Technology Scaling and Roadmap for 22nm CMOS logic and beyond

October 27, 2008

@Dalian University of Technology

Hiroshi Iwai

Tokyo Institute of Technology
Tokyo Institute of Technology

Founded in 1881, Promoted to Univ. 1929
Tokyo Institute of Technology
東京工業大学

Promoted to Univ. 1929
Total 10,000 students: 5,000 under graduate 5,000 graduate

(As of May. 1, 2005)
Outline

1. Scaling

2. ITRS Roadmap

3. Voltage Scaling/ Low Power and Leakage

4. SRAM Cell Scaling

5. Roadmap for further future as a personal view
1. Scaling
Scaling Method: by R. Dennard in 1974

**Wdep**: Space Charge Region (or Depletion Region) Width

- Wdep has to be suppressed
- Otherwise, large leakage between S and D

Potential in space charge region is high, and thus, electrons in source are attracted to the space charge region.

**K=0.7** for example

By the scaling, Wdep is suppressed in proportion, and thus, leakage can be suppressed.

**X, Y, Z**: K, **V**: K, **Na**: 1/K

**Wdep** ~ \( \sqrt{V/Na} \)

**Good scaled I-V characteristics**
# Downscaling merit: Beautiful!

<table>
<thead>
<tr>
<th>Geometry &amp; Supply voltage</th>
<th>$L_g$, $W_g$</th>
<th>$K$</th>
<th>Scaling $K$ : $K=0.7$ for example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive current in saturation</td>
<td>$I_d$</td>
<td>$K$</td>
<td>$I_d = v_{sat} \frac{W_g C_o (V_g-V_{th})}{W_g t_{ox}}$</td>
</tr>
<tr>
<td>$I_d$ per unit $W_g$</td>
<td>$I_d/\mu m$</td>
<td>1</td>
<td>$I_d$ per unit $W_g = I_d / W_g = 1$</td>
</tr>
<tr>
<td>Gate capacitance</td>
<td>$C_g$</td>
<td>$K$</td>
<td>$C_g = \varepsilon_o \varepsilon_{ox} L_g W_g / t_{ox}$, $KK/K = K$</td>
</tr>
<tr>
<td>Switching speed</td>
<td>$\tau$</td>
<td>$K$</td>
<td>$\tau = C_g V_{dd}/I_d$, $KK/K = K$</td>
</tr>
<tr>
<td>Clock frequency</td>
<td>$f$</td>
<td>$1/K$</td>
<td>$f = 1/\tau = 1/K$</td>
</tr>
<tr>
<td>Chip area</td>
<td>$A_{chip}$</td>
<td>$\alpha$</td>
<td>$\alpha$: Scaling factor, In the past, $\alpha&gt;1$ for most cases</td>
</tr>
<tr>
<td>Integration (# of Tr)</td>
<td>$N$</td>
<td>$\alpha/K^2$</td>
<td>$N \rightarrow \alpha/K^2 = 1/K^2$, when $\alpha=1$</td>
</tr>
<tr>
<td>Power per chip</td>
<td>$P$</td>
<td>$\alpha$</td>
<td>$f NC V^2/2 \rightarrow K^{-1}(\alpha K^{-2})K (K^1)^2 = \alpha = 1$, when $\alpha=1$</td>
</tr>
<tr>
<td>$k= 0.7$ and $\alpha = 1$</td>
<td>$k= 0.7^2 = 0.5$ and $\alpha = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single MOFET</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$V_{dd} \rightarrow 0.7$</td>
<td>$V_{dd} \rightarrow 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_g \rightarrow 0.7$</td>
<td>$L_g \rightarrow 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$I_d \rightarrow 0.7$</td>
<td>$I_d \rightarrow 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_g \rightarrow 0.7$</td>
<td>$C_g \rightarrow 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P$ (Power)/Clock</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow 0.7^3 = 0.34$</td>
<td>$\rightarrow 0.5^3 = 0.125$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$ (Switching time)</td>
<td>$\rightarrow 0.7$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow 0.5$</td>
<td>$\rightarrow 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Chip</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$ (# of Tr)</td>
<td>$N$ (# of Tr)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow 1/0.7^2 = 2$</td>
<td>$\rightarrow 1/0.5^2 = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f$ (Clock)</td>
<td>$f$ (Clock)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow 1/0.7 = 1.4$</td>
<td>$\rightarrow 1/0.5 = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P$ (Power)</td>
<td>$P$ (Power)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow 1$</td>
<td>$\rightarrow 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- The concerns for limits of down-scaling have been announced for every generation.

- However, down-scaling of CMOS is still the ‘royal road’* for high performance and low power.

- Effort for the down-scaling has to be continued by all means.

* Euclid of Alexandria (325BC?-265BC?)

‘There is no royal road to Geometry’

Mencius (Meng-zi), China (372BC?-289BC?)

孟子: 王道, 覇道 (Rule of right vs. Rule of military)
Actual past downscaling trend until year 2000

Past 30 years scaling
Merit:  N, f increase
Demerit:  P increase

Vdd scaling insufficient
Additional significant increase in Id, f, P

Change in 30 years

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Ideal scaling</th>
<th>Real Change</th>
<th>Ideal scaling</th>
<th>Real Change</th>
<th>Ideal scaling</th>
<th>Real Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lg</td>
<td>K</td>
<td>10^-2</td>
<td>10^-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tox</td>
<td>K(10^-2)</td>
<td>10^-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vdd</td>
<td>K(10^-2)</td>
<td>10^-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achip</td>
<td>α</td>
<td>10^1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Vd scaling insufficient, α increased → N, Id, f, P increased significantly

- Now, power and/or heat generation are the limiting factors of the down-scaling

- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.

- Growth rate in clock frequency and chip area becomes smaller.
2. ITRS Roadmap
   (for 22 nm CMOS logic)
ITRS Roadmap does change every year!

- 2007 Edition
- 2006 Update
- 2005 Edition
- 2004 Update
- 2003 Edition
- 2002 Update
- 2001 Edition
- 2000 Update

http://www.itrs.net/reports.html

The current latest version: ITRS 2007 Edition

ITRS 2008 Update will be published on the web at the end of Dec 2008 or Jan. 2009
HP, LOP, LSTP for Logic CMOS

- **HP CMOS** (High Performance)
  - Highest Ion, Lowest CV/I
  - High leakage
  - Medium Vdd

- **LOP CMOS** (Low Operation Power)
  - Lowest Vdd
  - Medium Ion, medium CV/I
  - Medium leakage

- **LSTP CMOS** (Low Standby Power)
  - Lowest leakage
  - Low Ion, high CV/I
  - High Vdd

Source: 2007 ITRS Winter Public Conf.

Operation Frequency (a.u.)

Subthreshold Leakage (A/µm)
What does *22 nm* mean in 22 nm CMOS Logic?

### 'XX nm CMOS Technology

<table>
<thead>
<tr>
<th>Technology name</th>
<th>Starting Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 nm</td>
<td>2007</td>
</tr>
<tr>
<td>32 nm</td>
<td>2009?</td>
</tr>
<tr>
<td>22 nm</td>
<td>2011?~ 2012?</td>
</tr>
<tr>
<td>16 nm</td>
<td>2013?~ 2014?</td>
</tr>
</tbody>
</table>

### ITRS (Likely in 2008 Update) for High Performance Logic

<table>
<thead>
<tr>
<th>Year</th>
<th>Half Pitch (1st Metal)</th>
<th>Physical Gate Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>68 nm</td>
<td>32 nm</td>
</tr>
<tr>
<td>2008</td>
<td>59 nm</td>
<td>29 nm</td>
</tr>
<tr>
<td>2009</td>
<td>52 nm</td>
<td>27 nm</td>
</tr>
<tr>
<td>2010</td>
<td>45 nm</td>
<td>24 nm</td>
</tr>
<tr>
<td>2011</td>
<td>40 nm</td>
<td>22 nm</td>
</tr>
<tr>
<td>2012</td>
<td>36 nm</td>
<td>20 nm</td>
</tr>
<tr>
<td>2013</td>
<td>32 nm</td>
<td>18 nm</td>
</tr>
<tr>
<td>2014</td>
<td>29 nm</td>
<td>16 nm</td>
</tr>
</tbody>
</table>

Source: 2008 ITRS Summer Public Conf.

### ‘XX nm’ CMOS Logic Technology:
- In general, there is no common corresponding parameter with ‘XX nm’ in ITRS table, which stands for ‘XX nm’ CMOS.
Definition of the Half Pitch

Logic 1st Metal Half Pitch

DRAM ½ Pitch
= DRAM Metal Pitch/2
MPU/ASIC M1 ½ Pitch
= MPU/ASIC M1 Pitch/2

Metal
Pitch

Typical DRAM/MPU/ASIC Metal Bit Line

Source: 2008 ITRS Summer Public Conf.
What does ‘22 nm’ mean in 22 nm CMOS Logic?

8µm → 6µm → 4µm → 3µm → 2µm → 1.2µm → 0.8µm → 0.5µm
- Originally, ‘XX’ means lithography resolution.
- Thus, ‘XX’ was the gate length, and half pitch of lines
- ‘XX’ had shrunk 0.7 in 3 years in average (0.5 in 6 years) those days.
- ‘XX’ value deviated among companies: example: 1.5µm, 1.2µm, 1µm

→ 350nm → 250nm → 180nm → 130nm → 90nm → 65nm → 45nm
- ‘XX’ values were established by NTRS* and ITRS with the term of ‘Technology Node**’ and ‘Cycle***’ using typical ‘half pitch value’.
  *NTRS: National Tech. Roadmap, **Term ‘Technology Node’ is not used now.
  ***Cycle: Period or year for which the half pitch becomes X0.71.
- The gate length of logic CMOS became smaller with one or two generations from the half pitch, and ‘XX’ names ahead of generations have been used for logic CMOS.
- Memory still keeps the half pitch as the value of ‘XX’

→ 32nm → 22nm → 16nm → 11nm → 8nm?? → 5.5nm ??
For example, Typical Half Pitches at ITRS 2007

Source: 2008 ITRS Summer Public Conf.  2007 - 2022 ITRS Range
Physical gate length in past ITRS was too aggressive. The dissociation from commercial product prediction will be adjusted.

Physical gate length of High-Performance logic will shift by 3-5 yrs.

Correspond to
45nm 32nm 22nm Logic CMOS

Source: 2008 ITRS Summer Public Conf.
EOT and Xj shift backward, corresponding to Lg shift

EOT: 0.55 nm → 0.88 nm, Xj: 8 nm → 11 nm @ 22nm CMOS

Likely in 2008 Update

Correspond to 22nm

Source: 2008/ITRS Summer Public Conf.

<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>2007 MPU/ASIC Lg (nm)</td>
<td>25</td>
<td>23</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6.3</td>
<td>5.6</td>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>2008 MPU/ASIC Lg (nm)</td>
<td>32</td>
<td>29</td>
<td>27</td>
<td>24</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>17</td>
<td>15</td>
<td>14.0</td>
<td>12.8</td>
<td>11.7</td>
<td>10.7</td>
<td>9.7</td>
<td>8.9</td>
<td>8.1</td>
</tr>
<tr>
<td>Shift/Interpolate Formula</td>
<td>2005</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>2009</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
<td>interp</td>
</tr>
</tbody>
</table>

EOT w/3E20 poly, bulk MPU (nm)

<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>1.2</td>
<td>0.71</td>
<td>0.54</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EOT w/3E20 poly, bulk MPU (nm)

<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>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>1</td>
<td>0.66</td>
<td>0.64</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EOT w/metal gate, bulk MPU (nm)

<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>0.9</td>
<td>0.75</td>
<td>0.65</td>
<td>0.55</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Likely in 2008 Update

EOT w/metal gate, bulk MPU (nm)

<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>1.0</td>
<td>0.95</td>
<td>0.88</td>
<td>0.75</td>
<td>0.60</td>
<td>0.60</td>
<td>0.53</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Likely in 2008 Update

Drain Ext. Xj bulk MPU (nm)

<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>12.5</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Likely in 2008 Update

Drain Ext. Xj bulk MPU (nm)

<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>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>8.5</td>
<td>7.7</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

non-steady trend corrected

filled in for metal gate EOT for 2009/10 based on latest conference presentations
Clock frequency does not increase aggressively anymore.

Advantage in RISC
Simple configuration

Advantage in SISC
Era for ‘out of order’

Even decreased!

Multi Core
Clock
Performance

Source: Mitsuo Saito, Toshiba
MPU “GHz” by “Cores” ITRS2007

Source: 2007 ITRS Winter Public Conf.

Cell Broadband Engine

6GHz capability for SRAM

Source: IBM, Toshiba, Sony
ISSCC2008 and 08
Structure and technology innovation (ITRS 2007)

Source: 2008 ITRS Summer Public Conf.
Technology innovation described in ITRS 2007

Alternative material (Ge, III-V) and structure (Nanowire) in channel region.

Source: 2007 ITRS Winter Public Conf.
Timing of CMOS innovations shifts backward.

**Bulk CMOS has longer life now!**

Correspond to 22nm Logic CMOS

Source: 2008 ITRS Summer Public Conf.
Roadmap towards 22nm technology and beyond

- Physical gate length downsizing rate will be less aggressive.

- Corresponding to the above, performance increase would slow down – Clock frequency, etc.

- Introduction of innovative structures – UTB SOI and DG delayed, and bulk CMOS has longer life than predicted by previous ITRS roadmaps.
3. Voltage Scaling
   / Low Power and Leakage
Difficulty in Down-scaling of Supply Voltage: $V_{dd}$

Because, $V_{th}$ cannot be down-scaled anymore, $V_{dd}$ down-scaling is difficult.

$V_{dd} - V_{th}$ determines the performance (High $I_d$) and cannot be too small.

$\Delta V_{th}$: $V_{th}$ variation

$> \Delta V_{th}$

Margin for $V_{th}$ variation is necessary
Subthreshold leakage current of MOSFET

Subthreshold leakage current of MOSFET

Subthreshold Current Is OK at Single Tr. level
But not OK For Billions of Trs.

Subthreshold region

Vg=0V

Subthreshold Leakage Current

Vth (Threshold Voltage)

Id

Ion

loff

OFF

ON
Vth cannot be decreased anymore

significant Ioff increase

Vth: 300mV → 100mV
Ioff increases with 3.3 decades

(300 – 100)mV/(60mv/dec) = 3.3 dec

Subthreshold slope (SS)
= (Ln10)(kT/q)(C_{ox}+C_{D}+C_{it})/C_{ox}
> ~ 60 mV/decade at RT

SS value:
Constant and does not become small with down-scaling
**ITRS for HP logic**

- **Vdd** will stay higher in 2008 update.
- **Vth-sat** will be around 0.1V.

**Saturated Vth**

Source: ITRS and 2008 ITRS Summer Public Conf.

---

2008 Values are from ITRS Public Conf. and still under discussion.
SS (Subthreshold Slope) becomes worse in the following cases

1. Improper down-scaling
   Ex. When $T_{ox}$, $W_{dep}$, or $V_{dd}$ is not scaled

2. High impurity doping in channel or substrate
   High impurity Conc.
   $\rightarrow C_D$ increase
   $\rightarrow$ SS increase
   $SS = (\ln 10) (kT/q)(C_{ox} + C_D + C_{it})/C_{ox}$

3. Enhanced Drain-Electric-field penetration through oxide
   Ex. High-k, SOI, Multi-gate (Double gate: DG)
   DG and SOI often show better SS, but be careful!
Comparison of Bulk and DG

Bulk

DG

Same parameter condition for both
(2006 ITRS Bulk parameters are used for both Bulk and DG)

Lg=16nm, tox(EOT)=0.5nm, Dopant@Channel=8.1X10^{18}cm^{-2}

\[ \Lambda : \text{Penetration Depth of DIBL} \]

\[ \frac{\partial V(x,y)}{\partial V_d} \] \( V_d=1V \)


DIBL: Drain Induced Barrier Lowering

\[ \Lambda : \text{DIBL penetration (nm)} \]

\[ \Delta : \text{DIBL at drain edge} \]

\[ \text{Sub-threshold Swing (mV/dec)} \]

\[ \text{DIBL@D Edge (mV/V)} \]

\[ \text{Fin Width (nm)} \]

\[ W_{\text{fin}} = 10.7 \text{ nm} \]
\[ W_{\text{fin}} = 30 \text{ nm} \]
\[ W_{\text{fin}} = 40 \text{ nm} \]
Comparison of High-k and SiO₂ MOSFETs

\[
I_d = 0.05 \\
V_d = 0.1V \\
EOT = 2nm
\]

- \( k = 390 \) SiO₂ Too large High-k
- \( K = 3.9 \) Too large high-k

Penetration of lateral field from Drain through high-k causes significant short channel effects

V_{dd} will stay higher than predicted by previous ITRS roadmaps.

Solution towards Low V_{dd}
Effort to reduce I_{sd-leak} and increase I_{d-sat} is important

- Scaling: Proper down-scaling
  - Introduction of Next generation high-k, S/D etc.
  - CD* variation control by lithography and etching techniques
    * CD: Critical dimension

- Structure: Bulk → UTB-SOI → DG → Nanowire
- Variation: Proper scaling by new tech. – High-k, litho. Etc.
  V_{th} adjustment by V_{sub} control
- Circuit techniques: Dynamic and local Multi-V_{dd}, etc.
Another concern for Low $V_{dd}$ besides $I_{sd\text{-leak}}$ increase

$\rightarrow$ Huge power loss for voltage conversion to such low $V_{dd}$
4. SRAM cell scaling
Intel’s **SRAM** test chip trend

*Source: B. Krzanich, S. Natrajan, Intel Developer’s Forum 2007*

http://download.intel.com/pressroom/kits/events/idffall_2007/Briefing
Silicon&TechManufacturing.pdf

<table>
<thead>
<tr>
<th>Process name</th>
<th>Lithography</th>
<th>1st production</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1264</td>
<td>65nm</td>
<td>2005</td>
</tr>
<tr>
<td>P1266</td>
<td>45nm</td>
<td>2007</td>
</tr>
<tr>
<td>P1268</td>
<td>32nm</td>
<td>2009</td>
</tr>
<tr>
<td>P1270</td>
<td>22nm</td>
<td>2011</td>
</tr>
</tbody>
</table>

Only schedule has been published

**SRAM down-scaling trend has been kept until 32nm and probably so to 22nm**

![Graph showing the trend of cell area over years](image)

<table>
<thead>
<tr>
<th>Technology</th>
<th>90 nm Process</th>
<th>65 nm Process</th>
<th>45 nm Process</th>
<th>32 nm Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell size</td>
<td>1.0 µm²cell</td>
<td>0.57 µm²cell</td>
<td>0.346 µm²cell</td>
<td>0.182 µm²cell</td>
</tr>
<tr>
<td>Capacity</td>
<td>50 Mbit</td>
<td>70 Mbit</td>
<td>153 Mbit</td>
<td>291 Mbit</td>
</tr>
<tr>
<td>Chip area</td>
<td>109 mm²</td>
<td>110 mm²</td>
<td>119 mm²</td>
<td>118 mm²</td>
</tr>
<tr>
<td>Functional Si</td>
<td>February '02</td>
<td>April '04</td>
<td>January '06</td>
<td>September '07</td>
</tr>
</tbody>
</table>
22 nm technology 6T SRAM Cell: Size = 0.1μm

Announced on Aug 18, 2008

Consortium: IBM (NYSE), AMD, Freescale, STMicroelectronics, Toshiba and the College of Nanoscale Science and Engineering (CNSE)

0.1μm cell size is almost on the down-scaling trend

New technologies introduced
- High-NA immersion lithography
- High-K metal gate stacks
- 25 nm gate lengths
- Thin composite oxide-nitride spacers
- Advanced activation techniques
- Extremely thin silicide
- Damascene copper contacts

Static noise margin of 220 mV at 0.9 V

Source: IEDM2008 Pre-conference Publicity
http://www.btbmarketing.com/iedm/
Cell size reduction trends

1/2 or 2/3 per cycle?

- **Intel**
  - 65nm: Apr. 2004
  - 45nm: Jan. 2006
  - 32nm: Sep. 2007

- **TSMC**
  - 45nm: Dec. 2007
  - 32nm: Dec. 2007

- **IBM Gr. (Consortium)**
  - 32nm: Dec. 2007
  - 22nm: Aug. 2008

**Functional Si**
- 65nm: Apr. 2004
- 45nm: Jan. 2006
- 32nm: Sep. 2007

**Conference (IEDM)**
- 45nm: Dec. 2007
- 32nm: Dec. 2007

**Press release**
- 22nm: Aug. 2008
\[ \sigma V_{\text{Tran}} = \left( \frac{4q^3 \varepsilon_{si} \phi_B}{2} \right) \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \left( \frac{4\sqrt{N}}{\sqrt{W_{\text{eff}} \cdot L_{\text{eff}}}} \right) \]

\[ = \frac{1}{\sqrt{2}} \left( \frac{c_2}{\sqrt{W_{\text{eff}} \cdot L_{\text{eff}}}} \right) \]

NMOS Mismatch Coefficient \( (C_2) \) improvement with technology scaling

Source: K.J.Kuhn
IEDM 2007
Mismatch improvement by layout (Intel)

“tall” design
90nm : 1.0 μm²

“wide” design
65nm : 0.57 μm²

“wide” design (Square endcaps)
45nm 0.346 μm²

Source: K. J. Kuhn
IEDM2007 Tech. Dig. pp.471
Double patterning for square endcap

- Pattern gate lines/spaces
- Pattern cut mask
- Final gate pattern
- Intel 45nm SRAM cell

Source: M. Bohr, ICSICT2008

TSMC 45nm
TSMC 32nm
IBM Gr. 32nm
IBM Gr. 22nm

TSMC 45nm
TSMC 32nm
IBM Gr. 32nm
IBM Gr. 22nm

Source: M. Bohr, ICSICT2008
Most Difficult part of SRAM down-scaling is Vdd down-scaling

Density of on-chip cache SRAM memory is high and thus, Vth cannot be down-scaled too much because of large Isd-leak

Also, under low Vdd, read- and write margin degrades, data retention degrade.

Thus, Vdd down-scaling is more severe in SRAM than logic part of the circuits
Intel® Xeon® 7400 Series (Dunnington)

Source: Intel Developer Forum 2008

45 nm high-k6 cores
16MB shared L3 cache

Cache occupies huge area
→ Cell size of SRAM should be minimized
→ Isd-leak should be minimized
   → Vth are often designed to be higher than logic Vth
   → Lg are often designed to be larger than logic Lg
Nehalem (Intel) 2, 4 or 8 Cores

Voltage/Frequency Partitioning
- DDR Vcc
- Core Vcc
- Uncore Vcc

Dynamic Power Management

Chip

Core

8T SRAMCell
- 32kB L1 I-cache
- 32kB L1 D-cache
- 256kB L2-cache

6T SRAMCell
- 8 MB L3 cache

Source: Intel Developer Forum 2008
5. Roadmap for further future as a Personal View
When?
At what size?
Approaching Downsizing limit!

What is waiting beyond the cloud?

Source: ITRS2005
Could we squeeze technologies for ultimate CMOS scaling?

Saturation of EOT thinning is a serious roadblock to proper down-scaling.

Is 0.5nm real limit?
EOT<0.5nm with Gain in Drive Current is Possible

La$_2$O$_3$ gate insulator

(a) EOT=0.37nm  \(V_{th}=-0.04V\)
(b) EOT=0.43nm  \(V_{th}=-0.03V\)
(c) EOT=0.48nm  \(V_{th}=-0.02V\)

W/L=2.5/50\(\mu\)m
PMA 300°C (30min)

Drain current (mA) vs Drain voltage (V)

Insufficient compensation region

EOT scaling below 0.5nm
Still useful for larger drain current


*Because \(L_g\) is very large (2.5\(\mu\)m), gate leakage is large in case (a). The gate leakage component was subtracted from measured data for case (a). However, if we make small gate length, the gate leakage current should become sufficiently small to be ignored compared with \(I_d\) as we verified with SiO$_2$ gate before (Momose et al.,IEDM 1994). The gate leakage could be suppressed by modifying material and process in future.
Thus, in future, maybe continuous development of new techniques could make more proper down-scaling possible.

It is difficult to say, but EOT and Vdd may become smaller than expected today.
Personal view:
- There will be still 4~6 cycles (or technology generations) left until we reach 11 ~ 5.5 nm technologies, at which we will reach down-scaling limit, in some year between 2020-30.

- Effort and passion to pursue the down-scaling until then are important.

Probably, still CMOS and Si. Maybe, Ge, III-V, C

*5.5nm? was added by Iwai*
After reaching the down-scaling limit, we could still continue R & D, seeking sufficiently higher Id-sat under low Vdd.

Two candidates have emerged for R & D for the above.

1. Nanowire/tube MOSFETs
2. Alternative channel MOSFETs (III-V, Ge)

Other Beyond CMOS devices are still in the cloud.
Our new roadmap

Source: H. Iwai, IWJT 2008

Extended CMOS: More Moore + CMOS logic

PJT (2007~2012)

Beyond the horizon

Natural direction of downsizing

Selection

Graphene

CNT

Tube, Ribbon

Nanowire

Ill-Vand Ge Nano wire

Si Nanop wire

Si Fin, Tri-gate

High conduction

By 1D conduction

Extended CMOS

More Moore

Cloud

ITRS

ITRS Beyond CMOS
Thank you for your attention!