Future Trends in Microelectronics – Impact on Detector Readout

P. O'Connor
Brookhaven National Laboratory, Upton, NY 11973, USA

Mainstream CMOS is now a well-established detector readout technology. We review technology scaling trends and limits, the implementation of analog circuits in digital CMOS processes, and radiation resistance. Emphasis is placed on the growing importance of power dissipation in ultra-scaled technologies.

1. CLASSICAL CMOS SCALING

The essential aspects of CMOS digital circuit behavior can be understood by examining the circuit of Fig. 1.

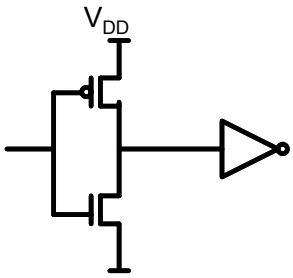

Figure 1 Illustrative CMOS digital circuit (inverter loaded by inverter).

The PMOS and NMOS here act as switches, charging and discharging the capacitance of the output node. The attractive features of this gate are:

- the switches work on complementary polarity, so a common control voltage can be used;
- the circuit is quite tolerant of supply voltage variation;
- most importantly, there is no power consumed except while switching is in progress.

In the early 70's it was recognized that the MOSFET was easily and predictably scaled to smaller lithographic dimensions provided a simple scaling rule was followed [1]. Because it is a majority carrier device, the MOSFET's current is almost pure drift; therefore the current density is proportional to electric field. Hence if one scales all dimensions and voltages by the same factor (α) the electric field and current density remain constant and the DC characteristics are unchanged. When this scaling rule is applied to a logic circuit one finds easily that transistor density goes up by α^2, speed goes up by α, and power density remains constant. Industry has capitalized on this scaling behavior, adopting a model where a new process generation is introduced into production every 2 years, with $\alpha \cong \sqrt{2}$. With each generation density doubles, speed increases by 40%, and performance – defined as the number of switching operations per unit area, time, and power – goes up by almost a factor of three.

Manuscript authored by employees of Brookhaven Science Associates, LLC under contract DE-AC02-98CH10886 with the U.S. Department of Energy.

Table 1 Classical constant-field scaling rules and consequences.

Voltages	V_{th}, V_{DS}	$1/\alpha$
Dimensions	Lateral and vertical	$1/\alpha$
Electric field	V/L, V/d	const.
Conductance	I/V	const.
Capacitance		$1/\alpha$
Speed	I/CV	α
Switching energy	CV^2	$1/\alpha^3$
Power/gate	CV^2f	$1/\alpha^2$
Transistor density	$1/L^2$	α^2
Power density		const.

This simple scaling relationship and ongoing improvements in photolithography quickly caused CMOS to become the dominant digital technology, and bore out Moore's 1965 prediction [2]. Figures 2-5 illustrate the trend over the roughly 14 generations of CMOS that have been put into production.

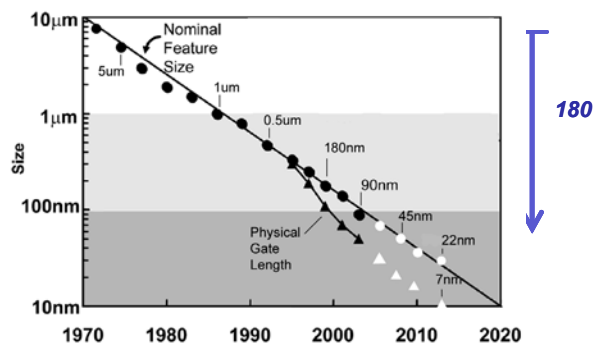

Figure 2. Feature size scaling trend. Black circles: actual mask dimensions in production as of 2005. White circles: planned future generations. Triangles: physical gate length.

Note that because of the etching processes used to fabricate the transistor gates, the physical gate length is less than the drawn feature size. Below 200nm this causes "super-scaling" of the physical gate length.

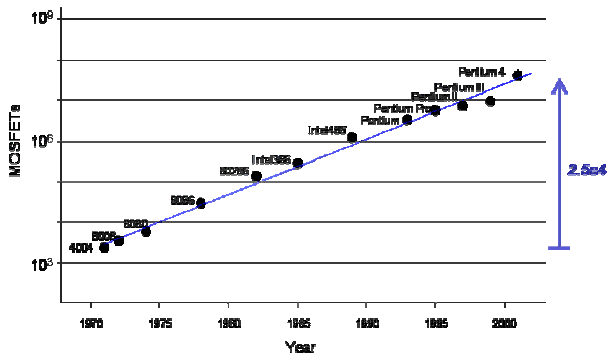

Figure 3. Transistors per chip for Intel© microprocessors. Reflects feature size shrink as well as growth in chip size. Large memory chips in 2006 integrate roughly 10^9 MOSFETs.

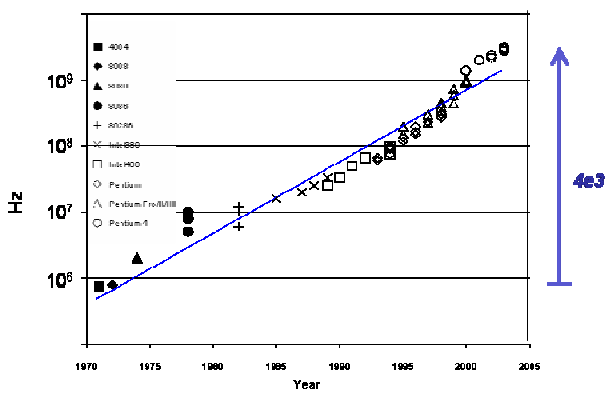

Figure 4. Clock frequency trend for Intel© microprocessors.

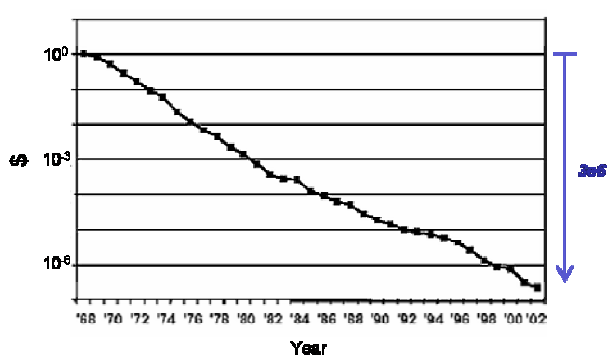

Figure 5. Cost per transistor. Roughly 10^{17} MOSFETs are produced per year in 2005.

2. DEPARTURE FROM CLASSICAL SCALING

Constant-field scaling rests on several simplified assumptions. In the early stages of CMOS technology scaling it was acceptable to overlook:

- Thermal energy of carriers
- Mobility degradation and velocity saturation
- Fringing capacitances
- Tunneling through the gate oxide
- Atomistic effects: discrete, random dopant and trap distribution, step-height oxide thickness variation, and line edge roughness
- Industry economics

After 30 years of exponential growth, these oversimplifications are no longer accurate. In particular, the effects of finite temperature and of tunneling have raised serious obstacles to continued scaling.

2.1. Thermal energy of carriers

The MOSFET is an effective switch as long as its conductance is effectively zero in the off state, while at the same time being capable of high on-state current to drive load capacitances. In the off state the gate is grounded and the residual current is proportional to $\exp(-qV_{th}/nkT)$, where V_{th} is the threshold voltage. According to classical constant-field scaling V_{th} as well as the supply voltage V_{DD} should scale down with the feature size. However, the thermal voltage kT/q does not scale, and so continued scaling of the threshold voltage produces significant off-state leakage current. This is related to the reduction in V_{DD}; the on-state current I_{on} is roughly proportional to $(V_{DD}-V_{th})$. Hence there is a tradeoff between speed and static power dissipation. Modern CMOS processes now offer multiple thresholds to allow designers more control over the speed-static power tradeoff. Fig. 6 shows the I_{on}-I_{off} characteristics of a batch of 90nm NMOS transistors intentionally fabricated to have a range of threshold voltages.

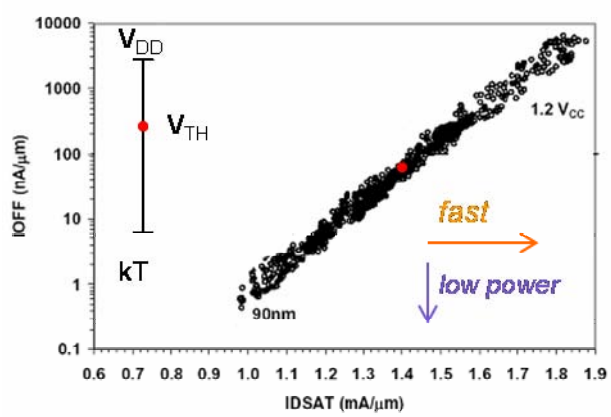

Figure 6. I_{off}-I_{on} characteristics of 90nm NMOS transistors showing effect of threshold voltage variation.

2.2. Oxide tunneling

When the gate insulator gets too thin carriers can tunnel between gate and channel. The tunneling current is highly sensitive to oxide thickness, increasing about 100X per generation. There is intensive effort to find a replacement material for the gate dielectric, one with a higher permittivity which would allow the same control

capacitance with a physically thicker insulating layer. Unfortunately the candidate materials to date all have difficulty integrating into the CMOS process flow. Many modern processes offer thick-oxide transistor options, mainly to create I/O circuits for off-chip communication and for some analog circuitry.

2.3. Scaling in practice

To adapt to these physical realities, the industry has adopted a modified scaling which has allowed a continued increase in transistor density and clock speed. Until the 180nm generation classical scaling was more or less followed with $\alpha=\sqrt{2}$ and 2.8X performance improvement per generation. Afterwards gate length continued to scale while supply voltage and threshold voltage essentially stopped scaling, having reached the limits dictated by on- and off-state current requirements. Oxide thickness (for the "high-performance" transistors in the process) continued to scale, but at a slower rate than lateral dimensions. As a result the electric fields and channel conductance have increased instead of staying constant, and the switching energy has decreased by $1/\alpha$ instead of $1/\alpha^3$. Consequently the switching *power density* has increased substantially. Moreover there has been a dramatic increase of *static* power consumption from subthreshold leakage and gate tunneling currents (Fig. 7), compounding the power dissipation problem.

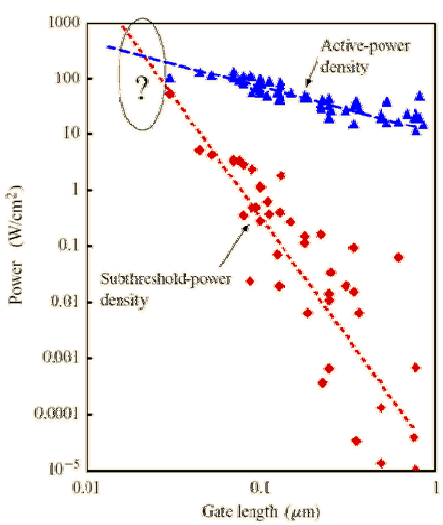

Figure 7. Power density trend caused by departure from constant-field scaling [3].

An additional cause of high power dissipation is the need to impose sufficient margin on the supply voltage V_{DD} to compensate for gate-to-gate variability. As transistors shrink down, random dopant fluctuations increase threshold mismatch between gates. The supply voltage must be margined to ensure that the worst-case mismatch still results in a working digital gate. With multi-million gate designs, the cost of supply voltage margining can be significant.

In practice, today's digital designs have become limited by the power budget rather than the number of available gates. Future processes will offer only modest energy scaling and advances in performance will have to come from power-conserving design techniques such as clock and supply gating, dynamic supply voltage adjustment, and architectural improvements. A discussion of these techniques can be found in [4].

2.4. Next-generation transistors

Continued scaling of standard bulk CMOS faces enormous difficulties. Several process and device modifications have been proposed to cope with the problems of threshold scaling, thin gate oxide, and short channel anomalies. Techniques which are nearing production include:

- Strained silicon – introducing uniaxial strain into the silicon lattice can increase effective carrier mobility, thereby allowing higher I_{on} for the same I_{off} and improving the speed-power tradeoff.
- Thin-body devices – by redesigning the MOSFET to have a thin channel with gates on both sides of the channel, or an insulating layer on the back, the electrostatics of the device are improved and short-channel effects are suppressed.
- High-permittivity gate dielectrics – in one example [5] a HfO_2 material was used which suppressed gate tunneling current by a factor of 10^6 compared to SiO_2 at the same effective thickness.
- Si_xGe_{1-x} heterojunction bipolar transistor – an npn transistor which can be integrated with CMOS and has roughly twice the high-frequency performance as a MOSFET at the same power level. This option has been available since the 0.5μm generation and finds frequent use in RF circuits.

3. MIXED-SIGNAL CIRCUITS

CMOS technology is driven by digital applications, which comprise about 90% of the total market, and was initially regarded as unsuitable for analog functions. However, the rapid progress in speed, integration density, and parametric control led designers to work around the drawbacks to produce a class of analog circuits that could piggy-back on low cost, mass-produced digital CMOS. This trend has accelerated as the market has shifted from primarily logic and memory to consumer, communications, and automotive applications which all demand some analog content on chip. Analog and digital functions now support each other as digital calibration and trimming correct for mismatch in analog circuits, and

digital circuits benefit from analog temperature sensors, adaptive clocking, adaptive power supply adjustment, phase locked loops, etc. On-chip power conversion and regulation is now gaining attention as a way to simplify the power delivery problems of big microprocessors.

Unlike logic, analog circuits can't be characterized by a single figure of merit. In general the important technological features for analog circuits are high gain and bandwidth, low noise, and good matching. Present trends in CMOS technology that benefit digital circuits may not have corresponding benefits for analog performance.

3.1. Gain and bandwidth

As gate length goes down, the current gain cutoff frequency f_T goes up proportionately. MOSFET f_T is also a strong function of drain current I_D. Fig. 8 shows calculated f_T values for minimum gate length devices in three scaled CMOS families for a range of power $P=I_D \times V_{DD}$. The high-frequency performance of recent CMOS generations is responsible for the upsurge of interest in RF circuits integrated on low-cost digital processes.

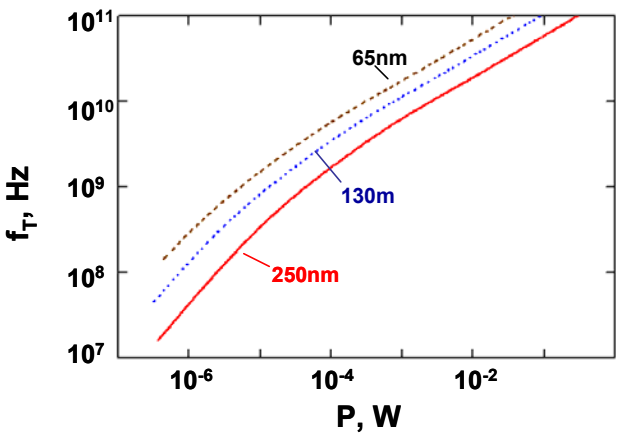

Figure 8. Cutoff frequency f_T vs. power dissipation for minimum gate length devices in three technologies.

Unfortunately the departures from classical scaling change the channel electric field components in such a way that the device output resistance degrades strongly, leading to a fundamental speed/gain tradeoff (Fig. 9). Moreover constraints imposed by the low supply voltage cause gain to degrade with technology scaling even at equal gate length.

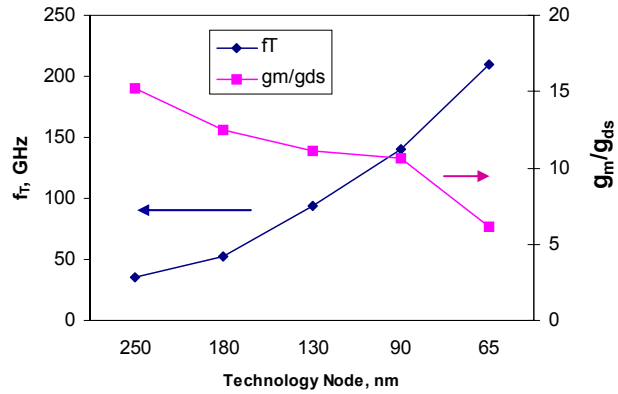

Figure 9. Speed and gain for minimum gate length devices dependence on technology node.

Gain is also compromised by the tunneling current, which changes the device input impedance from capacitive to resistive. The frequency f_{gate} where they cross over shoots up dramatically because of the exponential sensitivity of tunneling to oxide thickness. Over a few generations f_{gate} has gone from an unnoticeable sub-Hz range into the megahertz.

3.2. Noise and dynamic range

MOSFETs have two important noise mechanisms, and for the important case of a charge sensitive amplifier they have to be evaluated relative to the associated gate capacitance. For a unit-capacitance device, the Johnson noise of the channel improves roughly as $1/\sqrt{\alpha}$ due to higher f_T. On the other hand the 1/f noise associated with the interface is technology dependent and doesn't follow a consistent trend from generation to generation. Choosing the MOSFET length and width for optimum noise follows different rules in scaled technology where moderate-inversion operating points are more common [6]. The consequences of scaling for charge amplifiers are mixed. One can expect to achieve lower ENC with a scaled front end in situations where 1/f noise does not dominate, for instance with fast shaping and a low power budget. But the maximum output signal is constrained to stay within the reduced V_{DD} swing. In most cases the V_{DD} loss outweighs the ENC decrease leading to lower overall dynamic range. Examples of predicted noise, expressed as equivalent input noise charge (ENC), as a function of power and technology node are shown in Fig. 10.

Note that the preamplifier's dynamic range is approximately proportional to the ratio of V_{DD} to ENC. Taking the data of Fig. 10 and normalizing to the value of V_{DD}/ENC for the 0.35µm process at 1mW, we find that the relative dynamic range degrades by about 15% per generation (Table 2).

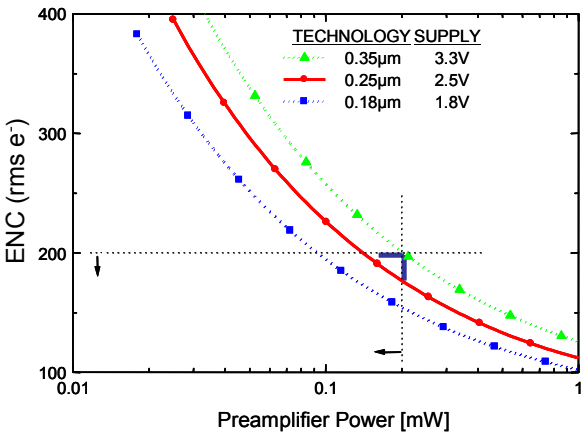

Figure 10. ENC vs. preamplifier power for three CMOS technology generations. Simulated for a detector capacitance of 1pF and shaper peaking time of 1µs.

Table 2 Dynamic range for the three processes of Fig. 10 normalized to the value of V_{DD}/ENC for the 0.35µm process at 1mW power.

PROCESS	VDD	NORMALIZED DYNAMIC RANGE @ POWER		
		1mW	0.2mW	0.05mW
0.35µm	3.3V	**1.00**	0.64	0.38
0.25	2.5	0.84	0.55	0.33
0.18	1.8	0.66	0.45	0.27

For other analog amplifiers it has been shown [7] that a similar relation between power and dynamic range exists, and that lowering the supply voltage increases the power consumption required to maintain the same SNR and bandwidth.

3.3. Transistor mismatch

In highly integrated analog circuits parallel processing is a necessity, and consequently the matching of the parallel paths is important. Unmatched paths lead to performance or yield loss in analog circuits or reduced robustness in digital circuits. Based on random uncorrelated dopant fluctuations, MOSFET threshold voltage mismatch scales as the inverse square root of the device area $(WL)^{-1/2}$ times a technology-dependent factor A_{VT} which in turn scales more slowly than the technology scale factor α [8]. As a result the threshold mismatch of minimum-size devices gets worse with each technology node. As the power supply comes down, designers need to use a larger-area device every generation to preserve the relative matching (σ_{Vth} / V_{DD}). The resulting high input capacitance requires larger load-driving currents, imposing a power penalty.

3.4. Constraints on analog circuit configuration

The low supply voltage headroom found in scaled CMOS processes imposes limits on analog circuit topologies. The increasing ratio of V_{TH}/V_{DD} rules out the use of many classical analog design topologies.

- The cascode connection, useful in providing high gain loads and current sources, becomes unusable once V_{DD} falls below ~1.2V.
- CMOS transmission gates, commonly found in sample/hold and switched capacitor circuits, perform poorly in scaled processes as the control gate voltage is limited and self-discharge rates increase.
- Source followers limit the available signal swing and have to be avoided in output stages and other large-signal circuit blocks.

4. RADIATION TOLERANCE

Thinner gate oxides in scaled CMOS technologies are more tolerant of total ionizing dose (TID), but thick field oxides or shallow trench isolation (STI) allow edge leakage in NMOS transistors and loss of isolation between n-doped regions upon irradiation. To circumvent these effects it has been shown to be effective to use enclosed-geometry gates and guard rings [9].

Single-event upset (SEU) data indicates competing effects of reduced device size and smaller critical charge, leading to inconclusive trends as devices scale. SEU can be mitigated by incorporating redundant or error-correcting logic. Other single-event effects, such as single-event latchup and gate rupture, are not reported for deep submicron CMOS operated at standard voltages.

5. COST AND AVAILABILITY

The increasing sophistication of deep submicron lithography tools and the large number of mask layers in advanced processes have caused a dramatic increase in mask and fabrication costs. Engineering run cost has increased by a factor of about 1.7 per generation and now exceeds $500K for a minimum run of 20 wafers. The design flow for error-free chip layout now requires a suite of complex, expensive tools. Even in technologies that lag the state of the art, the cost of developing a new design may exceed the resources available to smaller research projects. However, foundry support for older processes with mixed-signal emphasis (such as 0.35µm CMOS with analog enhancements) appears to be guaranteed for many years.

6. SUMMARY

- Power dissipation is the major obstacle to further CMOS scaling.

- Foundry and mask costs are increasing dramatically as process options (multi-V_{th}, multi-t_{ox}, HBT, passives, etc.) are added for system-on-chip applications.
- Analog design is compromised by the low supply rail in ultra-scaled CMOS.
- Radiation tolerance to total ionizing dose improves in scaled technologies provided proper layout techniques are used, while single event upset rates must be calculated and mitigated by appropriate error detection and correction.

Acknowledgments

The author thanks V. Radeka, G. De Geronimo, and G. Anelli for stimulating discussions.

This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy.

Disclaimer

References

[1] R. Dennard, F. Gaensslen, H. Yu, V. Rideout, E.Bassous, and A. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE J. Solid-State Circuits SC-9, 256 –268 (1974).

[2] G. E. Moore, "Cramming More Components onto Integrated Circuits", Electronics, vol. 38, no. 8, 1965.

[3] E.J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down", IBM J. Res. & Dev. 46, 169 (2002)

[4] M. Horowitz, E. Alon, D. Patil, "Scaling, power, and the future of CMOS", 2005 IEEE Intl. Electron Devices Meeting Technical Digest p. 9-15.

[5] E. Gusev et al., IBM, Proc. 2004 Intl. Electron Devices Meeting.

[6] P. O'Connor and G. De Geronimo, "Prospects for charge sensitive amplifiers in scaled CMOS", Nucl. Instrum. Methods A480, 713 (Mar. 2002).

[7] A.-J. Annema, B. Nauta, R. Van Langevelde, H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS", IEEE J. Solid-State Circuits 40(1), 132 (2005).

[8] M. Pelgrom, H. Tuinhout, M. Vertregt, "Transistor matching in analog CMOS applications", 1998 IEEE Intl. Electron Devices Meeting Technical Digest p.915-918.

[9] F. Faccio and G. Cervelli, "Radiation-induced edge effects in deep submicron CMOS transistors", IEEE Trans. Nucl. Sci. 52(6), 2413 (2005).