

3D ADVANCED TECHNOLOGIES FOR NEUROMORPHIC ARCHITECTURES – HUGHES METRAS - CEA/LETI

Alexandre Valentian, Pascal Vivet, Severine Cheramy, Amandine Jouve, Perrine Batude

LETI'S TECHNOLOGICAL: SILICON PLATFORM



World-class facilities for your future business needs

© Pierre JAYET/CEA

NICE '2019

March.26-29th.2019

Neuro-Inspired
Computational Elements
Workshop



300 mm



200 mm

1 CMOS / FDSOI



2 BEYOND CMOS

3 POWER



4 MEMS



5 MEMORIES



6 IMAGERS



7 DISPLAY/LED



8 Si PHOTONIC

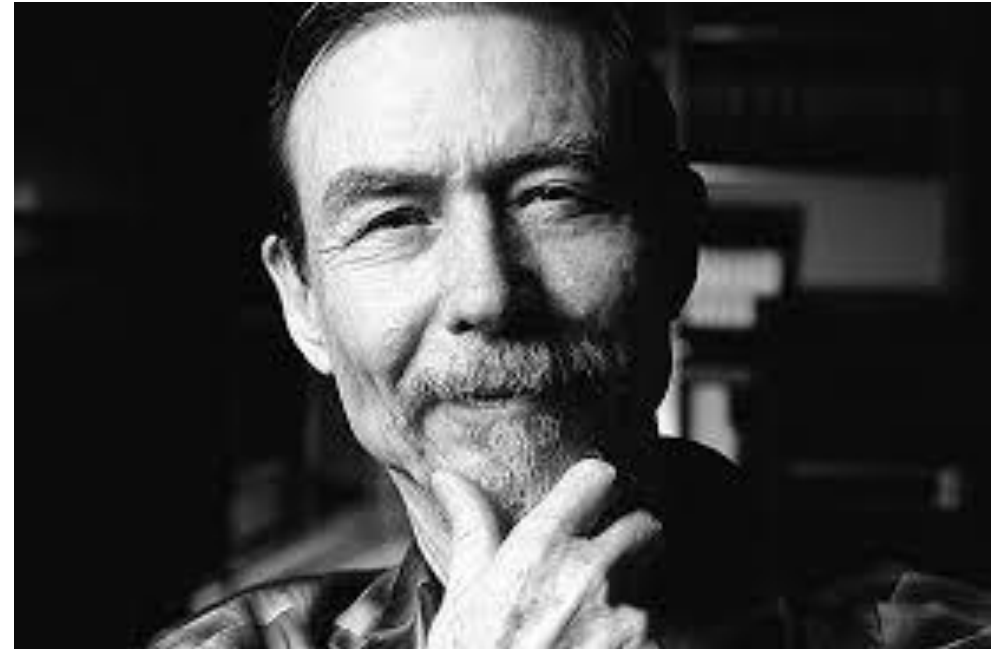


9 3D



10 SOI SUBSTRATE

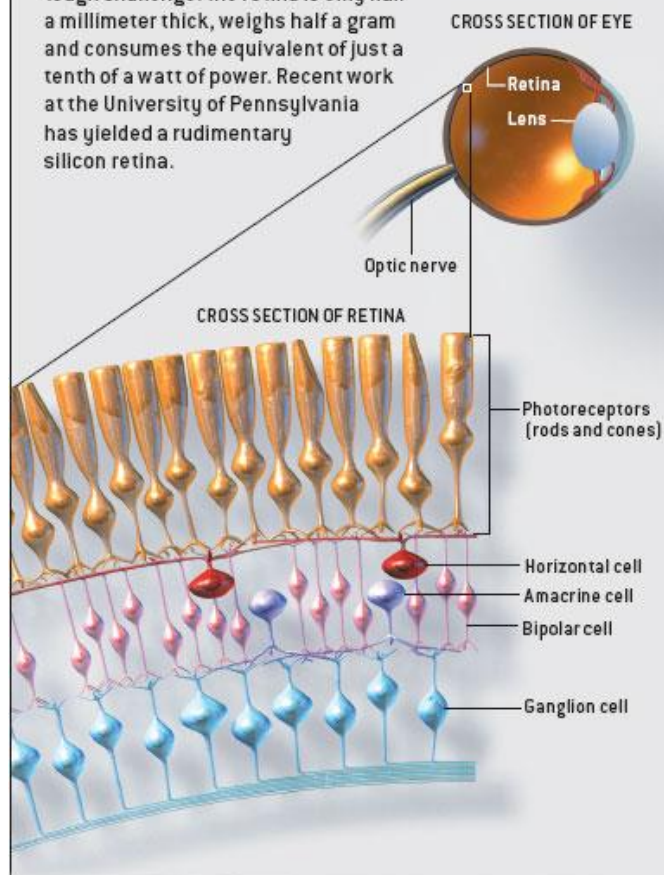




WHY GOING 3D ?

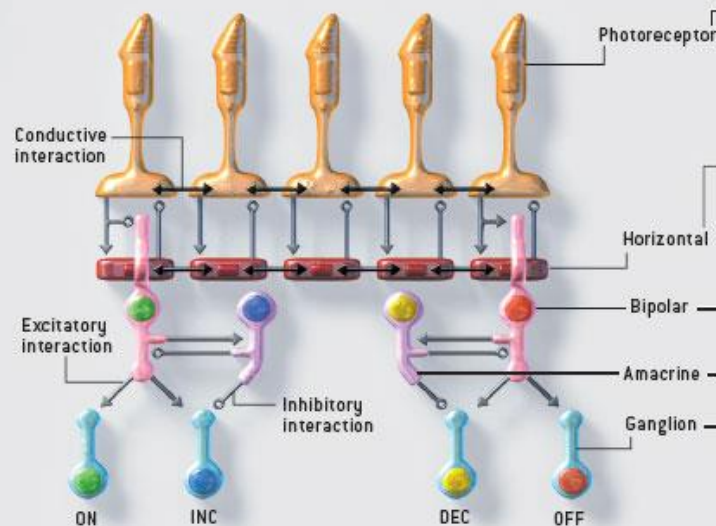
RETINAL NEURONS AND NEUROMORPHIC VISION CHIPS

Biological sensory systems provide compact, energy-efficient models for neuromorphic electronic sensors. Engineers attempting to duplicate the retina in silicon face a tough challenge: the retina is only half a millimeter thick, weighs half a gram and consumes the equivalent of just a tenth of a watt of power. Recent work at the University of Pennsylvania has yielded a rudimentary silicon retina.



BIOLOGICAL RETINA

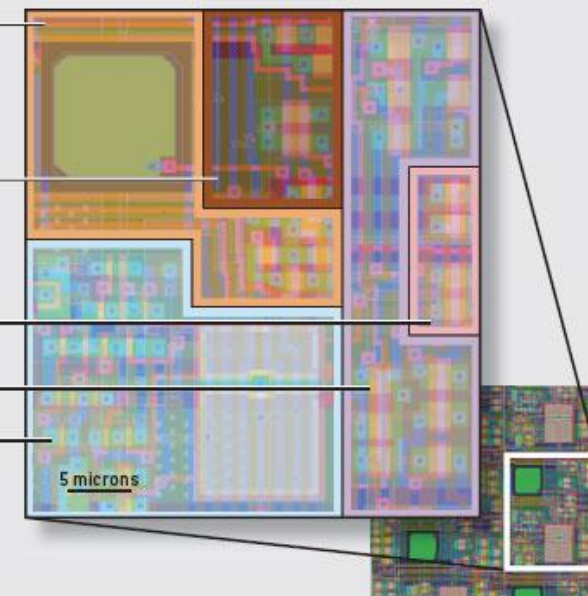
The cells in the retina, which are interconnected, extract information from the visual field by engaging in a complex web of excitatory (one-way arrows), inhibitory (circles on a stick), and conductive or bidirectional (two-way arrows) signaling. This circuitry generates the selective responses of the four types of ganglion cells (at bottom) that make up 90 percent of the optic nerve's fibers, which convey visual information to the brain. On (green) and Off (red) ganglion cells elevate their firing (spike) rates when the local light intensity is brighter or darker than the surrounding region. Inc (blue) and Dec (yellow) ganglion cells spike when the intensity is increasing or decreasing, respectively.



SILICON RETINA

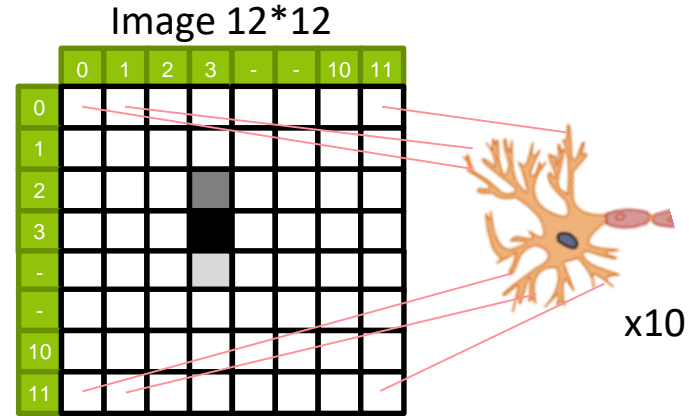
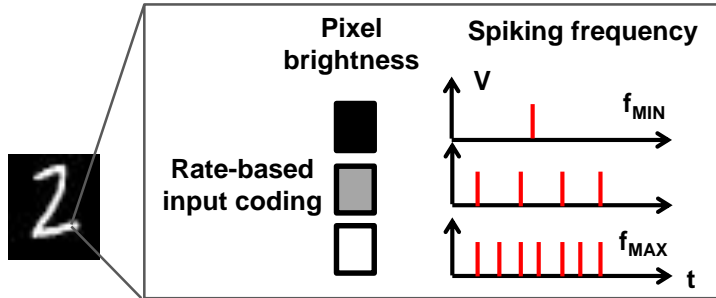
Neuromorphic circuits emulate the complex interactions that occur among the various retinal cell types by replacing each cell's axons and dendrites (signal pathways) with metal wires and each synapse with a transistor. Permutations of this arrangement produce excitatory and inhibitory interactions that mimic similar communications among neurons. The transistors and the wires that connect them are laid out on silicon chips. Various regions of the chip surface perform the functions of the different cell layers. The large green squares are phototransistors, which transduce light into electricity.

SILICON CHIP DETAIL



Source : Scientific American

BIOLOGICAL INSPIRED NEURONES USING OXRAM



- Classification of handwritten numbers
- Small resolution image
 - 12*12 pixels
- Fully-connected network
 - 10 neurones : 1 neurone / class
 - 144 synapses

- Si Real Estate: 1,8 mm²
- Clock frequency: 50 MHz
- 10 neurones
- 10*144 synapses = 11,5 kOxRAMs

→ NEED FOR BETTER INTEGRATION /3D
→ Relative cost of Oxrams vs Neurones

■ Neural network for vision processing

- A SNN layer is often a 2D structure
- Image recognition applications need at least two layers
 - This becomes inherently a 3D structure

■ It lends itself well to a 3D implementation

- Logical layers are mapped to physical tiers

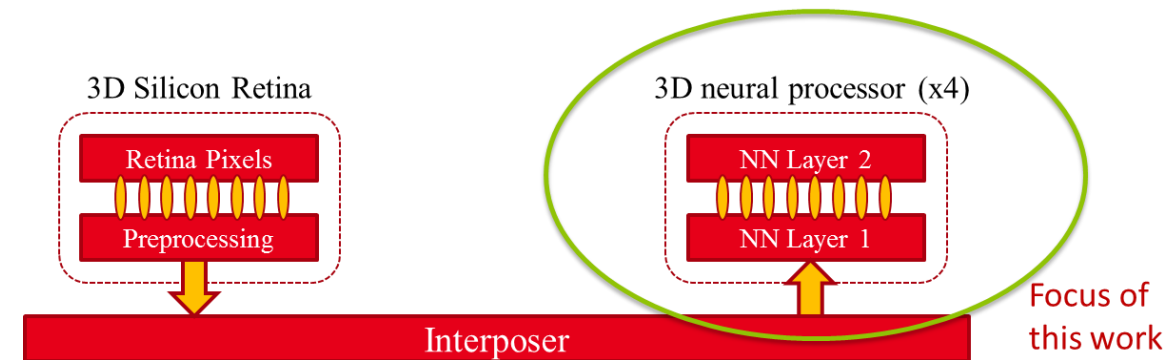
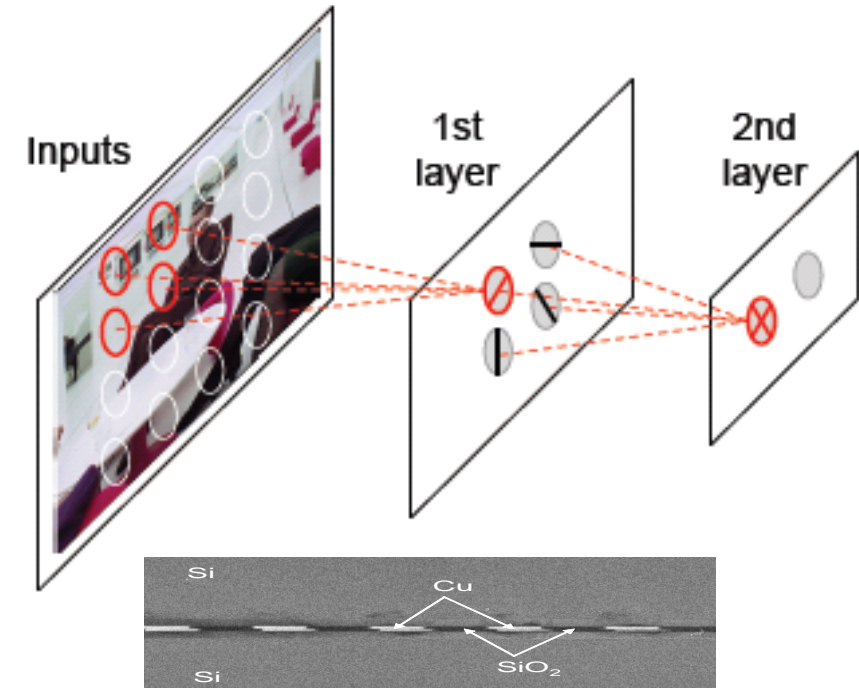
■ Two distinct building blocks

■ A Silicon Retina

- CMOS image sensor tier, with a 256x192 resolution
- Pre-processing tier, which generates spiking events corresponding to changes in pixel intensity

■ A Neural processor

- 'Layer 1' extracts *features*: horizontal, vertical, diagonal segments ...
- 'Layer 2' combines those *features* to extract complex shapes: leads to object recognition

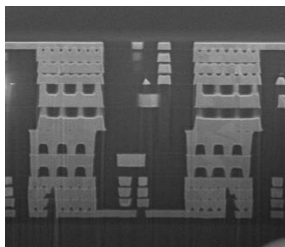


Computer Vision applications

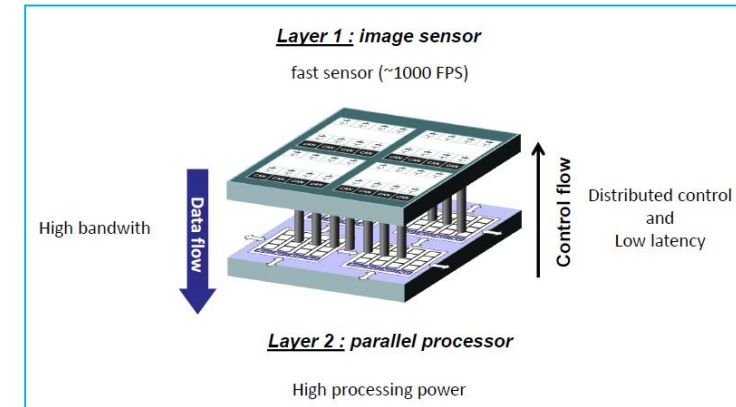
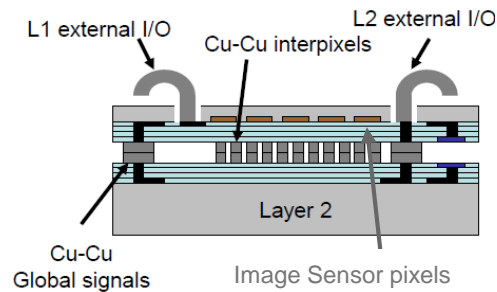


3D-stack : Image Sensor & Processor

- SIMD fully programmable accelerator
- Heterogenous technology, High Sensitivity
- Technology ALTIS 130nm
- CuCu Hybrid Bonding stacking



CuCu, pitch 7 μ m



- ALTIS 130 nm
- Die size : 17 mm x 11 mm
- Layer 1 – image sensor :
 - fill factor > 70 %
 - pixel 12 μ m
 - pixel dynamic 1 to 8 bit
 - > 1000 FPS @ 192x256
 - > 60 FPS @ 768x1024

RETINE - Layer 1 :

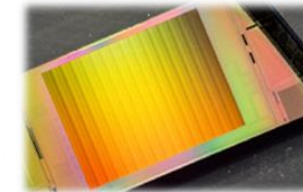
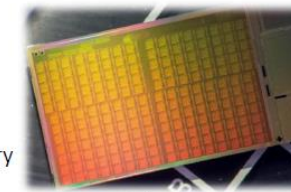


image sensor

- Layer 2 – parallel processor :
 - SIMD matrix of 3072 ALUs (16 x 12 x 16 ALUs)
 - Computing power : 161 GOPS
 - Target : 100 MHz – 175 Mhz – 210 MHz
 - 72 kB distributed memory + 96 kB shared RAM memory

RETINE - Layer 2 :

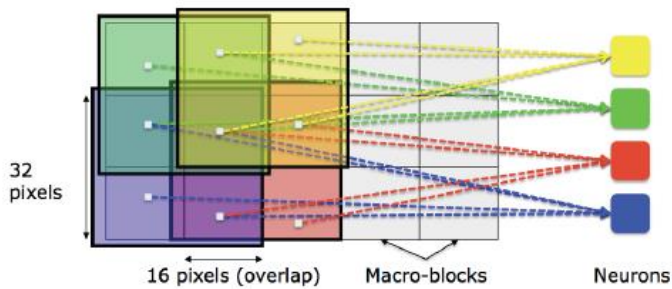


parallel processor

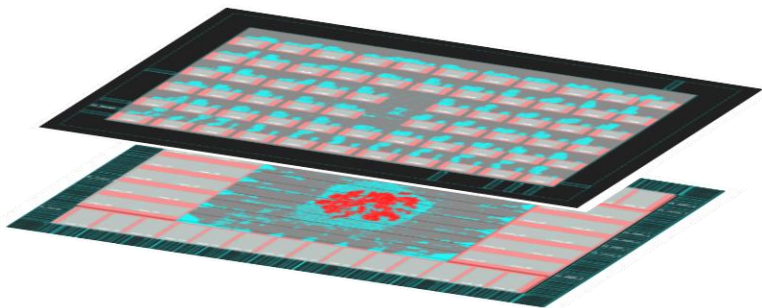
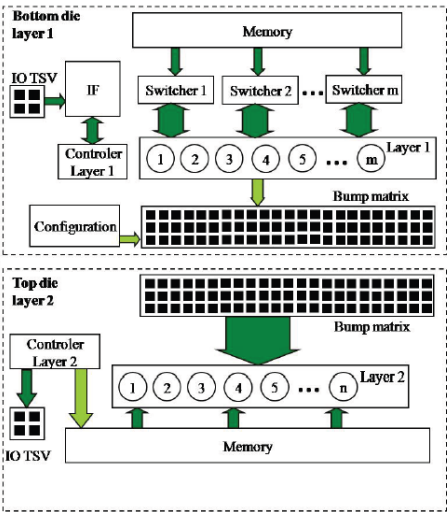
[S.Chevobbe, L. Millet, C. Andriamisaina, M. Duranton, D43D'15]

Neural Networks

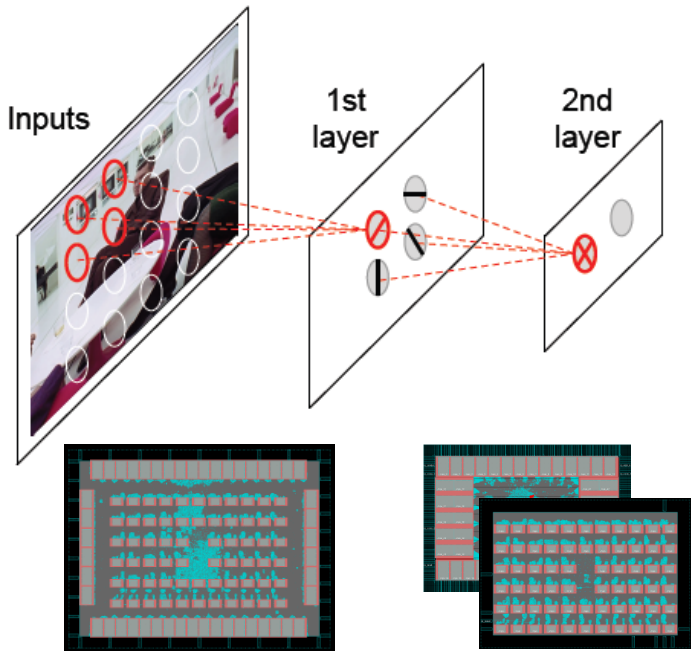
- Classically divided in two layers of computation
- Difficult to implement in 2D, due to high congestions
- Very well adapted to 3D : one neuron layer per die !



*Compared to 2D,
3D offers :
2x better total area
25% better in power*



**More layers ?
Tighter integration of Neuron, Memory, and NVM ?**



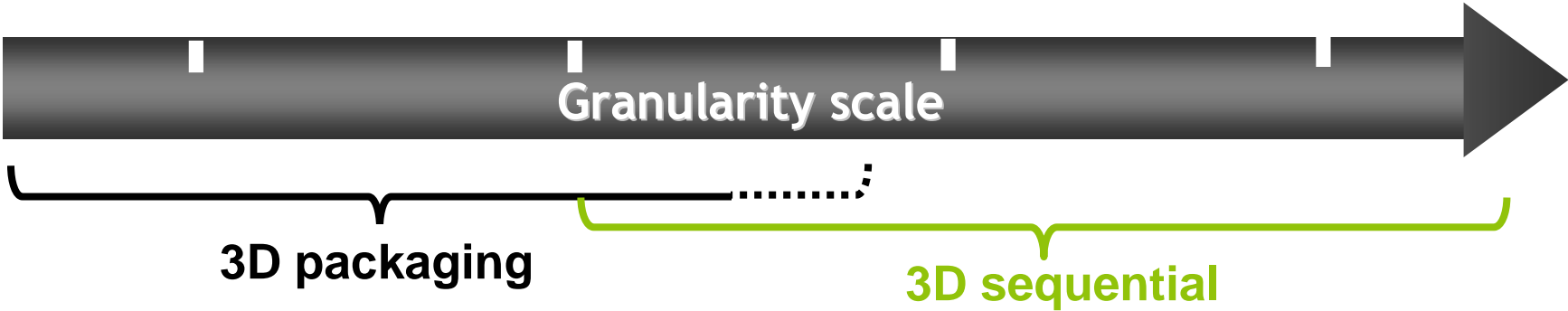
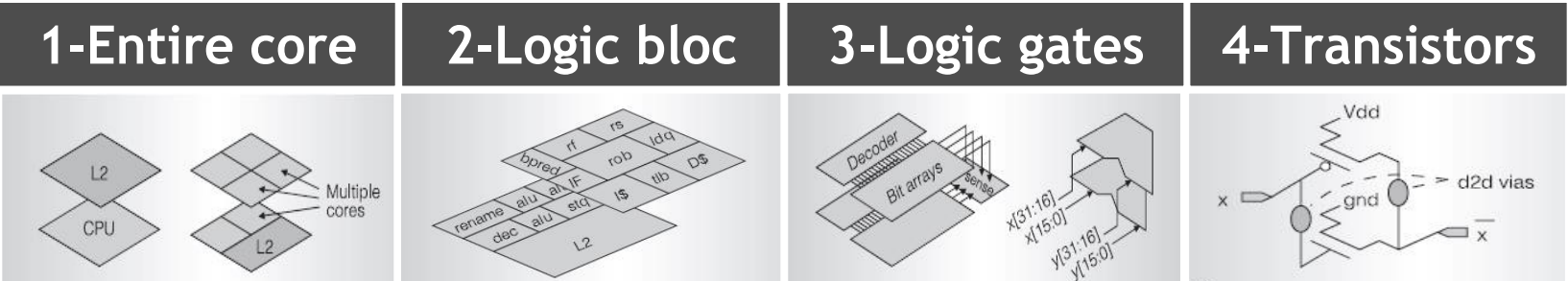
	2D circuit	3D circuit	Gain
Critical path (ns)	9	6.6	-26%
Power (mW)	430	350	-17%
Area (mm ²)	7.9	3.6	-54%
Wires (m)	19.9	15.6	-21%

[B. Belhadj, R. Heliot, A. Valentian, P. Vivet, CASSES'2014]

3D integration

Towards High Density
Interconnects

3D PARTITIONNING LEVELS



NEW ARCHITECTURES:

Partitioning
Alternative to scaling
Shorter Interconnects

DRIVERS:

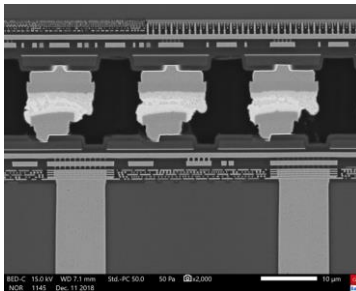
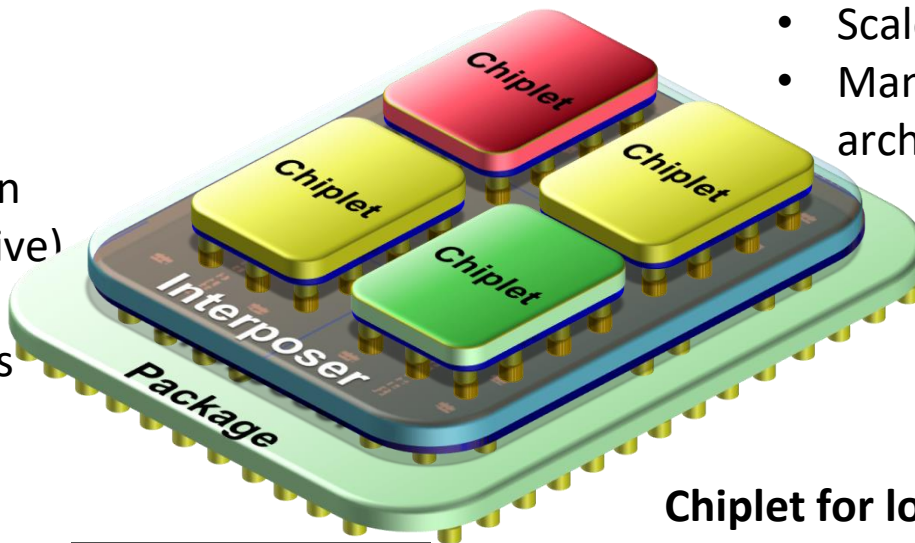
Form Factor
Yield / Cost
Perf / Interconnect / Density
Heterogeneity

3D INTEGRATION: A TOOL BOX FOR NOVEL COMPUTING ARCHITECTURES

Chiplet concept: Memory proximity – high bandwidth

Interposer for specialization:

- System-in-Package, Silicon (Passive or active) photonic
- Heterogeneous integration enablement
- Application specific



Intact

Integration for high performance:

- Scale-out
- Many-core architecture

Chiplet for low cost:

- Small to medium size chips (1 cm² max)
- Advanced technology node
- Generic
- High volume

TSV TECHNOLOGIES

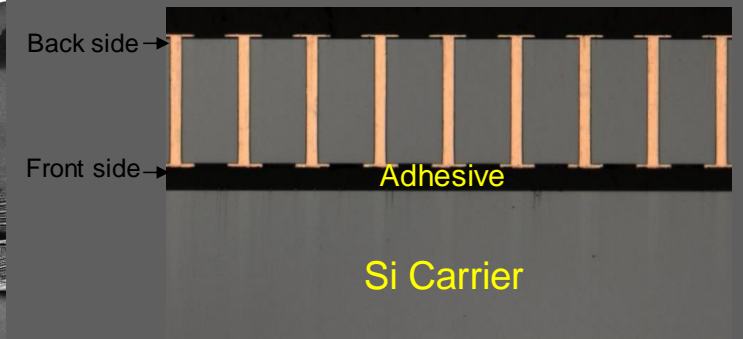
Smaller diameters



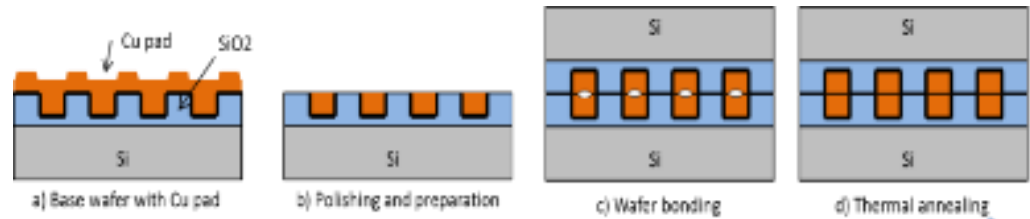
Higher Aspect Ratios



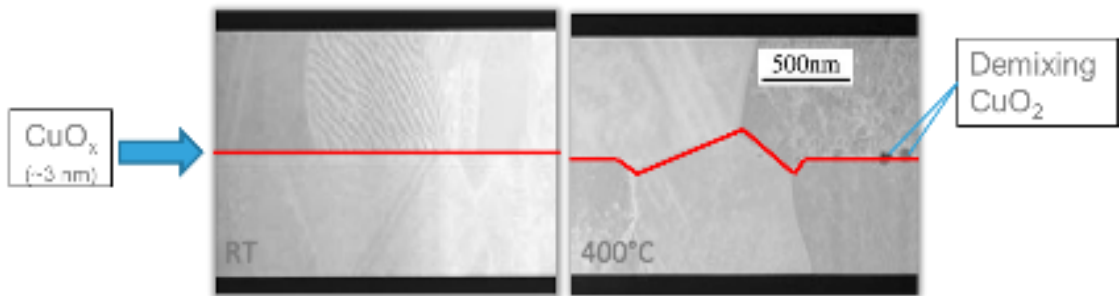
TSV 10*120μm



3D INTEGRATION: HIGH DENSITY HYBRID BONDING



PROCESS FLOW



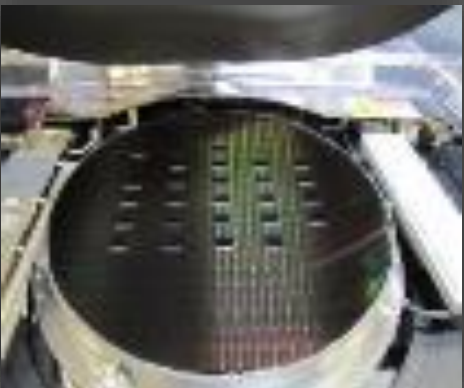
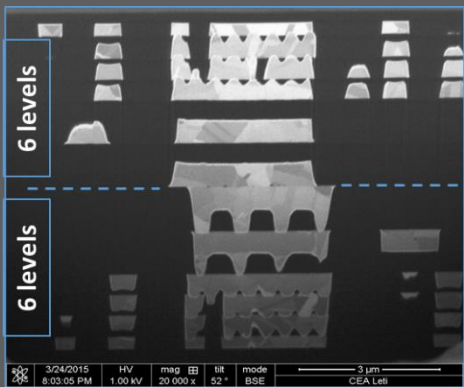
Solution to provide small pitch integration capability for application with high density interconnect requirements



High reliability solution for multi-stacks, high temperature & secure solutions



Best of Class Signal to Noise figures



MAIN CHALLENGES

Wafer to wafer

Towards smaller pitch & multi layers

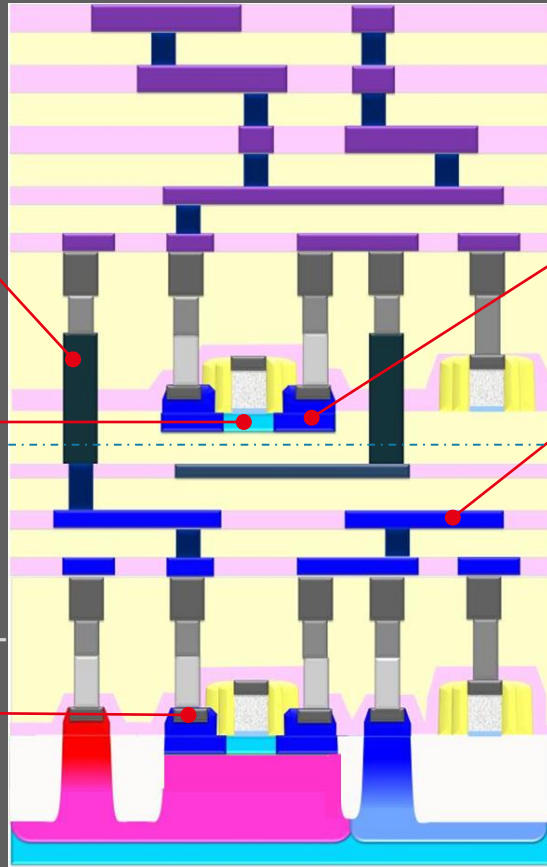
Die To Wafer

Single Die Handling

Mid Term Self Assembly

Increasing throughput

3D SEQUENTIAL INTEGRATION: COOLCUBE...



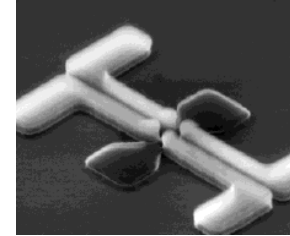
CHALLENGES

- Process Flow Validation
 - Low Temperature Epitaxy
 - Low K Spacer
- Cost Analysis
- Design Flow and Tools

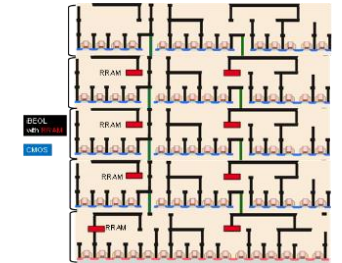
... A LARGE RANGE OF APPLICATIONS



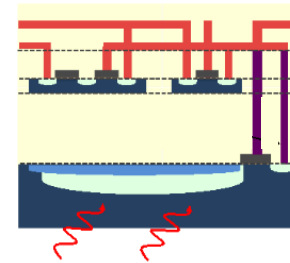
QUANTUM COMPUTING



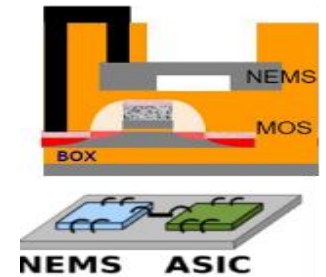
NEUROMORPHIC COMPUTING



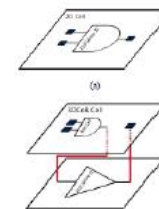
3D PIXEL



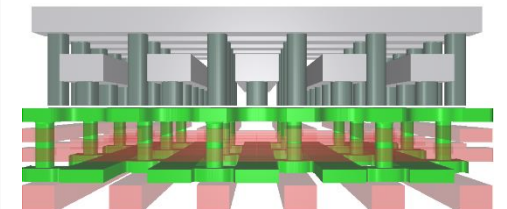
SENSORS



HV ANALOG



N/P PARTITIONNING



MEMORY: A UNIQUE VALUE PROPOSITION

DEFINITION OF TECHNOLOGY
SPECIFICATIONS

Large variety of
materials available

GeSbTe
SiOx
TaOx
ZrO2
AlOx
VOx
HfAlxOy

GeAsSbTe

Large variety of
Memories available



pSTT-Magnetic RAM
Conductive Bridge RAM
Oxide Resistive RAM
Ferro-electric RAM
Phase – Change Memory

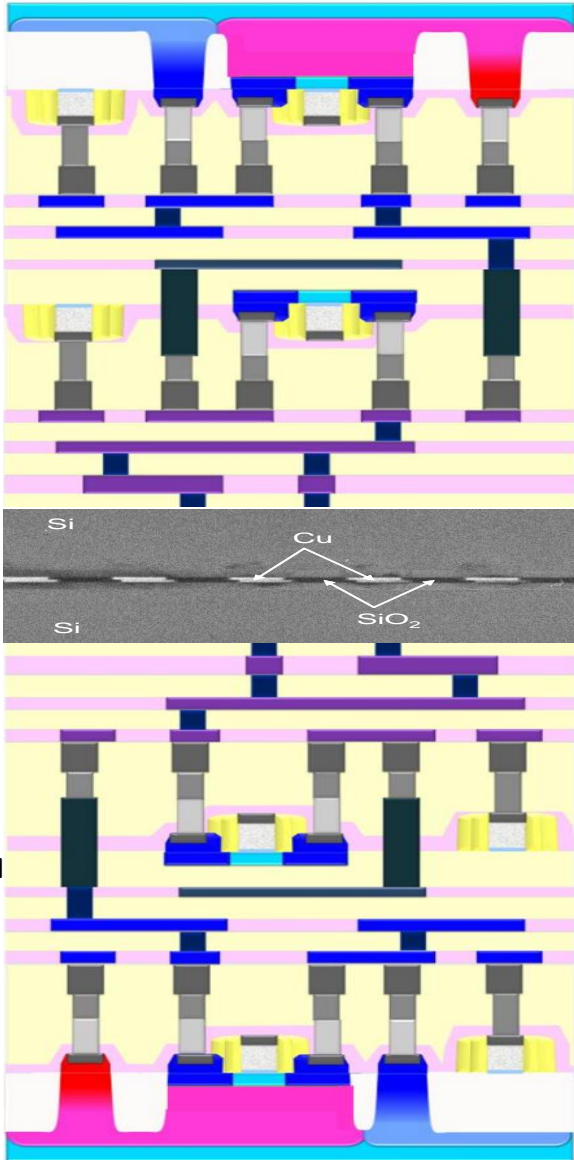
DESIGN ENABLEMENT

MODULE DEVELOPMENT

200/300 MM
INTEGRATION

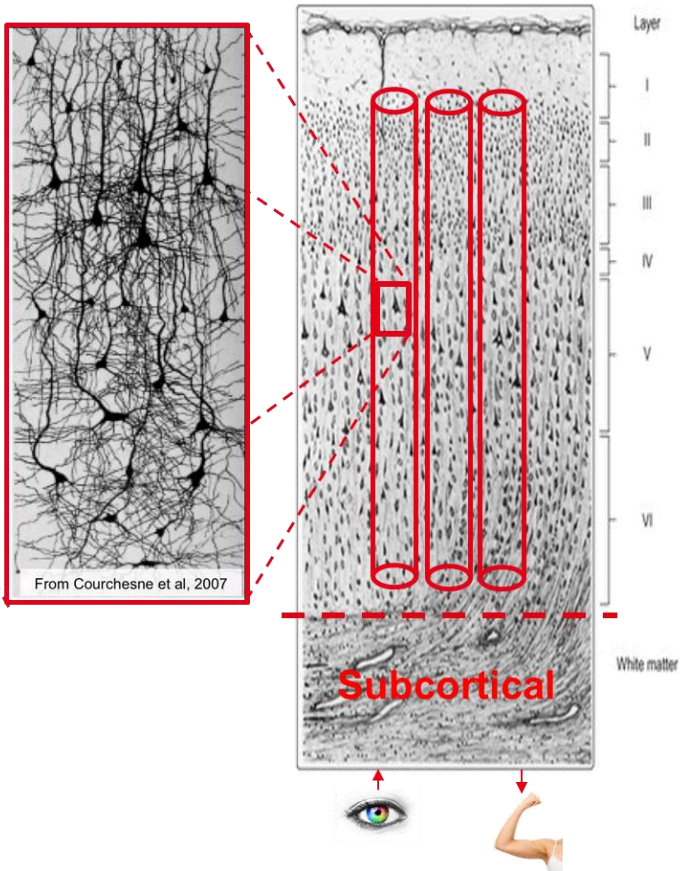
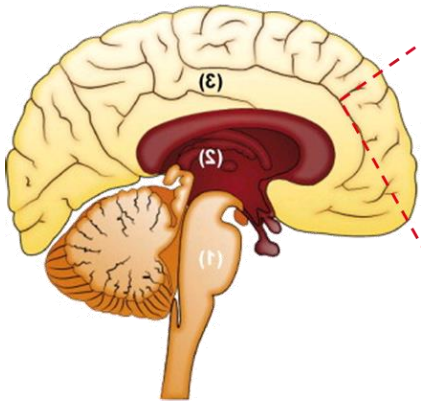
MODELING,
SIMULATION & NANO-
CHARACTERIZATION

TEST & CHARACTERIZATION



EXPLORING THE VALUE OF

- STACKING TWO DOUBLE LAYERS OF LOGIC + RRAM STACKS...



... TO MIMIC CORTICAL COLUMNS

■ DEEP NEURAL NETWORKS, AI ALGORITHMS

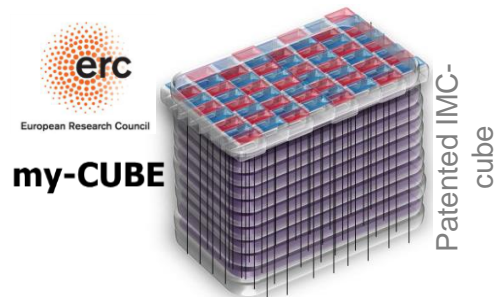
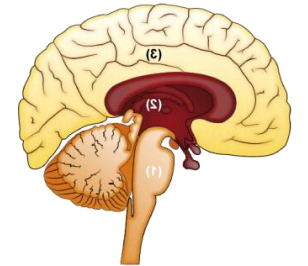
- Are already supported by 3D technologies
- Digital Architectures using Interposer and HBM memories



■ NEED TO GO FURTHER



to pursue integration and reduce power consumption for embedded applications

- Start with biological inspired systems
- Analog circuitry
- Non volatile memories
- High density 3D → Hybrid bonding fine pitch and/or Monolithic Integration




➔ ***Interaction Architecture, design, technology***



-  Exploration of New Materials & Process Integration
-  Joint Developments Programs with Equipment Manufacturers
-  Joint Technology Development & Exploration of New Concepts

Prototyping / MPW Shuttles

-  Technology Transfer to Fab
(IDMs, Foundries)

