Transistor and Logic Design for 5nm Technology Node
Semicon Taiwan 2016

Victor Moroz
September 9, 2016
Outline

• Scaling
• DTCO for 5nm and 3nm
• Nanowire design
• Variability analysis
• Variability repair
• Conclusions
Outline

• **Scaling**

• DTCO for 5nm and 3nm

• Nanowire design

• Variability analysis

• Variability repair

• Conclusions
### Why Scaling?

<table>
<thead>
<tr>
<th>When</th>
<th>What scales?</th>
<th>When does it end?</th>
</tr>
</thead>
</table>
| 1965  | **Moore’s Law (Fairchild):**
        | Double transistor density every couple of years                               | • By 2043, there will be 1 atom per transistor
        |                                                                             | • But you can go up (3D IC)
        |                                                                             | • Great for planning and aligning the industry                                  |
| 1999  | **Claasen’s Law (Philips CEO):**
        | Usefulness = \( \log(\text{Technology}) \), or:
        | Technology = \( \exp(\text{Usefulness}) \)                                  | Forever?                                                                        |
| 2010  | **Koomey’s Law (Stanford Professor):**
        | "at a fixed computing load, the amount of battery you need will fall by a factor of two every year and a half." | • By the [second law of thermodynamics](#) and [Landauer’s principle](#), irreversible computing cannot continue to be made more energy efficient forever. As of 2011, computers have a computing efficiency of about 0.000001%. The Landauer bound will be reached in 2048. Thus, after 2048, the law could no longer hold.
        |                                                                             | • With [reversible computing](#), however, Landauer’s principle is not applicable. With reversible computing, though, computational efficiency is still bounded by the [Margolus-Levitin theorem](#). By the theorem, Koomey’s law has the potential to be valid for about 125 years. |
Transistor Scaling Trend

**Industry scaling roadmap**

*Source: Scotten Jones, Semicon West 2016*

<table>
<thead>
<tr>
<th>Node, nm</th>
<th>CPP, nm</th>
<th>MP, nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>90</td>
<td>64</td>
</tr>
<tr>
<td>10</td>
<td>64</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>54</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
<td>32</td>
</tr>
<tr>
<td>3.5</td>
<td>32</td>
<td>24</td>
</tr>
</tbody>
</table>

CPP is Contacted Poly Pitch, a.k.a. Gate Pitch

MP is Metal Pitch

Scaling from 14nm to 10nm is on track

Scaling beyond 10nm is slower than 0.7x

Is that a problem?
Logic Area Scaling Factors Besides Transistors

- Fin depopulation enables cell height reduction

- SDB is Single Diffusion Break
- DDB is Double Diffusion Break
- IG is Isolating Gates
Logic Area Scaling Trend

Pitch scaling

Cell height

Isolation width

Technology Scaling Roadmap

Logic area scaling

CPP*MP scaling

Moore's law

Pretty much on track!
Outline

• Scaling

• **DTCO for 5nm and 3nm**

• Nanowire design

• Variability analysis

• Variability repair

• Conclusions
### 3nm 2-NAND Cell: Design Choices

<table>
<thead>
<tr>
<th>9 Track Tall, 2 Fins</th>
<th>9 Track Tall, 1 Fin</th>
<th>6 Track Tall, Rotated Fins</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP = 32nm</td>
<td>GP = 32nm</td>
<td>GP = 2 * MP</td>
</tr>
<tr>
<td>MP = 24nm</td>
<td>MP = 24nm</td>
<td>FP = 2 * MP</td>
</tr>
<tr>
<td>FP = 18nm</td>
<td></td>
<td>MP = 24nm</td>
</tr>
</tbody>
</table>

- **90° rotated fins**
- Relaxed gate pitch
- More stress

Rotate the fins
Proposed DTCO: Pre-Si Power Performance Area Evaluation

Multiple sets of DR

<table>
<thead>
<tr>
<th>DR</th>
<th>DR 1</th>
<th>DR 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fin pitch</td>
<td>24 nm</td>
<td>22 nm</td>
</tr>
<tr>
<td>MG ext.</td>
<td>15 nm</td>
<td>15 nm</td>
</tr>
<tr>
<td>Spacer</td>
<td>7 nm</td>
<td>6 nm</td>
</tr>
</tbody>
</table>

Multiple process conditions

GDS: Maxwell or Laker

Litho: Sentaurus

3D structure: Process Explorer

Switching behavior: TCAD

2-NAND cell

Process window

<table>
<thead>
<tr>
<th>Process condition</th>
<th>Design rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process condition</td>
<td>bad</td>
</tr>
<tr>
<td>Design rule</td>
<td>good</td>
</tr>
<tr>
<td>Process condition</td>
<td>bad</td>
</tr>
</tbody>
</table>

Table:

<table>
<thead>
<tr>
<th>Process</th>
<th>T</th>
<th>Time</th>
<th>Etch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process1</td>
<td>480 C</td>
<td>25 min</td>
<td>12 nm</td>
</tr>
<tr>
<td>Process2</td>
<td>475 C</td>
<td>23 min</td>
<td>13 nm</td>
</tr>
</tbody>
</table>
3D Library Cell in Process Explorer

FinFETs  Nano-Wires  Stacked Nano-Wires

Transistors
Power-Performance-Area Evaluation in TCAD

- Transient analysis of the switching behavior in Sentaurus-Device
- Time delay is the averaged pull-up and pull-down delays
- Rigorous current flow analysis in the 3D structure

![Diagram showing Potential vs Time with 3D current crowding]
2-NAND logic cell

Load: Fan-out of 2 plus 70 pitches long BEOL wire

3D current flow in TCAD

2 Fins: Pin capacitance = 0.796 fF
3nm Technology Evaluation: Power Performance Area @TCAD

2-NAND logic cell

Load: Fan-out of 2 plus 70 pitches long BEOL wire

3D current flow in TCAD

2 Fins:
- Pin capacitance = 0.796 fF
- $I_{on}$: -50%
- $C_{pin}$: -40%

1 Fin:
- Pin capacitance = 0.495 fF

Energy per switch, fJ

Switching delay, ps

© 2016 Synopsys, Inc.
3nm Technology Evaluation: Power Performance Area @TCAD

2-NAND logic cell

Load: Fan-out of 2 plus 70 pitches long BEOL wire

3D current flow in TCAD

Load: Fan-out of 2 plus 70 pitches long BEOL wire

Energy per switch, fJ

Switching delay, ps

- 2 Fins
- 1 Fin
- Rotated Fin

- 2 Fins: Pin capacitance = 0.796 fF
- 1 Fin: Pin capacitance = 0.495 fF
  - I_{on}: -50%
  - C_{pin}: -40%
- Rotated Fin: Pin capacitance = 0.432 fF
  - I_{on}: -40%
  - C_{pin}: -45%
Outline

• Scaling
• DTCO for 5nm and 3nm
• Nanowire design
• Variability analysis
• Variability repair
• Conclusions
5nm Nanowire Design

Same layout
(5nm 1x inverter)
5nm Nanowire Design

Same layout (5nm 1x inverter)

Wide SiGe

Narrow SiGe

Different technology options
5nm Nanowire Design

Same layout
(5nm 1x inverter)

Process Explorer

GAA Nano-sheet

Non-GAA Nano-sheet
5nm Nanowire Design

Same layout (5nm 1x inverter)
5nm Nanowire Design

Same layout (5nm 1x inverter)

Potential

Gate

Gate

Gate

Gate

On-Current density

Gate

On-Current density

GAA Nano-sheet

Non-GAA Nano-sheet

Current density

Gate

Gate

Gate

Gate

Gate

Gate

Gate

Gate

Gate
5nm Nanowire Design

Same layout
(5nm 1x inverter)

30% lower $I_{on}$

But 10% lower $C_{MOL}$

GAA Nano-sheet

Non-GAA Nano-sheet

Potential

Gate

Gate

Gate

Gate

D

S

On-Current density

Gate

HK

Gate

HK

Current density

© 2016 Synopsys, Inc.
Outline

• Scaling
• DTCO for 5nm and 3nm
• Nanowire design
• **Variability analysis**
• Variability repair
• Conclusions
Why Variability is Important

- What matters is nominal - 3σ
- Therefore variability affects chip area

Design spec: Nominal – 3σ

Technology A

Nominal

Performance A
Why Variability is Important

- What matters is nominal - $3\sigma$
- Therefore variability affects chip area
- There is no "good enough" variability - the target is zero!
Fin Depopulation Adds Pressure to Variability Scaling

\[ \sigma_{Vt} \sim 1 / \sqrt{\text{(number of fins)}} \]
Fin Depopulation Adds Pressure to Variability Scaling

\[ \sigma_{V_t} \sim \frac{1}{\sqrt{\text{# of fins}}} \]

- Other considerations:
  - Electromigration
  - Power density
  - Fin pitch
Planar to FinFET Transition

Variability Evolution

- FinFETs improve variability
- Planar MOSFETs suffered from RDF
- FinFETs are insensitive to channel doping RDF

Measured data for low Vt process from S. Natarajan et al., IEDM 2014
HKMG Grains Introduce Gate Workfunction Variation

- At 10nm and 7nm nodes, HKMG becomes the dominant variability mechanism
- Introduction of amorphous MG at 7nm would solve this issue
• Planar MOSFETs are insensitive to geometry
• FinFETs and NW are more sensitive to geometry
• NW are less sensitive to L than FinFET, but more sensitive to W
• This data is based on 1 geometry sigma staying at 5% of CD
RDF (Random Dopant Fluctuations)

• Planar MOSFETs suffered from RDF
• FinFETs are insensitive to channel doping RDF

Variability Evolution

Sigma $V_t$, mV vs Technology node, nm

- Total
- HKMG
- RDF
- Geometry
FinFET to Nanowire Transition: Counting Particles

Variability Evolution

- NW variability depends on how many S/D dopants get into the channel
Outline

• Scaling
• DTCO for 5nm and 3nm
• Nanowire design
• Variability analysis
• Variability repair
• Conclusions
Observation in a Recent Nanowire Research Paper

Nanowire Joule self-heating due to high current density when operating at higher than normal Vdd

Driving strength increases after Joule self-heating, but eventually nanowire breaks down

With longer Joule heating, on-state current increases, and subthreshold slope degrades, which is a signature of channel length getting shorter
Repairing Slow Transistors/Gates Removes Design Margin

Spec for repaired chip

Design spec: Nominal – $3\sigma$

Conventional spec

Nominal

• As-manufactured
• Repaired

• Selectively repair transistors that are in critical paths

• What matters is nominal – $3\sigma$

• For 14nm FinFETs, $3\sigma = \sim30\%$

• For 5nm Nanowires, $3\sigma = \sim50\%$
As-Manufactured Distribution of Transistor Properties

- As-manufactured

- This is a typical sample of transistors

- Individual Ion/Ioff points are stretched along the Ion/Ioff trade-off trend for a particular technology
Repairing Slow Part of Transistor Population

- Repair can be applied selectively, for example, only to critical paths
- Over time, some of the transistors will drift towards lower left corner due to NBTI and HCI degradation mechanisms that increase $V_t$
- Then, self-heating repair can be applied again to speed up such transistors
Summary

- Combination of CPP and MP pitch scaling and library cell design evolution provides on-track logic area scaling at least down to 3nm design rules

- 5nm and 3nm technologies have multiple trade-offs in transistor architecture and MOL RC that require holistic DTCO engineering

- Ideal variability is zero, and fin depopulation adds even more pressure

- Nanowire transistors can be selectively repaired to eliminate local variability