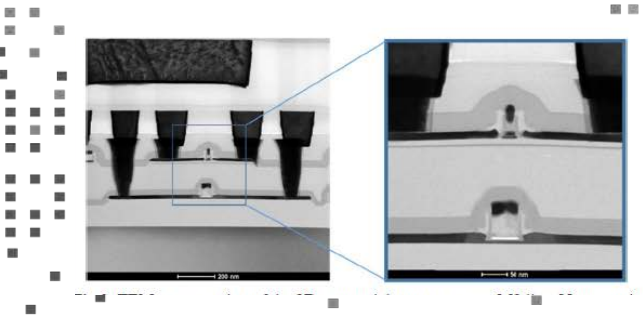
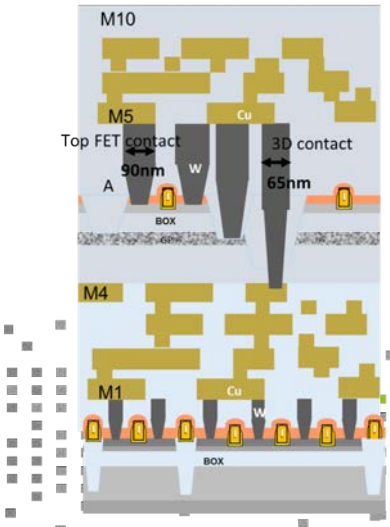
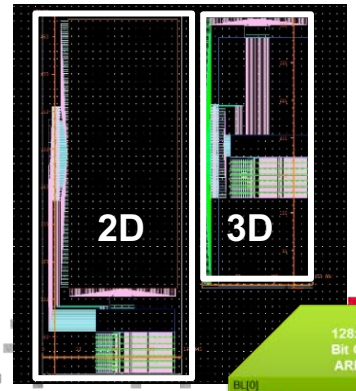
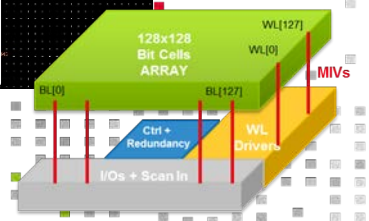
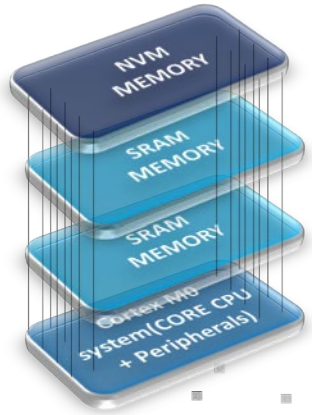








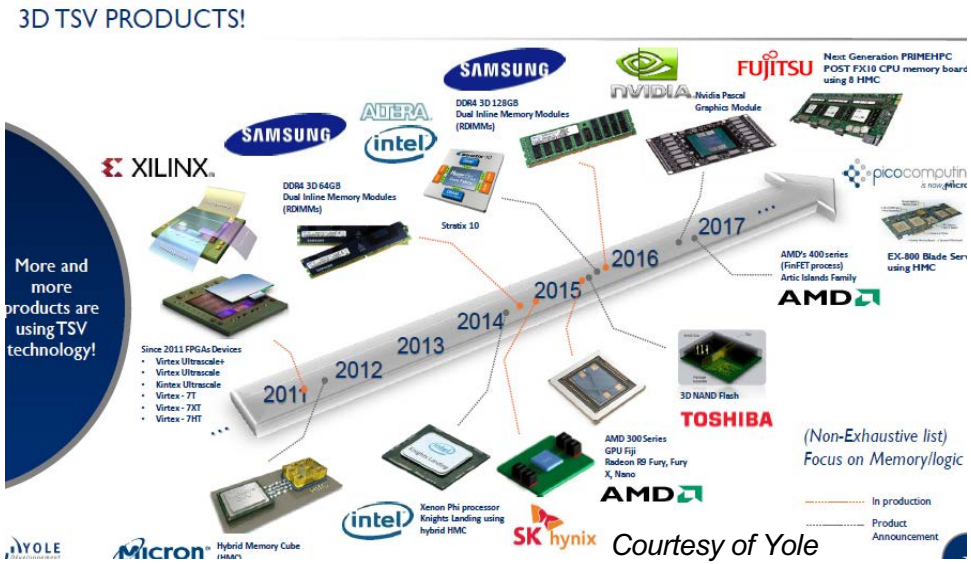


COOLCUBE™, A TRUE 3DVLSI ALTERNATIVE TO SCALING

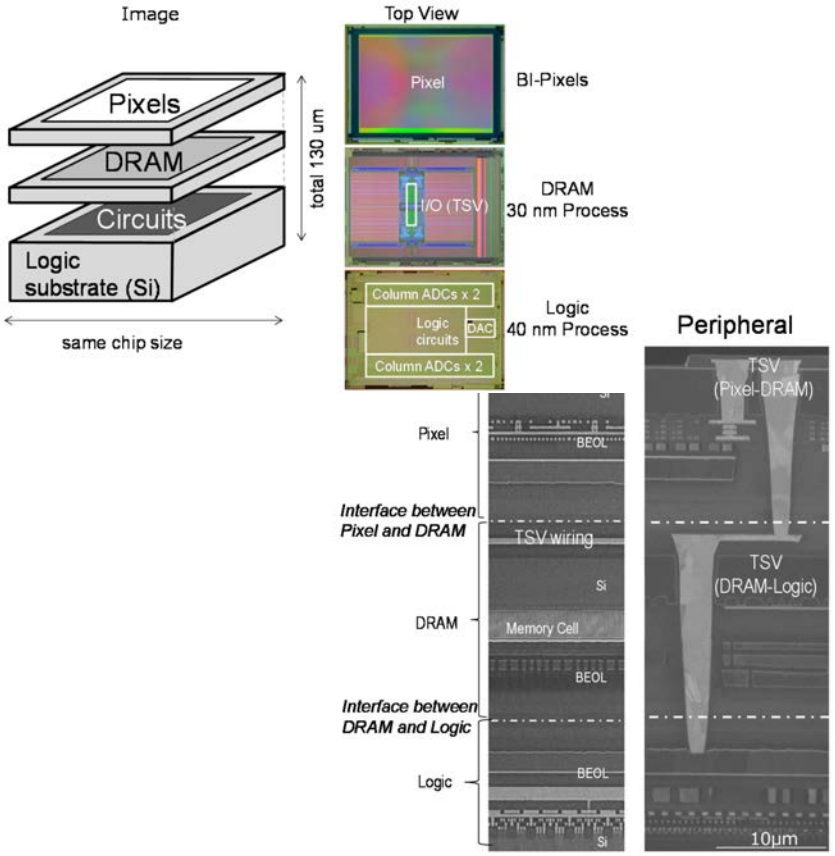
Sébastien THURIES, Olivier Billoint, Adam Makosiej, Pascal Vivet, Edith Beigné, Sylvain Choisnet, Francois Andrieu, Claire Fenouillet-Beranger, Laurent Brunet, Perrine Batude, Didier Lattard, Mehdi Mouhdach, Sebastien Martinie, Joris Lacord, Gerald Cibrario, Maud Vinet, Fabien Clermidy, Jean Eric Michallet

ARM RESEARCH SUMMIT – 09/17/2018

3D TSV IN PRODUCTION TODAY… AND PITCH SCALES DOWN QUICKLY



Courtesy of Yole



C2W - Computing : HBM, HMC Memory Cube, passive interposer … Coarse grain 3D (2.5D) – TSV & Cu-Pillars – 20-10 µm pitch

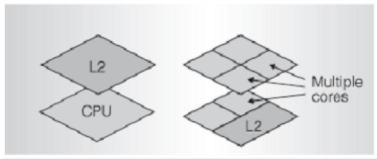
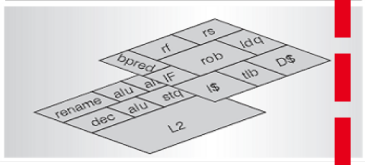
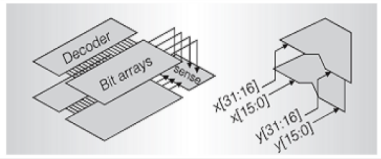
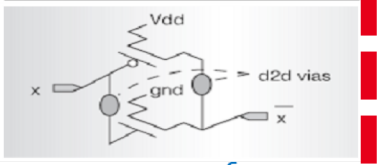
W2W - 3 Layers Stacked Imagers
19.3M pixels and a 1 Gbit DRAM
TSV 2.5 µm diameter - 6.3 µm pitch [1]
→ 3D Pitch scales down

[1] SONY, ISSCC 2017

3D TECHNOLOGIES PORTFOLIO – 3DVLSI ZOOM

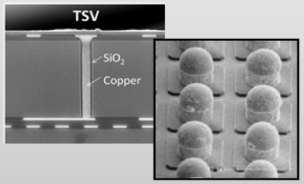
Global Interconnect

Local Interconnect

1-Entire core	2-Logic bloc	3-Logic gates	4-Transistors
			
2.5 x10^3 3D plugs / mm^2	**100 x10^3** 3D plugs / mm^2	**5 x10^6** 3D plugs / mm^2	**100 x10^6** 3D plugs / mm^2

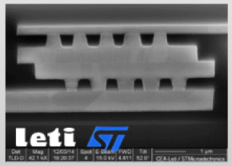
Granularity scale

TSV + Cu-pillar



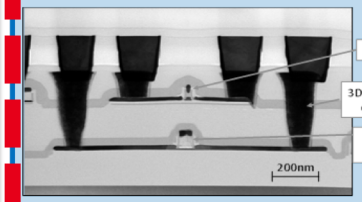
Ø	10 µm
pitch	20 µm

HD TSV + Direct bonding



Ø	1.7 µm
pitch	3.4 µm

3D Monolithic: 28nm FDSOI CoolCube™



3D monolithic integration

- Very thin active and inter-die layers
- Intermediate BEOL

Ø	0.05 µm
pitch	0.11 µm

MIV=3D-contact
(Monolithic Inter-tier Via)

Pitch Scale

TSV & Cu-Pillar > 10 µm

7µm > D2W > 3 µm

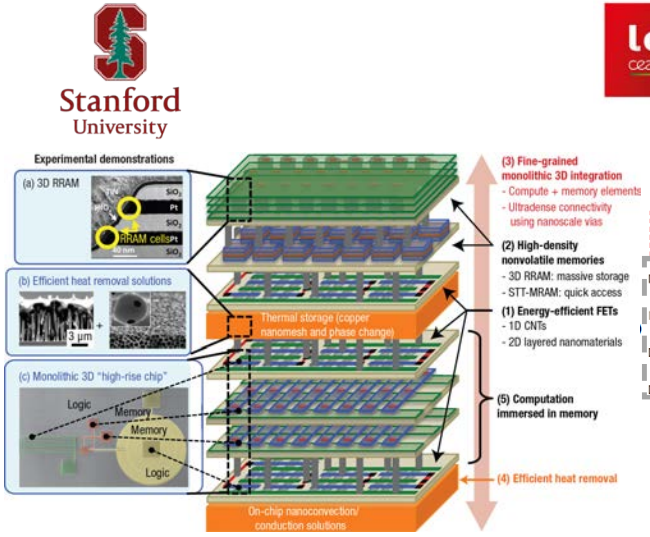
5 µm > W2W < 1 µm

Nano Scale 3D contact

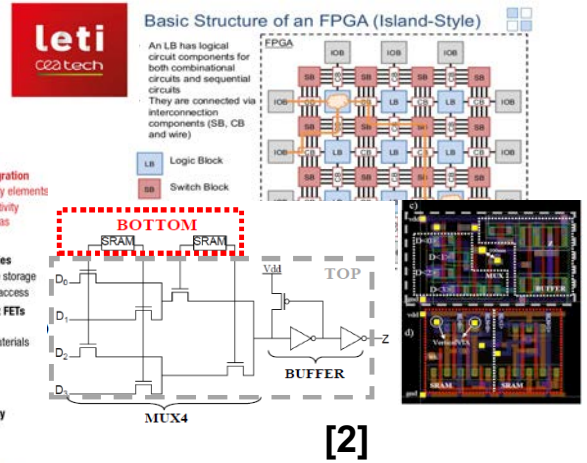
3D Sequential == 3D Monolithic == CoolCube™ … and 3DSoC in US

150 nm > 3D Monolithic

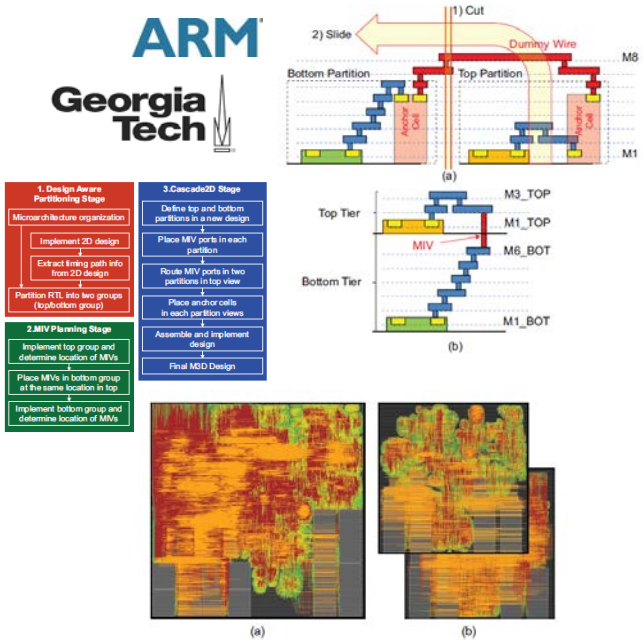
3DVLSI PPA STATE OF THE ART



[1]



[2]



[3]

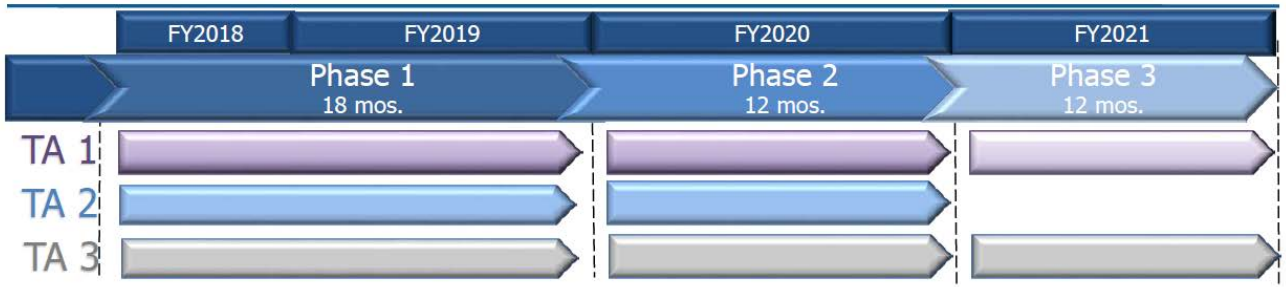
1. 1000x benefit thanks to MIV (+ CNT, NVM…) for Computation Immersed in Memory (N3XT)
2. -55% Area & -47% EDP for FPGA → Logic-Memory Partitioning
3. M3D to tackle scaling with 25% better Perf . Vs 7 nm node → PnR EDA Tool Focus

Multi Tiers Architecture for huge PPA gain (Increase Memory Bandwidth)
Need EDA tools for optimized PPA design

[1] M.M. Sabry, Energy-Efficient Abundant-Data Computing: The N3XT 1,000×, COMPUTER 2015. STANFORD
[2] O. Turkyilmaz., 3D FPGA using high-density interconnect Monolithic Integration. DATE 2014. CEA-Leti
[3] K. Chang, Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools. ARM & GaTech. ICCAD 2016

US-DARPA 3DSOC PROGRAM OVERVIEW

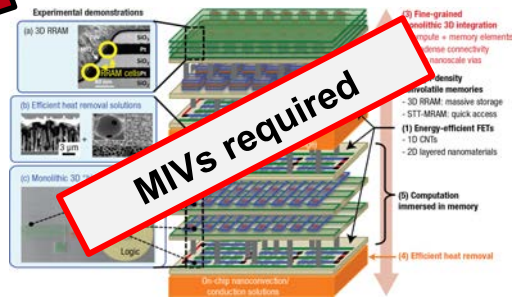
3DSoC Program Schedule



Phase 1 Outcomes
- Initial 3DSoC process defined
- PDK V0.5 defined
- 3DSoC technology benefits simulated
- First pass DEC fabricated and tested
- Initial 3DSoC EDA tools released

Phase 2 Outcomes
- 3DSoC process demonstrated
- PDK V1.0 released for design
- 3DSoC benefits demonstrated
- Final DEC design fabricated
- 3DSoC EDA tools released for targeted designs

Phase 3 Outcomes
- 3DSoC process used for external designs
- Final PDK released for design
- MPW runs successfully yielded
- 3DSoC EDA tools released for general designs



- TA-1: Developing the 3DSoC fabrication process
  - Establish unit processes and flow integration
  - Define the 3DSoC technology PDK
- TA-2: Designing and Implementing the DEC
  - Design 1st and 2nd pass DEC design
  - Foster use of the DEC to drive development and yield
- TA-3: Developing the 3DSoC EDA design flow
  - Develop EDA tools for 3DSoC compute/memory designs
  - Support tools for advanced 3DSoC designs

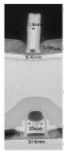
Metric	Goal
3DSoC Capability	> 50X 7nm 2D PaP
Hardware Accuracy	< 2% deviation from3DSoC technology targets
Yield	> 30% for full 3DSoC designs
EDA Tools	Successful use of EDA flow for a > 500M gate/4GB memory design

Overcome Scaling

EDA Tools to support 3DVLSI

CEA-LETI 3D CIRCUITS ROADMAP

3DVLSI



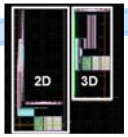
150 nm Pitch

Design Early Exploration

3D early PDKIT, Thermal study
Partitioning evaluation (CoC, CoB, NoP)
Device silicon demo.

1st MPW Q3 2017

PDKit v1,
SRAM test chip
Early EDA Tools



2nd MPW Q2 2019

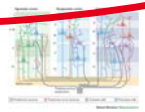
PDKit v2,
SoC test chip,
Signoff EDA Tools

Multi Layer Computing In Memory

Heterogeneous & Configurable computing
Non Volatile Memory Computing Cube

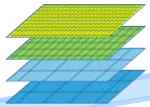
Computing in Memory Cube

Cortical Column



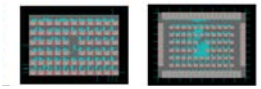
Bio-inspired 3D archi
Heterogeneous computing
3DHD and 3D Seq.

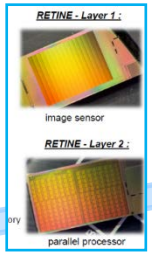
Smart Imager



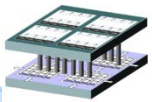
4 layer ISP
Analog & SIMD computing
Smart host interface
40nm/28nm, CuCu 2µm

Imager & Neuro Applications





Smart Imager

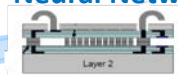


2 layer ISP
SIMD array
130nm, CuCu 7µm

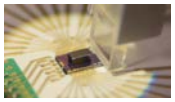
3D-HD

10-1 µm Pitch

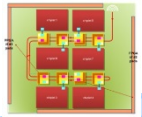
Neural Network



2 layer Neural Network
Neuron & Memory array
130nm, CuCu 7µm

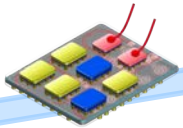


Photonic Interposer



Optical NoC
96 cores prototype
FDSOI 28nm + µbump 40µm

Photonic System



3D TSV & Photonic convergence
Photonic Memory-interface

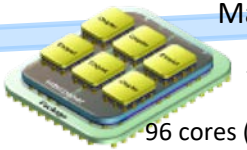
2.5D/3D

20-10 µm Pitch

Logic-on-Logic

Async robust links
DFT & Fault Tol.
CMOS 65nm
µbump 40µm pitch

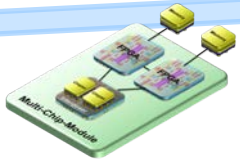
Passive & Active Interposer



Manycore

96 cores (32bit MIPS)
cache-coherent memory
FDSOI 28nm + CMOS 65nm
µbump 20µm pitch

Homogeneous & Heterogeneous Manycore



High-end server compute node
Heterogeneous chiplets
FDSOI 28nm + CMOS 65nm
µbump 20µm pitch

Manycore & Computing Roadmap

2015 2016 2017 2018 2019 2020

AGENDA

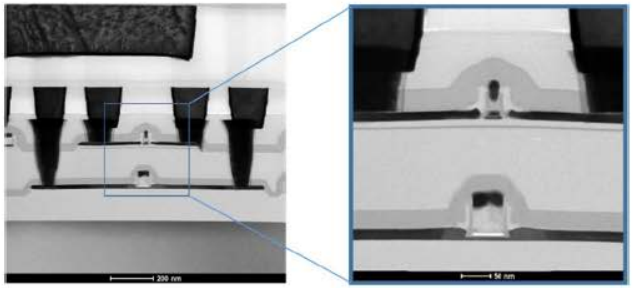
- 3D VLSI Silicon Device Measurements
- 28 nm 3DVLSI Silicon MPW specifications & PDKit add-on
- 3DVLSI Digital Design Methodologies (PnR) Overview
- 3DVLSI Architectures Breakthrough
- Conclusion

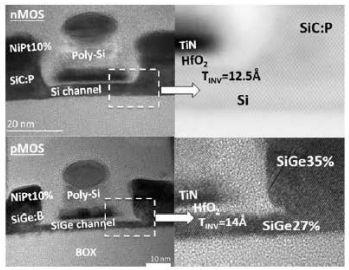
AGENDA

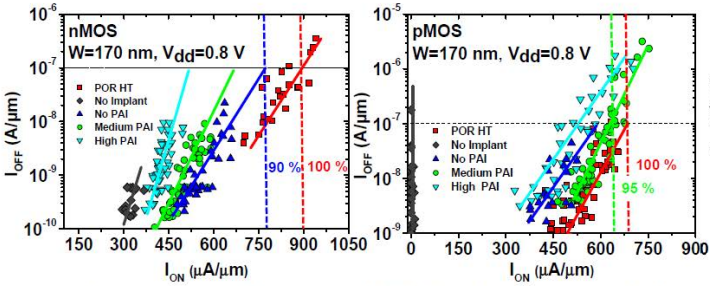
- 3D VLSI Silicon Device Measurements
- 28 nm 3DVLSI Silicon MPW specifications & PDKit add-on
- 3DVLSI Digital Design Methodologies (PnR) Overview
- 3DVLSI Architectures Breakthrough
- Conclusion

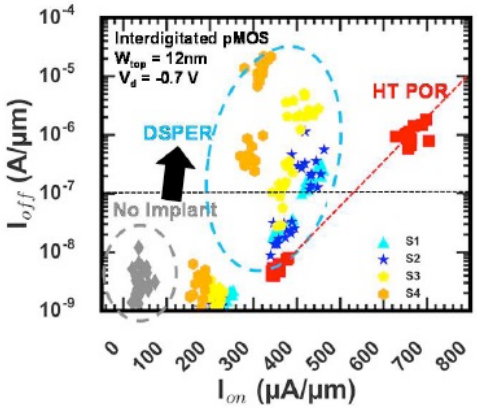
COOLCUBE™ - 300MM DEMONSTRATION

For the first time, a full 3D VLSI CMOS over CMOS CoolCube™ integration was demonstrated on 300mm wafers within an industrial environment[1] and Low T° FinFET process with performances close to those of the High Temperature Process Of Reference









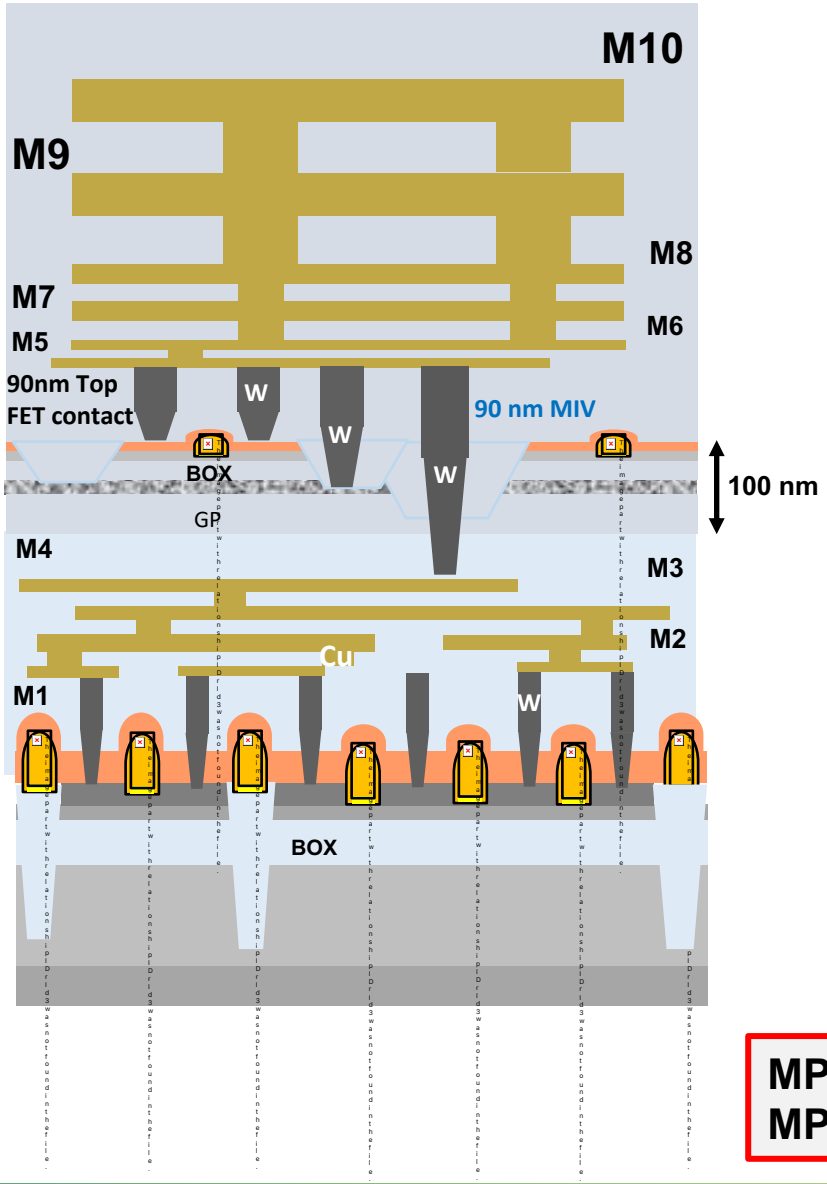
CMOS Cold Activation (SPER) with perf. very close (~5 to 10 %) from high temperature reference[1]

[1] CEA-Leti., VLSI 2016 & IEDM 2017

AGENDA

- 3D VLSI Silicon Device Measurements
- **28 nm 3DVLSI Silicon MPW specifications & PDKit add-on**
- 3DVLSI Digital Design Methodologies (PnR) Overview
- 3DVLSI Architectures Breakthrough
- Conclusion

28 NM 3DVLSI PROCESS FABRICATION (MPW)



Top BEOL M5 to M10 (+AluCap)

28nm pitch processed back to foundry for BEOL finishing

Top tier (Cold Process):

28nm like density **65nm like Processed at leti**

→ **Via4 becomes Monolithic Inter-tier Via (MIV)**

Bottom tier:

FDSOI 28nm foundry baseline process

4 metal layers in Copper

without any modification

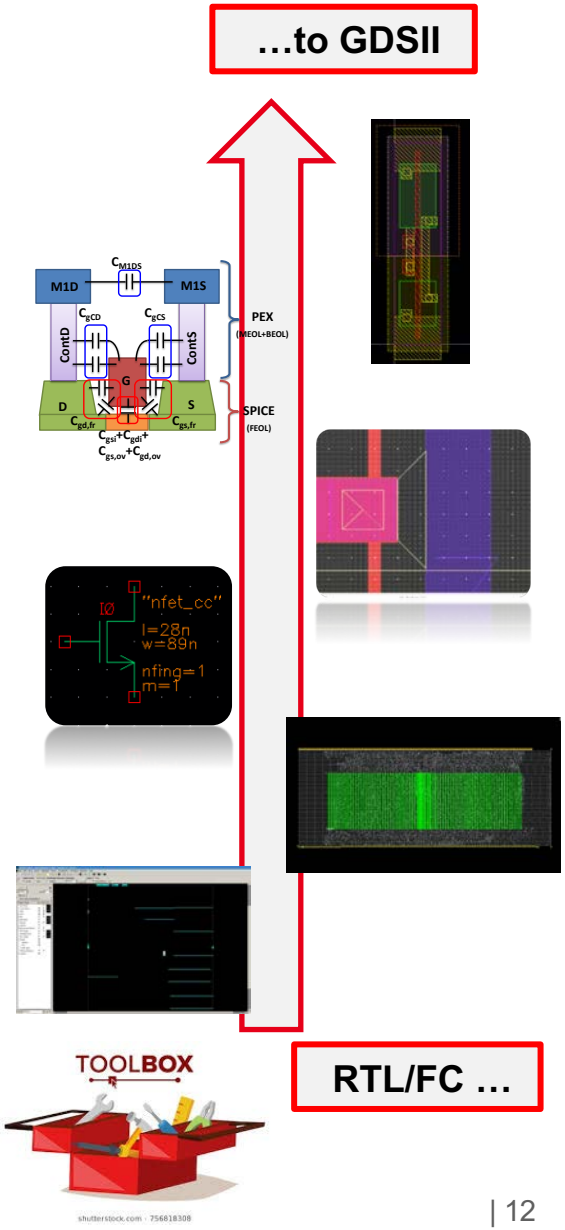
MPW #1 : TO Q3 2017 - PDKit 1 - RnD CEA/Leti
MPW #2 : TO Q2 2019 - PDKit 2 - H2020 3D MUSE

3DVLSI EDA TOOLBOX & PDKIT ADD-ON

Flow	CAD tools requirement (release aligned on ST)
Design Environment (schematic/layout)	Cadence Virtuoso & Analog Design Environment (ADE)
Post-Layout Simulation (PLS) Parasitic Extraction (PEX)	Mentor Calibre PEX (xACT) & interactive
Physical Verification (DRC/LVS)	Mentor Calibre DRC/LVS & interactive
Electric al Simulator	Eldo (Mentor G.), HSPICE (Synopsys)
Place and Route	Innovus (Cadence) + MIV Parasitics & Techno Lef
Synthesis	Design Compiler topo (Synopsys)

PDkit content :

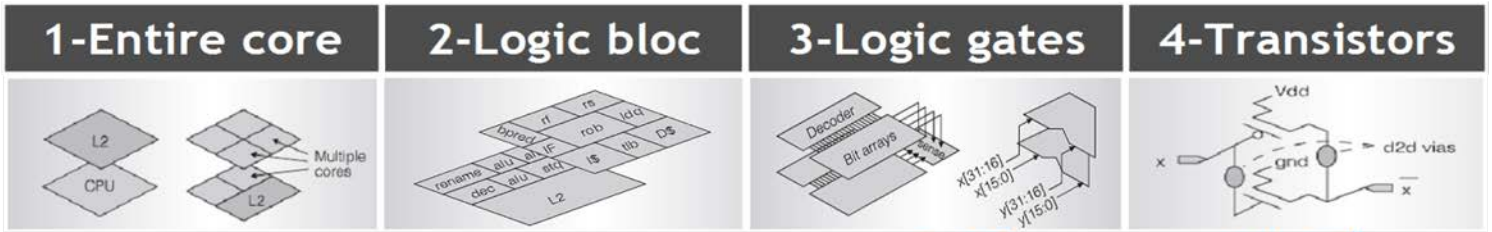
- **STm full design Platform for Bottom Die (Pcells, SRAM, SC…)**
- **CEA-Leti Pcells, Standard Cells and IO Pad for Top CoolCube™ Die**



AGENDA

- 3D VLSI Silicon Device Measurements
- 28 nm 3DVLSI Silicon MPW specifications & PDKit add-on
- **3DVLSI Digital Design Methodologies (PnR) Overview**
- 3DVLSI Architectures Breakthrough
- Conclusion

3DVLSI DESIGN CHALLENGES

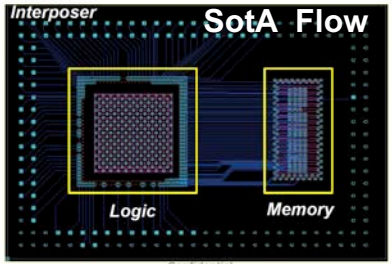


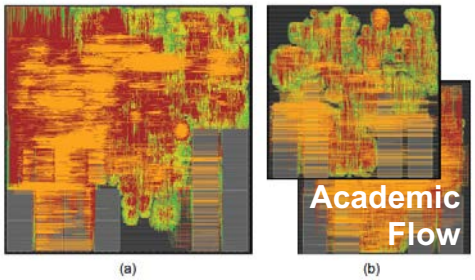
2.5 x10^3 3D plugs / mm^2 — **100 x10^3** 3D plugs / mm^2 — **5 x10^6** 3D plugs / mm^2 → **100 x10^6** 3D plugs / mm^2

Granularity scale

Global Interconnect → 3D Packaging

Local Interconnect → 3D Design





Georgia Tech DARPA leti ceatech ARM GLOBAL FOUNDRIES

Pitch Scale

TSV & Cu-Pillar > 10 µm

7µm > D2W > 3 µm

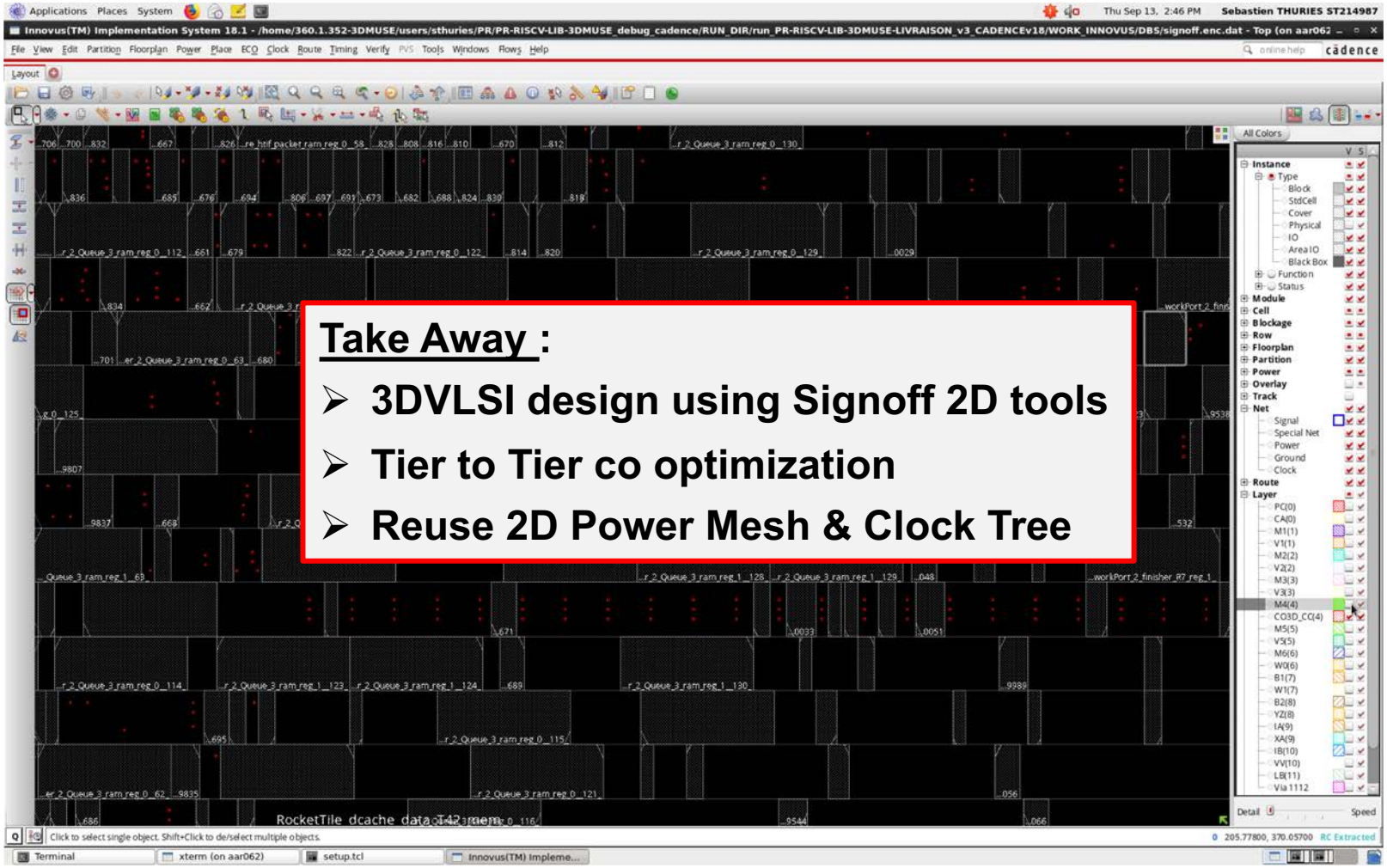
5 µm > W2W < 1 µm

Design flow needs to be upgraded for optimal PPA

Nano Scale 3D contact

150 nm > 3D Monolithic

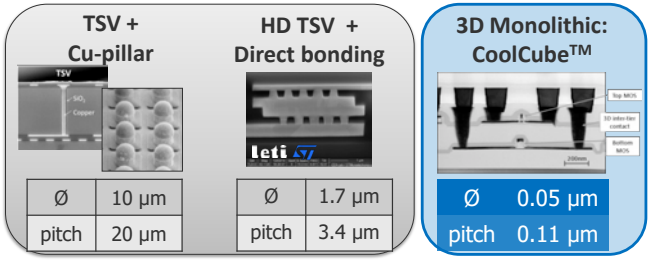
3DVLSI SIGNOFF PLACE AND ROUTE

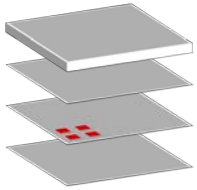


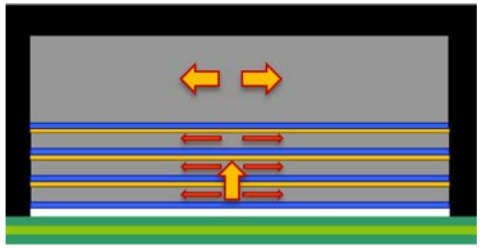
Top Standard Cells on Coolcube™ process …	Bottom SRAM Memories	Zoom on SRAM Pins and SC above	MIVs are managed by PnR tool

3D TECHNOLOGIES THERMAL COMPARISON [1]

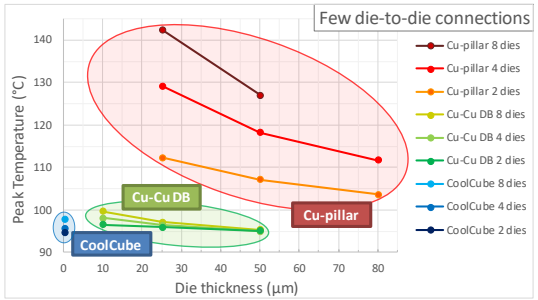
- **Objective**
  - Compare thermal performances of different 3D technologies
    - TSV + µ-bumps
    - Hybrid Bonding
    - 3DVLSI (CoolCubeTM)
- **Method**
  - Define a Set of Experiments :
    - Different technology parameters (#layers, 3D interco pitch, materials, etc)
    - Different power scenerio
    - Different thermal dissipation, etc
  - Thermal model using SAHARA tool





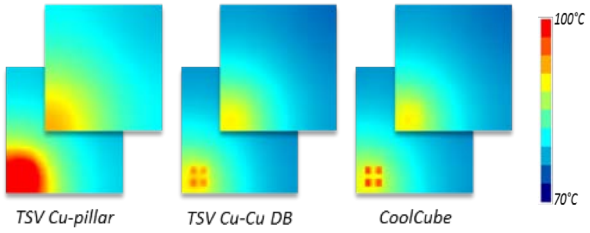


Hot spot dissipation



Results:

- Better thermal coupling for Hybrid Bonding & CoolCube
- Reduced hot spot effect
- Strong sensitivity to interconnect density, and die thickness



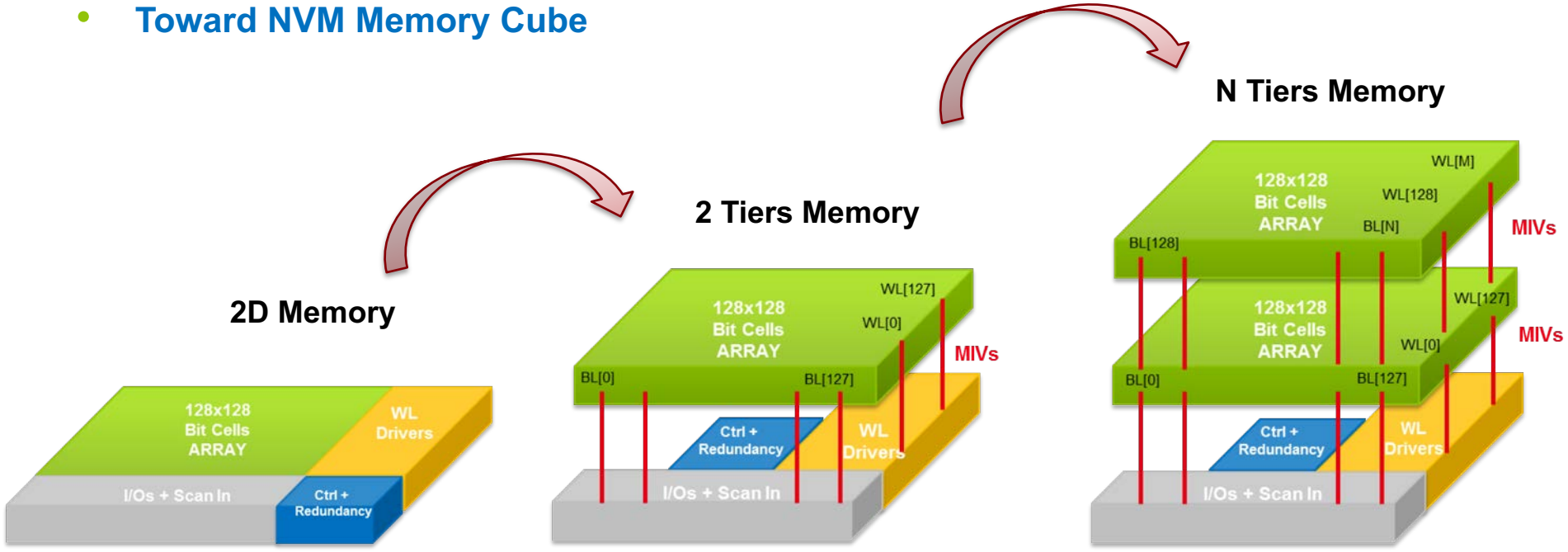
[1] C. Santos et al., 3DIC'16, CEA-Leti

AGENDA

- 3D VLSI Silicon Device Measurements
- 28 nm 3DVLSI Silicon MPW specifications & PDKit add-on
- 3DVLSI Digital Design Methodologies (PnR) Overview
- **3DVLSI Architectures Breakthrough**
- Conclusion

MULTI TIERS ON-CHIP MEMORY CUBE ARCHITECTURE

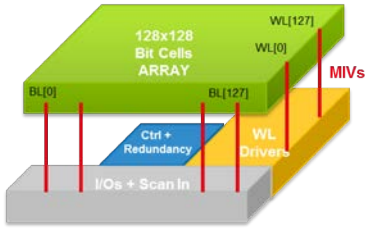
- **Objective : Multi Tiers embedded Memory → Multiple Array on Periphery Partitioning :**
  - Bottom level: all decoding logic, drivers, I/Os and redundancy
  - Top level is a 128x128 SRAM array connected thanks to MIVs
  - **Toward NVM Memory Cube**



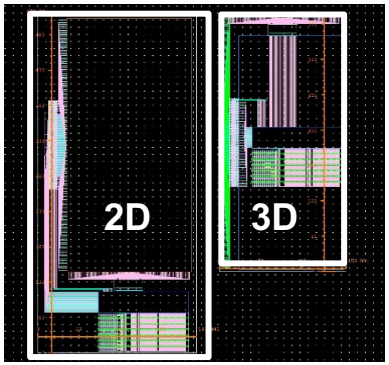
→ ~2x reduced footprint target
→ Reduced BEOL layers (#4) on Top process

→ ~Nx reduced foorptint

2 TIERS SRAM MPW #1 TAPED OUT



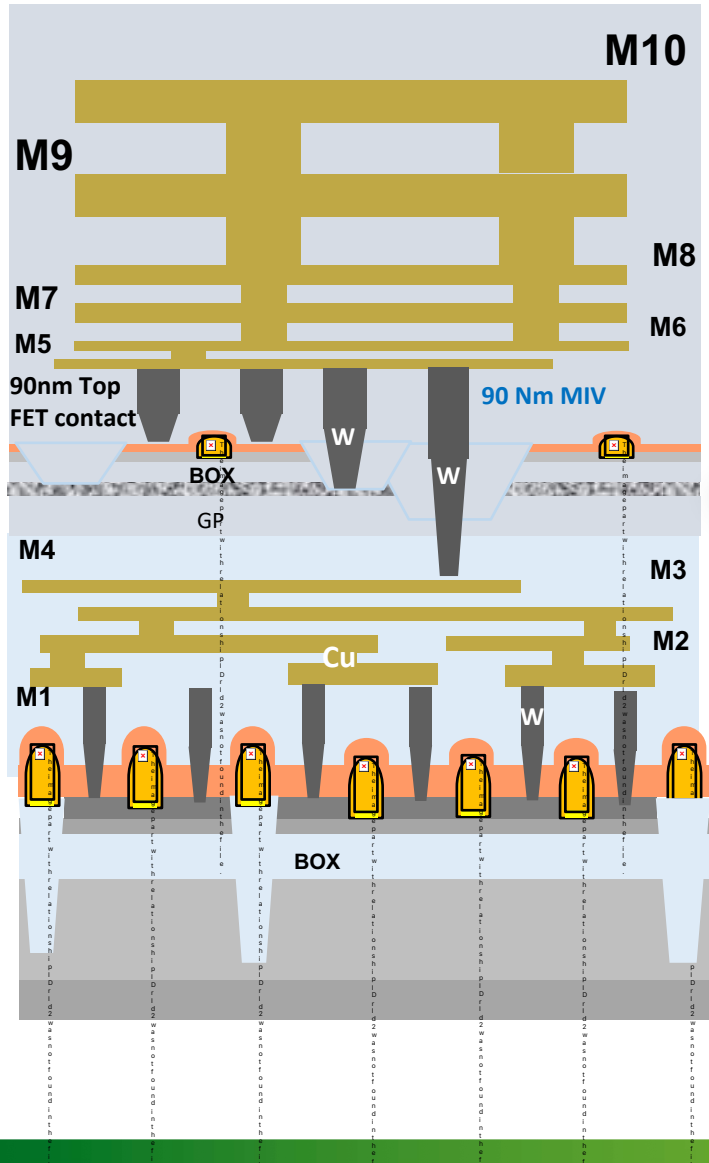
- **Objective:**
  - first tape out at IP level (building block)
- **Partitioning:**
  - Bottom die for IO, periphery & redundancy
  - Top die : bit cell array 128x128.

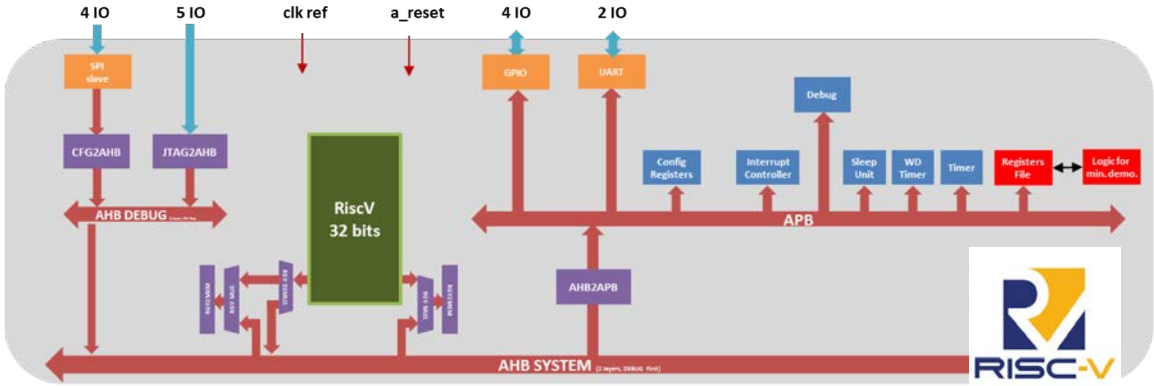


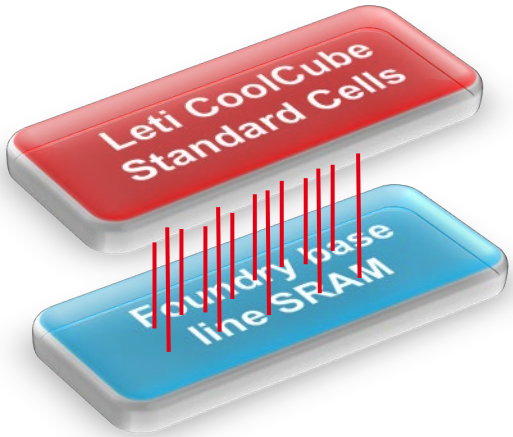
- **Early Results**
  - 128x128x6T = 98304 Transistors on top process
  - 2D Area : 193um x 478um= 92475um2 (same design rules than 3D)
  - 3D Area : 153um x 360um= 55421um2
  - Done on PDKit v1 → Perf. Power simulation on going with PDKit v2
  - Silicon Measurements planned for ~Q1 2019

- **40% Footprint reduction**
- **~100000 Transistors on Top (Cold) Process**
- **2068 MIVs – Density : 37600 MIVs / mm²**

2ND SILICON TESTCHIP (MPW #2)
32 BITS RISC-V SOC TARGET



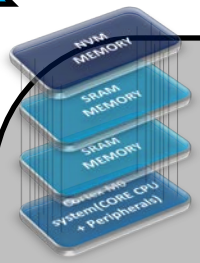




Next : PPA vs 2D (with full 28/28 PDKit) + early PPA on adv. node

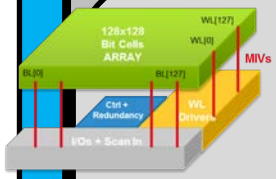
CONCLUSION & PERSPECTIVES

3D Seq. Architectures breakthrough



- Ultra Dense in **Memory Computing Cube** (**NVM & M3D**)
- Multi Layers stacked BSI (**Smart Imagers**) – Heterogeneous 3D Integration
- High memory BW DNN

we are here

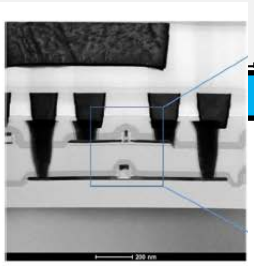


- 2 Layers Memory Cube
- Logic on Memory
- DNN Architectures
- Sensor Array

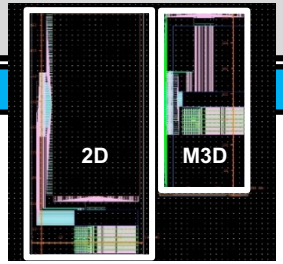
Next :

- **BeoL (ReRAM) NVM Memory**
- **Adv. node PPA study vs 2D**
- **SoC Silicon Demo.**

- Silicon demo at device level
- Design methodology based on 2D tools ready
- Thermal & Cost are good
- Early PDKit

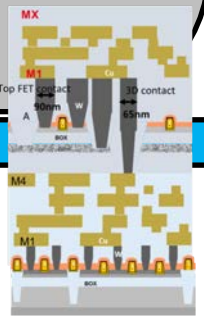


- IP Level Silicon demo. & measurements
- Advanced Device Maturity
- Prelim. Design Platform



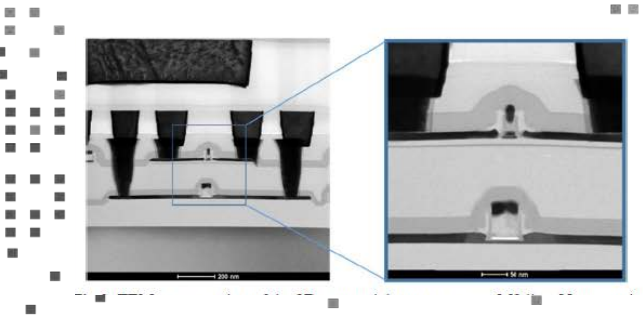
- SoC Level Silicon demo.
- Advanced Techno. maturity
- Full CC Design Platform to plug on customer technology targeted

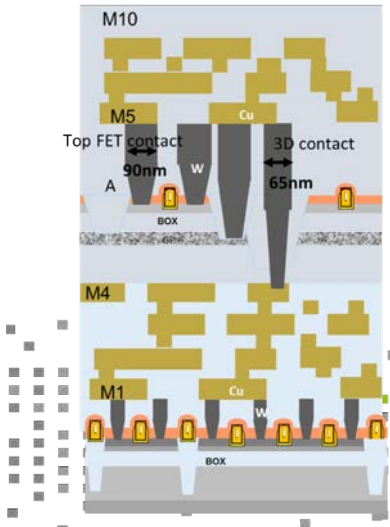
we are here

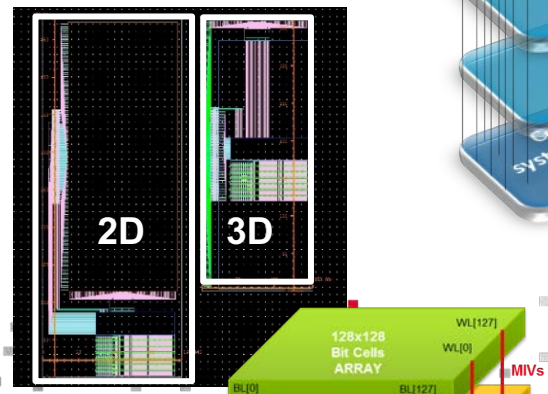


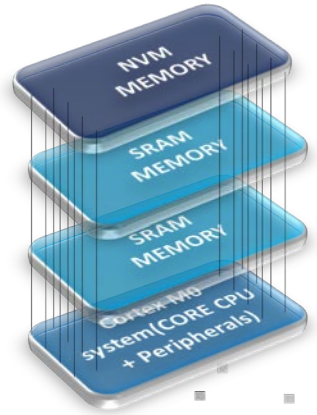
Silicon Readiness

Thank you for your attention







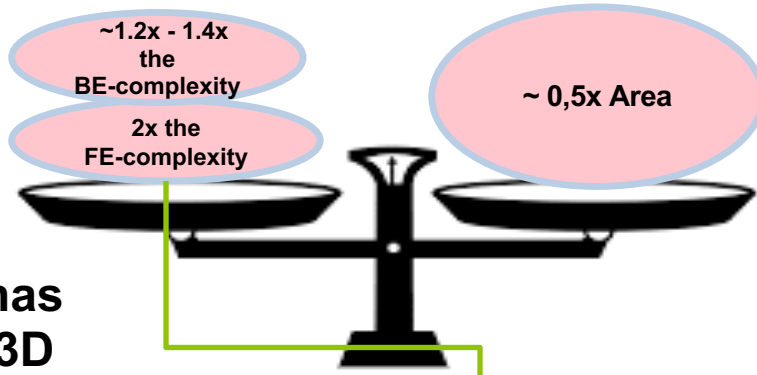


COOLCUBE™, A TRUE 3DVLSI ALTERNATIVE TO SCALING

Sébastien THURIES, Olivier Billoint, Adam Makosiej, Pascal Vivet, Edith Beigné, Sylvain Choisnet, Francois Andrieu, Claire Fenouillet-Beranger, Laurent Brunet, Perrine Batude, Didier Lattard, Mehdi Mouhdach, Sebastien Martinie, Joris Lacord, Gerald Cibrario, Maud Vinet, Fabien Clermidy, Jean Eric Michallet

ARM RESEARCH SUMMIT – 09/17/2018

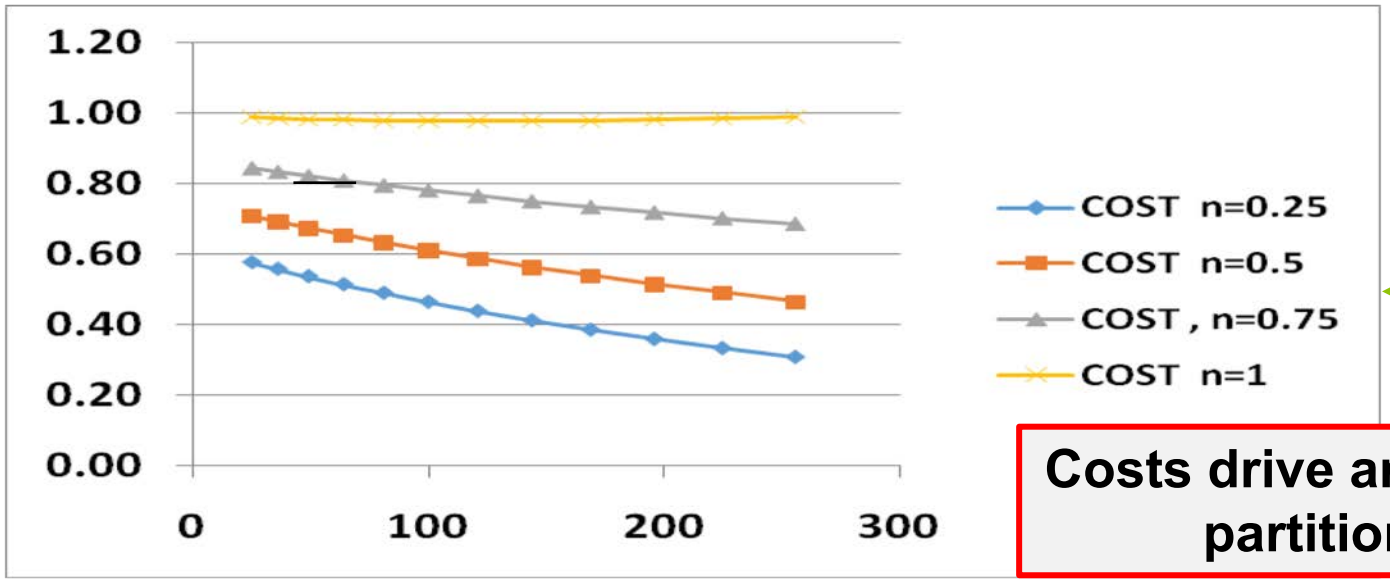
3D SEQUENTIAL COST IS GOOD



The Bose-Einstein yield model for 3D wafers has the additional factor Y3D. Namely, Y = Y2D * Y3D

$$Y = 1/(1 + DoA)^N * 1/(1 + D3DA)^{N3D}$$ [1]

Do = 0.1, N=24, a=2, d=1



Relative die cost CoolCube™ vs. traditional process

Cost = ~0.6

Costs drive architecture partitioning

Area (A) of chips in 3D wafers in mm^2

- NGD is # of good die per wfr
- a is the area reduction factor
- n is the normalized complexity # relative to 2D (N)
- d is the normalized defect density relative to 2D (Do)

[1] Gitlin D. et. al. Cost Model for Monolithic 3D Integrated Circuit, S3S 2016