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


MCSoC’16 Conference, Lyon, France.
CONVENTIONAL 2D SOC

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

Computing component

Performance bottleneck

Cost of new design

Manufacturing Cost
How to fit more?

... More cores
... More Memory
... Memory closer to core
... Computing Model
... Power Efficiency
... Thermal Dissipation & cost!
Going Forward – What Options Do We Have?

2D SOC
“All-in-One chip system integration”

More Moore

More than Moore

3D Packaging

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

[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
Through Si via (TSV)
Chip-to-chip interconnection (with underfill)
Chip-to-package interconnection

Redistribution layer (RDL)

**TSV: Via first, via middle and via last?**

<table>
<thead>
<tr>
<th></th>
<th>FEOL + Middle End</th>
<th>BEOL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CMOS process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Backside process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dicing Stacking</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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-WAFTER 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.)

Pitch
160 µm
40 µm
20 µm

5µm
10 µm pitch
2µm
<5 µm pitch

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

A. Garnier et al., ECTC 2014
TSV High Aspect Ratio, Metallization Challenges

Silicon thickness? ➔ key contributor for thermal & stress management
Need more aggressive 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

Application Examples

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

P. Vivet, 3DIC'2015, Sendai, Japon
3D TECHNOLOGY : DESIGN CHALLENGES?

- Co-design architecture/die/package
- Architecture partitioning
- TSV & Interconnect yield & testability
- Test Access DFT Reliability
- CAD tool maturity
- Link Performances & Energy efficiency
- Thermal Dissipation
- Power-In Thermal-Out
- Scalability and Modularity
- Heterogeneous technologies Cost & Yield

Mbps & pJ/bit
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
2.5/3D Commercial Announcements!

- DDR4 3D Dual Inline Memory Modules (RDIMMs)
- Altera 10 Generation FPGA using HMC
- Arria 10
- Stratix 10

More and more products using TSV

European 3D Summit – Grenoble, France

3D Packaging @ the Forefront of Semiconductor Industry
Most industrial players have adopted 3D Stacked BSI

- Optimized pixel array
- Optimized logic
- Smaller size
- Added functions
- With better performance

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

Sony IMX260 in Samsung Galaxy S7:
**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

Wide IO DRAM

Wide IO DRAM

Wide IO DRAM

Wide IO DRAM

Processor

Package

TSV: Through Silicon Via

(WideIO2 exemple)
# 3D DRAM: COMPARISON

<table>
<thead>
<tr>
<th></th>
<th>LPDDR4</th>
<th>WideIO/2</th>
<th>HBM</th>
<th>HMC</th>
<th>DiRAM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface type</td>
<td>parallel</td>
<td>wide data</td>
<td>wide data</td>
<td>serial</td>
<td>wide data or serial</td>
</tr>
<tr>
<td>Data bus</td>
<td>16b DDR</td>
<td>64b DDR</td>
<td>128b DDR</td>
<td>16 lanes</td>
<td>64b</td>
</tr>
<tr>
<td>Channel</td>
<td>2</td>
<td>4-8</td>
<td>8</td>
<td>4-8</td>
<td></td>
</tr>
<tr>
<td>I/O bandwidth</td>
<td>3.2Gbps @1600MHz</td>
<td>0.8Gbps @400MHz</td>
<td>1-2Gbps @500-1000MHz</td>
<td>10-15Gbps</td>
<td></td>
</tr>
<tr>
<td>Total bandwidth</td>
<td>12.8GBps</td>
<td>25.6-51.2GBps</td>
<td>128-256GBps</td>
<td>160-320GBps</td>
<td>2TBps</td>
</tr>
<tr>
<td>Capacity</td>
<td>16GB</td>
<td>16GB</td>
<td>32GB Currently 1GB (Gen1)</td>
<td>32GB Current 2-4GB</td>
<td>8GB</td>
</tr>
<tr>
<td>Total I/O</td>
<td>66</td>
<td>776</td>
<td>1616</td>
<td>256-512</td>
<td></td>
</tr>
<tr>
<td>Integration / Packaging</td>
<td>POP, MCP</td>
<td>3D</td>
<td>2.5D</td>
<td>MCP</td>
<td></td>
</tr>
<tr>
<td>Computing-In-Memory</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td></td>
</tr>
</tbody>
</table>
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

DRAM Stack
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

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

Integration is performed thanks to the EMIB (Embedded Multidie Interconnect Bridge)
Face-to-Back with TSV middle

- Face-to-Back stack configuration
- TSV Middle in MPSoc: for memory supplies and signals.

Chip-to-Chip Cu Pillars: 020 μm, Pitch 40 μm

TSV (#1016):
- Height 80 μm,
- Ø10 μm, AR 8,
- Pitch 40 μm

MPSoc-to-Substrate Cu Pillars (#932):
- Ø55 μm, Pitch 150 μm

Package balls (#459):
- Ø250 μm, Pitch 0.4 mm

Comparison with LPDDR3

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

<table>
<thead>
<tr>
<th>Memory Type</th>
<th>LPDDR3-1</th>
<th>WideIO - This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Package</td>
<td>PoP / Discrete</td>
<td>3D-IC</td>
</tr>
<tr>
<td>BW (Gbyte/s)</td>
<td>6.4 GB/s</td>
<td>12.8 GB/s</td>
</tr>
<tr>
<td>Total power</td>
<td>293 mW*</td>
<td>72.4 mW*</td>
</tr>
<tr>
<td>VDD-MPSoc</td>
<td>121 mW*</td>
<td></td>
</tr>
<tr>
<td>VDD-Mem</td>
<td>81 mW*</td>
<td></td>
</tr>
<tr>
<td>VDD-I/O</td>
<td>91 mW</td>
<td></td>
</tr>
<tr>
<td>I/O power</td>
<td>4.6 pJ/bit**</td>
<td></td>
</tr>
</tbody>
</table>

*4x gain

Source: Dutoit, 2013 Symposia on VLSI Technology and Circuits
SRAM-ON-LOGIC: 3DMAPS MULTI CORE

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

- 3D MAssively Parallel processor with Stacked memory
- 130nm GLOBALFOUNDRIES + Tezzaron F2F bonding
- 64 cores, 5-stage/2-way VLIW architecture
- 256KB SRAM, 1-cycle access
- 5mm X 5mm, 230 IO cells
- 277MHz Fmax, 1.5V Vdd
- 64GB/s memory BW @ 4W

- 2 logic tiers, face-to-face bonded
  - Top die thinned to 15um, bottom die is 150um
  - GLOBALFOUNDRIES 130nm technology + Artisan library

[ISSCC’2012, Georgia Tech]
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 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
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

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

<table>
<thead>
<tr>
<th>Architecture</th>
<th>GeorgiaTech ISSCC’2012</th>
<th>Kobe Univ. ISSCC’2013</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process &amp; 3D technology</td>
<td>130nm F2F CuCu</td>
<td>90nm F2B TSV</td>
<td>65nm F2B TSV</td>
</tr>
<tr>
<td>3D Bandwidth</td>
<td>277 Mbps</td>
<td>200 Mbps</td>
<td>326 Mbps</td>
</tr>
<tr>
<td>3D I/O Power</td>
<td>-</td>
<td>0.56 pJ/bit</td>
<td>0.32 pJ/bit</td>
</tr>
</tbody>
</table>

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

- **3D NOC 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

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

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

Fully implemented in asynchronous logic
Robust 3D interface, no clocking issues
### 3D TECHNOLOGY & 3DNOC CIRCUIT

![3D cross section](image)

**Bottom die photo (72 mm²)**

![Bottom die photo](image)

### 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 I/Os

### Die Size
- 8.5mm x 8.5mm = 72.2 mm²

### Power Supply
- Core = 1.2 Volts, I/O pads = 2.5 Volts

---

[Image: Bottom die photo (72 mm²)]

[Image: 3D cross section]

---

**BEOL top die**

**μ-bump**

**TSV AR 1:8**

**BEOL bottom die**

**C4 bump**

**Package ball**
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

Live demo presented @ ISSCC’2016, DAC’16
3DNOC scalability: from 2 layers to 8 layers?

- Is 3DNOC circuit scalable up to 8 layers?

Power Map & Budget ~ 800 mW / layer

- Voltage drop within the stack

APACHE/RedHawk 3D simulations

8 layers, Worst IR drop ~ 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

- **3DNOC Thermal Dissipation**
  
  **Steady-state thermal analysis**
  
  Thermal Dissipation with regular packaging
  (8 layers, Pmax=6 Watts, Tmax=94°C)

  - 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 : 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 partitioning for high energy efficiency and reduced cost

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

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]
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.
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?

<table>
<thead>
<tr>
<th>Technology</th>
<th>Metallic</th>
<th>Active</th>
<th>Photonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip bandwidth</td>
<td>≤ 250 Gb/s</td>
<td>≤2 Tb/s</td>
<td>&gt;4 Tb/s (2x)</td>
</tr>
<tr>
<td>Number of cores</td>
<td>≤ 16</td>
<td>≤ 36</td>
<td>&gt; 72 (2x)</td>
</tr>
<tr>
<td>Power for on-chip com</td>
<td>~ 1 W</td>
<td>~ 20 W</td>
<td>~ 20 W (~1x)</td>
</tr>
</tbody>
</table>

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

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

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)

MONOLITHIC 3D : COOLCUBE™ PROCESS & DESIGN

- Top layer @ low thermal budget (500/550°C) [1]
- High alignment precision process
- Up to $10^8$ 3D Vias per mm² => 10⁴ 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

**NON-VOLATILE MEMORY**

- **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

---

**Non-Volatile SRAM**

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

---

**Non-Volatile FF**

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

---

**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

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

<table>
<thead>
<tr>
<th>Component or Block</th>
<th>Power (mW)</th>
<th>Power (%)</th>
<th>Area (\mu m^2)</th>
<th>Area (%)</th>
<th>Critical path (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>353.90</td>
<td>100.00</td>
<td>3,634,195.44</td>
<td>100.00</td>
<td>6.63</td>
</tr>
<tr>
<td>Layer 1</td>
<td>247.62</td>
<td>69.97</td>
<td>913,385.45</td>
<td>25.08</td>
<td></td>
</tr>
<tr>
<td>Decoder</td>
<td>0.35</td>
<td>0.10</td>
<td>33,913.60</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Configuration</td>
<td>0.03</td>
<td>0.01</td>
<td>4,424.40</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Synapses (RAM)</td>
<td>208.10</td>
<td>58.80</td>
<td>63,659.64</td>
<td>11.88</td>
<td></td>
</tr>
<tr>
<td>Neuron</td>
<td>39.15</td>
<td>11.36</td>
<td>471,402.80</td>
<td>12.97</td>
<td></td>
</tr>
<tr>
<td>Layer 2</td>
<td>166.28</td>
<td>47.03</td>
<td>2,722,789.99</td>
<td>70.92</td>
<td></td>
</tr>
<tr>
<td>Decoder</td>
<td>0.42</td>
<td>0.12</td>
<td>7,495.99</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Configuration</td>
<td>0.04</td>
<td>0.01</td>
<td>3,220.20</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Synapses (RAM)</td>
<td>90.18</td>
<td>25.48</td>
<td>2,544,733.30</td>
<td>70.02</td>
<td></td>
</tr>
<tr>
<td>Neuron</td>
<td>15.64</td>
<td>4.42</td>
<td>167,361.80</td>
<td>4.61</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Component or Block</th>
<th>Power (mW)</th>
<th>Power (%)</th>
<th>Area (\mu m^2)</th>
<th>Area (%)</th>
<th>Critical path (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>220.24</td>
<td>100.00</td>
<td>7,974,742.84</td>
<td>100.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Decoder</td>
<td>1.03</td>
<td>0.24</td>
<td>2,1497.90</td>
<td>11.17</td>
<td></td>
</tr>
<tr>
<td>Configuration</td>
<td>4.32</td>
<td>0.19</td>
<td>4,506,958.60</td>
<td>55.52</td>
<td></td>
</tr>
<tr>
<td>Synapses (RAM)</td>
<td>298.28</td>
<td>69.85</td>
<td>2,976,359.14</td>
<td>37.32</td>
<td></td>
</tr>
</tbody>
</table>

More layers?
Tighter integration of Neuron, Memory, and NVM?
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, 0.66 pJ/bit
  - Demonstrated self-adaptation to temperature, can scale up to 8 dies,

Chiplet partitionning 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 continuously 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 opportunities for many applications
  - *Imagers, Neuro, Processing-In-Memory, Many other ones*
ACKNOWLEDGMENTS

CEA-LETI design & technology teams:
• 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