicular to the current flow



(R. S. Popovic, 2004)

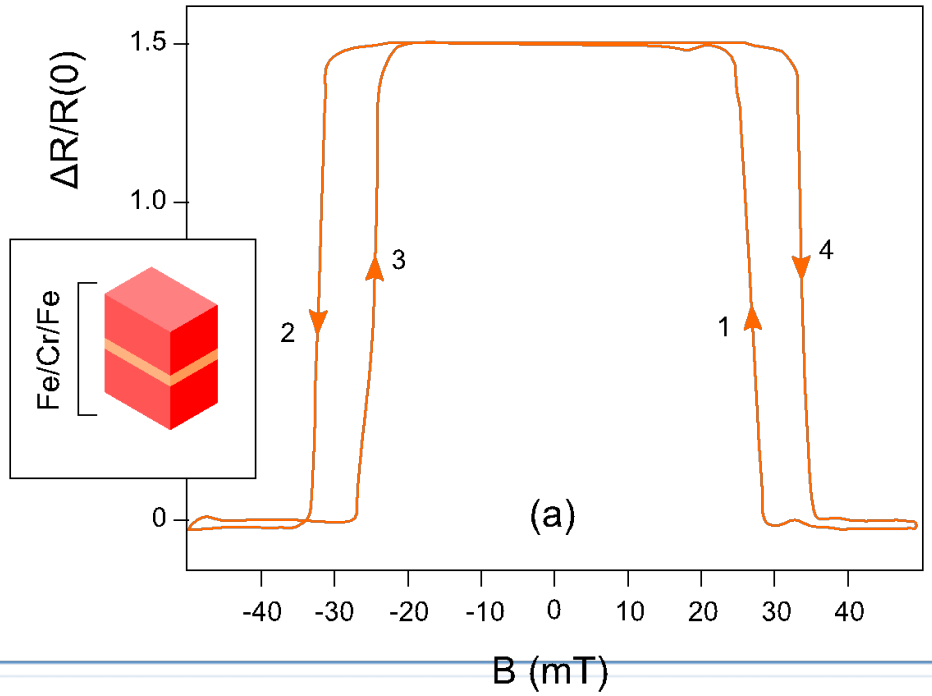
- Currently known as **Anisotropic Magnetoresistance (AMR)**
- Resistance variation attained: 2%-5% in RT

3 – Emerging memory technologies : MRAM

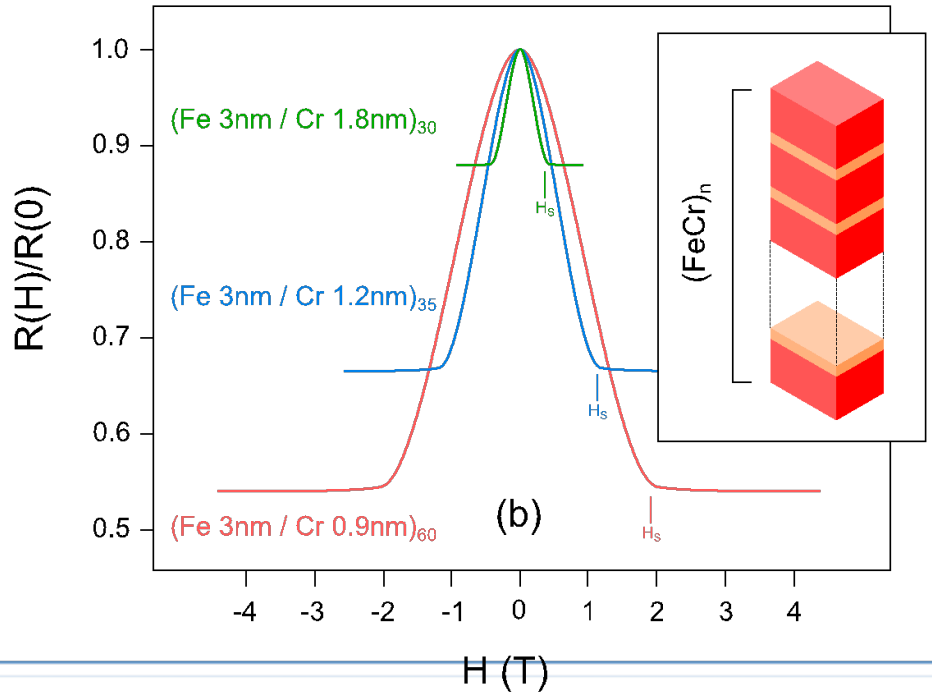
Peter Grünberg and Albert Fert
2007 Nobel Prize in Physics

- Thin stacks of FM/NM metals have seen a conductance increase of up to 100% when subjected to a magnetic field

B. Guinasch et al., 1989



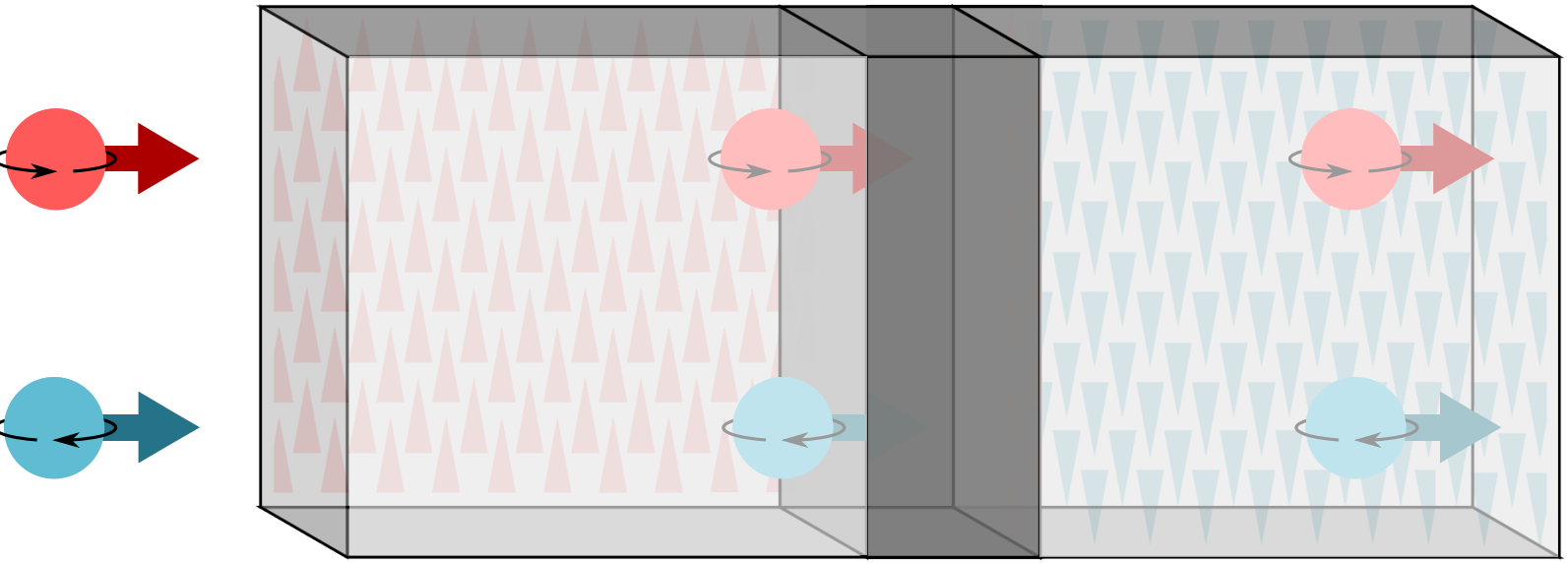
M. N. Baibich et al., 1988

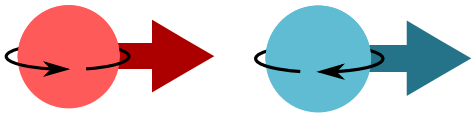


SPIN TECHNOLOGY OVERVIEW
GIANT MAGNETORISTANCE

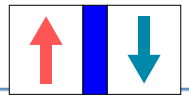
Peter Grünberg and Albert Fert
2007 Nobel Prize in Physics

- In FM/NM/FM structures, electrons are scattered as a result of interactions between the magnetic field and their spin





Spin-up Spin-down

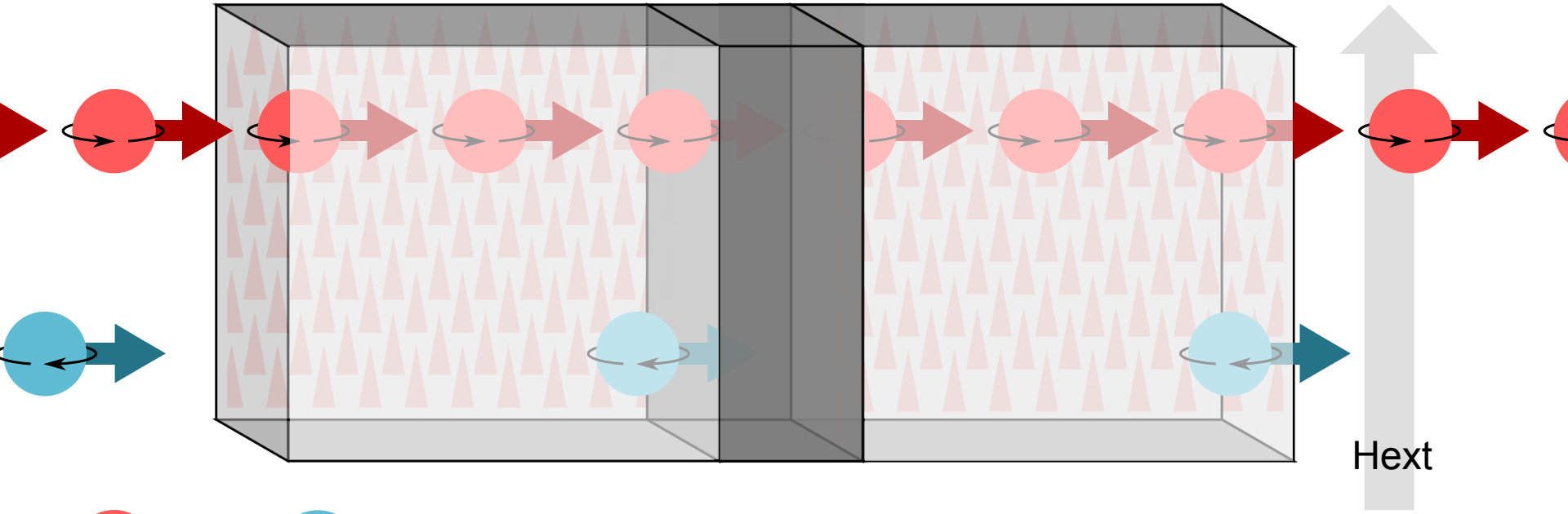


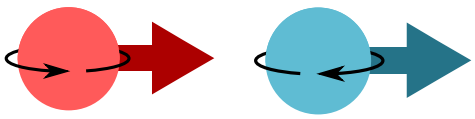
Anti-Parallel configuration

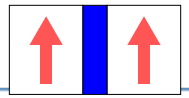
SPIN TECHNOLOGY OVERVIEW
GIANT MAGNETORISTANCE

Peter Grünberg and Albert Fert
2007 Nobel Prize in Physics

- In FM/NM/FM structures, electrons are scattered as a result of interactions between the magnetic field and their spin





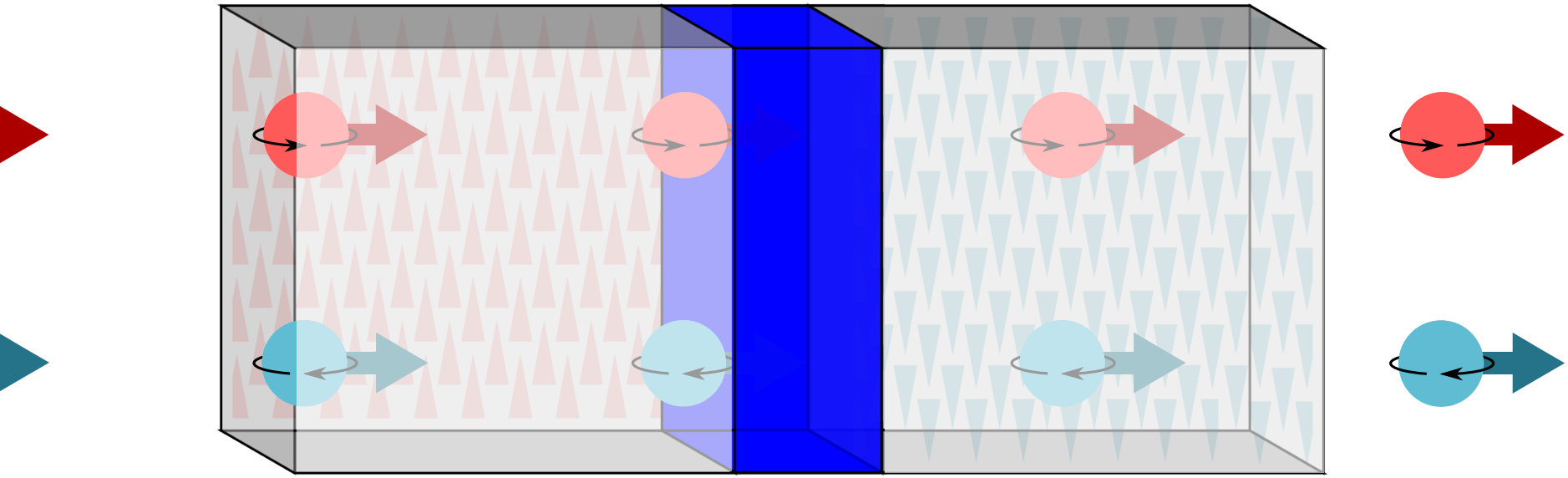


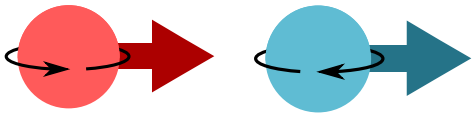
Parallel configuration

3 – Emerging memory technologies : MRAM

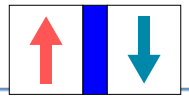
T. Miyazaki, J. Moodera, J. Slonczewski
(not in the pictures: M. Jullière)

- Spin-Dependent Transport (SDT): spin-up and spin-down have different probabilities of tunneling an FM/I/FM structure





Spin-up Spin-down

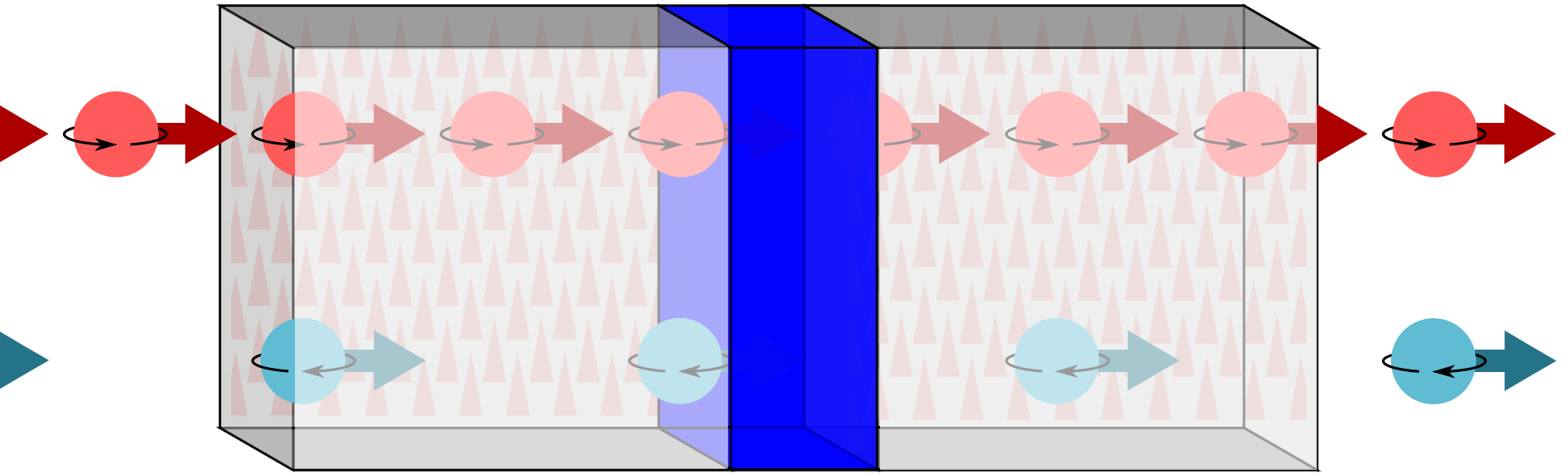


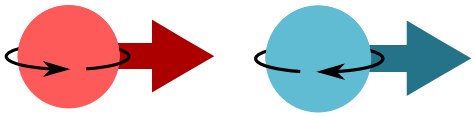
Anti-parallel configuration

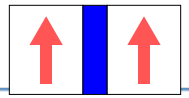
3 – Emerging memory technologies : MRAM

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

- Spin-Dependent Transport (SDT): spin-up and spin-down have different probabilities of tunneling an FM/I/FM structure



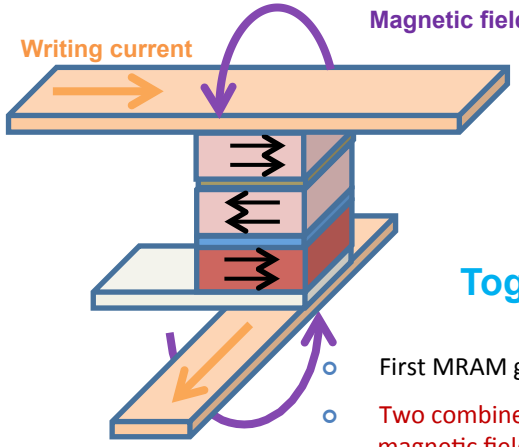




Parallel configuration

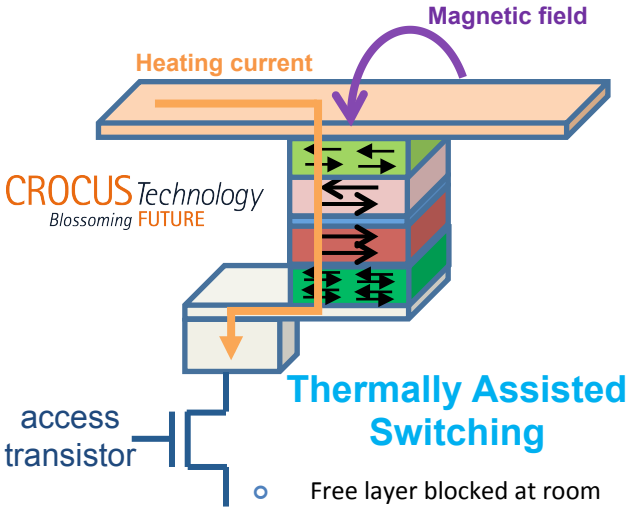
3 – Emerging memory technologies : MRAM

Field Induced Switching



Toggle

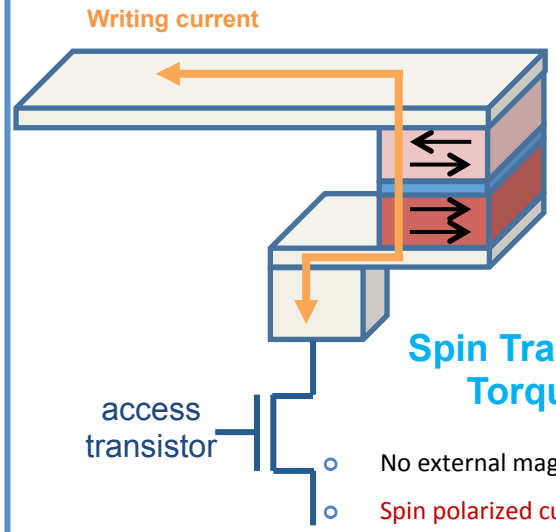
- First MRAM generation
- Two combined magnetic fields



Thermally Assisted Switching

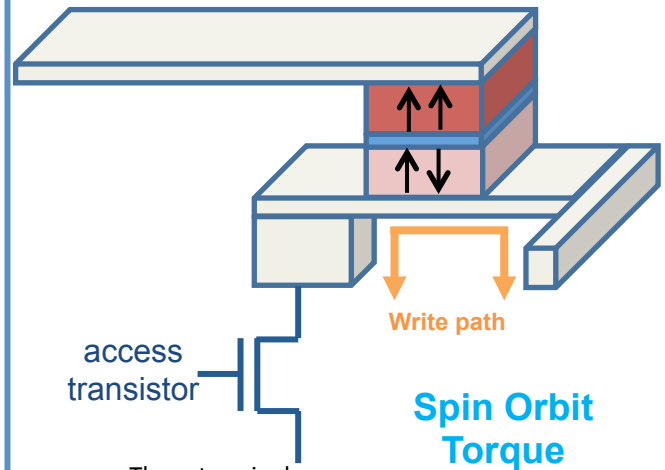
- Free layer blocked at room temperature
- Heat current + magnetic field

Current Induced Switching



Spin Transfer Torque

- No external magnetic field
- Spin polarized current

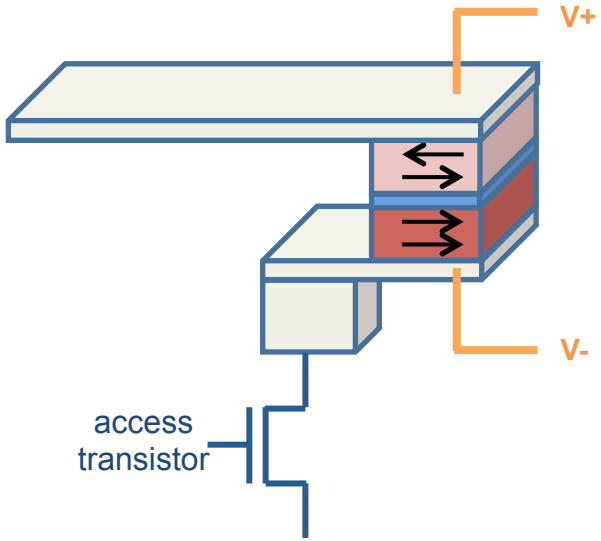


Spin Orbit Torque

- Three terminals
- SOT switching not fully understood (Rashba effect + Spin Hall Effect)

Voltage Induced Switching

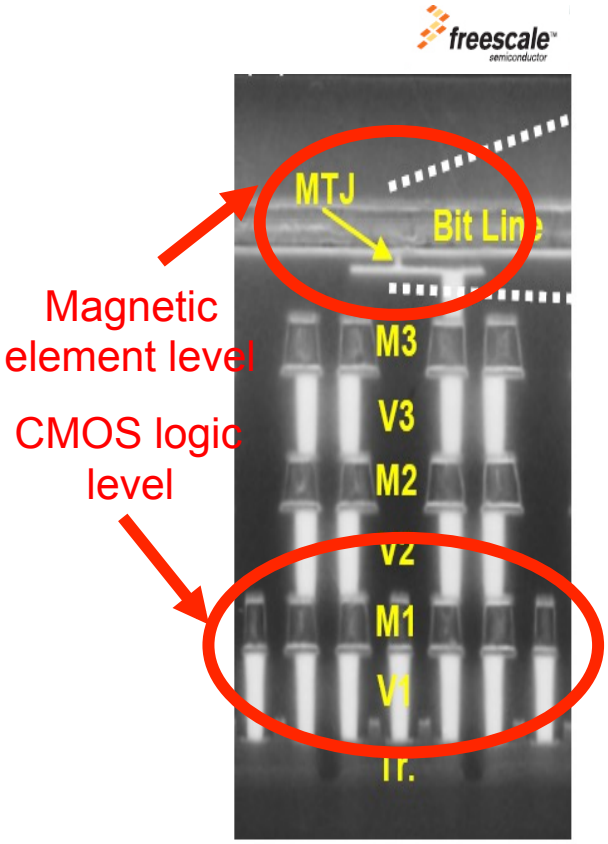
Magnetoelectric RAM (MeRAM)



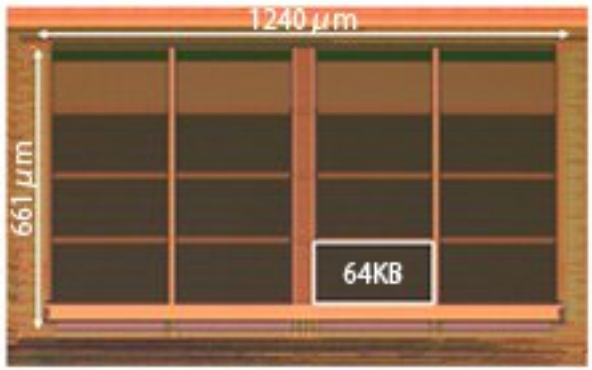
- Very low current
- Voltage-controlled magnetic anisotropy effect

3 – Emerging memory technologies : MRAM

- **Vitesse à l'écriture (~ns)**
- **Courant d'écriture ~5MA/cm² ⇔ 15uA pour 20nm, diminue avec la surface de la MTJ**
- **Endurance (>$10^{14\text{-}15}$)**
- **Intégration sur CMOS**
  - Tension d'écriture et résistance (~kΩ) compatible avec CMOS
  - ~3 masques supplémentaires
  - Température back-end faible (<350°C)
  - Matériaux magnétiques « exotiques »
- **Information non stockée sous forme de charge => immune aux radiations**
- **Compromis entre :**
  - Vitesse et consommation à l'écriture
  - Rétention et consommation à l'écriture
- **Scalabilité assurée jusqu'à 14 nm**
  - Stabilité proportionnelle au volume
  - Courants d'écriture très petits : risque d'écriture lors de la lecture

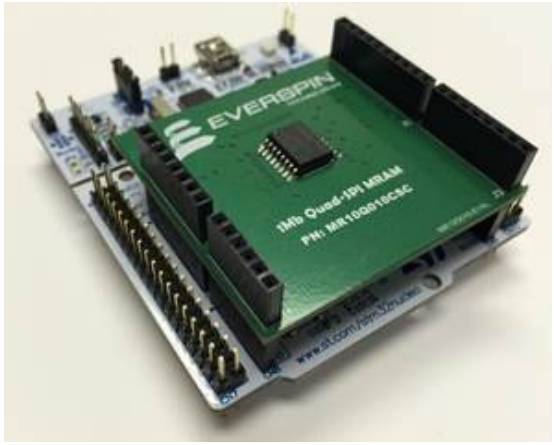


3 – Emerging memory technologies : MRAM

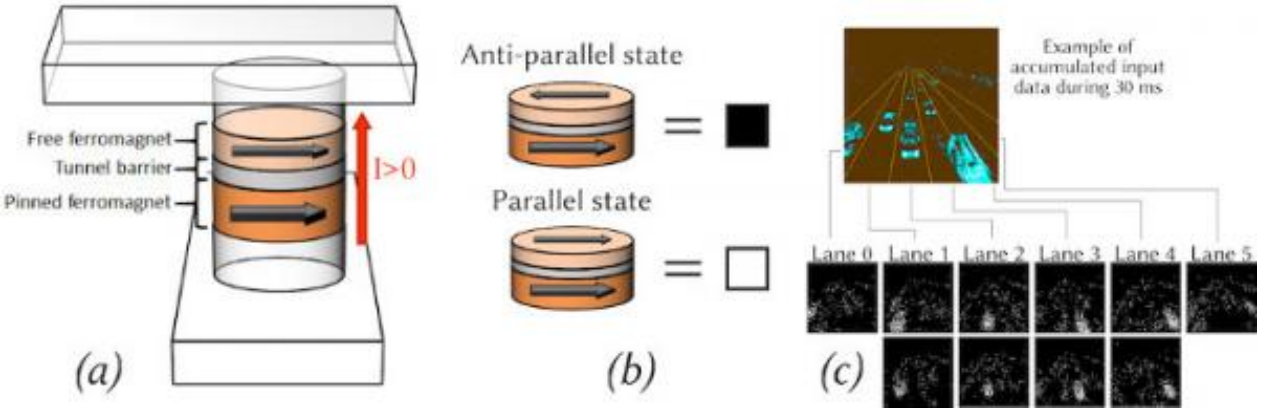


<STT-MRAM test chipt>

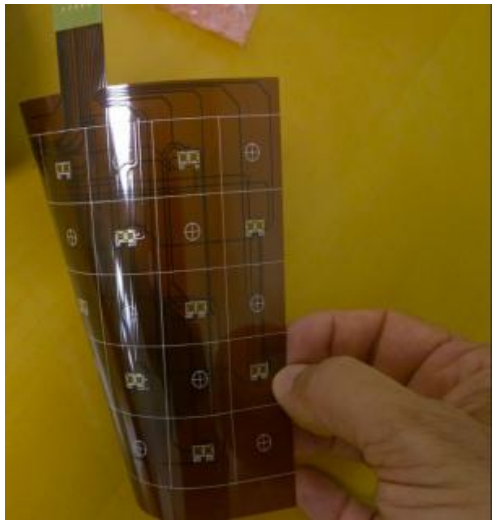
STT-MRAM Cache memory



Everspins MRAM Arduino Shield



MRAM based memresistor – Human Brain – IEF/CEA



Crocus MRAM flexible sensors

3 – Emerging memory technologies : Overview

- All these technologies are Non-volatile, based on resistance switching, with fast access time
- Two important aspects to consider
  - Technology maturity
  - Système integration (easy to replace actual memory)

Metricx	PCM	OxRAM	CBRAM	PSTT MRAM	FeFET	Flash
Endurance	4	4	2	5	1	3
Energy	2	3	3	4	5	1
Integration	3	5	4	2	5	3
Scalability	4	5	5	4	3	1
Retention	5	4	4	3	2	3
Speed	4	3	3	5	4	1
Maturité	3	3	2	2	1	5

Techno overview summary

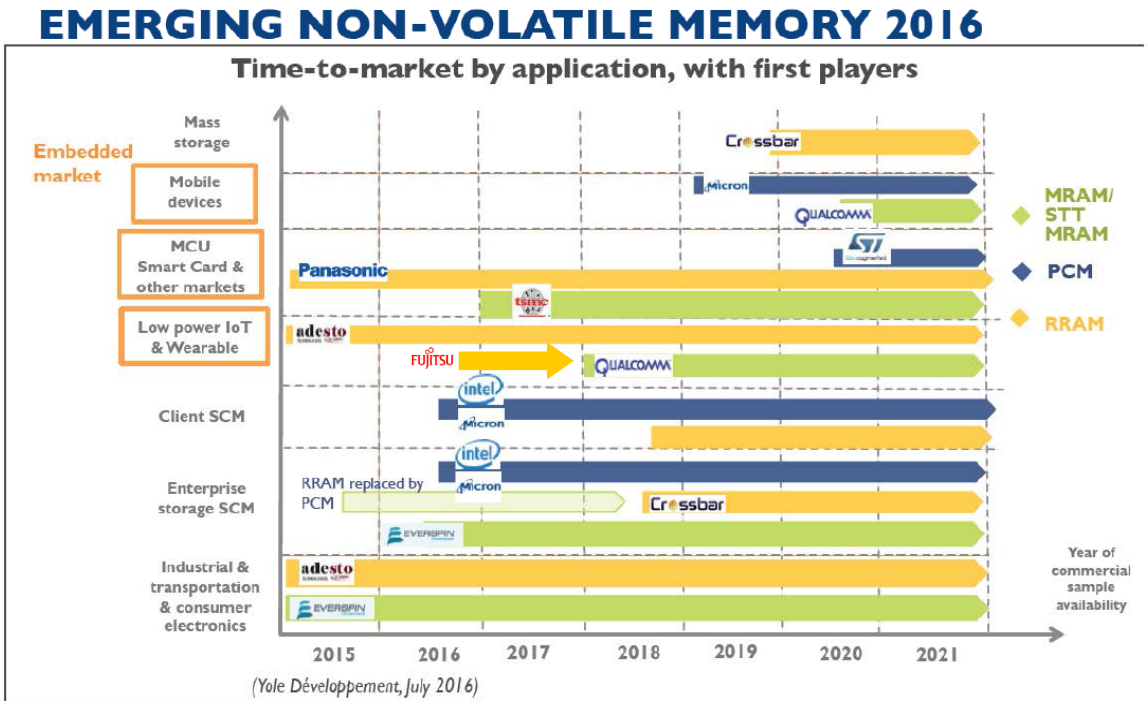
- **J. Joshua Yang, Dmitri B. Strukov & Duncan R. Stewart**

Nature Nanotechnology **8**, 13–24 (2013)

	Memristor	PCM	STTRAM	SRAM	DRAM	Flash (NAND)	HDD
		Prototypes		Commercialized technologies			
Reciprocal density (F^2)	<4	4–16	20–60	140	6–12	1–4†	2/3
Energy per bit (pJ)	0.1–3	2–25	0.1–2.5	0.0005	0.005	0.00002	1–10 × 10^9
Read time (ns)	<10	10–50	10–35	0.1–0.3	10	100,000	5–8 × 10^6
Write time (ns)	~10	50–500	10–90	0.1–0.3	10	100,000	5–8 × 10^6
Retention	years	years	years	As long as voltage applied	<<second	years	years
Endurance (cycles)	10^{12}	10^9	10^{15}	>10^{16}	>10^{16}	10^4	10^4

3 – Emerging memory technologies : Overview

- **Actual market of emerging technologies is about 50 Millions $ - 80 billions for DRAM and Flash**
- **2021: 4,6 billions $, a market growth of 110% per year**
- **All the majors semiconductors companies are leading this market – but lot of start-up too !**



Yole market report

Summary

1 – Context and objectives of the lecture

2 – Classical technologies and memory architecture overview (SRAM, DRAM, FLASH)

3 – Emerging memory technologies

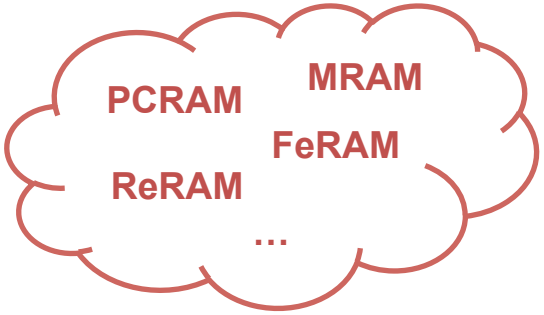
4 – Computing with Non-Volatile memory technologies

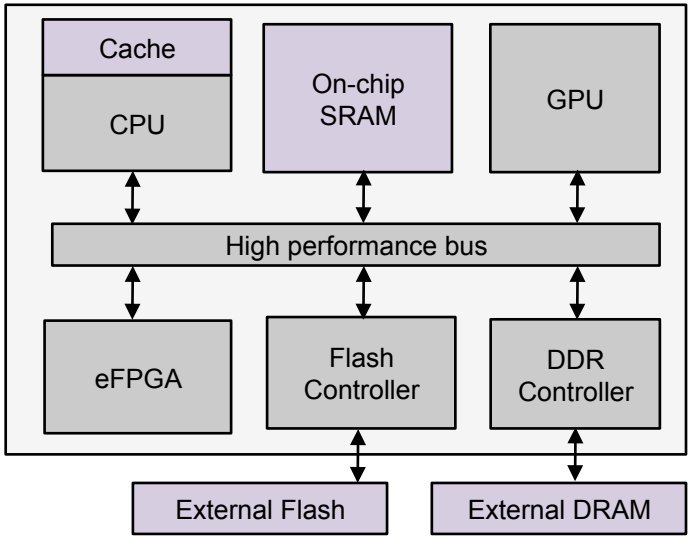
- For high performance computing applications
- For Embedded applications (Non-volatile processor)
- For secure applications

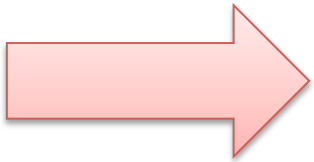
5 - Conclusions

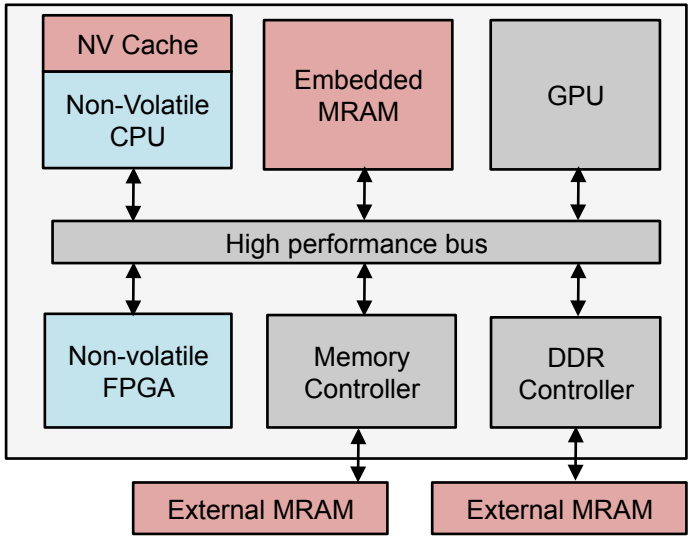
Motivation

- **Solution**
  - Go towards non-volatile systems using emerging NVMs
  - Current NVMs issues : Speed, Dynamic energy, Reliability









Where and how to place MRAM to:

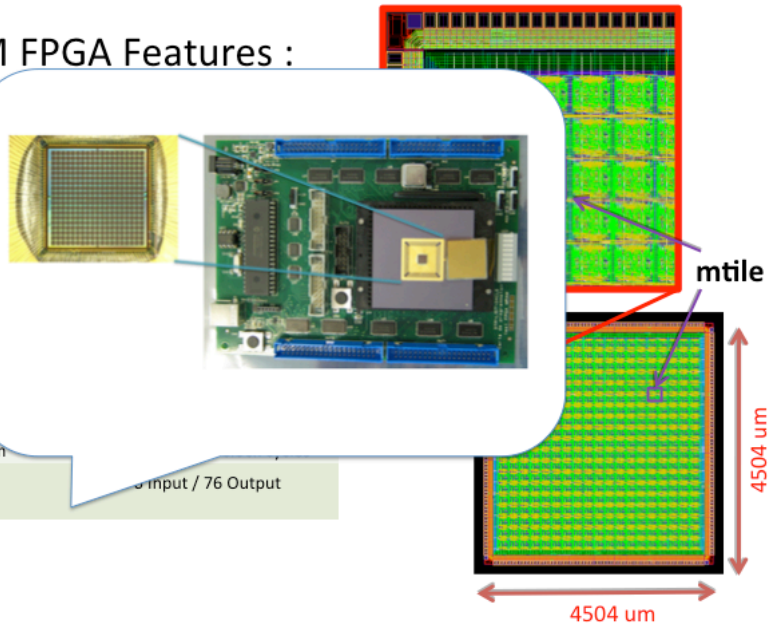
reduce total power consumption ?

keep same or get better performance ?

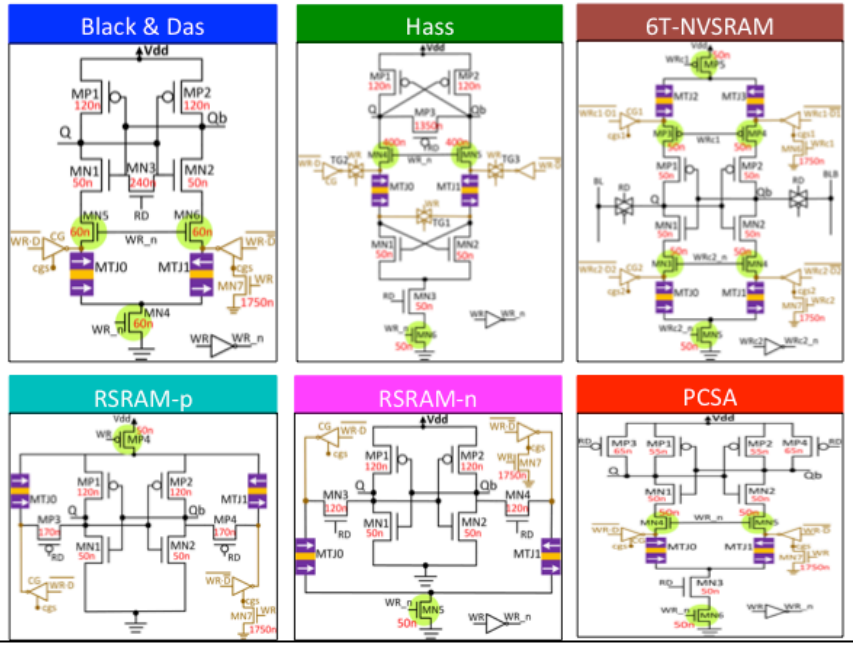
Hybrid CMOS/MRAM Blocks for FPGAs

- MRAM FPGA Features :

# LUT 4	
# TILES	
# Sequential el	
# of MTJs	
# of Transistors	
Silicon Area	
MRAM Reconfi Energy	
MRAM Restora	
Clock Frequenc	
Full configuratio	
Tile reconfiguration	
# Input/Output	Input / 76 Output



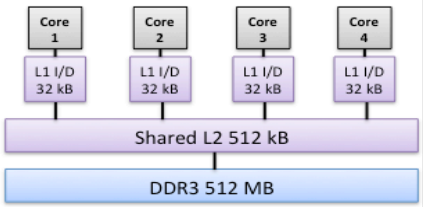
Self Referenced cells



MRAM-based cache

- MRAM technologies evaluated:
  - 130 nm TAS (L2)
  - 45 nm STT (L1/L2)

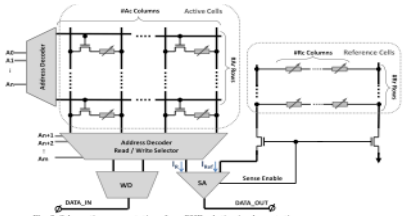
- Multi-core architecture:
  - ARMv7 ISA
  - Private L1 I/D
  - Shared L2



4- What about security …

Physically Unclonable Function

PUF solution exploits the differential sensing during read operation, based on read current comparison against a reference value.



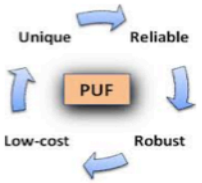


Fig. 7. Schematic representation of our PUF solution implementation

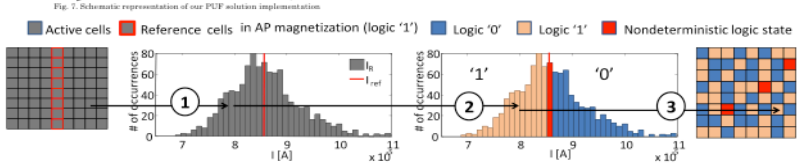


Fig. 4. The implementation strategy of the proposed PUF solution: 1) Write all cells to '1'; 2) Read each cell; 3) Use the read value

Elena Ioana Vatajelu, Giorgio Di Natale, Mario Barbareschi, Lionel Torres, Marco Indaco, Paolo Prinetto **STT-MRAM-Based PUF Architecture exploiting Magnetic Tunnel Junction Fabrication-Induced Variability**, ACM JETC, 2015

Contributions

1. Evaluation of MRAM-based cache memory hierarchy:
   - Exploration flow and extraction of memory activity
   - L1 and L2 caches based on STT-MRAM and TAS-MRAM

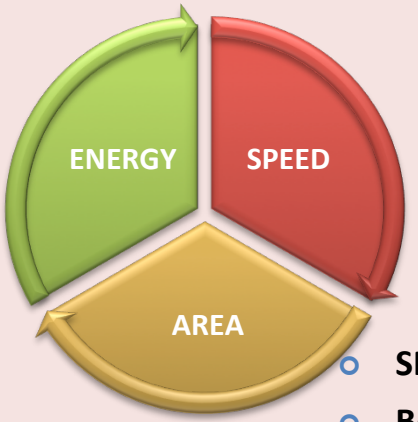
2. Non-volatile computing
   - *Instant-on/off* capability for embedded processor
   - Analysis and validation of *Rollback* mechanism

3. Secure applications with NVM

MRAM applied to cache

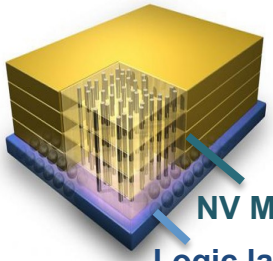
- Possible studies

Performance comparison

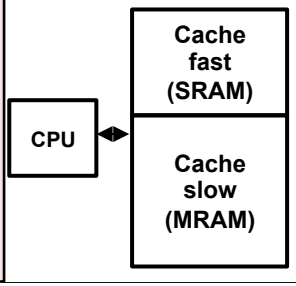


- **SRAM vs MRAM**
- **Benefits of MRAM**
  - Low leakage
  - High density
  - Non-volatility

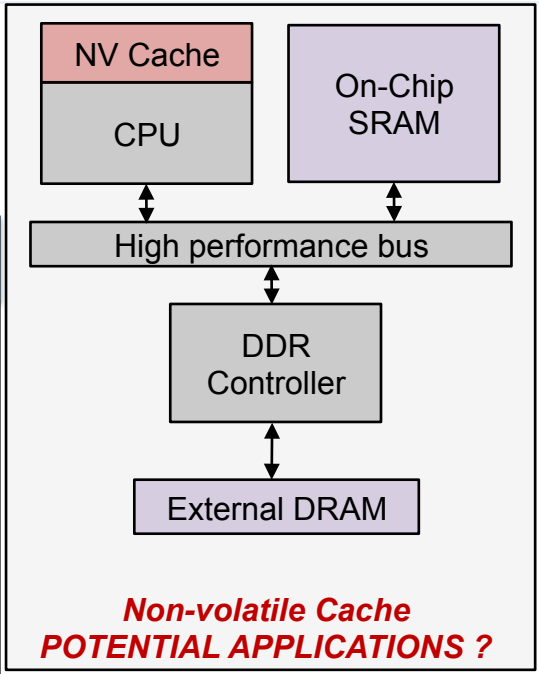
New architectures



3D-Stacking capability of MRAM



Hybrid SRAM/ MRAM cache



Non-volatile Cache POTENTIAL APPLICATIONS ?

Take advantages of MRAM

Low leakage
High density
Non-volatility

Mitigate drawbacks of MRAM

write latency
write energy

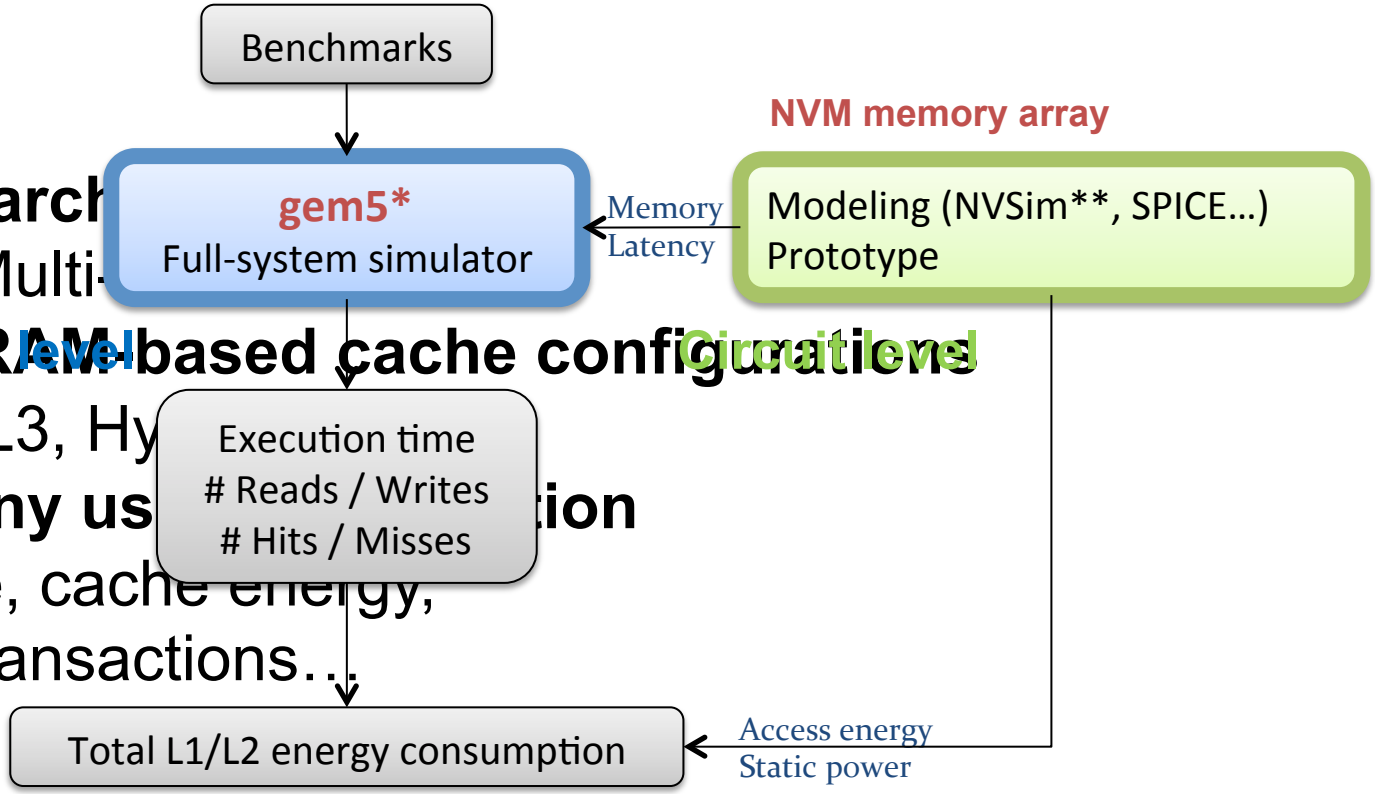
MRAM applied to cache

NVM exploration flow

1. **Define the arch**
   - Single/Multi-
2. **Explore MRAM-based cache configuration**
   - L1, L2, L3, Hy
3. **Extract many us tion**
   - Runtime, cache energy, cache transactions…

Architecture level

Circuit level



* N. Binkert et al., "**The gem5 simulator**," ACM SIGARCH Computer Architecture News, Aug. 2011.

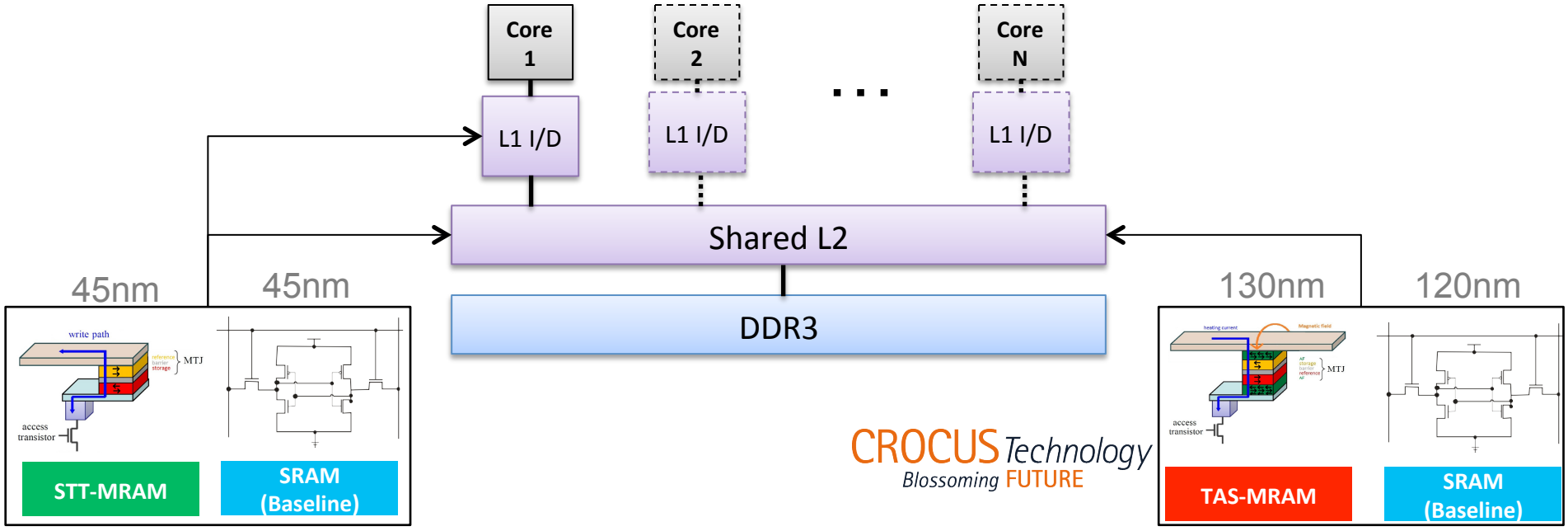
** X. Dong et al., "**NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory**," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Jul. 2012 .

MRAM applied to cache

Experimental setup

From single to multi-core architecture

ARMv7 ISA
Private L1 instruction/Data
Shared L2
(Additional levels of caches possible)
Main Memory

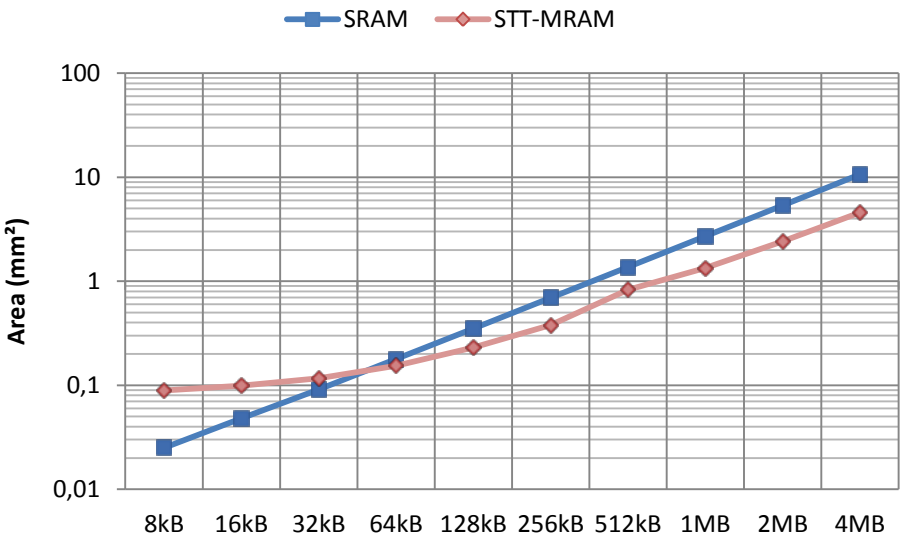


CROCUS Technology
Blossoming FUTURE

MRAM applied to cache

Circuit-level analysis: Models (NVSim) & Prototype

Area



Node	Technology	512kB L2 (mm²)	32kB L1 (mm²)
45nm	SRAM	1.36	0.091
	STT-MRAM	0.82	0.117
120nm	SRAM	9.7	-
	TAS-MRAM	11.7	

- MRAM is denser for large cache capacity
- MRAM cell size smaller than that of SRAM
- MRAM needs large transistors for write
- TAS-MRAM cache larger due to field lines

MRAM applied to cache

Circuit-level analysis: Models (NVSim) & Prototype

512kB L2 cache

		Read		Write		Standby
Node	Technology	Latency (ns)	Energy (nJ)	Latency (ns)	Energy (nJ)	Leakage (mW)
45nm	SRAM	4.28	0.27	2.87	0.02	320
	STT-MRAM	2.61	0.28	6.25	0.05	23
120nm	SRAM	5.95	1.05	4.14	0.08	82
	TAS-MRAM	35	1.96	35	4.62	10

/2.2 ≈ (Read) · 2.1 2.5 (Write) · /14 (Standby) · /8 (Standby)

STT-MRAM ≈ SRAM
TAS-MRAM > SRAM

MRAM > SRAM

MRAM << SRAM

32kB L1 cache

Node	Technology	Latency (ns)	Energy (nJ)	Latency (ns)	Energy (nJ)	Leakage (mW)
45nm	SRAM	1.25	0.024	1.05	0.006	22
	STT-MRAM	1.94	0.095	5.94	0.04	3.3

/7 (Leakage)

MRAM > SRAM

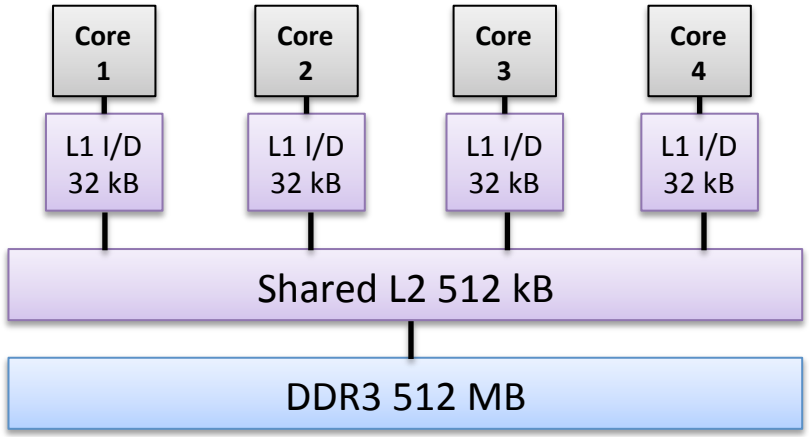
MRAM > SRAM

MRAM << SRAM

MRAM applied to cache

Case study

- Quad-core architecture:
  - Frequency 1GHz
  - ARMv7 ISA
  - Private L1 I/D
  - Shared L2
  - DDR3 Main memory

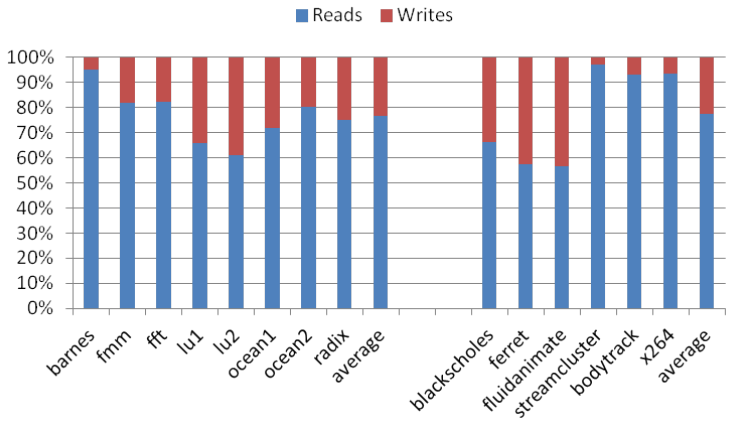


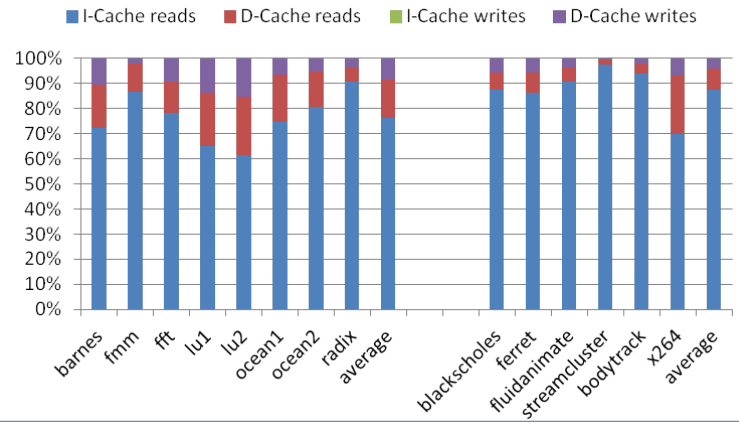
- Benchmarks
  - SPLASH-2
    - Mostly high performance computing
  - PARSEC
    - Animation, data mining, computer vision, media processing

MRAM applied to cache

Architecture-level analysis: gem5

Read/Write ratio





L2/L1 access ratio

Benchmark	Number of accesses	
	L1 cache	L2 cache
SPLASH-2	~2 billions (0.5 billions/CPU)	~26 millions
PARSEC	~12 billions (3 billions /CPU)	~16 millions

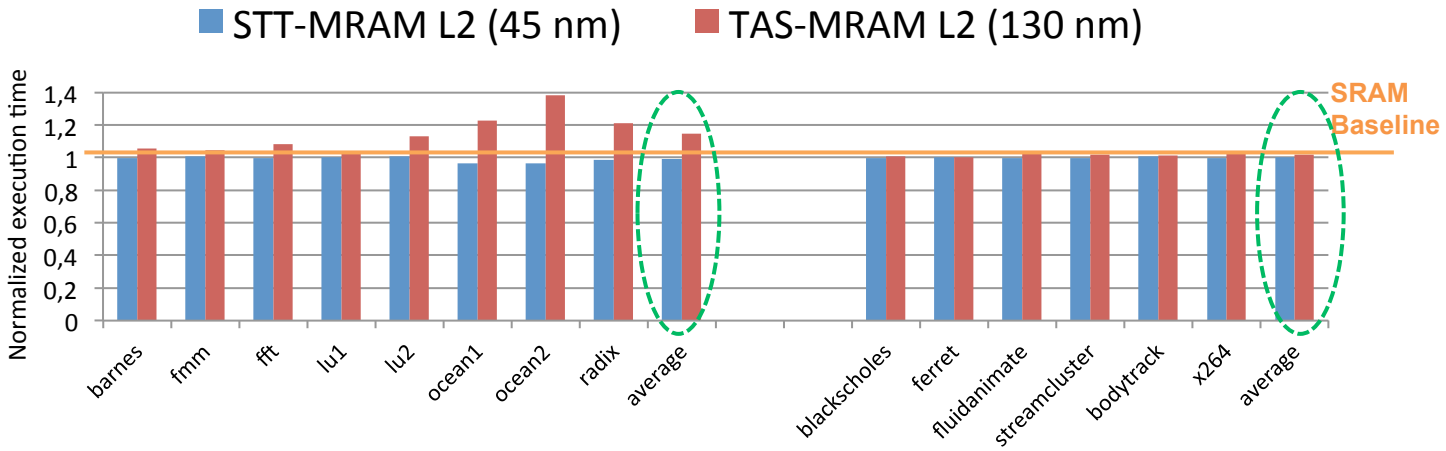
Static/Dynamic energy ratio

Static energy

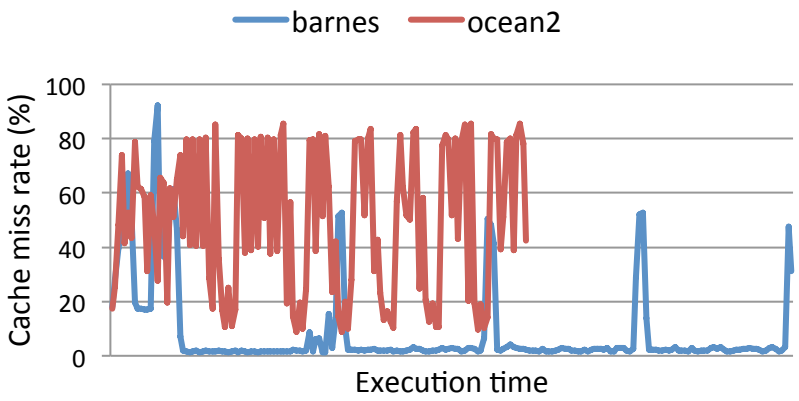
L2 → 90%
L1 → 80%

MRAM-based L2

Execution time

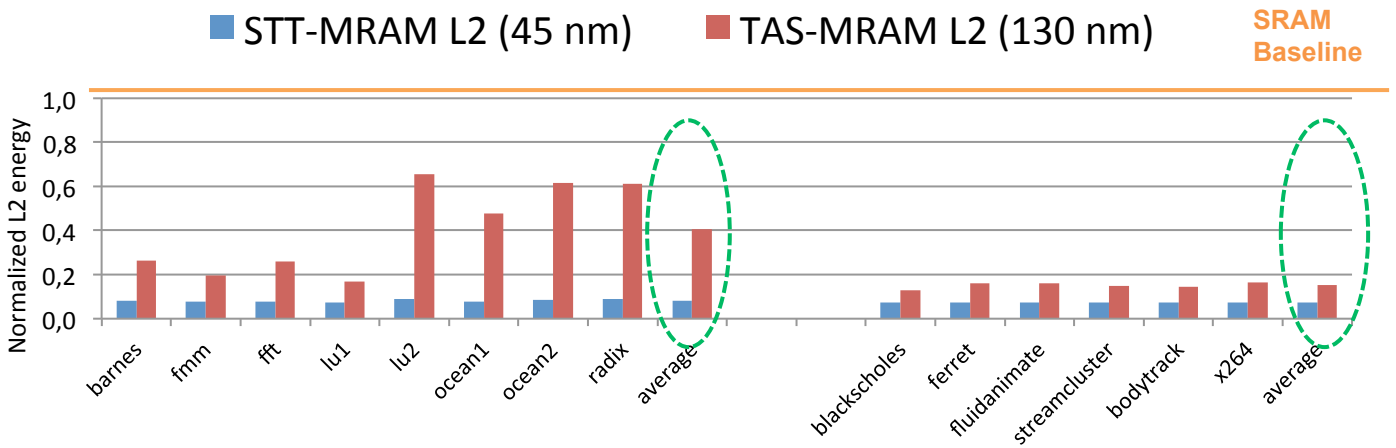


- Observations:
  - STT shows good performance
    - L2 has small impact in overall performance
  - For TAS, 14% of penalty in average (SPLASH-2)
    - Depends on applications
      (Cache miss rate, L1/L2 access ratio)

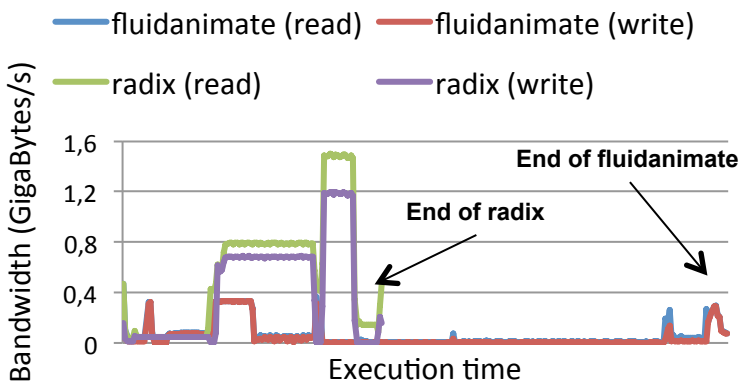


MRAM-based L2

Total L2 cache energy consumption



- Observations:
  - Up to 90% of gain for STT
  - From 40% to 90% for TAS
    - Due to the very low leakage of MRAM-based cache



MRAM-based cache

Summary

- **NVM exploration flow available**
  - Input from models or silicon chip
  - Memory activity analysis
- **Is MRAM suitable for cache ?**
  - Good candidate for lower level of cache (L2 or last level cache)
    - Up to 90% of energy gain
    - No or small performance penalty
    - More memory capacity using MRAM
    - Cache L2 is up of 20% energy consumption of overall system
  - Not suitable for upper level of cache (L1) for high performance – but depending of the application some gain in energy
    - Micro-architectural modifications required to mask latency
    - Not detailed in this presentation but full evaluation of cache L1 done too

Contributions

1. Evaluation of MRAM-based cache memory hierarchy:
   - Exploration flow and extraction of memory activity
   - L1 and L2 caches based on STT-MRAM and TAS-MRAM

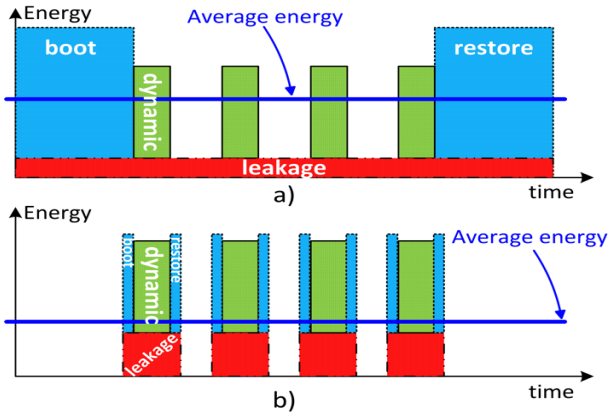
2. Non-volatile computing
   - *Instant-on/off* capability for embedded processor
   - Analysis and validation of *Rollback* mechanism

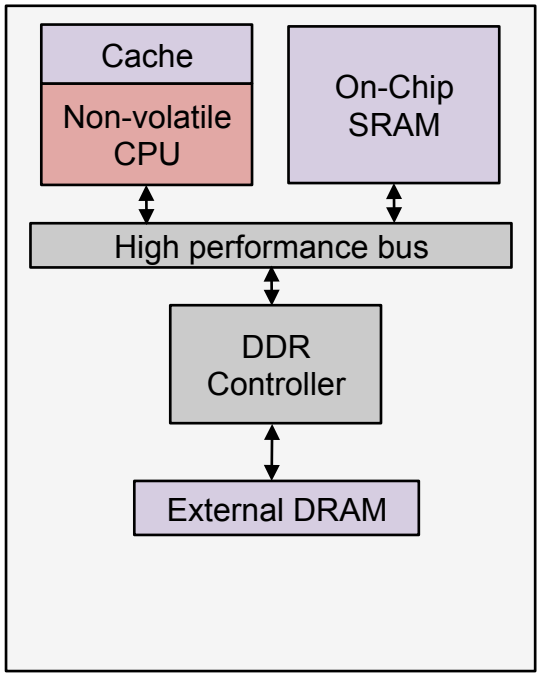
3. Secure applications with NVM

MRAM-based processor

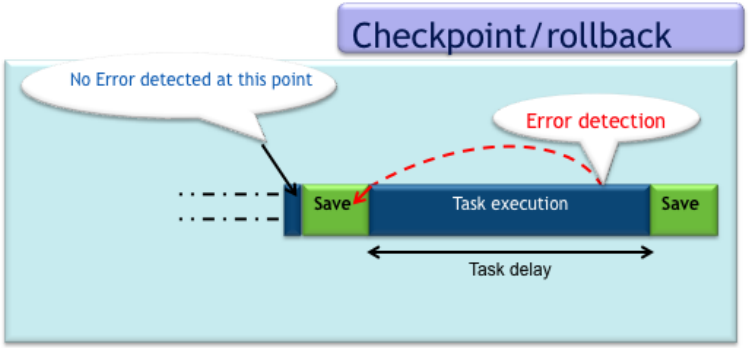
Normally-off computing

- Two concepts:
  - Instant on/off
    - Restore processor state

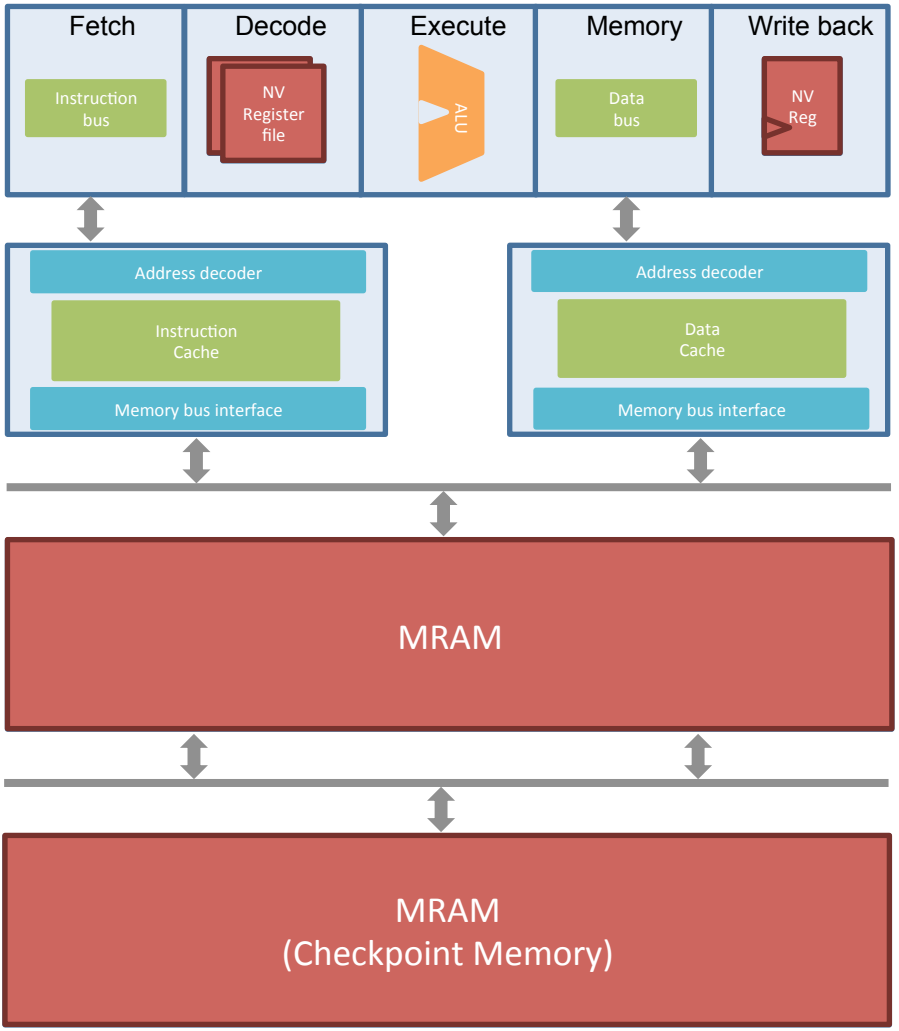




  - Backward error recovery (Rollback)
    - Restore previous valid state



MRAM-based processor



Instant on/off & Rollback

Non-volatile register

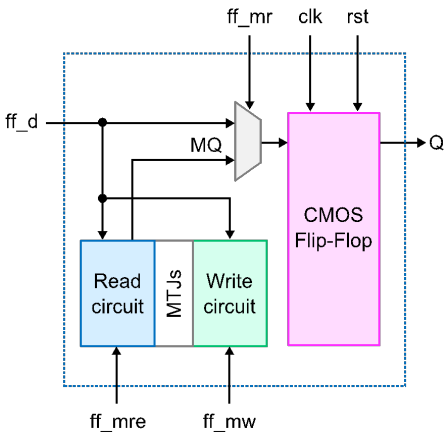
+

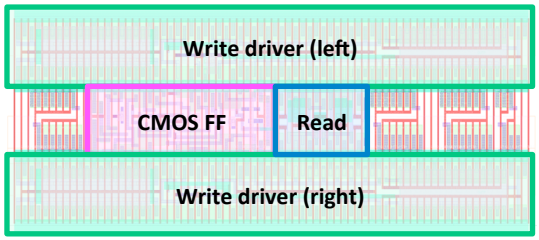
Non-volatile Memory

+

Checkpoint Memory (Rollback)

MRAM-based register



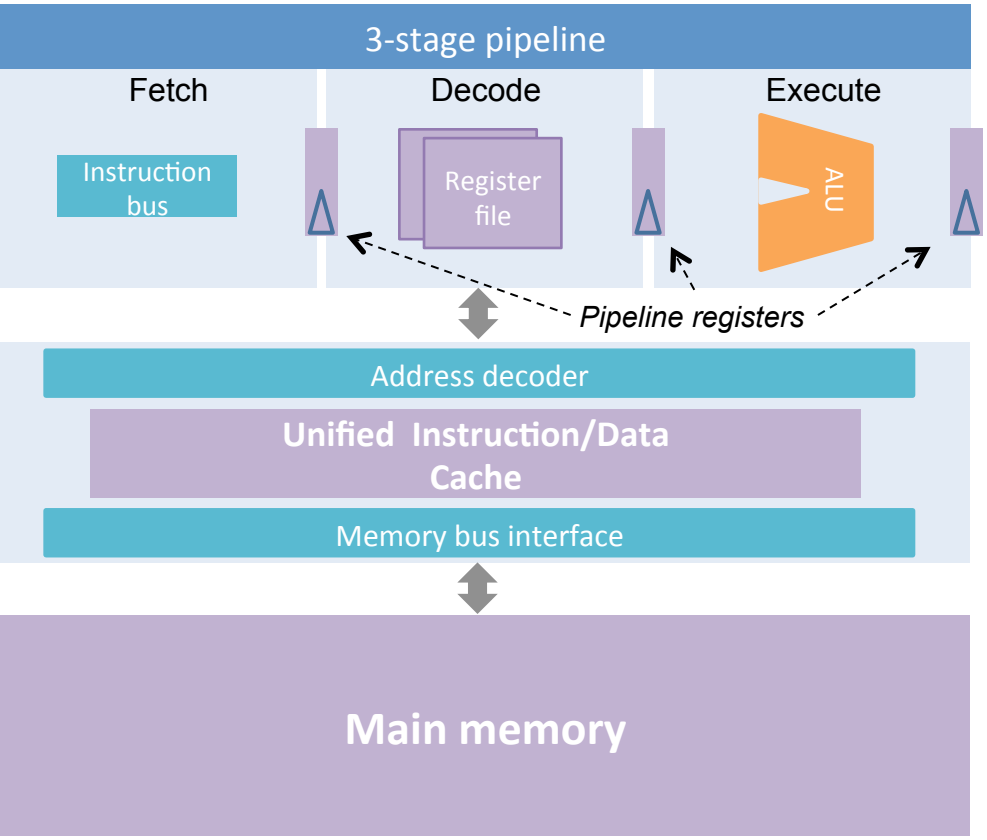


Layout of NV Flip-Flop
(28nm FDSOI, 90nm STT)

B. Jovanovic, R. Brum, L. Torres, *Comparative Analysis of MTJ/CMOS Hybrid Cells based on TAS and In-plane STT Magnetic Tunnel Junctions*, **IEEE Transactions on Magnetic**,, 2014.

MRAM-based processor

Case study: Amber 23 processor (ARM based instruction)



FEATURES

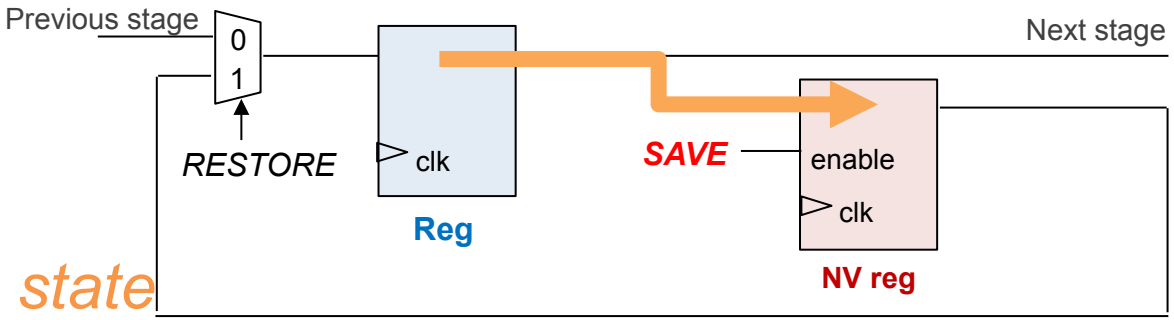
- 3-stage pipeline
- 16x32-bit register file
- 32-bit wishbone system bus
- Unified instruction/data cache (16 kBytes)
  - Write through
  - Read-miss replacement policy
- Main memory (> Mbytes)
- Multiply and multiply-accumulate operations

➢ Implementation of both instant-on/off and rollback (Verilog code modified)

➢ Duplication of the registers to emulate the non-volatility

Instant on/off

Instant on/off



1. *Save the register's state*

2. *POWER DOWN*

Main memory based on MRAM

↓

Data preserved

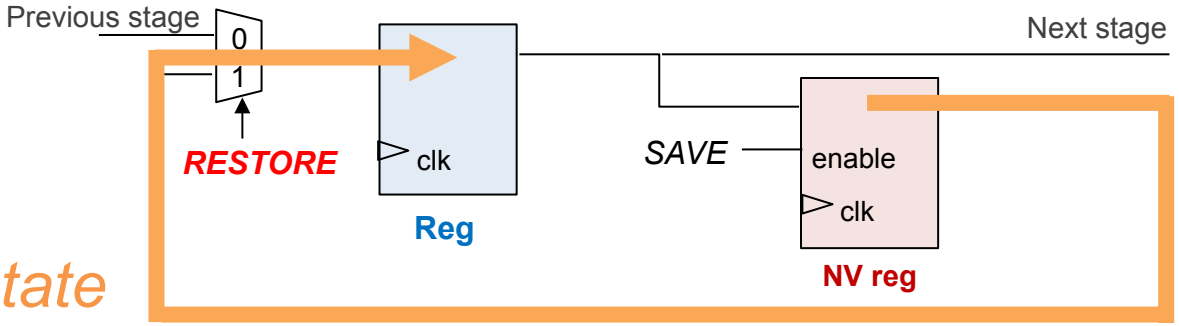
3. *POWER UP*

Main memory based on MRAM

↓

Data available

4. *Restore the register's state*



Instant on/off

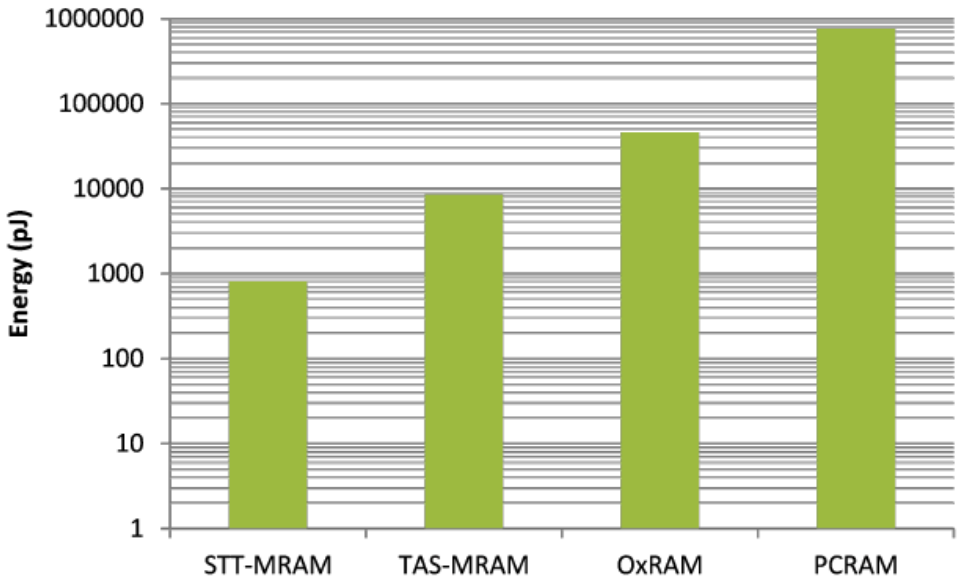
Instant-on/off: backup energy

Non-volatile flip-flops performance

Technology	Latency (ns)		Energy (pJ)	
	Restore	Back-up	Restore	Back-up
STT-MRAM [Chabi et al. 2014]	0.2	4	0.012	0.5
TAS-MRAM [Jovanovic et al. 2015]	0.13	16	0.012	5.2
OxRAM [Jovanović et al. 2014]	6	70	1.4	28
PCRAM [Choi et al. 2013]	370	370	7.4	463

- 1644 Flip-Flops saved
- Flip-Flops are backed-up in parallel

- Backup energy:
  → less than 1nJ for STT-MRAM
  → less than 10nJ for TAS-MRAM
- [1] The required current to erase and program flash can vary from 4 to 12 mA



[1] “Benchmarking mcu power consumption for ultra-low-power applications,” White paper, Texas Instruments

Instant on/off

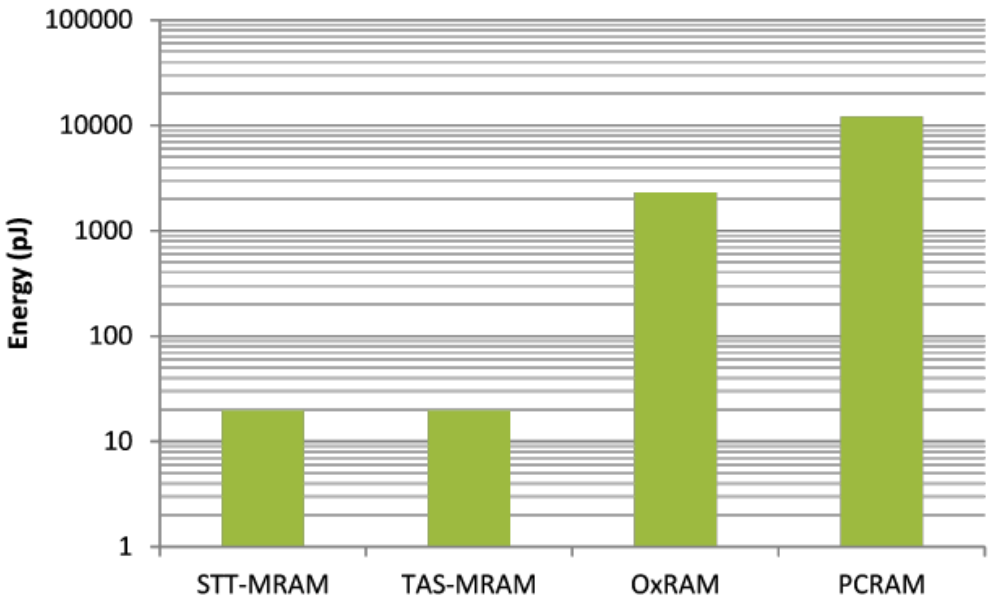
Instant-on/off: Restore energy

Non-volatile flip-flops performance

Technology	Latency (ns)		Energy (pJ)	
	Restore	Back-up	Restore	Back-up
STT-MRAM [Chabi et al. 2014]	0.2	4	0.012	0.5
TAS-MRAM [Jovanovic et al. 2015]	0.13	16	0.012	5.2
OxRAM [Jovanović et al. 2014]	6	70	1.4	28
PCRAM [Choi et al. 2013]	370	370	7.4	463

- 1644 Flip-Flops restored
- Flip-Flops are restored in parallel

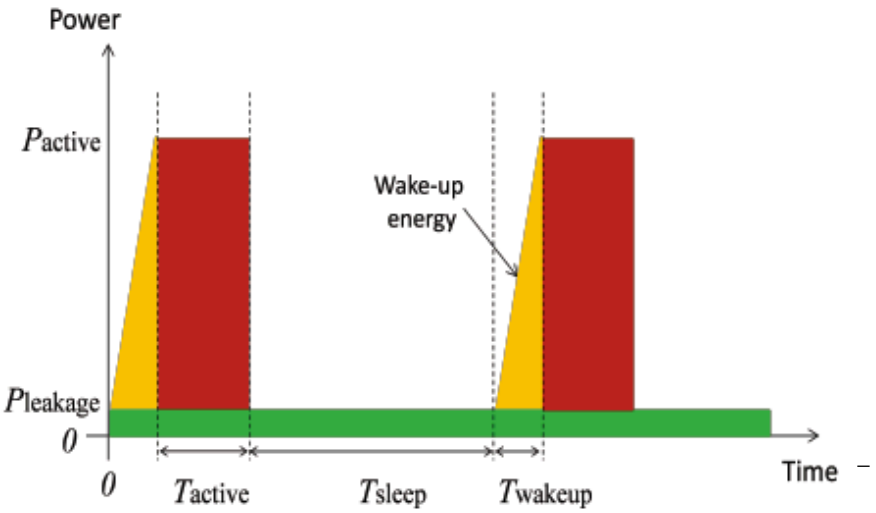
- Restore energy:
  → 20pJ for both STT-MRAM and TAS-MRAM
- [1] Wang et al. showed that the energy consumption to restore 1607 Flip-Flops from off-chip flash (on-chip flash) is 1.3µJ (0.6µJ)



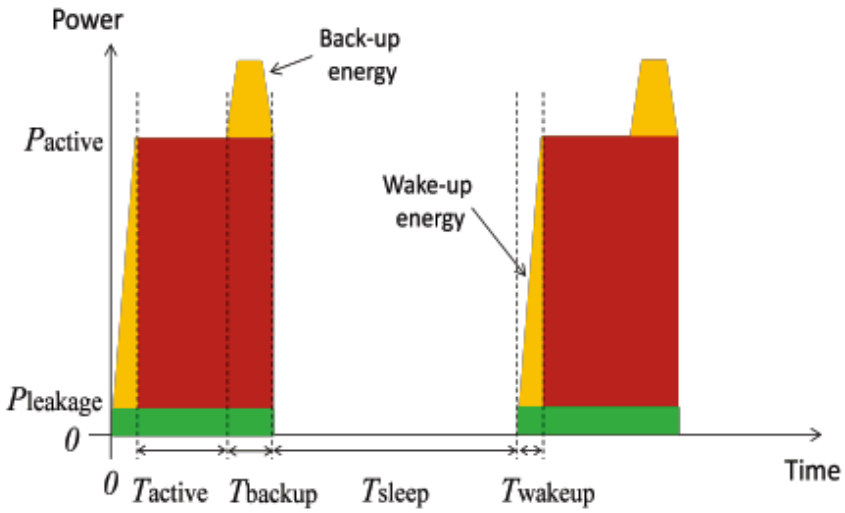
[1] “A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops,” in ESSCIRC (ESSCIRC), Proceedings of the. IEEE, 2012

Instant on/off

Instant-on/off: sleep mode



Without instant-on/off



With instant-on/off

$$(P_{active} + P_{leakage}) \times T_{backup} + E_{backup} < P_{leakage} \times T_{sleep}$$

Minimum $Tsleep$ required to be more energy efficient

$$T_{sleep} > \frac{(P_{active} + P_{leakage}) \times T_{backup} + E_{backup}}{P_{leakage}}$$

Instant on/off

Instant-on/off: sleep mode

Switching activity → 0.5/cycle

Synthesis of the Amber 23
(65nm CMOS low-power HVT process)

P_{active} = 173 mW (40 MHz)
$P_{leakage}$ = 12 mW

Technology	Minimum *Tsleep*
STT-MRAM	130 ns
TAS-MRAM	968 ns
OxRAM	4.9 µs
PCRAM	69 µs

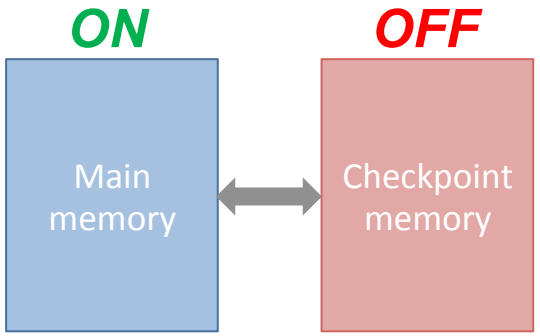
- Not considering the power down/up circuitry
- Cache warm-up penalty to consider
- Area overhead to consider

Rollback

Rollback

NORMAL EXECUTION

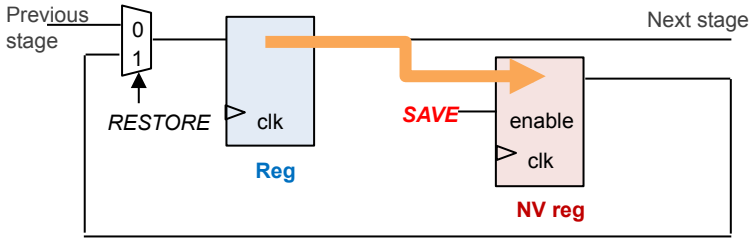
- ***Only the main memory contents are modified***
- ***The checkpoint memory is powered off***



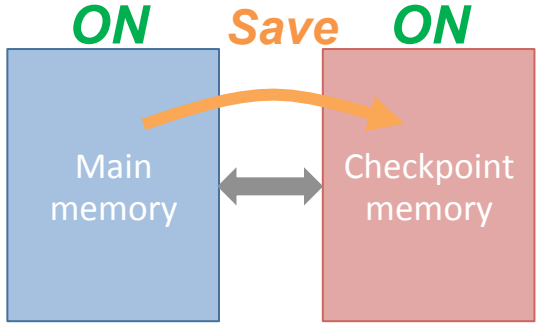
CHECKPOINT

- ***Save registers***
- ***Save memory***

=

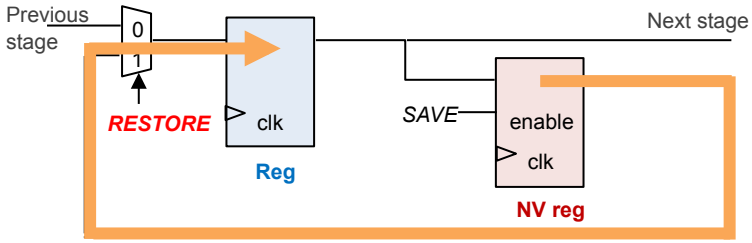


+

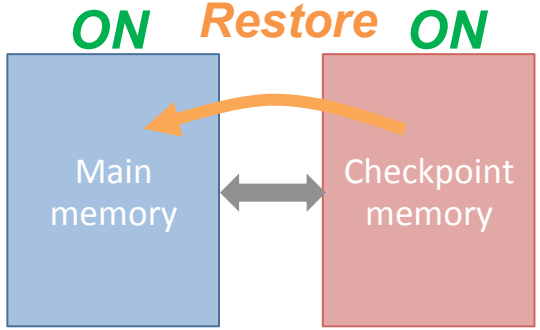


ROLLBACK

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

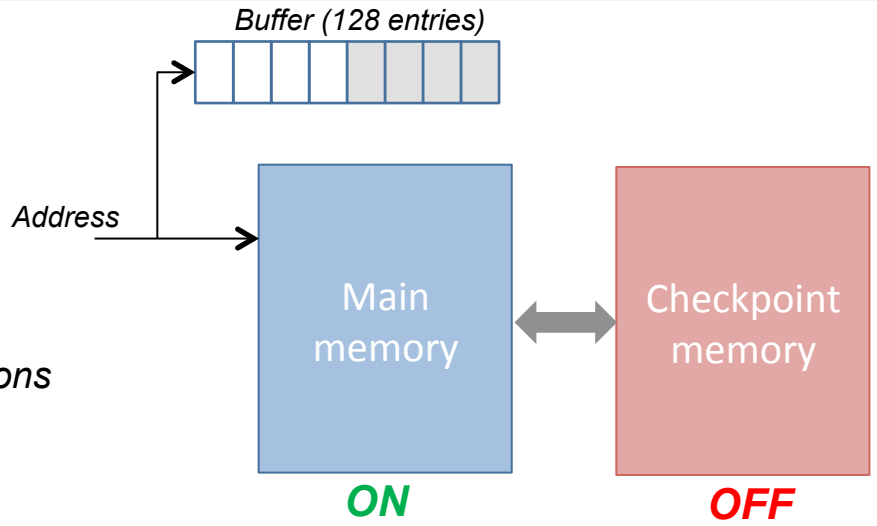


+



Rollback

Rollback (Memory part)

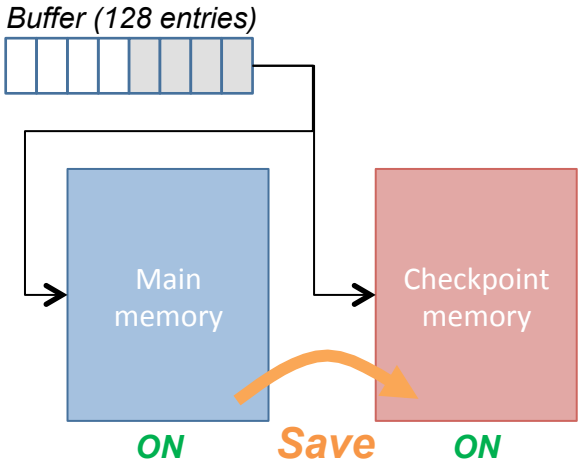


NORMAL EXECUTION

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

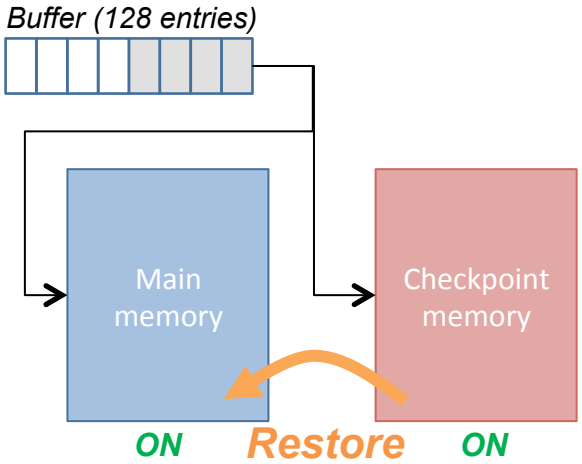
CHECKPOINT

- *Only the modified memory locations are copied*



ROLLBACK

- *Only the modified memory locations are restored*



Rollback

Rollback: validation

Checkpoint created here

```
Dhrystone Benchmark, Version 2.1 (Language: C)

Program compiled without 'register' attribute

Execution starts, 256 runs through Dhrystone
Execution ends

Final values of the variables used in the benchmark:

Int_Glob:            5
        should be:   5
Bool_Glob:           1
        should be:   1
Ch_1_Glob:           A
        should be:   A
Ch_2_Glob:           B
        should be:   B
Arr_1_Glob[8]:       7
        should be:   7
Arr_2_Glob[8][7]:    266
        should be:   266
```

●●●

```
Int_3_Loc:           7
        should be:   7
Enum_Loc:            1
        should be:   1
Str_1_Loc:           DHRYSTONE PROGRAM, 1'ST STRING
        Execution ends

Final values of the variables used in the benchmark:

Int_Glob:            5
        should be:   5
Bool_Glob:           1
        should be:   1
Ch_1_Glob:           A
        should be:   A
Ch_2_Glob:           B
        should be:   B
Arr_1_Glob[8]:       7
        should be:   7
Arr_2_Glob[8][7]:    266
        should be:   266
```

Checkpoint restored here

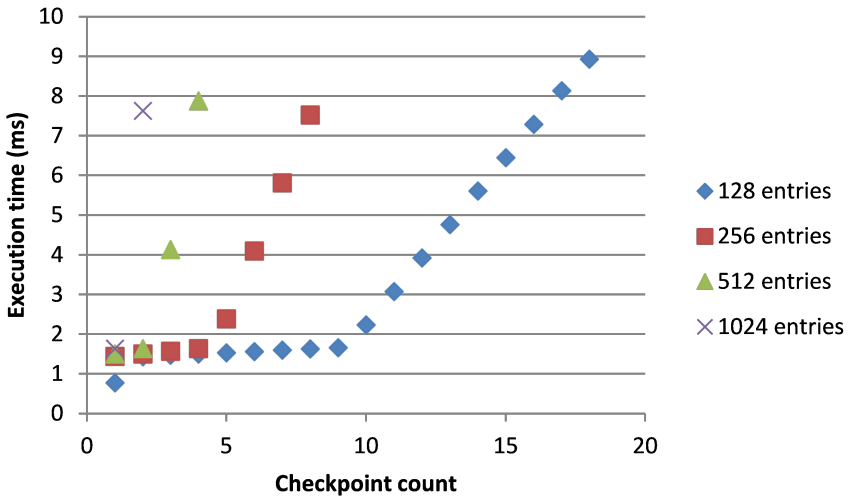


Fig. 13: Checkpoint count for different address buffer size (blowfish application)

Rollback

Rollback: validation

Checkpoint created here →

```
Dhrystone Benchmark, Version 2.1 (Language: C)

Program compiled without 'register' attribute

Execution starts, 256 runs through Dhrystone
Execution ends

Final values of the variables used in the benchmark:

Int_Glob:            5
        should be:   5
Bool_Glob:           1
        should be:   1
Ch_1_Glob:           A
        should be:   A
Ch_2_Glob:           B
        should be:   B
Arr_1_Glob[8]:       7
        should be:   7
Arr_2_Glob[8][7]:    266
        should be:   266
```

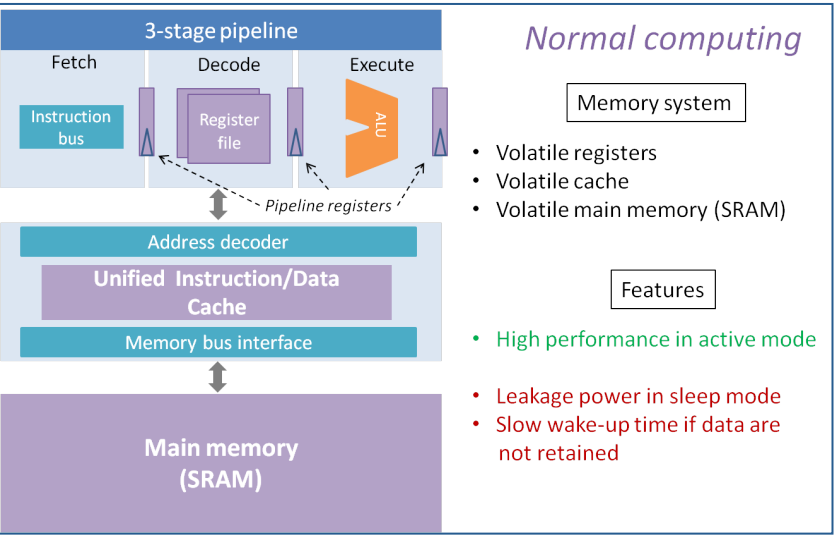
...

```
Int_3_Loc:           7
        should be:   7
Enum_Loc:            1
        should be:   1
Str_1_Loc:           DHRYSTONE PROGRAM, 1'ST STRING
        Execution ends

Final values of the variables used in the benchmark:

Int_Glob:            5
        should be:   5
Bool_Glob:           1
        should be:   1
Ch_1_Glob:           A
        should be:   A
Ch_2_Glob:           B
        should be:   B
Arr_1_Glob[8]:       7
        should be:   7
Arr_2_Glob[8][7]:    266
        should be:   266
```

Checkpoint restored here →

- Dhrystone 2.1 application
- Register part:
  - Same time/energy as intant-on/ off to backup/restore
  - Area overhead to consider
- Memory part:
  - To be evaluated more precisely
  - We know how to evaluate checkpoint memory size
  - Penalty due to cache warm-up to consider

Summary



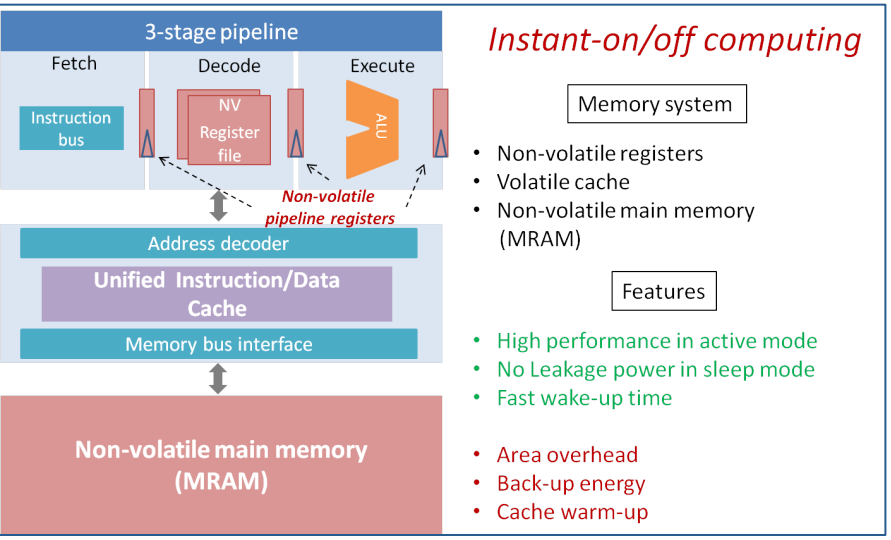
Normal computing

Memory system

- Volatile registers
- Volatile cache
- Volatile main memory (SRAM)

Features

- High performance in active mode
- Leakage power in sleep mode
- Slow wake-up time if data are not retained



Instant-on/off computing

Memory system

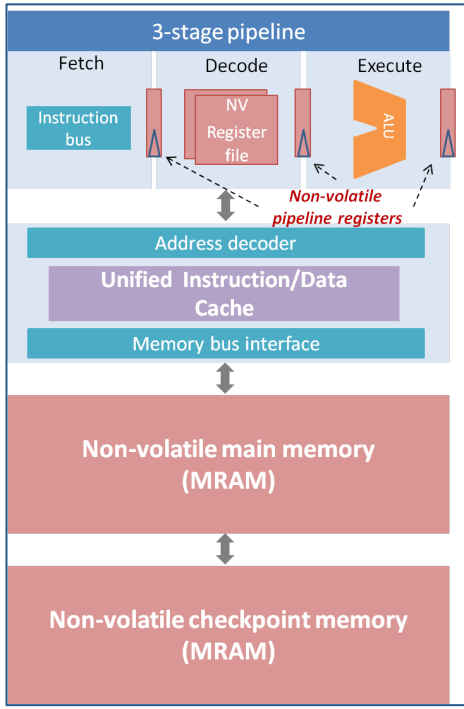
- Non-volatile registers
- Volatile cache
- Non-volatile main memory (MRAM)

Features

- High performance in active mode
- No Leakage power in sleep mode
- Fast wake-up time

- Area overhead
- Back-up energy
- Cache warm-up

Instant-on/off & Rollback Architectural changes



Instant-on/off + Rollback

Memory system

- Non-volatile registers
- Volatile cache
- Non-volatile main memory (MRAM)
- Additional checkpoint memory (MRAM)

Features

- High performance in active mode
- No Leakage power in sleep mode
- Fast wake-up time
- Fault tolerant (Rollback)

- Area overhead
- Back-up energy
- Cache warm-up
- Execution time penalty (checkpoints)

Contributions

1. Evaluation of MRAM-based cache memory hierarchy:
   - Exploration flow and extraction of memory activity
   - L1 and L2 caches based on STT-MRAM and TAS-MRAM

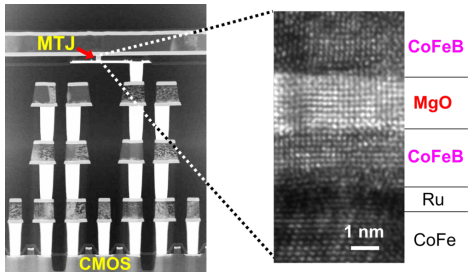
2. Non-volatile computing
   - *Instant-on/off* capability for embedded processor
   - Analysis and validation of *Rollback* mechanism

3. Secure applications with NVM

WHAT ABOUT SECURITY …

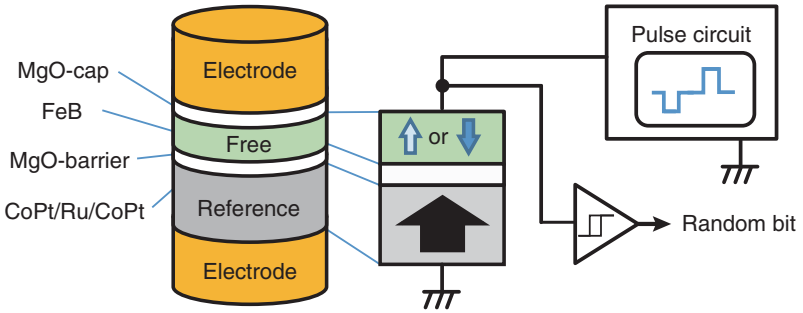
True Number Generator

Smart Efficient TRNG based on MRAM



Physically Unclonable Function

Physically Unclonable Function using MRAM



Fukushima & al (2015)

Secure elements

Dedicated logic for secure Elements based on MRAM

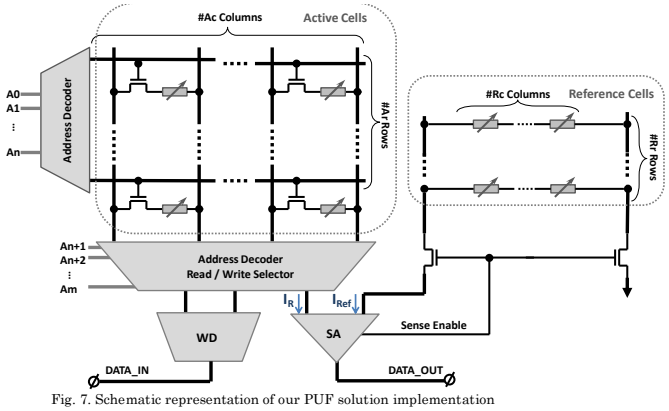


Fig. 7. Schematic representation of our PUF solution implementation

Vatajelu & al (2015)

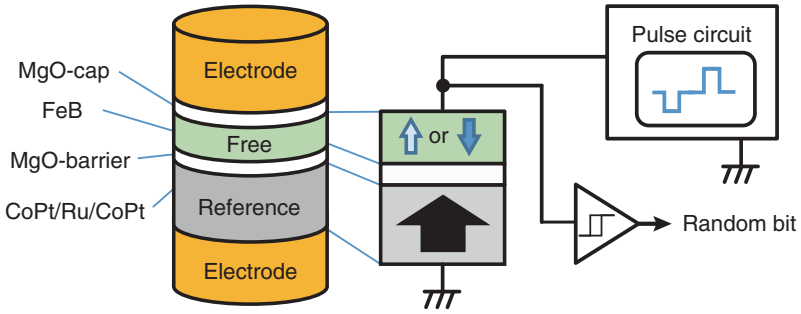
Side Channel Analysis

Side Channel Analysis of MRAM memories

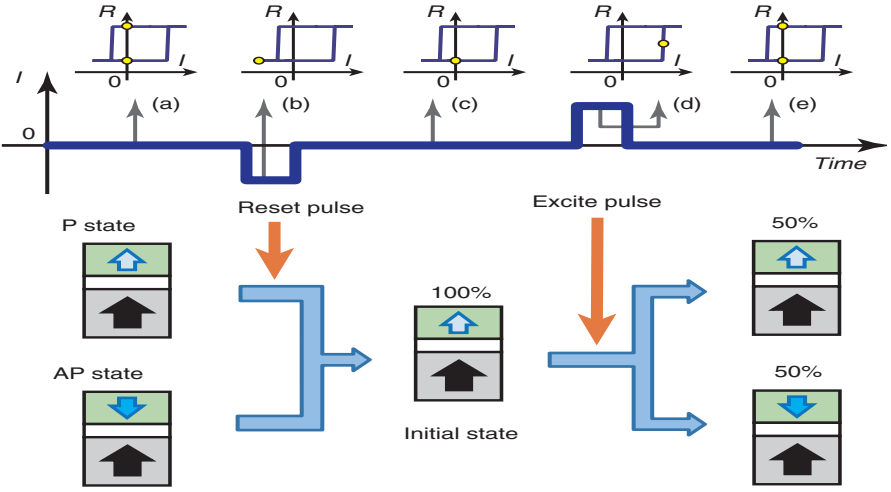
WHAT ABOUT SECURITY …

True Number Generator

Smart Efficient TRNG based on perpendicular STT-MRAM



Main principle



$$P_{sw}(I) = 1 - \exp\left\{-\frac{t}{\tau_0}\exp\left[-\Delta\left(1-\frac{I}{I_{c0}}\right)^2\right]\right\}, \quad (1)$$

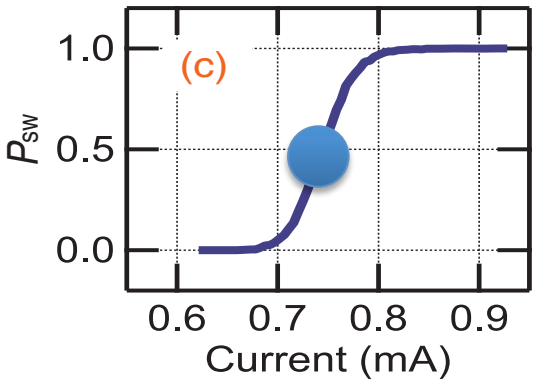


Table II. Pass rate of the randomness tests of NIST SP-800.[18]

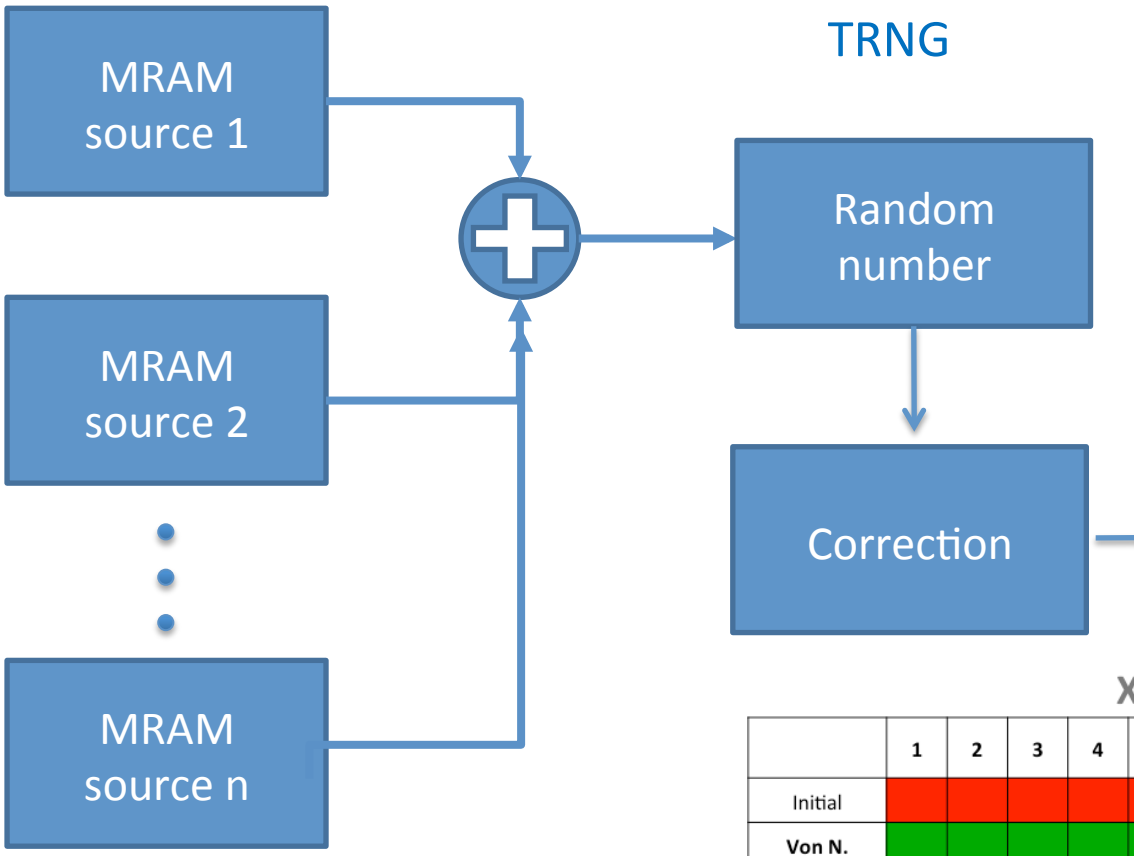
	Pass rate			
	Raw	XOR	XOR^2	XOR^3
MTJ1	0.000	0.000	0.417	0.467
MTJ2	0.000			
MTJ3	0.000	0.167		
MTJ4	0.000			
MTJ5	0.000	0.058	0.475	
MTJ6	0.000			
MTJ7	0.000	0.075		
MTJ8	0.000			

Akio Fukushima*, Takayuki Seki, Kay Yakushiji, Hitoshi Kubota, Hiroshi Imamura, Shinji Yuasa, and Koji Ando Spin dice: A scalable truly random number generator based on spintronics, Applied Physics Express **7**, 083001 (2014)

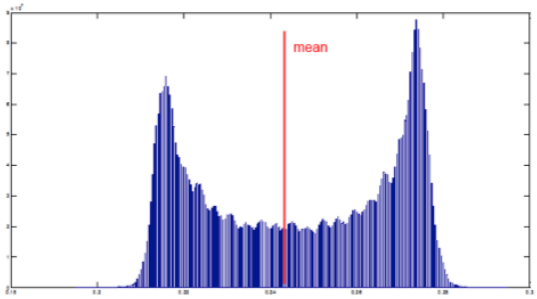
WHAT ABOUT SECURITY …

True Number Generator

Smart Efficient TRNG based on perpendicular STT-MRAM
(instead current pulse, external field is used)



50 Millions of random bits



Threshold = Raw sequence mean = 0.2464

Converted sequence mean = 0.51165

XORed sequence

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Initial															
Von N.															
Parity 5th							12/147								
Parity 7th															
Parity 9th							3/147								

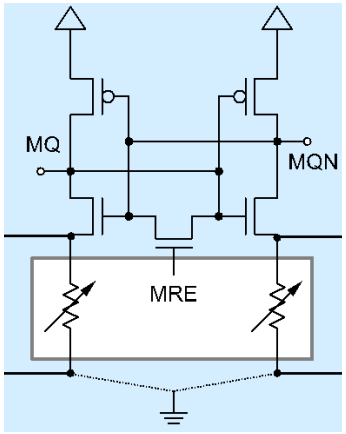
Experiment done with IEF, Thibault DEVOLDER

NIST Test

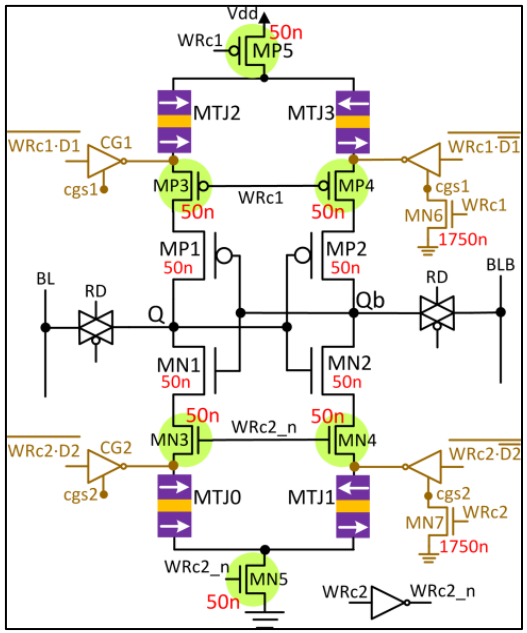
WHAT ABOUT SECURITY …

Secure elements

- Non-Volatility help security (and also Energy !)
  - Persistent data storage
  - Authentication
  - Battery backed-memories
  - Secure CPU Boot



Non-Volatile SRAM/MRAM cell



NV – SRAM
2 NV state
1 Volatile State

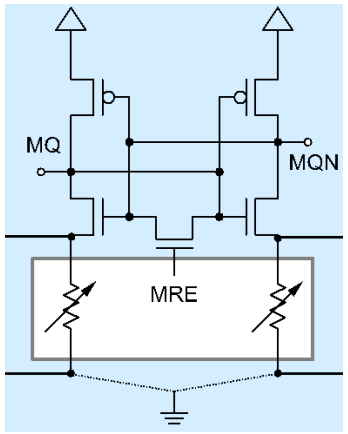
Read speed	39 ps (1.5 f)	**Density**	25.000nm2/3bits
Read Energy	5.8 fJ	**SNM**	314 mV
Write Energy	56 fJ	**DR**	Yes (1)
Static Power	396 nW	**WVD**	Yes (1)

B. Jovanovic, R. Brum, L. Torres, *Comparative Analysis of MTJ/CMOS Hybrid Cells based on TAS and In-plane STT Magnetic Tunnel Junctions*, **IEEE Transactions on Magnetic**,

WHAT ABOUT SECURITY …

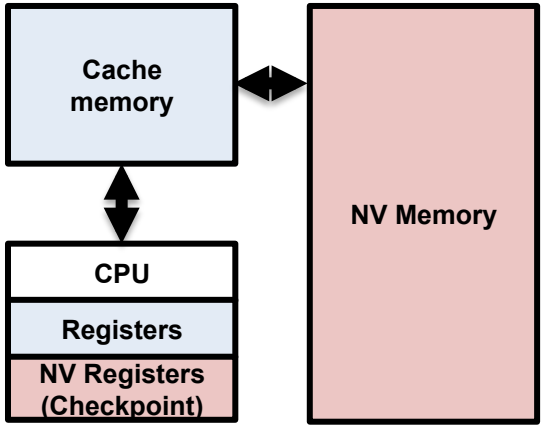
Secure elements

- Non-Volatility help security (and also Energy !)
  - Persistent data storage
  - Authentication
  - Battery backed-memories
  - Secure CPU Boot
  - ….



Non-Volatile SRAM/MRAM cell

- First evaluation of NV CPU @ LIRMM :
  - 32-bit RISC like processor
  - Validation of checkpoint/rollback capability
  - Non-Volatile register bank (instead Volatile)
  - Low performance overhead
  - Non-volatile memory from register level to main memory



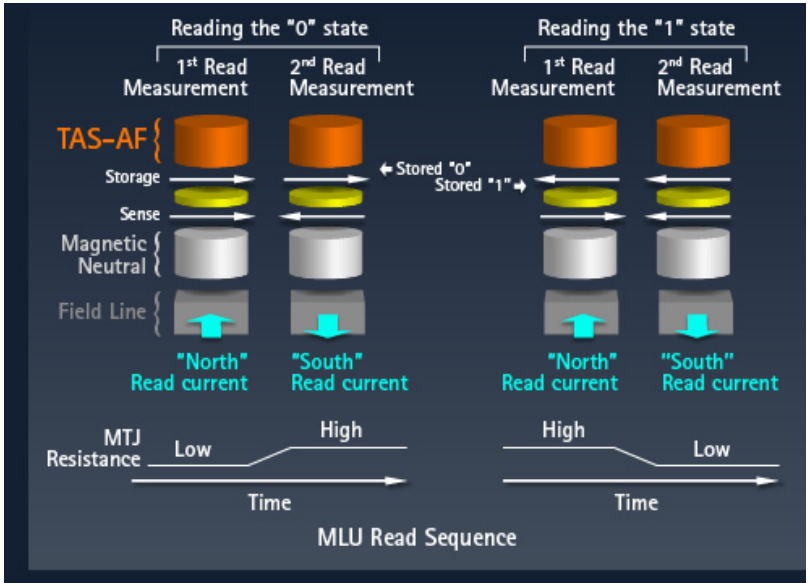
B. Jovanovic, R. Brum, L. Torres, *Comparative Analysis of MTJ/CMOS Hybrid Cells based on TAS and In-plane STT Magnetic Tunnel Junctions*, **IEEE Transactions on Magnetic**,

WHAT ABOUT SECURITY …

Secure elements

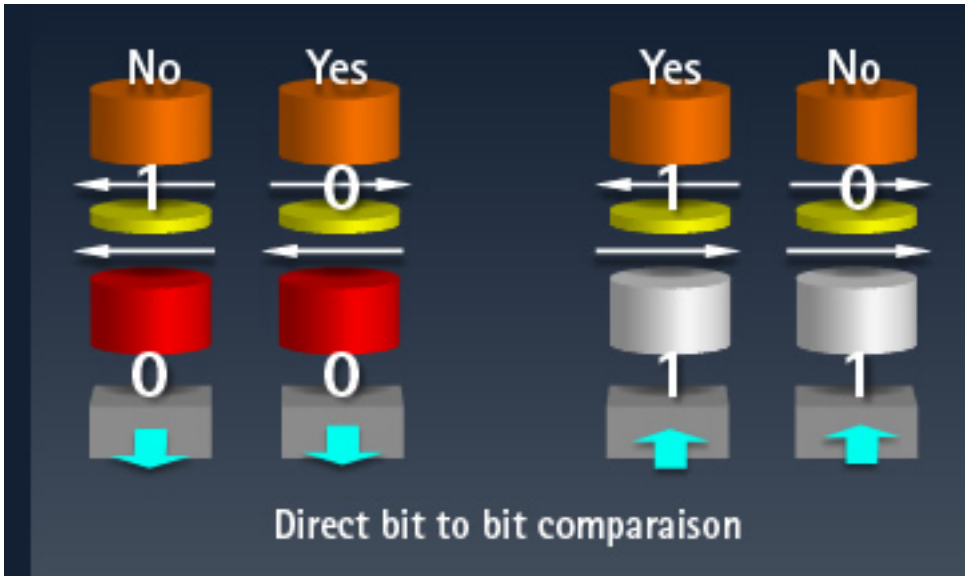
CROCUS Technology
Blossoming future

Magnetic Logic Unit



XOR Function

Match In place



Authentication function/comparison

Symetric Cryptography → Elementary operations XOR, Substitution, Shift

WHAT ABOUT SECURITY …

Physically Unclonable Function

PUF solution exploits the differential sensing during read operation, based on read current comparison against a reference value.

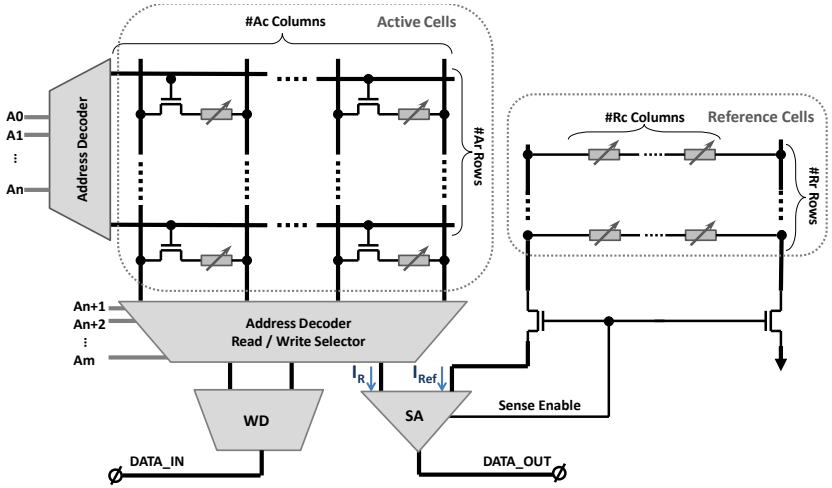
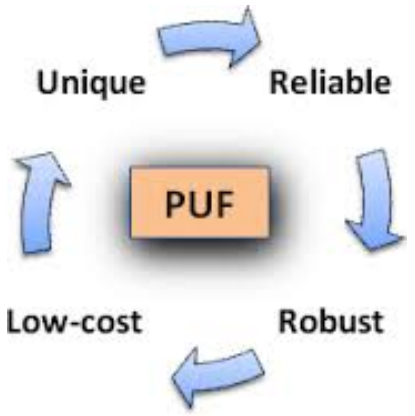


Fig. 7. Schematic representation of our PUF solution implementation



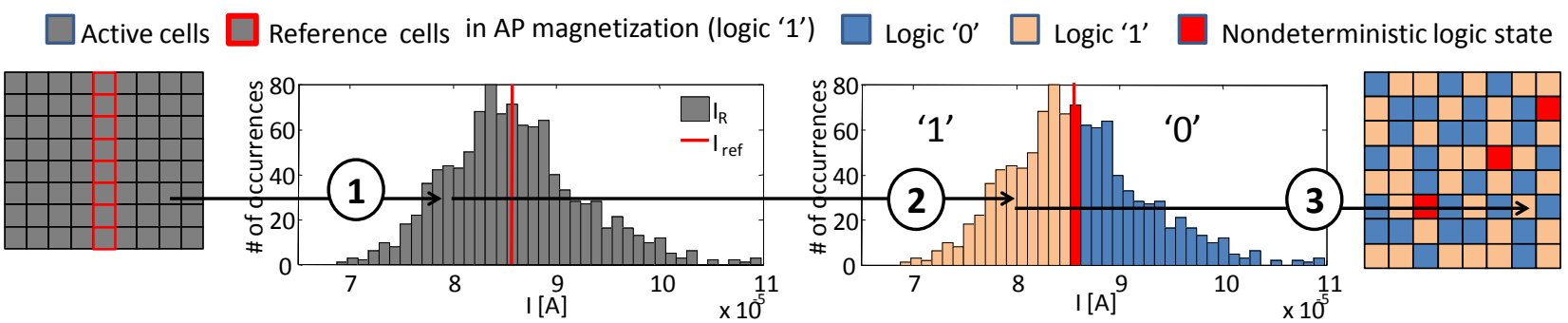
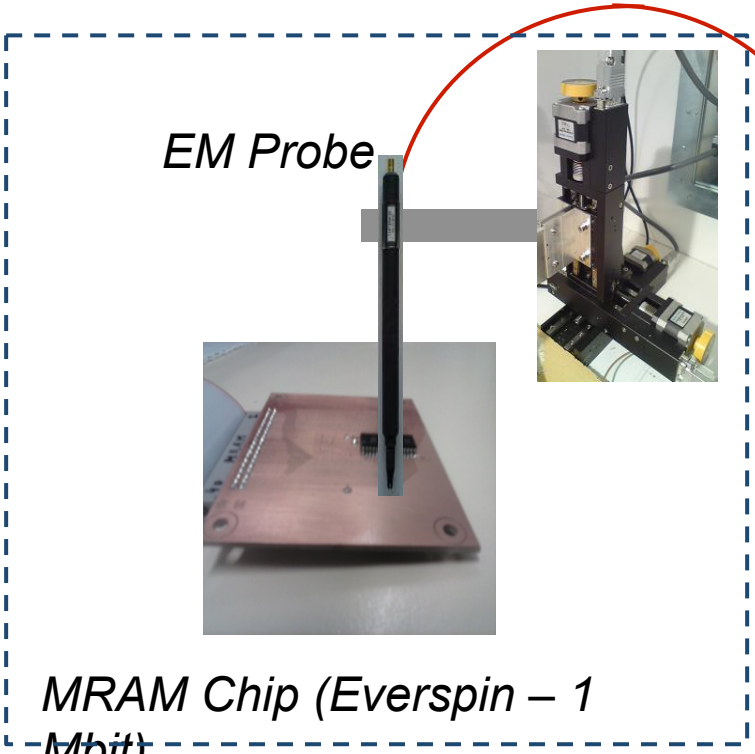


Fig. 4. The implementation strategy of the proposed PUF solution: 1) Write all cells to '1'; 2) Read each cell; 3) Use the read value

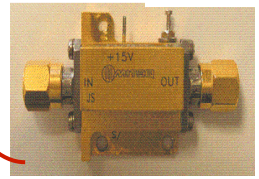
WHAT ABOUT SECURITY …

Side Channel Analysis

Test Bench Overview



EM waves probing

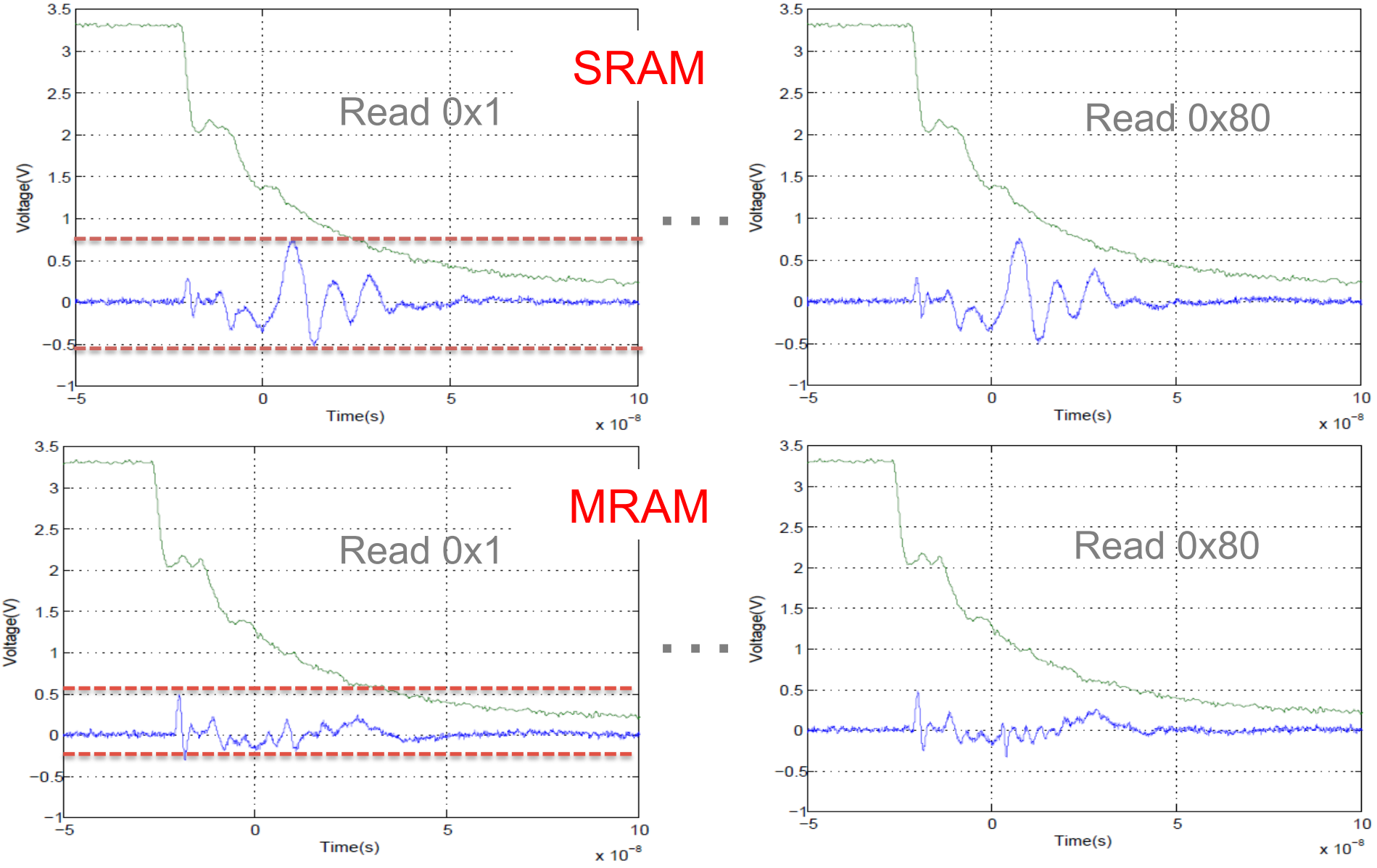


EM waves plotting

oscilloscope settings & data acquisition

Active DSO control

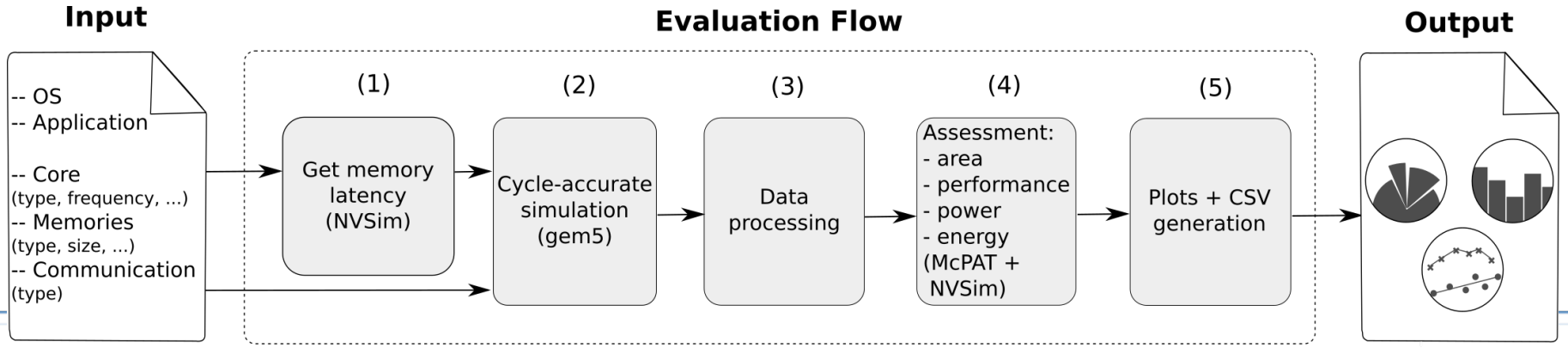
WHAT ABOUT SECURITY …



MRAM Conclusion

MRAM has a high potential to:

- **Certainly Reduce energy consumption**
  - At cache level (sure and proven)
  - Normally-off computing (to be confirmed)
- **Can facilitate some features**
  - Normally-off computing / Instant on-off
  - Backward error recovery (Rollback)
- **Results should be confirmed through measurements on silicon prototype !**
- **Link with compilation and OS (→ National project started, Non-volatility)**
- **Under development : a complete flow including power consumption estimation of processor + memory hierarchy**

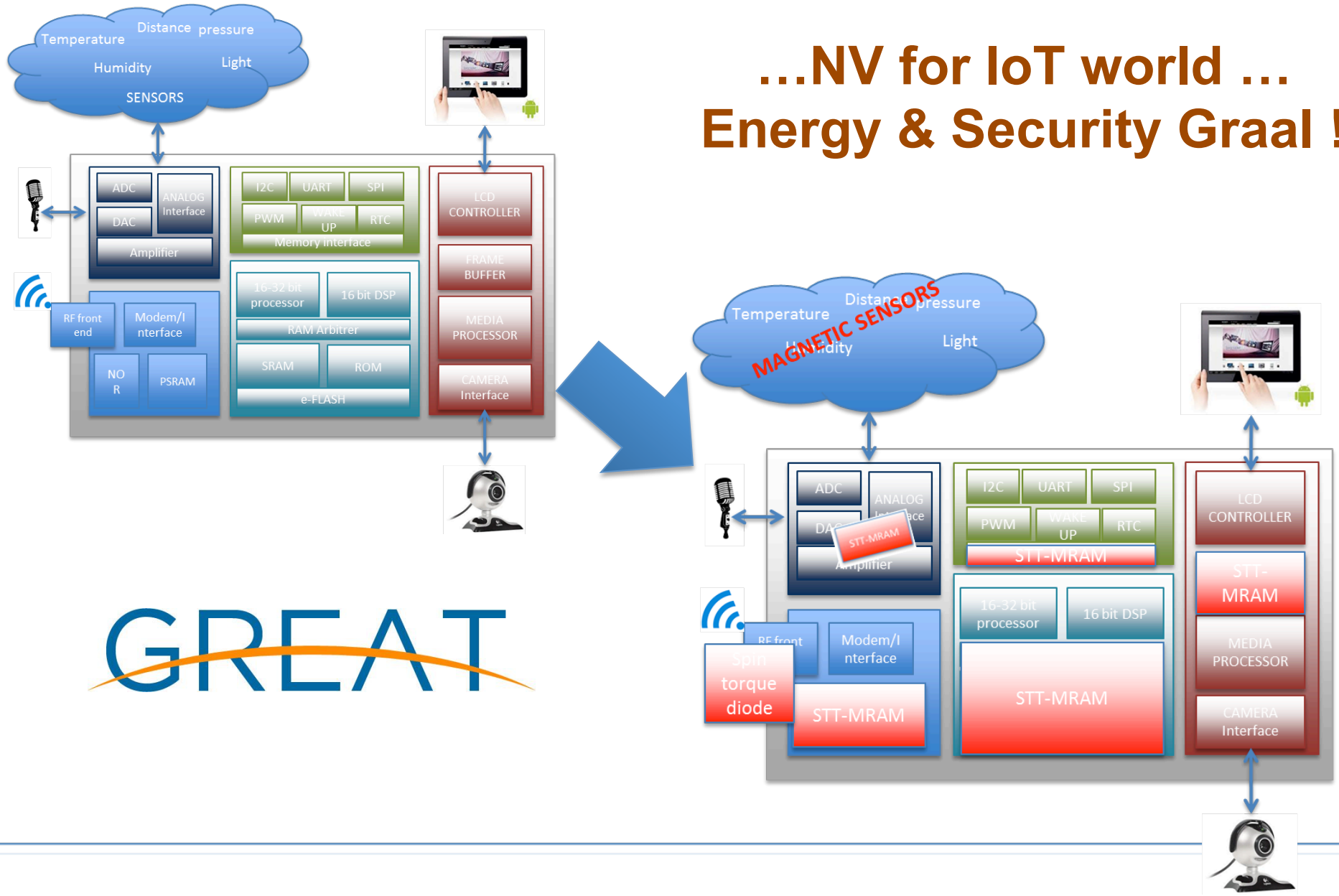


Overall Conclusion

- No Ideal memory technology – really depending of the targeted application
- Normally-off computing will be tomorrow the key element for SoC design (for Energy !)
- But Non-Volatile memory could change the way to imagine the memory hierarchy

- Rather than improved the memory hierarchy… rethink it!
  - Distributed NV elements/memories
  - Security (with IoT trend) will be everywhere
  - Better understand NV technologies for security issues
  - Use NV Technologies for security !

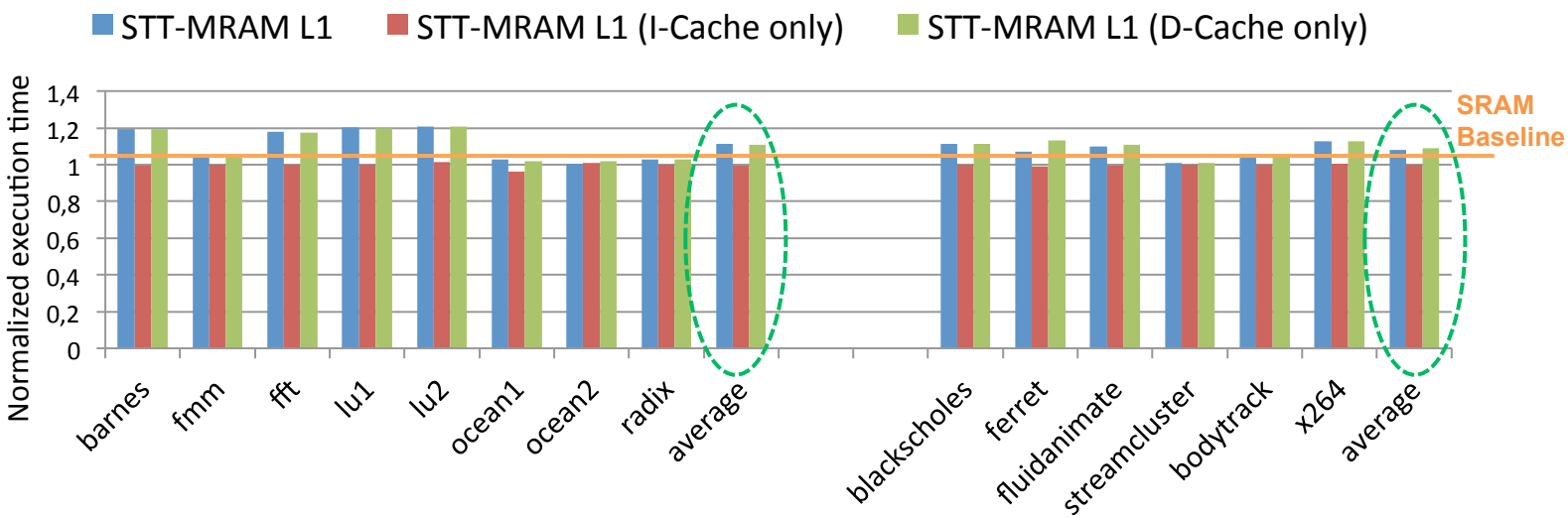
NV future …


Temperature
Distance
pressure
Humidity
Light
SENSORS
ADC
DAC
ANALOG Interface
Amplifier
I2C
UART
SPI
PWM
WAKE UP
RTC
Memory interface
LCD CONTROLLER
FRAME BUFFER
MEDIA PROCESSOR
CAMERA Interface
RF front end
Modem/I nterface
NO R
PSRAM
16-32 bit processor
16 bit DSP
RAM Arbitrer
SRAM
ROM
e-FLASH
…NV for IoT world …
Energy & Security Graal !
MAGNETIC SENSORS
STT-MRAM
Spin torque diode
GREAT

THANK YOU FOR YOUR ATTENTION

MRAM-based L1

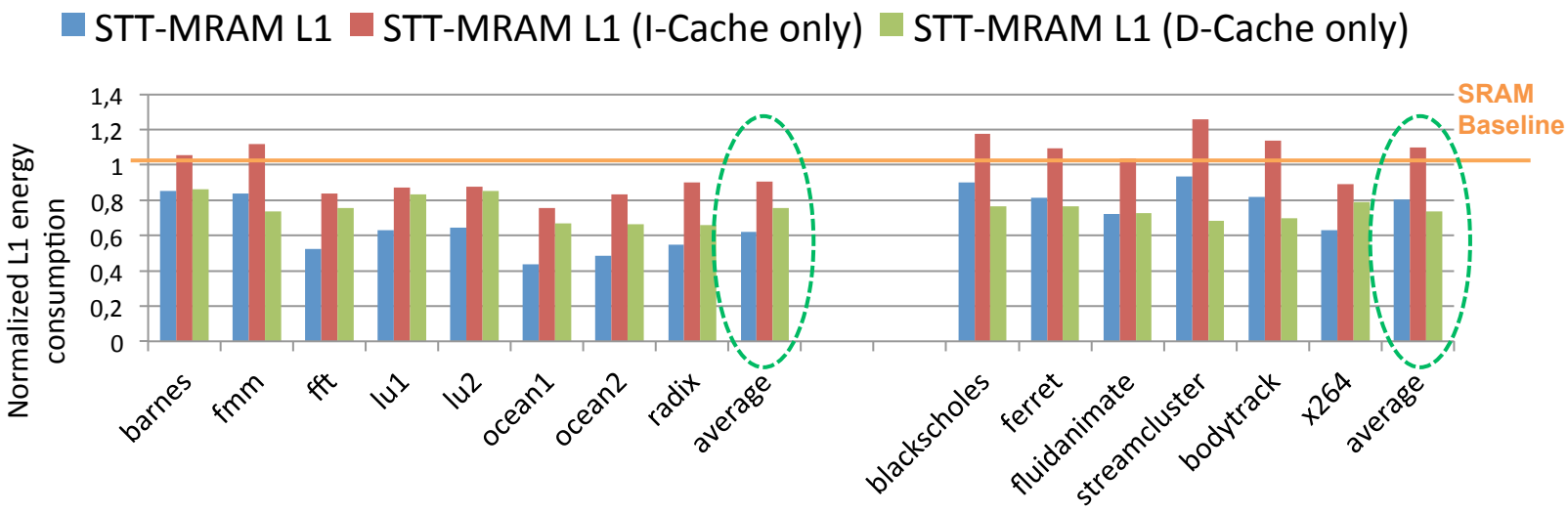
Execution time



- Observations:
  - Up to 21% of runtime penalty
    - L1 much more accessed than L2

MRAM-based L1

Total L1 energy consumption



- Observations:
  - Low leakage does not always compensate the high dynamic energy of MRAM