Analog Design in ULSI CMOS Processes

Giovanni Anelli

CERN, 1211 Geneva 23, Switzerland
Giovanni.Anelli@cern.ch

Abstract

This paper discusses some of the advantages and of the disadvantages of using a CMOS process in the 180 – 100 nm range for the design of analog blocks in mixed-mode integrated circuits.

I. INTRODUCTION

The microelectronics industry is moving to more and more downscaled processes, driven by the communication and computer markets. People developing Application Specific Integrated Circuits (ASICs) for High Energy Physics (HEP) experiments have a strong motivation to follow this technological trend, because ULSI processes have the following advantages:

- these are the processes which will be certainly available in the future;
- they have extremely good performance for digital circuits;
- they are intrinsically more radiation (Total Ionizing Dose) tolerant than their less advanced counterparts.

On the other hand, analog design becomes more difficult, since the power supply voltages become smaller and smaller.

This paper focuses on some of the issues which are encountered by an analog designer when he wants (or has to!) use a ULSI process for the design of analog circuit blocks. Section II briefly illustrates the concept of scaling, sections III and IV deal respectively with the impact of scaling on the performance of a single transistor and an analog circuit, section V shows an example of the options available for analog design in a typical 130 nm process.

II. THE CONCEPT OF SCALING [1, 2, 3, 4, 5, 6]

In the following we assume, if not otherwise stated, to apply the rules for *constant field scaling* to the transistor. The gate oxide thickness (t_{ox}), the gate length (L), the gate width (W) and the depletion region widths within the device (x_D) are all scaled down by a factor $\alpha > 1$. This increases the gate capacitance per unit area by a factor α but decreases the gate capacitance of each transistor. To keep the electric fields inside the device constant all the bias voltages (and therefore the power supply voltage V_{DD}) have to scale down by the same factor α.

The impact of scaling on digital circuits is extremely beneficial: smaller transistors improve the density, go faster and at the same time consume less power (the power-delay product for digital circuits improves indeed by a factor α^3). But it has to be stressed that the dynamic power dissipation per unit area does not scale.

In reality, industry does not follow *constant field scaling* but a more complex scaling scenario called *generalized scaling*, in which the transistor dimensions are scaled down as in the constant field scaling case, but the voltages are scaled down at a slower pace, to limit the subthreshold currents increase and to have a less constrained voltage swing for the signals.

Several of the trends which will be highlighted in the following are deduced applying the constant field scaling rules, but similar trends can be found in the generalized scaling case.

III. SCALING IMPACT ON TRANSISTOR PERFORMANCE

In this section analyze the impact of scaling on some of the characteristics of a single MOS transistor. How many electrons can we squeeze out of these tiny devices?

A. Transconductance

In strong inversion, the transconductance g_m is given by

$$g_m = \sqrt{2\frac{\beta}{n}I_{DS}} \tag{1}$$

where n is the subthreshold slope factor, I_{DS} is the transistor bias current and β is

$$\beta = \mu C_{ox}\frac{W}{L} \tag{2}$$

where μ is the electron (or hole) mobility, C_{ox} is the gate capacitance per unit area and W and L are the width and the length of the transistor, respectively. C_{ox} is inversely proportional to the gate oxide thickness t_{ox}, and therefore it increases scaling down a process.

Scaling seems therefore to have a beneficial effect on the transconductance, since it increases C_{ox}.

The advantage given by the increase in the μC_{ox} product is unfortunately partially diminished by the fact that the range of currents for which a transistor works in strong inversion becomes smaller, and the transistor enters sooner the velocity

saturation region. Figure 1 shows the g_m/I_{DS} ratio as a function of I_{DS} for two transistors 10 µm wide and 0.13 µm long, one NMOS and one PMOS. The weak inversion (WI), strong inversion (SI) and velocity saturation (VS) regions can be identified in this log-log graph by means of the slope of the curves, which is respectively 0, - 1/2 and – 1. The three slopes are indicated in the plot by segments. It can be seen here that the width of the WI region covers several decades of current, and that the transition from WI to VS through the SI region takes place very quickly. When the transistor enters the VS region, increasing the current (and therefore the power dissipation) does not practically increase the transconductance anymore.

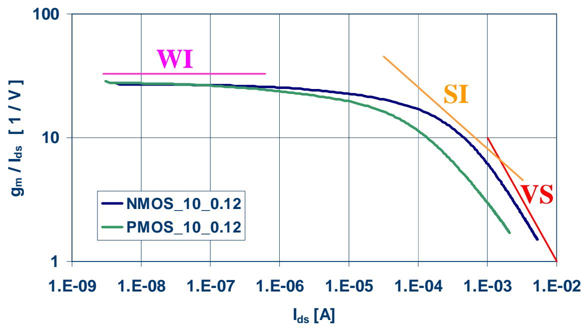

Figure 1: Measurement of the power efficiency of a NMOS and a PMOS transistor as a function of the bias current in a 130 nm process.

Nevertheless, it can be seen from the two curves above that even with a small device (10 µm/ 0.13 µm) remarkably high values of transconductance can be obtained.

B. *Output conductance, intrinsic gain*

Another very important parameter for analog design is the transistor output conductance. The drain-to-source current I_{DS} in an MOS transistor working in strong inversion and in saturation is

$$I_{DS} = \frac{\beta}{2n}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \qquad (3)$$

where V_{GS}, V_{DS} and V_T are the gate-to-source, drain-to-source and threshold voltages, respectively, and λ is given by

$$\lambda = \frac{1}{V_{DS} - V_{DS_SAT}} \cdot \frac{\Delta L}{L - \Delta L} \approx \frac{1}{V_{DS} - V_{DS_SAT}} \cdot \frac{\Delta L}{L} \qquad (4)$$

where V_{DS_SAT} is the drain-to-source voltage needed for the transistor to work in saturation and ΔL is the distance between the pinch-off point and the drain diffusion. As a first approximation, ΔL can be expressed as

$$\Delta L \approx \sqrt{\frac{2\varepsilon_{Si}}{qN_a}(V_{DS} - V_{DS_SAT})} \qquad (5).$$

In eq. (5), ε_{Si} is the dielectric constant of silicon, q is the electron charge and N_a is the doping density in the channel region. The output conductance g_{out} of the transistor, defined as the slope of the I_{DS} vs V_{DS} characteristic in the saturation region, is therefore given by

$$g_{out} = \frac{\partial I_{DS}}{\partial V_{DS}} = \lambda \cdot I_{DS_SAT} \qquad (6)$$

where I_{DS_SAT} is the transistor drain-to-source current flowing when $V_{DS} = V_{DS_SAT}$.

The inverse of the output conductance is called output resistance r_0.

In analog design, we would always like to have the possibility to have the smallest possible output conductance (or, conversely, the highest output resistance). This means that λ should be as small as possible. As we can see from (4), increasing the gate length is an easy way to decrease λ (but also to slow down the circuit!).

The product g_m*r_0, called *intrinsic gain*, is an extremely useful figure of merit to understand the impact of scaling on the analog performance of a single transistor. The intrinsic gain represents the maximum voltage gain which can be obtained from a single transistor.

How does the intrinsic gain vary with scaling? This depends also on the way we decide to size the transistor. In fact, a designer is interested not only in understanding what is the impact of changing the technology but also in what happens with different scaling options for the transistor size (W, L), which is something the designer can (freely) choose in a given process.

Table 1 shows how several quantities vary when scaling a technology by a factor $\alpha > 1$. Each row represents a different scaling choice for the transistor dimensions.

Table 1: Intrinsic gain scaling.

W	L	β	g_m	I_{DS_SAT}	r_0	g_m*r_0
$1/\alpha$	$1/\alpha$	α	1	$1/\alpha$	1	1
1	$1/\alpha$	α^2	α	1	$1/\alpha$	1
$1/\alpha$	1	1	$1/\alpha$	$1/\alpha^2$	α	α
1	1	α	1	$1/\alpha$	α	α
α	$1/\alpha$	α^3	α^2	α	$1/\alpha^2$	1

The first row of Table 1 shows what happens when both W and L are scaled down by α. In the second row only L is scaled, in the third only W, in the fourth row the device dimensions are left unchanged. The fifth and last row of the table illustrates a special case which will be further discussed in section III.

From Table 1 we can notice that the way the intrinsic gain scales depends only on the scaling of the gate length! If we do not scale the gate length we have an improvement in the intrinsic gain by a factor α, if we scale L by the same factor the intrinsic gain does not change. We can write this as a pseudo-formula

$$g_m * r_0 \propto \alpha * L \qquad (7).$$

The above exercise has been done assuming that the non-zero output conductance is caused mainly by the Channel

Length Modulation (CLM) effect (and not Drain Induced Barrier Lowering – DIBL -), that the transistor is working in strong inversion and that we apply the constant field scaling rules to the technology. It can be shown that (7) still holds even dropping the above mentioned assumptions.

C. Gate leakage

One important effect of scaling on the transistor geometry is the reduction of the gate oxide thickness. This has a quite severe impact on the current which can tunnel through the gate oxide. In the 180 – 100 nm range the gate oxide thickness is between 3.5 nm and slightly less than 2 nm.

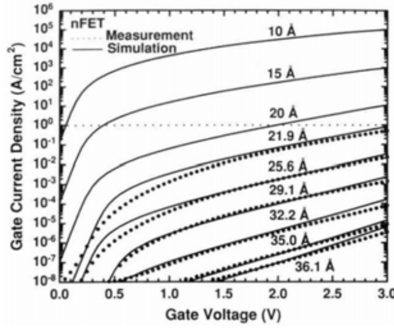

Figure 2: Gate current density as a function of the gate bias and of the oxide thickness.

The gate current has a detrimental effect on digital circuits, because it severely increases the static power dissipation. For analog circuits, one of the major issues is the shot noise associated to this current.

For processes down to 100 nm the gate leakage is still, in my opinion, a manageable problem, but it will become a more serious problem in the future, unless new gate dielectric materials are introduced.

D. Noise and matching

The two most important sources of noise in an MOS transistor are the 1/f and the white noise. The input referred voltage noise for these two sources can be expressed as

$$\frac{\overline{v_{in}^2}}{\Delta f} = 4kTn\gamma\frac{1}{g_m} + \frac{K_a}{C_{ox}WL}\frac{1}{f^{\alpha}} \quad (8)$$

where k is the Boltzmann's constant, T the absolute temperature, n the subthreshold slope factor, γ a parameter which ideally should be comprised between 1/2 (WI) and 2/3 (SI), g_m the gate small-signal transconductance, K_a the technology-dependent 1/f noise parameter, $C_{ox}WL$ is the gate capacitance, f is the frequency and α is a parameter usually close to 1.

For the same device dimensions and the same bias current, scaling seems to have a positive impact on the white noise, as we can expect to have an improvement in the transconductance. In reality, one must be careful since several effects linked to the very short channel can generate a noise somewhat higher than what is foreseen by (8).

For what concerns the 1/f noise, for the same device dimensions scaling increases the gate capacitance, and therefore the noise should decrease with scaling if we assume the parameter K_a to be constant. This parameter varies in fact quite considerably from process to process. Figure 3 illustrate this spread for several processes and several technology nodes. It is quite difficult to extract a trend from figure 3, apart from the well known difference between NMOS and PMOS transistors. It seems, from these data, that reducing the gate oxide thickness does not necessarily increase K_a in well mastered processes.

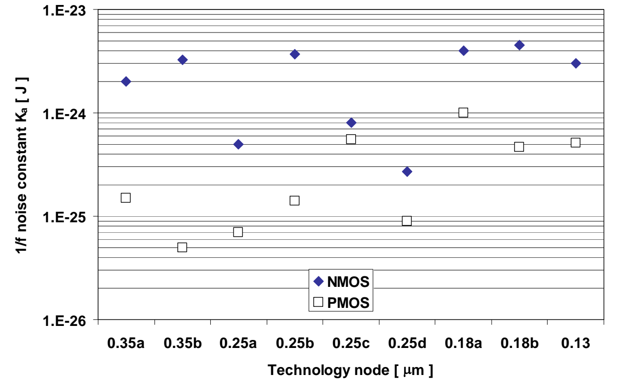

Figure 3: 1/f noise parameter K_a taken from several different literature sources, except for the 0.13 node and one of the 0.25 nodes, which are measurement done at CERN.

The differences (mismatch) between the threshold voltages of two identically designed transistors, biased in the same way and having the same environment follow a Gaussian distribution, and the standard deviation is given by [7]

$$\sigma_{\Delta V_{th}} = \frac{A_{V_{th}}}{\sqrt{WL}} \quad (9)$$

where WL is the transistor area and A_{Vth} is a technology-dependent parameter (matching parameter), generally different for NMOS and PMOS. A similar expression holds for the mismatch of β (eq. 2).

The threshold voltage mismatch is caused by several sources: the statistical fluctuation of the number of dopant atoms in the channel of the transistor [8] is the most important. The component of A_{Vth} related to this phenomenon is given by

$$A_{V_{th},N} = \sqrt{2}\cdot C\cdot\frac{t_{ox}}{\varepsilon_{ox}}\cdot\sqrt[4]{N} \quad (10)$$

where C is a scaling-independent constant, t_{ox} is the gate oxide thickness, ε_{ox} is the oxide dielectric constant and N is the dopant density in the transistor channel. With scaling, t_{ox} decreases and N increases (but at a slower pace), and therefore $A_{Vth,N}$ (and A_{Vth}) decrease.

This means that <u>for the same device dimensions</u> scaling has a beneficial impact on transistor threshold voltage matching. On the other hand, for minimum size transistors

mismatch will become more important (since the number of dopant atoms in the channel becomes smaller).

Concerning the mismatch of β (eq. 2), there is not a clear trend about what will happen in the future.

IV. Scaling Impact on Analog Circuit Performance

This section focuses on some of the problems encountered in the design of analog blocks in mixed-mode ICs fabricated in a ULSI process.

A. *Power, speed, SNR*

Table 2 shows what happens applying the constant field scaling rules to the power consumption, the speed and the Signal-to-Noise Ratio (SNR), for several different choices of the device dimensions (the same as in Table 1).

The power consumption is defined as $PWR = I_{DS_SAT} * V_{DD}$, the speed is defined as $1/\Delta t = I_{DS_SAT} / (C_g * V_{DD})$, the Noise is the white noise (square root of the first term in equation 8) and the SNR is here very simply given by $SNR = V_{DD}$ / Noise. These definitions are not general, and we saw that constant field scaling is not the scaling procedure used today, but it is nevertheless interesting to see what happens to the above defined quantities with scaling.

The first four rows in Table 2 indicate that there is a saving in power with scaling, and a gain in speed when the gate length is scaled. On the other hand, the SNR is also reduced, since the signal (assumed here as a voltage) scales to fit in the reduced power supply, but the noise does not scale at the same rate. To keep the SNR constant with scaling we have to scale down L but increase the W (fifth line of the table). This still gives a gain in speed, but there is no gain in power (and intrinsic gain) any longer. As we have already said, this consideration is not general, but it holds for many different analog circuits. Moreover, a more precise analysis [9] shows that in reality things become even worse going to ULSI processes, and the power dissipation increases to keep the SNR constant.

Table 2: Impact of scaling by a factor α and as a function of different transistor dimensions (W and L) on power consumption (PWR), transistor gate capacitance (C_g), speed ($1/\Delta t$), white noise (Noise) and Signal to Noise ratio (SNR).

W	L	PWR	C_g	$1/\Delta t$	Noise	SNR
$1/\alpha$	$1/\alpha$	$1/\alpha^2$	$1/\alpha$	α	1	$1/\alpha$
1	$1/\alpha$	$1/\alpha$	1	α	$1/\alpha^{1/2}$	$1/\alpha^{1/2}$
$1/\alpha$	1	$1/\alpha^3$	1	$1/\alpha$	$\alpha^{1/2}$	$1/\alpha^{3/2}$
1	1	$1/\alpha^2$	α	$1/\alpha$	1	$1/\alpha$
α	$1/\alpha$	1	α	α	$1/\alpha$	1

B. *Low voltage issues*

One of the most important issues for analog design in ULSI processes is the dramatic reduction in the power supply voltage, which complicates or renders impossible the use of many architectures employed in less advanced CMOS technologies. Very far for making an exhaustive discussion of the problem here, I will just give some hints about the possible solutions which can be used when V_{DD} drops below 1.5 V:

- Use of PMOS and NMOS differential pairs in parallel to obtain rail-to-rail op-amps;
- Low V_{DS_SAT}'s can be obtained increasing, for a given current and technology, the W/L ratio of a transistor. This has the detrimental effect of reducing the speed;
- Use of low-V_T or zero-V_T transistors;
- Use of multi-gain system for high dynamic range;
- Use devices working in weak inversion. This has the double advantage of having low V_{DS_SAT} voltages and the highest power efficiency (g_m/I_{DS} ratio);
- Current mode architectures are suitable for low-voltage architectures, and should be further developed.

C. *Substrate noise [10, 11, 12, 13]*

The main push for analog design in ULSI processes is the possibility of having a complex system made up by analog and digital blocks coexisting in one single silicon die. While analog design becomes trickier in advanced CMOS processes, digital circuits profit enormously from the device scaling.

One serious problem which comes from the common substrate for analog and digital blocks is that the noise generated from the digital blocks (digital noise) can propagate to the sensitive analog blocks. This can happen mainly through three channels:

1. Power and ground interconnection lines;
2. Parasitic capacitance between interconnection lines;
3. Common substrate (in this case we talk about substrate noise). Any noise in the substrate couples in the transistors through the bulk transconductance and the parasitic capacitances between the substrate and the source, drain and channel.

The first two problems are relatively easy to solve: special care must be used for the layout of sensitive lines (shielding can be used, exploiting the several metal levels available), and digital blocks and analog blocks should have separate power lines.

The problem coming from the common substrate is much more difficult to solve, and what can be done depends a lot on the kind of substrate available. For high resistivity substrates, guard rings around the analog and digital blocks (biased separately) and separation between the blocks can help reducing the problem. In the case of low resistivity substrates this is by far less effective. Most (if not all!) foundries have high resistivity substrates for sub 180 nm processes, which as just explained helps in fighting the substrate noise problem.

As a conclusive remark for this section, it should be stressed that, since none of the above techniques is 100 % effective, great care should be taken choosing the architecture and designing the analog blocks. In particular, fully differential architectures should be used, and parameters like Power Supply Rejection Ratio (PSRR) and Common Mode Rejection Ratio (CMRR) should be maximized.

V. Process Options for Analog Design

There are several different manufacturers offering CMOS technologies, but generally comparing the characteristics and the several options available for processes of the same generation show that they are all quite similar. An example of some of the typical characteristics for 130 nm follows:

- 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;
- Multiple gate oxide thicknesses (and therefore multiple supply voltages);
- Several different metal options (number of layers, copper or aluminium);
- Different kind of resistors: diffusions, polysilicon, metal;
- Triple well NMOS (allows having the bulk contact separate from the common substrate);
- Low-V_T, High-V_T, Zero-V_T devices (both with thin and thick oxides);
- Metal-to-metal capacitors;
- Electronic fuses;
- Inductors.

Between all the above mentioned points, the most beneficial for analog design are the possibility of having multiple power supply voltages and multiple threshold voltages.

VI. 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 many analog circuits 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. 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.

VII. Acknowledgements

I would like to thank Roberto Dinapoli, 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 many useful discussions.

VIII. References

[1] B. Davari et al., "CMOS Scaling for High Performance and Low Power - The Next Ten Years", Proc. of the IEEE, vol. 87, no. 4, Apr. 1999, pp. 659-667.

[2] Y. Taur et al., "CMOS Scaling into the Nanometer Regime", Proc. of the IEEE, vol. 85, no. 4, Apr. 1997, pp. 486-504.

[3] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, 1998, p. 186.

[4] T. N. Theis, "The future of interconnection technology", IBM Journal of Research and Development, vol. 44, no. 3, May 2000, pp. 379-390.

[5] G. A. Sai-Halasz, "Performance trends in high-end processors", Proceedings of the IEEE, vol. 83, no. 1, January 1995, pp. 20-36.

[6] D. J. Frank et al., "Device Scaling Limits of Si MOSFETs and Their Application Dependencies", Proc. IEEE, vol. 89, no. 3, March 2001, pp. 259-288.

[7] M.J.M. Pelgrom et al., "Transistor matching in analog CMOS applications", Technical Digest of the International Devices Meeting 1998, pp. 915-918.

[8] P.A. Stolk et al., "Modeling Statistical Dopant Fluctuations in MOS Transistors", IEEE Trans. Elect. Dev., vol. 45, no. 9, Sept. 1998 , pp. 1960-1971.

[9] A.-J. Annema, "Analog Circuit Performance and Process Scaling", IEEE Transactions on Circuit and System II, vol. 46, no. 6, June 1999, pp. 711-725.

[10] A. Samavedam et al., "A Scalable Substrate Noise Coupling Model for Design of Mixed-Signal IC's", IEEE JSSC, vol. 35, no. 6, June 2000, pp. 895-904.

[11] N. K. Verghese and D. J. Allstot, "Computer-Aided Design Considerations for Mixed-Signal Coupling in RF Integrated Circuits", IEEE JSSC, vol. 33, no. 3, March 1998, pp. 314-323.

[12] M. Ingels and M. S. J. Steyaert, "Design Strategies and Decoupling Techniques for Reducing the Effects of Electrical Interference in Mixed-Mode IC's", IEEE Journal of Solid-State Circuits, vol. 32, no. 7, July 1997, pp. 1136-1141.

[13] N. K. Verghese, T. J. Schmerbeck and D. J. Allstot, "Simulations Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits", Kluwer Academic Publishers, Boston, 1994.