

Analog Circuit Design in Nanoscale CMOS Technologies

Opportunities and Challenges

Trond Ytterdal

Circuits and Systems Group

Department of Electronics and Telecommunication

**Norwegian University of Science and Technology
(NTNU)**

2006-2007: Sabbatical at UofT

Room: BA5106

trond@eecg.toronto.edu

October 11, 2006

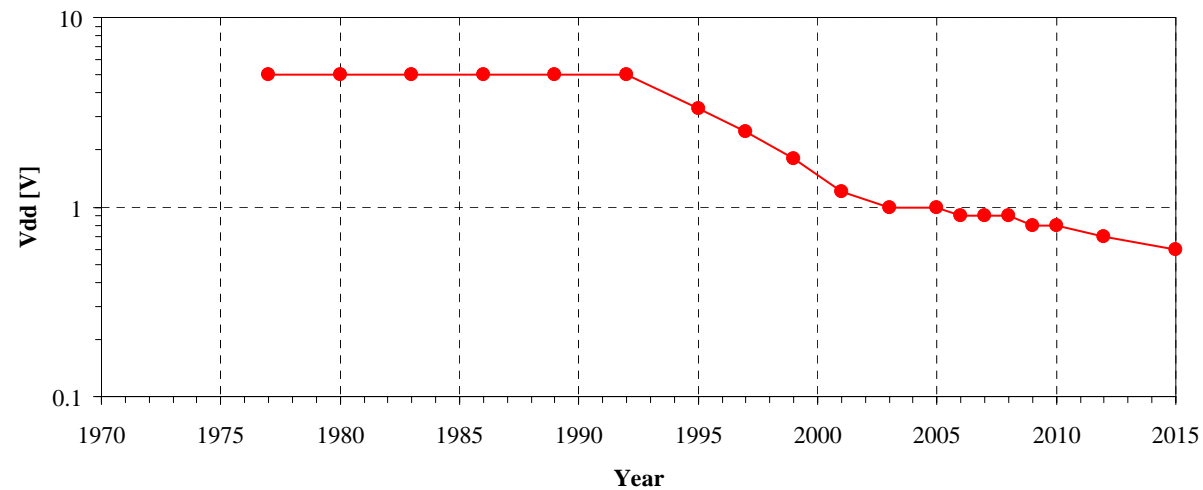
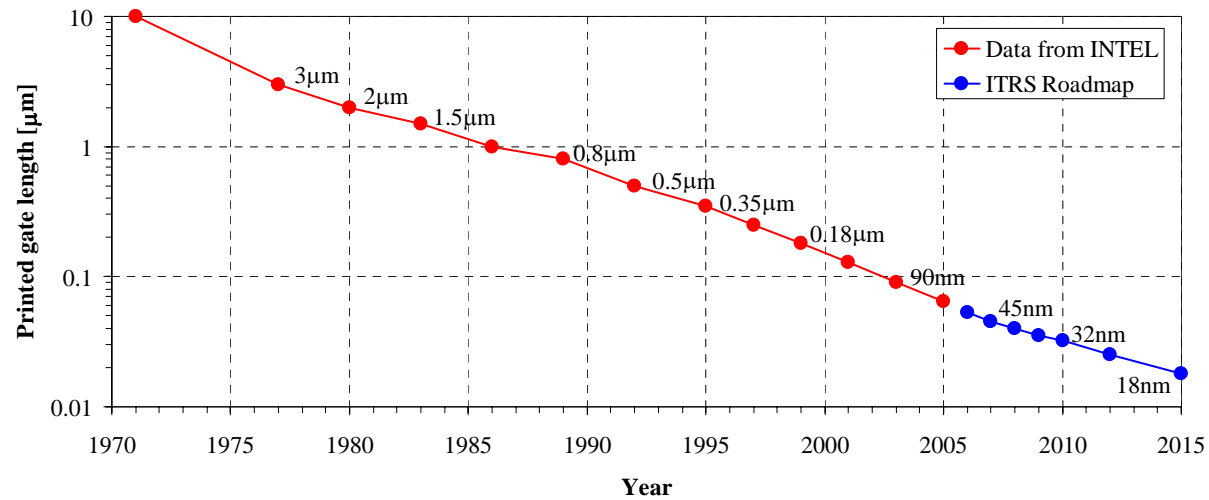


Outline

- **Introduction**
- **Transistor performance**
 - **How does transistor performance change as technology is scaled down to nanoscale* dimensions**
- **Analog circuit performance**
 - **How does analog circuit performance change as technology is scaled down**

*The term nanoscale is used for dimensions less than 100nm

CMOS technology downscaling

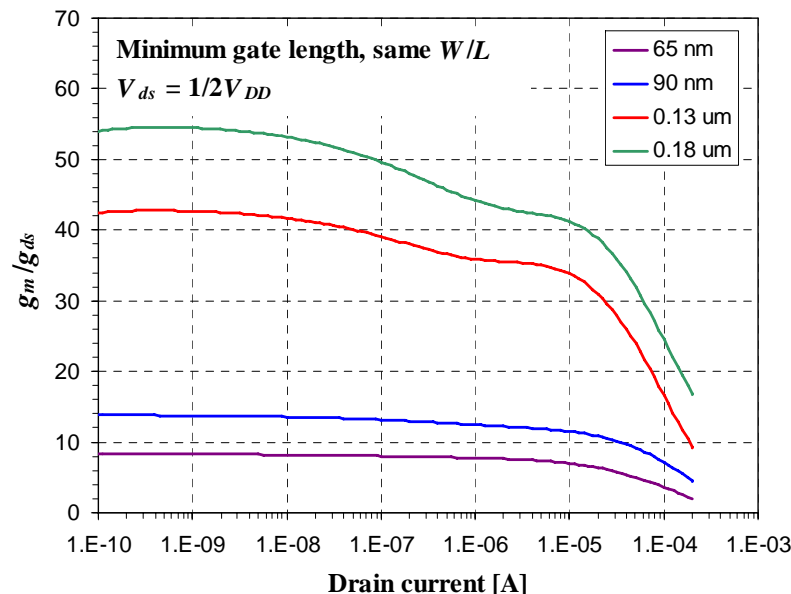
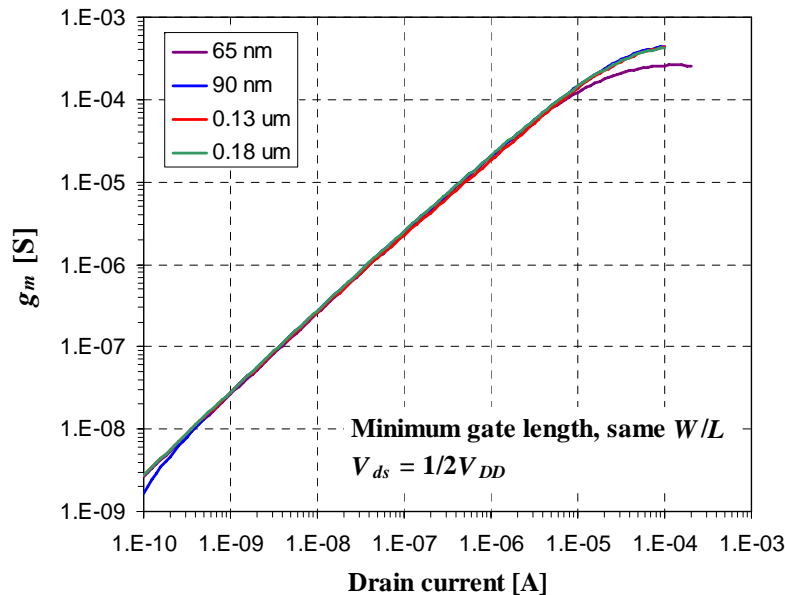


Gate length divided by two approximately every 5 years

Transistor performance

- Transistor performance metrics for analog and RF design:
 - Transconductance
 - Intrinsic gain (g_m/g_{ds})
 - Capacitances
 - Maximum operating frequency
 - Efficiency (g_m/I_d)
 - Linearity
 - Noise
 - Mismatch
 - Gate leakage

Transconductance and intrinsic gain



- Transconductance is, for all practical purposes, independent of the technology node

For a velocity saturated channel:

$$I_d = Wc_{ox}V_{eff}v_s$$

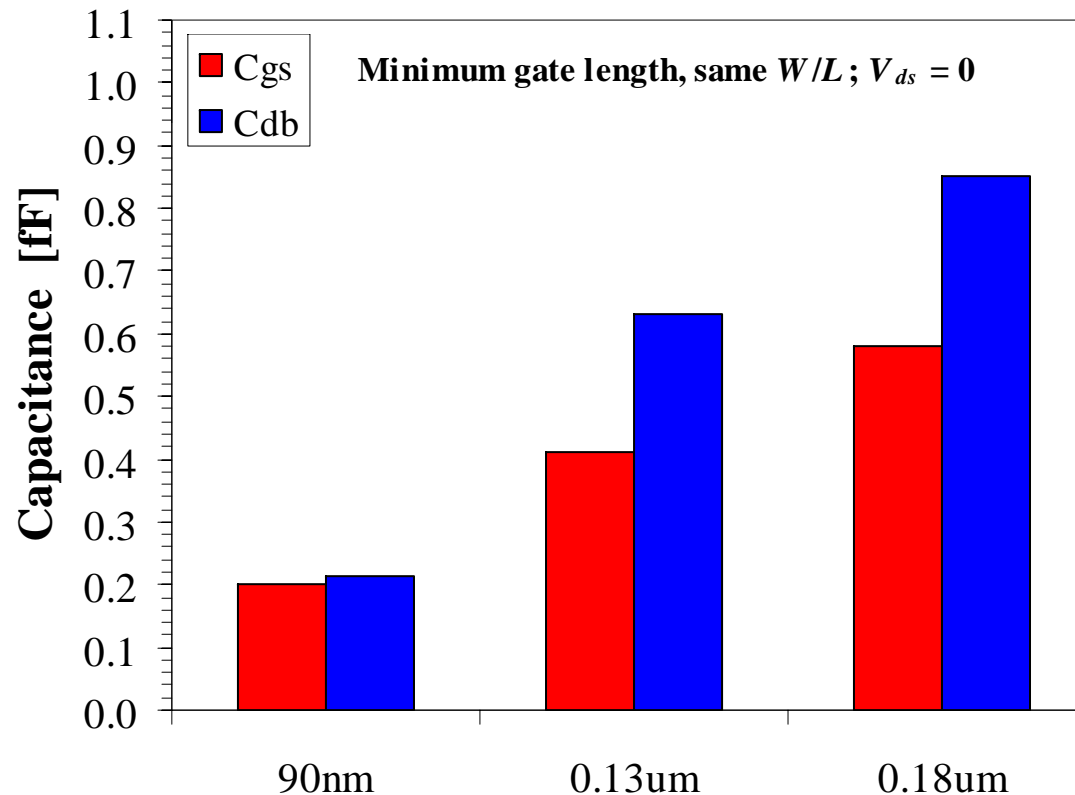
$$\Rightarrow g_m = Wc_{ox}v_s$$

$$W \propto L; \quad c_{ox} \propto \frac{1}{L}$$

$\Rightarrow g_m$ does not scale

- Intrinsic gain is reduced as technologies are scaled down. At the 65nm node the maximum intrinsic gain is reduced by more than 80% compared to the gain at the 0.18 μ m node.

Transistor capacitances

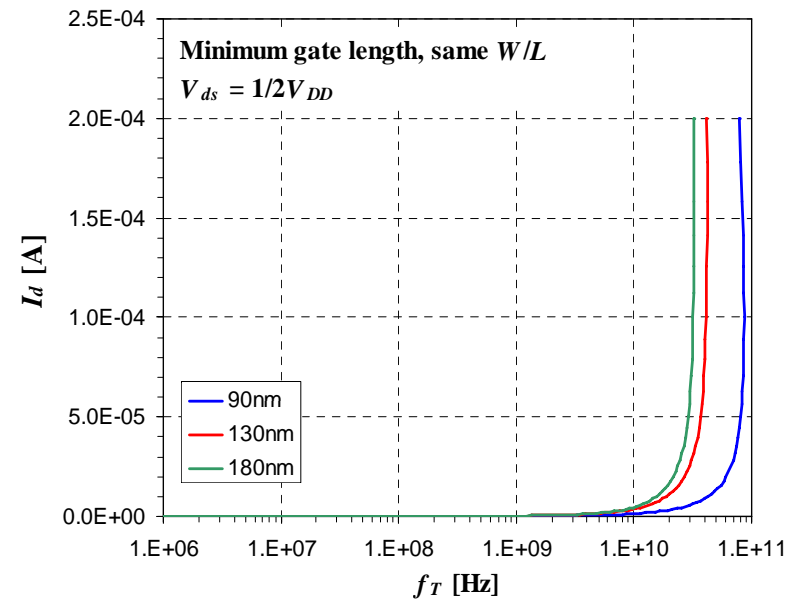
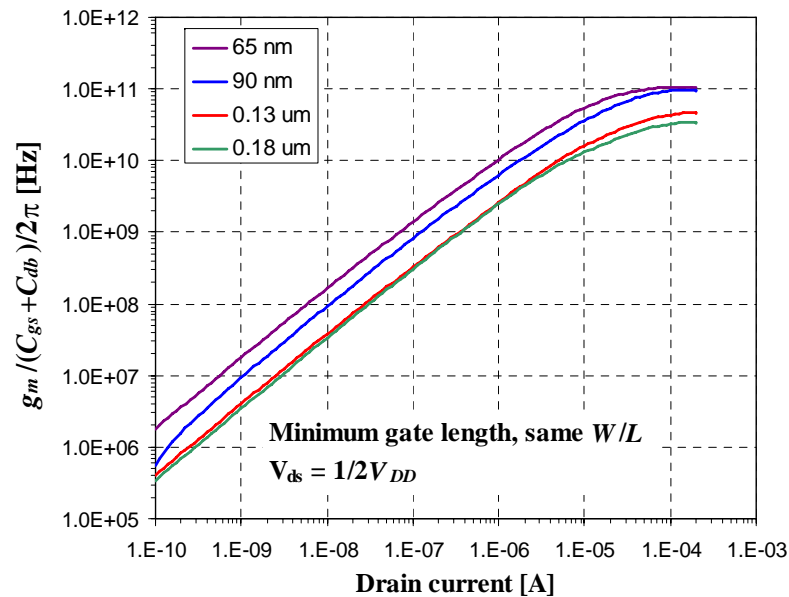


Capacitances extracted from NMOS having minimum drain and source areas for the given W/L (~3).

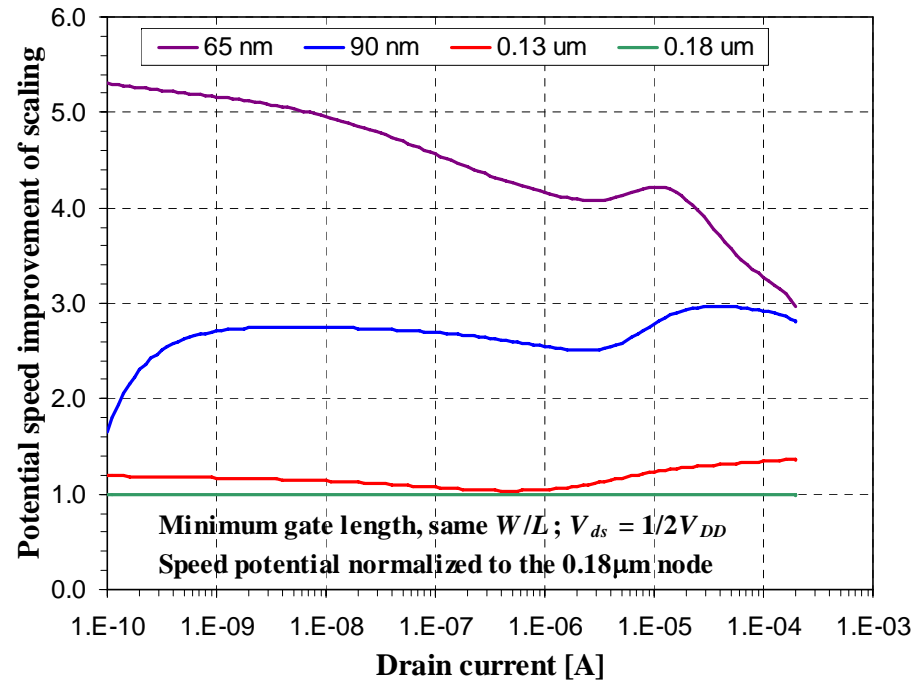
Maximum operating frequency

- The maximum speed of an amplifier is limited by the ratio of the transconductance and the capacitance of a transistor:

$$f_T = \frac{g_m}{2\pi(C_{gs} + C_{gd} + C_{db})} \quad (1)$$

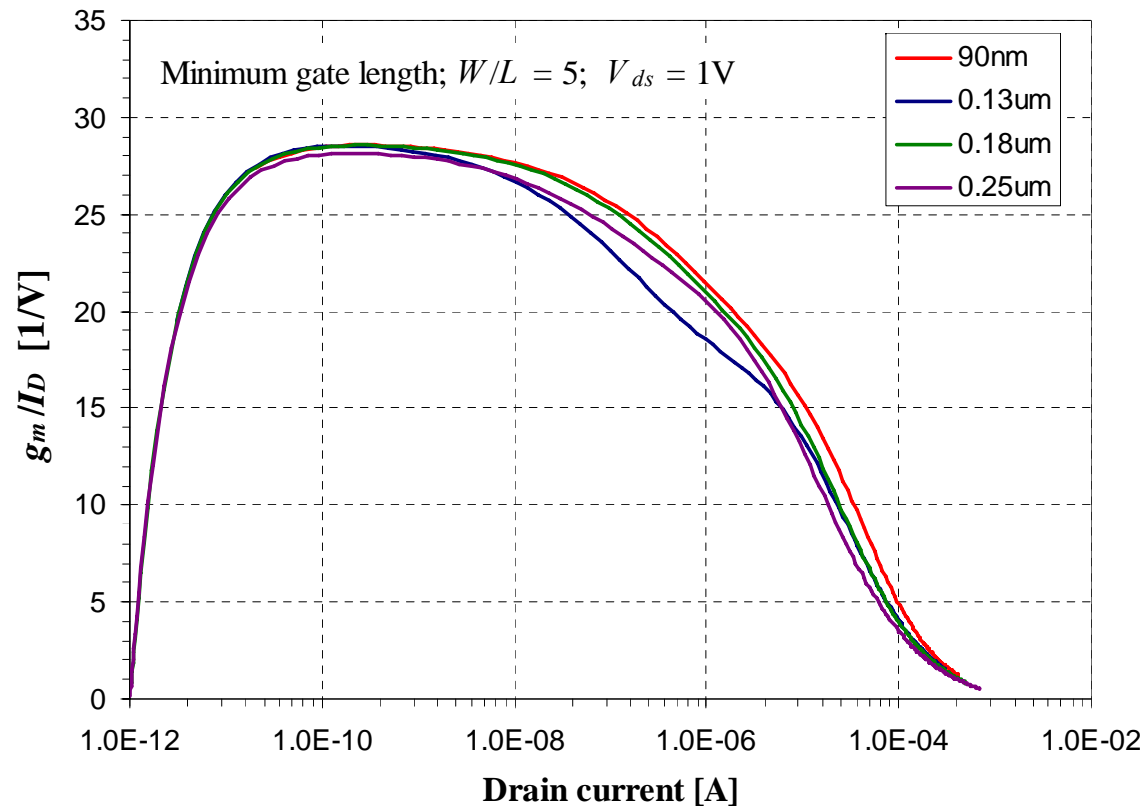


Potential speed improvement



- **By scaling down a potential improvement in speed can be obtained compared to realization in a 0.18 μ m technology**

Transistor efficiency (g_m/I_D)



- The transistor is most efficient (maximum bang for the buck) in weak inversion
- Maximum value is nearly independent of technology

Linearity (1)

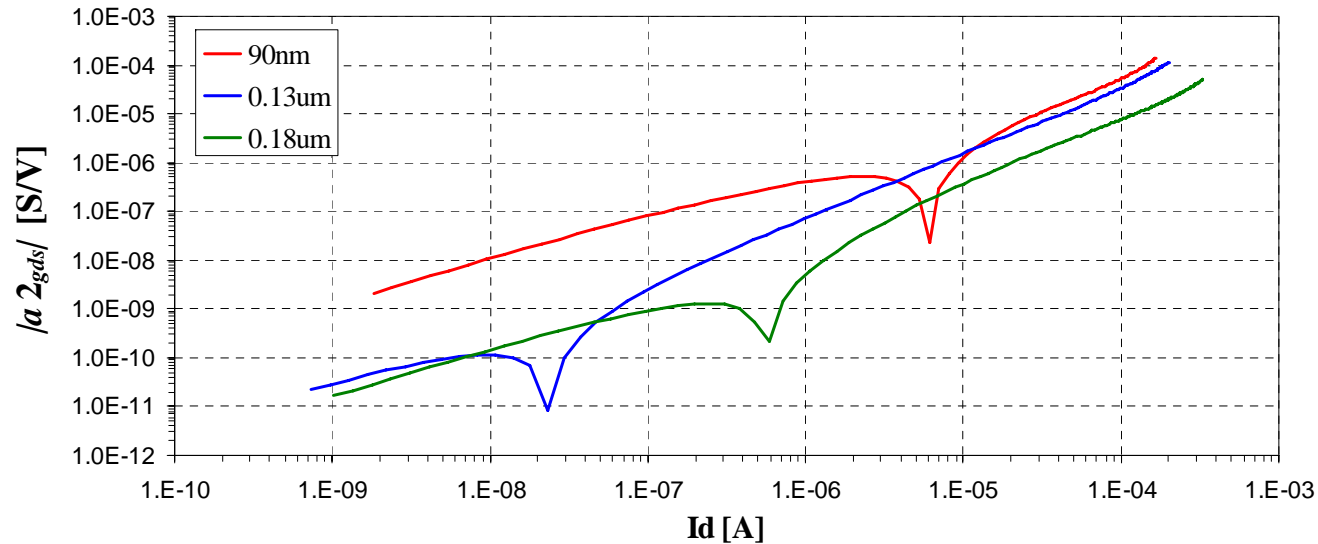
The Taylor expansion of the small signal drain current of a MOS transistor can be written as:

$$\begin{aligned}i_d = & g_m v_{gs} + a2_{g_m} v_{gs}^2 + a3_{g_m} v_{gs}^3 + \dots \\ & + g_{ds} v_{ds} + a2_{g_{ds}} v_{ds}^2 + a3_{g_{ds}} v_{ds}^3 + \dots \\ & + g_{mb} v_{bs} + a2_{g_{mb}} v_{bs}^2 + a3_{g_{mb}} v_{bs}^3 + \dots \\ & + a2_{g_m g_{ds}} v_{gs} v_{ds} + a3_{2g_m g_{ds}} v_{gs}^2 v_{ds} + a3_{g_m 2g_{ds}} v_{gs} v_{ds}^2 + \dots \\ & + a2_{g_m g_{mb}} v_{gs} v_{bs} + a3_{2g_m g_{mb}} v_{gs}^2 v_{bs} + a3_{g_m 2g_{mb}} v_{gs} v_{bs}^2 + \dots \\ & + a2_{g_{ds} g_{mb}} v_{ds} v_{bs} + a3_{2g_{ds} g_{mb}} v_{ds}^2 v_{bs} + a3_{g_{ds} 2g_{mb}} v_{ds} v_{bs}^2 + \dots \\ & + a3_{g_m g_{ds} g_{mb}} v_{gs} v_{ds} v_{bs} + \dots\end{aligned}$$

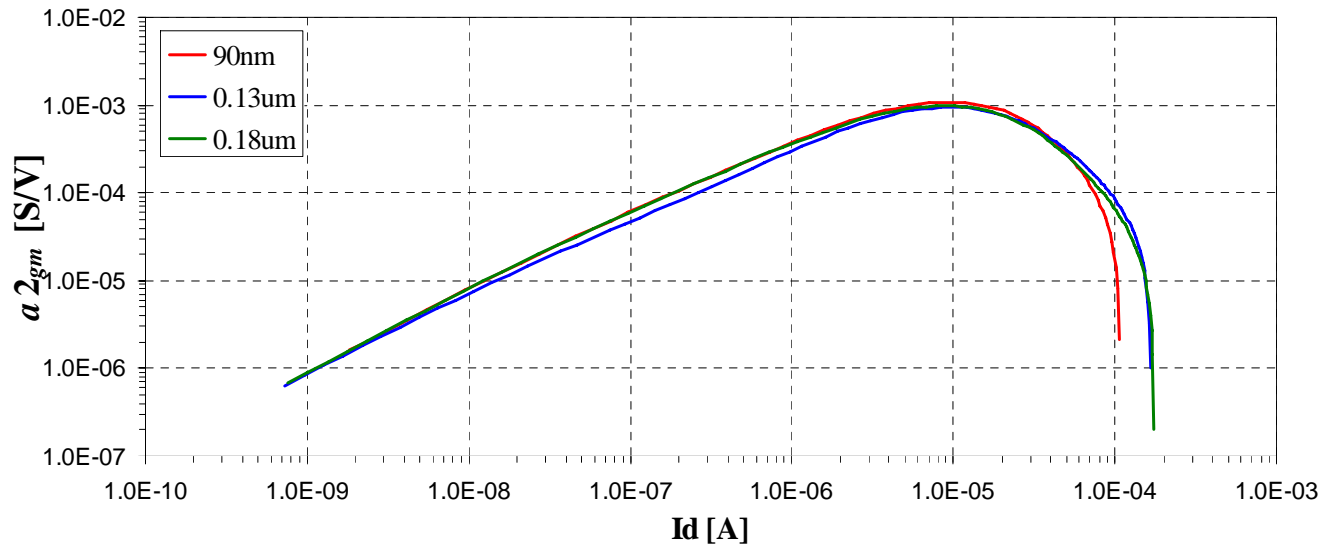
Here, the coefficients are the higher order derivatives of the total drain current with respect to one or more of the control voltages. For example:

$$a2_{g_m} = \frac{1}{2} \frac{\partial^2 I_d}{\partial V_{gs}^2}$$

Linearity (2)



Second order nonlinearity in channel conductance (g_{ds}) increases as technology is scaled down.



Second order nonlinearity in transconductance almost independent of technology

Noise

- Equivalent noise PSD (power spectral density) referred to the gate of a transistor connected in a common-source configuration*:

$$V_{ng}^2 = 4k_B T \left(\frac{\gamma}{g_m} + \frac{K}{W L C_{ox} f} \right) ; K \text{ process dependent}$$

Channel noise \uparrow \swarrow Si-SiO₂ interface noise

- Increases at low current
- Minimum in weak inversion for given current

Traditionally [1], the following expression is used for γ :

$$\gamma = \begin{cases} 2n/3 & \text{in strong inversion} \\ n/2 & \text{in weak inversion} \end{cases}$$

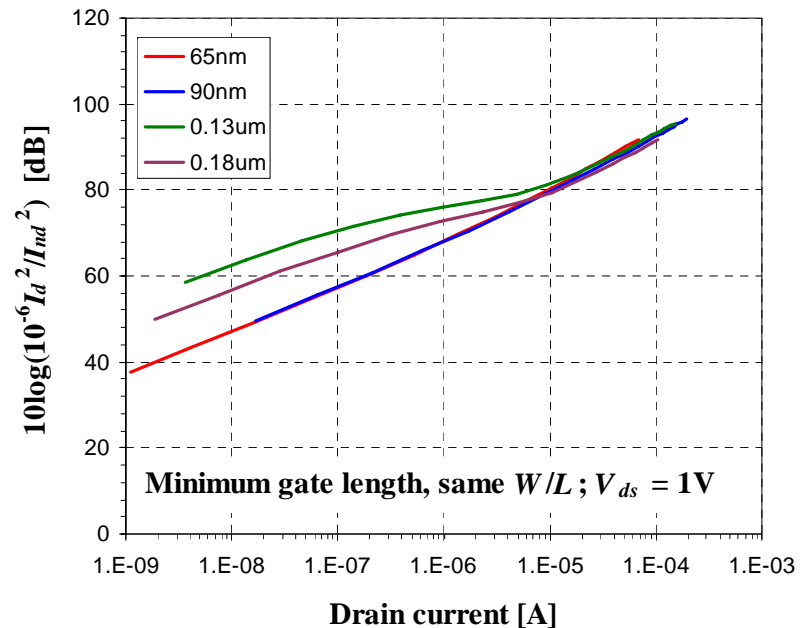
*Transform valid up to a frequency $\omega \sim g_m/C_{gd}$

Potential SNR (1)

Thermal noise

$$\frac{I_d^2}{I_{nd}^2} = \frac{1}{4k_B T \frac{\gamma}{I_d} \frac{g_m}{I_d}}$$

- **Decreases at low current**
- **Minimum in weak inversion for a given current**

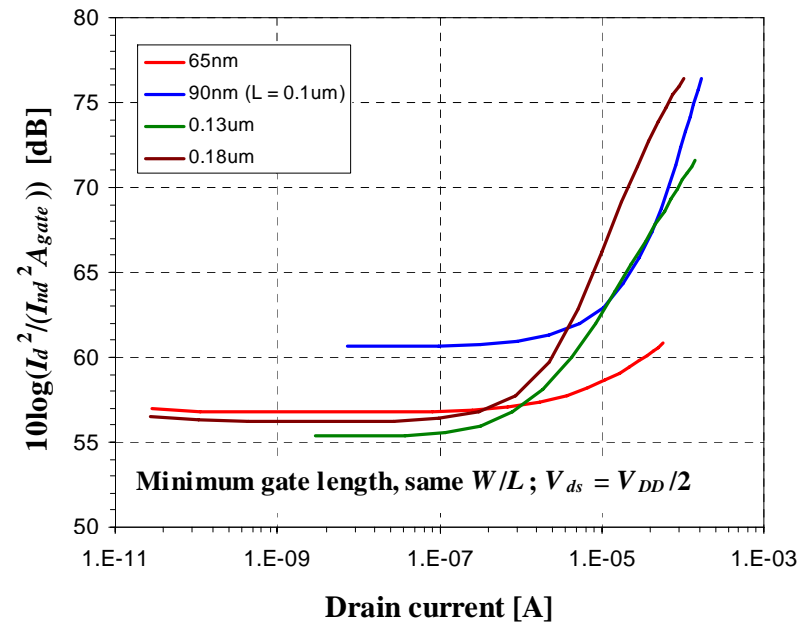


Potential SNR (2)

Flicker noise

$$\frac{I_d^2}{I_{nd}^2} = \frac{WLf}{4k_BTK \left(\frac{g_m}{I_d} \right)^2}$$

- **Decreases at low current**
- **Minimum in weak inversion for a given current**



Mismatch (1)

- **Mismatch of two matched transistors identical by design can be characterized by:**

$$\Delta V_T = V_{T1} - V_{T2} \quad \text{with} \quad \sigma_{V_T} = \sigma(\Delta V_T) = A_{V_T} / \sqrt{WL}$$

$$\Delta\beta/\beta = 2(\beta_1 - \beta_2)/(\beta_1 + \beta_2) \quad \text{with} \quad \sigma_\beta = \sigma(\Delta\beta/\beta) = A_\beta / \sqrt{WL}$$

A_{V_T} and A_β depends on process, layout and more ...

- **Based on the quadratic MOS model, one can derive the following well known formulas:**

Same gate voltage:

$$\sigma\left(\frac{\Delta I_d}{I_d}\right) = \sqrt{\sigma_\beta^2 + \left(\frac{g_m}{I_d} \sigma_{V_T}\right)^2}$$

Same drain current:

$$\sigma(\Delta V_G) = \sqrt{\sigma_{V_T}^2 + \left(\frac{I_d}{g_m} \sigma_\beta\right)^2}$$

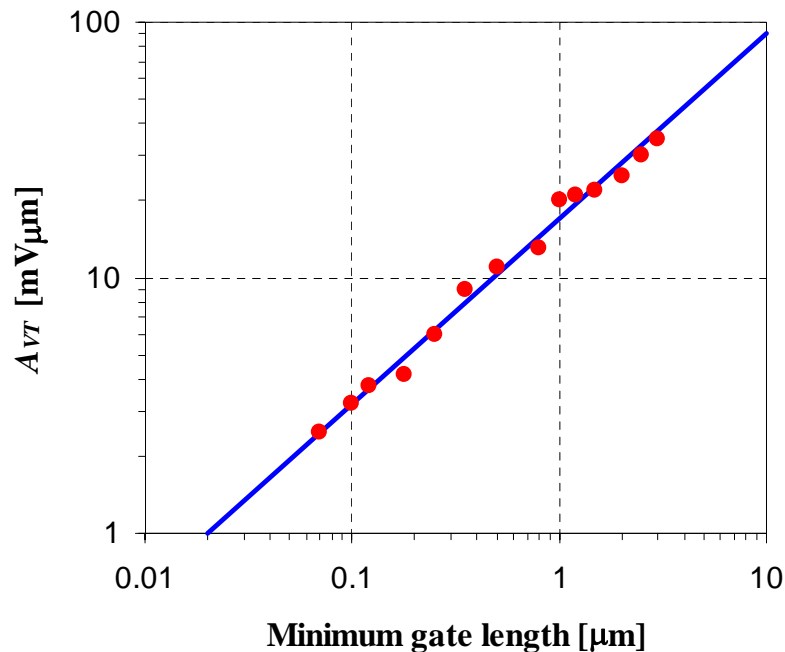
Mismatch (2)

- Mismatch per unit area decreases as technologies are scaled down

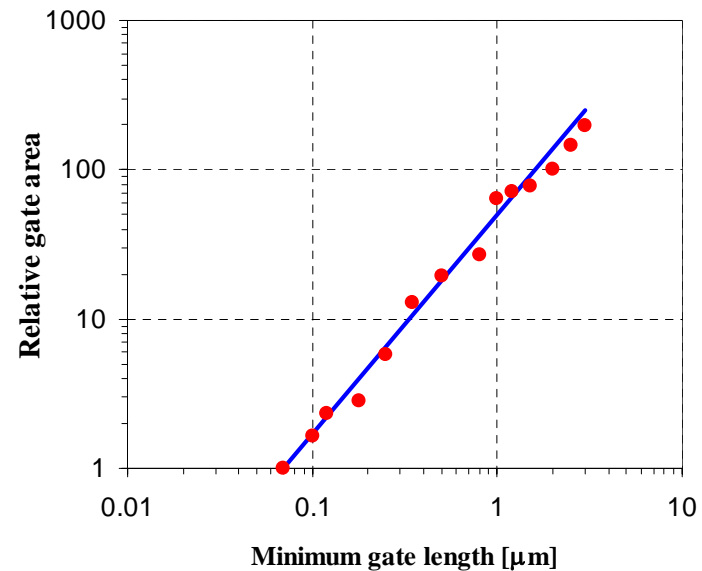
From theory and ideal scaling:

$$A_{VT} \propto t_{ox} \propto L_{min}$$

In reality A_{VT} does not scale as fast as L_{min}



Gate area required for any given $\sigma(\Delta V_T)$ normalized to the $L_{min} = 70\text{nm}$

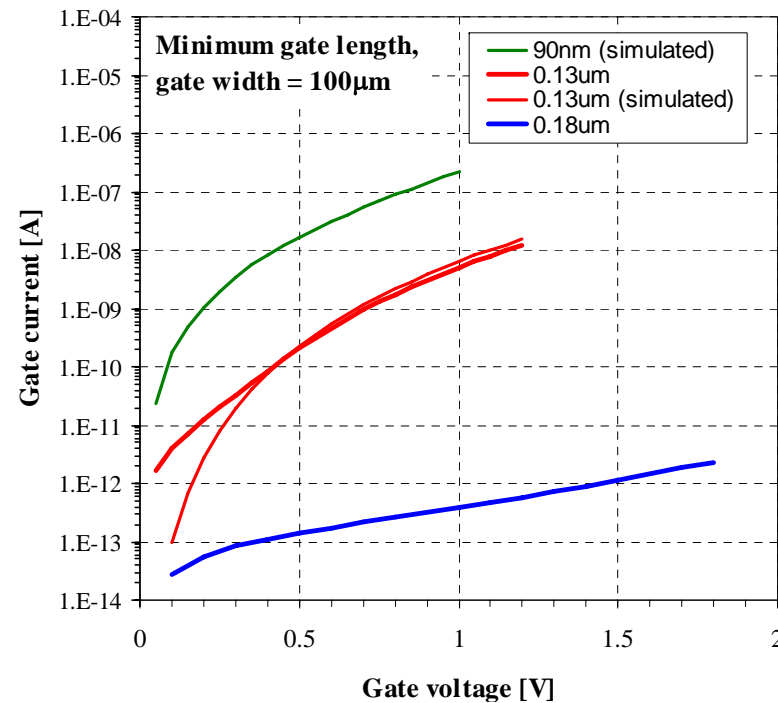
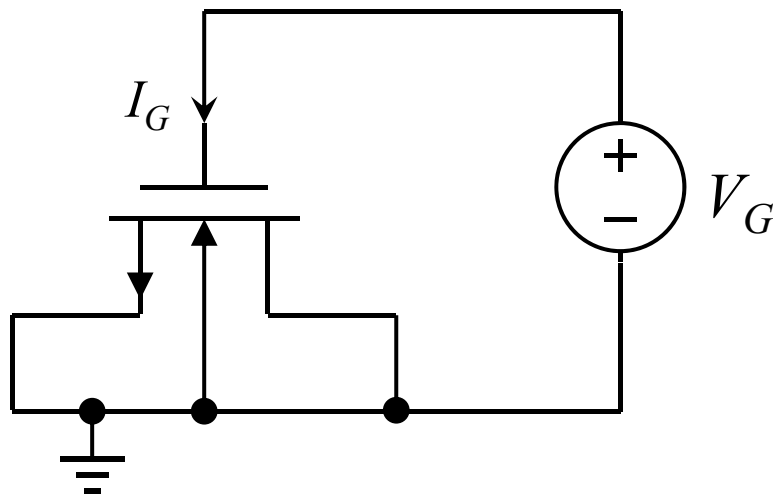


Experimental Data from [2]

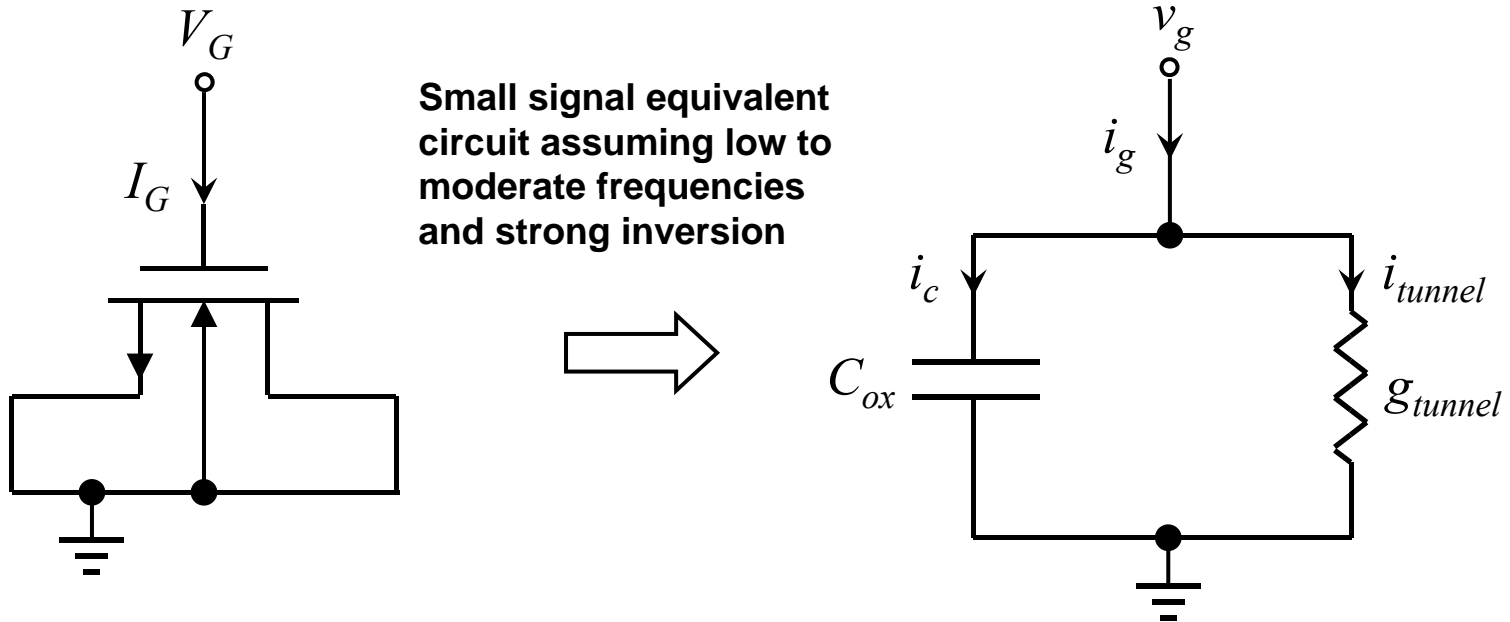
Gate leakage current (1)

Gate current can NOT be neglected in scaled down technologies.

Dominated by tunneling current => Relatively insensitive to device temperature.



Gate leakage current (2)



Small signal equivalent circuit assuming low to moderate frequencies and strong inversion

At a given frequency the impedance of the two branches have the same magnitude and the currents i_c and i_{tunnel} will be equal in magnitude. This frequency is given by

$$\omega_{gate} C_{ox} = g_{tunnel}$$

$$\Rightarrow f_{gate} = \frac{1}{2\pi} \frac{g_{tunnel}}{C_{ox}}$$

Above this frequency, the total gate current is dominated by i_c as in the traditional case.

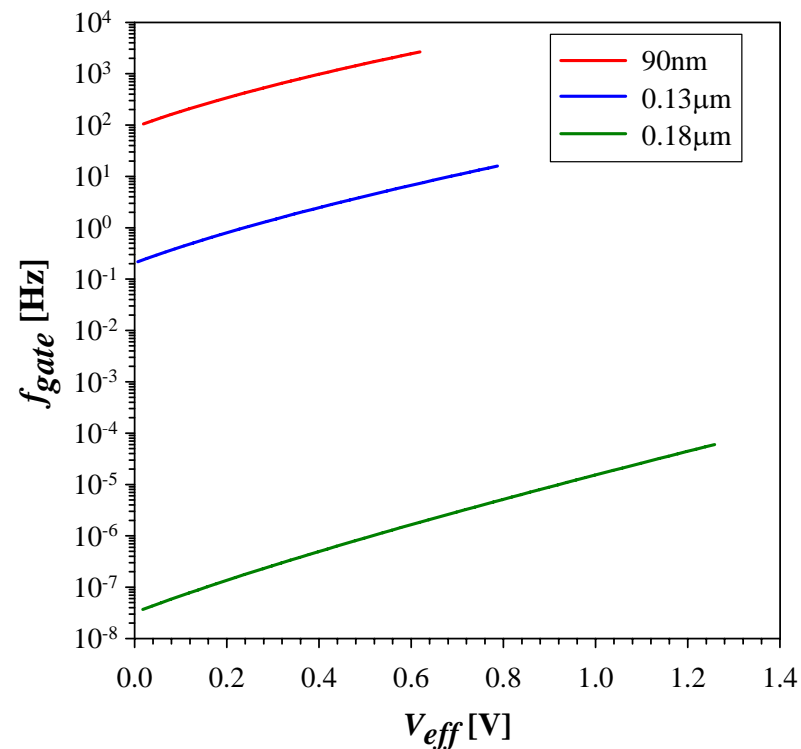
f_{gate} versus technology

f_{gate} is almost gate area independent and an approximate expression is given in [3] as:

$$f_{gate} \cong K_{fg} V_{gs}^2 e^{t_{ox}(V_{gs}-13.6)}$$

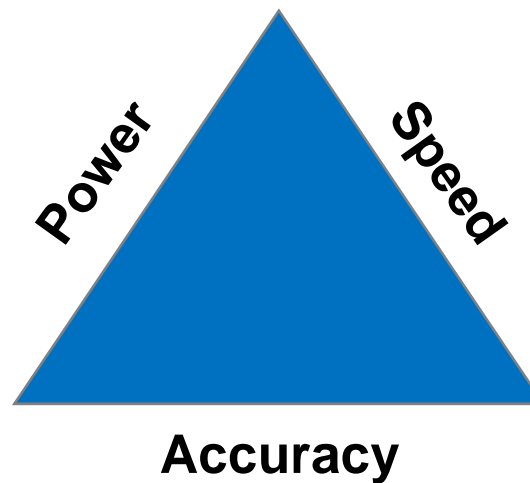
Here, t_{ox} is the oxide thickness in nm and K_{fg} is $1.5 \cdot 10^{16}$ for NMOS and $0.5 \cdot 10^{16}$ for PMOS.

A plot of this expression is shown to the right for three different technologies.



Analog circuit performance

- Analog and RF circuit performance metrics:
 - Accuracy (Dynamic range)
 - Speed
 - Power consumption



FOM definition

- One commonly used Figure-of-merit (*FOM*) definition:

$$FOM = \frac{P}{DR^2 \cdot f}$$

Here, P is the power consumption, DR is the dynamic range and f is the highest signal frequency that can be processed by the circuit

- Small is **GOOD**.

Dynamic range (1)

- Here, we define the dynamic range (DR) as the ratio of the largest to the smallest possible signal amplitudes.*
- If we assume that DR is limited by noise on the lower side and the power supply voltage on the upper side, we can write

$$DR = \frac{V_{pp}/2}{V_{n_rms}} = \frac{\eta_v V_{DD}}{2V_{n_rms}} \quad (2)$$

where, V_{pp} is the maximum peak-to-peak value of the signal, V_{n_rms} is the RMS noise voltage, η_v is the voltage efficiency defined by V_{pp}/V_{DD} .

*The ratio of squared amplitudes is also a common definition

Dynamic range (2)

- If we assume that matching limits DR on the lower side, we can write

$$DR = \frac{V_{pp}/2}{\kappa\sigma(V_{os})} = \frac{\eta_v V_{DD}}{2\kappa\sigma(V_{os})} \quad (3)$$

- where κ is a yield parameter and $\sigma(V_{os})$ is the standard deviation of a critical offset voltage in the circuit.
- We observe for both (2) and (3) that DR scales with V_{DD} and η_v

For the rest of the talk we assume that DR is limited by noise on the lower side

Signal-to-noise ratio (SNR)

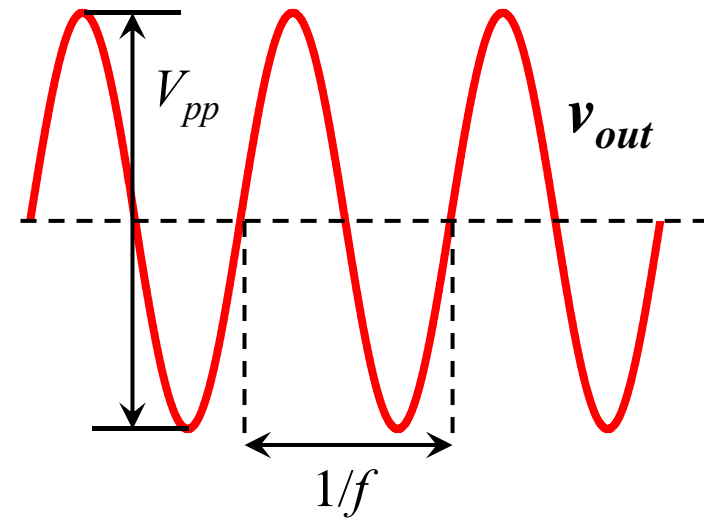
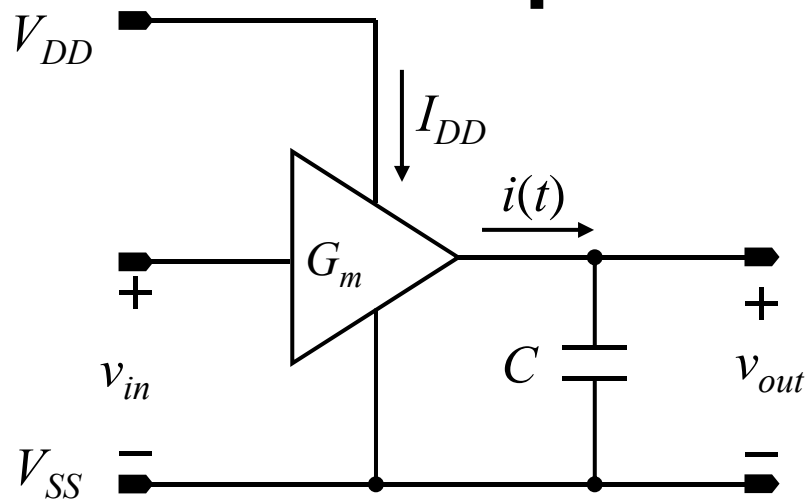
- SNR is defined as the ratio of the signal power to the noise power. For a sinusoidal signal, the maximum SNR given by

$$SNR = \frac{V_{pp}^2}{8V_{n_rms}^2} = \frac{(\eta_v V_{DD})^2}{8V_{n_rms}^2} \quad (4)$$

- If we compare (4) with (2), we see that

$$DR = \sqrt{2 \cdot SNR} \quad (5)$$

SNR versus power



Power:
$$P = V_{DD} \overline{I_{DD}} = V_{DD} \frac{1}{\eta_i} f \Delta q = V_{DD} \frac{1}{\eta_i} f V_{PP} C = \frac{1}{\eta_v} \frac{1}{\eta_i} f V_{PP}^2 C$$

Here, η_i is the current efficiency of the transconductor and $\Delta q / \eta_i$ is the charge transferred from the power supply in one period and. We have assumed that $V_{SS} = 0$

SNR:
$$SNR = \frac{V_{PP}^2 / 8}{k_B T / C} \Leftrightarrow V_{PP}^2 = 8 \frac{k_B T}{C} SNR$$
 Only thermal noise is considered.

Combining the two equations we get the power required per pole versus SNR :

$$P = \frac{8k_B T \cdot f \cdot SNR}{\eta_i \eta_v} \quad (6)$$

Based on [1] and [4]

Theoretical FOM

- Rewriting (6) in terms of *FOM*:

$$\frac{P}{f \cdot SNR} = \frac{8k_B T}{\eta_i \eta_v} \Rightarrow \frac{P}{f \cdot DR^2} = \frac{4k_B T}{\eta_i \eta_v}$$
$$\Rightarrow FOM = \frac{4k_B T}{\eta_i \eta_v}$$

- Ideal world ($\eta_v = \eta_i = 1$):

$$FOM_i = 4k_B T \quad (7)$$

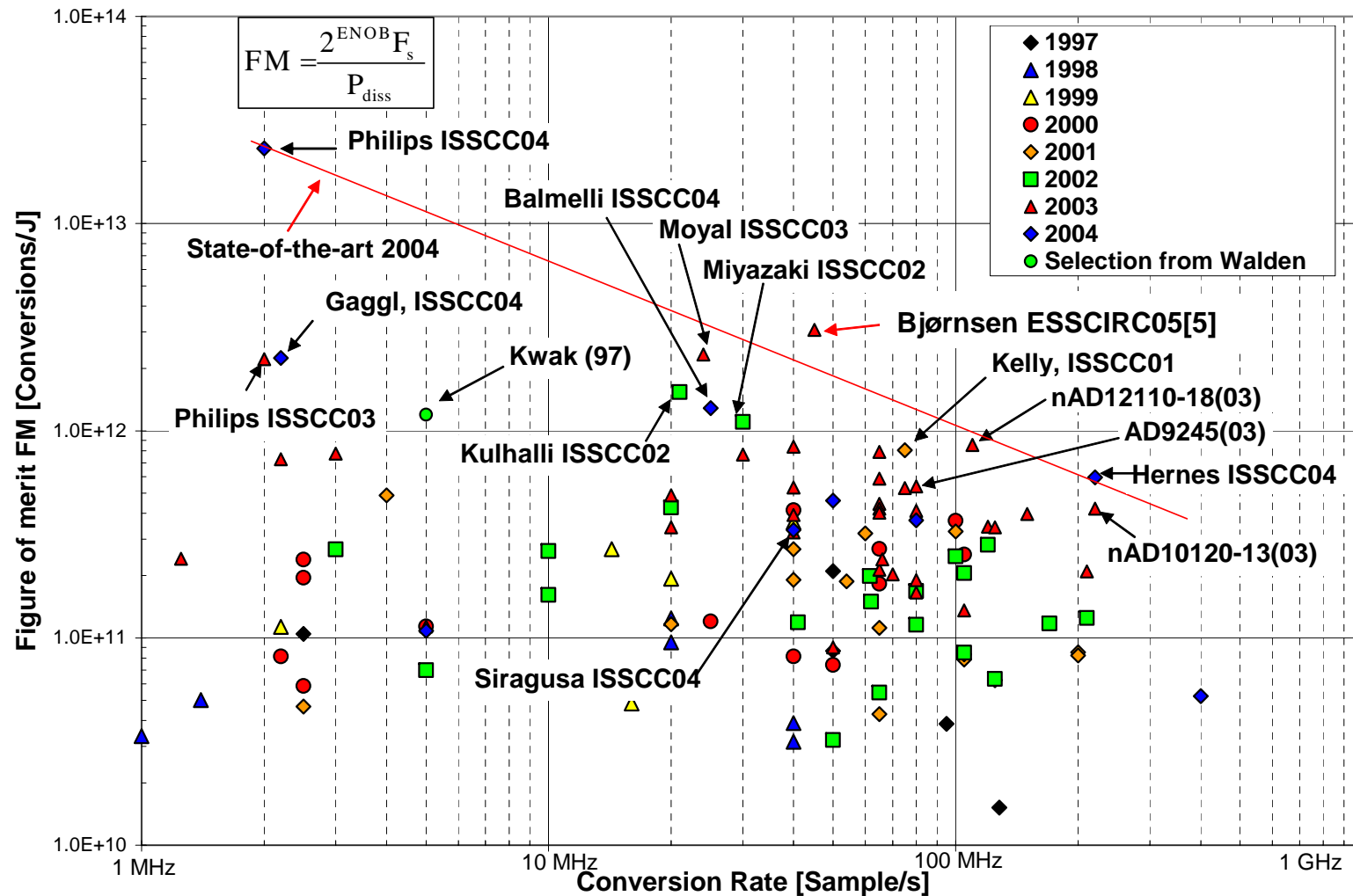
- Hence, one important question becomes:

How does η_i and η_v depend on technology?

FOMs based on measurements

- ***FOMs* calculated based on measured performance of CMOS circuits are always larger than values obtained using (7) due to one or more of the following:**
 - ***DR* limited by distortion**
 - **Additional noise sources (e.g. flicker noise)**
 - **Clock power in switched-capacitor circuits**
 - **Poor current efficiency of MOS transistors (it is even bias dependent)**
 - **Parasitic capacitances**
 - **Matching requirements**
 - **Smaller mismatch => larger dimensions => larger parasitic capacitors**

Reported FOMs for ADCs



From [5]

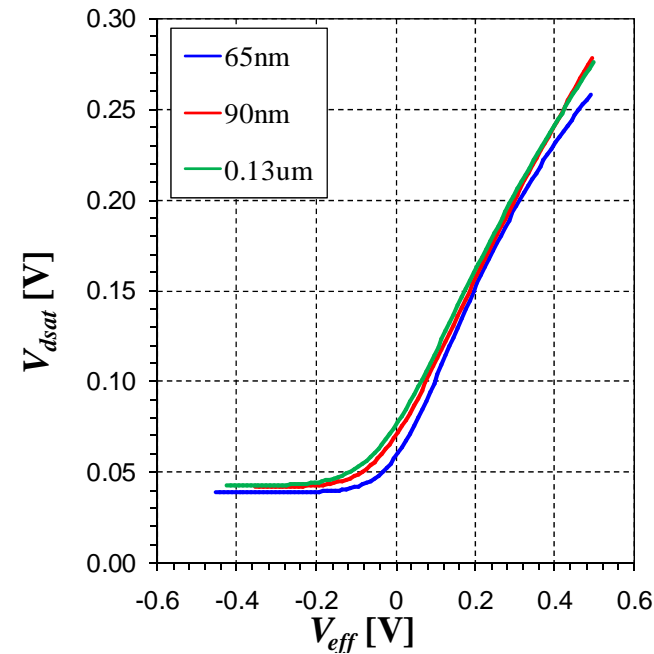
Voltage efficiency (η_v)

- Assume that

$$V_{pp} = V_{DD} - 2V_{dsat}$$

where V_{dsat} is the saturation voltage of our transistors

- To keep η_v unchanged when we scale down technology, V_{dsat} must be scaled down at the same rate as V_{DD} .



Not possible to scale V_{dsat} much lower than 75mV

**η_v expected to decrease
→ *FOM* degrades**

Current efficiency (η_i)

Assuming CMOS implementation:

The gain bandwidth (GBW) product of a single stage OTA is given by

$$GBW = \frac{g_m}{2\pi C} \quad (8)$$

Here g_m is the transconductance of the input transistor and C is the load capacitance.

Writing the square of the noise RMS voltages as:

$$V_{n_rms}^2 = \frac{k_B T}{C},$$

Inserting this equation in (2), solving for C and inserting in (8) yields

$$GBW = \frac{g_m (\eta_v V_{DD})^2}{8\pi k_B T \cdot DR^2} \quad (9)$$

Hence, if DR and g_m is kept constant, the speed of the circuit decreases with the square of the supply voltage.

Scenario 1: Dividing V_{DD} by 2

while keeping GBW and DR :

To keep it simple, assume that η_v does not change. From (9) we note that to keep GBW when supply voltage is reduced by a factor of 2, g_m has to be increased by a factor of 4.

It is common to change the W/L ratio and the drain current by the same relative amount in order to keep V_{eff} unchanged (adding devices in parallel). Hence, to increase g_m by a factor of 4, bias current and W/L are increased by a factor of 4. If we assume that all stages of the amplifier are scaled in the same way, the current consumption will be increased by a factor of four.

→ **Power consumption doubles**

→ η_i **decreases** → **FOM degrades**

Downscaling will not help

Scenario 2: Increasing speed (1)

while keeping DR and V_{DD} :

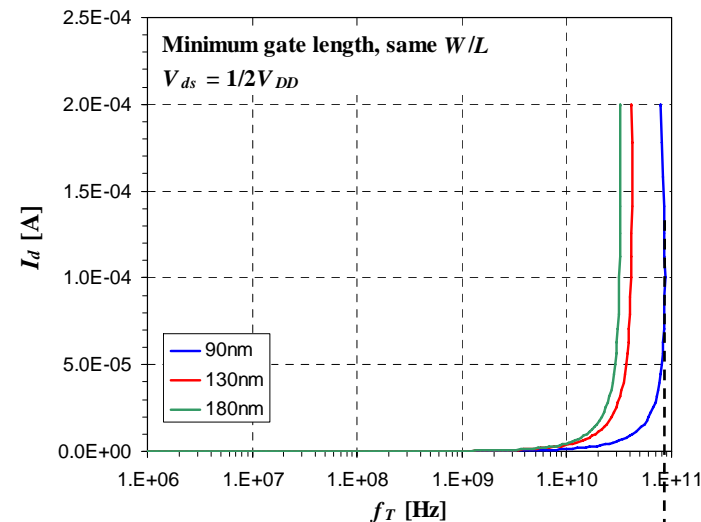
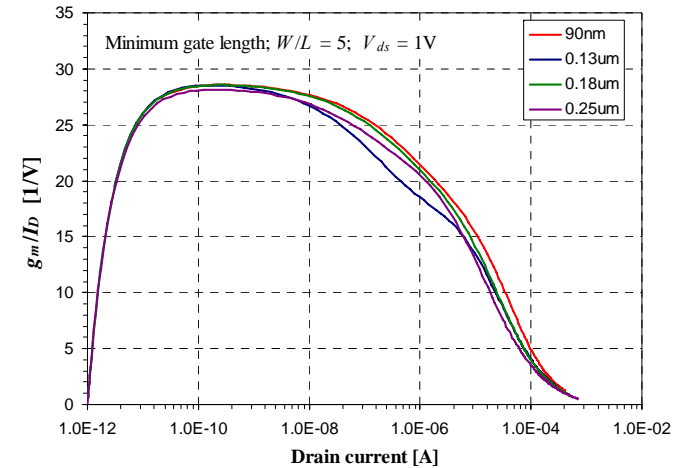
Assume that the frequency of the first non-dominant pole is given by (1) and that phase margin requirements forces a fixed ratio between f_T and GBW .

Assume that the required speed is such that we have to push the transistor deep into strong inversion.

→ η_i decreases

→ FOM degrades

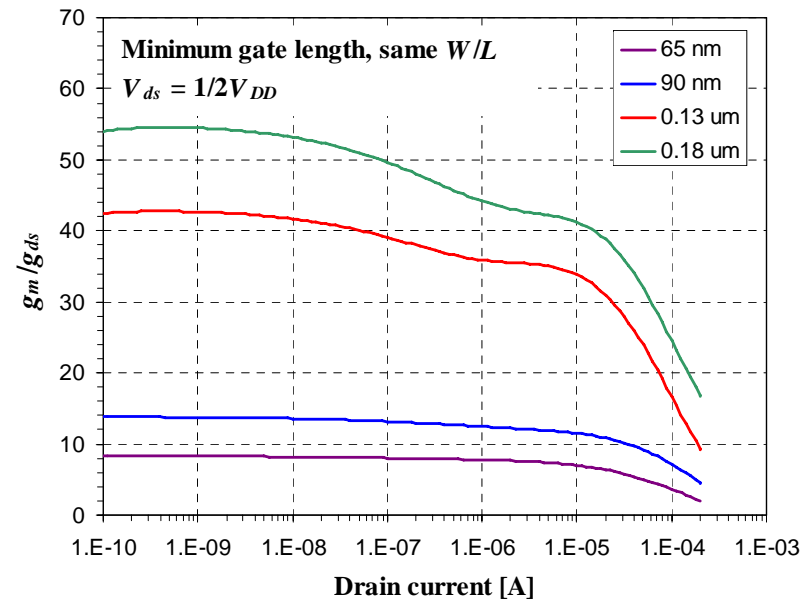
Downscaling will probably help



f_{Tmax} @ 90nm

Scenario 2: Increasing speed (2)

- Watch out for gain degradation when increasing speed
- If speed increase is obtained by increasing V_{eff} , gain will decrease



→ η_i decreases

→ FOM degrades

Downscaling will not help

Maximizing FOM

Some guidelines for maximizing *FOM* in nanoscale CMOS circuits:

- **Circuit level:**
 - Maximize η_v by using as low V_{eff} as possible
 - Identify f_{Tmax} in your technology. Stay well below this frequency to maximize η_i^*
 - Use as high supply voltage as possible (maybe even higher than allowed by the foundry)
- **Architecture level:**
 - Avoid high gain requirements (use, for example, gain calibration)
 - Go to interleaved architectures if speed requirements cause you to move dangerously close to f_{Tmax}^*

*or switch to a finer technology

Summary

- **Back to near constant voltage scaling (at least for a while) 😊**
- **Back to near constant voltage scaling 😞**
 - **Reliability issues: Have to verify circuits versus age**
- **Speed of transistors 😊**
- **SNR/DR 😞**
- **Mismatch 😊**
- **Linearity 😞**
- **Potential for FOM improvement in some circuits by scaling down technology* 😊**

*One example is ADCs where transistors are biased in regions of low current efficiency

List of symbols

Parameter	Description
A_{gate}	Gate area
β	$\mu c_{ox} W/L$
C_{gs}	Gate-source capacitance
C_{gd}	Gate-drain capacitance
C_{db}	Drain-bulk capacitance
c_{ox}	Oxide capacitance per gate area
C_{ox}	Oxide capacitance
f	Frequency
g_m	Transconductance
g_{ds}	Channel conductance
I_d	Drain current
k_B	Boltzmann's constant
L	Gate length
L_{min}	Minimum gate length

Parameter	Description
μ	Mobility
n	Subthreshold ideality factor
T	Absolute temperature
t_{ox}	Oxide thickness
V_{DD}	Supply voltage
V_{ds}	Drain-source voltage
V_{eff}	$V_{gs} - V_T$
V_{gs}	Gate-source voltage
v_s	Saturation velocity
V_T	Threshold voltage
W	Gate width

References

- [1] E.A. Vittoz, “Low Power Design: Ways to Approach the Limits”, *Solid-State Circuits Conference, 1994. Digest of Technical Papers. 41st ISSCC*, pp. 14-18, Feb. 1994.
- [2] K. Bult, “Analog Design in Deep Sub-Micron CMOS,” in *Proc. ESSCIRC 2000*, pp. 11-17.
- [3] A. Annema *et al.*, “Analog Circuits in Ultra-Deep-Submicron CMOS,” *IEEE J. of Solid-State Circuits*, vol. 40, no. 1, Jan. 2005.
- [4] E.A. Vittoz, “Low-power Low-Voltage Limitations and Prospects in Analog Design”, in *Analog Circuit Design*, Editors R.S. van de Plaasche, W.M.C. Sansen, and J.H. Huijsing, Kluwer Academic Publishers, 1994.
- [5] J. Bjørnsen, Ø. Moldsvor, T. Sæther, T. Ytterdal, “A 220mW 14b 40MSPS Gain Calibrated Pipelined ADC,” in *Proc. European Solid-State Circuits Conference (ESSCIRC) 2005*, Grenoble, Sept. 12-14, pp. 165-168, 2005.