ADVANCED MEMORY SOLUTIONS FOR EMERGING CIRCUITS AND SYSTEMS

L. Ciampolini[1], B. Giraud[1], M. Kooli[1], A. Makosiej[1], R. Boumchedda[1,2], and J.-P. Noel[1]

[1]Univ. Grenoble Alpes, CEA, LETI, MINATEC Campus, Grenoble, France,
[2]STMicroelectronics, 850 rue Jean Monnet, 38920 Crolles, France

Acknowledgements to: A. Levisse[3], N. Gupta[4], E. Vianello[1], F. Andrieu[1], K. C. Akyel[5], M. Brocard[6]
[3]now at EPFL [4]now at Minima Processor [5]now at NanoXplore [6]now at Dolphin

OUTLINE

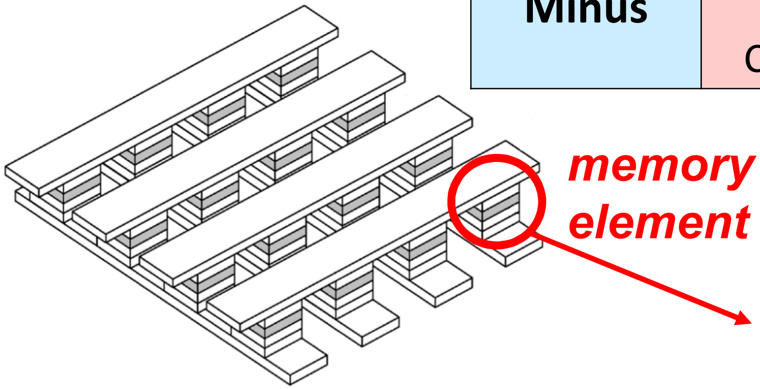
- Emerging Non-Volatile Memory Landscape
- In-Memory Computing Promises
- 4T SRAM in CoolCube
- E : IMC^3
- An SRAM-to-CAM Transformer
- Modeling 1 ppm Yield (and $) Losses in SRAM

EMERGING NON-VOLATILE MEMORY LANDSCAPE

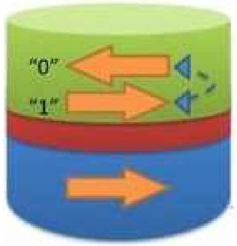
EMERGING NVM LANDSCAPE

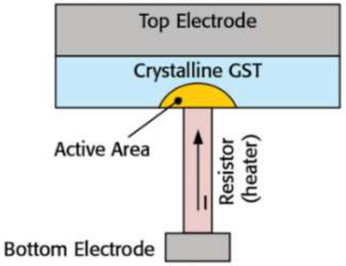
- **Which technology to replace Flash memories?**

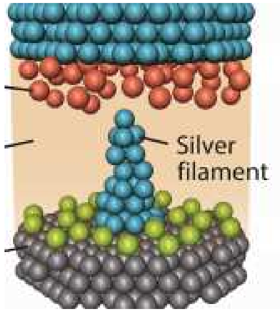
	Magneto-resistive RAM	**Phase-Change RAM**	**Resistive RAM**
Plus	Endurance	Maturity	Density, CMOS compatibility
Minus	Costly technology, Density, Read, CMOS compatibility	Consumption Thermal stability	Maturity, Forming



Crossbar memory



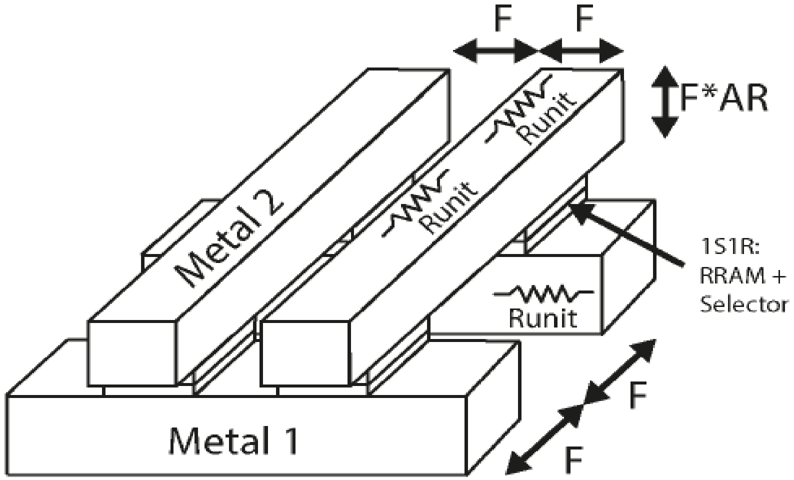




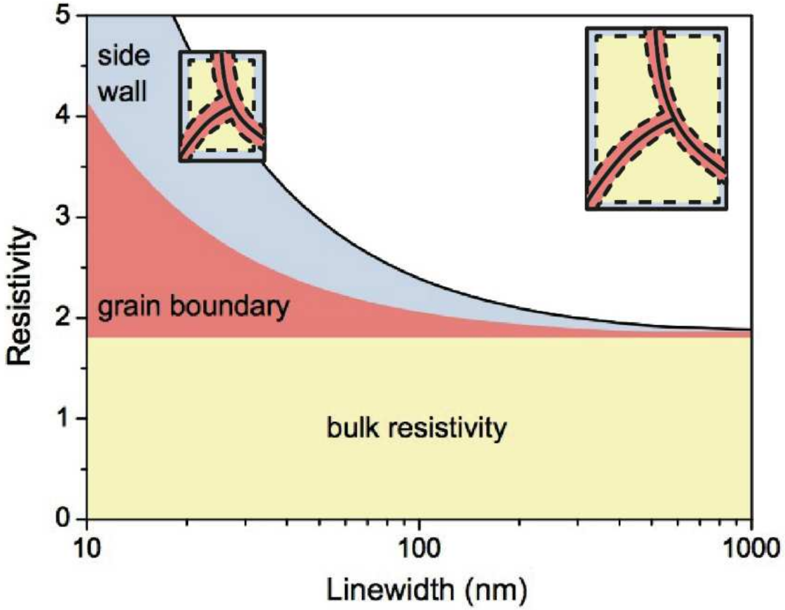
Memory function by resistance switching
Programmed with Current, Voltage and Time
BEOL process → no FEOL masks→ Low-cost solution

NVM INTERCONNECT VOLTAGE DROP

- **Unavoidable Scaling Effects:**
  - Shrink of conductive section, therefore increase of metal line resistance
  - Increase of the metal resistivity (right, data from ITRS)
- **Large-bank effect:**
  - Series connection of multiple resistances



$$R_{unit} = \sigma_{metal} * \frac{L}{W * H} \, \alpha \, \frac{\sigma_{metal}}{F}$$



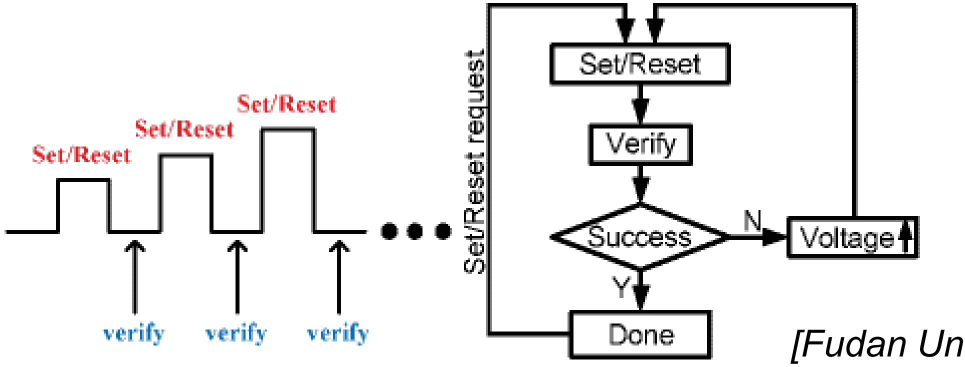
Metal resistivity becomes critical

[A. Levisse, LASCAS 2017]

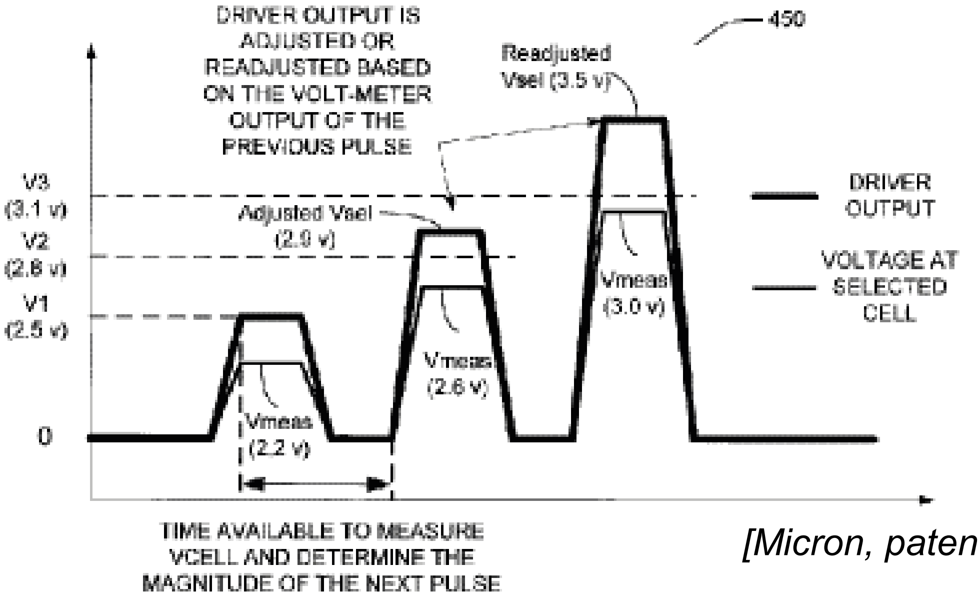
[ITRS]

NVM VOLTAGE DROP COMPENSATION

- **Standard approach by trial-and-error**



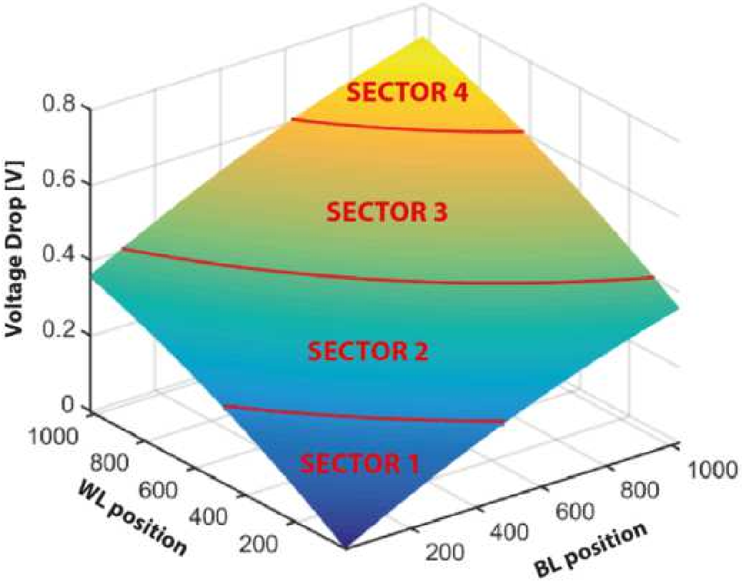
[Fudan University, ISCAS 2015]


DRIVER OUTPUT IS ADJUSTED OR READJUSTED BASED ON THE VOLT-METER OUTPUT OF THE PREVIOUS PULSE
450
Readjusted Vsel (3.5 v)
V3 (3.1 v)
V2 (2.8 v)
V1 (2.5 v)
0
Adjusted Vsel (2.9 v)
Vmeas (3.0 v)
Vmeas (2.6 v)
Vmeas (2.2 v)
DRIVER OUTPUT
VOLTAGE AT SELECTED CELL
TIME AVAILABLE TO MEASURE VCELL AND DETERMINE THE MAGNITUDE OF THE NEXT PULSE

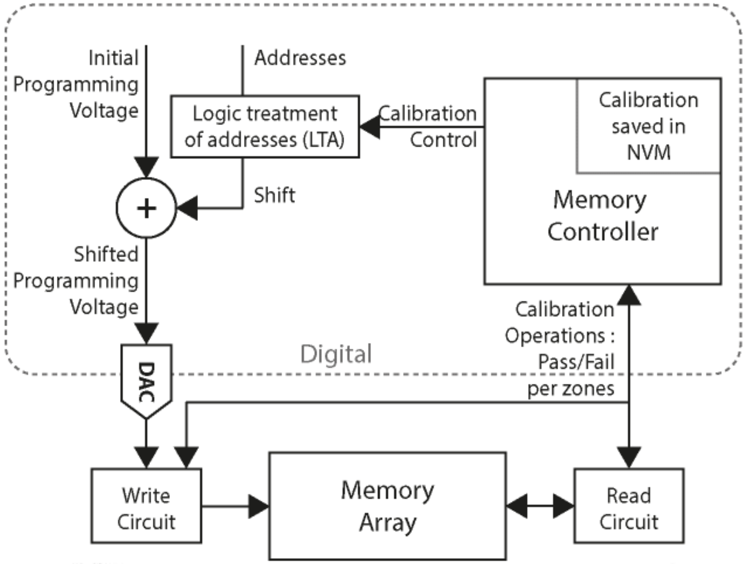
[Micron, patent filed 2015]

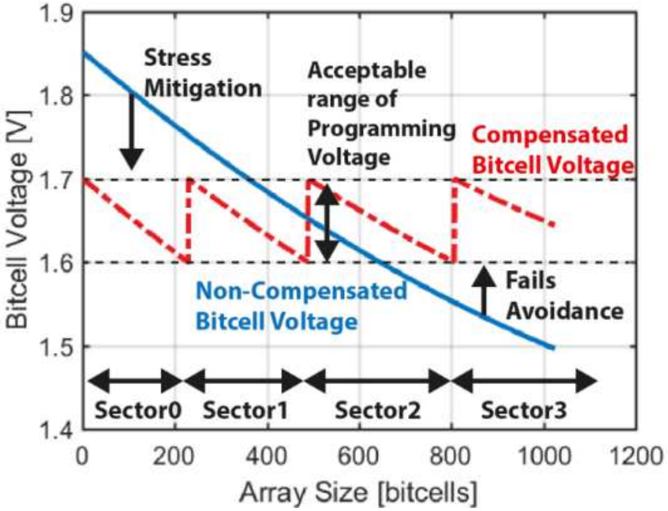
NVM VOLTAGE DROP COMPENSATION

- **Proposed solution takes into account that voltage drop is tied to the cell position**



[A. Levisse et al., NVMTS 2015]


Initial Programming Voltage
Addresses
Logic treatment of addresses (LTA)
Calibration Control
Calibration saved in NVM
Memory Controller
+
Shift
Shifted Programming Voltage
Digital
Calibration Operations : Pass/Fail per zones
DAC
Write Circuit
Memory Array
Read Circuit

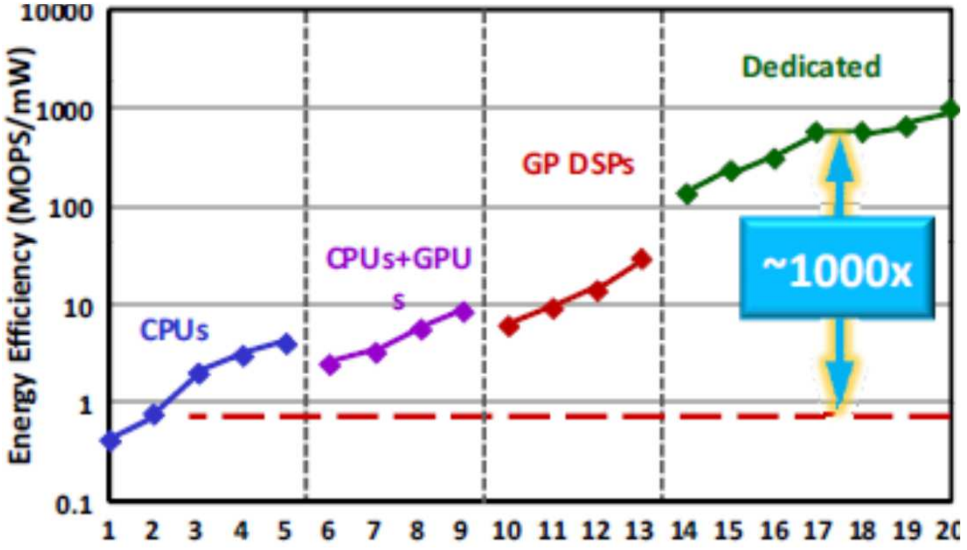


Efficient Calibration, Compatible with other compensation techniques

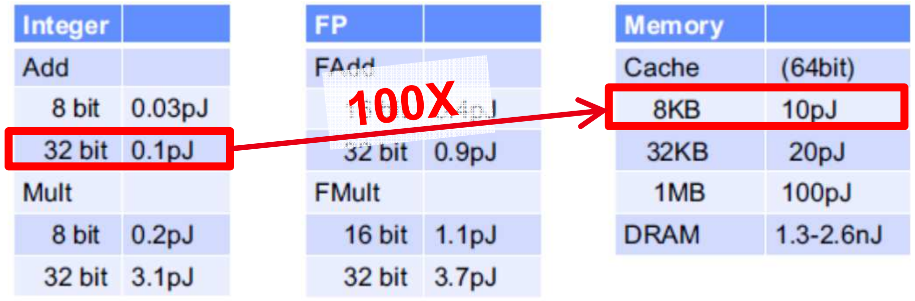
IN-MEMORY COMPUTING PROMISES

IN-MEMORY COMPUTING

- **The largest part of power consumption of logic and arithmetic operations in some kinds of ICs is due to the memory access**



[M. Horowitz (Stanford), ISSCC 2014]

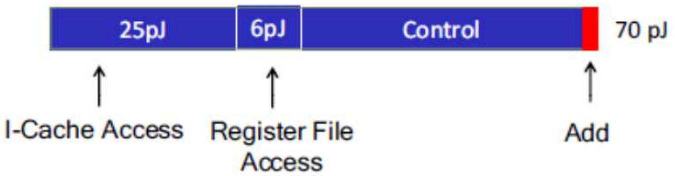


Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ

Instruction Energy Breakdown



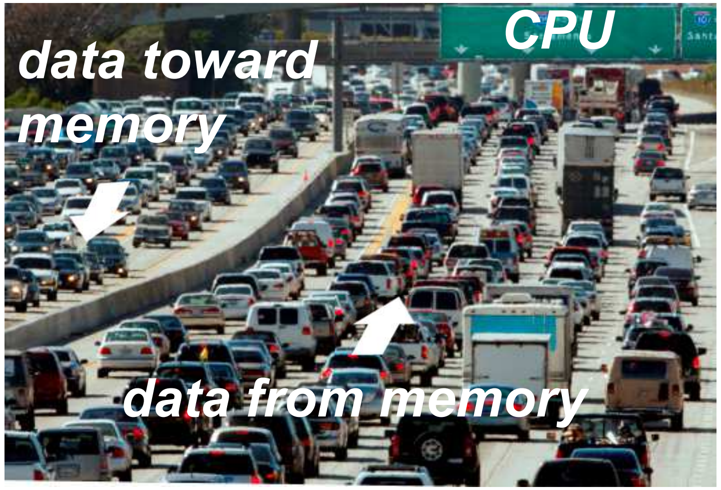
logic/arith. op. vs memory access energy

→ **100-1000X**

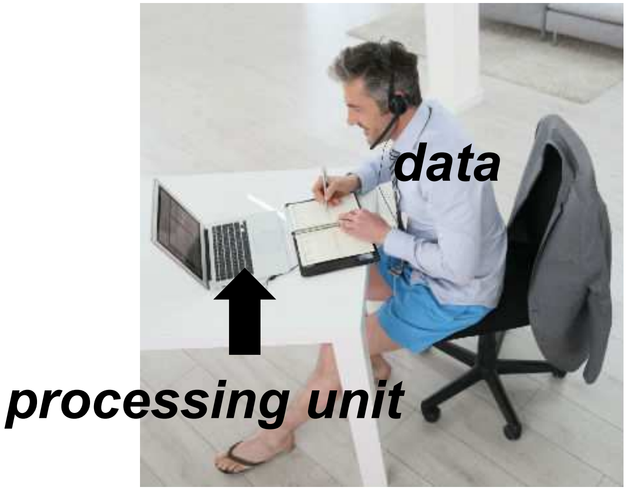
IN-MEMORY COMPUTING

- *In-Memory Computing* (IMC) consists in performing computation tasks where the data is stored, *i.e.* in memories, to counter the heavy data traffic between CPU and cache

Today…



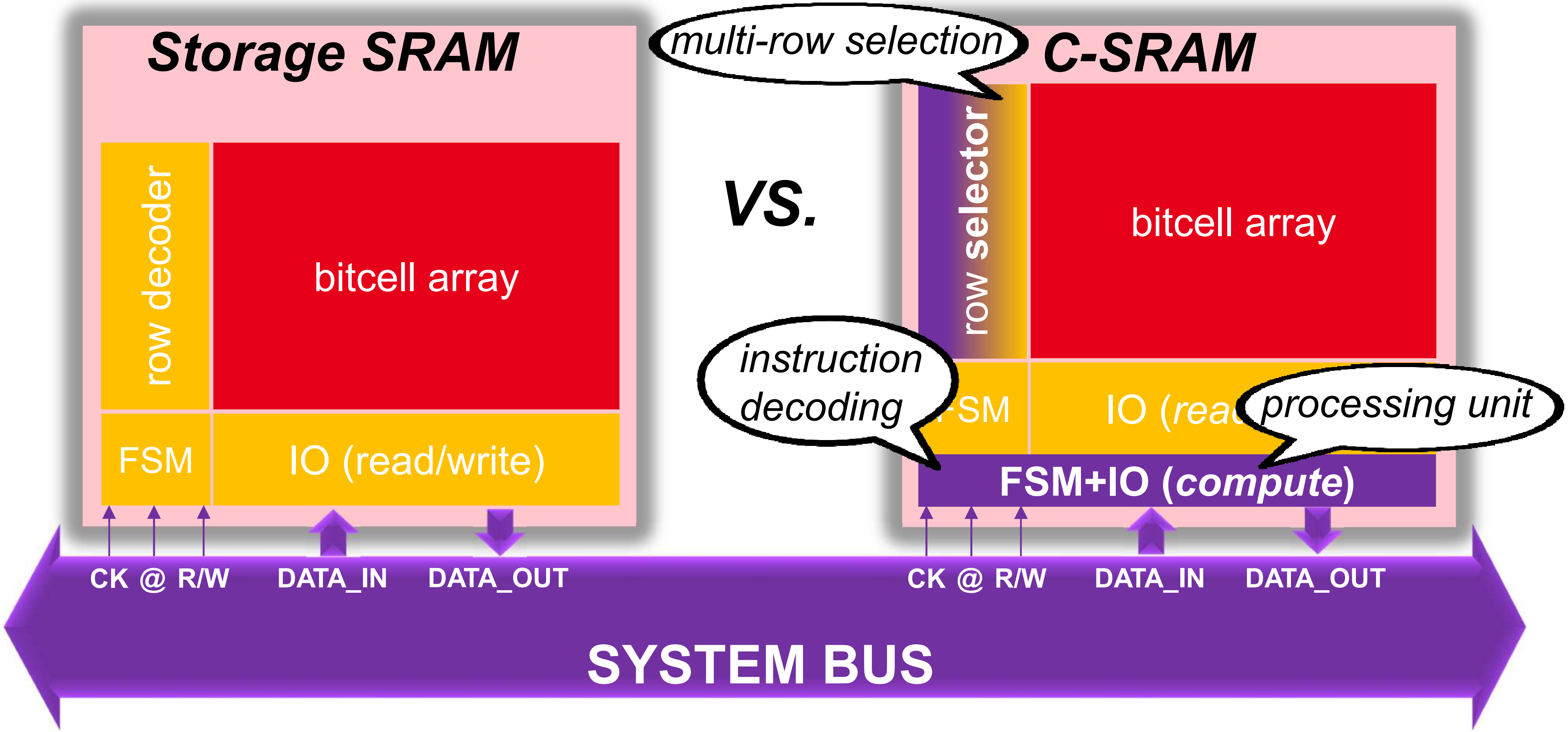
…tomorrow with IMC!



- This solution makes sense when processing data in the CPU becomes very heavy or inadequate (*data-centric apps, AI…*)

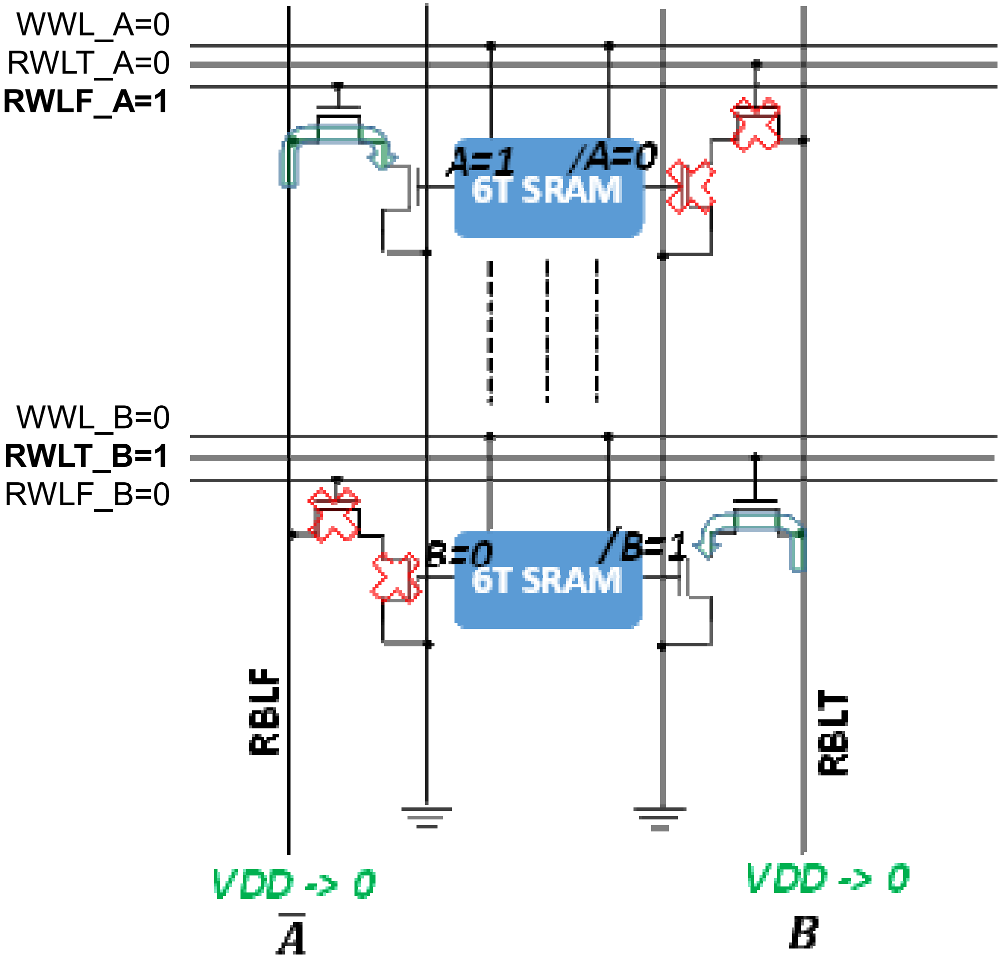
IMC COMPUTATIONAL MEMORIES

- A Computational SRAM (C-SRAM) **executes *in-situ* micro-instructions**



IMC : HOW DOES IT WORK?

- **Conventional 2R operations of a 10T, three-port (1RW2R) SRAM**



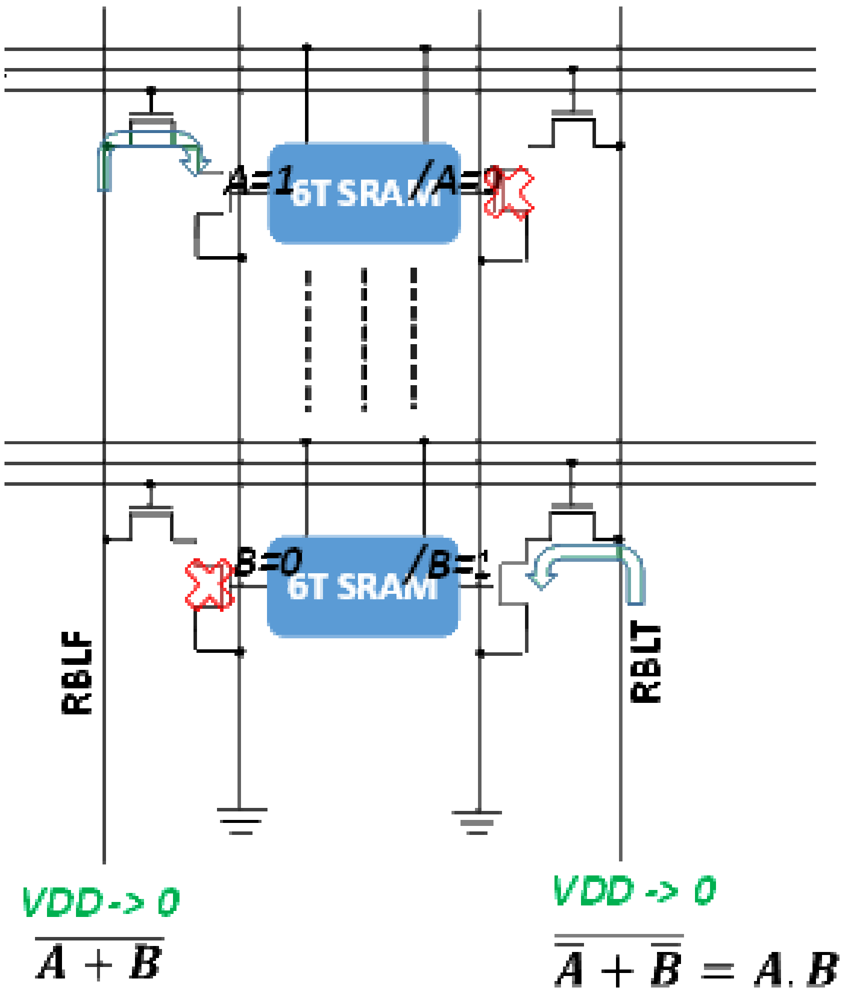
[K. Akyel et al. , ICRC 2016]

MULTI-ROW SELECTION

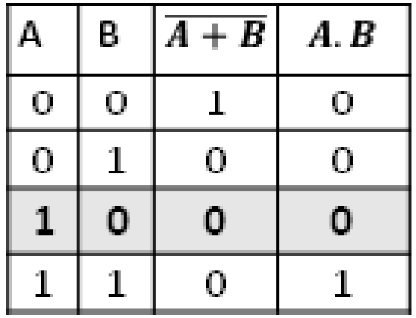
- **Multi-row selection yields a Boolean function of data**

WWL_A=0
RWLT_A=1
RWLF_A=1

WWL_B=0
RWLT_B=1
RWLF_B=1



A=1
6T SRAM
/A=0
B=0
6T SRAM
/B=1
RBLF
RBLT
VDD-> 0
$\overline{A+B}$
VDD -> 0
$\overline{\overline{A}+\overline{B}} = A.B$

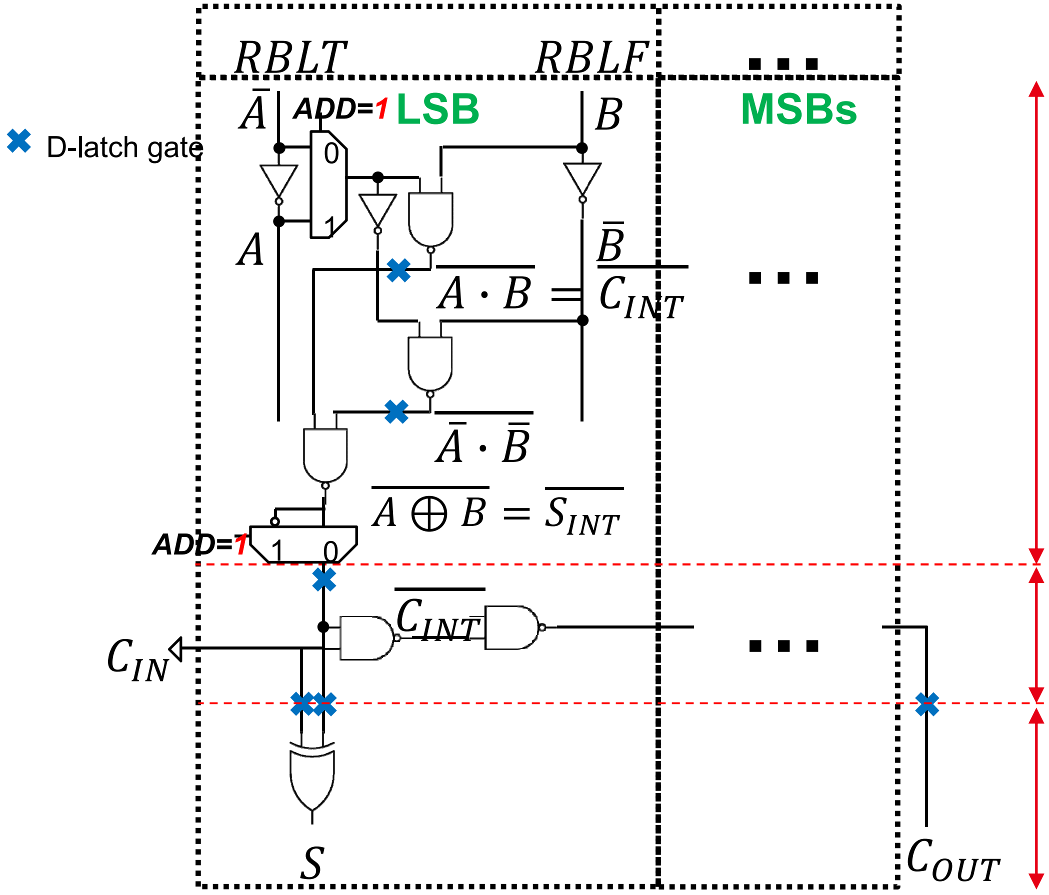


A	B	$\overline{A+B}$	$A.B$
0	0	1	0
0	1	0	0
1	0	0	0
1	1	0	1

[K. Akyel et al. , ICRC 2016]

EXPECTED GAIN: EVALUATION MODEL

- **Boolean functions of data are the bricks to obtain additions, subtractions, multiplications**



1st cycle
Determine, at bit level, the level between S_{INT} and C_{INT} (half adder outputs)

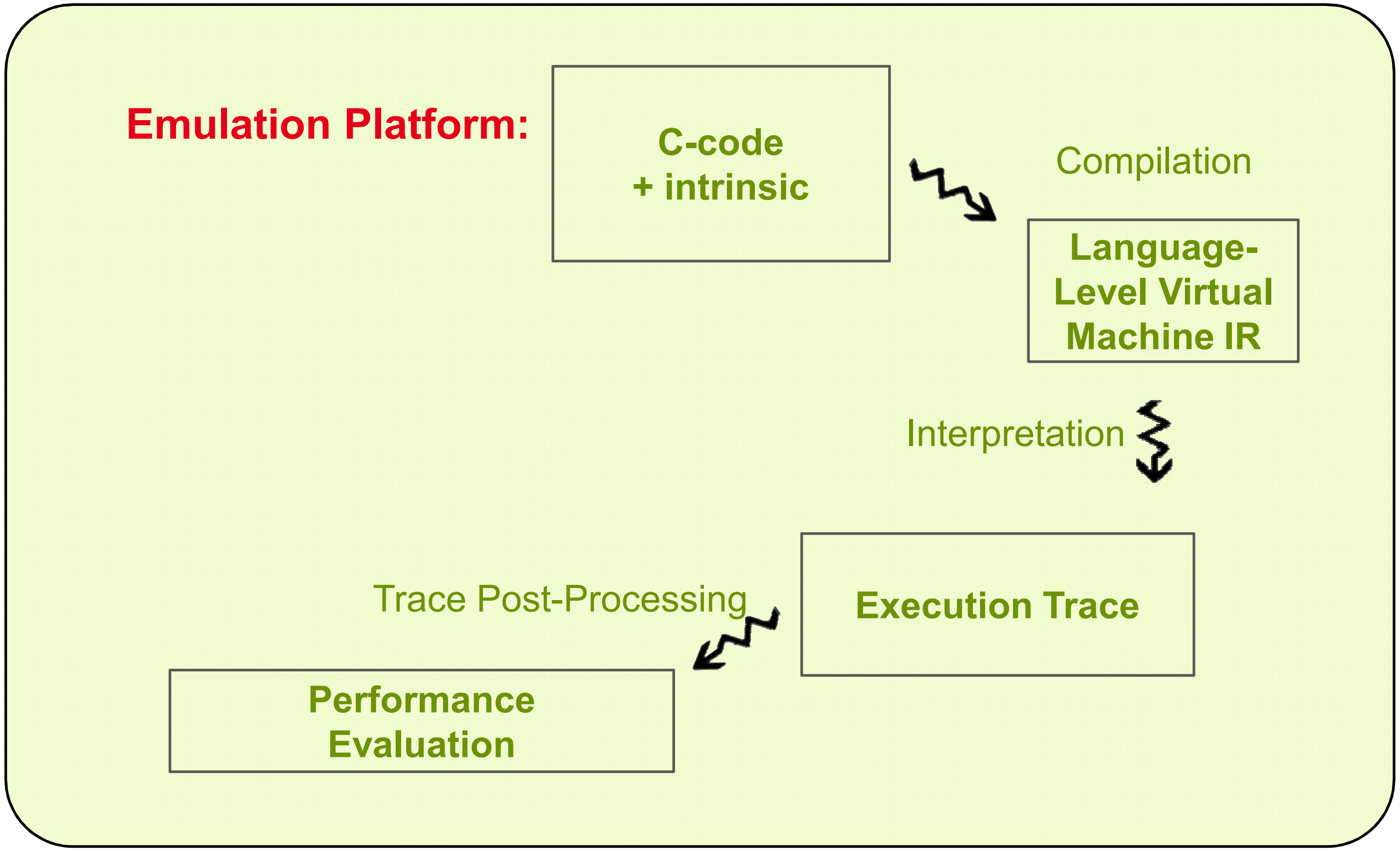
2nd cycle
Propagate computed carry at each bit level

3rd cycle
Compute the final result (S+C_{OUT})

[K. Akyel et al. , ICRC 2016]

EXPECTED GAIN: EVALUATION MODEL

- The current emulation platform (IMPACT) allows to roughly estimate the benefits of using IMC operations instead of standard ALU operations



[M. Kooli et al., DATE 2018]

EXPECTED GAIN: PRELIMINARY RESULTS

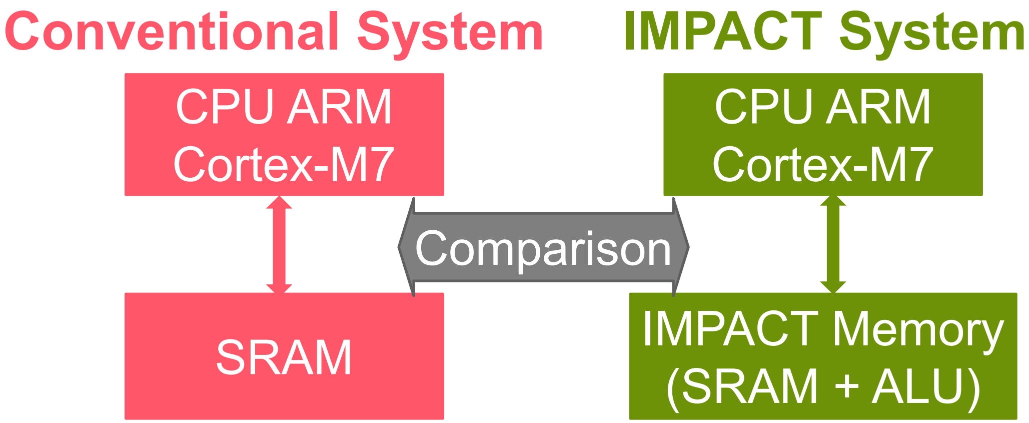
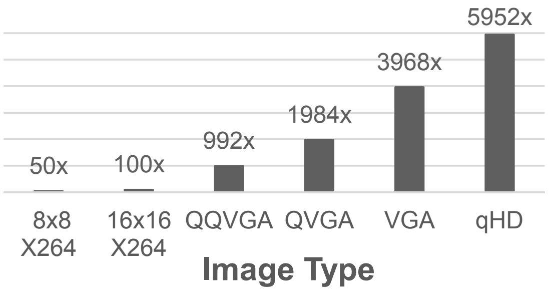
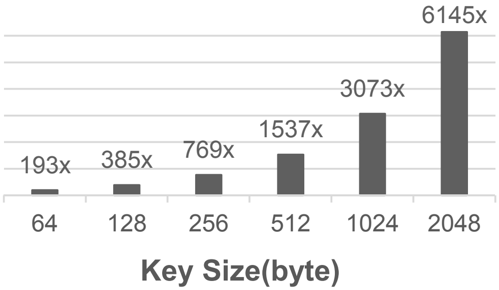


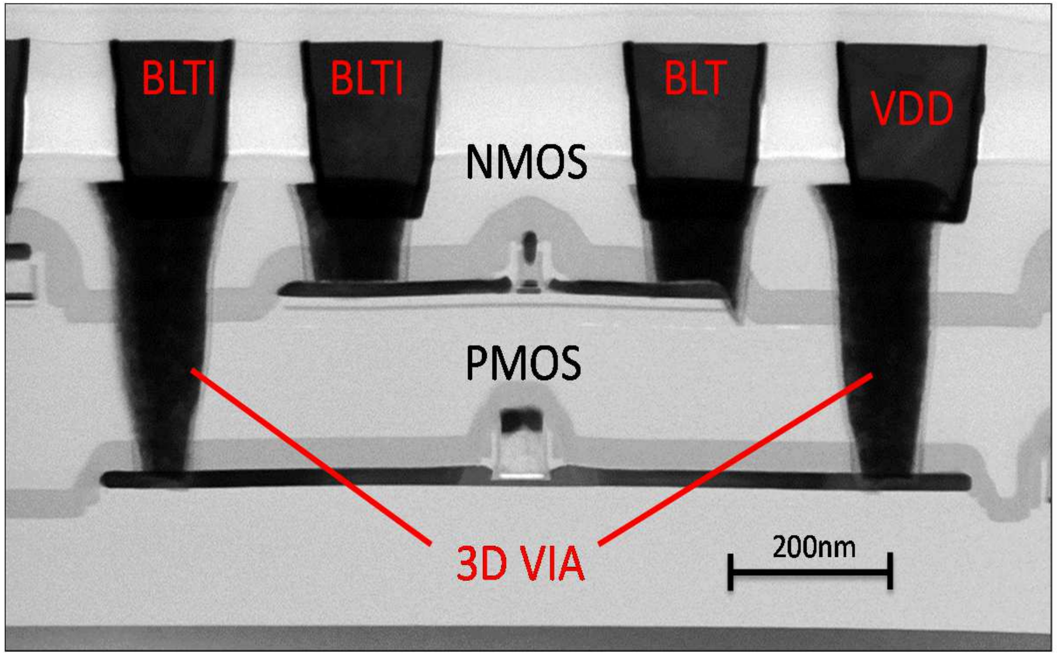
	Image Processing	**Cryptography**
Application	**Motion detection algorithm:** (subtraction between two images)	**One Time Pad**: (bitwise between message and key)
Timing Evaluation **Speed Factor** **(x nb of cycles)**		
Energy Consumption	12,4x reduction	12,9x reduction

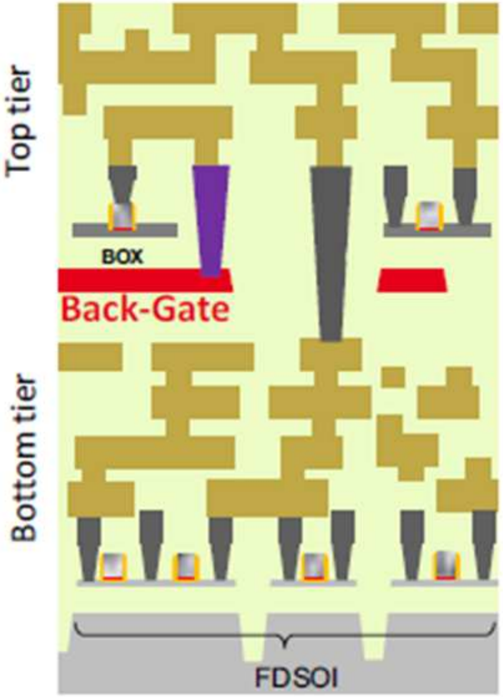
[M. Kooli et al., RSP 2017]

4T SRAM IN MONOLITHIC 3D COOLCUBE TECHNOLOGY

COOLCUBE^{TM}: LETI'S MONOLITHIC 3D

- Monolithic 3D consists in manufacturing a second layer (Top tier) of active MOSFETS over a first layer (Bottom tier) where active MOSFETS already exist

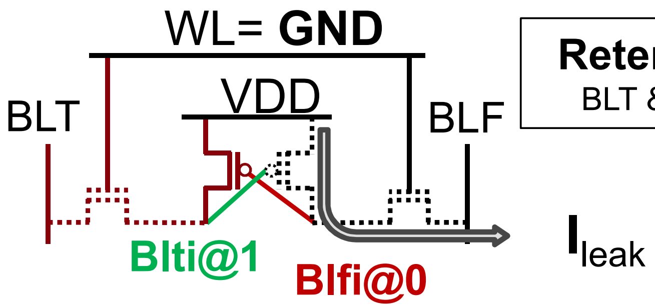




[F. Andrieu et al., IEDM 2017]

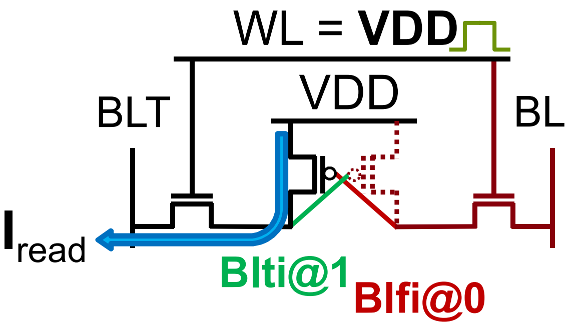
4T SRAM IN 3D COOLCUBE^{TM} TECHNOLOGY

- **4T Driver-Less SRAM bitcell:**



Retention mode
BLT & BLF @ GND

Stability issue: Voltage divider causes Blfi to raise in a time dependent on leakage intensity



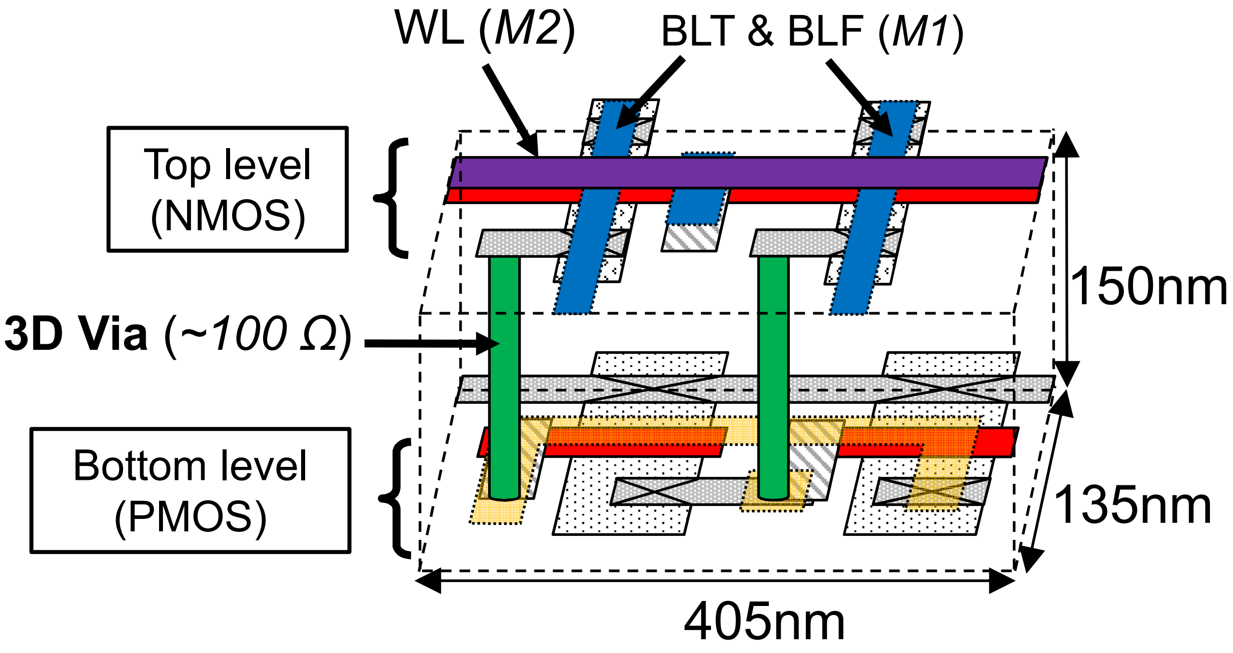
Read mode
BLT & BLF Floating

The GND-precharged bitline connected to the node storing 1 (here Blt) raises in a time dependent on read current. *Stability issue: Voltage divider causes Blti to drop.*

- **Stability depends on the PMOS/NMOS threshold voltage gap ΔV_{th}**

4T SRAM IN 3D COOLCUBE^{TM} TECHNOLOGY

- **Design split across tiers:**

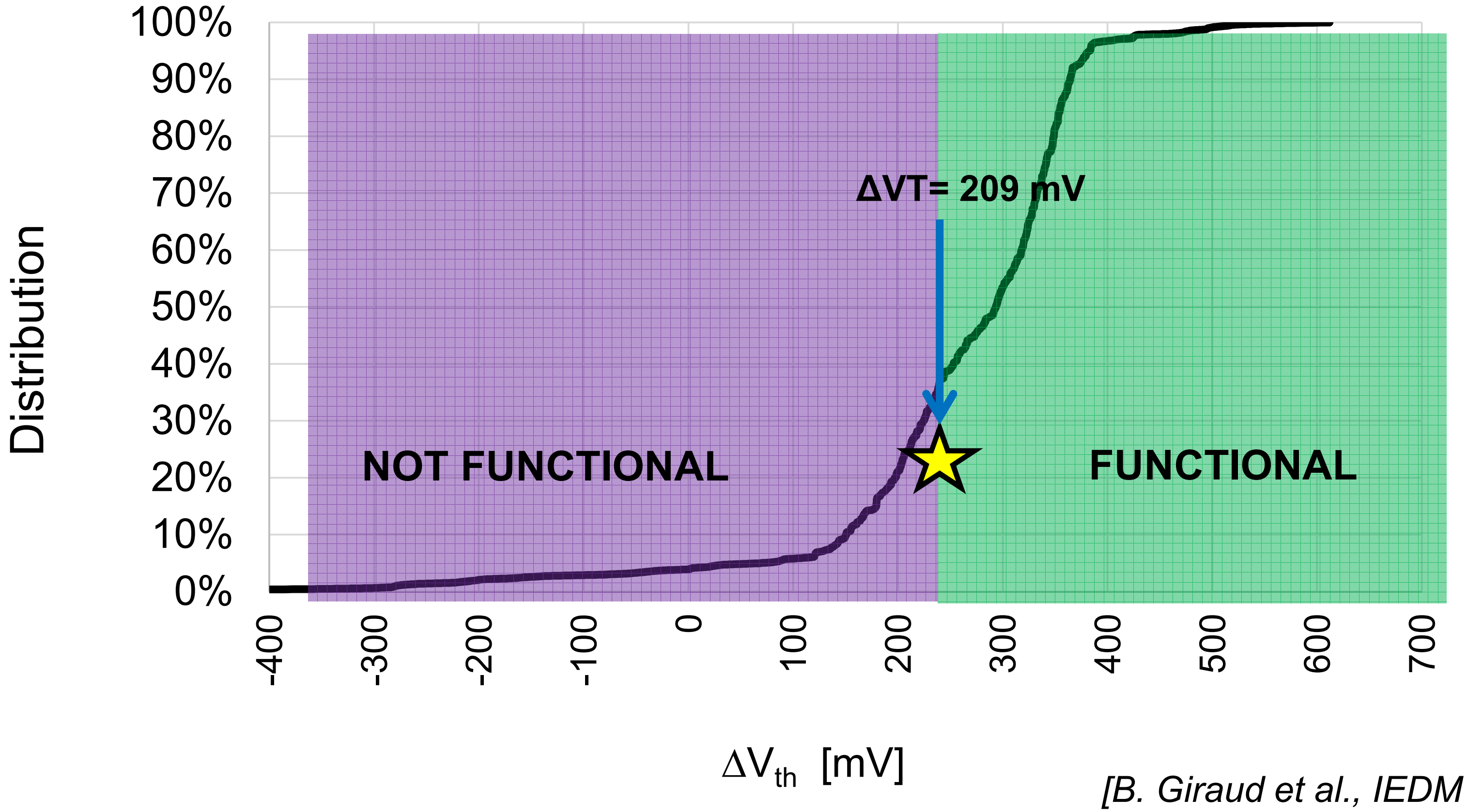


Bitcell area: **0.054µm² (-30% versus SPHD)**
Manufactured on both tiers in former STM **14nm FD-SOI**

[M. Brocard et al., S3S, 2016]

4T SRAM IN 3D COOLCUBE^{TM} TECHNOLOGY

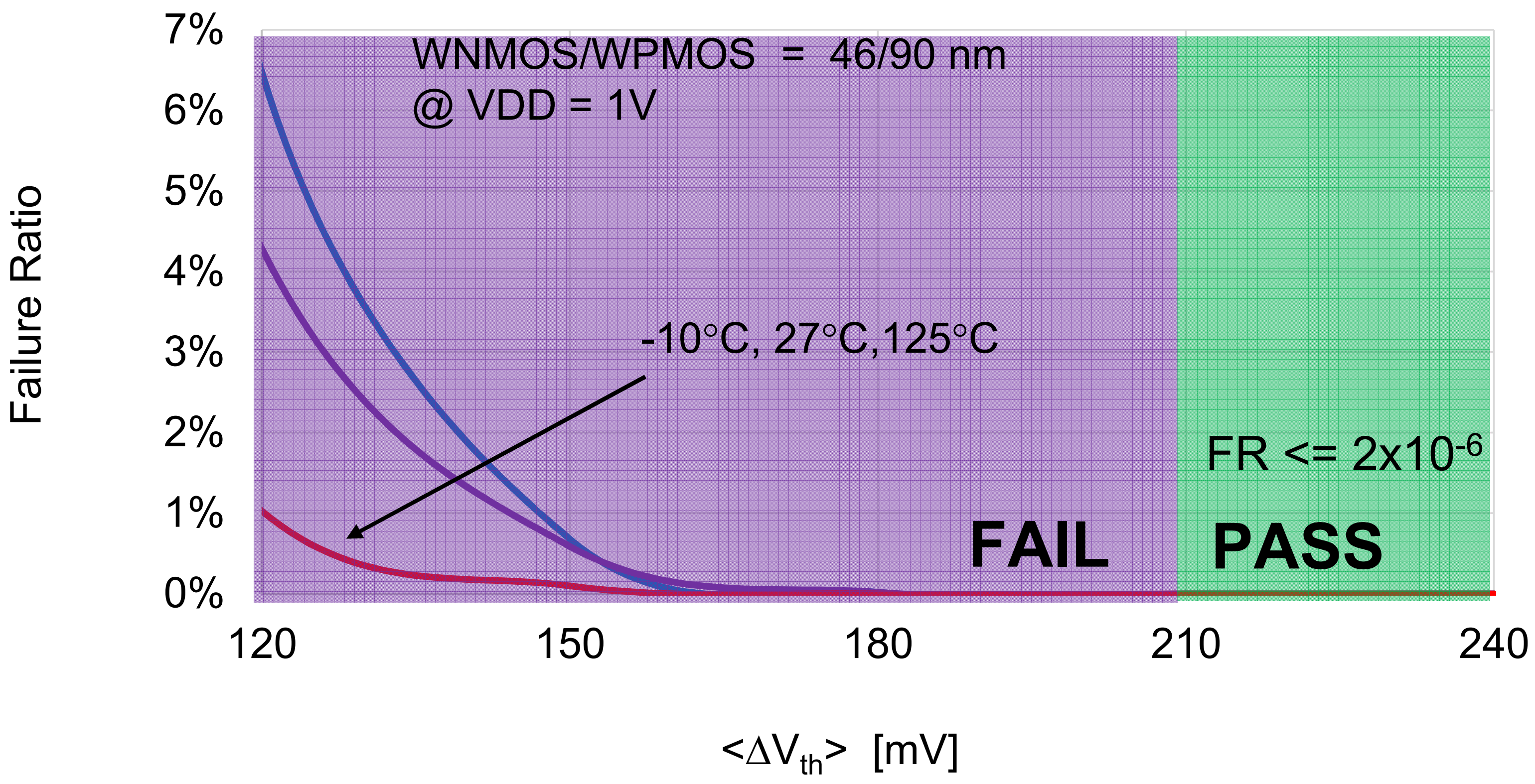
- **Device threshold voltage gap ΔV_{th} = ($V_{th\ PMOS}$ - $V_{th\ NMOS}$) distribution from 507 pairs available in a wafer manufactured at 2014 (non-mature process)**



[B. Giraud et al., IEDM 2017]

4T SRAM IN 3D COOLCUBE^{TM} TECHNOLOGY

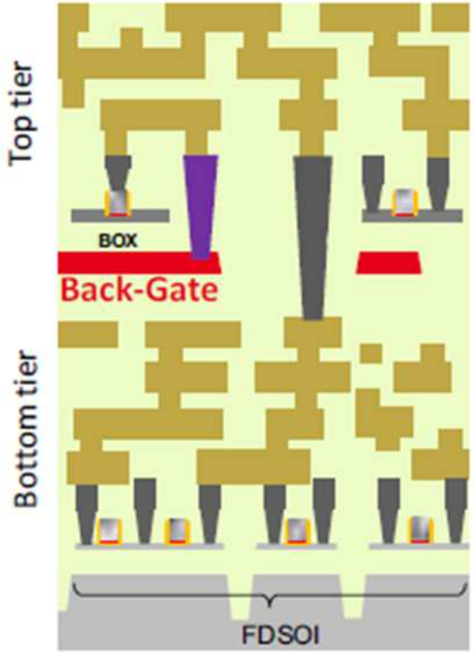
- **The Functional/ Non Functional regions have been found through Spice MC simulations around a variable, mean device threshold voltage gap $<\Delta V_{th}>$**

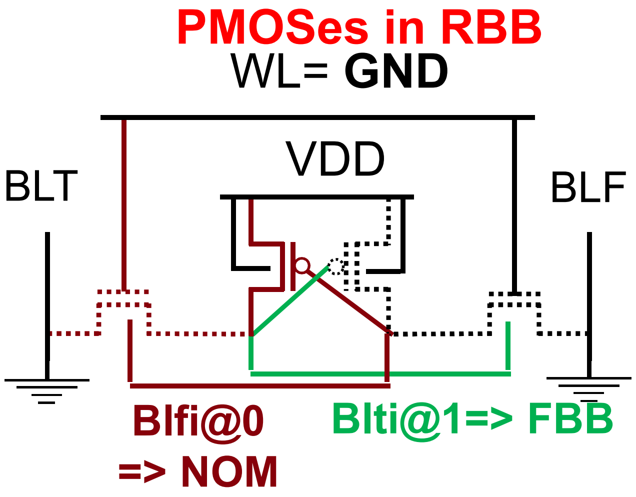


[B. Giraud et al., IEDM 2017]

4T SRAM IN 3D COOLCUBE™ TECHNOLOGY

- **The requirements over the mean device threshold voltage gap $<\Delta V_{th}>$ can be relaxed by using Data – Dependent Back-Biasing :**
  - PMOS V_{th} is increased statically
  - Top-Tier NMOS V_{th} is modified dynamically and dependent on the stored value

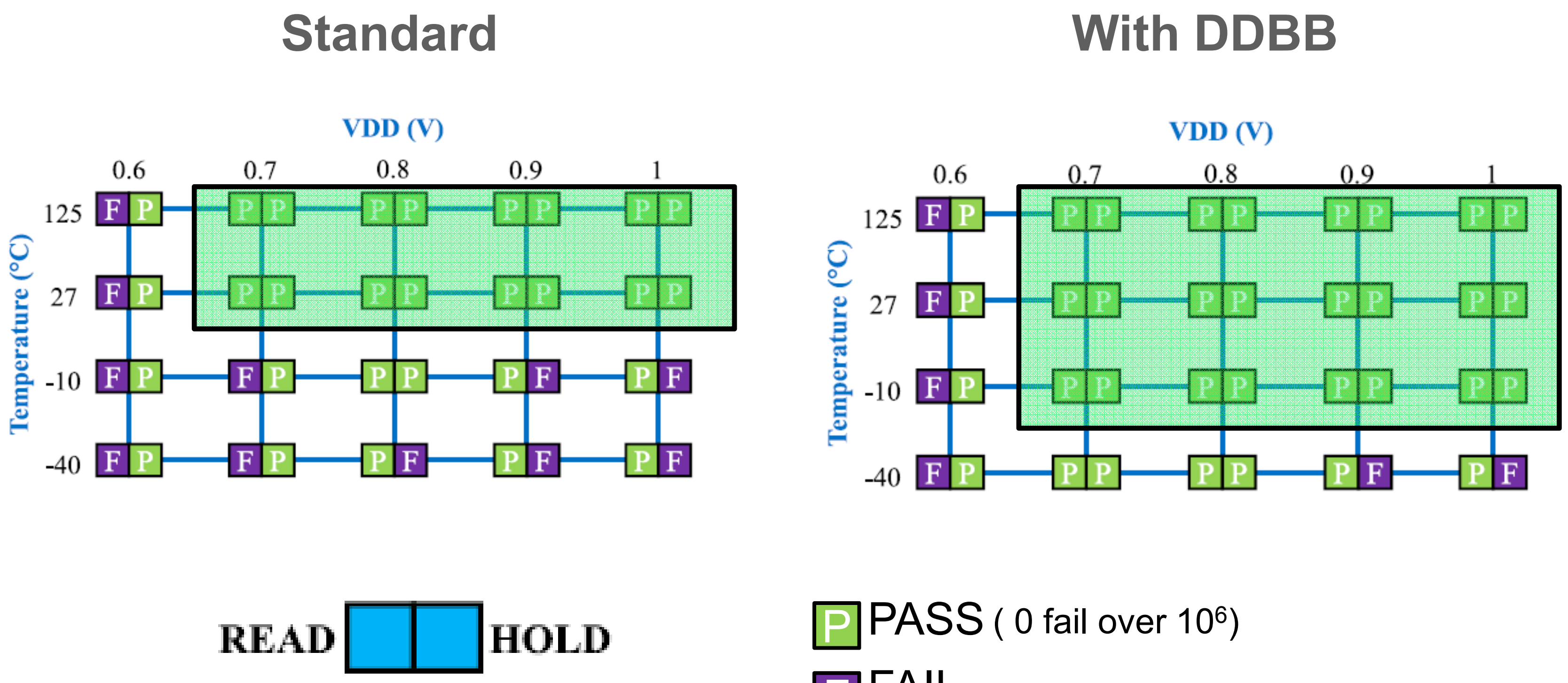




[R. Boumchedda et al., TVLSI, 2017]

4T SRAM IN 3D COOLCUBE^{TM} TECHNOLOGY

- **Spice MC analysis of DBB effect on single-cut functionality:**
  - 32 bitcell / column
  - ΔV_{th} =180 mV @ T = -10°C
  - Process = TT



Standard

Temperature (°C) \ VDD (V)	0.6	0.7	0.8	0.9	1
125	F P	P P	P P	P P	P P
27	F P	P P	P P	P P	P P
-10	F P	F P	P P	P F	P F
-40	F P	F P	P F	P F	P F

With DDBB

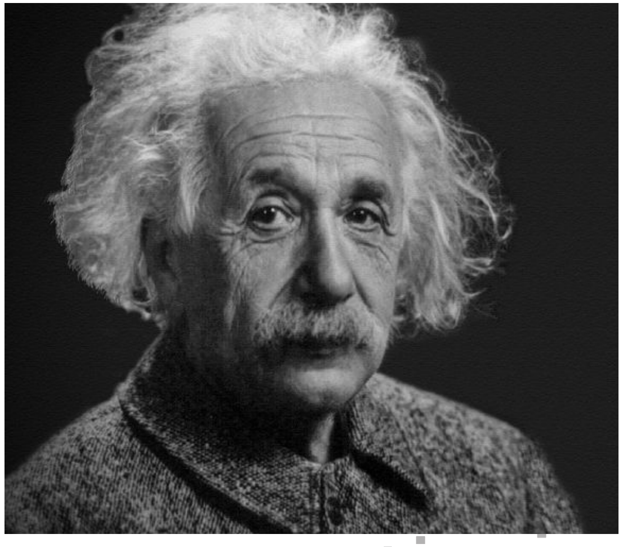
Temperature (°C) \ VDD (V)	0.6	0.7	0.8	0.9	1
125	F P	P P	P P	P P	P P
27	F P	P P	P P	P P	P P
-10	F P	P P	P P	P P	P P
-40	F P	P P	P P	P F	P F

P PASS (0 fail over 10^6)

F FAIL

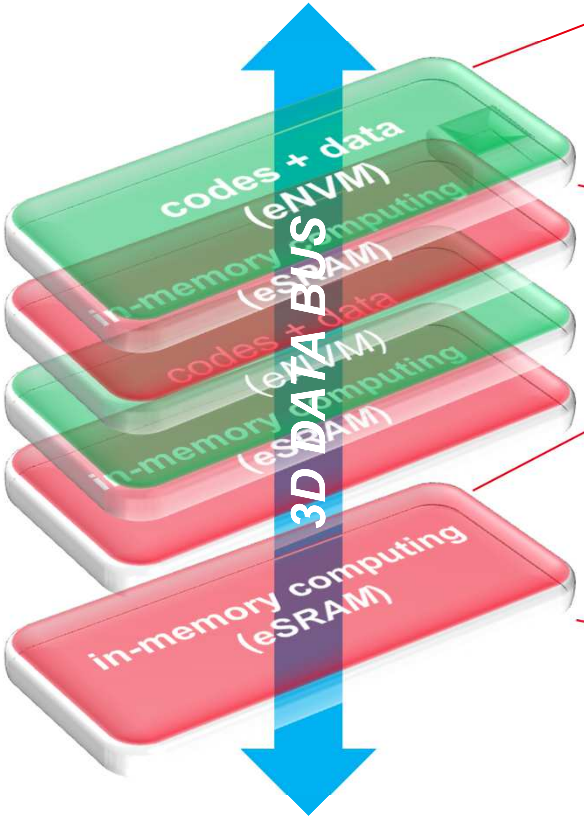
[R. Boumchedda et al., TVLSI, 2017]

E*[XPECTING]* : IMC^3

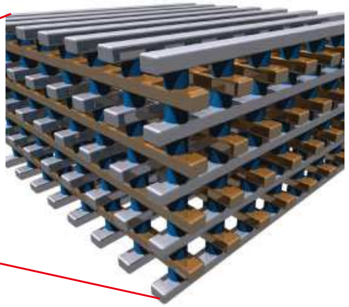


LOOKING TO THE FUTURE: IMC^3?

- **Memory Computing ultimate unification…**



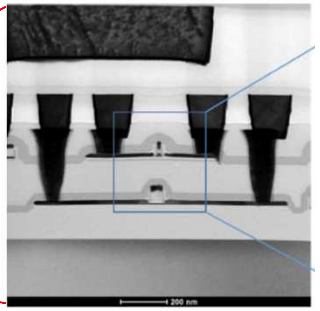
ReRAM technology (*OxRAM*, *PCM*, …)



CEA embedded NVM:

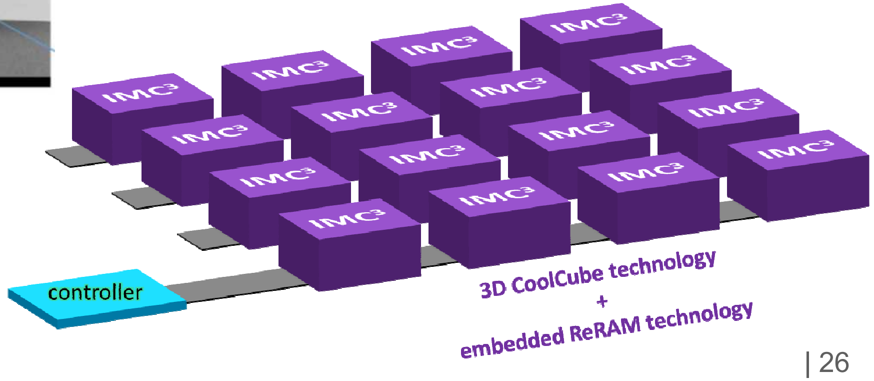
- *high density*
- *fast access/low voltage*
- *CMOS compatible*

3D CoolCube^{TM} technology



CEA manufacturing process:

- *transistor layer stacking*
- *high density 3D interconnections*



AN SRAM-TO-CAM TRANSFORMER

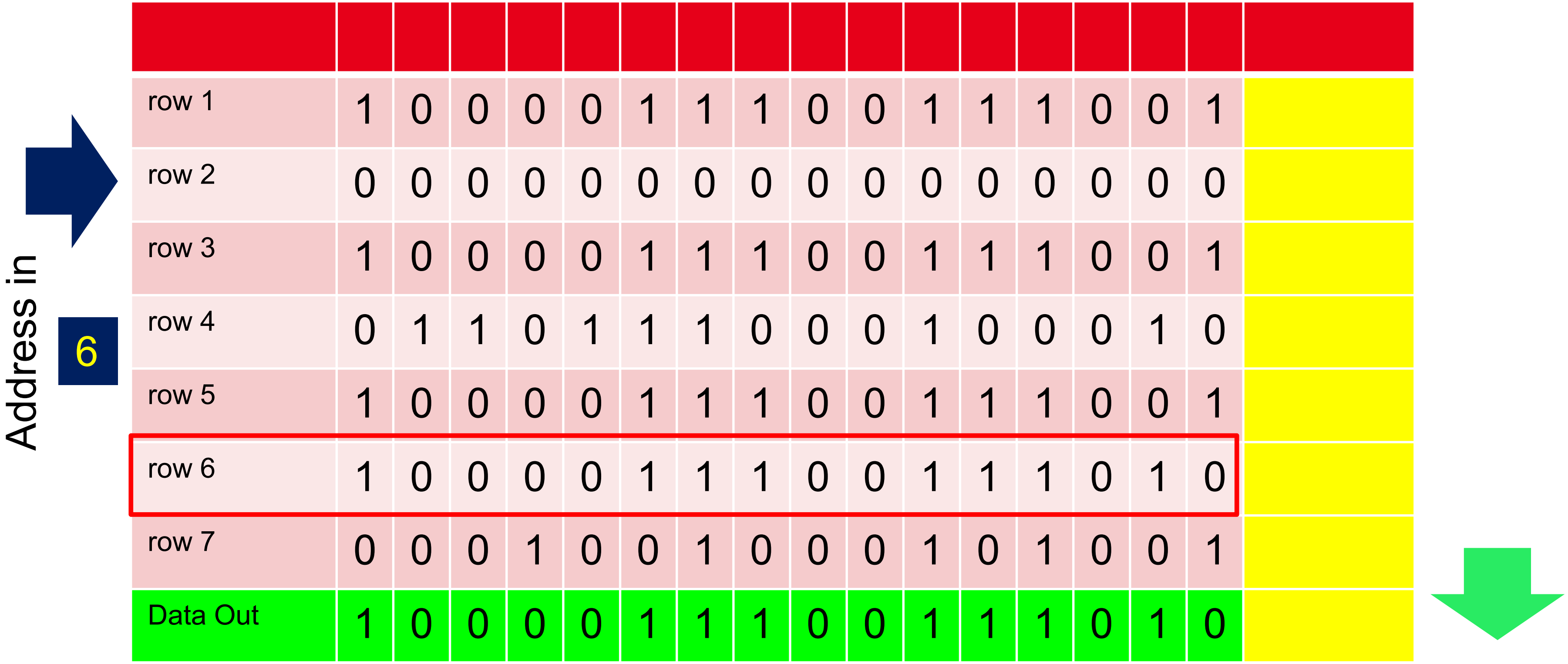
CAM-SENSIBLE MARKETS

- CAMs are Content-Addressable Memory that are able to search quickly for a particular stored key
- Typical application is searching internet addresses in huge tables in router devices

- If by any chance you know about possible other companies that could be interested about searching data quickly…

A SEARCH VIEW OF SRAM

- Standard memory (SRAM/DRAM) contents are indexed by a numerical key: the memory address (~ row number)
- A memory readout provides the word stored at the correspondent location (one-word hit of the address search)



row 1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	
row 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
row 3	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	
row 4	0	1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	
row 5	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	
row 6	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	0	
row 7	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	
Data Out	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	0	

CONTENT-ADDRESSABLE MEMORY

- A CAM allows a one-cycle search of a given search key amongst all stored words
- A memory readout provides the address where the word is stored (one-word hit of the content search)
- If multiple hits, one might be chosen

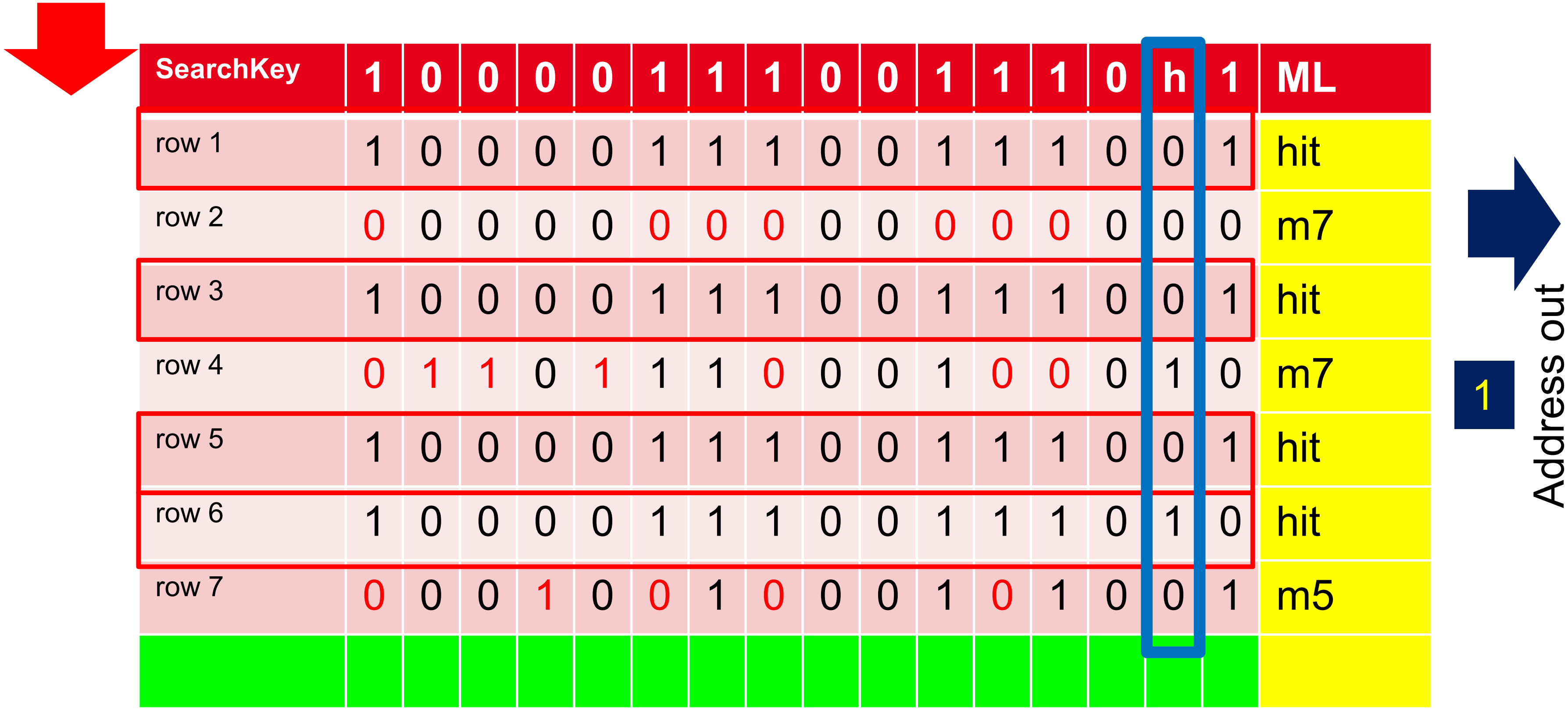
SearchKey	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	ML
row 1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	m7
row 3	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 4	0	1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	m8
row 5	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 6	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	0	m1
row 7	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	m5

1

Address out

CONTENT-ADDRESSABLE MEMORIES

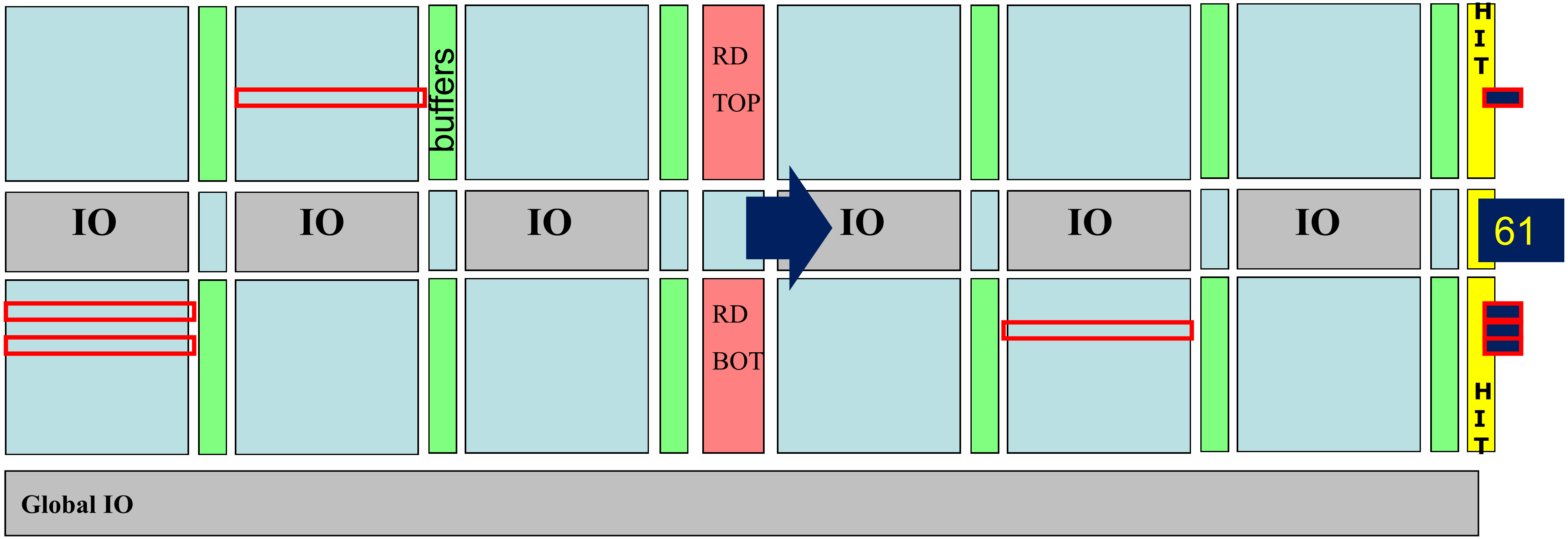
- Some key parts can be masked (e.g. search for “Albert *instein”)
- We say that the correspondents key bits are “h” meaning that they are in always-hit



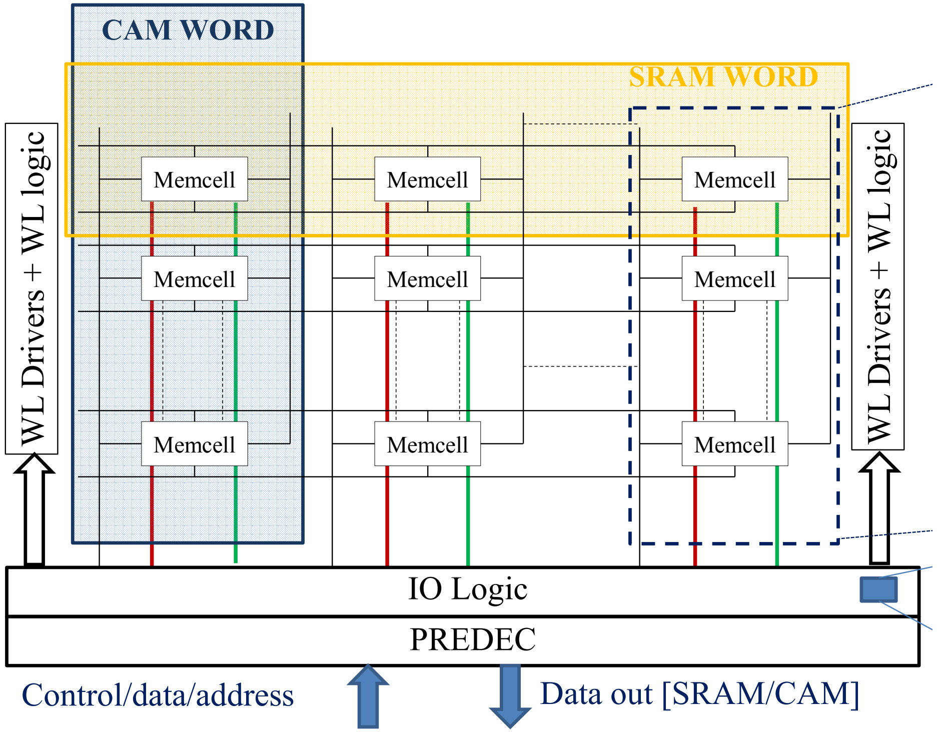
SearchKey	1	0	0	0	0	1	1	1	0	0	1	1	1	0	h	1	ML
row 1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	m7
row 3	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 4	0	1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	m7
row 5	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	1	hit
row 6	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	0	hit
row 7	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	m5

CONTENT-ADDRESSABLE MEMORIES

- CAM designs might dissipate enormous amounts of power due to large capacity, hierarchical architecture, high-perf requirements and parallel operations on all rows



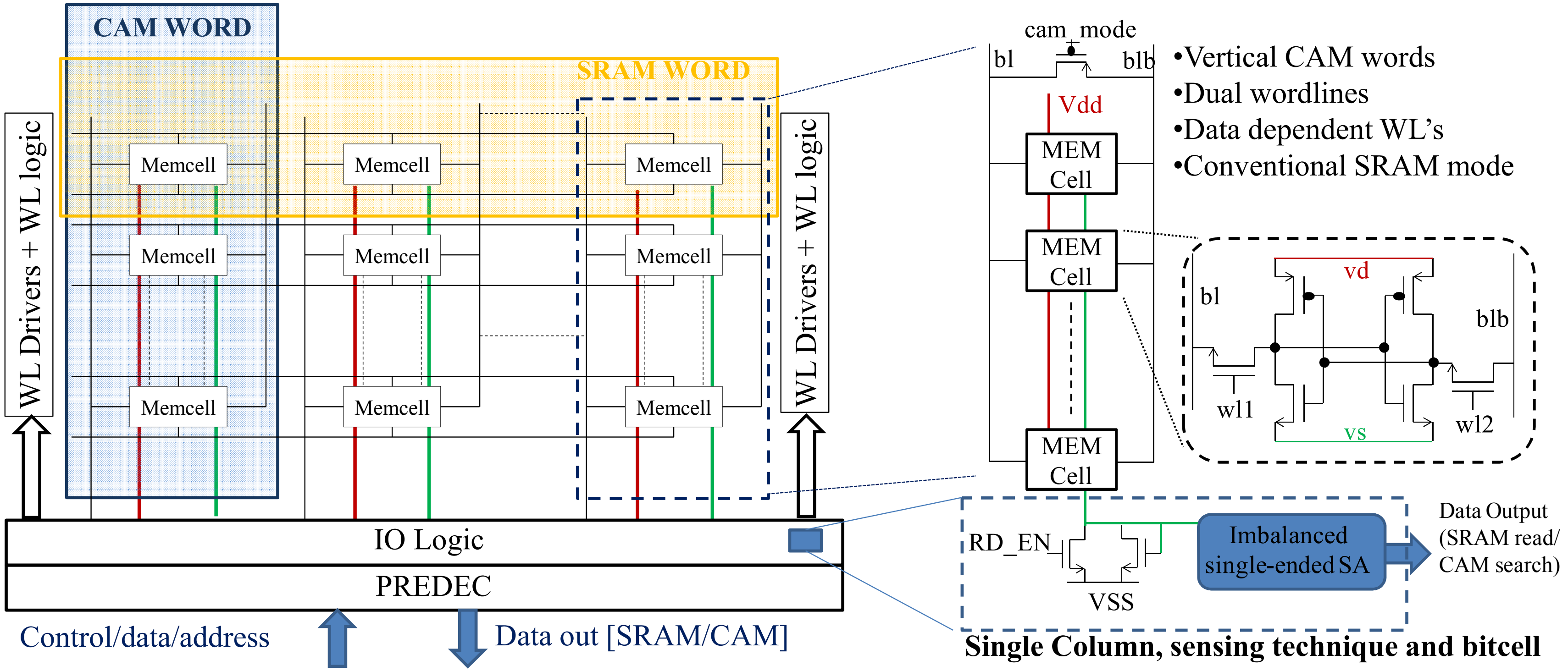
RECONFIGURABLE SRAM/CAM ARCHITECTURE



- **Key features:**
  - Use only SRAM cells (large density)
  - Allow R/W SRAM operations on horizontal words
  - Allow CAM operations on vertical words
  - This is obtained by routing 2 WLs per row, and with additional digital/Row Dec circtuitry

[N. Gupta et al., ESSCIRC, 2017]

RECONFIGURABLE SRAM/CAM ARCHITECTURE



- Vertical CAM words
- Dual wordlines
- Data dependent WL's
- Conventional SRAM mode

Single Column, sensing technique and bitcell

- **Key features:**
  - Read on ground-line for both CAM and SRAM
  - SA on Vss is single-ended, imbalanced

- CAM/SRAM read : single WL
- CAM Write : single-ended SRAM write
- SRAM Write : standard with differential BL's

[N. Gupta et al., ESSCIRC, 2017]

COMPARISON WITH PREVIOUS WORKS

	This work	[1]	[2]	[3]	[4]
Technology	28nm FDSOI	28nm FDSOI	32nm	65nm	0.13μm
Transistors/cell	6T	6T	11T	10T	9T+Read
Area/cell [μm2]	0.197μm^{2} $^{\alpha}$	0.152μm^{2}	-	3.3	20
Array Size	128x64	64x64	(64x64) *4	128x128	128x32
Frequency (VDD)	**1.56 GHz@0V9** $^{\beta}$ **8.9MHz@0V38** $^{\gamma}$	370 MHz (1V)		500MHz (1V)	250MHz (1V)
Energy/Search/ bit [fJ]	**0.13 (0.9V)**	0.6 (1V) 0.41 (0.75V)	1.07 (1V) 0.3 (0.5V)	0.77 (1.2V)	1.87 (1V)
Match-line Technique	1-Single-ended imbalanced SA	2-Single Ended SA	Wide AND	NOR	Differential
Memory Modes	BCAM/SRAM/ Pseudo-TCAM	BCAM/ TCAM/SRAM	BCAM	BCAM	BCAM

[1] Jeloka, S. et al. VLSI-C 2015, [2] Agarwal, A. et al. ESSCIRC 2011, [3] Do, A. T. et al. ESSCIRC 2013 [4], Wang, C.C., et al. TCAS-II 2010

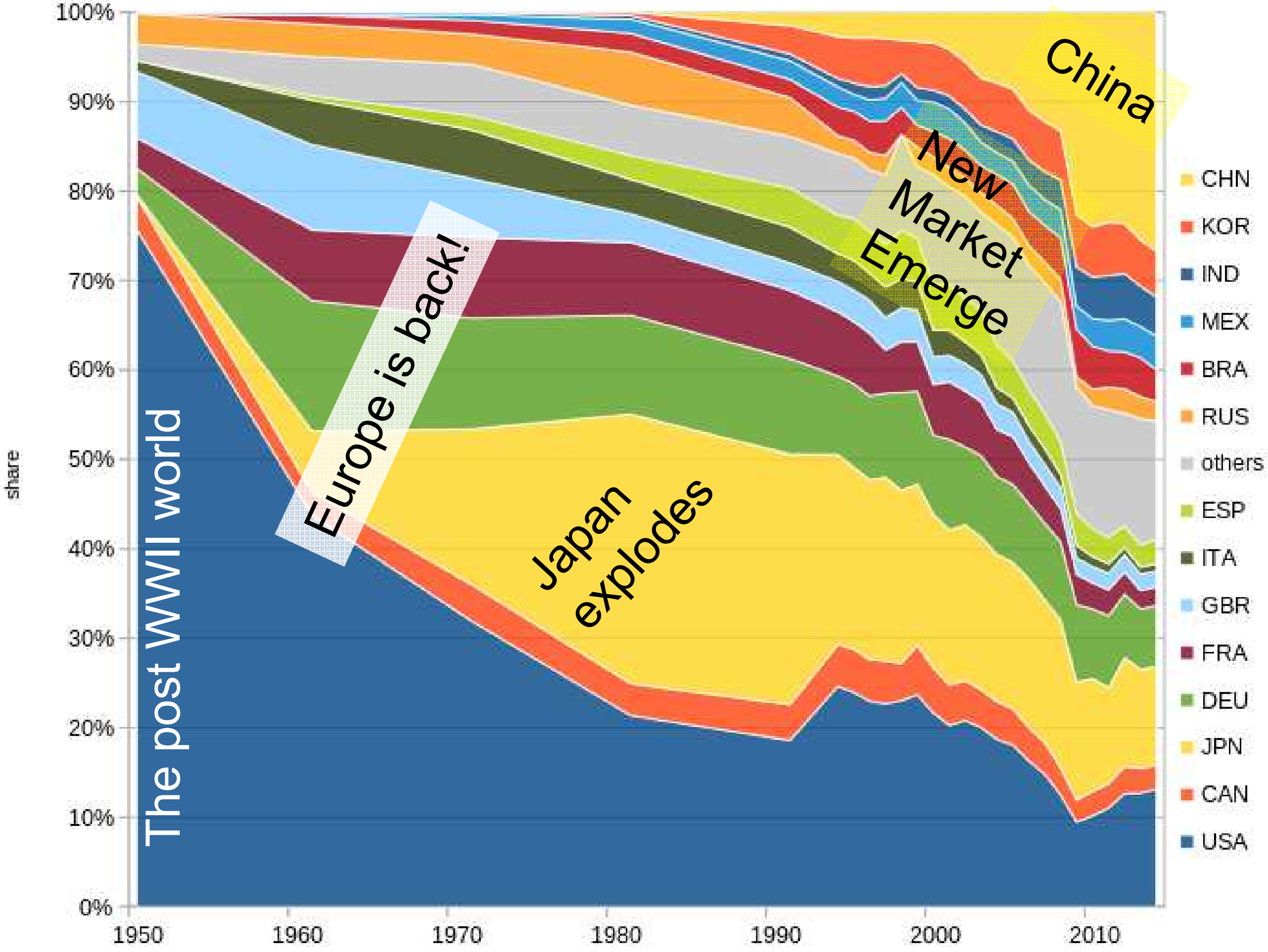
α Area with compact-design rules (with waiver on metal routing) β meas. WLMIN +300ps periphery delay (estimated) γ Assuming cycle time is 120% of meas. WLMIN

[N. Gupta et al., ESSCIRC, 2017]

MODELING 1 PPM YIELD (AND $) LOSSES IN SRAM

A GLANCE TO AUTOMOTIVE MARKET

- World Motor Vehicle Production per year / by country



[Image from Wikipedia]

HOW DEFECT TRACKING LEADS TO $

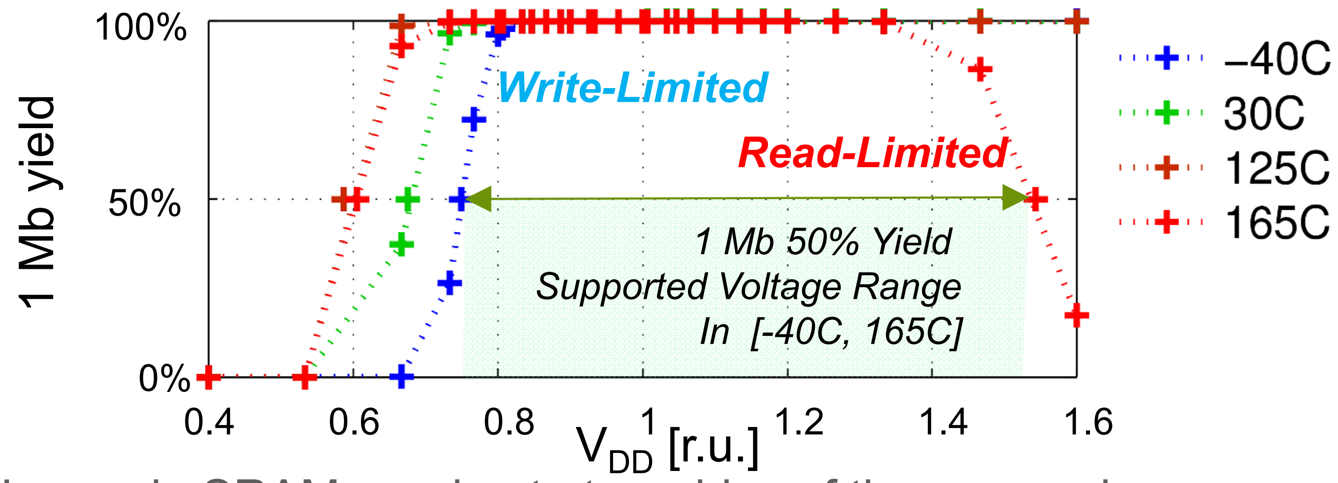
- **Impressive how Japan took over the car market in the 80s…**
  - Might be related to the **Toyota** Quality Management
- “Six Sigma (6σ) is a set of techniques and tools for process improvement. It was introduced by engineer Bill Smith while working at **Motorola** in 1986. Jack Welch made it central to his business strategy at **General Electric** in 1995.”
- “A six-sigma process is one in which 99.99966% of all opportunities to produce some feature of a part are statistically expected to be free of defects (3.4 defective features per million opportunities)”
- “… **Johnson and Johnson**, with $600 million of reported savings, **Texas Instruments**, which saved over $500 million as well as **Telefónica de Espana**, which reported $30 million euros of revenue in the first 10 months.”

- In circuits, SRAM is one of the highest Yield Detractors. With Emerging NVMs, other kinds of memory will assume this role.

[Quotes from Wikipedia]

YIELD MODELING OF MEMORY DEVICES

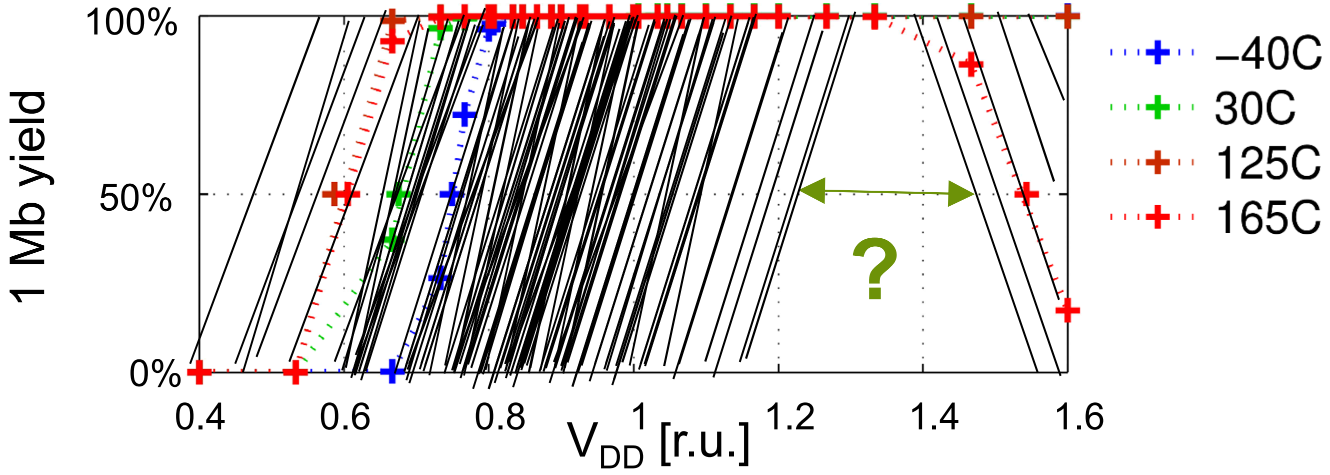
- **Four Classic Yield-Vs-Vdd curves**



- Yield losses in SRAM are due to two sides of the same phenomenon:
  - ❖ Either content is lost during read (cell can be written easily): Read-Limited
  - ❖ Or new content cannot be written: Write-Limited
- **Cell *limitation* changes with temperature**
- **Yield is monitored during technology developement on test vehicles of various capacity (e.g. 6) over the temperature range ~24statistics for both fresh and aged silicon**

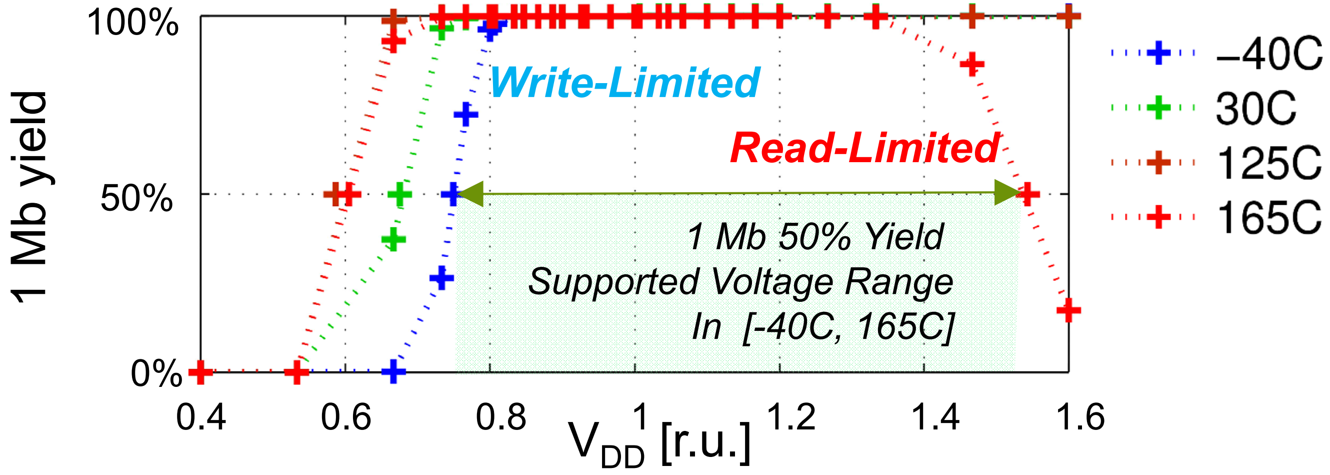
YIELD MODELING OF MEMORY DEVICES

- **In FD28SOI, Body Bias can be an effective performance booster for digital circuits. SRAM can be excluded from such boosting at the cost of:**
  - Increased area due to block isolation
  - No improvements in memory operations when digital is accelerated
- **Adding BB: (6 capacities) x (4 temperatures) x (6 BB voltages) = 144 Classic Yield Vs Vdd curves !!!!**

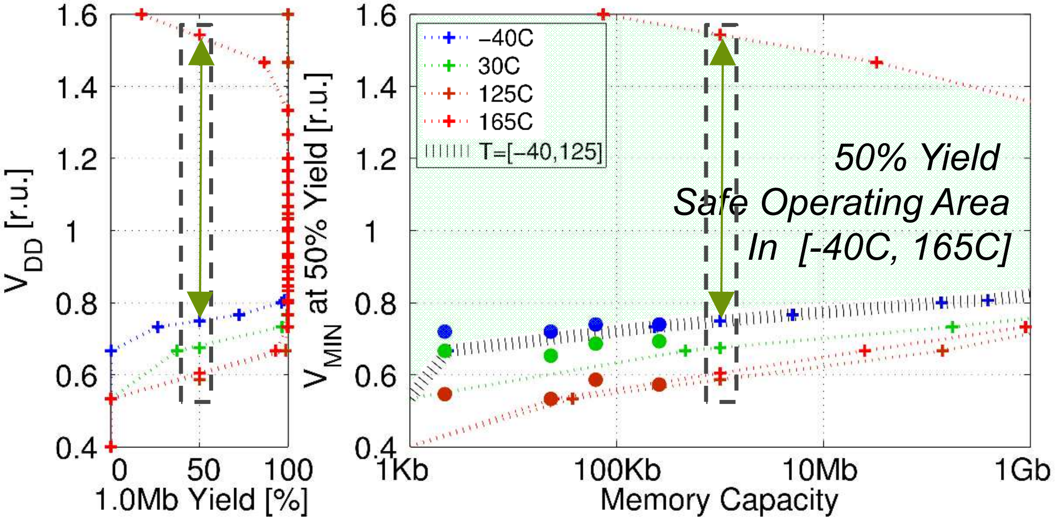


YIELD MODELING OF MEMORY DEVICES

- **Yield Vs Vdd curves are temperature-dependent, capacity dependent, Body Bias-dependent (here only T shown)**



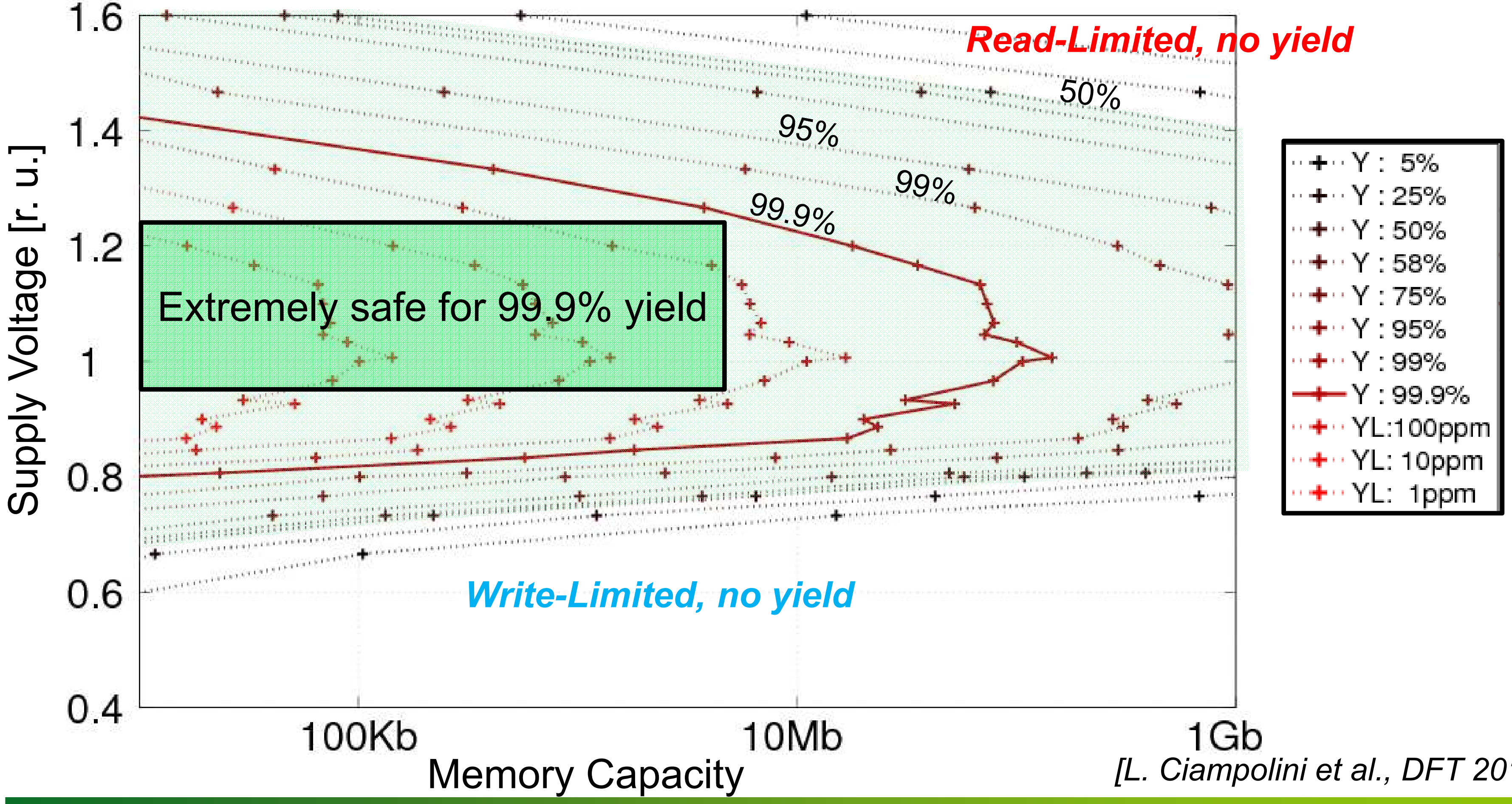
- **It is possible to add and densify information with the so-called *Yieldograms***



[L. Ciampolini et al., DFT 2017]

YIELD MODELING OF MEMORY DEVICES

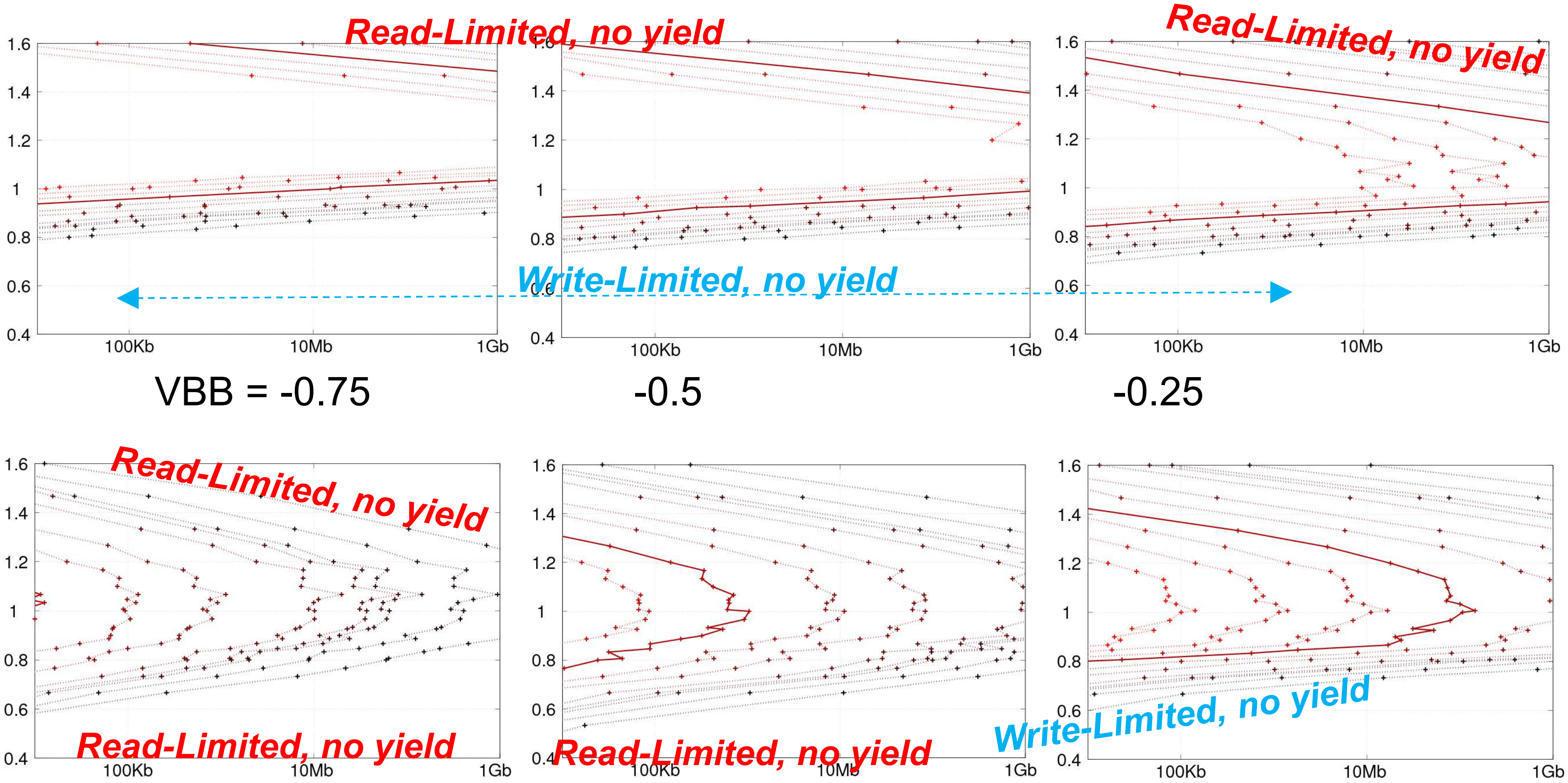
- Yieldograms represent yield levels (YL = Yield Loss) for any cut size and show how the bitcell performs at various voltages
- They allow to understand how far are we from yield losses



[L. Ciampolini et al., DFT 2017]

YIELD MODELING OF MEMORY DEVICES

- **Body-Bias effects on yield in FD28SOI SRAM**



Read-Limited, no yield

Read-Limited, no yield

Write-Limited, no yield

VBB = -0.75

-0.5

-0.25

Read-Limited, no yield

Read-Limited, no yield

Read-Limited, no yield

Write-Limited, no yield

VBB = 0.5

0.25

0

(read clockwise)

CONCLUSIONS

CONCLUSIONS

- Voltage drop calibration techniques open up options for designing compact and reliable high-density crossbar memories
- Computational memory opens the way to energy-efficient data-centric applications
- High-density 4T SRAM bitcell in 3D CoolCube technology demonstrated on silicon with 30% area gain
- Reconfigurable SRAM/CAM offers high performance in both operation modes with very low CAM search energy/bit
- Effective Yield modeling through yieldograms allows to monitor complex runtime use of SRAMs with dynamic Body-Bias