Design-Technology Co-Optimization for 5nm Node and Beyond
Semicon West 2016

Victor Moroz
July 12, 2016
## Why Scaling?

<table>
<thead>
<tr>
<th>When</th>
<th>What scales?</th>
<th>When does it end?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>Moore’s Law (Fairchild): Double transistor density every couple of years</td>
<td>• By 2043, there will be 1 atom per transistor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• But you can go up (3D IC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Great for planning and aligning the industry</td>
</tr>
<tr>
<td>1999</td>
<td>Claasen’s Law (Philips CEO): Usefulness = ( \log(\text{Technology}) ), or: Technology = ( \exp(\text{Usefulness}) )</td>
<td>Forever?</td>
</tr>
<tr>
<td>2010</td>
<td>Koomey’s Law (Stanford Professor): &quot;at a fixed computing load, the amount of battery you need will fall by a factor of two every year and a half.&quot;</td>
<td>• 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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 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.</td>
</tr>
</tbody>
</table>
Semiconductor Technology

Scaling
Capacitance
Transistor strength
Existing Early Design Rule 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>

GDS: Maxwell or Laker

Litho: Sentaurus

2-NAND cell

Process window

Design rule: bad, good, bad
Existing Early Design Rule 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>

GDS: Maxwell or Laker

Litho: Sentaurus

- Missing process proximity effects outside of litho
- Missing process interaction with design
- Gives design window, but no guidance within the window
Proposed DTCO: Pre-Si \( P_{\text{ower}}P_{\text{erformance}}A_{\text{rea}} \) 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>

**Process conditions**

**GDS: Maxwell or Laker**

**Litho: Sentaurus**

**3D structure: Process Explorer**

**Switching behavior: TCAD**

**Process window**

2-NAND cell

<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>
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**

<table>
<thead>
<tr>
<th>Process</th>
<th>T</th>
<th>Time</th>
<th>Etch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process 1</td>
<td>480 C</td>
<td>25 min</td>
<td>12 nm</td>
</tr>
<tr>
<td>Process 2</td>
<td>475 C</td>
<td>23 min</td>
<td>13 nm</td>
</tr>
</tbody>
</table>

GDS: Maxwell or Laker

Litho: Sentaurus

3D structure: Process Explorer

Switching behavior: TCAD

2-NAND cell

Process window

**iterations**

Best PPA
Proposed DTCO: Pre-Si PowerPerformanceArea 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

<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>

GDS: Maxwell or Laker

Litho: Sentaurus

3D structure: Process Explorer

Switching behavior: TCAD

- Provides quick PPA estimate
- Includes process effects
- Enables process-design feedback
- Reasonable TAT

• Design rule
  • Bad
  • Good

2-NAND cell

Process window

iterations

Best PPA
2-input NAND library cell with a load of:
- Fan-out of 2
- Metal wire that is 70 metal pitches long

\[ C_{\text{load}} = 2 \times C_{\text{pin}} + C_{\text{wire}} \]

- \( C_{\text{wire}} = 0.34 \text{ fF} \)
- Typical \( C_{\text{load}} \) is 1 fF to 2 fF
Layout: 5nm 2-NAND Cell, 9 Tracks Tall

- Dummy gate
- Fins
- Gates
- S/D contact
- Gate contact

PMOS

NMOS

M2 (PWR)

M1

Via2

Via1

M2 (GND)

3 Gate Pitches wide

9 Metal Pitches tall

GP = 32nm
MP = 24nm
FP = 18nm
3D Library Cell in Process Explorer

**FinFETs**

**Nano-Wires**

**Stacked Nano-Wires**

Transistors

M2

M1

M0

© 2016 Synopsys, Inc.
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
5nm Technology Evaluation: $P_{\text{ower}}P_{\text{erformance}}A_{\text{rea}} @\text{TCAD}$

- 2-NAND logic cell
- Load: Fan-out of 2 plus 70 pitches long BEOL wire
- 3D current flow in TCAD

Graph:
- Energy per switch, fJ on the y-axis
- Switching delay, ps on the x-axis

- Reference 2-fin FF cell
- FF 2x
5nm Technology Evaluation: \( P_{\text{ower}} \)\( P_{\text{erformance}} \)\( A_{\text{rea}} \) @TCAD

![Diagram showing energy per switch and switching delay with 2-NAND logic cell and low MG R reduction.](image)

- Low MG resistance: 10% power reduction
- 2-NAND logic cell
- Load: Fan-out of 2 plus 70 pitches long BEOL wire
- 3D current flow in TCAD

© 2016 Synopsys, Inc.
Insight Into Metal Gate Resistance Effect

• Due to metal resistance, the input signal takes time to get to the fins

• Different parts of the gate experience different biases at any given time

• This is a new effect, due to the lack of space inside MG for tungsten fill, so MG resistivity increases from $\sim 20 \text{ mW cm}$ to $\sim 200 \text{ mW cm}$

• It gets worse for 3 and 4 fins

Electrostatic potential map

2D cut across the gate

Input signal arrives here first
5nm Technology Evaluation: Power \* Performance \* Area @ TCAD

- FF 2x
- FF 2x, low MG R
- FF 1x

Energy per switch, \( fJ \)

Switching delay, \( ps \)

Fin depopulation from 2 to 1: 30% power reduction

2-NAND logic cell

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

3D current flow in TCAD
5nm Technology Evaluation: P_{Power}P_{Performance}A_{Area} @TCAD

Energy per switch, fJ

Switching delay, ps

- FF 2x
- FF 2x, low MG R
- FF 1x
- NW 2x2

Nano-wires: 44% better

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

3D current flow in TCAD
5nm Technology Evaluation: Power Performance Area @TCAD

2-NAND logic cell

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

Nano-wire depopulation: 50% power reduction
5nm Technology Evaluation: Power, Performance, Area @TCAD

2-NAND logic cell

Switching delay, ps

Energy per switch, fJ

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

3D current flow in TCAD

FinFET

NW

FF 2x

FF 2x, low MG R

FF 1x

NW 2x2

NW 2x1

NW 1x2

10%

30%

44%

50%
5nm Technology Evaluation: **P**ower, **P**erformance, **A**rea @ TCAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Energy per switch, fJ</th>
<th>Switching delay, ps</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF 2x</td>
<td>0.796</td>
<td>10%</td>
</tr>
<tr>
<td>FF 1x</td>
<td>0.495</td>
<td>30%</td>
</tr>
<tr>
<td>NW 2x2</td>
<td>0.541</td>
<td>44%</td>
</tr>
<tr>
<td>NW 2x1</td>
<td>0.368</td>
<td>0.54</td>
</tr>
<tr>
<td>NW 1x2</td>
<td>0.364</td>
<td>0.54</td>
</tr>
</tbody>
</table>

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

3D current flow in TCAD
5nm Technology Evaluation: P_owerP_ermanenceA_rea @TCAD

Energy per switch, fJ

- FF 2x
- FF 2x, low MG R
- FF 1x
- NW 2x2
- NW 2x1
- NW 1x2

Switching delay, ps

0 5 10 15 20 25

0 0.5 1 1.5 2

2-fins

FinFET

NW

1 fin

2-NAND logic cell

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

3D current flow in TCAD

Better

Cpin, fF

Ion, normalized

FF 2x
FF 1x
NW 2x2
NW 2x1
NW 1x2

0.796
0.495
0.541
0.368
0.364

0.40%
0.50%
0.54
0.54
50%
50%
0.50
0.368
0.364
5nm Technology Evaluation: \( P_{\text{ower}} \) \( P_{\text{erformance}} \) \( A_{\text{rea}} \) @TCAD

2-NAND logic cell

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

3D current flow in TCAD

Better

Energy per switch, fJ

Switching delay, ps

\( \sim CV^2 \)

\( \sim CV/I \)

\( C_{\text{pin}}, \, \text{fF} \)

\( I_{\text{on}}, \, \text{normalized} \)
5nm Technology Evaluation: Power, Performance, Area @TCAD

- MOL capacitance engineering rules!

Energy per switch, fJ

Switching delay, ps

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

3D current flow in TCAD

© 2016 Synopsys, Inc.
Why Variability is Important

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

\[ \sigma_{V_t} \sim \frac{1}{\sqrt{\text{# of fins}}} \]
Fin Depopulation Adds Pressure to Variability Scaling

$$\sigma_{Vt} \sim 1 / \sqrt{\text{(\# of fins)}}$$

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

- Encouraging trend
- Several “reset buttons”
- There is nothing that can be done to eliminate RDF, so it kept getting worse for planar
- The FinFETs are more sensitive to geometry, which can be better controlled by the equipment

V. Moroz, WMED 2013
Variability Evolution: Planar to FinFET

The lower Sigma Vt values here are due to low Vt process.

Measured data from S. Natarajan et al., IEDM 2014

V. Moroz, WMED 2013
Planar to FinFET Transition

• FinFETs improve variability
• Planar MOSFETs suffered from RDF
• FinFETs are insensitive to channel doping RDF
HKMG Grains Introduce Gate Workfunction Variation

Variability Evolution

- 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 Vt, mV vs Technology node, nm
FinFET to Nanowire Transition: Counting Particles

Variability Evolution

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

- 5nm technology has multiple trade-offs in transistor architecture and MOL RC that require holistic engineering

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

- Several key factors suggest fin depopulation towards 1 fin and beyond (i.e. fractional fins a.k.a. nano-wires):
  - PPA
  - MOL RC
  - Electrostatics (DIBL)
  - Electromigration and power density