Monolithic 3D Integration
using Standard Fab & Standard Transistors

Zvi Or-Bach
CEO
MonolithIC 3D Inc.
3D Integration
Through Silicon Via ("TSV"), Monolithic

- Increase integration
  - Reduce interconnect total length
  - Not subject to improved litho. and process
- Heterogeneous integration
  - Silicon, Optoelectronic, Sensors, ...
  - I/O, Logic, Memory, ...
- New functionality utilizing the third dimension
  - Redundancy (WSI)
  - Shared Litho. And processing (BiCS)
Monolithic 3D
10,000x the Vertical Connectivity of TSV

<table>
<thead>
<tr>
<th>TSV</th>
<th>Monolithic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer Thickness</td>
<td>~50μ</td>
</tr>
<tr>
<td>Via Diameter</td>
<td>~5μ</td>
</tr>
<tr>
<td>Via Pitch</td>
<td>~10μ</td>
</tr>
</tbody>
</table>

High density of vias

- 3D TSV
- D=10,000/mm²
- D=100,000/mm²
- D=5,000,000/mm²
- Sequential 3D
The Monolithic 3D Challenge

Why is it not already in wide use?

- **Thermal Budget of Upper Layer Processing**
  - < 550°C over advanced node transistors
  - < 400°C over Copper or Aluminum interconnect
    - How to bring mono-crystallized silicon on
    - How to fabricate state-of-the-art transistors on top

- **Misalignment of pre-processed wafer to wafer bonding step**
  used to be ~1µm
  - How to achieve 100nm or better connection pitch
  - How to fabricate inter-layer vias of ~50nm
Outline: Layer Formation

- **Deposit**
  - Poly silicon + crystallize by laser (Taiwan NAR Lab)
  - Non silicon

- **Layer Transfer**
  - Bond SOI wafer and etch/grind back
    - IBM, MIT Lincoln Lab
  - Ion-Cut (Smart Cut®)
  - ‘Smart’ Substrate
    - ELTRAN (Epitaxial Layer Transfer)
    - *New Breakthrough*
Outline: Overcoming the Process Heat

- Non silicon transistor
- Modified process to less than $550^\circ C$ (CoolCube)
- Break transistor formation into two phases, Hot-Cold
- Isolate Top layer from base—“Crème Brulee”
- ‘nano-TSV’ – Precise ultra thin stratum transfer
Modified the process for low temp

Top FET critical thermal budget

- Gate oxide stabilization
  - 800°C
- Dopant activation
  - 1000°C
  - Si @ 750°C
  - SiGe @ 650°C
- Epitaxy
  - 630°C
- Spacer deposition

Bottom strata stability (Hours duration)

Rmk: Bottom strata stability window is wider for shorter anneal durations (up to 800°C, ms)
Precision Bonder – Multi-Stratum M3D

- Utilizing the existing front-end process !!!
  - <200 nm (3σ)
  - Achieving 10,000x vertical connectivity as the transferred strata will be thinner than 100 nm
    - Mix – Sequential/Parallel M3D
    - Low manufacturing costs
Fusion Bonding – Overlay Evolution

2014

Translation y-direction (nm)

Translation x-direction (nm)

2015

Translation y-direction (nm)

Translation x-direction (nm)

200 nm Overlay

2016

Translation y-direction (nm)

Translation x-direction (nm)

[nm]

2016

Translation x-direction (nm)
2nd Stratum Processed on ‘Cuttable’ wafer

Substrate

- Silicon layer
- Etch selective - ‘cut’ layer
- Base wafer – (reused)

Transferable stratum

- ~700 µm Donor Wafer
- NMOS
- PMOS
- ‘Cut Layer’
Flip and Bond 2\textsuperscript{nd} Stratum on Top of 1\textsuperscript{st} Stratum

Oxide to oxide or hybrid bonding

‘Cut Layer’
‘Cut’ Carrier Wafer and Connect with Nano-TSV
The Breakthrough –
SiGe – Fab Compatible ‘Cut Layer’

- Use SiGe as etch stop

- Use selective etch of SiGe (Dry, ~1:100) for reuse

Samsung’s

Applied Material’s Selectra
Monolithic 3D using ‘Cuttable’ Substrate

UTILIZES EXISTING TRANSISTOR PROCESS
(first ever)

LEVERAGE PROVEN SILICON PROCESSES

VERY COMPETITIVE COST STRUCTURE
The Monolithic 3D Advantages

- Enables increased integration without size reduction
- Reduce average wire length by ~0.7x

Heterogeneous Integration
- Type of wafer, Si, GaAs, GaN,..
- Mix fabs and nodes, 7nm, 28nm, 130nm, ...
- Mix process, logic, memory, analog, RF, ...
- Multiple layers Processed Simultaneously - ‘BiCS’

New functionality leveraging rich 3D connectivity
- Redundancy
- Allow connectivity from both sides

The 1,000x
EDA for Monolithic 3D

- Commercial EDA is not available – yet.
- Preliminary tools and flow has been developed
  - Georgia Tec
  - Qualcomm
  - CEA Leti…
- 2D are sufficient for the 1,000x applications !!!
Summary

- We have reached an inflection point
- Monolithic 3D would provide unparalleled advantage for most market segment
- Multiple practical paths to monolithic 3D exist
  - Monolithic 3D could be now implemented by any fab without changes to front-line
  - In 2017 30% of all transistors made will be M3D (3D NAND)

=> Use correctly MonolithIC 3D would provide 1,000x better devices
Monolithic 3D Integration
for
1,000x Better Computer
Processor Memory Gap

Gap grew 50% per year

Source: Computer Architecture, A Quantitative Approach by John L. Hennessy and David A. Patterson
The Problem with Wires:

Energy to move data proportional to distance

- Cost to move a bit on copper wire:
  - power = bitrate * Length / cross-section area

- Wire data capacity constant as feature size shrinks
- Cost to move bit proportional to distance
System Level Interconnect Gap Grow

Fig. 7 System level interconnect gaps

Source: VLSI 2013
Dr. Jack Sun, CTO of TSMC
Growing interconnect challenge

Source: ITRS
Devices enabling high-density, fine-grained, monolithic 3D systems for reducing data movement and communication cost:

- High-performance devices for logic and memory suitable for low-temperature processing used for stacked layer integration; the resulting energy-delay product should be 1000X better than state-of-the-art conventional technology. Candidate materials for next-generation logic.

**Experimental demonstrations**

1. **3D RRAM**
   - TIN
   - Pt
   - SiO₂
   - HfO₂
   - RRAM cells

2. **Efficient heat removal solutions**
   - 3 μm

3. **Monolithic 3D “high-rise chip”**
   - Logic
   - Memory

4. **Efficient heat removal**

5. **Computation immersed in memory**

(3) **Fine-grained monolithic 3D integration**
   - Compute + memory elements
   - Ultradense connectivity using nanoscale vias

(2) **High-density nonvolatile memories**
   - 3D RRAM: massive storage
   - STT-MRAM: quick access

(1) **Energy-efficient FETs**
   - 1D CNTs
   - 2D layered nanomaterials
Monolithic 3D for 1,000x better compute

- Processor to memory connectivity breakthrough:
  - Number of wires: \(\sim 100\) \(\rightarrow\) \(\sim 100,000\)
  - Length of wires: \(\sim 20\text{mm}\) \(\rightarrow\) \(\sim 20\mu\text{m}\)

- Applications:
  - Smart-Phones
  - IoT
  - Computers
Replacing few mm long bus of <100 wires with few microns long bus of 2,500 x 100 wires

Array of 200µ x 200µ units each with its peripherals on top
1,000x Better Compute System Peripherals over Memory Array

- Solving the ‘Memory Wall’
  - Close proximity of memory to the logic control
  - Allows a different process for the memory matrix than the peripherals

[Diagram of the process involving Donor Wafer, Memory Peripherals, Memory Array, and the "Cut" Layer, with Flip & Bond steps shown.]
Summary

1,000x better compute Systems using M3D

- Multi thin layers each with thousands memory units
  => ~100,000 wires @ ~20μ length
  => 1,000x better computer system
- Utilizes existing memory (DRAM) and logic process
- Far cheaper than any other form of DRAM module
- Yield repaired in fabrication and in operation (field)
- Memory stack could serve as a generic end product
- Low-power (1/10), low size, low weight
New Memory Product – Generic 3D Wafer

- Thermal isolation + pads
- Unit Decoder
- Memory Stack
- Controller +

Generic Memory Wafer
New Memory Product – Generic 3D Wafer

- Thermal isolation + pads
- Unit Decoder
- Memory Stack
- Controller +

Generic Memory Wafer

I/O

Custom Processors

Generic Memory Wafer

3D Integrated Computer
Other Approaches

- HBM, HMC, Tezzaron – DiRAM4
- 2-8x more expensive
- Only solve the memory bank integration
- Only for DRAM

<table>
<thead>
<tr>
<th>Interface type</th>
<th>LPDDR4</th>
<th>WideI/O/2</th>
<th>HBM</th>
<th>HMC</th>
<th>DiRAM4</th>
</tr>
</thead>
<tbody>
<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 Currently 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>
# Alternatives for Processor Memory Connectivity

<table>
<thead>
<tr>
<th></th>
<th>Processor</th>
<th>PCB</th>
<th>Module</th>
<th>DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>5mm</td>
<td>10-30mm</td>
<td>20mm</td>
<td>5mm</td>
</tr>
<tr>
<td>2.5D-TSV</td>
<td>5mm</td>
<td>5-10mm</td>
<td>0</td>
<td>1mm</td>
</tr>
<tr>
<td>M3D</td>
<td>0.5mm</td>
<td>0</td>
<td>0</td>
<td>0.2mm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>#of Wires</th>
<th>M Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2.5D</td>
<td>1,000</td>
<td>2-8</td>
</tr>
<tr>
<td>M3D</td>
<td>100,000</td>
<td>0.5</td>
</tr>
</tbody>
</table>