Beyond CMOS computing

1. CMOS Scaling

Dmitri Nikonov

Thanks to Kelin Kuhn

Dmitri.e.nikonov@intel.com
Outline

- Moore’s law = scaling
- Performance improvement with scaling
- Latest: tri-gate transistors
- Fundamental limits to scaling
Moore’s Law

Transistor size becoming smaller
Moore’s Law

MORE MOORE

Gordon Moore becoming wiser
Scaling Falsifies Predictions

In the limit, microscope objectives with 0.95 N.A. are available and, provided very small fields (200μ x 200μ) are adequate, linewidths < 0.4μ should be achievable under carefully controlled laboratory conditions, and in very thin resist layers.

Depth of field will be reduced to about ± 0.2μ. Deep U.V. (λ = 200nm - 260nm) lenses will be difficult to build because of the lack of materials that are transparent at these wavelengths and yet have relatively high refractive indices.

1980: Optical Lithography Limit ~ 400nm

IEDM Plenary Session 1980 (Broers)
How Far Scaling Went

1980 SRAM Cell: 1700 $\text{um}^2$

22nm SRAM Cell: 0.092 $\text{um}^2$

Small enough that a 2011 22nm SRAM cell is dwarfed by a 1980 SRAM cell CONTACT

K. Kuhn, MIT invited seminar (MTL), 45nm High-k + Metal Gate Logic Technology, 5-19-08 (images from archives Mark Bohr, 2007)
Metal Interconnects

9 levels of metal
Transistor gate pitch continues to scale 0.7x every 2 years. Proves to be $4^*F$. $F$ is the label for process generations.

M. Bohr, ISCC, 2009.
Economics of Moore’s Law

“Doubling of number of transistors per chip every 2 years”. Lowers cost per transistor. Self-fulfilling prophesy.

Original paper: G.E. Moore, Electronics 19, 114 (1965)
Classical MOSFET scaling was first described by Dennard in 1974

Dennard, IEEE JSSC, 1974
Process Evolution Over Time

90nm – TALL
1.0 \(\mu m^2\)

65nm – WIDE 0.57 \(\mu m^2\)

45nm – WIDE 0.346 \(\mu m^2\)

32nm – WIDE 0.171 \(\mu m^2\)

22nm – WIDE 0.092 \(\mu m^2\)

Bohr
Intel Press release
2012

130 nm 2001
90 nm 2003
65 nm 2005
45 nm 2007
32 nm 2009
22 nm 2011
# Inflection in Scaling

## THEN

- Scaling drove down cost
- Scaling drove performance
- Performance constrained
- Active power dominates
- Independent design-process

## NOW

- Scaling drives down cost
- **Materials** drive performance
- **Power** constrained
- **Standby power** dominates
- **Collaborative** design-process

Images from Bai, Mistry, Natarajan, Auth IEDM/VLSI, 2004/7/8/12 (see course required reading)
Short Channel Effects (SCE)

Degradation of short channel effects

~ 1 / Subthreshold Slope (mV/dec)

~ DIBL (mV/V)

V2 > V1

Decreasing L
Electrostatics Benefits

Basic ID-VG

Gate Voltage (V)

"On" Current

"Off" Current

Channel Current (normalized)

Poor SCE

Reduced Threshold Voltage

Good SCE

Basic ID-VG
Threshold Voltage

~Subthreshold Slope (mV/dec)
~VTlin
~DIBL (mV/V)

http://en.wikipedia.org/wiki/Threshold_voltage
On Current, Off Current

~Subthreshold Slope (mV/dec)

~VTlin

~DIBL (mV/V)

$\text{Idsat}$

$\text{Ioff}$

$\text{Idsat} (=\text{Ion})$

$\text{loff}$

$\text{Drain Current (µA/µm)}$

$\text{Drain Voltage (V)}$
Performance from Scaling

Natarajan, Intel, IEDM, 2008
Where Performance Comes From

Strain and High-k + metal gate are key enablers past the 90nm node

Intel Announced First Tri-Gate (2011)

Mark Bohr, Kaizad Mistry: Intel, April 25th, press release
Close-Up on Tri-gate Transistors

The 22nm process technology was the first to exploit fin-based Tri-Gate devices and combine their benefits with strained silicon and high-k/metal-gate
Nomenclature of Non-Planar Devices

Extremely Thin-SOI
Double-Gate
Pi-Gate
Omega-FET
FinFET
TriGate
Gate-All-Around

Tsi
Wsi
Where Non-Planar Can Go?

UTB

FinFET/Trigate

Nanowire

SS=80
DIBL=95

SS=74
DIBL=68

SS=75
DIBL=55

SS=73
DIBL=63

SS=64
DIBL=32

Cheng – IEDM 2009 (IBM)

Yeh – IEDM 2010 (TSMC)

Tachi – IEDM 2010 (CEA-LETI)
Working With Atomic Dimensions
# International Technology Roadmap for Semiconductors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MPU/ASIC Metal 1 (M1) ½ Punch (nm) (contacted)</td>
<td>38</td>
<td>32</td>
<td>27</td>
<td>24</td>
<td>21</td>
<td>18.9</td>
<td>16.9</td>
<td>15.0</td>
<td>13.4</td>
<td>11.9</td>
<td>10.6</td>
<td>9.3</td>
<td>8.4</td>
<td>7.5</td>
<td>6.7</td>
<td>6.0</td>
</tr>
<tr>
<td><em>Lg</em>: Physical Logic (nm) [1]</td>
<td>24</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>17</td>
<td>15.3</td>
<td>14.0</td>
<td>12.8</td>
<td>11.7</td>
<td>10.6</td>
<td>9.7</td>
<td>8.9</td>
<td>8.1</td>
<td>7.4</td>
<td>6.6</td>
<td>5.9</td>
</tr>
<tr>
<td><em>Vdd</em>: Power Supply Voltage (V) [2]</td>
<td>0.90</td>
<td>0.87</td>
<td>0.85</td>
<td>0.82</td>
<td>0.80</td>
<td>0.77</td>
<td>0.75</td>
<td>0.73</td>
<td>0.71</td>
<td>0.68</td>
<td>0.66</td>
<td>0.64</td>
<td>0.62</td>
<td>0.61</td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td><em>EOT</em>: Equivalent Oxide Thickness (nm) [3]</td>
<td>0.38</td>
<td>0.34</td>
<td>0.39</td>
<td>0.37</td>
<td>0.36</td>
<td>0.37</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td><em>Idss</em>: NMOS Drive Current (μA/μm) [14]</td>
<td>1,320</td>
<td>1,367</td>
<td>1,422</td>
<td>1,496</td>
<td>1,582</td>
<td>1,670</td>
<td>1,775</td>
<td>1,887</td>
<td>1,942</td>
<td>1,847</td>
<td>1,916</td>
<td>1,976</td>
<td>2,030</td>
<td>2,087</td>
<td>2,152</td>
<td>2,228</td>
</tr>
<tr>
<td><em>Vth</em>: Equivalent Injection Velocity, vth (10^7 cm/s) [15]</td>
<td>1.03</td>
<td>1.09</td>
<td>1.11</td>
<td>1.13</td>
<td>1.24</td>
<td>1.33</td>
<td>1.39</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
</tr>
<tr>
<td><em>Cg</em>: Fringing Capacitance (fF/μm) [16]</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td><em>Cext</em>: Total Gate Capacitance for Calculation of CV/1 (fF/μm) [17]</td>
<td>0.936</td>
<td>0.898</td>
<td>0.872</td>
<td>0.851</td>
<td>0.834</td>
<td>0.816</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
</tr>
<tr>
<td><em>Cv</em>: NMOSFET Dynamic Power Indicator (fF/μm) [18]</td>
<td>0.76</td>
<td>0.68</td>
<td>0.63</td>
<td>0.57</td>
<td>0.53</td>
<td>0.49</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td><em>τ</em>: NMOSFET Intrinsic Delay (ps) [19]</td>
<td>0.64</td>
<td>0.57</td>
<td>0.52</td>
<td>0.47</td>
<td>0.42</td>
<td>0.38</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
</tr>
</tbody>
</table>

## Nikonov 1. CMOS
Limits of Computing

Moving from fundamental limits given by the laws of physics to practical limitations. These limits tend to be broken.

Equations for Fundamental Limits

**Thermodynamics:** if energy is less than thermal – bit errors

\[ E_{sw} = 4kT \]

**Quantum Mechanics:** energy time uncertainty

\[ E_{sw} t_{sw} \geq h \]

**Relativity:** Signal no faster than the speed of light

\[ \frac{L}{\tau} \leq c_0 \]

Better Fundamental Limits

Thermodynamic limit on bit error ratio

\[ E_{sw} = 3kT \]

Higher energy = faster switching

\[ E_{sw} t_{sw} = \pi \hbar \]

Energy of electron in a transistor is limited by quantum confinement

\[ E_{sw} = \frac{3\pi^2 \hbar^2}{2ma^2} \]

Gate raised = confined in the source.
Gate lowered = can travel to drain.

Solving equations together gives limits

\[ a = 3.8\text{nm} \quad t_{sw} = 27\text{fs} \quad E_{sw} = 1.2 \times 10^{-20}\text{J} = 78\text{meV} \]

MOSFET Scales Towards the Limit

Current CMOS device scaling close to the ideal limits

* Data courtesy of Robert Chau (Intel)
How long is left for Moore’s law

Intel’s generation to HVM

<table>
<thead>
<tr>
<th>Year</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>14nm</td>
</tr>
<tr>
<td>2017</td>
<td>7nm</td>
</tr>
<tr>
<td>2021</td>
<td>3.5nm</td>
</tr>
</tbody>
</table>

ITRS start of production

<table>
<thead>
<tr>
<th>Year</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>32nm</td>
</tr>
<tr>
<td>2018</td>
<td>15nm</td>
</tr>
<tr>
<td>2024</td>
<td>7.5nm</td>
</tr>
<tr>
<td>2030</td>
<td>3.8nm</td>
</tr>
</tbody>
</table>

Scaling might end between 2021 and 2030
But it is NOT the end of Moore’s law:
better architectures, 3D circuits.
Summary

- Moore’s law = 0.7 size every 2 years
- Despite trends, Intel developers manage to improve performance
- Tri-gate transistors = major advance
- Fundamental laws limit size scaling to ~4nm