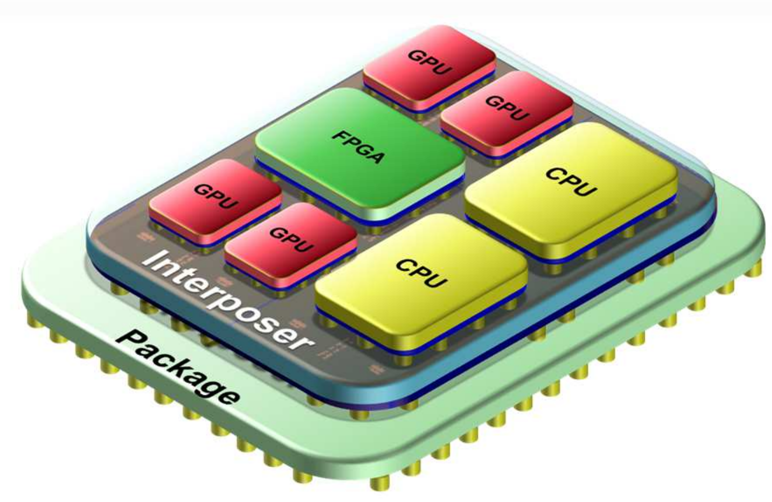


FROM 3D TECHNOLOGY TO 2.5D AND 3D MANY-CORE ARCHITECTURES

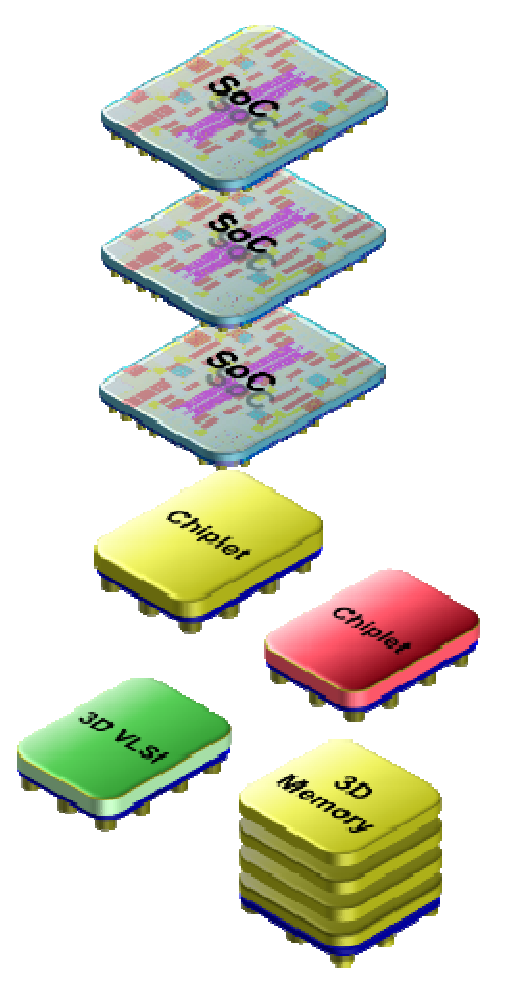
Pascal Vivet | Cea-Leti | 21-22 Sept 2016

MCSoC'16 Conference, Lyon, France

CONVENTIONAL 2D SOC

SINGLE DIE :
- MEMORY WALL
- POWER WALL
- COMPLEXITY WALL

CHALLENGES OF HIGH PERFORMANCE COMPUTING



How to fit more ?

… More cores

… More Memory

… Memory closer to core

… Computing Model

… Power Efficiency

… Thermal Dissipation

& cost !

Computing Applications

Going Forward – What Options Do We Have?

More Moore

Number I/O per cm²

20.000

10.000

1500

500

Scaling of Transistor Nodes => I/Os Density Increase

Technology Node CMOS (nm)

100 80 60 40 20

CMOS90 CMOS65 CMOS45 CMOS28 CMOS16 CMOS14

FinFET

2D SOC

"All-in-One chip system integration"

High Speed Logic | eDRAM eNVM | Critical Leakage | Analog/RF | High Voltage | SRAM

High Voltage Operation, data transfert

Chip area ↑, Cost ↑, Time to Market ↑

More than Moore

3D Packaging

Time to Market ↓, Cost ↓, Performance ↑

Size ↓, Flexibility ↑

European 3D Summit – Grenoble, France

3D Packaging @ the Forefront of Semiconductor Industry

10

YOLE Développement

CHIPLET PARTITIONING AND 3D INTEGRATION :
➔ POWER EFFICIENCY, SCALABILITY, MODULARITY FOR COMPUTING

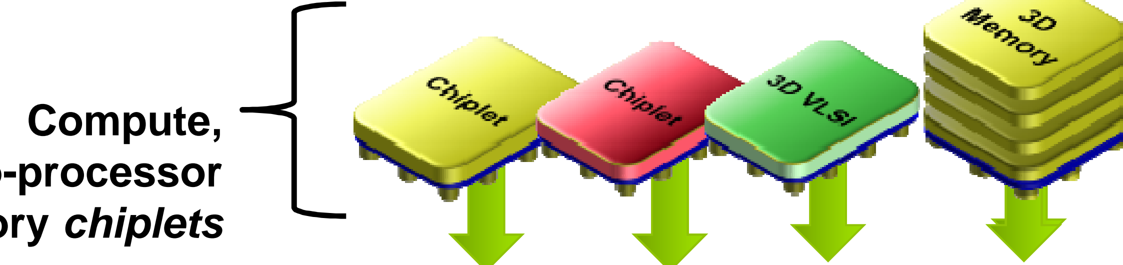
Chiplets in FDSOI Technology :

- Power Efficiency
- Ultra Wide Voltage Range
- Body Biasing for logic boost and leakage control
- Reduced cost

Computing is highly segmented:

➢ **computing continuum**

Compute, co-processor memory *chiplets*



Interposer Integration

Photonic Interposer
Active Interposer
Passive Interposer
System-in-Package

Performance

Server
HPC
Computing Continuum
Embedded
Micro-server

Specialization

[D. Dutoit, VLSI Symp'16 tutorial]

OUTLINE

- **Introduction**
- **3D Technology : an introduction**
- **State-of-Art on Circuits & Applications**
- **3D Circuit Demonstrators**
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- **New Trends with High Density 3D technologies**
- **Conclusions & Perspectives**

OUTLINE

- Introduction
- 3D Technology : an introduction
- State-of-Art on Circuits & Applications
- 3D Circuit Demonstrators
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- New Trends with High Density 3D technologies
- Conclusions & Perspectives

3D technology : Some Definitions

3-D chip stacking

With the possibility of backside routing

Passivation

BEOL

FEOL

Redistribution layer (RDL)

Si

Chip 2

Through Si via (TSV)

Si

Chip 1

Chip-to-chip interconnection (with underfill)

Chip-to-package interconnection

pitch

back-end / "Advanced Packaging"

BEOL

front-end

FEOL

Legend

lead-free solder bump

seal layer (nitride or oxide)

PSG

SiN seal layer

Cu 5

SOD

Cu 4

Cu 4

Cu 4

Ta/TaN barrier layer

SiC etch stop layer

Cu 3

SiC etch stop layer

Cu 2

Cu 2

SOD

PE-TEOS

SiC seal layer

SOD

tungsten

Source : wikipedia

TSV : Via first, via middle and via last ?

	CMOS process		Backside process	Dicing Stacking
	FEOL + *Middle End*	BEOL		
Via First (W, Poly Si)	*TSV process*; *Transistors + W contacts*	*Cu interconnects + passivation*	*Temp carrier*	
Via Middle (Cu)		*TSV process*	*Temp carrier*	
Via Last (Cu)			*TSV process*; *Temp carrier*	

3D Stacking strategy : Wafer ? Die ?

Wafer-to-wafer (WtW)	***Easier to process*** but require the same die size with very good yields
Die-to-Wafer (DtW)	***Possibility to select the known good dice***
Die-to-Die (DtD)	***More flexible***

3D Si technologies – focus on interconnection

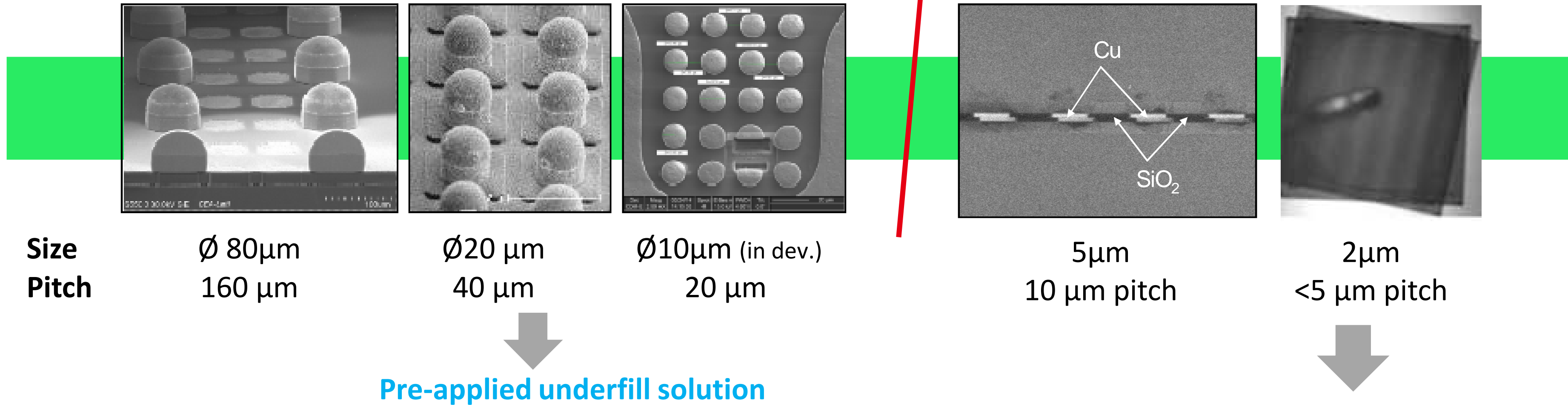
3D SILICON TECHNOLOGY / FINE PITCH CHIP-TO-WAFER ROADMAP

Cu/Sn solder µbumps

with pre-applied underfill

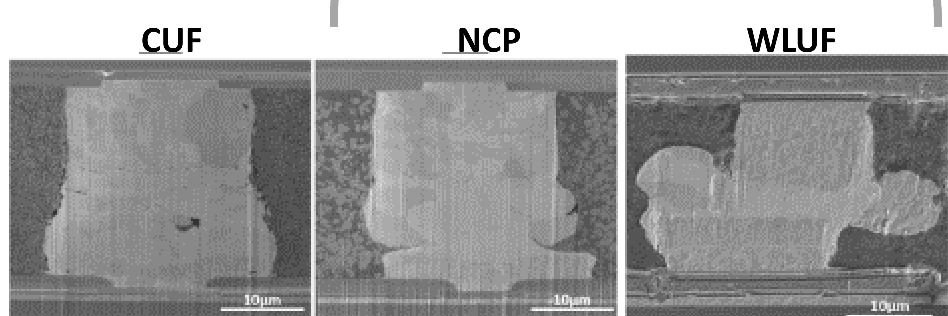
Hybrid Cu-SiO2 bonding

Glue-less **and self-alignment**

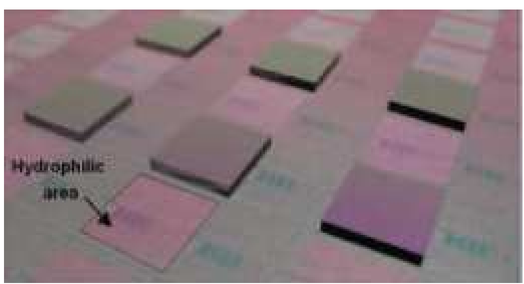


Size	Ø 80µm	Ø20 µm	Ø10µm (in dev.)	5µm	2µm
Pitch	160 µm	40 µm	20 µm	10 µm pitch	<5 µm pitch

Pre-applied underfill solution



A. Garnier et al., ECTC 2014



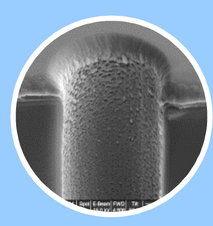
<1 µm alignment accuracy using self-assembling with hybrid bonding

TSV High Aspect Ratio, Metallization Challenges

Silicon thickness ? → key contributor for thermal & stress management
Need more agressive TSV aspect ratio for trading-off perf & thermal/stress

Barrier

MoCVD TiN promising 30% step coverage @ 20:1

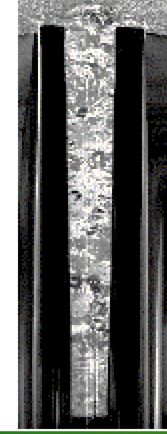


Seed

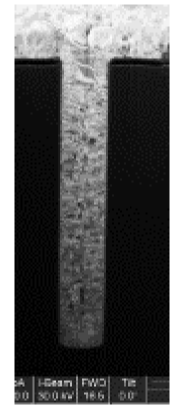
Positive evaluation of electrografting process @ 15:1

Filling

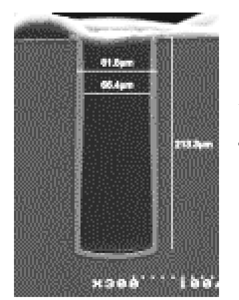
Gen IV chemistry for AR > 12:1



TSV middle
Target : Interposer
10x120 µm TSV



TSV middle
Target : High Density
2 x 15 µm TSV



Via Last High AR
Target : Heterog. Integrat.
60 x 200 µm TSV
Courtesy of Th. Mourier

3D TECHNOLOGY : DESIGN CHALLENGES ?

Co-design architecture/ die/package

Sensor Chip
MEMS
RF-IC
MM IC
Analog LSI
Power IC
Control IC
Logic IC
Flash Memory
DRAM
Cache SRAM
CPU

Architecture partitionning

TSV & Interconnect yield & testability

3D Challenges

Scalability and Modularity

Test Access
DFT
Reliability

Heterogeneous technologies
Cost & Yield

CAD tool maturity

Link Performances & Energy efficiency

Thermal Dissipation

Sign-off
Place & Route
Power Analysis

Mbps & pJ/bit

Power-In
Thermal-Out

OUTLINE

- Introduction
- 3D Technology : an introduction
- State-of-Art on Circuits & Applications
- 3D Circuit Demonstrators
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- New Trends with High Density 3D technologies
- Conclusions & Perspectives

2.5/3DIC Commercial Announcements!

SAMSUNG
DDR4 3D
Dual Inline Memory Modules (RDIMMs)

ALTERA
Altera 10 Generation FPGA using HMC
Arria 10
Stratix 10

FUJITSU
Next Generation PRIMEHPC POST FX10 CPU memory board using 8 HMC

intel

SK hynix
HBM

Micron
HMC

2011
2012
2013
2014
2015
2016
2017
...

ST
mCube

TOSHIBA
SONY
XILINX
BOSCH
Invented for life

First Heterogeneous 3D FPGA Virtex-7 H580T

Tezzaron
DiRAM4
picocomputing
EX-800 Blade Server using HMC

AMD
AMD R9 390X
Graphics product with HBM built with 20nm technology

TOSHIBA
Stacked NAND Flash

matrox
Graphics for Professionals
AMD

NVIDIA
Nvidia Pascal Graphics Module

New second generation Xenon Phi processor "Knights Landing" using HMC

- More and more products using TSV

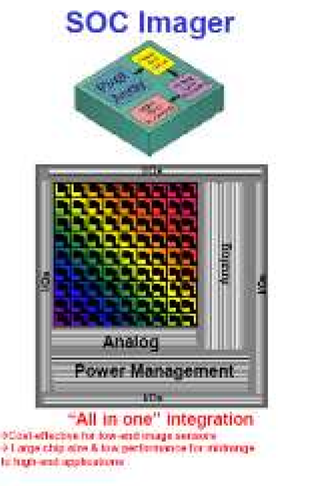
42

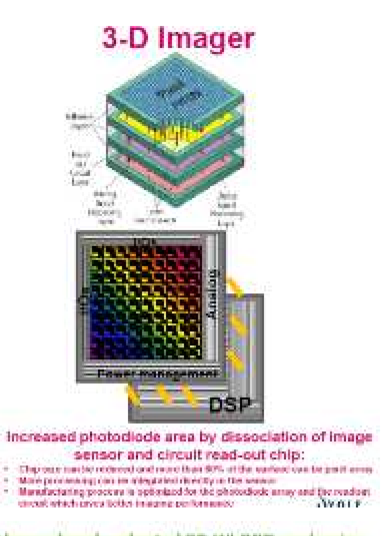
European 3D Summit – Grenoble, France
3D Packaging @ the Forefront of Semiconductor Industry

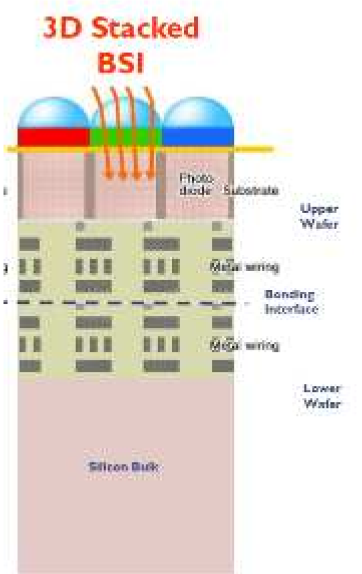
YOLE Développement

3D STACKED BACKSIDE IMAGERS

➔ Most industrial players have adopted 3D Stacked BSI





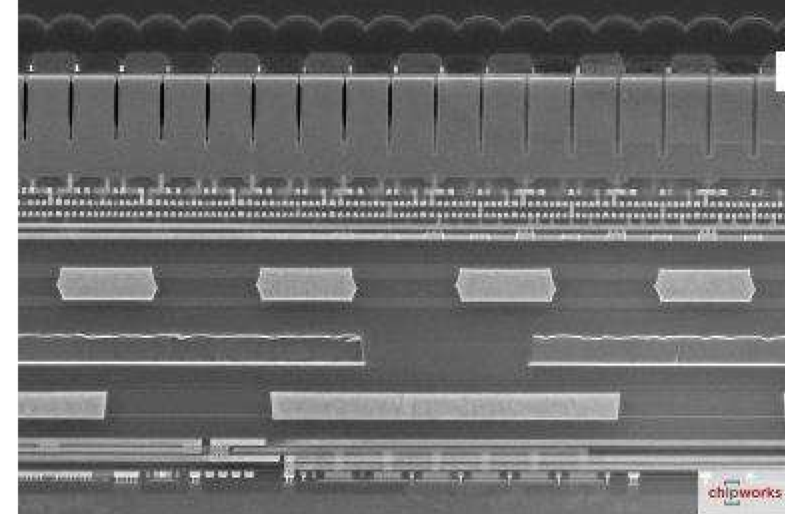


Optimized pixel array

Optimized logic

Smaller size

Added functions
With better performance

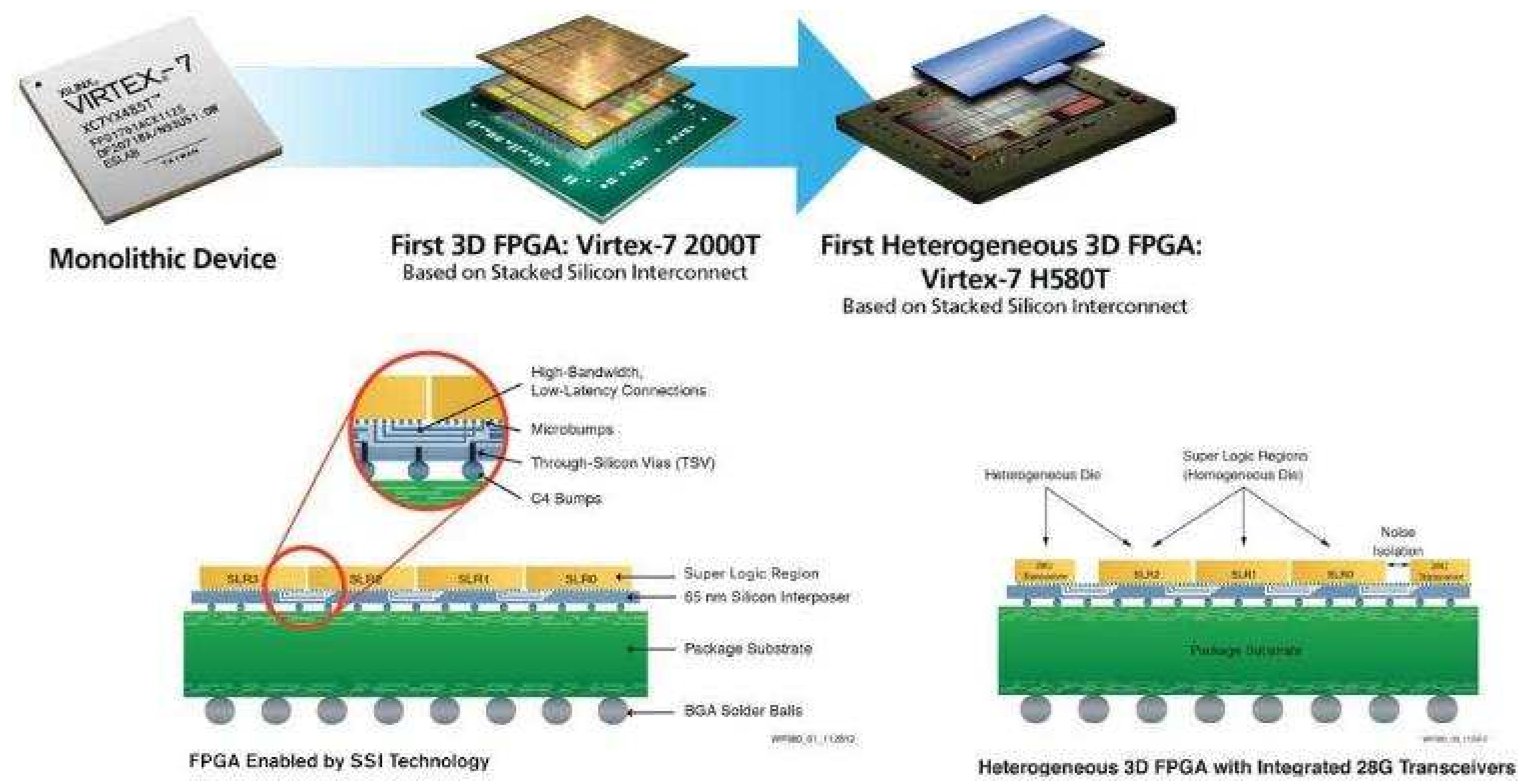


Sony IMX260 in Samsung Galaxy S7:

Source : JL Jaffard, Imaging Technologies and applications: Pioneers of TSV and 3D technologies, TSV Summit 2016

INTERPOSER (OR 2.5D) : XILINX VIRTEX 7 SERIE

- **XILINX: The first 2.5D interposer product**
  - FPGA is split in slices, stacked onto an interposer
  - Main advantages : gain in yield for very large dies
  - **A full product family & roadmap is available**
  - Xilinx is now going to heterogeneous dies (for fast IO's)



FROM SINGLE DRAM USING POP …. TO ….. 3D DRAM !

LP DRAM

Package

Processor

Package

(WideIO2 exemple)

Wide IO DRAM

Wide IO DRAM

Wide IO DRAM

Wide IO DRAM

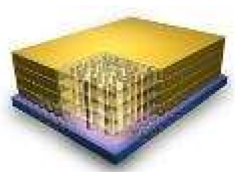
TSV

Processor

Package

TSV: Through Silicon via

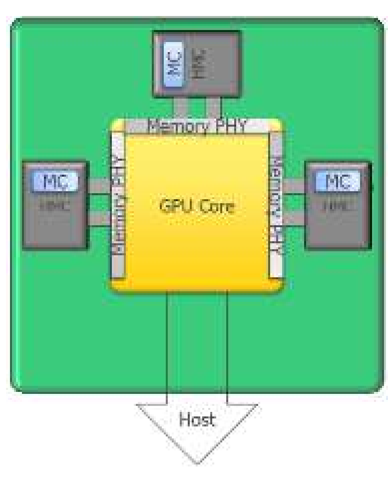
3D DRAM : COMPARISON

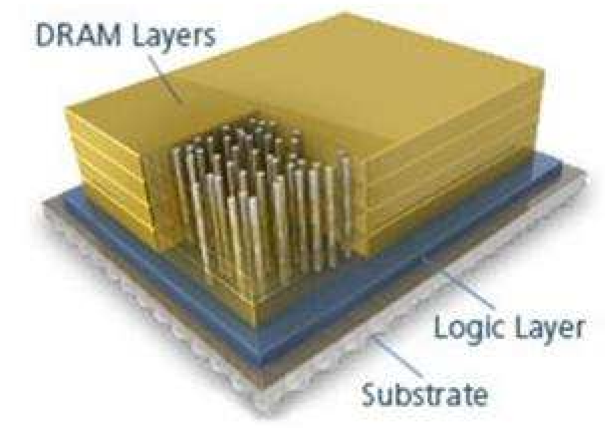


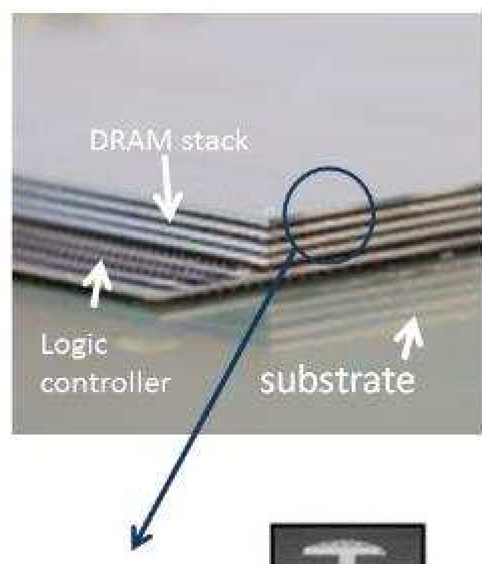
	LPDDR4	WideIO/2	HBM	HMC	DiRAM4
		JEDEC; SK hynix	JEDEC; SK hynix, SAMSUNG	Hybrid Memory Cube Consortium; Micron	Tezzaron Semiconductor; Tezzaron Semiconductor
Interface type	parallel	wide data	wide data	serial	wide data or serial
Data bus	16b DDR	64b DDR	128b DDR	16 lanes	64b
Channel	2	4-8	8	4-8	
I/O bandwidth	3.2Gbps @1600MHz	0.8Gbps @400MHz	1-2Gbps @500-1000MHz	10-15Gbps	
Total bandwidth	12.8GBps	25.6-51.2GBps	128-256GBps	160-320GBps	2TBps
Capacity	16GB	16GB	32GB Currently 1GB (Gen1) Next 4-8GB (Gen2)	32GB Currently 2-4GB	8GB
Total I/O	66	776	1616	256-512	
Integration / Packaging	POP, MCP	3D	2.5D	MCP	
Computing-In-Memory		NO	NO	**YES**	NO

3D DRAM MEMORY STACKING : HMC VS HBM

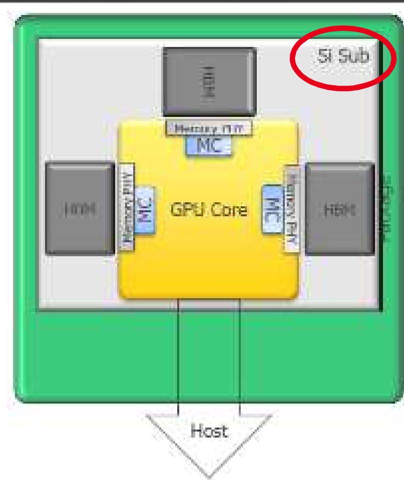
HMC (Hybrid Memory Cube), ex : Micron
3D stack only, no passive silicon interposer

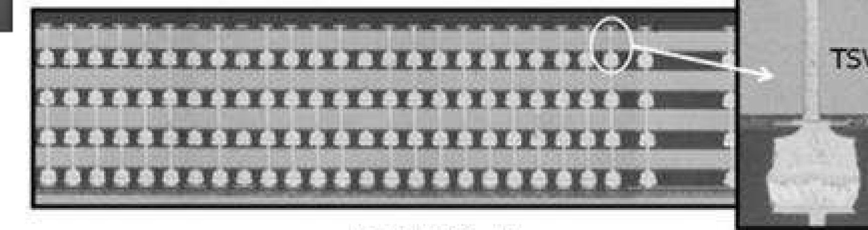






HBM (High Bandwidth Memory), ex : SK Hynix
3D stack + High Density passive silicon interposer



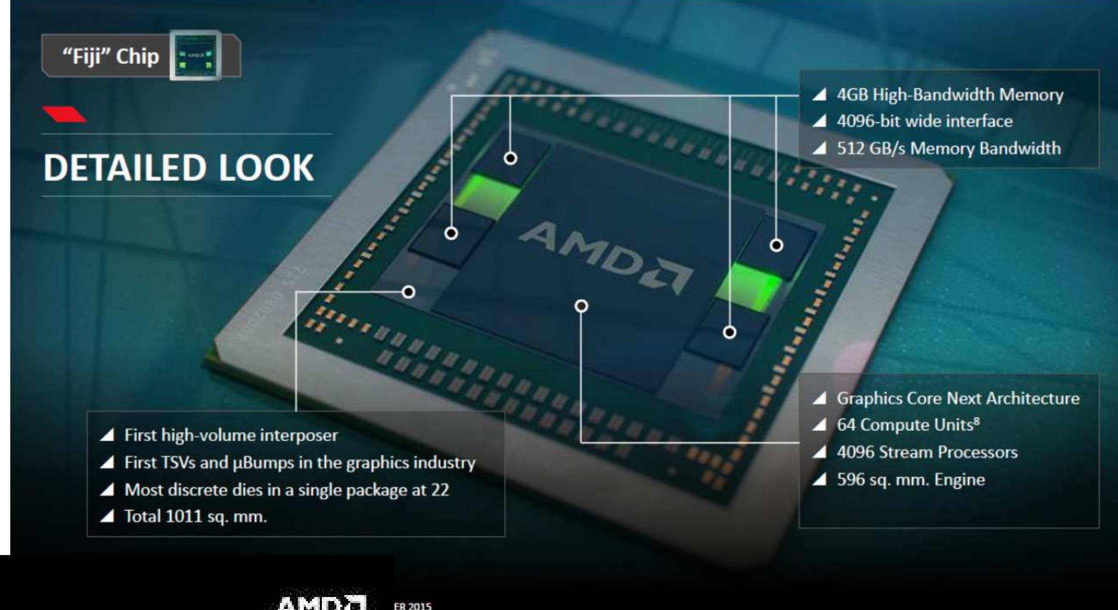


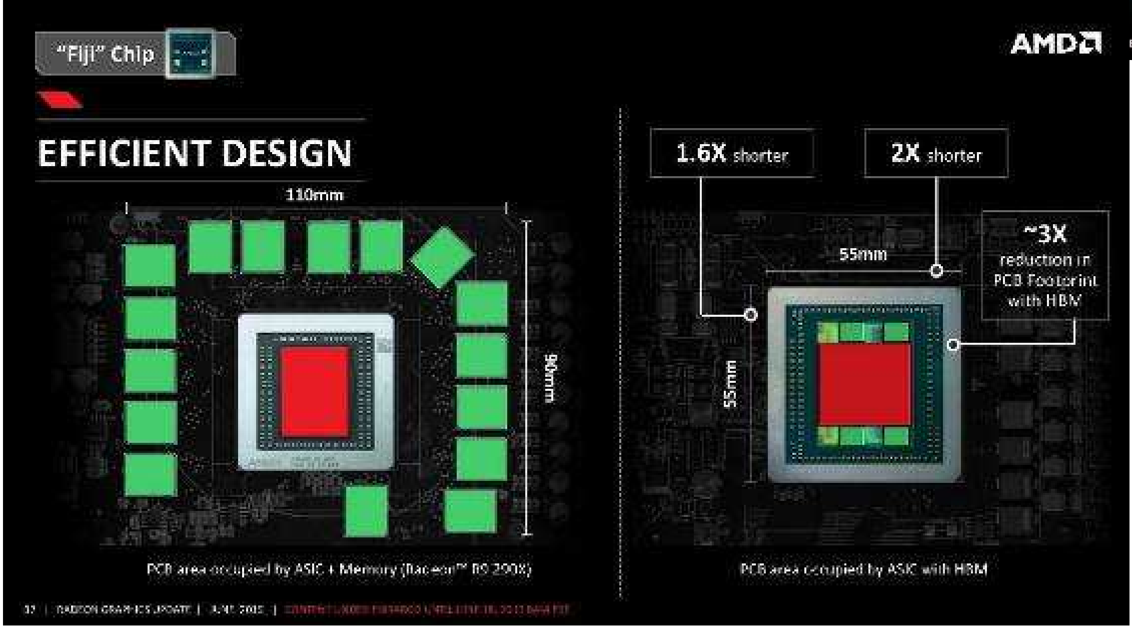
Source : Micron

HBM PRODUCT EXAMPLES (1/2)

AMD

- AMD has presented in 2015 the first commercial GPU product including HBM Gen1 memories
- "Fiji" chip is part of the Radeon Fury graphics card series





Combination of:
- HBM DRAM memory (3D)
- Silicon interposer (2.5D)

- x3 Performance per Watt
- 60% gain in Memory BW
- 95% less PCB area versus GDDR5

HBM PRODUCT EXAMPLES (2/2)

FPGA

- Altera integrates HBM2 memories from SK hynix in Stratix 10 products

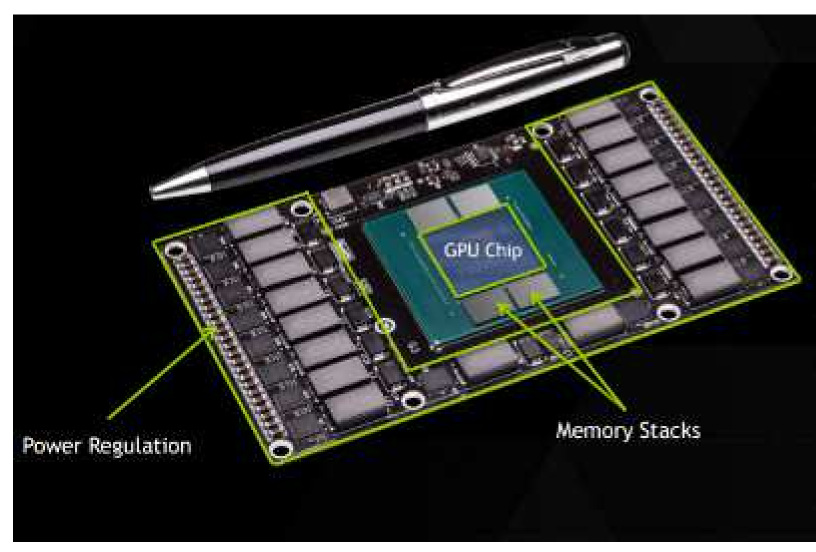
- Integration is performed thanks to the EMIB (Embedded Multidie Interconnect Bridge)

Heterogeneous Integration using EMIB Technology

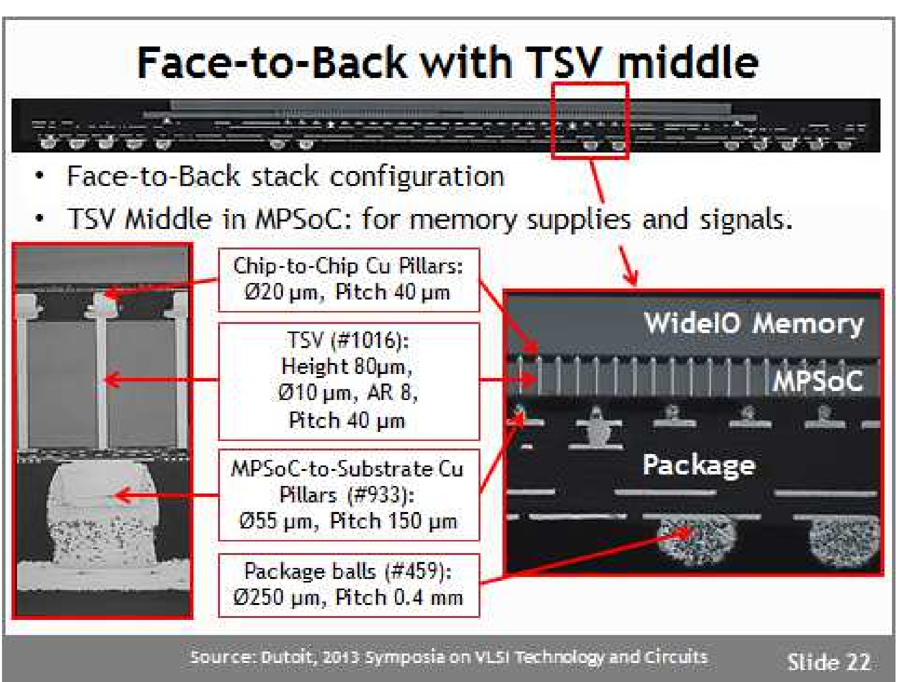
Standard Package Traces
(2) Transceiver Die
Package
Bumps
(5) Package Lid
(1) Stratix 10 FPGA
(4) Package Balls
(3) Package Substrate
(6) EMIB
Circuit Board

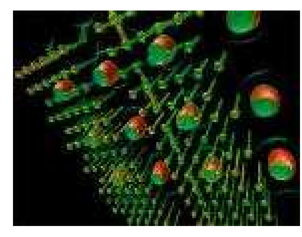
GPU

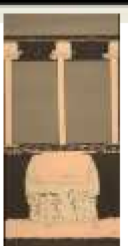
- NVIDIA will integrate HBM2 memory from Samsung in the “Pascal” GPU module expected in 2017.



DRAM-ON-LOGIC : WideIO1 DRAM EXEMPLE







Comparison with LPDDR3

➢ 4x gain in power efficiency with 3D-TSV interconnect

Memory Type	LPDDR3 - [1]	WideIO - This work
Package	PoP / Discrete	3D-IC
BW (Gbyte/s)	6.4 GB/s	12.8 GB/s
Total power		293 mW*
VDD-MPSoC MPSoC power		121 mW*
VDD-Mem Memory Power		81 mW*
VDD-I/O I/O power		91 mW*
I/O power efficiency	3.7 pJ/bit**	0.9 pJ/bit*

4x gain

[1] Yong-Cheol Bae, et al. "A 1.2V 30nm 1.6Gb/s/pin 4Gb LPDDR3 SDRAM with input skew calibration and enhanced control scheme," ISSCC-2012.

Measurements conditions: * at speed (200MHz) 13N MBIST, 80°C
** Read, 5pF load without ODT

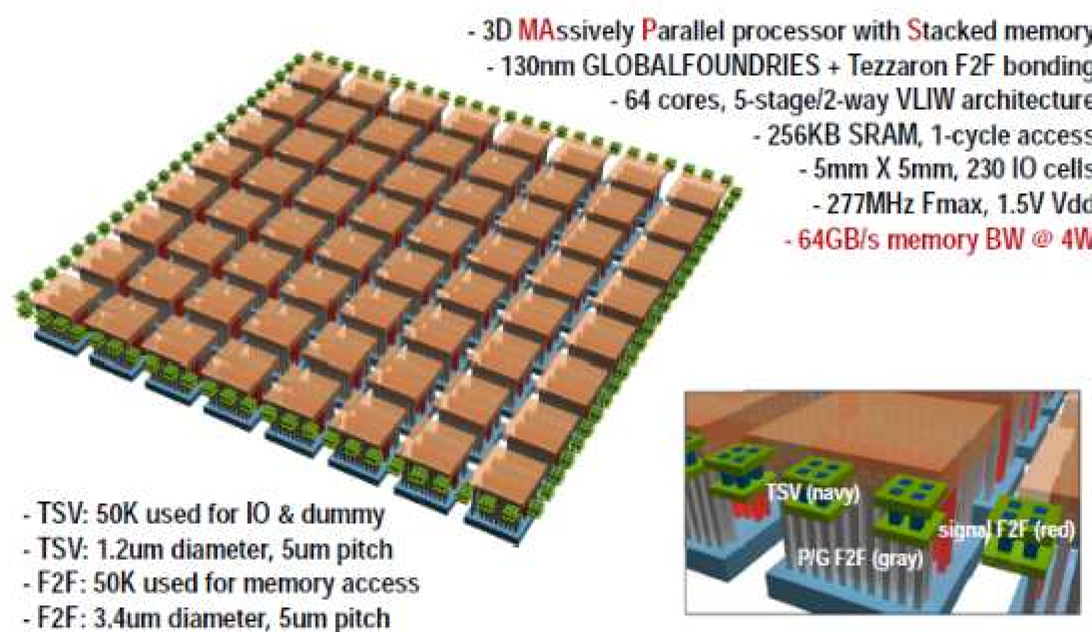
Source: Dutoit, 2013 Symposia on VLSI Technology and Circuits — Slide 24

3D Integration

Source: D. Dutoit, VLSI'13

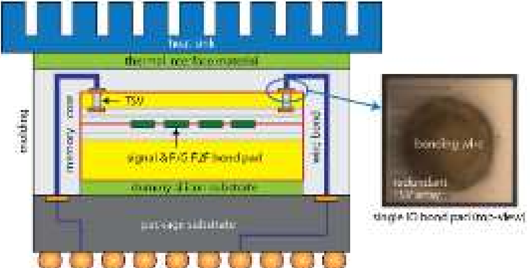
Comparison with LPDDR3

SRAM-ON-LOGIC : 3DMAPS MULTI-CORE



64 Cores,
Split in 2 layers
CPU ⇔ SRAM,
5 stage VLIW pipeline,

- 2 logic tiers, face-to-face bonded
  - Top die thinned to 12um, bottom die is 765um
  - GLOBALFOUNDRIES 130nm technology + Artisan library/IP



I/O bond pad
dummy TSV
TSVs
core die
F2F bond pad
memory die

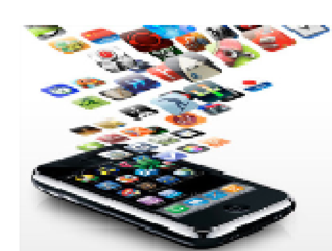
[ISSCC'2012, GeorgiaTech]

OUTLINE

- Introduction
- 3D Technology : an introduction
- State-of-Art on Circuits & Applications
- 3D Circuit Demonstrators
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- New Trends with High Density 3D technologies
- Conclusions & Perspectives

A 3D ASYNCHRONOUS NOC
FOR ENERGY EFFICIENT MULTI-CORE ARCHITECTURES

Energy Efficient Multi-Core



- Performances adaptation wrt. application requirements
- High energy efficiency : cores & system communications

Design Challenge ?

- *high bandwidth & energy efficient communication infrastructure*

Use 3D technology for :

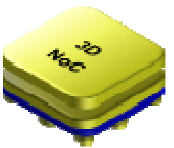
- Logic-on-Logic partitioning,
  to scale delivered performances
- Reduce inter-chip communication
  power consumption

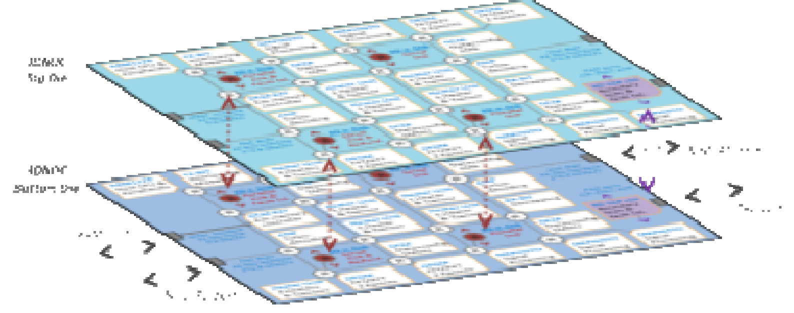
From 2D to 3D NoC

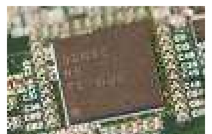
From 2D to 3D Network-on-Chip

- Scalable and modular chip-to-chip communication
- **Target both homogeneous & heterogeneous cores & technologies**
- **Asynchronous logic** avoids global clocking, robust to thermal variations

3DNOC CIRCUIT : A LOGIC-ON-LOGIC MULTI-CORE







- **3D Network-on-Chip based multi-core**
  - Heterogeneous multi-core,
    MIMO 4G-Telecom application
  - Stack 2 similar dies on top of each others
  - No global clock, **robust asynchronous 3D links**
  - Serial link for throughput / #TSV trade-off
  - 3D-DFT & Fault Tolerance Scheme

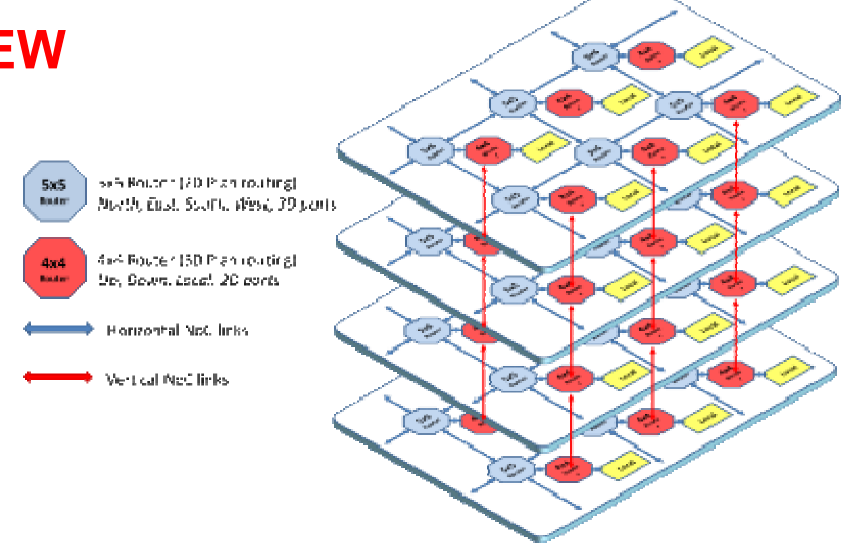
	GeorgiaTech ISSCC'2012	*Kobe Univ. ISSCC'2013*	*This Work*
Architecture	Cache-on-CPU Manycore	Memory-on-Logic 1 layer DRAM	Logic-on-Logic 2 layers 3DNOC
Process & 3D technology	130nm F2F CuCu	90nm F2B TSV	65nm F2B TSV
3D Bandwidth	277 Mbps	200 Mbps	**326 Mbps**
3D I/O Power	-	0.56 pJ/bit	**0.32 pJ/bit**

- **3D Link Performances**
  - Fastest link, +20% (326 Mflit/s)
  - Best Energy Efficiency, +40% (0.32 pJ/bit)
  - Self-Adaptation to Temperature, a strong 3D concern

[P. Vivet et al. ISSCC'16]

An efficient 3DPlug (asynchronous 3DNOC including test & fault tolerance**):**
a first step towards **3D-based computing architectures**

3D NOC & 3D LINK : OVERVIEW

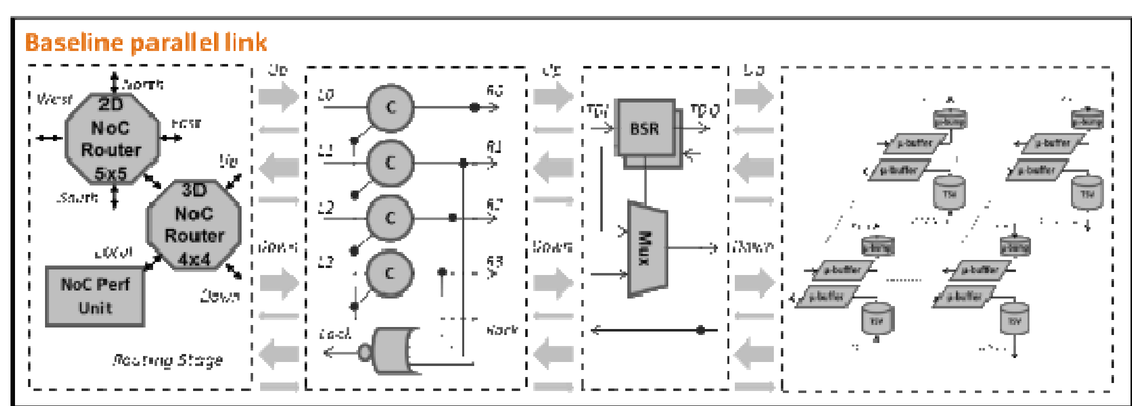


- **3DNOC router & topology**
  - No use of 7x7 ports router : too large & slow !
- **Hierarchical router**
  - –5x5 routers for intra-die com.
  - –4x4 router for inter-die com. and cores
- **Performances**
  - One-hop latency for intra-die com.
  - Two-hop latency for inter-die com.
  - Preserve throughput
  - Better area than 7x7 router

Fully implemented in asynchronous logic
Robust 3D interface, no clocking issues

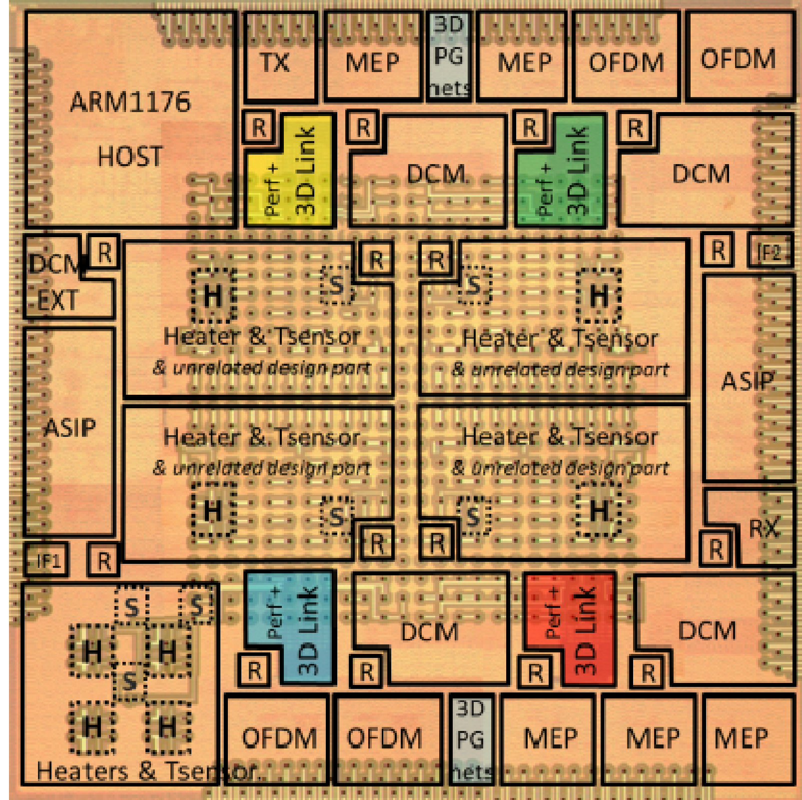
Each bi-directional up/down 3D link composed of:

- 3D Routing stage
- Pipeline stage
- DFT stage
- µbuffer & physical stage



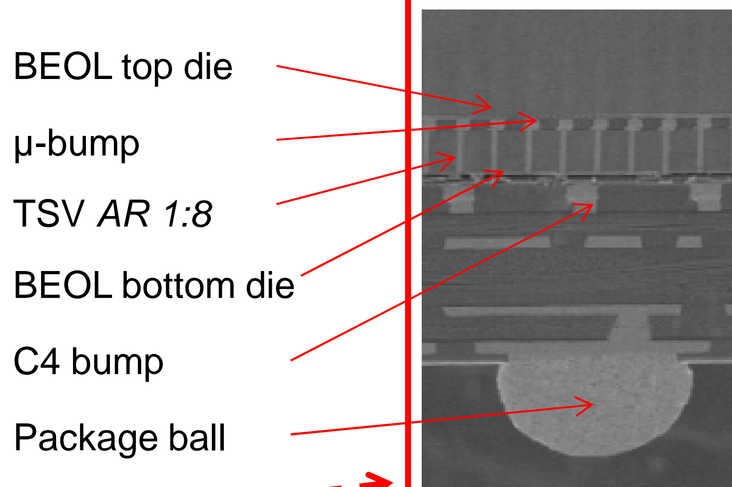
3D TECHNOLOGY & 3DNOC CIRCUIT

[P. Vivet et al. ISSCC'16]



Bottom die photo (72 mm²)

Cmos Process	65nm STMicroelectronics, Low-Power, Multi-VT
3D Tech.	TSV middle (AR 1:8) CEA-LETI µ-bumps, 50µm x 40µm pitch Face2Back stacking, Die2Die assembly
Package	12x12x1.2 flip-chip package, 4 layers substrate
Complexity	1.63 Mbyte SRAM, 228 Mtransistors, 276 IOs
Die Size	8.5mm x 8.5mm = 72.2 mm^2
Power Supply	Core = 1.2 Volts, I/O pads = 2.5 Volts



3D cross section

Molding
Top die
Bottom die
Package Substrate

3DNOC CIRCUIT DEMONSTRATION : SELF-ADAPTATION OF ASYNCHRONOUS LINK PERFORMANCES WRT. TEMPERATURE

Thermal impacts in 3D ?

- Due to 3D, increased power density, use of thin die (TSVs),
- Thermal impact on package, cost, reliability, & circuit performances

Live demo of 3DNOC circuit

- Thermal throttling using active heaters
- On-chip thermal measurements
- 3D NoC asynchronous link performance measurements with traffic generators showing self-adaptation

3D 32b parallel & fault tol.

3D 16b serial link

eSW execution of:
- thermal control
thermal measurements
- 3DNoC perf. measurements

3D 16b serial link & 2:1 red.

3D 32b parallel link

Heater

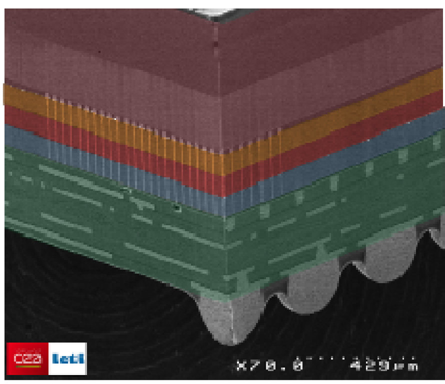
Tsensor

Network

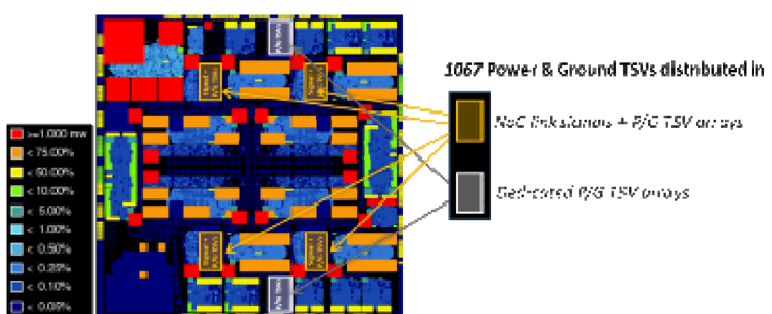
Demonstration of an Adaptive Asynchronous 3D Network-on-Chip Circuit

Live demo presented @ ISSCC'2016, DAC'16

3DNOC scalability : from 2 layers to 8 layers?



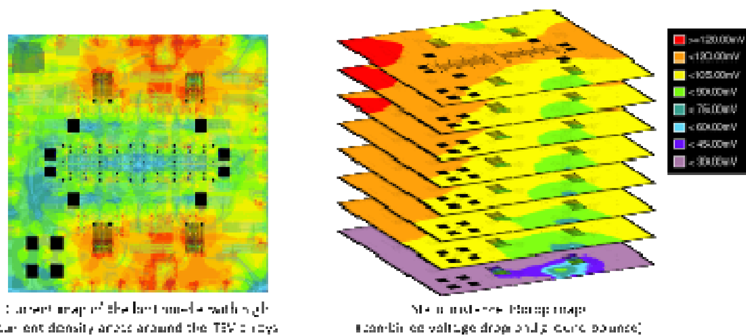
- **Is 3DNOC circuit scalable up to 8 layers ?**



Power In ?
Thermal Out ?

Power Map & Budget ~ 800 mW / layer

- **Voltage drop within the stack**



Static IR-drop analysis

Worst instance IR-drop (mV)

Total current (A) @ 1.2V

Impact of TSVs and µ-bumps

Impact of the current hooked up to the top dies on the IR-drop over the P/G network in bottom die

Single die | 2-die stack | 4-die stack | 8-die stack

Die stack configuration

Bottom-die, Top-die 1, Top-die 2, Top-die 3, Top-die 4, Top-die 5, Top-die 6, Top-die 7, Total current

APACHE/RedHawk 3D simulations

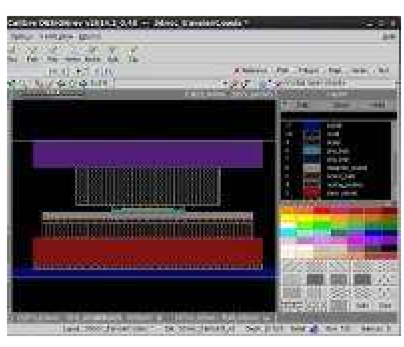
8 layers, Worst IRdrop ~ 125 mV

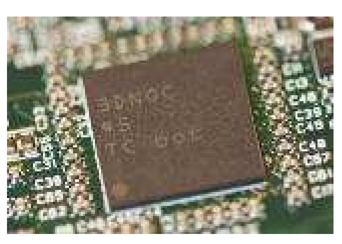
[P. Vivet, to appear in JSSC'17-01]

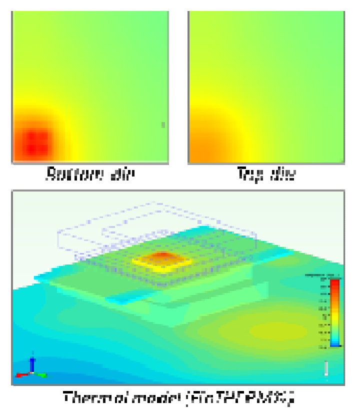
3DNOC scalability : from 2 layers to 8 layers?

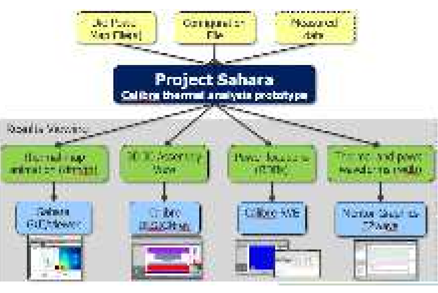
- **Thermal Model & Study**

Power Map & Budget ~ 800 mW / layer
Thermal model : 3D dies + package + socket + PCB



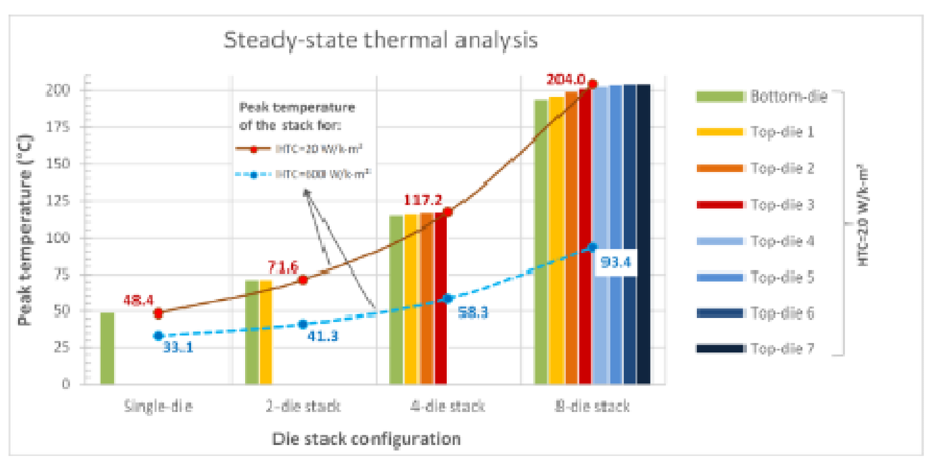






Thermal analysis
using SAHARA & FloTHERM

- **3DNOC Thermal Dissipation**



Thermal Dissipation with regular packaging
(8 layers, Pmax=6 Watts, Tmax=94°)

For limited power budget :
- Power delivery is sufficient (< 10% IRdrop)
- Max temperature < 100°C

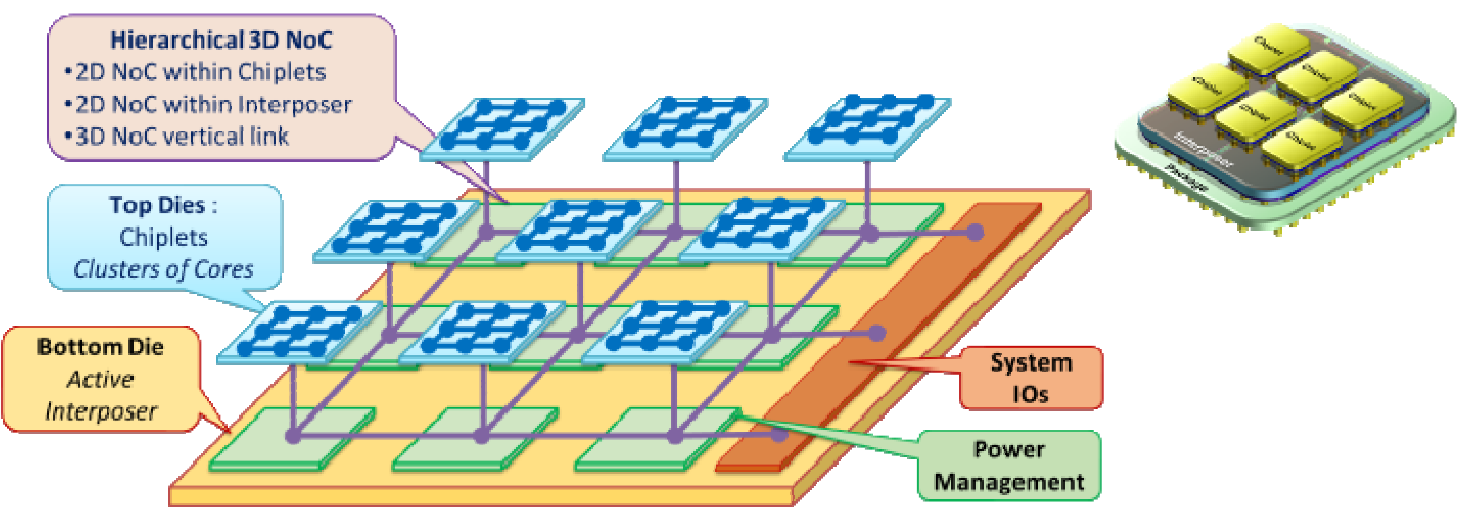
➔ multilayer 3DNOC is feasible up to **8 layers**

[P. Vivet, to appear in JSSC'17-01]

OUTLINE

- **Introduction**
- **3D Technology : an introduction**
- **State-of-Art on Circuits & Applications**
- **3D Circuit Demonstrators**
  - 3DNOC : A logic-on-logic multi-core
  - ***INTACT : An Active Interposer for computing***
  - HUBEO : Photonic Interposer
- **New Trends with High Density 3D technologies**
- **Conclusions & Perspectives**

ACTIVE INTERPOSER PARTITIONING FOR MANY-CORE



« Active » Interposer : which added value ?

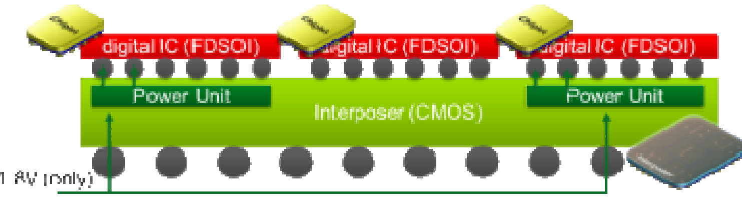
- Heterogeneous 3D - Advanced tech node for computation within chiplets
  - Mature tech node for communication/power/DFT/etc
- Chip-to-Chip Interconnect - Hierarchical NoC, for energy efficient communications
- System IOs - On Interposer, for off-chip memory accesses
- Power Management - Chiplet power supply, without any external passives

- And most of all ... preserve (active) interposer cost !

Target low logic density (eg < 10%) to preserve interposer yield & cost

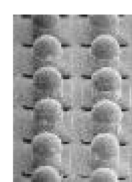
Source Vivet, ISVLSI'15

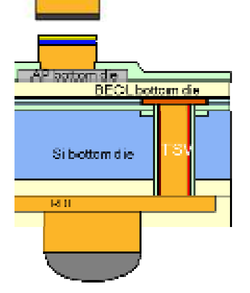
ACTIVE INTERPOSER FOR COMPUTING : 28FDSOI CHIPLETS 3D-STACKED ON A 65NM ACTIVE INTERPOSER OFFERING A 96 CORES COMPUTE FABRIC

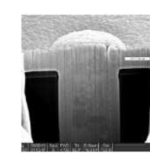


Heterogeneous 3D partitionning for high energy efficiency and reduced cost

μ-bumps
Ø 10 μm
Pitch 20 μm







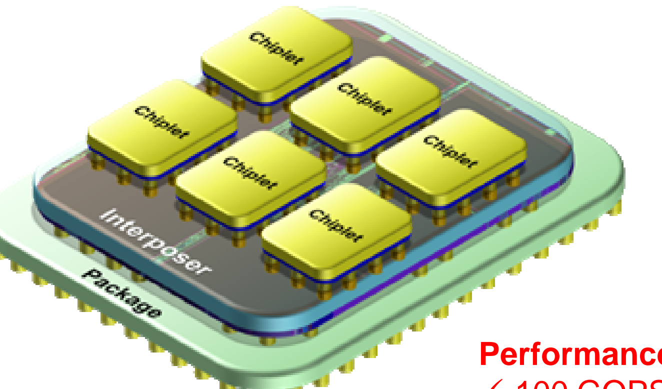
TSV
Ø 10μm
Height 100μm

28nm FDSOI chiplets (x6)

- Low Power Compute Fabric
- Wide Voltage Range (0.6V – 1.2V)
- Body Biasing for logic boost & leakage ctrl

65nm Active Interposer

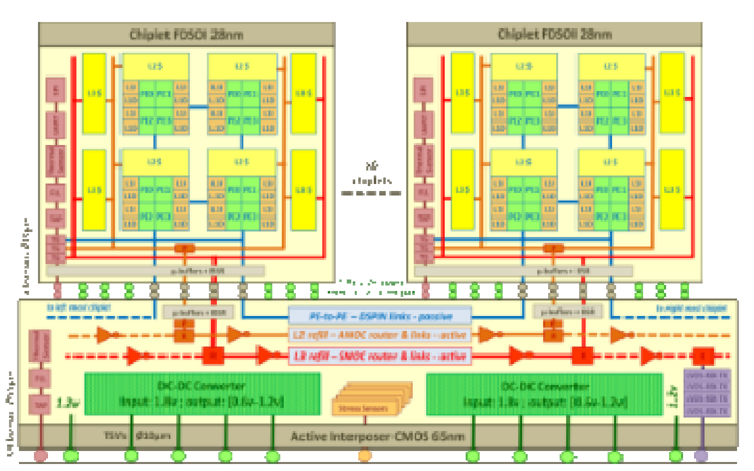
- Power unit (Switched Cap DC-DC conv.)
- Interconnect (Network-on-Chip)
- Test, clocking, thermal sensors, etc



Performance Targets

- ✓ 100 GOPS
- ✓ 10 GOPS/Watt
- ✓ 25 Watts total

NANOELEC

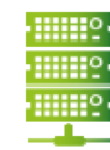


Cache Coherent Compute Fabric

- 96 cores (MIPS-32bit)
- L1/L2/L3 coherent caches
- Implemented with 3D-Plugs
- Full support of Linux OS

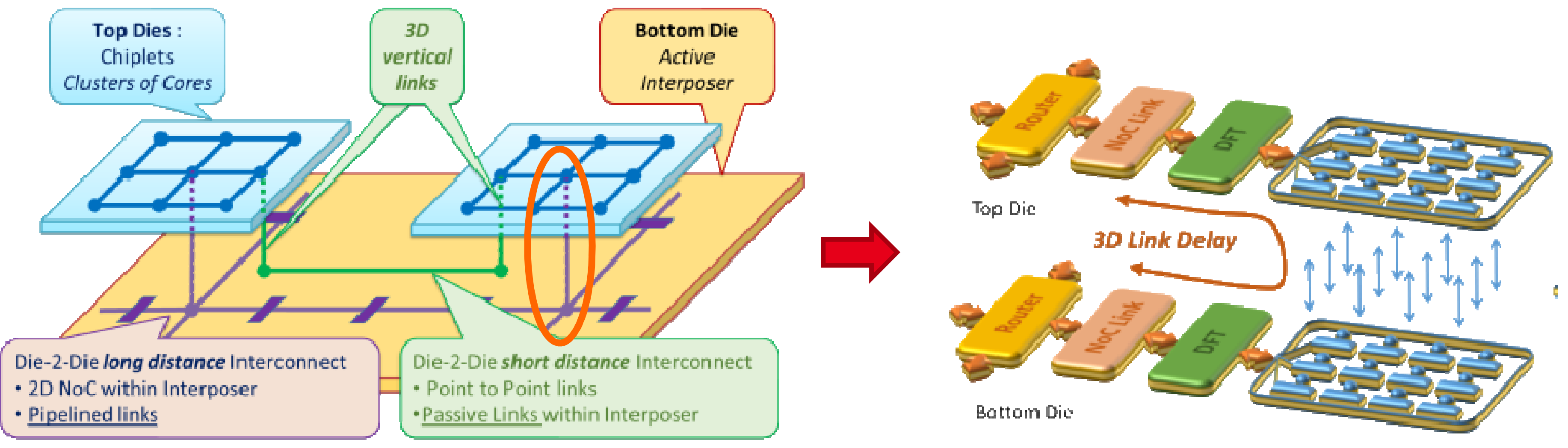
Application Targets

- ✓ Big Data
- ✓ Networking
- ✓ High Performance Computing



[D. Dutoit, VLSI-Symposium'2016]
[P. Vivet, S. Cheramy, 3DIC'2015]
[P. Vivet, E. Guthmuller, ISVLSI'2015]

3D COMMUNICATION: 3D-PLUG DESIGN CHALLENGES



- **Chip-to-Chip Active or Passive NoC links**
  High throughput, Low latency, robust interface

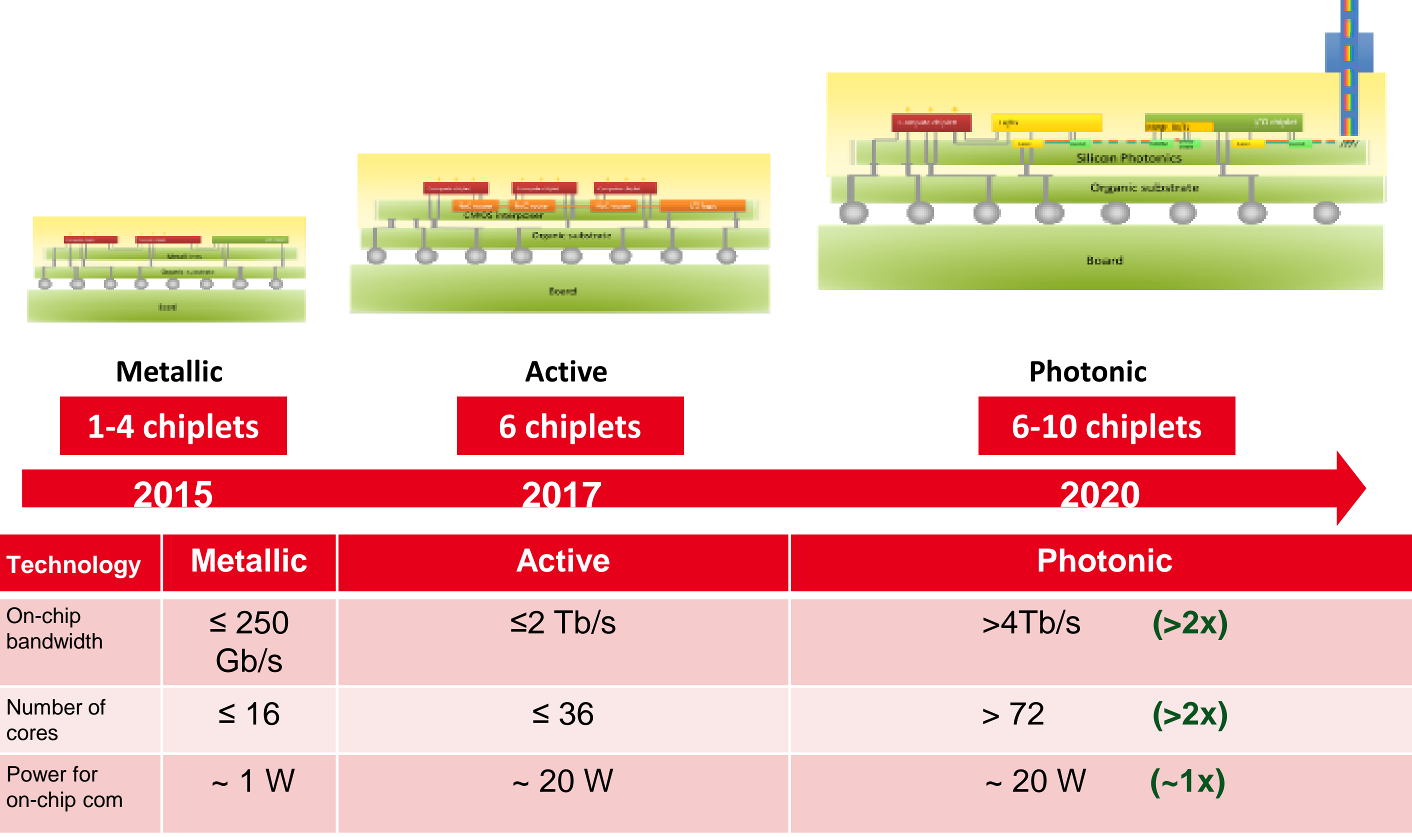
- **3D-Plug need to cope with :**
  - DFT interface : muxes for Boundary Scan cells
  - Electrical Interface : µ-buffer cell design
  - Physical interface : layout constraints of µ-bump/TSV array, PG grid, etc.
  - Logical interface : protocol signalling, timing margins, etc.

Source Vivet, ISVLSI'15

OUTLINE

- **Introduction**
- **3D Technology : an introduction**
- **State-of-Art on Circuits & Applications**
- **3D Circuit Demonstrators**
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - ***HUBEO : Photonic Interposer***
- **New Trends with High Density 3D technologies**
- **Conclusions & Perspectives**

ON-CHIP COMMUNICATION ON INTERPOSER : PASSIVE, ACTIVE OR PHOTONIC ?



Technology	Metallic	Active	Photonic
On-chip bandwidth	≤ 250 Gb/s	≤2 Tb/s	>4Tb/s **(>2x)**
Number of cores	≤ 16	≤ 36	> 72 **(>2x)**
Power for on-chip com	~ 1 W	~ 20 W	~ 20 W **(~1x)**

➔ Photonic : The Scale-up/Scale-out Technology !
For a given power envelop, it will offer larger traffic bandwidth, & integrate more cores onto a single package

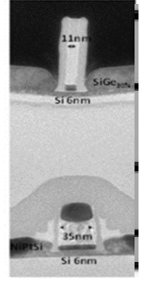
Source: Thonnart, Y., Zid, M. "Technology assessment of silicon interposers for manycore SoCs: Active, passive, or optical? " NoCS 2014

OUTLINE

- **Introduction**
- **3D Technology : an introduction**
- **State-of-Art on Circuits & Applications**
- **3D Circuit Demonstrators**
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- **New Trends with High Density 3D technologies**
- **Conclusions & Perspectives**

HIGH DENSITY 3D : A REAL ALTERNATIVE TO SCALING

3D Density of Integration



Monolithic 3D (CoolCube™) [3]
Diameter: 0.05 µm
Pitch : 0.11 µm

10^8 3D Contacts / mm²
=> Gate-/Transistor- Level Integration

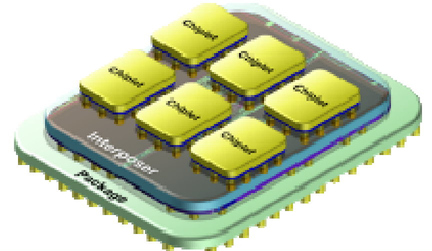
HD-TSV [2]
Diameter : 0.85 µm
Pitch : 1.75 µm

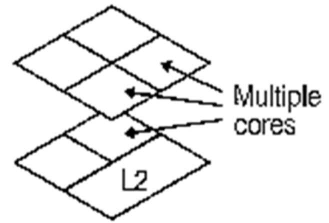
~10^5 3D Contacts / mm²
=> Core-/Block-Level Integration

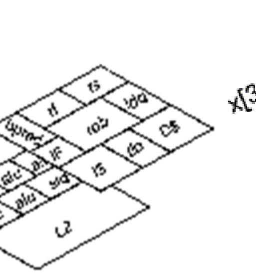
~10^3 3D Contacts / mm²
=> SoC Level Integration

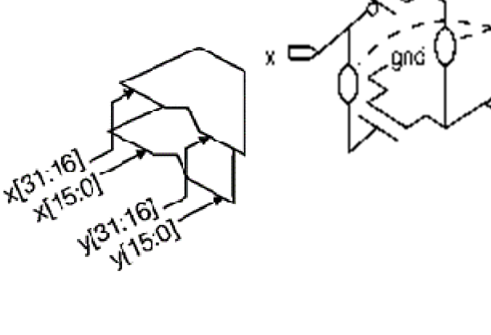
TSV + µBump[1]
Diameter: 10 µm
Pitch : 20 µm

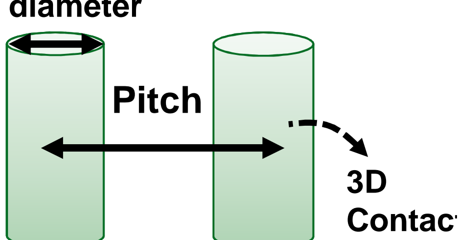
Cu-Cu [2]
Diameter : 1.7 µm
Pitch : 3.4 µm





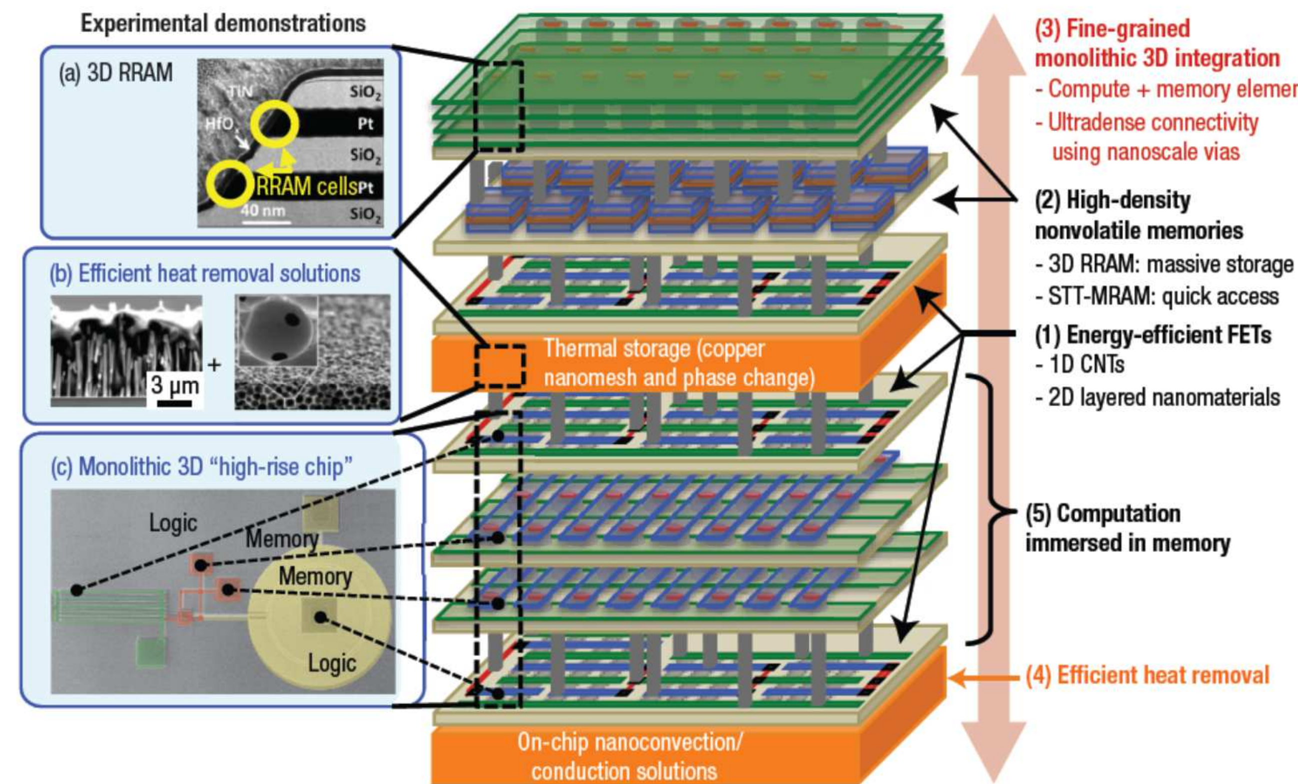






[1] Cheramy, S., et al. “Advanced Silicon Interposer”, C2MI Workshop, 2015
[2] Patti, B., “Implementing 2.5D and 3D Devices”, In AIDA workshop in Roma, 2013
[3] Batude, P., et al. "3DVLSI with CoolCube process: An alternative path to scaling ." VLSI technology symposium 2015

3D TECHNOLOGY AND NEW COMPUTING PARADIGM



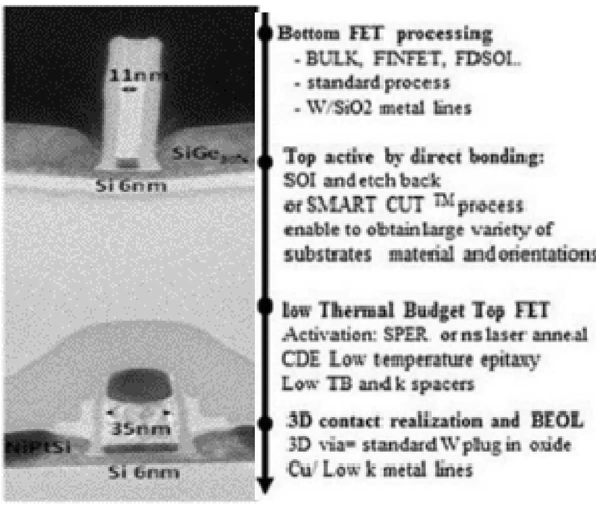
N3XT Architecture

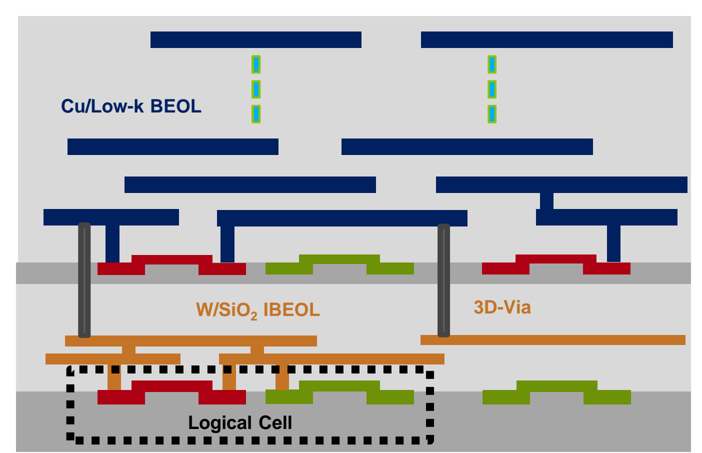
- Monolithic 3D
- 3D RRAM
- CNT FET
- Tight memory-computing integration

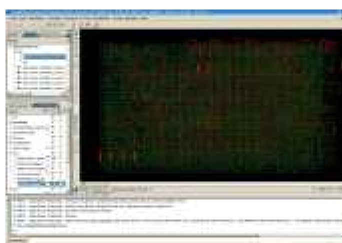
Claim a ~ x1000 gain in energy efficiency gain (from technology, architecture)

« Energy-Efficient Abundant-Data Computing: The N3XT 1,000x », M. Sabry & al, Computer, 2015, Volume: 48, Issue: 12

MONOLITHIC 3D : COOLCUBE^{TM} PROCESS & DESIGN







- **Top layer @ low thermal budget (500/550°C)** [1]
- **High alignment precision process**
- **Up to 10^8 3D Vias per mm^2 => 10^4 x than Cu-Cu or HD-TSV**

- **EDA collaboration : Architecture level (Atrenta) ; Signoff DRC/LVS (Mentor)**
- **EDA tools for 3D High Density Place and Route : *required !***

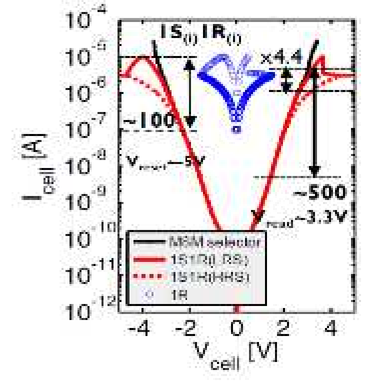
- **Up to 60% Area reduction & 25% better perf vs 2D 28 nm @ preliminary result** [2]

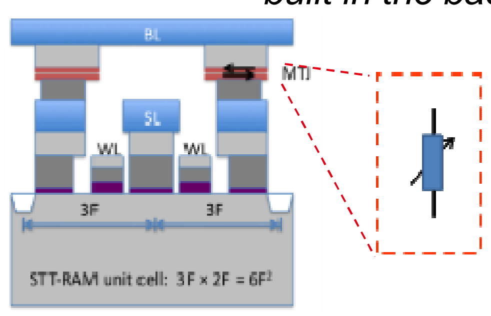
  ➔ ***Objective : 1 node gain without scaling : 28nm / 28 nm ⇔ 14 nm***

[1] P. Batude, et al., “3DVLSI with CoolCube process: An alternative path to scaling”, VLSI technology symposium 2015.
[2] H. Sarhan, et al., “An Unbalanced Area Ratio Study for High Performance Monolithic 3D Integrated Circuits”, ISVLSI 2015.

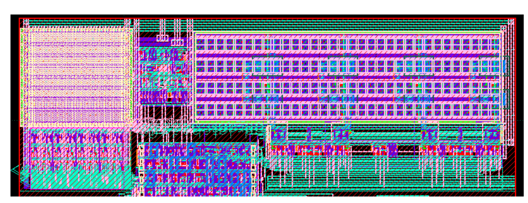
NON-VOLATILE MEMORY

Metal-Insulator-Metal structure built in the back-end-of-line

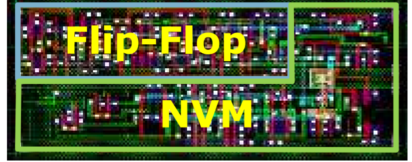




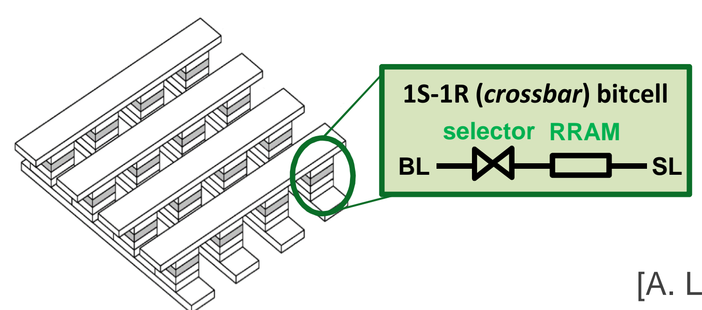
Non-Volatile SRAM

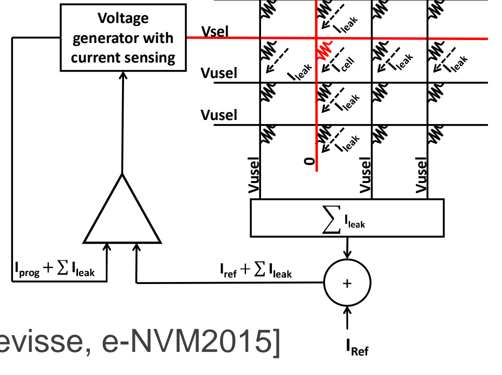


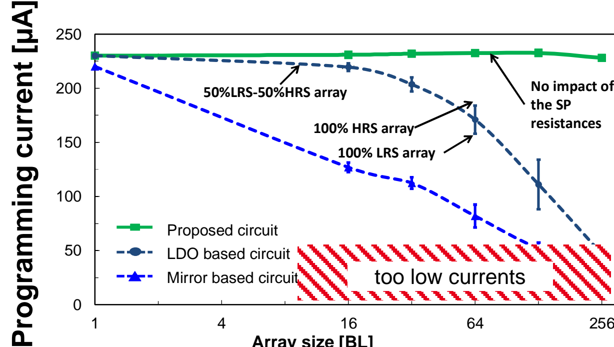
Non-Volatile FF



[TED14 J. Zhou] [TED13 Y. Deng] [IEDM14 L. Zhang]







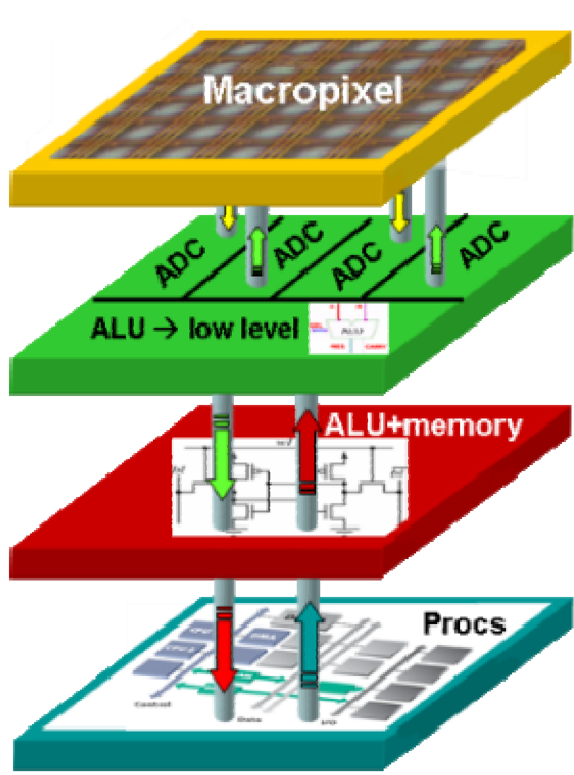
[A. Levisse, e-NVM2015]

- **RRAM ? it is a kind of 3D device – post-processed within regular technology process**
- Co-design between Circuit Architecture & Technology is mandatory
- **Circuit Design : Crossbar exploration & Sneak Path compensation**
- **System Design : non-volatile processor for IoT : fast wake-up, NV-FF, NV-SRAM, NV-REG**
- **Going Further ? Advanced research on-going : Logic-in-Memory, Neuromorphic**

Device

System

4 LAYERS SMART IMAGER



L1 : image capture

- BSI (Back Side Illumination)

L2 : read out circuit

- ADC (analog & digital)
- Analog processing

L2 : low level processing

- SIMD digital processing array
- Distributed Memory, 1st level

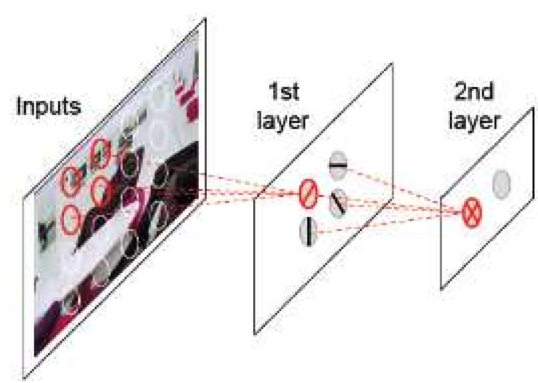
L3 : medium level processing

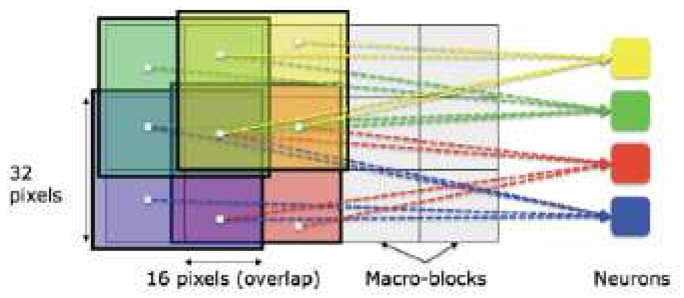
- Distributed Memory, 2nd level
- Host interface, System Communication
- Image processing

LOGIC-ON-LOGIC : 3D NEURAL NETWORK CIRCUIT

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 !

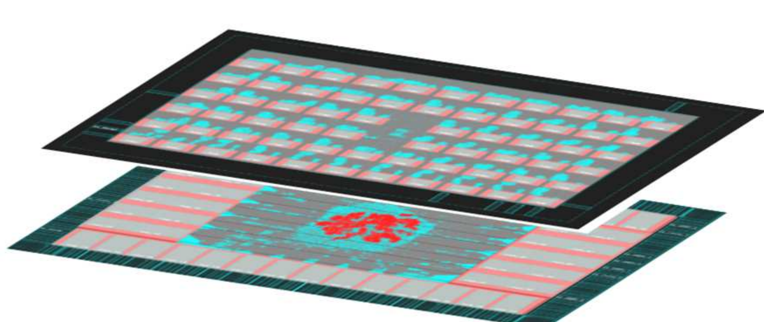




Compared to 2D,
3D offers :
2x better total area
25% better in power

Component or Block	Power (mW)	Power (%)	Area (μm^2)	Area (%)	Critical path (ns)
TOTAL	353.90	100.00	3,634,195.44	100.00	6.63
Layer 1	247.62	69.97	911,395.45	25.08	
Decoder	0.35	0.10	5,913.60	0.16	
Configuration	0.03	0.01	2,442.40	0.07	
Synapses (RAM)	208.10	58.80	431,636.64	11.88	
Neuron	39.15	11.06	471,402.80	12.97	
Layer 2	106.28	30.03	2,722,799.99	74.92	
Decoder	0.42	0.12	7,495.99	0.21	
Configuration	0.04	0.01	3,219.20	0.09	
Synapses (RAM)	90.18	25.48	2,544,723.20	70.02	
Neuron	15.64	4.42	167361.60	4.61	

Table 1. *Characteristics and breakdown of (two-layer) 3D circuit.*



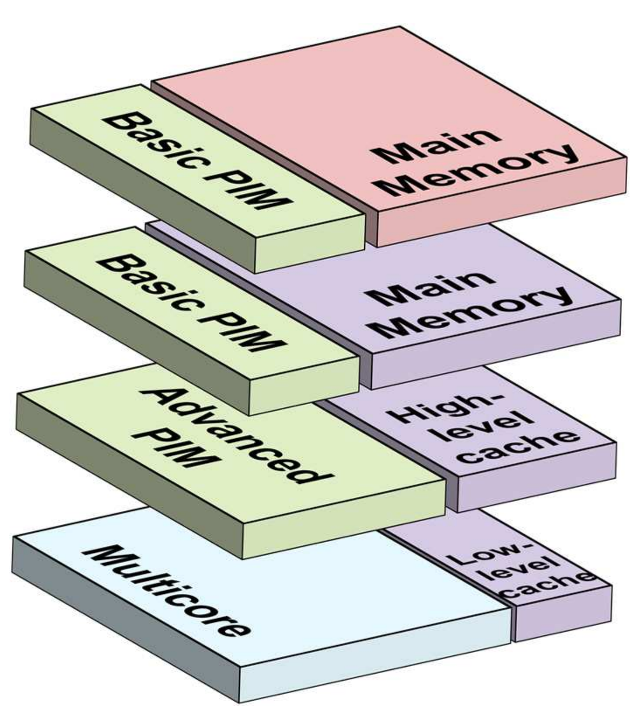
Component or Block	Power (mW)	Power (%)	Area (μm^2)	Area (%)	Critical path (ns)
TOTAL	428.24	100.00	7,974,762.94	100.00	9.00
Decoder	1.05	0.24	13,497.90	0.17	
Configuration	4.32	1.01	4,506,958.60	56.52	
Synapses (RAM)	298.28	69.65	2,976,359.84	37.32	
Neuron	124.60	29.09	477,946.59	5.99	

Table 2. *Characteristics and breakdown of (two-layer) 2D circuit.*

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

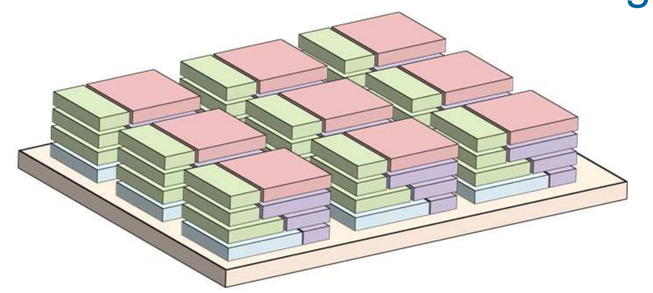
More layers ?
Tighter integration of Neuron, Memory, and NVM ?

ARCHITECTURE “DATA CENTRIC”: A 3D VISION ?



Re-visit Processing-In-Memory thanks to new technologies ?

Interposer integration for scaling



Distribute the processing within the memory hierarchy

- Memory hierarchy ? programming model ? some level of coherency ?

Heterogeneous 3D integration

- Active Interposer, Non Volatile Memory technology, advanced node for computing

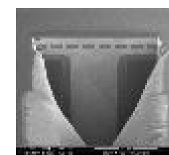
Scalability

- Vertically : more memory layers
- Horizontally : more chiplets

OUTLINE

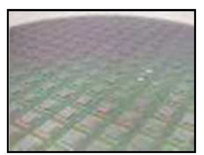
- Introduction
- 3D Technology : an introduction
- State-of-Art on Circuits & Applications
- 3D Circuit Demonstrators
  - 3DNOC : A logic-on-logic multi-core
  - INTACT : An Active Interposer for computing
  - HUBEO : Photonic Interposer
- New Trends with High Density 3D technologies
- Conclusions & Perspectives

CONCLUSIONS & PERSPECTIVES



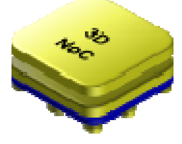
3D technology is mature and is already on the market !

- Imagers (Sony), MEMS
- Memory Cubes (Samsung, Hynix), with HMC, HBM, WideIO
- Xilinx Virtex7 (Passive Interposer)
- AMD & NVIDIA (GPU & HBM cubes on interposer)



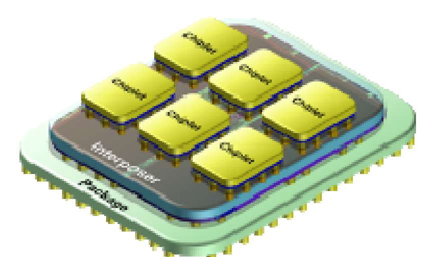
➔ *3D Technology and Value chain are ready and available*

➔ *3D CAD tools are getting mature*



Logic-on-Logic partitionning

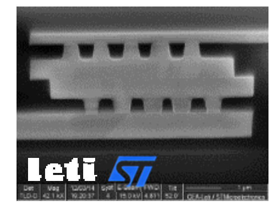
- Many number of demonstrators …
- **3DNOC : a first large scale 3D Network-on-Chip architecture & circuit**
  - Energy efficient 3D communication, 326 Mbit/s, 0.66pJ/bit
  - Demonstrated self-adapation to temperature, can scale up to 8 dies,

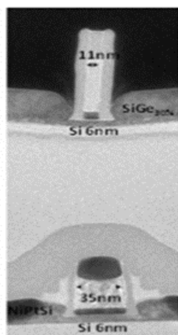


Chiplet partionning for scale-out architectures

- Cost effective, heterogeneous technologies,
- **Active Interposer, INTACT, offering 96 cores, target 100 GOPS, 25 Watts**
- **Photonic Interposer, for future large scale many-core**

CONCLUSIONS & PERSPECTIVES



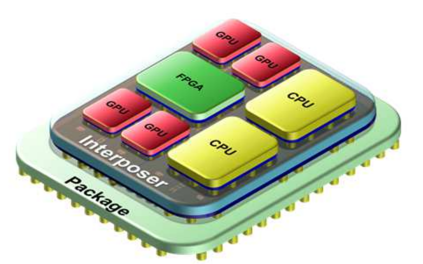


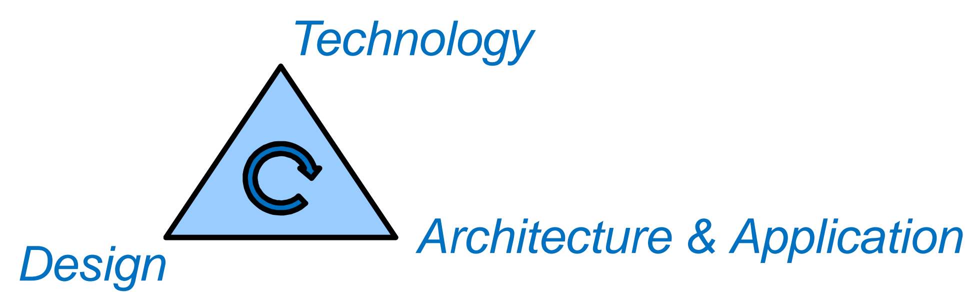
3D technology is continously evolving !

- Smaller pitch, new technologies
- *Copper-Copper Hybrid bonding*
- *Monolithic 3D (CoolCube™)*

An architecture R-evolution

- **Smaller & Denser** 3D interconnects will be available soon,
- Many design & CAD challenges
- **Need to re-think system and computer architecture**
- New opportunites for many applications
  - *Imagers, Neuro, Processing-In-Memory, Many other ones*





ACKNOWLEDGMENTS

CEA-LETI design & technology teams :

- S. Thuriès, Y. Thonnart, R. Lemaire, C. Santos, B. Giraud, D. Dutoit, F. Clermidy, J. Martin, E. Guthmuller, C. Bernard, I. Miro-Panadès, F. Darve, J. Durupt, G. Pillonnet, J. Pontès, D. Varreau,
- S. Cheramy, D. Lattard, L. Arnaud, F. Bana, A. Garnier, A. Jouve, T.Mourier

IRT-3D project

- *Part of this work was funded thanks to the French national program “Programme d’Investissements d’Avenir, IRT Nanoelec” ANR-10-AIRT-05*

Our Partners