3D-FPGA

Abbas El Gamal
Joint work with: Mingjie Lin, Yi-Chang Lu, Simon Wong
Work partially supported by DARPA 3D-IC program

Stanford University
3D-IC Technologies

- **Chip stacking**
  - Vertical interconnect density < 20/\(\text{mm}\)

- **Wafer Stacking**
  - Through Silicon Via (TSV) pitch: 3-5X Via 3
    - 0.54\(\mu\text{m}\)-0.9\(\mu\text{m}\) in 45nm CMOS
  - TSV pitch today 2-5\(\mu\text{m}\)

- **Monolithic Stacking**
  - Roughly same size via as CMOS
  - Can only add few monolithic layers to CMOS
Stanford 3D-IC Program

- Funded through DARPA 3D-IC program
- Interdisciplinary team of nanotechnologists, device physicists, IC technologists, circuit designers, architects
- Goals:
  1. Develop technology to monolithically stack active layers on top of CMOS
     - Challenge: Low temperature processing
  2. Investigate 3D-IC architectures
     - Demonstrate the performance gain of 3D-IC in logic density, delay, and power consumption
     - Demonstration vehicles: FPGA, SRAM
A Modern FPGA

- SRAM programmable
- Fabricated in 65nm CMOS
- Integrates up to 0.33M logic blocks (each roughly 20 gates)
- Up to 10Mb of SRAM block memory
- Embedded microprocessor
- DSP slices
- 550MHz internal clock speed
- 1200 user I/Os including many types of high speed I/Os
FPGA: The Good and the Bad

- FPGAs are becoming increasingly attractive for digital system design:
  - Escalating cell-based ASIC design and prototyping costs in deep submicron
  - Prototyping and field re-programmability
- But, FPGA performance is much lower than cell-based ASIC [Kuon et al. 07]:
  - 10-40X lower logic density
  - 3-4X higher delay
  - 5-12X higher dynamic power
How 3D Can Help Digital

Shorter interconnects → Lower delay and power consumption
Wafer Stacking: Scenario 1

- Stack block memory and hard IPs on top of FPGA logic fabric
  - Relaxed TSV pitch requirement
  - Delay-power benefits depend on size of IPs included
Wafer Stacking: Scenario 2

- Homogeneous stacking (true 3D):
  - Good news: TSV pitch of 3-5x Via 3 allows for stacking several layers with small footprint overhead
  - Issues:
    - Performance benefits may diminish with increased number of layers
      - Programming overhead for vertical dimension
      - TSV parasitics
      - Heat dissipation in intermediate layers
    - Significant modifications to routing architecture and CAD tools needed
Logic Block, Connection Box, switch box: Connect LB I/Os to segments

Segmented Routing Channel

Switch Box: Connect segments to each other

Logic block, connection box, switch box are SRAM programmable
Why is FPGA Inefficient?

<table>
<thead>
<tr>
<th>Logic Block (LB)</th>
<th>Routing Resources (RR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14%</td>
<td>43%</td>
</tr>
<tr>
<td>Logic</td>
<td>buffers + MUXs</td>
</tr>
<tr>
<td>Memory</td>
<td>Memory</td>
</tr>
</tbody>
</table>

Around 80% of the area is overhead
Idea: Stack FPGA Overhead

- Potential benefit: reduce footprint by up to 6 times
Delay:
- HSPICE simulations with BPT model used to extract device and metal wire parameters
- Elmore delay used to determine buffer and switch point sizes
- VPR from University of Toronto used to perform P&R on 20 largest MCNC benchmark designs, extract delay
- Delay improvement: geometric average of all pin-to-pin net delays/critical path delays relative to a baseline 2D-FPGA in 65nm CMOS

Dynamic power:
- Assume typical breakdown of FPGA power consumption between logic block, routing fabric, clock network
- Logic block power doesn’t change with 3D
- Sum up capacitances of routed nets in benchmark designs
- Power improvement: total capacitance relative to baseline 2D-FPGA, factor in fixed logic power consumption
Digression: Baseline 2D-FPGA

LB

CB

Switch Box

Switch Point

SB

<table>
<thead>
<tr>
<th>LB</th>
<th>Single</th>
<th>Double</th>
<th>Length-3</th>
<th>Length-6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Idea: Stack FPGA Overhead

- Potential benefit: reduce footprint by up to 6 times
- Can we do this using wafer stacking?
### Wafer Stacking: Scenario 3

#### Stack configuration memory

<table>
<thead>
<tr>
<th>Configuration Memory</th>
<th>TSV Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>59%</td>
<td>41%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logic</th>
<th>buffers + MUXs</th>
<th>Unoccupied</th>
</tr>
</thead>
<tbody>
<tr>
<td>19%</td>
<td>60%</td>
<td>21%</td>
</tr>
</tbody>
</table>

72% of baseline footprint

- Assume TSV pitch = 4x Via 3
  - TSV area ≈ 1k $\lambda^2$ vs 1.5k $\lambda^2$ for SRAM cell
  - Delay improvement = **1.21X**
  - Dynamic power improvement = **1.14X**
  - Relative to baseline 2D-FPGA in 65nm CMOS
Wafer Stacking: Scenario 3

- Stack logic blocks and some interconnect
  - Length-3/Length-6
  - TSV - 2%
  - Logic blocks: 31% Unoccupied: 10%
  - Programming+Single/Double: 51% Memory: 49%
  - 71% of baseline footprint

- Many fewer TSVs needed
- Delay improvement = 1.21X
- Dynamic power improvement = 1.14X
  - Relative to baseline 2D-FPGA in 65nm CMOS
- Unoccupied area may be used for block memory, IPs
Wafer Stacking Conclusions

- Stacking IPs on top of FPGA fabric feasible but marginally beneficial to delay and power
- Homogeneous stacking promising---needs more investigation
- Stacking FPGA overhead on top of logic provides marginal improvements in delay and power
  - Need significantly finer vertical via pitch to realize potential --> Use monolithic stacking
Monolithically Stacked 3D-FPGA

- 2-T flash [Cao’94]
- 3D SRAM cells [Hitachi’04]
- Reprogrammable via

Ge PMOS:
- Ge rapid growth
- Metal induced Si Epitaxy
- Template based Epitaxy
- Ge nanowire

- Standard CMOS
Monolithically Stacked 3D-FPGA [Lin et al 06]

- LB-SRAM + RR-SRAM
  - Memory
- PT 84% 16%
  - Switch
- LB 45% RR 55%
  - CMOS

- Can achieve:
  - 3.2X higher logic density
  - 1.7X lower geometric mean delay
  - 1.7X less dynamic power consumption

Relative to baseline 2D-FPGA in 65nm CMOS

- Improvements achieved using very few layers on top of CMOS
- How much better can we do by optimizing the FPGA architecture?
Utility of long segments (length 3 and 6) decreases with technology scaling and 3D
Design tool for segmented routing channel

Input:
- FPGA architecture with initial (random) segmentation
- CMOS technology parameters
- A set of benchmark circuits

Cost function:
- Product of average net delay and interconnect power consumption averaged over benchmarks

Use simulated annealing and incremental routing to incrementally optimize cost function--computation easily parallelized

Output:
- An optimized segmentation
Results: Technology Scaling

(a)

(b)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Single</th>
<th>Double</th>
<th>HEX-3</th>
<th>HEX-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>32nm</td>
<td>25</td>
<td>22</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>45nm</td>
<td>26</td>
<td>18</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>65nm</td>
<td>26</td>
<td>19</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>90nm</td>
<td>23</td>
<td>18</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>130nm</td>
<td>20</td>
<td>21</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Avg. Segment Length

<table>
<thead>
<tr>
<th>Technology</th>
<th>Avg. Segment Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>130nm</td>
<td>2.20</td>
</tr>
<tr>
<td>90nm</td>
<td>2.10</td>
</tr>
<tr>
<td>65nm</td>
<td>1.95</td>
</tr>
<tr>
<td>45nm</td>
<td>1.91</td>
</tr>
<tr>
<td>32nm</td>
<td>1.88</td>
</tr>
</tbody>
</table>
Observation 2

Switch point is the bottleneck of interconnect delay/power

<table>
<thead>
<tr>
<th>Technology Generation</th>
<th>Single Interconnects</th>
<th>Double Interconnects</th>
</tr>
</thead>
<tbody>
<tr>
<td>65nm</td>
<td>62%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>18%</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>51%</td>
<td>34%</td>
</tr>
<tr>
<td>90nm</td>
<td>64%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>17%</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td>56%</td>
<td>30%</td>
</tr>
<tr>
<td>130nm</td>
<td>69%</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>28%</td>
</tr>
<tr>
<td>180nm</td>
<td>76%</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>12%</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>21%</td>
</tr>
</tbody>
</table>

- Switch Point
- Side Loading
- Metal Wire
New Routing Fabric [Lin et al. 07]
Highlights

- Horizontal and vertical routing channels
  - Each comprised only of single, double interconnects
  - Longer segments formed by directly connecting several single and double segments without going through routing blocks

- Routing block (RB)
  - Integrates connection and switch box functionalities
  - Provides direct connects between neighboring LBs
  - Provides extended switching capability
Performance Benefits [Lin et al. 07]

- **3.30X** higher logic density
- **2.35X** lower delay (vs. 1.7x)
- **2.82X** less dynamic power (vs. 1.7x)

Relative to baseline 2D-FPGA

- Further logic density improvement can be obtained by optimizing the logic block [Lin 08]
Conclusion

- 3D can help close the performance gap between FPGA and ASIC
- Wafer stacking:
  - Stack IPs on top of FPGA logic fabric
  - Homogeneous stacking--true 3D
- Monolithic stacking:
  - Stack programming overhead on top of logic
  - Significant performance advantages relative to baseline 2D
  - Need few monolithic layers on top of CMOS
Ultimate 3D-FPGA

Monolithic stacking
Homogeneous+ IP
Wafer stacking

Has potential to approach 2D-ASIC performance
Thank You!