Dependence on Feature Size

Gary Drake
Argonne National Laboratory

University of Chicago Workshop
Apr. 29, 2011
Outline

- Basic Notation
- Survey of Feature Sizes
- How Feature Size Affects Speed – Scaling Laws
- Consequences of Deep Submicron Technologies
- Practical Considerations
- Summary
Basic Notations & Principles in CMOS Design

- **Geometry**
  - \( W \) = Gate Width
  - \( L \) = Gate Length
  - \( t_{ox} \) = thickness of gate oxide
  - \( W/L \) – defines transistor size
  - \( L_{min} \) = min. feature size

- **Semiconductor Construction**
  - **N-type** – mobile charge carriers are electrons (negative charge)
  - **P-type** – mobile charge carriers are holes (positive charge)
  - **Doping** – add impurities to pure silicon to make material N-type or P-type
    - \( N_a \) – density of acceptor atoms - holes
    - \( N_d \) – density of donor atoms - electrons
  - **NMOS or N-channel** - electrons
    - Conductive channel is N-type
  - **PMOS or P-channel** - holes
    - Conductive channel is P-type
  - **Source & Drain** – doping implants into substrate to create \( n^+ \) & \( p^+ \) regions
  - **Oxide Layer** - Gate is isolated from substrate, which makes it high impedance, except for leakage

- **Operation**
  - Gate-Source voltage \( V_{GS} \) creates conduction channel under gate, allowing current to flow between drain & source
  - Minimum gate voltage to create channel called the Threshold Voltage \( V_{TH} \)
Survey of Feature Sizes or Size **DOES** Matter!

- **Big is Good!**
- **Big is Good?**
- **Big is Good!**
- **Small is Good. But not too small...**
- **Big is Good?**
Survey of Feature Sizes
IBM & TSMC CMOS Processes Offered Through MOSIS Today

Data from MOSIS
The Basic Question: How to Achieve Faster (Higher BW) ASICs?

- The Basic Answer:
  - General trend: 
    *The smaller the feature size, the faster the chip can operate.*
  - Why? A few high-level reasons:
    - Smaller size = **shorter distance** that signals need to propagate
    - Smaller size = generally **lower parasitic capacitance**
      (But watch out for larger capacitance/unit area…)
    - Smaller size \(\rightarrow\) use **lower voltage rails** \(\rightarrow\) reach logic levels faster

⇒ *More technical answers later…*
The Basic Question:
How to Achieve Faster (Higher BW) ASICs? (Cont.)

- **Measurements Tell the Tale:**
  - One measure of speed in CMOS: Ring Oscillator Frequency
    - Basic test structure used by MOSIS to measure fabrication run acceptance
    - Generally configured as an odd number of inverters, minimum-sized devices
      \[ F_0 = \frac{1}{2N \cdot \text{Inverter}_\text{delay}}, \text{N = # Inverters (MOSIS uses 31 stages.)} \]

![Ring Oscillator Diagram](Courtesy Wikipedia)

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 nm</td>
<td></td>
</tr>
<tr>
<td>28 nm</td>
<td></td>
</tr>
<tr>
<td>32 nm</td>
<td>32SOI</td>
</tr>
<tr>
<td>45 nm</td>
<td>12SOI</td>
</tr>
<tr>
<td>65 nm</td>
<td>10SF</td>
</tr>
<tr>
<td>65 nm</td>
<td>10LPe/RF</td>
</tr>
<tr>
<td>90 nm</td>
<td>9SF</td>
</tr>
<tr>
<td>90 nm</td>
<td>9LP/RF</td>
</tr>
<tr>
<td>0.130 (\mu\text{m})</td>
<td>8RF-LM</td>
</tr>
<tr>
<td>0.180 (\mu\text{m})</td>
<td>7SF</td>
</tr>
<tr>
<td>0.250 (\mu\text{m})</td>
<td>7RF</td>
</tr>
<tr>
<td>0.350 (\mu\text{m})</td>
<td>7RFSOI</td>
</tr>
<tr>
<td>0.500 (\mu\text{m})</td>
<td>7HV</td>
</tr>
</tbody>
</table>

Data from MOSIS
How Feature Size Affects Speed

- Robert Dennard’s Scaling Law (1974)
  - If scale the physical parameters of an integrated circuit equally by factor $K$, then performance parameters scale as follows:

<table>
<thead>
<tr>
<th>Geometry &amp; Supply voltage</th>
<th>$L_g$, $W_g$, $T_{ox}$, $V_{dd}$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scaling</strong> $K$: $K=0.7$ for example</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drive current in saturation</th>
<th>$I_d$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_d$ per unit $W_g$</td>
<td>$I_d/\mu m$</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gate capacitance</th>
<th>$C_g$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_g = \varepsilon_0 \varepsilon_{ox} L_g W_g / t_{ox}$</td>
<td>$KK/K = K$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Switching speed</th>
<th>$\tau$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = C_g V_{dd} / I_d$</td>
<td>$KK/K = K$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clock frequency</th>
<th>$f$</th>
<th>$1/K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f = 1/\tau = 1/K$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chip area</th>
<th>$A_{chip}$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$: Scaling factor</td>
<td>In the past, $\alpha&gt;1$ for most cases</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Integration (# of Tr)</th>
<th>$N$</th>
<th>$\alpha/K^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N \rightarrow \alpha/K^2 = 1/K^2$, when $\alpha=1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power per chip</th>
<th>$P$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P \rightarrow $ fNCV$^2$/2</td>
<td>$K^{-1}(\alpha K^2)K (K^4) = \alpha = 1$, when $\alpha=1$</td>
<td></td>
</tr>
</tbody>
</table>

Table Courtesy of Mark Bohr, IBM

Slide Courtesy of Hirosi Iwai, Tokyo Institute of Technology
How Feature Size Affects Speed (Cont.)

- **Dennard’s Scaling Law & Moore’s Law**
  - **Gordon Moore’s Law**: “The number of transistors that can placed inexpensively on an integrated circuit doubles approximately every two years.”
    → Basis for Industry Roadmap (ITRS)
  - **Dennard**: If scale physical dimensions by K, then area of chip $\sim K^2$
    - For $K \sim 0.7 \sim 1/\sqrt{2}$, $K^2 \sim \frac{1}{2}$ → $X2$ more transistors for a given die size
  - **Industry**: Ratios between successive technology steps:
    $350/500 = 0.7$; $250/350 = 0.71$; $90/130 = 0.69$; $65/90 = 0.722$

Moore’s Law will come to an end in ~2014, when the feature size approaches a few atomic layers.

Plots Courtesy of Mark Bohr, IBM
How Feature Size Affects Speed (Cont.)

- Examining Speed versus Technology
  - A technical answer from semiconductor physics  
    (Courtesy of Willy Sansen):

\[
f_T = \frac{\mu}{2\pi L^2} \left( \frac{V_{GS}-V_T}{0.2 \ldots 1 V} \right)
\]

(f\(_T\) = Freq. where current gain = 1)

Critical Parameters:
- Mobility
- Channel Length
- Thresh. Voltage
- \(V_{GS}-V_T\)

⇒ These are functions of the technology…
How Feature Size Affects Speed (Cont.)

- Examining Speed versus Technology (Cont.)
  - Unity Gain Frequency:
    \[ f_T = \frac{\mu}{2\pi L^2} (V_{GS} - V_T) \]
  - Length L
    \[ \Rightarrow \text{Decreases as feature size decreases, } L \propto K \text{ scaling, for } K < 1 \]
    \[ \Rightarrow \text{Causes } f_T \text{ to increase as } 1/K^2 \text{ for } K < 1 \]
  - Carrier Mobility
    \[ \mu = \mu_o + \frac{\mu_1}{1 + \left(\frac{N}{N_{ref}}\right)^\alpha} \]
    \[ \mu_n(N_D) = 65 + \frac{1265}{1 + \left(\frac{N_D}{8.5 \times 10^{16}}\right)^{0.72}} \]
    \[ \mu_p(N_A) = 48 + \frac{447}{1 + \left(\frac{N_A}{6.3 \times 10^{16}}\right)^{0.76}} \]
    \[ \Rightarrow N_D \propto 1/K \Rightarrow \text{Increases for } K < 1 \]
    \[ \Rightarrow f_T \text{ decreases for } K < 1, \text{ but slowly...} \]
  - \( V_T \) & \( (V_{GS} - V_T) \)?
    ♦ Want \( (V_{GS} - V_T) \) as large as possible
    ♦ Want \( V_T \) as small as possible
How Feature Size Affects Speed (Cont.)

- Examining Speed versus Technology (Cont.)
  - Generally, threshold voltage $V_{TH}$ gets smaller with feature size
    - Partly a consequence of decreasing $V_{dd}$
    - Partly a consequence of increasing $C_{OX}$ (aF/$\mu$m² gate area)
  - But, are approaching a limit with submicron technologies… Why?

$$V_{th} = V_{fb} + 2\psi_B + \sqrt{2\varepsilon_{Si}qN_a (2\psi_B + V_{bs})}$$

where
- $V_{fb}$ is the flatband voltage,
- $\psi_T = kT/q$
- $C_{OX}$ is the gate oxide capacitance/area
- $\psi_B$ is the zero bias mobility
- $V_{bs}$ is the max. depletion layer width
- $N_a$ is the acceptor doping density

Data from MOSIS (Line Not a Fit...
How Feature Size Affects Speed (Cont.)

- Examining Speed versus Technology (Cont.)
  - $V_{dd}$ scales with $K$, so $V_{dd}$ decreases with smaller feature size
    - 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 Id) and cannot be too small.

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

$> \Delta V_{th}$ Margin for $V_{th}$ variation is necessary

Slide Courtesy of Hiroshi Iwai, Tokyo Institute of Technology
How Feature Size Affects Speed (Cont.)

- Examining Speed versus Technology (Cont.)
  - Bottom Line - Unity Gain Frequency of FETs:

\[ f_T = \frac{\mu}{2\pi L^2} (V_{GS} - V_T) \]

- Decreases as \((K)^{0.7}\)
- Scales with \(K\)
- Approaching a constant

\(\Rightarrow\) So, \(f_T \propto 1/K\) (approximate)

\(\Rightarrow\) i.e., \(f_T\) increases as feature size decreases
Consequences of Deep Submicron Technologies

- Scaling Laws worked well until ~ 100 nm and smaller
  - Performance generally well-predicted > 100 nm
  - Below 100 nm, start to see second-order and higher-order effects that affect performance
    - Especially for analog IC or mixed-signal ICs

- Direct Performance Effects
  - *Increased leakage currents*
    - Increased power consumption
  - Transistor size mismatch

- Other Performance Factors → Especially for Analog
  - Reduced voltage rails
  - Decreased signal-to-noise
  - Decreased dynamic range
  - Increased power density

Graphic Courtesy of Kaushik Roy, Purdue University
Consequences of Deep Submicron Technologies (Continued)

- **Leakage Currents**
  - Gate oxide tunneling leakage ($I_G$)
    - Thin gate oxides allow electron tunneling from gate to substrate
    - **Major source of leakage current in sub-micron CMOS**
    - The thinner the oxide, the worse the leakage current, i.e. the smaller the feature size, the worse the leakage current, since $t_{ox}$ scales
Consequences of Deep Submicron Technologies
(Continued)

- **Leakage Currents (Cont.)**
  - **Subthreshold leakage ($I_{SUB}$)**
    - Drain-source current during weak inversion, $V_{GS} < V_{TH}$
    - Dominated by diffusion current of minority carriers, rather than drift current that dominates in strong inversion
    - Carriers move along surface
    - Very sensitive to process parameters, device size, supply voltage, and temperature

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L} (m - 1)(v_T)^2 \times e^{(V_g - V_{th})/m v_T} \times \left(1 - e^{-v_{ds}/v_T}\right)$$

where

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{e_{ox}}{\frac{e_{ox}}{W_{dm}}} = 1 + \frac{3\mu_{ox}}{W_{dm}}$$

where

- $V_{th}$ is the threshold voltage,
- $v_T = kT/q$
- $C_{ox}$ is the gate oxide capacitance
- $\mu_{ox}$ is the zero bias mobility
- $W_{dm}$ is the max. depletion layer width

- **Subthreshold Leakage**
  - $I_{OFF}$ decreases as feature size decreases
  - $I_{OFF} = 390 \text{ pA} = 20 \text{ pA/um} @ (V_{gs}=2.5\text{V})$
  - $80.7 \text{ pA} = 4 \text{ pA/um} @ (V_{gs}=0.1\text{V})$

Graphics Courtesy of Kaushik Roy, Purdue University
Consequences of Deep Submicron Technologies (Continued)

- **Leakage Currents (Cont.)**
  - PN Junction Reverse-Bias Leakage ($I_{\text{REV}}$)
    - 2 mechanisms
      - Minority-carrier drift/diffusion
      - Electron-hole generation
    - Occurs when P & N junctions are heavily doped
      - Doping ($N_a$ & $N_d$) increases as feature size decreases
        → Leakage increases as feature size decreases
      - Causes band-to-band tunneling (BTBT) due to electric field from the doping

\[
E = \sqrt{\frac{2qN_aN_d(V_{\text{app}} + V_{bi})}{\varepsilon_{\text{Si}}(N_a + N_d)}}
\]
Leakage Currents (Cont.)

- Gate Induced Drain Leakage ($I_{GIDL}$)
  - Caused by narrowing of depletion region around drain when FET is off
    - Due to electric field between gate and drain
  - Get tunneling of minority carriers from drain to substrate
  - Strong function of
    - Doping profile at drain edge
    - Oxide thickness
    - $V_{dd}$
Consequences of Deep Submicron Technologies (Continued)

- **Matching Problems**
  - Generally, device matching becomes worse as feature size decreases
  - Limitations of lithography & processing

Threshold voltage mismatch $A_{VT}$

- $A_{VT} \sim t_{ox}^{4} \sqrt{N_B}$
- Plot Courtesy of Hiroshi Iwai, Tokyo Institute of Technology
- Plot Courtesy of Willy Sansen

- **Slope is gentle ...**
- **Many designs may not be affected ...**
Other Consequences of Deep Submicron Technologies

- **Reduced Headroom**
  - Circuits with more than ~2-3 transistors between $V_{dd}$ and $V_{ss}$ run out of head room due to reduced $V_{dd}$ (scales with $K$…)

- **Reduced Signal-to-Noise**
  - Reduced voltage rails mean smaller signals
  - Noise may not improve with smaller feature size
    - Depends on design…

- **Reduced Dynamic Range**
  - Also a consequence of reduced voltage rails and higher noise…

- **Increased Power Density**
  - Consequence of leakage currents

⇒ *For digital, these are mostly OK*
⇒ *For analog, can be a problem…*
Practical Considerations

- **Availability (through MOSIS)**
  - Currently, technologies between 40-100 nm only offered by TSMC
    - Even then, models sparse or not available
  - IBM: Only trusted vendors below 130 nm

![IBM & TSMC CMOS Processes - Runs per Year by Technology](chart)

<table>
<thead>
<tr>
<th>Fab House</th>
<th>Feature Size</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>22 nm</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>28 nm</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>32 nm</td>
<td>32 SOI</td>
</tr>
<tr>
<td>TSMC</td>
<td>40 nm</td>
<td>CLN40/C MN40</td>
</tr>
<tr>
<td>IBM</td>
<td>45 nm</td>
<td>12 SOI</td>
</tr>
<tr>
<td>TSMC</td>
<td>45 nm</td>
<td>CLN45/C MN45</td>
</tr>
<tr>
<td>IBM</td>
<td>65 nm</td>
<td>10 SF</td>
</tr>
<tr>
<td>IBM</td>
<td>65 nm</td>
<td>18 LP/RF</td>
</tr>
<tr>
<td>TSMC</td>
<td>65 nm</td>
<td>CLN65/C MN65</td>
</tr>
<tr>
<td>TSMC</td>
<td>65 nm</td>
<td>CMN65T</td>
</tr>
<tr>
<td>IBM</td>
<td>90 nm</td>
<td>9 SF</td>
</tr>
<tr>
<td>IBM</td>
<td>90 nm</td>
<td>18 SF/RF</td>
</tr>
<tr>
<td>TSMC</td>
<td>90 nm</td>
<td>CLN90/C MN90</td>
</tr>
<tr>
<td>TSMC</td>
<td>90 nm</td>
<td>CMN90</td>
</tr>
<tr>
<td>IBM</td>
<td>0.130 μm</td>
<td>8RF-LM</td>
</tr>
<tr>
<td>IBM</td>
<td>0.130 μm</td>
<td>8RF-DM</td>
</tr>
<tr>
<td>TSMC</td>
<td>0.130 μm</td>
<td>CL013/C MN013</td>
</tr>
<tr>
<td>TSMC</td>
<td>0.130 μm</td>
<td>CL013LVP</td>
</tr>
<tr>
<td>TSMC</td>
<td>0.130 μm</td>
<td>CL013LV</td>
</tr>
<tr>
<td>IBM</td>
<td>0.180 μm</td>
<td>7 SF</td>
</tr>
<tr>
<td>IBM</td>
<td>0.180 μm</td>
<td>7 RF</td>
</tr>
<tr>
<td>IBM</td>
<td>0.180 μm</td>
<td>17RF/FSO I</td>
</tr>
<tr>
<td>IBM</td>
<td>0.180 μm</td>
<td>7HV</td>
</tr>
<tr>
<td>IBM</td>
<td>0.260 μm</td>
<td>6 RF</td>
</tr>
<tr>
<td>IBM</td>
<td>0.350 μm</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>0.500 μm</td>
<td></td>
</tr>
</tbody>
</table>
Practical Considerations

- **Cost**
  - (Costs unknown for every technology at press time…)
  - Generally, leading-edge technologies very expensive, even through MOSIS

- **Models**
  - Models are poor – to – non-existent for leading-edge technologies
  - Customer support is poor for leading-edge technologies
  - MOSIS data not available for leading-edge technologies

- **General Difficulty**
  - Expect long learning curve with the more advanced technologies
Summary

- For high-speed and high-bandwidth, generally want technologies with smaller feature size
  - Smaller distances
  - Smaller parasitic capacitances
  - Smaller voltage swings
  - Intrinsically faster devices (from scaling law principles)

- Below ~ 100 nm, start to have higher-order problems
  - Leakage currents of various kinds can affect designs
    - Higher internal power
    - May significantly affect analog circuits, in particular sample & hold circuits
  - Matching of transistors is worse
  - Signal headroom is reduced
  - Dynamic range is reduced

- Prospects for near future
  - Scaling laws have worked beautifully for 40 years, but we’re reaching a hard limit
  - Industry has kept pace, but will have difficulty below 10 nm
  - Leading-edge technologies difficult to use right now, but this will improve
  - Not clear if higher-order process problems can be solved for analog use…
  - Perhaps clever circuit design can be employed, but this will take time…

 Refuge: 130 nm technology is not a bad place to be right now for analog work…

