Technology Scaling and Roadmap for 22nm CMOS and beyond

March 13, 2009

@Kyongpook National University

Hiroshi Iwai

Tokyo Institute of Technology
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

$W_{dep}$: 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, $W_{dep}$ is suppressed in proportion, and thus, leakage can be suppressed.

$W_{dep} \propto \sqrt{V/Na}$

$X, Y, Z: K, \quad V: K, \quad Na: 1/K$

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>
</tr>
</thead>
<tbody>
<tr>
<td>Drive current in saturation</td>
<td>$I_d$</td>
<td>$K$</td>
</tr>
<tr>
<td>$I_d$ per unit $W_g$</td>
<td>$I_d/\mu m$</td>
<td>1</td>
</tr>
<tr>
<td>Gate capacitance</td>
<td>$C_g$</td>
<td>$K$</td>
</tr>
<tr>
<td>Switching speed</td>
<td>$\tau$</td>
<td>$K$</td>
</tr>
<tr>
<td>Clock frequency</td>
<td>$f$</td>
<td>$1/K$</td>
</tr>
<tr>
<td>Chip area</td>
<td>$A_{\text{chip}}$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Integration (# of Tr)</td>
<td>$N$</td>
<td>$\alpha/K^2$</td>
</tr>
<tr>
<td>Power per chip</td>
<td>$P$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

#### Scaling $K$: $K=0.7$ for example

- $I_d = \frac{V_{\text{sat}} W_g C_o (V_g-V_{\text{th}})}{W_g (t_{ox}^{-1}) (V_g-V_{\text{th}})}$
- $C_o$: gate $C$ per unit area

- $I_d = W_g t_{ox}^{-1} (V_g-V_{\text{th}}) = KK^{-1}K = K$
- $I_d$ per unit $W_g = I_d / W_g = 1$

- $C_g = \frac{\varepsilon_o \varepsilon_{ox} L_g W_g / t_{ox}}{K K / K} = K K / K = K$

- $\tau = C_g V_{\text{dd}} / I_d$

- $f = 1 / \tau = 1 / K$

- $\alpha$: Scaling factor

- In the past, $\alpha>1$ for most cases

- $N \rightarrow \alpha / K^2 = 1 / K^2$, when $\alpha=1$

- $f N C V^2 / 2 \rightarrow K^{-1} (\alpha K^{-2}) K (K^1)^2 = \alpha = 1$, when $\alpha=1$
<table>
<thead>
<tr>
<th>k = 0.7 and ( \alpha = 1 )</th>
<th>k = 0.7^2 = 0.5 and ( \alpha = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single MOFET</strong></td>
<td></td>
</tr>
<tr>
<td>( V_{dd} \rightarrow 0.7 )</td>
<td>( V_{dd} \rightarrow 0.5 )</td>
</tr>
<tr>
<td>( L_g \rightarrow 0.7 )</td>
<td>( L_g \rightarrow 0.5 )</td>
</tr>
<tr>
<td>( I_d \rightarrow 0.7 )</td>
<td>( I_d \rightarrow 0.5 )</td>
</tr>
<tr>
<td>( C_g \rightarrow 0.7 )</td>
<td>( C_g \rightarrow 0.5 )</td>
</tr>
<tr>
<td>( P ) (Power)/Clock</td>
<td>( P ) (Power)/Clock</td>
</tr>
<tr>
<td>( \rightarrow 0.7^3 = 0.34 )</td>
<td>( \rightarrow 0.5^3 = 0.125 )</td>
</tr>
<tr>
<td>( \tau ) (Switching time)</td>
<td>( \tau ) (Switching time)</td>
</tr>
<tr>
<td>( \rightarrow 0.7 )</td>
<td>( \rightarrow 0.5 )</td>
</tr>
<tr>
<td><strong>Chip</strong></td>
<td></td>
</tr>
<tr>
<td>( N ) (# of Tr)</td>
<td>( N ) (# of Tr)</td>
</tr>
<tr>
<td>( \rightarrow 1/0.7^2 = 2 )</td>
<td>( \rightarrow 1/0.5^2 = 4 )</td>
</tr>
<tr>
<td>( f ) (Clock)</td>
<td>( f ) (Clock)</td>
</tr>
<tr>
<td>( \rightarrow 1/0.7 = 1.4 )</td>
<td>( \rightarrow 1/0.5 = 2 )</td>
</tr>
<tr>
<td>( P ) (Power)</td>
<td>( P ) (Power)</td>
</tr>
<tr>
<td>( \rightarrow 1 )</td>
<td>( \rightarrow 1 )</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

V_{dd} scaling insufficient
Additional significant increase in I_d, f, P

Change in 30 years

<table>
<thead>
<tr>
<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>L_g</td>
<td>K</td>
<td>10^{-2}</td>
<td>K(10^{-2})</td>
<td>10^{-2}</td>
<td></td>
</tr>
<tr>
<td>t_{ox}</td>
<td>K(10^{-2})</td>
<td>10^{-1}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V_{dd}</td>
<td>K(10^{-2})</td>
<td>10^{-1}</td>
<td>K(10^{-2})</td>
<td>10^{-1}</td>
<td></td>
</tr>
<tr>
<td>A_{chip}</td>
<td>\alpha</td>
<td>10^{1}</td>
<td>1</td>
<td>10^{1}</td>
<td>10^{4}</td>
</tr>
</tbody>
</table>

V_{dd} scaling insufficient, \alpha increased \rightarrow N, I_d, 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)
What is a roadmap? What is ITRS?

Roadmap: Prediction of future technologies
ITRS: International Technology Roadmap for Semiconductors made by SIA (Semiconductor Industry Association) with Japan, Europe, Korea and Taiwan
1992 - 1997: NTRS (National Technology Roadmap)
1998 - : ITRS (International Technology Roadmap)
ITRS Roadmap does change every year!

2006 Update   2002 Update
2004 Update   2000 Update

http://www.itrs.net/reports.html
Subthreshold Leakage ($A/\mu m$)

Operation Frequency (a.u.)

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.
**What does ‘22 nm’ mean in 22 nm CMOS Logic?**

**‘XX nm CMOS Technology**
Commercial Logic CMOS products

<table>
<thead>
<tr>
<th>Technology name</th>
<th>Starting Year</th>
<th>ITRS (Likely in 2008 Update)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Year</td>
</tr>
<tr>
<td>45 nm</td>
<td>2007</td>
<td>2007</td>
</tr>
<tr>
<td>32 nm</td>
<td>2009?</td>
<td>2008</td>
</tr>
<tr>
<td>22 nm</td>
<td>2011?~</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>2012?</td>
<td>2010</td>
</tr>
<tr>
<td>16 nm</td>
<td>2013?~</td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td>2014?</td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2014</td>
</tr>
</tbody>
</table>

- In general, there is no common corresponding parameter with ‘XX nm’ in ITRS table, which stands for ‘XX nm’ CMOS.

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

- ‘XX nm’ does not correspond to the ‘Half Pitch’ nor ‘Physical Gate Length’ of ITRS.

- ‘XX nm’ is now just a commercial name for CMOS Logic generation of size and its technology.

- Actual parameter values and starting years for commercial products are somewhat different from the above ITRS table, depending on semiconductor companies.

- In 22 and 16 nm technologies, physical gate lengths of high-performance logic device may be close to XX nm.
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 ??
What does ‘22 nm’ mean in 22 nm CMOS Logic?

Gate length of Logic CMOS became significantly smaller than lithography resolution or half-pitch using special technique such as resist aching (or trimming) method since 350 nm CMOS.

Source: ITRS 2001 Update
Some Problem: Number of most advanced logic CMOS companies is decreasing in generations.
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

Flash Poly Gate Half Pitch

FLASH Poly Silicon ½ Pitch
= Flash Poly Pitch/2

Poly Pitch

8-16 Lines
Typical flash Un-contacted Poly

Source: 2008 ITRS Summer Public Conf.
For example, Typical Half Pitches at ITRS 2007

2007 ITRS Product Technology Trends - Half-Pitch, Gate-Length

Source: 2008 ITRS Summer Public Conf.
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

Source: 2008/ ITRS Summer Public Conf.
Clock frequency does not increase aggressively anymore. 

Even decreased!

Advantage in SISC
Era for ‘out of order’

Advantage in RISC
Simple configuration

Multi Core
Clock
Performance

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

Cell Broadband Engine

6GHz capability for SRAM

Source: IBM, Toshiba, Sony ISSCC2008 and 08

Source: 2007 ITRS Winter Public Conf.
Clock frequency Change in the past ITRS (Max on chip frequency or ‘Core clock’)

- **26%** per year
- **8%** per year
- **15%** per year

**Design Max On-Chip Clock Frequency**

- Including 2005 ITRS and Final (Aug'07)

- **22 nm:** 6 GHz?

**New Design TWG**
- 2007 ITRS Final “IS”
- Ave 8% CAGR

**Past <—— Future**

- **2007 Des TWG**
- Actual History of Average On-Chip
- ~ 21% CAGR

Source: 2008 ITRS Summer Public Conf.
Structure and technology innovation (ITRS 2007)

- **poly**
  - SiON

- **metal**
  - high k

- **gate stack**

- **bulk**

- **planar**

- **PDSOI**

- **FDSOI**

- **3D**
  - MuGFET
  - MuCFET

- **stressors**

- **+ substrate engineering**

- **+ high μ materials**

- **65nm**

- **45nm**

- **32nm**

- **22nm**

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**

- **2007**
  - HP bulk CMOS (ITRS 2007)
  - LOP bulk CMOS (ITRS 2007)

- **2008**
  - HP bulk CMOS (ITRS 2008)
  - LOP bulk CMOS (ITRS 2008)

- **2009**
  - HP UTB/FDSOI (ITRS 2007)

- **2010**
  - HP UTB/FDSOI (ITRS 2008)

- **2011**
  - HP Multi-Gate (ITRS 2007)

- **2012**
  - HP Multi-Gate (ITRS 2008)

- **2013**
  - Multi G delays 4 years!

- **2014**
  - HP Multi-Gate (ITRS 2008)

- **2015**
  - HP Multi-Gate (ITRS 2008)

- **2016**
  - HP Multi-Gate (ITRS 2008)

- **2017**
  - HP Multi-Gate (ITRS 2008)

- **2018**
  - HP Multi-Gate (ITRS 2008)

- **2019**
  - HP Multi-Gate (ITRS 2008)

**Source:** 2008 ITRS Summer Public Conf.
Wafer size (ITRS 2007)

<table>
<thead>
<tr>
<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 Pitch (nm)/contacted</td>
<td>68</td>
<td>59</td>
<td>52</td>
<td>45</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>MPU Physical Gate Length (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>
</tr>
<tr>
<td>MPU High-Performance Total Chip Area (mm²)</td>
<td>310</td>
<td>246</td>
<td>195</td>
<td>310</td>
<td>246</td>
<td>195</td>
<td>310</td>
<td>246</td>
<td>195</td>
</tr>
<tr>
<td>MPU High-Performance Active Transistor Area (mm²)</td>
<td>31.7</td>
<td>25.1</td>
<td>20.0</td>
<td>31.7</td>
<td>25.1</td>
<td>20.0</td>
<td>31.7</td>
<td>25.1</td>
<td>20.0</td>
</tr>
<tr>
<td>General Characteristics * (99% Chip Yield)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Substrate Diameter (mm)—High-volume Production (&gt;20K wafer starts per month)**</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>450</td>
<td>450</td>
<td>450</td>
<td>450</td>
</tr>
</tbody>
</table>

Source: ITRS 2007

Correspond to 22nm

Maybe delay??
Gate CD (Critical Dimension) Control

**ITRS 2007**

<table>
<thead>
<tr>
<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 Pitch (nm)(contacted)</td>
<td>68</td>
<td>59</td>
<td>52</td>
<td>45</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>MPU Physical Gate Length (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>
</tr>
<tr>
<td>Total maximum allowable etch 3σ (nm), including photoresist trim and gate etch [AA1]</td>
<td>1.5</td>
<td>1.38</td>
<td>1.2</td>
<td>1.08</td>
<td>0.96</td>
<td>0.84</td>
<td>0.78</td>
<td>0.66</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Source: ITRS 2007

**2008 Update**

<table>
<thead>
<tr>
<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 Physical Gate Length (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>
</tr>
<tr>
<td>$L_{gate} \ 3\sigma \ variation \ (nm) \ [\AA]</td>
<td>5.84</td>
<td>3.49</td>
<td>3.16</td>
<td>2.9</td>
<td>2.66</td>
<td>2.42</td>
<td>2.21</td>
<td>2.02</td>
<td>1.84</td>
</tr>
</tbody>
</table>

Source: 2008  ITRS Summer Public Conf.

Gate CD control color changed to ‘white’ through 2011 and to ‘yellow’ for 2012 reflecting the new Lg scaling.
## ITRS2008 Low-k Roadmap Update

Correspond to 22nm Logic

<table>
<thead>
<tr>
<th></th>
<th>Year of Production</th>
<th>Near-term</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2008</td>
<td>2009</td>
<td>2010</td>
<td>2011</td>
<td>2012</td>
</tr>
<tr>
<td><strong>ITRS 2007</strong></td>
<td>Interlevel metal insulator – effective dielectric constant (k)</td>
<td>2.7-3.0</td>
<td>2.5-2.8</td>
<td>2.5-2.8</td>
<td>2.5-2.8</td>
<td>2.1-2.4</td>
</tr>
<tr>
<td><strong>Update 2008</strong></td>
<td>Interlevel metal insulator – effective dielectric constant (k)</td>
<td>2.9-3.3</td>
<td>2.6-2.9</td>
<td>2.6-2.9</td>
<td>2.6-2.9</td>
<td>2.4-2.8</td>
</tr>
<tr>
<td><strong>ITRS 2007</strong></td>
<td>Interlevel metal insulator – bulk dielectric constant (k)</td>
<td>2.3-2.7</td>
<td>2.1-2.4</td>
<td>2.1-2.4</td>
<td>2.1-2.4</td>
<td>1.8-2.1</td>
</tr>
<tr>
<td><strong>Update 2007</strong></td>
<td>Interlevel metal insulator – bulk dielectric constant (k)</td>
<td>2.5-2.8</td>
<td>2.3-2.6</td>
<td>2.3-2.6</td>
<td>2.3-2.6</td>
<td>2.4-2.4</td>
</tr>
</tbody>
</table>

Source: 2008 ITRS Summer Public Conf.

$k$ value increases by $0.1 \sim 0.3$
Historical Transition of ITRS Low-k Roadmap

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

$Vg=0V$

$Vth$ (Threshold Voltage)
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)
= \((\ln 10)(kT/q)(C_{ox}+C_D+C_{it})/C_{ox}\)
> ~ 60 mV/decade at RT

SS value:
Constant and does not become small with down-scaling
S-D leakage current

I\(_{sd}\)-leak has to be stay less than 1\( \mu \text{A/\(\mu\text{m}\)} \)

Drain current

Id-sat growth will be modest in 2008 update

Source: ITRS and 2008 ITRS Summer Public Conf.

2008 Values are from ITRS Public Conf. and still under discussion
ITRS for HP logic

Ion/loff ratio

Source: ITRS and 2008 ITRS Summer Public Conf.

2008 Values are from ITRS Public Conf. and still under discussion
ITRS for HP logic

Vdd will stay higher in 2008 update

Vth-sat will be around 0.1V

Source: ITRS and 2008 ITRS Summer Public Conf.

2008 Values are from ITRS Public Conf. and still under discussion
ITRS for HP logic

2008 Values are from ITRS Public Conf. and still under discussion

Vth-sat / Vdd

Source: ITRS and 2008 ITRS Summer Public Conf.
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.
   → $C_D$ increase
   → 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!
Could we squeeze technologies for ultimate CMOS scaling?

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

**Improper down-scaling**

- **Interfacial C** @Metal gate and Gate oxd. (EOT=0.2~0.3nm?)
- **Inversion C** (EOT=0.3~0.5nm?)

**Is 0.5nm real limit?**

**Delay**

**Saturation**

EOT(C₁) + EOT(C₃) > 0.5nm
Small effect to decrease EOT(C₂) beyond 0.5nm?
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=2.5/50µm
PMA 300°C (30min)

EOT scaling below 0.5nm
Still useful for larger drain current


$^*$ Because Lg is very large (2.5µm), 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 Id 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.

$^{**}$ Estimated by Id value
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.
SS (Subthreshold Slope) becomes worse in the following cases

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

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

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!
Enhanced D-Electric-field

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}

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

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

\[ V_d=1V \]

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

Source: ECS Fall Meeting, Oct 2008, Honolulu,
Y. Kobayashi, A. B. Sachid, K. Tsutsui, K. Kakushima,

Comparison of Bulk and DG

DIBL at drain edge

Wfin = 10.7 nm
Wfin = 30 nm
Wfin = 40 nm

Sub-threshold Swing (mV/dec)

Fin Width (nm)

DIBL@D Edge (mV/V)

Electric Potential

Linear (V)

0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

\[ \Lambda = 7.6\text{nm} \]
\[ \Lambda = 17.1\text{nm} \]
\[ \Lambda = 13.2\text{nm} \]
Enhanced D-Electric-field

Comparison of High-k and SiO\textsubscript{2} MOSFETs

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 $\rightarrow$ UTB-SOI $\rightarrow$ DG $\rightarrow$ 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.
Random Variability Reduction Scenario in ITRS 2007

Assumption: Random dopant fluctuation is Main source of Random Variability. Line width roughness of Lg and Wg is not considered in this study.

Source: 2007 ITRS Winter Public Conf.
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

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

<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

<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 = 0.1 µm

Announced on Aug 18, 2008

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

0.1 µm 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?

Cell area ($\mu m^2$)

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

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

- IBM Alliance (Consortium)
  - 32nm Dec. 2007

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

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

**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

Cell evolution is similar

TSMC 45nm

TSMC 32nm

IBM Alliance 32nm

IBM Alliance 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)

- 45 nm high-k6 cores
- 16MB shared L3 cache

Source: Intel Developer Forum 2008

Cache occupies huge area

→ Cell size of SRAM should be minimized
→ Isd-leak should be minimized
  → Vth are often designed to be higher than Min. logic Vth
  → Lg are often designed to be larger than Min. logic Lg
Future Directions For Improving Vmin

- **Application**
  - Improvement in voltage and temperature tolerance
- **Package**
  - Separated array / logic voltage to minimize logic noise effect on SRAM
- **Design**
  - Higher array VDD and improved on-chip supply robustness
  - Increased redundancy
  - Improved timings
  - Cells per BL hierarchical BL structure
  - Write/Read assist and sense-amp design
- **Cell and Process**
  - Improved bit cell optimization
- **NFET/PFET centering and Beta/Gamma control**
- **Minimize device fluctuation by limiting device-geometry scaling larger cell**
- **Lpoly, Weff, LER**
  - Leakage / defect mechanisms

*Source: Harold Pilo IEDM2006 Short Course*
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
6T and 8T Cell

6T Cell

- Cell size is small
- For high density use

8T Cell

- Add separate read function
- Cell size increase 30%
- For low voltage use

Source: Morita et. al, Symp. on VLSI Circ. 2007
5. Roadmap for further future as a 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 (H. Iwai, IWJT2008).

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

- Other Beyond CMOS devices are still in the cloud.

Source: 2008 ITRS Summer Public Conf.

*5.5nm? was added by Iwai
Si nanowire FET with Semi-1D Ballistic Transport

Merit of Si-nanowire

- Reduction in Ioff (I_{sd-leak})
  - Good control of I_{sd-leak} by surrounding gate

- Increase in Ion (I_{d-sat})

Trade off

Carrier scattering probability
- Small vs. Large
  - # of quantum channel
    - Small vs. Large

High Conduction (1D)
- Go=77.8\mu S/wire

Multiple quantum channel (QC) used for conduction

High-density lateral and vertical integration


Our roadmap for R & D
Source: H. Iwai, IWJT 2008

Current Issues

Si Nanowire
- Control of wire surface property
- Source Drain contact
- Optimization of wire diameter
- Compact I-V model

III-V & Ge Nanowire
- High-k gate insulator
- Wire formation technique

CNT:
- Growth and integration of CNT
- Width and Chirality control
- Chirality determines conduction types: metal or semiconductor

Graphene:
- Graphene formation technique
- Suppression of off-current
  - Very small bandgap or no bandgap (semi-metal)
- Control of ribbon edge structure which affects bandgap
Long term roadmap for development

Source: H. Iwai, IPFA 2006

Miniaturization of Interconnections on PCB (Printed Circuit Board)

Some time in 2020 - 2030

5 nm?

Scaling proceeds

We do not know how?

Long term roadmap for development

We do know system and algorithms are important!

But do not know how it can be by us for use of bio?

After 2050?

Dragonfly brain has even further higher performance

Sensor
Infrared
Humidity
CO₂

Bite
Ultrimill name
Small number of neuron cells
Extremely low power
Real time image processing
(Artificial) Intelligence
3D flight control

Mosquito
Acknowledgement

I would like to express deep appreciation to the following people for the useful advice and support for material preparation. Special thanks to ITRS committee for the permission to refer roadmap and Public conference.

**ITRS Committee:**  Hidemi Ishiuchi (Toshiba), Paolo Gargini (Intel)

**Toshiba Corporation:**  Mitsuo Saito, Yukihiro Urakawa, Tomoaki Yabe

**Tsukuba University:**  Kenji Shiraishi, Kenji Natori

**Intel Corporation:**  Mark Bohr

**IBM Alliance:**  B.S. Haran et al,

**Tokyo Institute of Technology:**  Kuniyuki Kaukshima, Parhat Ahmet, Takamasa Kawanago, Yeonghun Lee
Thank you for your attention!