Analog Design in ULSI CMOS Processes

Giovanni Anelli
CERN - European Organization for Nuclear Research
Physics Department
Microelectronics Group
CH-1211 Geneva 23 – Switzerland
Giovanni.Anelli@cern.ch
Outline

• Motivation
• How scaling works for devices and interconnections
• Scaling impact on the transistor performance
• Scaling impact on analog circuits performance
• Noise in mixed-mode integrated circuits
• ULSI processes: which options for analog?
• Conclusions
Motivation

• The microelectronics industry is moving to ULSI CMOS processes, and we have interest to follow the trend because of:
  - Technology availability issues
  - Clear advantages for digital designs
  - Improved radiation tolerance

• The performance of detector electronics for future High Energy Physics experiments will still be strictly related to the analog front-end

What are the advantages and disadvantages of using a process in the 180 – 100 nm range for analog design? What do we gain? And what do we loose? And are there new problems and phenomena which have to be considered?
Outline

• Motivation
• How scaling works for devices and interconnections
  ➢ Why scaling?
  ➢ Transistor scaling
  ➢ Interconnection scaling
• Scaling impact on the transistor performance
• Scaling impact on analog circuits performance
• Noise in mixed-mode integrated circuits
• ULSI processes: which options for analog?
• Conclusions
Why scaling?

Example: CMOS inverter

\[
\begin{align*}
\text{P}_{\text{static}} &= I_{\text{leakage}} \cdot V_{DD} \\
\text{P}_{\text{dynamic}} &= C_L \cdot V_{DD}^2 \cdot f \\
\text{PDP} &= C_L \cdot V_{DD}^2
\end{align*}
\]

Power-delay product

Scaling improves density, speed and power consumption of digital circuits

LECC 2004

Giovanni Anelli, CERN
• more than 200,000 gates per mm²
• speed > 1 GHz
• power gate dissipation < 4 nW / MHz @ 1.2 V
• 8 metal levels, all copper, low K (FSG or BlackDiamond™)
• pitches: M1 0.34 µm, M2 to M7 0.41 µm, M8 0.9 µm
• embedded memory (single transistor, SRAM, Non-volatile)

VERY GOOD FOR System-on-Chip
Constant field scaling

The aim of constant field scaling is to reduce the device dimensions (to improve the circuit performance) without introducing effects which could disturb the good operation of the device.

Summary of the scaling factors for several quantities

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Scaling Factor</th>
<th>Quantity</th>
<th>Scaling factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimensions (L, W, t_{ox}, x_D)</td>
<td>1/\alpha</td>
<td>Capacitances</td>
<td>1/\alpha</td>
</tr>
<tr>
<td>Area</td>
<td>1/\alpha^2</td>
<td>Capacitances per unit area</td>
<td>\alpha</td>
</tr>
<tr>
<td>Devices per unit of chip area (density)</td>
<td>\alpha^2</td>
<td>Charges</td>
<td>1/\alpha^2</td>
</tr>
<tr>
<td>Doping concentration (N_A)</td>
<td>\alpha</td>
<td>Charges per unit area</td>
<td>1</td>
</tr>
<tr>
<td>Bias voltages and V_T</td>
<td>1/\alpha</td>
<td>Electric field intensity</td>
<td>1</td>
</tr>
<tr>
<td>Bias currents</td>
<td>1/\alpha</td>
<td>Body effect coefficient (\gamma)</td>
<td>1/\sqrt{\alpha}</td>
</tr>
<tr>
<td>Power dissipation for a given circuit</td>
<td>1/\alpha^2</td>
<td>Transistor transit time (\tau)</td>
<td>1/\alpha</td>
</tr>
<tr>
<td>Power dissipation per unit of chip area</td>
<td>1</td>
<td>Transistor power-delay product</td>
<td>1/\alpha^3</td>
</tr>
</tbody>
</table>

\[ \alpha > 1 \]

\[ C_{ox} = \frac{\varepsilon_{\text{SiO}_2}}{t_{ox}} \]
Constant field scaling problem

Subthreshold slope and width of the moderate inversion region do not scale. This can have a devastating impact on the static power consumption of a digital circuit.
Generalized scaling

- The dimensions in the device scale as in the constant field scaling
- \( V_{dd} \) scales to have reasonable electric fields in the device, but slower than \( t_{ox} \), to have a useful voltage swing for the signals
- The doping levels are adjusted to have the correct depletion region widths
- To limit the subthreshold currents, \( V_T \) scales more slowly than \( V_{dd} \)

An accurate scaling of the interconnections is needed as well, so that we can profit at the circuit level of the improvements made at the device level. Interconnections are becoming more and more important in modern technologies because the delay they introduce is becoming comparable with the switching time of the digital circuits.


"Reverse" scaling

The scaling method is different from the one applied to devices

If $W$, $L$, $t_m$ and $t_{ox}$ are decreased by $\alpha$

- Current density increases by $\alpha$
- $R$ increases by $\alpha$, $C$ decreases by $\alpha$
- $RC$ (delay) does not scale!!!

In practice, wires dimensions are reduced only for local interconnections (but not $t_m$). At the chip scale, $t_m$ and $t_{ox}$ are increased (reverse scaling).

Hierarchical scaling

Figure 36 Cross-section of Hierarchical Scaling


LECC 2004

Giovanni Anelli, CERN
Outline

• Motivation
• How scaling works for devices and interconnections
  • Scaling impact on the transistor performance
    ➢ Weak inversion, strong inversion, velocity saturation
    ➢ Transistor intrinsic gain
    ➢ Gate leakage and noise
• Scaling impact on analog circuits performance
• Noise in mixed-mode integrated circuits
• ULSI processes: which options for analog?
• Conclusions
Weak inversion (w.i.)

- \( I_{DS\_w.i.} = I_{D0} \frac{W}{L} e^{\frac{V_{GS}}{n\phi_t}} \)
- \( g_{m\_w.i.} = \frac{I_{DS}}{n\phi_t} \)

Strong inversion (s.i.)

- \( I_{DS\_s.i.} = \frac{\beta}{2n} (V_{GS} - V_T)^2 \)
- \( g_{m\_s.i.} = \sqrt{\frac{2\beta}{n}} I_{DS} \)

Velocity saturation (v.s.)

- \( I_{DS\_v.s.} = W C_{ox} v_{sat} (V_{GS} - V_T) \)
- \( g_{m\_v.s.} = W C_{ox} v_{sat} \)

- \( \beta = \mu C_{ox} \frac{W}{L} \)

\( V_{w.i.\_to\_s.i.} = 2n\phi_t \)

\( V_{s.i.\_to\_v.s.} = 2nL \frac{v_{sat}}{\mu} \)

\( V_{s.i.\_to\_v.s.} \) decreases with scaling!!!
Measurement example

NMOS, $W = 10 \, \mu m$, $L = 0.12 \, \mu m$

$V_{DS} = 1.2 \, V$, $V_{GS}$ swept from $0 \, V$ to $1.2 \, V$
Measurement example (2)

\[
\left( \frac{g_m}{I_{DS}} \right)_{\text{w.i.}} = \frac{1}{n\phi_t} \\
\left( \frac{g_m}{I_{DS}} \right)_{\text{s.i.}} = \sqrt{2} \frac{\beta}{n} \frac{1}{I_{DS}} \\
\left( \frac{g_m}{I_{DS}} \right)_{\text{v.s.}} = \frac{W C_{ox} v_{sat}}{I_{DS}}
\]

\[V_{DS} = 1.2 \text{ V, } V_{GS} \text{ swept from 0 V to 1.2 V}\]
The quantity $g_m r_0$ is called intrinsic gain of the transistor. It represents the maximum gain obtainable from a single transistor, and it is a very useful figure of merit in analog design.
Output resistance

\[ I_{DS} = \frac{\beta}{2n} (V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \]

\[ g_{out} = \frac{\partial I_{DS}}{\partial V_{DS}} = \lambda \cdot I_{DS_{\text{SAT}}} \]

\[ r_0 = \frac{1}{g_{out}} = \frac{1}{\lambda \cdot I_{DS_{\text{SAT}}}} = \frac{V_E \cdot L}{I_{DS_{\text{SAT}}}} \]

\[ \lambda = \frac{1}{V_{DS} - V_{DS_{\text{SAT}}}} \cdot \frac{\Delta L}{L - \Delta L} \approx \frac{1}{V_{DS} - V_{DS_{\text{SAT}}}} \cdot \frac{\Delta L}{L} \]

Dashed lines: ideal behavior
Scaling impact on the intrinsic gain

\[ \lambda \approx \frac{1}{V_{DS} - V_{DS\_SAT}} \cdot \frac{\Delta L}{L} \quad \Delta L \approx \sqrt{\frac{2\varepsilon_{Si}}{qN_a}} (V_{DS} - V_{DS\_SAT}) \]

\[ g_{out} = \lambda \cdot I_{DS\_SAT} \quad r_0 = \frac{1}{g_{out}} = \frac{1}{\lambda \cdot I_{DS\_SAT}} \quad \text{Intrinsic Gain} = g_m \cdot r_0 \]

Supposing to have constant field scaling for the technology, we obtain:

<table>
<thead>
<tr>
<th>W</th>
<th>L</th>
<th>( \beta )</th>
<th>( V_{GS}_V_T )</th>
<th>( g_m )</th>
<th>( V_{DS} )</th>
<th>( \Delta L )</th>
<th>( \lambda )</th>
<th>( I_{DS_SAT} )</th>
<th>( g_{out} )</th>
<th>( r_0 )</th>
<th>( g_m r_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1/( \alpha )</td>
<td>( \alpha^2 )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>1</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1/( \alpha )</td>
<td>1</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha^2 )</td>
<td>( \alpha^2 )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha^3 )</td>
<td>1/( \alpha )</td>
<td>( \alpha^2 )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha^2 )</td>
<td>1/( \alpha^2 )</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
The intrinsic gain is proportional to “$\alpha \cdot L$”: if $L$ is kept constant $g_m r_0$ increases by the scaling factor, if $L$ is decreased by $\alpha$ then $g_m r_0$ stays constant.

$$g_m \cdot r_0 \propto \alpha \cdot L$$

This result is based on the following assumptions:

1. We consider Channel Length Modulation and not Drain Induced Barrier Lowering
2. The transistor is working in Strong Inversion
3. We applied the Constant Field Scaling rules

It can be shown that the result obtained is true even dropping the assumptions above.
Gate leakage current

Implications:
Static power consumption for digital circuits and shot noise for analog

Scaling impact on noise

\[ \frac{v_{in}^2}{\Delta f} = 4kTn_\gamma \frac{1}{g_m} + \frac{K_a}{C_{ox}WL} \frac{1}{f^\alpha} \]

\[ g_m = \sqrt{\frac{2}{n} \mu C_{ox} \frac{W}{L} I_{DS}} \]

\[ C_{ox} = \frac{\varepsilon_{SiO_2}}{t_{ox}} \]

**White noise:** keeping the same W/L ratio and the same current, we have an improvement in the noise since \( C_{ox} \) (and therefore \( g_m \)) increases with scaling.

**1/f noise:** if we suppose that the constant \( K_a \) does not change with scaling, we have an improvement in the noise if we keep the same device area (WL). Data taken from the Roadmap foresee that \( K_a \) will remain more or less constant even for the most advanced CMOS processes. This must, of course, be verified…
Data taken from the literature except from the 0.13 µm node and one of the 0.25 µm node points, which are our measurements.
Outline

• Motivation
• How scaling works for devices and interconnections
• Scaling impact on the transistor performance
  • Scaling impact on analog circuits performance
    ➢ Signal to Noise Ratio (SNR)
    ➢ Analog power consumption
    ➢ Low voltage issues
• Noise in mixed-mode integrated circuits
• ULSI processes: which options for analog?
• Conclusions
Scaling impact on power, speed, SNR

\[ \text{PWR} = I_{DS} \cdot V_{DD} \]

\[ \frac{V_{n\_white}^2}{g_m} = 4kTn_\gamma \frac{1}{1/\alpha} \]

\[ \text{SNR}_w = \frac{V_{DD}}{\sqrt{V_{n\_white}^2}} \]

Assuming constant field scaling and strong inversion:

<table>
<thead>
<tr>
<th>W</th>
<th>L</th>
<th>( \beta )</th>
<th>( I_{DS} )</th>
<th>PWR</th>
<th>( C_{ox} \cdot W \cdot L )</th>
<th>Q</th>
<th>( \Delta t = Q/I )</th>
<th>( \sqrt{V_{n_white}^2} )</th>
<th>SNR(_w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1/( \alpha^2 )</td>
<td>1/( \alpha^2 )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1/( \alpha )</td>
<td>( \alpha^2 )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>1/( \alpha^{1/2} )</td>
<td>1/( \alpha^{1/2} )</td>
<td></td>
</tr>
<tr>
<td>1/( \alpha )</td>
<td>1</td>
<td>1</td>
<td>1/( \alpha^2 )</td>
<td>1/( \alpha^3 )</td>
<td>1</td>
<td>1/( \alpha )</td>
<td>( \alpha^{1/2} )</td>
<td>1/( \alpha^{3/2} )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1/( \alpha^2 )</td>
<td>( \alpha )</td>
<td>1</td>
<td>( \alpha )</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>( \alpha^3 )</td>
<td>( \alpha )</td>
<td>1</td>
<td>( \alpha )</td>
<td>1/( \alpha )</td>
<td>1/( \alpha )</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

To maintain the same SNR we do not gain in Power !!!
Analog power consumption

Min. power consumption for class A analog circuits:

\[ P_{\text{min}} = 8 \pi kT \cdot \text{SNR} \cdot f_{\text{sig}} \cdot \frac{V_{\text{DD}}}{V_{\text{DD}} - \Delta V} \]

\(\Delta V\) is the fraction of the \(V_{\text{DD}}\) not used for signal swing

Optimal analog power/performance trade-off for 0.35 - 0.25 \(\mu m\) technologies

Low voltage issues

• Use rail-to-rail input stages
• Low $V_{DS_{SAT}} \rightarrow$ Big transistors $\rightarrow$ Low speed
• Use low-$V_T$ or 0-$V_T$ transistors
• Use multi-gain systems to have high dynamic range
• Use devices in W.I. ($V_{DS_{SAT}}$ and high $g_m/I_D$)
• Use current-mode architectures
• Use bulk-driven MOS
• If very low-power is needed, this can also be obtained at the system level
In all the solutions that we have seen up to now, the common-mode input voltage range is about $V_{DD} - V_{GS} - V_{DS_{SAT}}$. This can cause some problems, especially if we want to use the op amp as a buffer or if the power supply voltage is quite low.

This solution has the drawback of having a variable total transconductance.
Outline

• Motivation
• How scaling works for devices and interconnections
• Scaling impact on the transistor performance
• Scaling impact on analog circuits performance
• Noise in mixed-mode integrated circuits
  - Digital noise
  - Substrate noise
• ULSI processes: which options for analog?
• Conclusions
Digital noise in mixed-signal ICs

Integrating analog blocks on the same chip with digital circuits can have some serious implications on the overall performance of the circuit, due to the influence of the “noisy” digital part on the “sensitive” analog part of the chip.

The switching noise originated from the digital circuits can be coupled in the analog part through:

- The power and ground lines
- The parasitic capacitances between interconnection lines
- The common substrate

The substrate noise problem is the most difficult to solve.

Noise reduction techniques

• Quiet the Talker. Examples (if at all possible !!!):
  ➢ Avoid switching large transient supply current
  ➢ Reduce chip I/O driver generated noise
  ➢ Maximize number of chip power pads and use on-chip decoupling

• Isolate the Listener. Examples:
  ➢ Use on-chip shielding
  ➢ Separate chip power connections for noisy and sensitive circuits
  ➢ Other techniques depend on the type of substrate. See next slide

• Close the Listener’s ears. Examples:
  ➢ Design for high CMRR and PSRR
  ➢ Use minimum required bandwidth
  ➢ Use differential circuit architectures
  ➢ Pay a lot of attention to the layout

Different types of substrates

There are mainly two types of wafers:

1. Lightly doped wafers: “high” resistivity, in the order of $10 \, \Omega \cdot \text{cm}$.

2. Heavily doped wafers: usually made up by a “low” resistivity bulk ($\sim 10 \, \text{m}\Omega/\text{cm}$) with a “high” resistivity epitaxial layer on top.

TSMC, UMC, IBM and STM (below 180 nm) offer type 1
Substrate noise reduction techniques

- In the case of a lightly doped substrate we can:
  - Use guard rings around the sensitive circuits to isolate them from the noisy circuits. Guard rings (biased separately) can also be used around the noisy circuits
  - Separate the sensitive and the noisy circuits
- For a heavily doped substrate, the above mentioned techniques are not very effective. The best option in this case is to have a good backside contact to have a low impedance connection to ground.
- In both cases, but especially with heavily doped substrates, it is a good idea to separate the ground contact from the substrate contact in the digital logic cells, to avoid to inject the digital switching current directly into the substrate.
Outline

• Motivation
• How scaling works for devices and interconnections
• Scaling impact on the transistor performance
• Scaling impact on analog circuits performance
• Noise in mixed-mode integrated circuits
  • ULSI processes: which options for analog?
• Conclusions
Available features and devices

- Shallow Trench Isolation (STI)
- Cobalt salicided N\(^+\) and P\(^+\) polysilicon and diffusions
- Low K dielectrics for interconnections
- Vertical Parallel Plate (VPP) capacitors and MOS varactors

Options:
- Multiple gate oxide thicknesses (→ supply voltages)
- Several different metal options
- Resistors: diffusion, poly, metal
- Triple well NMOS
- Low-\(V_T\), High-\(V_T\), Zero-\(V_T\) devices (thin and thick oxides)
- Metal-to-metal capacitors
- Electronic fuses
- Inductors
Conclusions

- The future of analog design in deep submicron processes in the 180 nm – 100 nm range looks quite promising. But it will not be straightforward for analog circuit to have the required SNR and speed without increasing the power dissipation.
- For analog applications in which speed and density are important, scaling can be very beneficial.
- It is clear that scaling brings some very important benefits for digital circuits. Digital circuits are profiting more from scaling than analog circuits. Example: in a mm² we can fit 200,000 gates running at 1 GHz and dissipating 0.8 W, or we could fit a full ARM microprocessor.
- This suggests that, within an ASIC, the position of the ideal separation line between analog and digital circuitry will have to be reconsidered.
- The problem of the substrate noise will have to be studied in detail.
Acknowledgements

I would like to especially thank:

• The conference organizers for giving me the opportunity to give this talk
• Federico Faccio and Alessandro Marchioro for many useful comments
• Alessandro La Rosa for the 0.13 µm noise measurements
• Silvia Baldi for the 0.13 µm static measurements
• Gianluigi De Geronimo, Paul O’Connor and Veljko Radeka for providing a very good working environment during my visit at BNL and for many useful discussions
SPARE
SLIDES
Constant field scaling

Width of a depleted zone as a function of the bias V

\[ x_d = \sqrt{\frac{2\varepsilon_{Si}}{qN_A}} \left( \phi_{bi} + V \right) \]

Threshold voltage of a MOS transistor

\[ V_T = V_{FB} + \phi_0 + \frac{\sqrt{2 \cdot q \cdot \varepsilon_{Si} \cdot N_A}}{C_{ox}} \left( \phi_0 + V_{SB} \right) \]

L ↓ → \( x_d \) ↓ → \( N_A \) ↑ and V ↓ → \( V_{DD} \) ↓

\( N_A \) ↑ → \( V_T \) ↑ → \( t_{ox} \) ↓
**Generalized selective scaling**

<table>
<thead>
<tr>
<th>Physical parameter</th>
<th>Constant-Electric Field Scaling Factor</th>
<th>Generalized Scaling Factor</th>
<th>Generalized Selective Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length, Insulator thickness</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha_d$</td>
</tr>
<tr>
<td>Wiring width, channel width</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha_w$</td>
</tr>
<tr>
<td>Electric field in device</td>
<td>1</td>
<td>$\varepsilon$</td>
<td>$\varepsilon$</td>
</tr>
<tr>
<td>Voltage</td>
<td>$1/\alpha$</td>
<td>$\varepsilon/\alpha$</td>
<td>$\varepsilon/\alpha_d$</td>
</tr>
<tr>
<td>On-current per device</td>
<td>$1/\alpha$</td>
<td>$\varepsilon/\alpha$</td>
<td>$\varepsilon/\alpha_w$</td>
</tr>
<tr>
<td>Doping</td>
<td>$\alpha$</td>
<td>$\varepsilon\alpha$</td>
<td>$\varepsilon\alpha_d$</td>
</tr>
<tr>
<td>Area</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha_d^2$</td>
</tr>
<tr>
<td>Capacitance</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha_w$</td>
</tr>
<tr>
<td>Gate delay</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha_d$</td>
</tr>
<tr>
<td>Power dissipation</td>
<td>$1/\alpha^2$</td>
<td>$\varepsilon/\alpha^2$</td>
<td>$\varepsilon/\alpha_w\alpha_d$</td>
</tr>
<tr>
<td>Power density</td>
<td>1</td>
<td>$\varepsilon^2$</td>
<td>$\varepsilon^2\alpha_w/\alpha_d$</td>
</tr>
</tbody>
</table>

$\alpha$ is the dimensional scaling parameter, $\varepsilon$ is the electric field scaling parameter, and $\alpha_D$ and $\alpha_W$ are separate dimensional scaling parameters for the selective scaling case. $\alpha_D$ is applied to the device vertical dimensions and gate length, while $\alpha_W$ applies to the device width and the wiring.


LECC 2004

Giovanni Anelli, CERN
Weak inversion region width

$t_{ox}$ scales for the same device dimensions the boundary between weak inversion and strong inversion moves towards higher currents.

\[ I_{DS\_w.i.\_to\_s.i.} = 2\mu C_{ox} \frac{W}{L} n\phi_t^2 \]
Scaling impact on $\mu C_{\text{ox}}$

Due to the scaling of the gate oxide thickness, the specific gate capacitance $C_{\text{ox}}$ increases with scaling. This increases the transistor driving capability. For a given $W/L$ ratio and a fixed bias current, the transconductance also increases with scaling.

$$g_m = \sqrt{\frac{2}{n} I_{DS}} = \frac{\beta}{n} (V_{GS} - V_T)$$

$$\beta = \mu C_{\text{ox}} \frac{W}{L}$$

$$C_{\text{ox}} = \frac{\varepsilon_{\text{SiO}_2}}{t_{\text{ox}}}$$

<table>
<thead>
<tr>
<th>$L_{\text{min}}$ [(\mu\text{m})]</th>
<th>$t_{\text{ox_physical}}$ [nm]</th>
<th>$t_{\text{ox_effective}}$ [nm]</th>
<th>$C_{\text{ox}}$ [fF/(\mu\text{m}^2)]</th>
<th>$\mu C_{\text{ox}}$ [(\mu\text{A}/\text{V}^2)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>17</td>
<td>---</td>
<td>2.03</td>
<td>~ 90</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>---</td>
<td>3.45</td>
<td>~ 134</td>
</tr>
<tr>
<td>0.25</td>
<td>5.5</td>
<td>6.2</td>
<td>5.5</td>
<td>~ 250</td>
</tr>
<tr>
<td>0.18</td>
<td>4.1</td>
<td>---</td>
<td>---</td>
<td>~ 340</td>
</tr>
<tr>
<td>0.13</td>
<td>2.2</td>
<td>3.15</td>
<td>10.9</td>
<td>~ 490</td>
</tr>
</tbody>
</table>

The values above are taken from measurements, design manuals or obtained from simulations. The $\mu C_{\text{ox}}$ values are for NMOS transistors with low vertical field.

Output conductance

\[ G_{\text{out}} = \frac{\Delta I}{\Delta V} = \frac{I_D}{\Delta V} \cdot \frac{\Delta L}{L - \Delta L} \]
Output resistance $r_0$

\[
I_{DS} = \frac{\beta}{2n} \left( V_{GS} - V_T \right)^2 \left( 1 + \lambda V_{DS} \right)
\]

\[
V_{DS\_SAT} = \frac{V_{GS} - V_T}{n}
\]

\[
I_{DS\_SAT} = \frac{\beta}{2n} \left( V_{GS} - V_T \right)^2 = \frac{\beta}{2} n V_{DS\_SAT}^2
\]

\[
g_{out} = g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}} = \lambda \cdot I_{DS\_SAT}
\]

\[
r_0 = \frac{1}{g_{ds}} = \frac{1}{\lambda \cdot I_{DS\_SAT}} = \frac{V_E \cdot L}{I_{DS\_SAT}}
\]

\[
\lambda = \frac{1}{V_{DS} - V_{DS\_SAT}} \cdot \frac{\Delta L}{L - \Delta L} \approx \frac{1}{V_{DS} - V_{DS\_SAT}} \cdot \frac{\Delta L}{L}
\]
Matching will have a very important impact on the performance of deep submicron CMOS circuits


LECC 2004

Giovanni Anelli, CERN

\[ \sigma_{\Delta V_{th}} = \frac{A_{Vth}}{\sqrt{WL}} \]
The ion implantation process follows Poisson statistics. Therefore, the uncertainty in the number of dopant implanted is given by the square root of the number. The error becomes proportionally more important for smaller devices! ($=1/\sqrt{N}$)
Scaling & dopant fluctuations

\[
\sigma_{\Delta V_{th}} = C \cdot \frac{t_{ox} \cdot 4\sqrt{N}}{\sqrt{W L}}
\]

- For the same device dimensions, matching improves
- For minimum size devices, matching might be worse

<table>
<thead>
<tr>
<th>L_{min} [\mu m]</th>
<th>t_{ox} [nm]</th>
<th>N_a [cm^{-3}]</th>
<th>A_N / t_{ox} [mV\cdot\mu m / nm]</th>
<th>A_N [mV\cdot\mu m]</th>
<th>\sigma_{\Delta V_{th}} [mV]</th>
<th>6 \cdot \sigma_{V_{th}} [mV]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>25</td>
<td>5 \cdot 10^{16}</td>
<td>0.328</td>
<td>8.2</td>
<td>6.84</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>6 \cdot 10^{16}</td>
<td>0.344</td>
<td>6.9</td>
<td>6.89</td>
<td>29.2</td>
</tr>
<tr>
<td>0.8</td>
<td>15</td>
<td>7.5 \cdot 10^{16}</td>
<td>0.365</td>
<td>5.5</td>
<td>6.84</td>
<td>29</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>1.2 \cdot 10^{17}</td>
<td>0.414</td>
<td>4.1</td>
<td>8.28</td>
<td>35.1</td>
</tr>
<tr>
<td>0.25</td>
<td>5.5</td>
<td>2.4 \cdot 10^{17}</td>
<td>0.498</td>
<td>2.7</td>
<td>11</td>
<td>46.5</td>
</tr>
<tr>
<td>0.18</td>
<td>4</td>
<td>3.3 \cdot 10^{17}</td>
<td>0.542</td>
<td>2.2</td>
<td>12</td>
<td>51.1</td>
</tr>
</tbody>
</table>

Matching data from the Roadmap

Data taken from The International Technology Roadmap for Semiconductors (2001 Edition)
Fig. 8. Minimum power consumption for a voltage follower circuit (with constant topology, SINAD and maximum signal frequency). (a) Made in various processes, operated at their corresponding supply voltages. (b) Made in various processes, all operated at $V_{dd} = 3.3$ V. (c) All made in 0.5-$\mu$m CMOS, operated at various supply voltages.
Speed-power-accuracy trade off

\[
\text{Speed} \propto \frac{g_m}{WLC_{\text{ox}}}
\]

\[
\sigma_{\Delta V_{\text{th}}} = \frac{A_{V_{\text{th}}}}{\sqrt{W L}}
\]

\[
\text{Accuracy} \propto \frac{V_{DD}}{\sigma_{\Delta V_{\text{th}}}} = \frac{V_{DD}}{A_{V_{\text{th}}}} \cdot \sqrt{WL}
\]

\[
\text{Power} \propto I \cdot V_{DD}
\]

\[
\text{Speed} \cdot \text{Accuracy}^2 \propto \left( \frac{g_m}{I} \right) \cdot \frac{V_{DD}}{C_{\text{ox}} \cdot A_{V_{\text{th}}}^2}
\]
Multi-metal-layer capacitors

This solution is a possibility, but it does not exploit the fact that in deep submicron processes the highest parasitic capacitance can be obtained “horizontally” rather than vertically, i.e. $t_{ox} > s$

Multi-metal-layer capacitors


LECC 2004

Giovanni Anelli, CERN