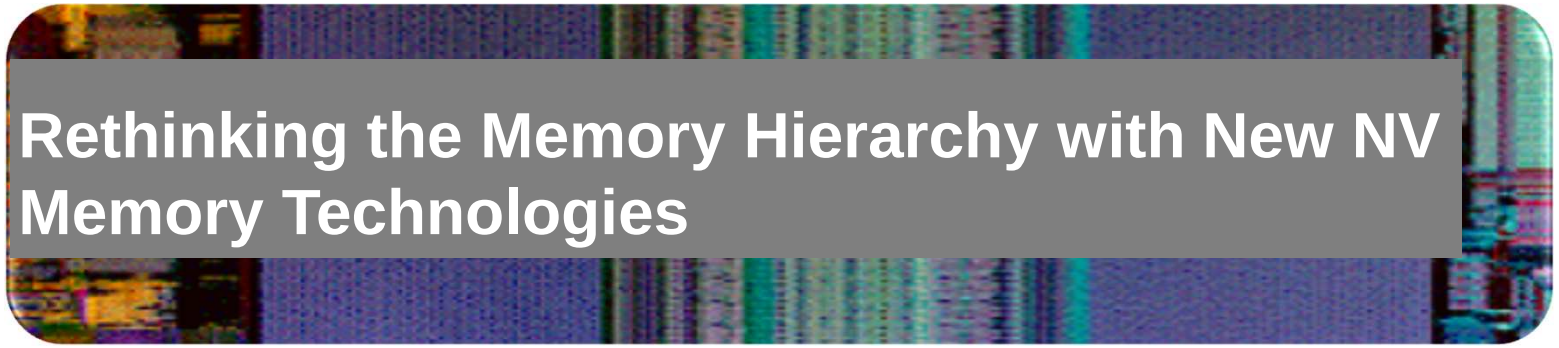
REVISITING THE MEMORY HIERARCHY FOR TOMORROW COMPUTING SYSTEMS

Leti Devices Workshop | Elisa Vianello | December 4, 2016

OUTLINE

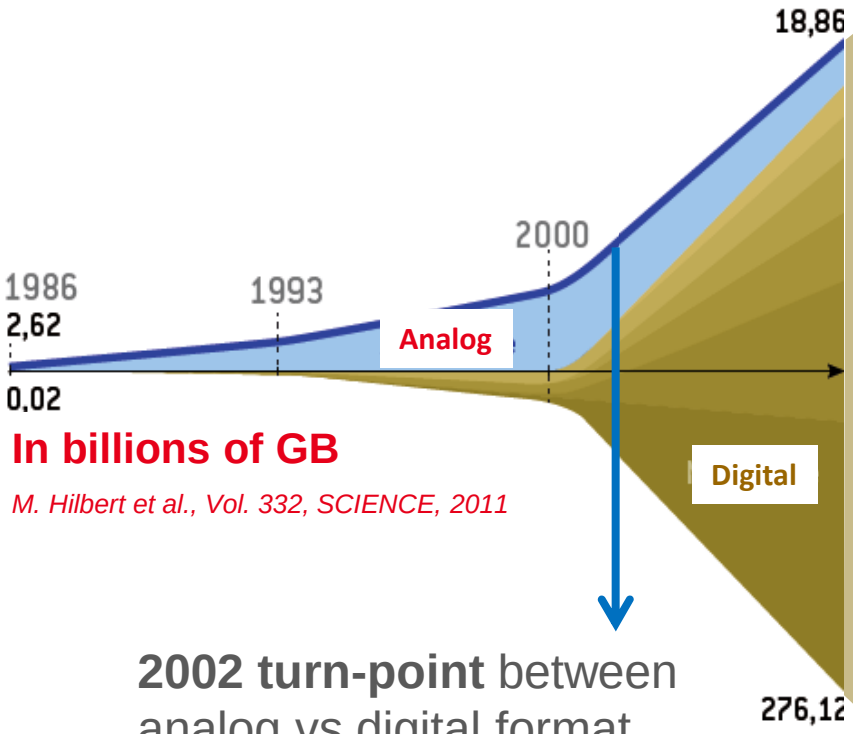
- Ever Increasing Need for More Memory

- Rethinking the Memory Hierarchy with New NV Memory Technologies



- Rethinking the System Architecture?

EVER INCREASING NEED FOR MORE MEMORY



In billions of GB

M. Hilbert et al., Vol. 332, SCIENCE, 2011

2002 turn-point between analog vs digital format

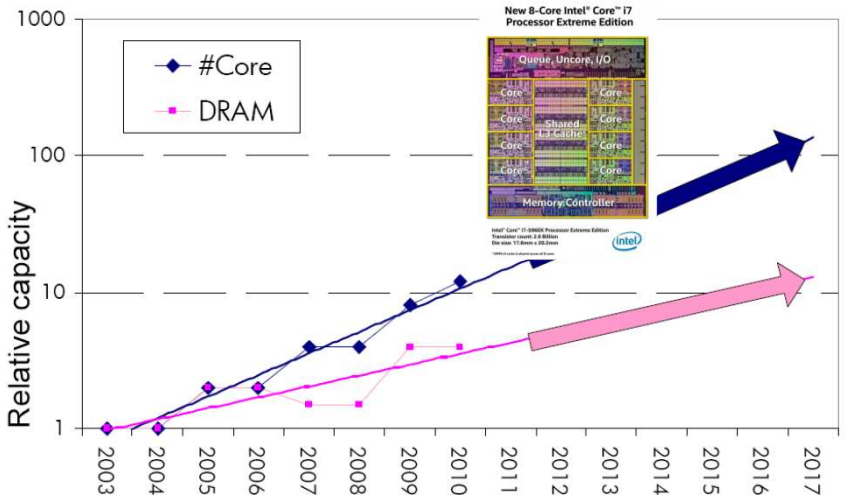
The amount of data being created is growing 40% a year into the next decade

By 2020 the digital universe is expected to contain nearly as many digital bits as there are Stars in the Universe…

EVER INCREASING NEED FOR MORE MEMORY

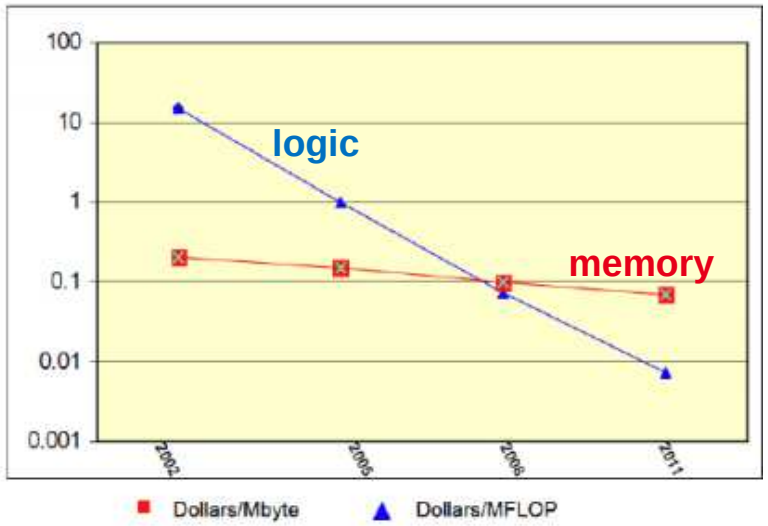
The growth in data produced is outpacing the improvements in the density and cost of storage technologies

Core count doubling ~ every 2 years
DRAM capacity doubling ~ every 3 years



Source: Liam et al ISCA 2009

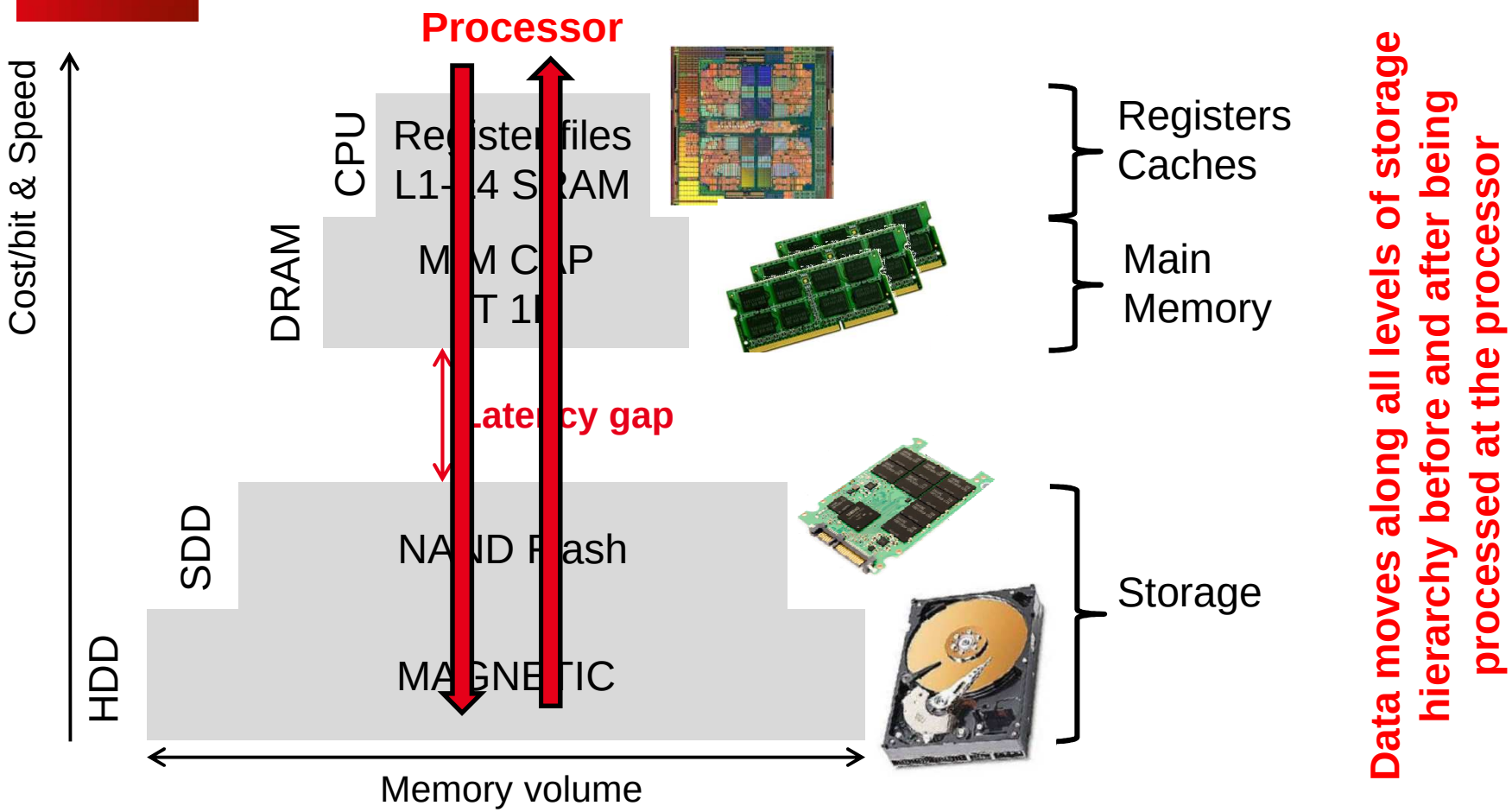
Compute costs dropping faster than memory costs



Source: IBM Deep Computing

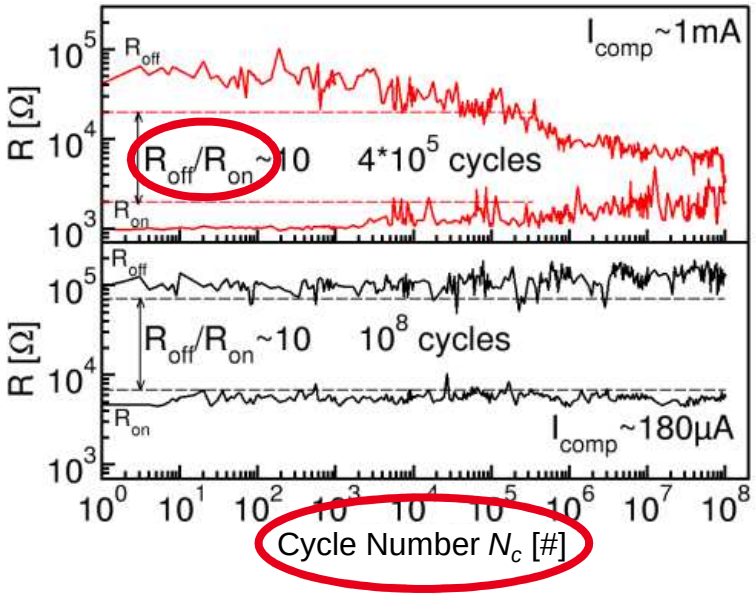
New memory technologies and rethinking of system architecture focused on data storage and management are needed!

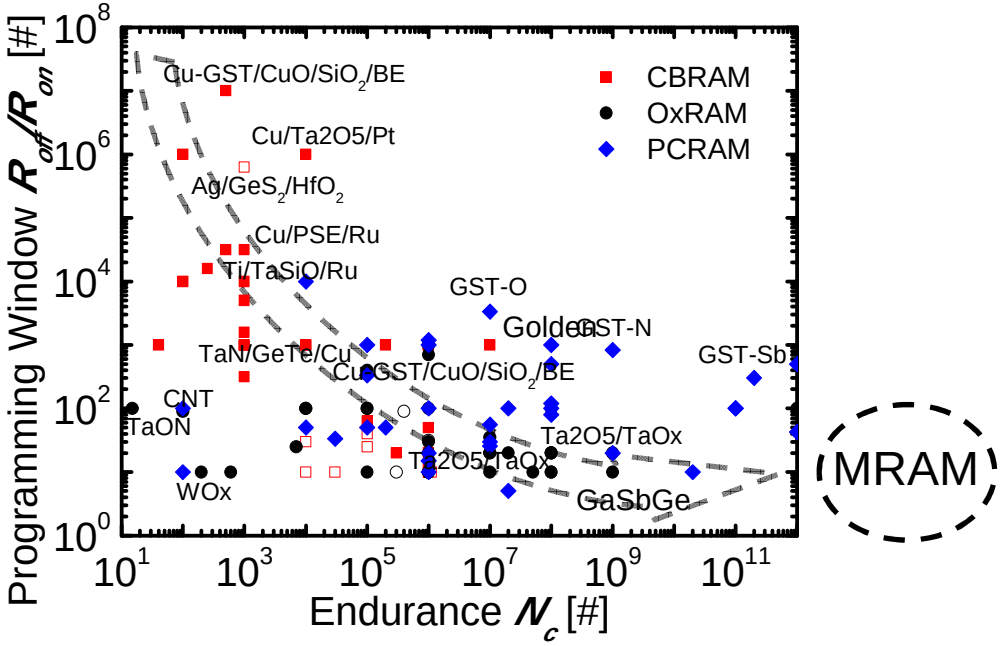
CONVENTIONAL TYPES OF MEMORIES AND TECHNOLOGIES



Can new NVMs (RRAM, PCM, MRAM…) help to reduce the latency gap and limit data movements across the Memory Hierarchy?

RRAM MEMORIES: PROGRAMMING WINDOW VS. ENDURANCE





- CBRAM → largest R_{off}/R_{on} chalco-based and/or bilayers; best endurance oxide-based
- OxRAM → largest R_{off}/R_{on} non-polar; best endurance bipolar
- PCRAM → best endurance GST-based
- MRAM → outlier..

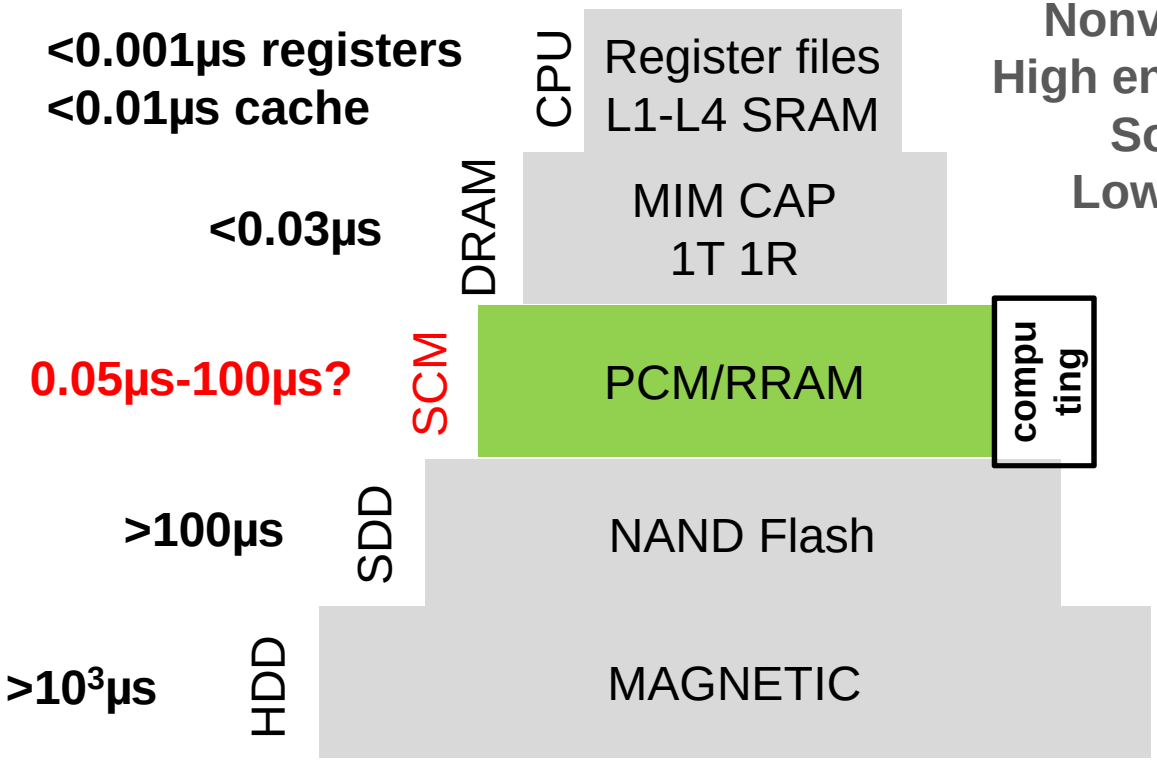
E. Vianello IEDM 2014
L. Perniola IMW 2016

Universal Memory does Not Exist!

4.5 paper on Monday afternoon on RRAM Endurance, Retention and Window Margin Trade-off

STORAGE CLASS MEMORIES

Latency of access

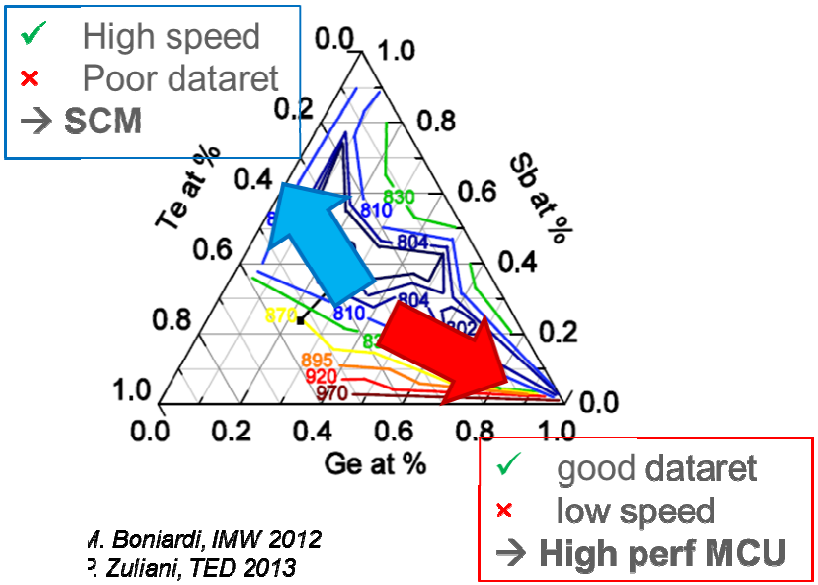


Requirements:

Fast access speed approaching DRAM
Nonvolatile retain data at power off
High endurance (program/erase cycles)
Solid state (no moving parts)
Low cost/bit (approaching HDD)

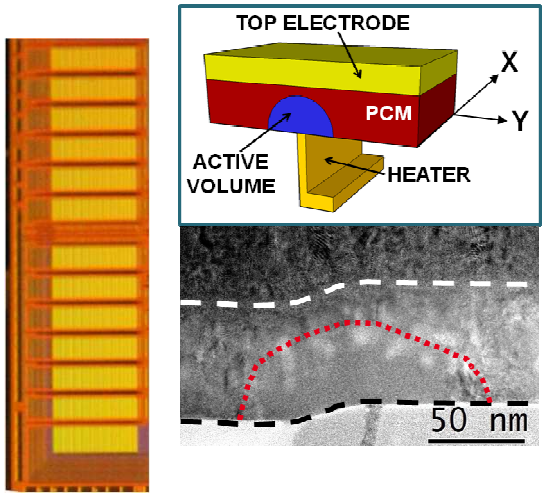
processing @ SCM: from data compression to query processing

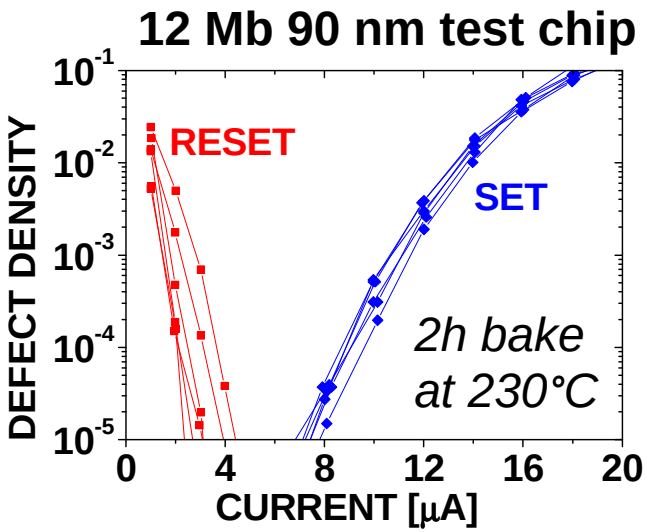
« TRANSFORMATIONAL » ABILITY OF PCM

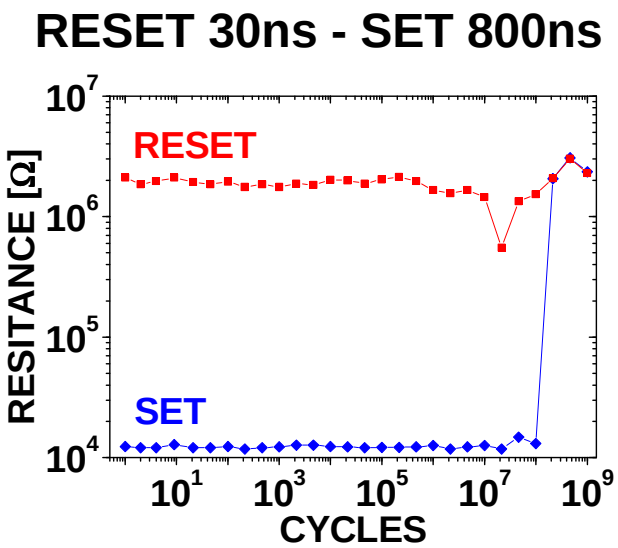


M. Boniardi, IMW 2012
P. Zuliani, TED 2013
G. Navarro, IEDM 2013

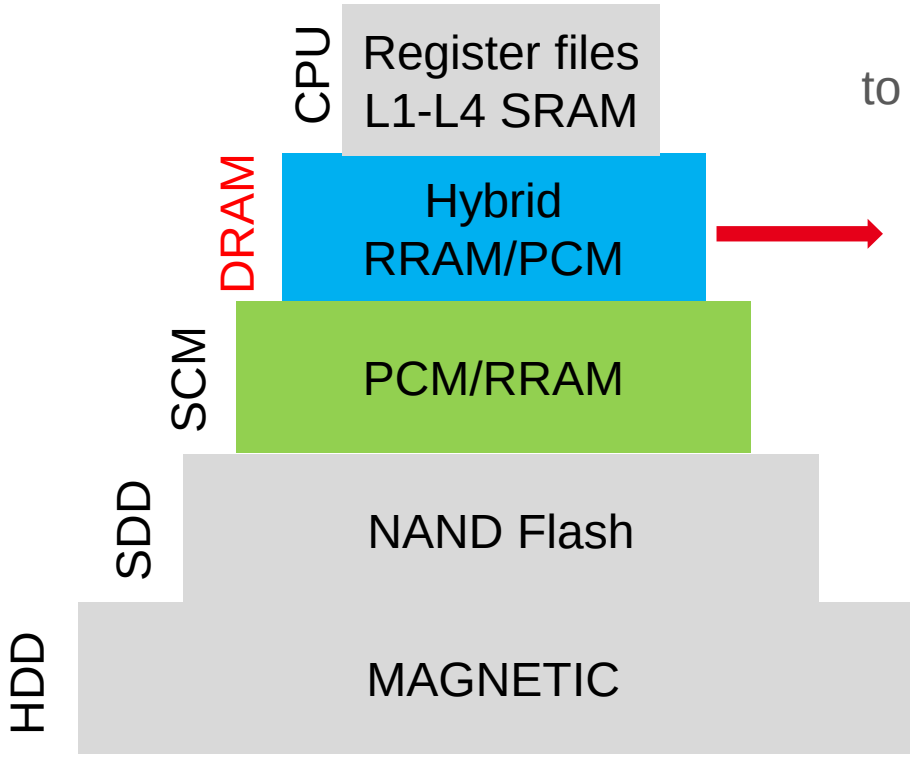
H. Y. Cheng et al., IEDM 2012
G. Navarro et al., IEDM 2013
H.Y. Cheng et al., JAP 2014
P. Zuliani et al., SSE 2015
V. Sousa et al., VLSI 2015







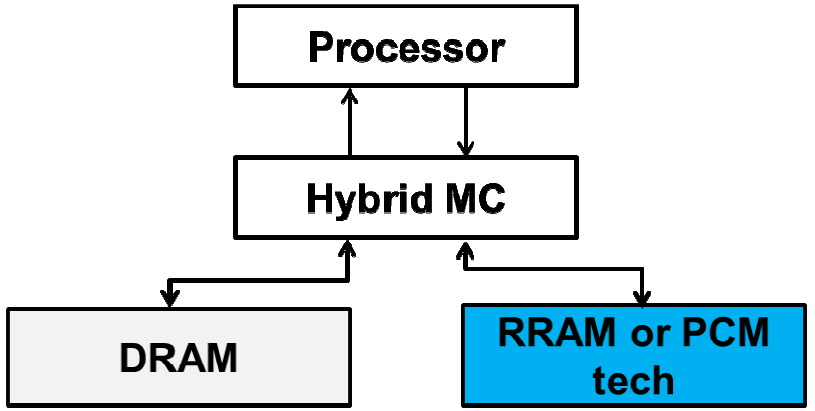
HYBRID MAIN MEMORY



High capacity working memory

Hybrid memory: DRAM as a cache to RRAM or PCM tech

to achieve the best of multiple technologies

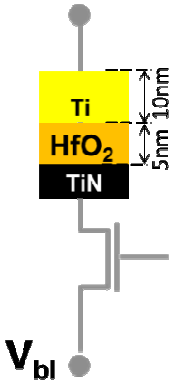


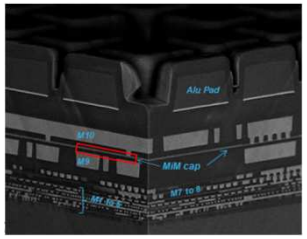
- fast
- durable

- non-volatile
- high-density
- low-cost

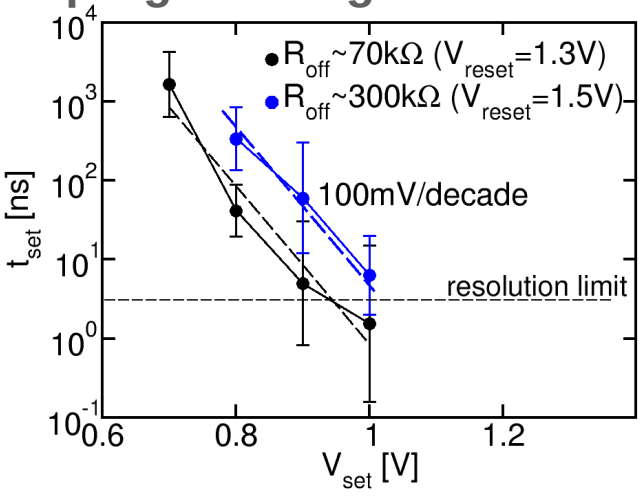
Ti/HfO_2 BASED-OxRAM

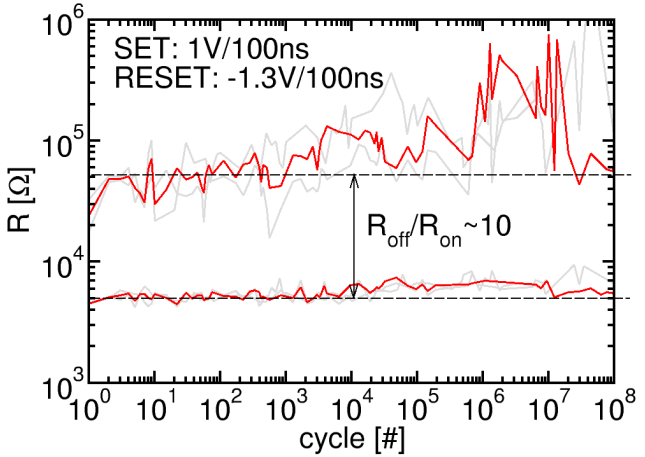
<100ns programming time at 1V with up to 100M cycles





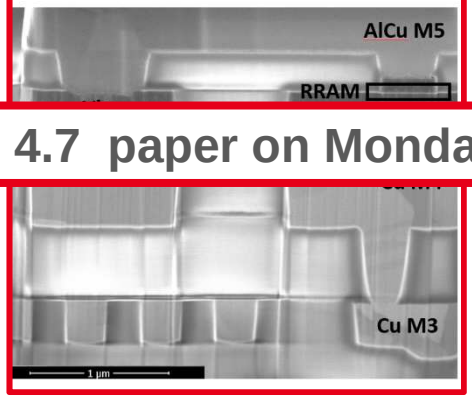
28nm CMOS



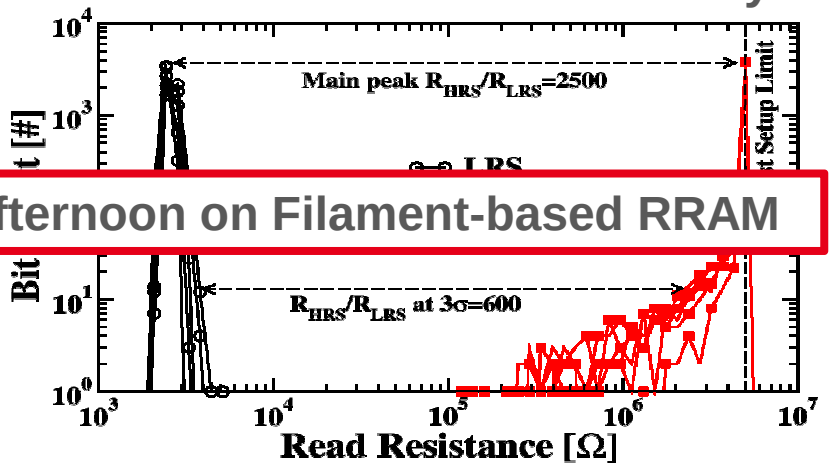


E. Vianello IEDM 2014
A. Benoist IRPS 2014 ST/Leti

MAD Leti OxRAM

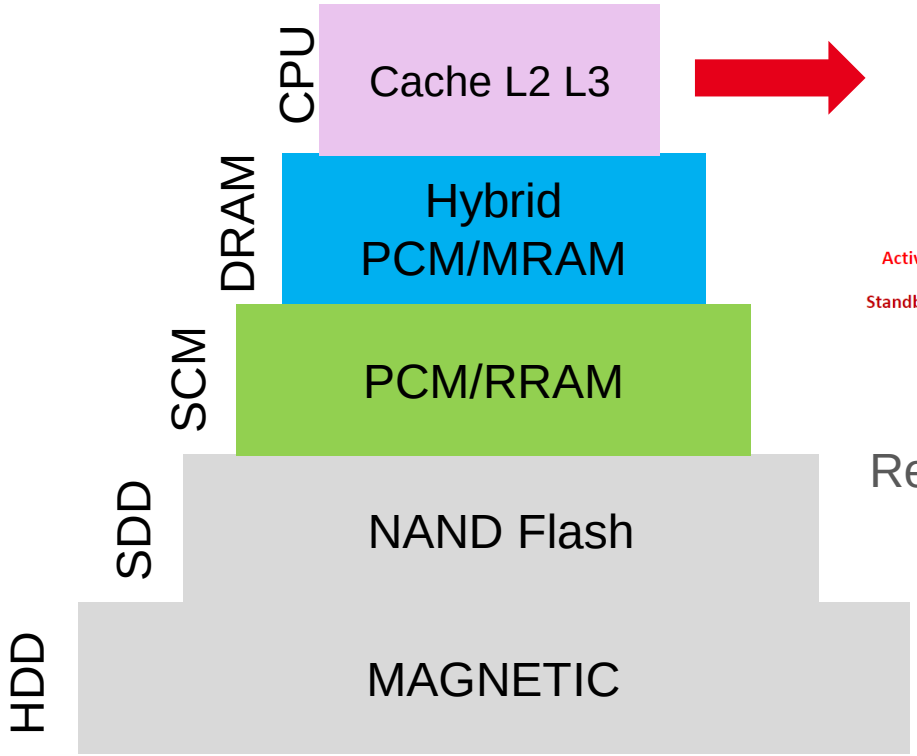


LRS and HRS on 4kbits array

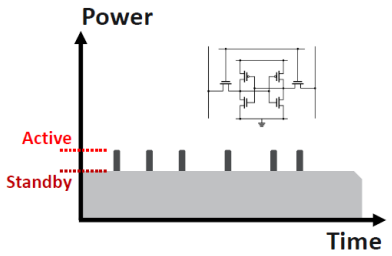


4.7 paper on Monday afternoon on Filament-based RRAM

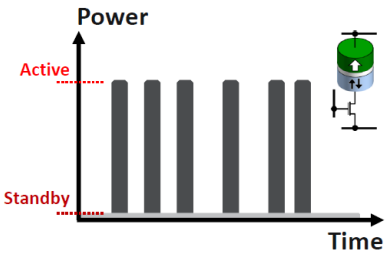
L2-L3 CACHE



Replacing SRAM (L2-L3)
best candidate is STT-MRAM
(fast and high endurance)

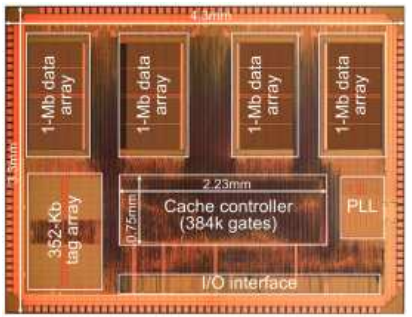


SRAM Cache Operation



STT-MRAM Cache Operation

Reduction of stand-by power consumption

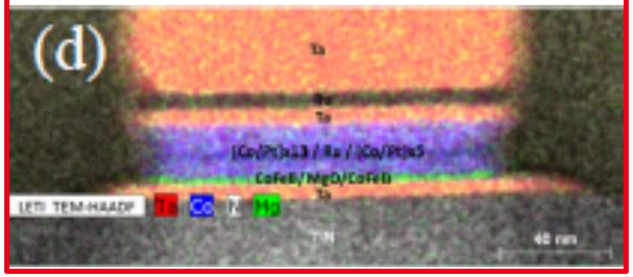


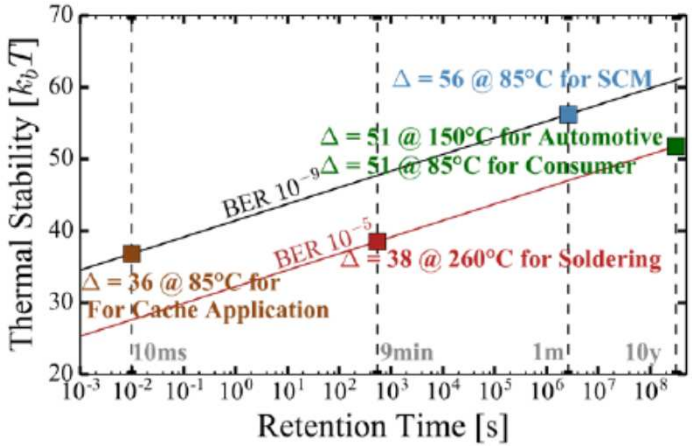
4Mbit STT MRAM
Read 3.3ns @ 1.25V
65nm CMOS
[Noguchi, Toshiba, ISSCC 2016]

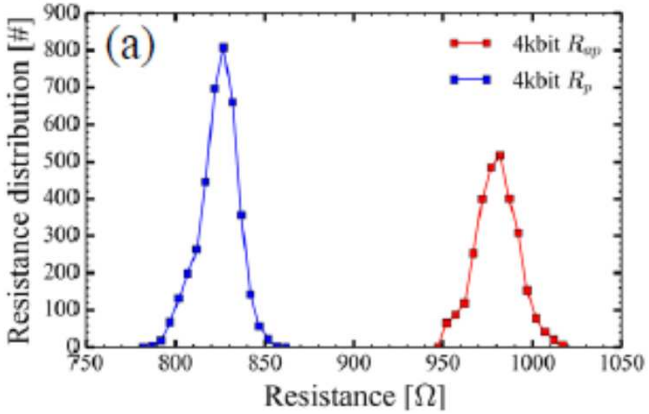
REPLACING SRAM WITH STT-MRAM

STT-MRAM demonstrated fast writing speed and high endurance, however retention still to be fully characterized

MAD Leti pSTT-MRAM







4kbit array

different retention extraction methods are compared
a trade off exists between thermal stability and programming speed

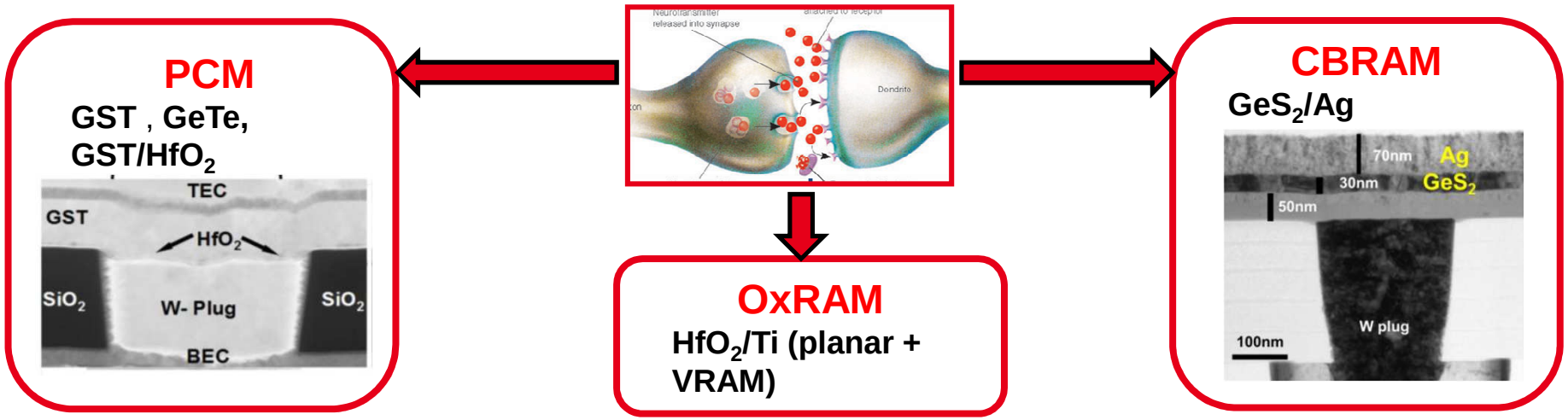
27.3 paper on Wednesday morning on Data Retention Extraction Methodology pSTT-MRAM

RETHINKING THE SYSTEM ARCHITECTURE?

Neuromorphic systems

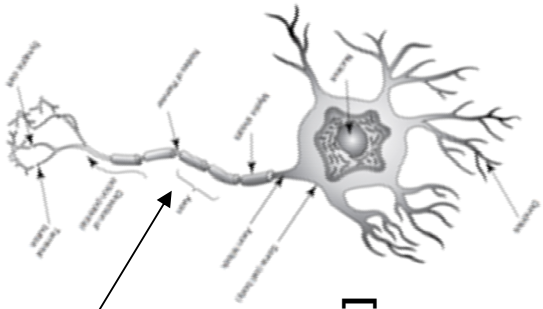
- Detect and predict patterns in complex data
- visual or auditory data analysis

RRAM has been promoted by Leti (and many others!) to emulate synaptic plasticity: the ability of synapses to strengthen or weaken over time



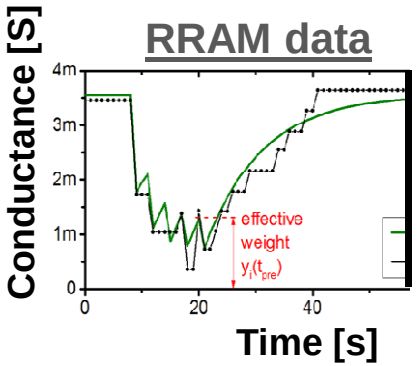
RRAM TO IMPLEMENT ON-LINE LEARNING

In nature synapses have a volatile component, is it useful for the learning process?

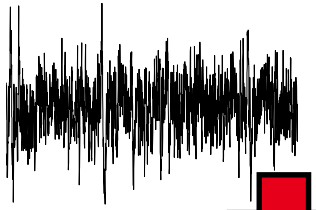


biological data

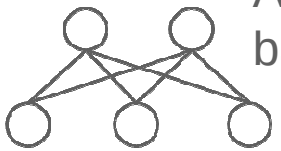
RRAM data



synaptic change has a volatile component



Highly noisy neurological data



ANN with RRAM based synapses

decoding of neural signals

The volatile component allows to improve detection in highly-noisy input data

16.6 paper on Tuesday morning on Short and Long-term Synaptic Plasticity Using OxRAM

CONCLUSION

The data explosion is leading to a corresponding growth in data centric applications (capture, classify, archive…). The adoption of new NVMs enable a rethinking of system architecture-based on data storage and management.

However universal memory does not exist, different NVM technologies have to be introduced in the storage hierarchy

Challenges:

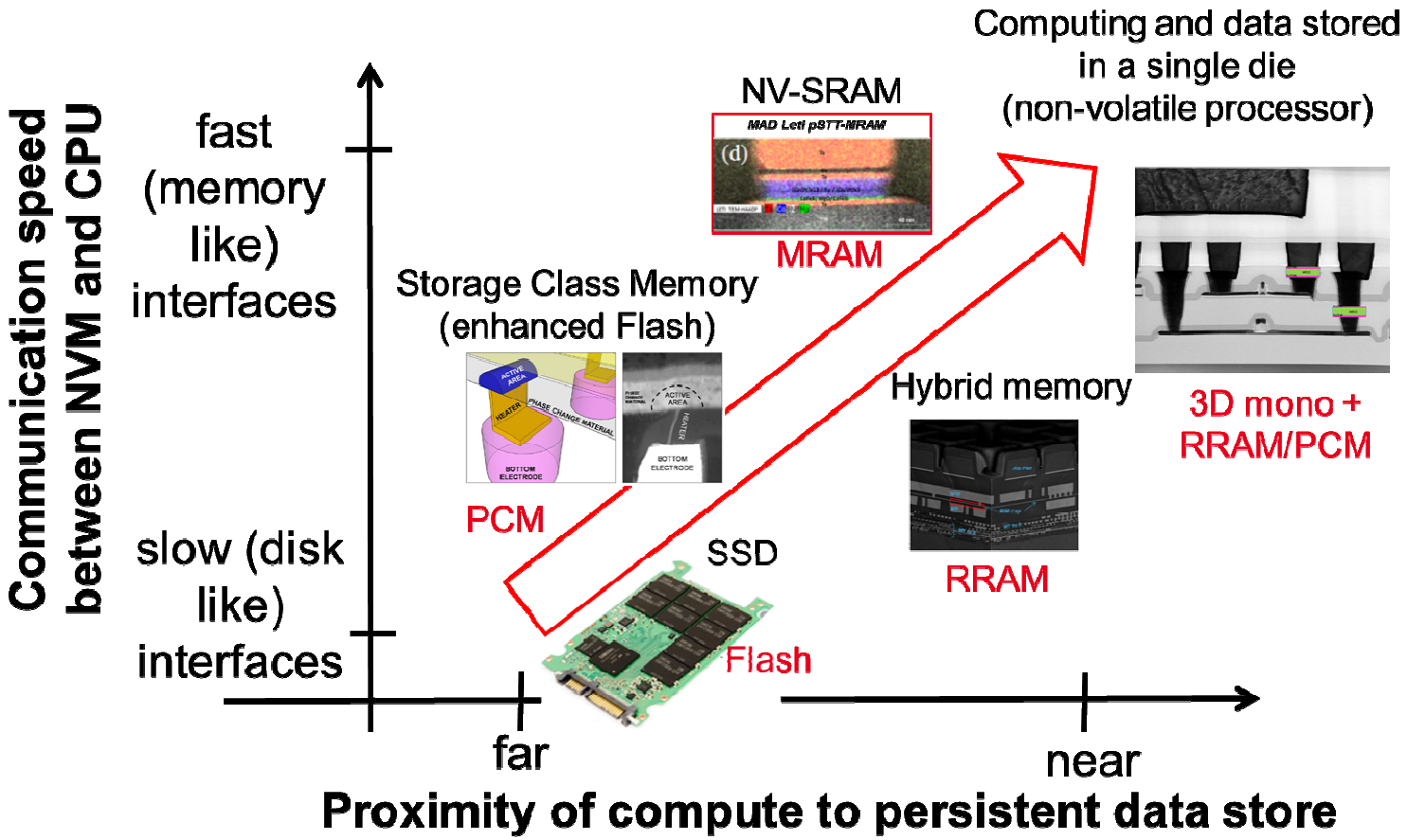
→ **Design matched to the specific memory technology features**

→ **Non volatility does non come for free**

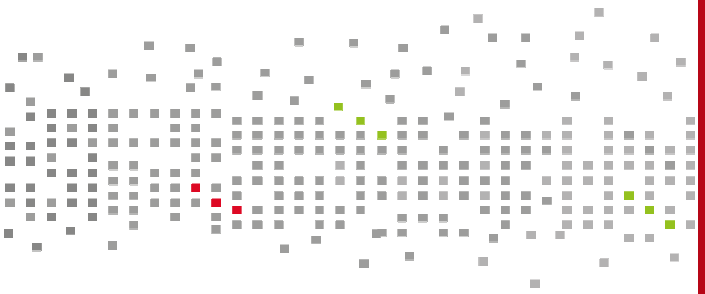
- static power vs. active power
- memory window vs. endurance
- memory window vs. data retention.

CONCLUSION

Thanks to the new NVMs technologies (PCM, RRAM, MRAM) the persistent memory is getting closer to the compute center avoiding wasting energy in the movement



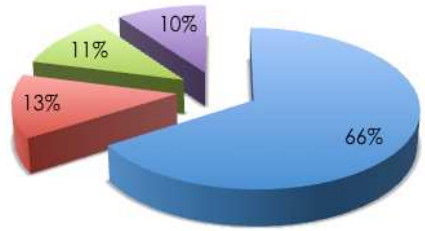
Thank you for your attention



Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 rue des Martyrs | 38054 Grenoble Cedex | France
www.leti.fr

INSTITUT CARNOT CEA LETI

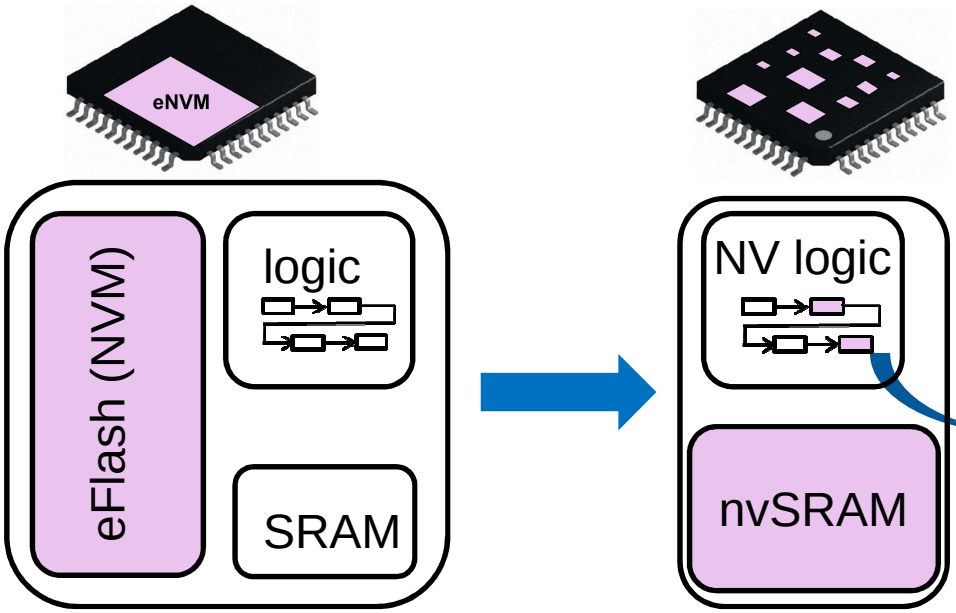
EMBEDDED SYSTEMS: TOWARD DISTRIBUTED MEMORY



- MCU (Stand-by)
- Sensor
- Radio
- MCU (active)

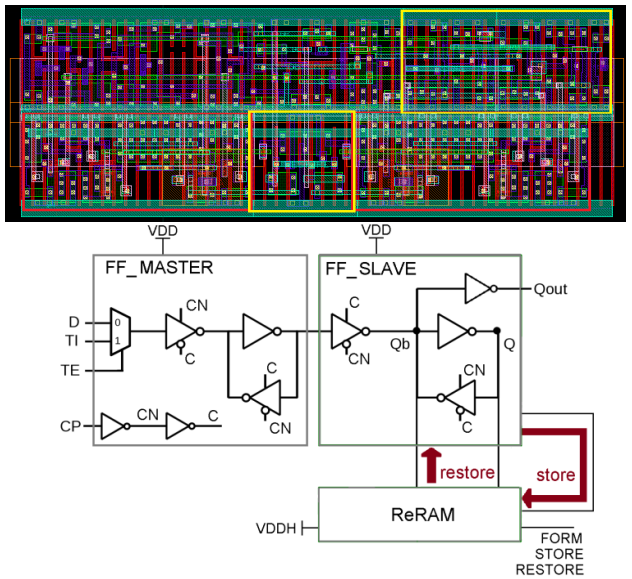
Source: Renesas ASP-DAC 2014

MCU power challenge: need for Ultra-Low Stand-by Power design solution



NVM merges with SRAM and logic
shutdown SRAM & registers thanks to distributed NVM

NV Flip-Flop
[N. Jovanović ISCAS 2016]



28nm FDSOI + HfO_2 based OxRAM

- **thermal stability for smart card & automotive**
- **easy integration with advance CMOS**

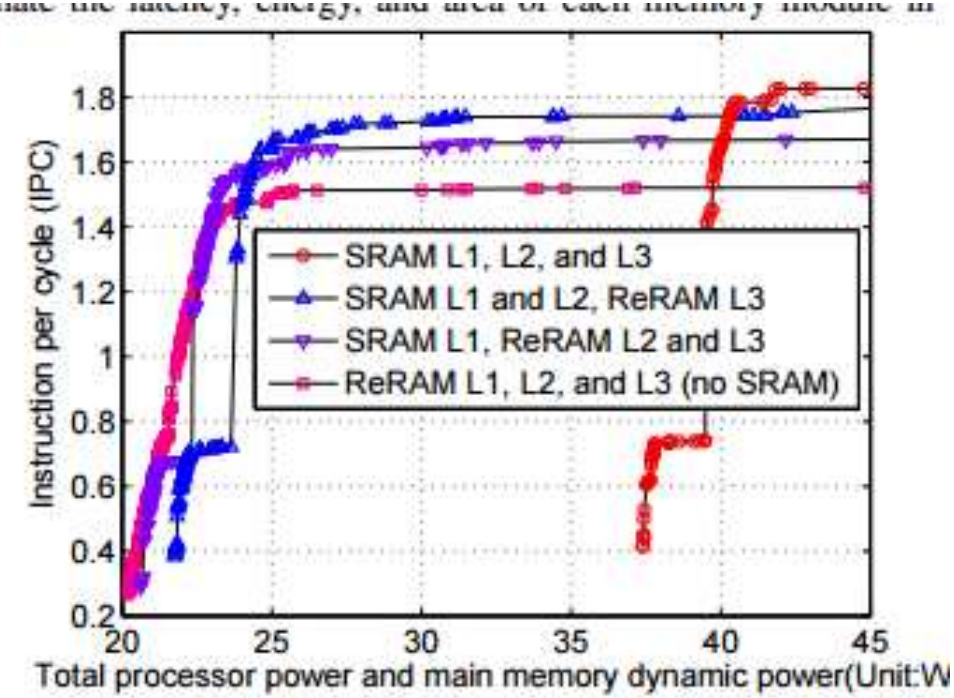


Fig. 2. Pareto curves: energy and performance trade-off under

A circuit-architecture co-optimization framework for evaluating emerging memory hierarchies

ISPASS.2013 X. Dong et al.