A DISRUPTOR TO THE SEMICONDUCTOR INDUSTRY

THE MONOLITHIC 3D-IC
Agenda:

- The Semiconductor future is exciting
- But we are reaching an inflection point
- Monolithic 3D IC – the emerging path for the next generation technology driver
- The challenge and solution for the fabrication of monolithic 3D IC
$15-34 trillion, annual =>~$5T Semi /year

Source: McKinsey Global Institute Analysis 2013
Cisco sees $19 Trillion opportunity in IoT

“CES LIVE: Cisco's Chambers Says Internet of Everything, $19 Trillion Opportunity, Is Next Big Thing” 1/7/14

$19 trillion: that’s the opportunity he says for the Internet of Everything in the private and public sector combined. Breakout is $14.4 trillion in private sector and $4.6 trillion in public sector of new revenue generation or new savings. That’s a conservative number he says for public sector.

“This will be bigger than anything done in high tech in a decade.”

“As many as 50 billion devices will be connected to the Internet by 2020, creating a $14.4 trillion business opportunity” said Rob Lloyd, president of sales and development at Cisco, <http://www.eetimes.com/electronics-news/4409928/Cisco-sees--14-trillion-opportunity-in-Internet-of-Things>
Semiconductor Industry is Facing an Inflection Point

Dimensional Scaling has reached Diminishing Returns
The Current 2D-IC is Facing Escalating Challenges - I

- On-chip interconnect is
  - Dominating device power consumption
  - Dominating device performance
  - Penalizing device size and cost

![Graph showing interconnect delay creates the timing closure problem](image-url)
Interconnect Delay
A Big Issue with Scaling

- Transistors improve with scaling, interconnects do not
- Even with repeaters, 1mm wire delay ~50x gate delay at 22nm node

Source: ITRS
Connectivity Consumes 70-80% of Total Power @ 22nm
Repeaters Consume Exponentially More Power and Area

- At 22nm, on-chip connectivity consumes 70-80% of total power
- Repeater count increases exponentially
- At 45nm, repeaters are > 50% of total leakage

Source: IBM POWER processors
R. Puri, et al., SRC Interconnect Forum, 2006
The Current 2D-IC is Facing Escalating Challenges - II

- Lithography is
  - Dominating Fab cost
  - Dominating device cost and diminishing scaling’s benefits
  - Dominating device yield
  - Dominating IC development costs

Subwavelength Lithography Challenge

![Graph showing subwavelength lithography challenge with timelines from 1980 to 2010 and silicon feature sizes from 436nm to 157nm.](synopsys.png)
A Challenge: Lithography

- Quad-patterning next year → costly. EUV delayed, costly.
- Can we get benefits of scaling **without relying on lithography**?
Cost becomes a concern post 28 nm

Sources: nVidia, ITPC, nov, 2011
Broadcom, IMEC, may 2012
GF, ISS, jan 2013
Embedded SRAM isn’t Scaling Beyond 28nm (1.1x instead of 4x)

*eSRAM > 60% of Die Area => End of Dimension Scaling !

---

**SRAM DENSITY - 16nm vs 28nm**

Memory density at 1500MHz and above scales by ~1.1x or less from 28nm to 16nm

---

For 400G ASSP/ASIC, need to double the SRAM density

But density improvement from process node N to N+1 is not 2x anymore but by 1.1x

500mm² die in 28nm for 200G with 60% SRAM ported to 16nm for 400G will be ~ 745mm²

Die size close to reticle limit - exacerbates yield & cost of lower end segments

At > 400G, embedding all the SRAM would make the die size bigger than reticle limit

---

Dinesh Maheshwari, CTO, Memory Products Division at Cypress Semiconductors, ISSCC2014
Embedded SRAM isn’t Scaling Beyond 28nm

eSRAM > 60% of Die Area => End of Dimensional Scaling!

* imec’s 2013 International Technology Forum,
Moore's Law Dead by 2022*
Bob Colwell, Director MTO, DARPA

My model: During and After Moore’s Law

1. COTS is both problem & opportunity for DoD for next 10 years.
2. Then COTS stalls out. (But DoD doesn’t have to!)

1968  
Moore’s Law in effect

Better, cheaper, faster

2012  
COTS is King

2020  
Post-Moore’s Law specialization & cleanup

Faster bigger hotter and more expensive

COTS stalled

2030  
New switch? Back to the races!

*CRA/CCC & ACM SIGDA, Pittsburgh, March 2013
Conclusions:

- Dimensional Scaling ("Moore’s Law") is already exhibiting diminishing returns
- The road map beyond 2017 (7nm) is unclear
- While the research community is working on many interesting new technologies (see below), none of them seem mature enough to replace silicon for 2019
  - Carbon nanotube
  - Indium gallium arsenide
  - '2D' devices: MoS2, etc
  - Graphene
  - Spintronics
  - Nanowire
  - Molecular computing
  - Nanowire
  - Molecular computing
  - Photonicics
  - Quantum computing
- 3D IC is considered, by all, as the near term solution, and Monolithic 3D IC is well positioned to be so, as it uses the existing infrastructure
  - It is safe to state that Monolithic 3D is the only alternative that could be ready for high volume in 2019
CMOS is the Best Device Option
3D and EDA need to make up for Moore’s Law, says Qualcomm*

- “Qualcomm is looking to monolithic 3D and smart circuit architectures to make up for the loss of traditional 2D process scaling as wafer costs for advanced nodes continue to increase. .. Now, although we are still scaling down it’s not cost-economic anymore”

- “Interconnect RC is inching up as we go to deeper technology. That is a major problem because designs are becoming interconnect-dominated. Something has to be done about interconnect. What needs to be done is monolithic three-dimensional ICs.”

- “TSV...are not really solving the interconnect issue I’m talking about. So we are looking at true monolithic 3D. You have normal vias between different stacks.”

* Karim Arabi Qualcomm VP of engineering Key Note DAC 2014

[link]
“CEA-Leti Signs Agreement with Qualcomm to Assess Sequential (monolithic)3D Technology”

Business Wire December 08, 2013

“Monolithic 3D (M3D) is an emerging integration technology poised to reduce the gap significantly between transistors and interconnect delays to extend the semiconductor roadmap way beyond the 2D scaling trajectory predicted by Moore’s Law.”

Geoffrey Yeap, VP of Technology at Qualcomm, Invited paper, IEDM 2013

Fig. 17: BEOL performance/area/cost scaling is the foremost issue for 10nm/7nm nodes.
Device technology roadmap

**Evolutionary scaling**: technology driven performance improvement

- 28FDSoI
- 14FDSoI
- 10FDSoI

**Early material and process coupling**

- Non planar/trigate/NW
- High mobility materials

**FinFET**

- 28nm
- 10nm

**Disruptive scaling**

- 14nm

**Alternative to scaling**

- 7nm
  - Steep slope devices
  - Mechanical switches
  - Single Electron Transistor

**Monolithic3D – M3D**

**Early design coupling**
V-NAND Era for the Future

3D V-NAND / No Patterning Limitation

16Gb  128Gb  1Tb
8 stack  24 stack

Paradigm Shift from Drive to Fly

Innovative Technology

Material Structure Integration
Two Types of 3D Technology

3D-TSV
Transistors made on separate wafer @ high temperature, then thin + align + bond

Monolithic 3D
Transistors made monolithically atop wiring (@ sub-400°C for logic)

TSV pitch > 1μm*

TSV pitch ~ 50-100nm

* [Reference: P. Franzon: Tutorial at IEEE 3D-IC Conference 2011]
**MONOLITHIC**

10,000x the Vertical Connectivity of TSV

<table>
<thead>
<tr>
<th>Enables:</th>
<th>TSV</th>
<th>Monolithic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer Thickness</td>
<td>~50μ</td>
<td>~50nm</td>
</tr>
<tr>
<td>Via Diameter</td>
<td>~5μ</td>
<td>~50nm</td>
</tr>
<tr>
<td>Via Pitch</td>
<td>~10μ</td>
<td>~100nm</td>
</tr>
<tr>
<td>Wafer (Die) to Wafer Alignment</td>
<td>~1μ</td>
<td>~1nm</td>
</tr>
</tbody>
</table>

- **Overall Scale**:
  - microns
  - nano-meters
Monolithic/ Sequential 3D

- Wafers Processed separately
- Stacking and Contacting
- 1 TSV/ Block of 10,000 FETs

- Bottom transistor processing
- Top FET processing
- Contacting
- 1 vertical contact/ FET

- 32nm High-K CMOS
- 11 level metal
- Deep trench capacitor
- Cu Through Silicon Via (TSV)

---

Advances, Challenges and Opportunities in 3D CMOS Sequential Integration.
P. Batude et al, IEDM 2011

---

High density 3D integration solutions at the wafer scale.

GSA Silicon Summit 2012 (G.S. Iyer)
3D ICs in older process (65\textit{nm}) is better than 2D ICs built with a newer process (32\textit{nm})

Fig. 6. Cross comparison among various 2D and 3D technologies. Dashed lines are wirelengths of 2D ICs. # dies: 4.

*IEEE IITC11 Kim
Interest for M3D

Without scaling avoid fab and process costs increase

Stack 2 layers: 25% die cost reduction

- Average gain benchmark for 6 circuits/planar

1 node gain without scaling

Source: G. Bartlett, Global Foundries, SMC 2013

Source: R. Gilmore, Qualcomm VP, ESSIRC 2012
The Monolithic 3D Challenge

Why is it not already in wide use?

- **Processing on top of copper interconnects should not make the copper interconnect exceed 400°C**
  - How to bring **mono-crystallized** silicon on top at less than 400°C
  - How to fabricate state-of-the-art transistors on top of copper interconnect and keep the interconnect below at less than 400°C

- **Misalignment of pre-processed wafer to wafer bonding step is ~1µm**
  - How to achieve 100nm or better connection pitch
  - How to fabricate thin enough layer for inter-layer vias of ~50nm
MonolithIC 3D – Breakthrough
3 Classes of Solutions (3 Generations of Innovation)

- **RCAT (2009)** – Process the high temperature on generic structures prior to ‘smart-cut’, and finish with cold processes – Etch & Depositions

- **Gate Replacement (2010) (=Gate Last, HKMG)** - Process the high temperature on repeating structures prior to ‘smart-cut’, and finish with ‘gate replacement’, cold processes – Etch & Depositions

- **Laser Annealing (2012)** – Use short laser pulse to locally heat and anneal the top layer while protecting the interconnection layers below from the top heat
Layer Transfer ("Ion-Cut"/"Smart-Cut")

The Technology Behind SOI

Hydrogen implant of top layer

Flip top layer and bond to bottom layer

Cleave using 400°C anneal or sideways mechanical force. CMP.

Similar process (bulk-to-bulk) used for manufacturing all SOI wafers today
MonolithIC 3D - 3 Classes of Solutions

- **RCAT** – Process the high temperature on generic structure prior to ‘smart-cut’, and finish with cold processes – Etch & Depositions

- Gate Replacement (=Gate Last, HKMG) - Process the high temperature on repeating structure prior to ‘smart-cut’, and finish with ‘gate replacement’, cold processes – Etch & Depositions

- **Laser Annealing** – Use short laser pulse to locally heat and anneal the top layer while protecting the interconnection layers below from the top heat
**Step 1. Donor Layer Processing**

*Step 1* - Implant and activate unpatterned N+ and P- layer regions in standard donor wafer at high temp. (~900°C) before layer transfer. Oxidize (or CVD oxide) top surface.

![Diagram of Step 1]

SiO₂ Oxide layer (~100nm) for oxide-to-oxide bonding with device wafer.

*Step 2* - Implant H+ to form cleave plane for the ion cut

![Diagram of Step 2]

H+ Implant Cleave Line in N+ or below
Step 3 - Bond and Cleave: Flip Donor Wafer and Bond to Processed Device Wafer

Cleave along H+ implant line using 400°C anneal or sideways mechanical force. Polish with CMP.

SiO$_2$ bond layers on base and donor wafers (alignment not an issue with blanket wafers)
Step 4 - Etch and Form Isolation and RCAT Gate

- Litho patterning with features aligned to bottom layer
- Etch shallow trench isolation (STI) and gate structures
- Deposit SiO$_2$ in STI
- Grow gate with ALD, etc. at low temp
  ($<350^\circ$ C oxide or high-K metal gate)

**Advantage:** Thinned donor wafer is transparent to litho, enabling direct alignment to device wafer alignment marks: no indirect alignment. (common for TSV 3D IC)
Step 5 – Etch Contacts/Vias to Contact the RCAT

- Complete transistors, interconnect wires on ‘donor’ wafer layers
- Etch and fill connecting contacts and vias from top layer aligned to bottom layer
MonolithIC 3D - 3 Classes of Solutions

- **RCAT** – Process the high temperature on generic structure prior to ‘smart-cut’, and finish with cold processes – Etch & Depositions

- **Gate Replacement (=Gate Last, HKMG)** - Process the high temperature on repeating structure prior to ‘smart-cut’, and finish with ‘gate replacement’, cold processes – Etch & Depositions

- **Laser Annealing** – Use short laser pulse to locally heat and anneal the top layer while protecting the interconnection layers below from the top heat
A Gate-Last Process for Cleave and Layer Transfer

- Fully constructed transistors attached to each other; no blanket films
- Proprietary methods align top layer atop bottom layer

Donor wafer

Device wafer
A Gate-Last Process for Cleave and Layer Transfer

**Step 3.**
Implant H for cleaving

**Step 4.**
- Bond to temporary carrier wafer (adhesive or oxide-to-oxide)
- Cleave along cut line
- CMP to STI
A Gate-Last Process for Cleave and Layer Transfer

Step 5.
- Low-temp oxide deposition
- Bond to bottom layer
- Remove carrier
A Gate-Last Process for Cleave and Layer Transfer

**Step 6.** On transferred layer:
- Etch dummy gates
- Deposit gate dielectric and electrode
- CMP
- Etch tier-to-tier vias thru STI
- Fabricate BEOL interconnect

Remove (etch) dummy gates, replace with HKMG
Path 2 – Leveraging Gate Last + Innovative Alignment

- Misalignment of pre-processed wafer to wafer bonding step is ~1um
- How to achieve 100nm or better connection pitch
- How to fabricate thin enough layer for inter-layer vias of ~50nm
Novel Alignment Scheme using Repeating Layouts

- Even if misalignment occurs during bonding, repeating layouts allow correct connections.
- Above representation simplistic (high area penalty).
A More Sophisticated Alignment Scheme

Bottom layer layout

Landing pad

Top layer layout

Through-layer connection

Oxide

MonolithIC 3D™ Inc. Patents Pending
MonolithIC 3D - 3 Classes of Solutions

- **RCAT** – Process the high temperature on generic structures prior to ‘smart-cut’, and finish with cold processes – Etch & Depositions

- **Gate Replacement** (=Gate Last, HKMG) - Process the high temperature on repeating structures prior to ‘smart-cut’, and finish with ‘gate replacement’, cold processes – Etch & Depositions

- **Laser Annealing** – Use short laser pulse to locally heat and anneal the top layer while protecting the interconnection layers below from the top heat
Annealing Trend with Scaling

Thermal Annealing Evolution

- Reduced Diffusion
  - Silicon Melt Temperature
- Increased Activation

Anneal Peak Temperature (°C)

- Batch Furnace Annealing >180nm
- Single Wafer RTP Spike Annealing 180-65nm
- Submelt Laser LSA: 65-14nm & beyond
- Ultrafast Laser LXA: 14nm & beyond

Ultratech’s Laser Process

Anneal Time

- Hours
- Minutes
- Seconds
- Milli Seconds
- Micro Seconds
- Nano Seconds
Planar UTBB FD-SOI Scalability: $T_{SOI}$ & $T_{BOX}$

- Electrostatic control improved by thinning $T_{BOX}$
- Scalability down to 10nm node
- Devices already processed with 3.5nm SOI film!

O. Faynot et al, IEDM 2010
K. Cheng et al, VLSI 2011
LSA 100A – Short Pulse, Small Spot

Dwell time ~ 275μs
Two Major Semiconductor Trends help make Monolithic 3D Practical NOW

- As we have pushed dimensional scaling:
  - The volume of the transistor has scaled
    - Bulk \textit{um-sized} transistors \rightarrow \textit{nm} FDSOI & FinFet transistors
  - Processing \textit{times} have trended lower
    - Shallower & sharper junctions, tighter pitches, etc.

=> Much less to heat and for much shorter time
The Top Layer has a High Temperature >1000°C without Heating the Bottom Layers (<400°C) !!!
Process Window Set to Avoid Damage

Temperature variation at the 20 nm thick Si source/drain region in the upper active layer during laser annealing. Note that the shield layers are very effective in preventing any large thermal excursions in the lower layers.
The Monolithic 3D Advantage

II. Reduction die size and power – doubling transistor count
   - Extending Moore’s law

   Monolithic 3D is far more than just an alternative to 0.7x scaling !!!

III. Significant advantages from using the same fab, design tools

IV. Heterogeneous Integration

V. Multiple layers Processed Simultaneously - Huge cost reduction (Nx)

VI. Logic redundancy => 100x integration made possible

VII. Enables Modular Design

VIII. Naturally upper layers are SOI

IX. Local Interconnect above and below transistor layer

X. Re-Buffering global interconnect by upper strata

XI. Others
   A. Image sensor with pixel electronics
   B. Micro-display
Monolithic 3D Provides an Attractive Path to…

- **LOGIC**
  - 3D-CMOS: Monolithic 3D Logic Technology
  - 3D-FPGA: Monolithic 3D Programmable Logic
  - 3D-GateArray: Monolithic 3D Gate Array
  - 3D-Repair: Yield recovery for high-density chips

- **MEMORY**
  - 3D-DRAM: Monolithic 3D DRAM
  - 3D-RRAM: Monolithic 3D RRAM
  - 3D-Flash: Monolithic 3D Flash Memory

- **OPTOELECTRONICS**
  - 3D-Imagers: Monolithic 3D Image Sensor
  - 3D-MicroDisplay: Monolithic 3D Display
  - 3D-LED: Monolithic 3D LED

MonolithIC 3D™ Inc. Patents Pending
Summary

- Monolithic 3D is now practical and well positioned to keep Moore’s Law alive for many years
- Multiple paths to process mono-crystal transistors over copper interconnect
- Monolithic 3D IC provides many opportunities for existing products and for new products & architectures
Back Ups
The Operational Thermal Challenge

- Upper tier transistors are fully surrounded by oxide and have no thermal path to remove operational heat away.

Good Heat Conduction ~100 W/mK

Poor Heat Conduction ~1 W/mK
Use Power Delivery (Vdd, Vss) Network (“PDN”) also for heat removal
Add heat spreader to smooth out hot spots
Add thermally conducting yet electrically non-conducting contacts to problem areas such as transmission gates
Cooling Three-Dimensional Integrated Circuits using Power Delivery Networks (PDNs)

Hai Wei, Tony Wu, Deepak Sekar+, Brian Cronquist*, Roger Fabian Pease, Subhasish Mitra

Stanford University, Rambus+, Monolithic 3D Inc.*
Monolithic 3D Heat Removal Architecture
(Achievable with Monolithic 3D vertical interconnect density)

- Global power grid shared among multiple device layers, local power grid for each device layer
- Local $V_{DD}$ grid architecture shown above
- Optimize all cells in library to have low thermal resistance to $V_{DD}/V_{SS}$ lines (local heat sink)

Monolithic 3D IC

![Graph showing temperature comparison between Monolithic 3D IC with and without power grid]

Patented and Patent Pending Technology
Power Delivery (Vdd, Vss) Network
Provide effective Heat Removal Path

- Stack layer PDN
- Thin $\text{xxxnm}$ Si stack layer with transistors & $\text{xxnm}$ diameter Inter-Layer Vias
- Main Substrate PDN
- Main Si Substrate (thick...xxx um)
- Heat Sink

Graphs showing:
- Max chip temp. versus ILVs/mm²
- Application power density versus Silicon thickness (µm)

- No PDN in model
- Thermally-aware PDN
- System thermal constraint

- All feasible
- Infeasible

Graphs comparing power density at 60°C, 70°C, 80°C for different silicon thicknesses.