CONTRIBUTED PAPER

Analog Circuit Design in Nanoscale CMOS Technologies

Classic analog designs are being replaced by digital methods, using nanoscale digital devices, for calibrating circuits, overcoming device mismatches, and reducing bias and temperature dependence.

By Lanny L. Lewyn, *Life Senior Member IEEE*, Trond Ytterdal, *Senior Member IEEE*, Carsten Wulff, *Member IEEE*, and Kenneth Martin, *Fellow IEEE*

ABSTRACT | As complementary metal–oxide–semiconductor (CMOS) technologies are scaled down into the nanometer range, a number of major nonidealities must be addressed and overcome to achieve a successful analog and physical circuit design. The nature of these nonidealities has been well reported in the technical literature. They include hot carrier injection and time-dependent dielectric breakdown effects limiting supply voltage, stress and lithographic effects limiting matching accuracy, electromigration effects limiting conductor lifetime, leakage and mobility effects limiting device performance, and chip power dissipation limits driving individual circuits to be more energy-efficient. The lack of analog design and simulation tools available to address these problems has become the focus of a significant effort with the electronic design automation industry. Postlayout simulation tools are not useful during the design phase, while technology computer-aided design physical simulation tools are slow and not in common use by analog circuit designers. In the nanoscale era of analog CMOS design, an understanding of the physical factors affecting circuit reliability and performance, as well as methods of mitigating or overcoming them, is becoming increasingly important. The first part of the paper presents factors affecting device matching, including those relating to single devices as well as local and long-distance matching effects. Several reliability effects are discussed, including physical design limitations projected for future downscaling. In some cases, it may be helpful to exceed foundry-specified drain-source voltage limits by a few hundred millivolts. Models are presented for achieving this, which include the dependence on the shape of the output waveform. The condition $V_{sb} > 0$ is required for cascode circuit configurations. The role of other terminal voltages is discussed, as $V_{sb} > 0$ increases both hot and cold carrier damage effects in highly scaled devices. The second part of the paper focuses on trends in device characteristics and how they influence the design of nanoscale analog CMOS circuits. A number of circuit design techniques employed to address the major nonidealities of nanoscale CMOS technologies are discussed. Examples include techniques for establishing on-chip accurate and temperature-insensitive bias currents, digital calibration of analog circuits, and the design of regulator and high-voltage circuits. Achieving high energy efficiency in ICs capable of accommodating 10^9 devices is becoming critically important. This paper also presents a survey of the evolution of figure of merit for analog-to-digital converters.

KEYWORDS | ADC survey; analog; analog-to-digital converters; CHE; CHISEL; CMOS; design for manufacturing; design for reliability; device matching; device scaling; DFM; DFR; dielectric breakdown; electromigration (EM); excess overdrive (EOD); figure-of-merit; GNR; HCI; hot carrier; hot electron; HP process; ITRS; litho-friendly design (LFD); LP process; nanoscale; nanoscale CMOS technology; NBTI; overdrive (OD); physical design; reliability; SOC; STI; stress; SWCN; TDDB; voltage headroom; WPE

Manuscript received February 5, 2009; revised April 28, 2009. Current version published September 16, 2009.

L. L. Lewyn is with Lewyn Consulting Inc., Laguna Beach, CA 92651 USA (e-mail: lanny@pacbell.net).
T. Ytterdal and **C. Wulff** are with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway (e-mail: ytterdal@iet.ntnu.no; carsten@wulff.no).
K. Martin is with Granite SemiCom, Toronto, ON, Canada (e-mail: martin@granitesemi.com).

Digital Object Identifier: 10.1109/JPROC.2009.2024663

NOMENCLATURE

ACLV — Across-chip length variation: variation in CMOS device length across a chip caused by lithographic errors such as gradients in photoresist and image defocus.

CHE	Channel hot electron: impact-ionization-produced electrons resulting from the vector sum of lateral and transverse electric fields near the drain end of the channel. In the CHE mode of hot electron damage, the source-bulk voltage is zero.
CHISEL	Channel-initiated secondary-electron generation: secondary hot electrons produced when carriers resulting from a first impact ionization near the drain are accelerated toward the bulk, generating hole-electron pairs in a secondary impact ionization event. In the CHISEL-mode of hot electron damage, the source-bulk voltage is not zero.
DFM	Design for manufacturability: design and verification methodology employed to assure that production silicon yields a suitable percentage of die meeting the design specifications.
DFP	Design for performance: design and verification methodology employed to assure the production silicon meets the performance objectives of the original circuit design.
DFR	Design for reliability: design and verification methodology employed to assure that the IC performance does not substantially degrade over the anticipated lifetime of the product.
DIBL	Drain-induced barrier lowering (short-channel effect): the effect of drain field reaching through the channel and lowering the source-channel energy barrier, thereby resulting in reductions in threshold and increases in output conductance with drain voltage.
DRC	Design rule check: a check of physical layout using a foundry-determined set of rules or more complex computations, to include the effect of nearby lithographic patterns.
EM	Electromigration: migration of conductive material resulting from current flow producing narrowing or "necking" to increase resistance, or bridging to create shorts.
EOD	Excess overdrive: an operating condition where devices may, depending on waveshape, exceed voltages beyond overdrive limits specified by the foundry.
F	Minimum-gate-length feature size: minimum gate feature size in the physical layout.
FOM	Figure-of-merit: a normalized quantity calculated based on several analog circuit performance parameters that enables a comparison of the "quality" of circuits having different performance parameters.
GBW	Gain-bandwidth product: gain-bandwidth is defined simply as gm/C_{ox} and is used here for the purpose of determining how GBW varies, rather than a numerical value.
GNRs	Graphene nanoribbons: deep-submicrometer conductor technology that is a candidate to replace dual-damascene copper. This technology uses planar fabrication methods and relies on long electron scattering lengths to reduce resistance and increase EM limits.
HCI	Hot carrier injection: carrier injection into the channel or gate insulator produced by impact ionization near the drain end of the channel creating interface and oxide trap damage.
Hi-K	High dielectric constant: term used in nanoscale technology nodes referring to gate materials that have higher dielectric constants than oxynitrde, such as hafnium and zirconium.
HP	High-performance process (platform) device: HP processes (platforms) incorporate thinner oxides than low power platforms to provide significantly better transconductance, lower thresholds, lower saturation voltage, higher saturation currents, higher gain-bandwidths, and, unfortunately, lower operating voltages.
I-DAC	Current DAC: current DACs in CMOS technology are formed by summing output currents of many CMOS devices, which are subject to significant stress-induced errors.
LFD	Litho-friendly design: a physical design strategy incorporating special rules or guidelines, intended to mitigate several adverse stress or lithographic effects and thereby avoid the relayout required when such effects are first revealed by postlayout verification tools.
Lo-K	Low dielectric-constant insulation material: intermetal insulation material having lower dielectric constant than the SiO_2 used prior to nanoscale technology nodes.
LP	Low power process (platform) device: low-power platforms incorporate thicker oxides than the high-performance platforms to reduce logic power using higher thresholds and lower leakages. LP platforms have higher saturation voltages, lower transconductance, higher output conductance and significantly lower gain-bandwidths than HP platforms.
M1, Mx, Mz	First metal, intermediate metal, thick metal: acronyms for first, thinner, and

	finer-pitch metal (M1), intermediate metal (Mx), and thick (Mz) top metal.
MiM	Metal–insulator–metal capacitor: technology using a thin insulator between intermediate interconnect metal and an added metal to create a precise capacitor in an analog process.
NBTI	Negative bias temperature instability: threshold voltage instability in PMOS devices that is dependent on temperature and device geometry, particularly W.
OD	Overdrive: an operating condition permitted by design rules where some devices, such as core devices, may exceed normal operating parameters such as voltages.
OPC	Optical proximity correction: a correction to the reticle patterns to mitigate distortion in the intended image shape resulting from diffraction or processing effects.
PiP	Poly–insulator–poly: technology using a thin insulator between interconnect poly and an added poly to create a precise capacitor in an analog process.
PSM	Phase shift mask: a photomask that uses either variable thickness (alternating PSM) or variable contrast in the dark regions (attenuated PSM) to improve image resolution using the interference of the phase-shifted light to improve image definition on the wafer.
RET	Resolution-enhancement technology: technology used to modify illumination source, contrast, or dimensions of the patterns on the reticle from their shape in the original layout to increase the accuracy of the image on the final silicon.
RIE	Reactive-ion etching: lithographic patterning technique using an ionized gas (plasma) to achieve high etch rates at low temperatures.
SOC	System-on-chip: technology that enables integration of all necessary electronic circuits for a complete system on a single integrated circuit.
SWCN	Single wall carbon nanotubes: deep-submicrometer conductor technology that is a candidate to replace dual-damascene copper. This technology relies on long electron scattering lengths to reduce resistance and mitigate electromigration limits.
TCAD	Technology (aware) computer-aided design: processing-technology-aware simulation technology used to define the physical configuration of devices within the silicon and then determine device (or simple circuit) performance using physical models of carrier propagation.
TDDB	Time-dependent dielectric breakdown: the breakdown of a gate dielectric insulator as a result of the application of a high electric field (such as $>$ 5 Mv/cm) over a period of time.
VCC_{MAX}	Maximum supply voltage: the maximum gate to source, drain to source, or source to bulk voltage for simulation purposes.
VCC_{OD}	Maximum overdrive supply voltage: the maximum drain-to-source voltage for core devices that are not operated at high drive currents.
VCC_{EOD}	Maximum excess overdrive supply voltage: the maximum drain-to-source voltage for devices with drain voltage waveforms that allow safe operation at voltages above VCC_{OD}.
WPE	Well proximity effect: effect producing threshold voltage increases where well implant ions reflect off the photoresist edges and increase bulk doping under the gate.

I. INTRODUCTION

The exponential evolutionary trend in nanoscale complementary metal–oxide–semiconductor (CMOS) technologies predicted by Moore's law has been fueled by a seemingly unending demand for ever better performance and by fierce global competition over the past three decades. A driving force behind this very swift progress is the long-term commitment to a steady downscaling of CMOS technologies needed to meet the requirements on speed, complexity, circuit density, and power consumption posed by the many advanced applications.

The degree of scaling is measured as the half-pitch of the first-level interconnect in DRAM technology. The degree of scaling is also termed the "technology node" by the International Technology Roadmap for Semiconductors (ITRS). According to the 2006 ITRS update, the 2007 production technology node has reached 45 nm, while the smallest features, the effective MOS field-effect transistor (FET) electrical gate lengths, are merely 25 nm. These numbers are expected to approach 12 and 7 nm, respectively, during the next ten years.

With the introduction of nanoscale (sub-100 nm) CMOS technologies, analog designers are faced with many new challenges at different phases of analog design. These challenges include severe degradation in device matching characteristics as a result of device and lithographic quantum limits, resolution-enhancement techniques that result in the device image on the reticle's no longer having a 1-to-1 correspondence with the layout because feature sizes are less than one-quarter of the wavelength of deep ultraviolet (UV) ($\lambda = 193$ nm) illumination, high leakage currents resulting from low thresholds and thin gate

oxide tunneling currents, electromigration limits requiring more parallel levels of metals to connect shared-drain analog devices, limitations on accurate prelayout simulation, and device counts greater than 100 million on a single chip, forcing new methods for system partitioning to accommodate the increased power densities.

The first part of this paper presents factors affecting device matching, including those relating to single devices as well as local and long-distance matching effects. Several reliability effects are discussed, including physical design limitations projected for future downscaling. In some cases, it may be helpful to exceed foundry-specified drain-source voltage limits by a few hundred millivolts. Models are presented for achieving this, which include the dependence on the shape of the output waveform. The second part of this paper focuses on trends in device characteristics and how they influence the design of nanoscale analog CMOS circuits. A number of circuit design techniques employed to address the major nonidealities of nanoscale CMOS technologies are discussed. Examples include techniques for establishing on-chip accurate and temperature-insensitive bias currents, digital calibration of analog circuits, and the design of regulator and high-voltage circuits. This paper also presents a survey of the evolution of FOM for analog-to-digital converters (ADCs).

II. NANOSCALE CMOS PHYSICAL DESIGN AND DEVICE TRENDS

Scaling MOS devices down below 100 nm has produced little improvement in device transconductance (g_m), has increased the dominance of wiring parasitics in predicting circuit gain bandwidth, and has brought into play a plethora of lithographic, stress, quantum, and process variability effects that make the problem of good analog device matching more difficult. There is no longer a 1-to-1 correspondence between geometric patterns in the layout and the predistorted patterns in the reticle as a result of RETs. These techniques, including PSM and OPC, are required to bring the critical dimensions in the fabricated ICs closer to the intended layout dimensions.

Digital design tools are presently far more capable than the limited tool set currently available to the analog IC designer for predicting circuit performance at the prelayout stage of circuit design and simulation. Analog simulation tools are gradually incorporating more prelayout effects and additional physical effects with postlayout parameter extraction. Yet, several matching effects resulting from local and long-distance lithography factors can only be included in the circuit simulations with fabrication-technology-aware simulation tools. These "TCAD-like" simulation tools are not normally used by analog circuit designers. Unfortunately, the gap between the semiconductor fabricator's ability to keep pace with Moore's law and the ability of DFP, DFR, and DFM tools to adequately predict and verify analog circuit performance in nanoscale CMOS technologies seems to be widening [1]. Using DFP/DFR/DFM tools helps overcome many nanoscale physical and circuit design problems but does not eliminate them.

Analog circuit simulation and verification techniques have never been proper substitutes for good analog system, circuit, and physical layout design. Given the widening gap between analog design and accurate circuit prelayout simulation, success in the nanoscale analog CMOS design era will depend heavily on understanding the many physical layout factors impacting circuit performance and taking them into account early in the design activity. A combination of old and new physical design strategies will be described in this section. They may be helpful in the mitigation of several effects that may become apparent only after postlayout extraction and simulation. Redesign and re-layout resulting from pretapeout DFR, DFP, and DFM check errors are becoming a significant cost factor and substantially increasing time-to-market.

A. Device Physical Scaling and Operating Conditions

The ITRS working group has published many device design targets for nanoscale technology nodes through the next decade. Many device dimensions and design rules remain relatively constant fractions of the minimum-gate-length feature size (F), over several past and future process technology nodes. In Table 1, ITRS SiO_2-equivalent gate oxide thickness (T_{ox}) targets for high-performance (HP) devices in column 3 are compared with targets for low-power (LP) devices in column 4 and a general prediction that T_{ox} will scale as a constant fraction of F ($0.02 \cdot F$) in column 5 [2]. At technology nodes beyond 180 nm, gate leakage as a result of direct tunneling becomes significant. At 100 nm and beyond, thin T_{ox} values are realized by using thicker dielectrics with higher dielectric constants that have SiO_2-equivalent capacitance.

For the purpose of predicting future technology node gate capacitance, the $0.02 \cdot F$ predicted values agree well with the ITRS target T_{ox} values for HP devices. HP

Table 1 Oxide thickness scaling predictions. Per the ITRS 2006 update, manufacturing solutions for metal-gate insulators at 32 nM and beyond are not yet known, and at the 16 nM node are presently undefined. LP devices do not follow HP scaling trends. LP devices provide higher values of T_{ox}, and therefore higher operating voltages, reduced β (from lower μC_{ox}), and reduced GBW performance

Node F (nm)	Year	ITRS Tox HP	ITRS Tox LP	0.02*F Tox (nm)
65	2007	1.30		1.30
45	2010	0.90	1.40	0.90
32	2013	0.65	1.20	0.64
22	2016	0.50	0.60	0.44
16	2019	?	?	0.32

device T_{ox} has been scaled to provide tight control of drain-induced-barrier lowering effects (short-channel effect, or DIBL) at the source-channel boundary. HP devices can therefore be constructed with short channel lengths and have high gain-bandwidths (GBW $\approx g_m/C_{ox}$) and the reasonably flat I–V characteristics that are required of analog CMOS amplifier devices in the saturation region. LP devices have thicker oxides to allow higher operating voltages, higher threshold voltages to achieve lower leakage currents, and longer effective channel lengths. LP processes do not follow device scaling trends closely, have significantly lower GBW performance than HP technologies, and will not be considered in the discussion on nanoscale HP device scaling trends and design strategies.

Operating current densities play an important role in design strategies for optimizing nanoscale CMOS W/L circuit values. NMOS current densities for medium-high GBW devices have been, and will remain, fairly constant at 10 μA/sq. A similar value for PMOS devices is 3 μA/sq. High-frequency (HF) devices are generally operated at 1.5 times these current densities. Where power supply headroom is severely constrained, current density is generally reduced by half and saturation voltage (V_{dsat}) drops to approximately 1/8 of maximum supply voltage (VCC_{MAX}) [3]. Because device current density for any performance target is generally constant under scaling, total device current I_{ds} therefore varies in direct proportion to the number of squares (W/L) in the device.

B. Device Matching

Device matching has always been a major concern for analog CMOS circuit designers. Matching between closely spaced pairs of devices is influenced primarily by local random and systematic error factors. The matching of devices such as precise capacitors in a large capacitor-DAC (C-DAC) array can be influenced by both local and systematic long-distance effects from lithographic patterning several tens of micrometers outside the array, regardless of the technology node. Physical design approaches intended to mitigate these effects in nanoscale technologies are commonly referred to as litho-friendly design.

The ITRS roadmap specifies a target value for gate lithographic accuracy in terms of the 3σ variation (ΔL) in minimum gate length. The target value for 3σ $\Delta L/F$ is holding fairly constant at $0.047 +/- 0.002$ for the 65 down through the 16 nm technology node. While it is of primary importance to examine the precision of key dimensions in minimum-area devices, it must not be forgotten that lithographic patterning for devices of equal area is constantly improving as process linewidths are scaled down.

As devices become smaller, threshold variations from quantum effects in substrate (N-well or P-tub) and threshold-setting implants play a stronger role. The total number of implant ions (N_d) under the gate at 65 nm is on the order of 10^2 and threshold variations > 10 mV in minimum-area devices resulting from the quantum variations $\approx N_d^{-0.5}$ cause a significant threshold voltage (V_t) mismatch factor. However, substrate (well or tub) doping density N_b is increasing under scaling, requiring a depletion capacitance increase, in order to mitigate short-channel effects. A consequence of increasing N_b is that devices of *equal area* have smaller V_t variation resulting from substrate doping quantum effects. V_t matching for equal area devices is improving, but historical data suggest that not much improvement should be expected beyond the 65 nm technology node.

At the 65 nm technology node, the gate insulator thickness is little more than ten molecular layers of SiO_2 with local variations of $+/-$ one layer. Fortunately, minimum device W and L values are approximately 50 times greater than T_{ox} ($T_{ox}/0.02$) and therefore orders of magnitude larger than molecular layer spacing. At 65 nm, device photoresist edge roughness effects tend to average out for devices with larger than minimum area, and industry-average gate insulator thickness values used in worst case SPICE models are still held to approximately the same percentage variation of $+/-3\%$, typical of earlier technology nodes.

While quantum effects in gate insulator thickness and lithography combine to result in significant V_t matching effects for small devices, good V_t matching has never been expected from minimum-area devices in any technology node. However, since the sum of minimum-feature width and space ($2F$) is significantly less than the wavelength (λ) of the 193 nm UV light used to expose the photoresist, lithography-induced optical proximity and defocus effects can cause significant systematic local and across-chip device-length variations (ACLV).

Good device V_t, g_m, and saturation current (I_{dsat}) matching has always depended on the precision of a number of process-driven variables such as bulk doping concentration, V_t-shifting implant dose and range, carrier mobility, device W-L dimensional accuracy, and gate-oxide thickness. At the 100 nm technology node, additional local effects have become significant factors influencing device matching. These effects include V_t shifts resulting from the proximity of gate to the well edge [4], and V_t and mobility shifts resulting from the distance to the local trench isolation, i.e., the shallow-trench effect (STI) [5].

While issues of device pattern uniformity used to be critical for pairs or several matched analog devices, device uniformity issues are now of concern in relation to the variability in performance of small, isolated, analog or digital circuits, such as inverters. Digital circuit performance at the 45 nm node, and beyond, is significantly affected by the lithography and stress effects from neighboring circuits or areas of low circuit density at distances greater than 1 μm. The analog designer who does not pay attention to the physical design of the circuits surrounding the layout of a local group risks the consequence of poor yield.

Because nanoscale analog devices are capable of operation at frequencies well beyond the 10 GHz range,

the location of the substrate ties is a concern in matching differential pairs used in HF amplifiers. Good device matching and operation at HF requires distances to the well or substrate ties that are uniform and in close proximity to the source-channel boundary.

Matching errors resulting from traps created in the device during RIE patterning (the antenna effect [6]) may be significantly larger than all the random V_t errors described above. Antenna-effect design rule checks are typically based on a maximum V_t shift of 50 mV that is acceptable for digital circuit design. DFM/DRC tools are therefore presently unsuitable as a pass–fail check for match-critical analog circuits such as comparators and differential amplifiers. Local and long-distance lithographic, STI, WPE, and radio-frequency (RF) systematic device matching errors, as well as methods for mitigating their effects, will be discussed in this section.

1) Closely Spaced Device Matching: Prior to the 100 nm technology node, all factors affecting closely spaced device threshold matching were addressed by the designer using a simple formula for V_t, g_m, or I_{dsat} variations such as $\sigma(V_t) = A(V_t)/(W \cdot L)^{-0.5}$, where desired values of $\sigma(V_t)$, $\sigma(g_m)$, etc., were determined by using values of A such as $A(V_t)$ or $A(g_m)$ that were empirically determined by the foundry. Foundries currently supply Monte Carlo simulation models for determining DFP limits on threshold variation for technologies past the 100 nm node; however, designers still rely primarily on foundry-supplied A factors for initial circuit design. It is therefore appropriate to consider A-factor trends, such as $A(V_t)$ in current and advanced technology nodes. The factor $A(V_t)$ is commonly expressed in dimensions of mV·μm.

In Fig. 1, historical values of $A(V_t)$ for minimum-length PMOS devices are plotted and compared to the recent trend where $A(V_t)$ is scaling in proportion to F. It is apparent from the figure that other factors are overtaking lithographic accuracy. These factors include random components of local process variables such as implant range (depth) variability that are becoming more significant as device vertical dimensions shrink. Stress induced by irregular overlying metal and nearby device patterns within the local stress field ($< 2\ \mu$m) has been an increasing cause of random threshold variations for several technology nodes prior to the nanoscale era. More recent stress factors include STI and features added to overcome reduced carrier mobility from higher channel doping. These include cap layers over the gate and Ge implants into raised S-D structures, as shown in Fig. 2 [7]. Note the following.

1) Low-threshold devices have a lower bulk implant dose, resulting in perhaps 10% better matching accuracy.
2) High-threshold devices generally have an implant added to the standard V_t implant, resulting in less V_t match accuracy.

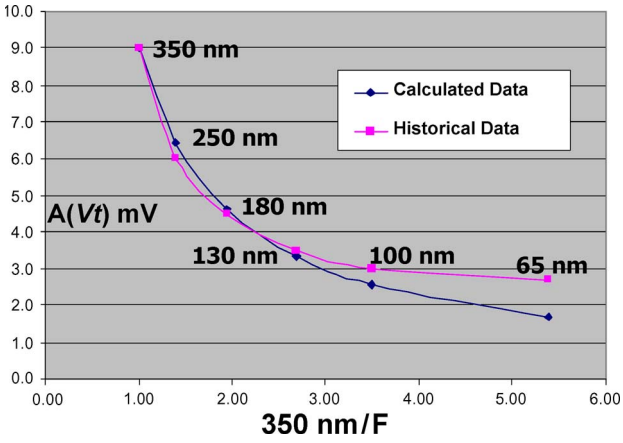

Fig. 1. ***Historical values of $A(V_t)$ for minimum-length PMOS devices (courtesy of Prof. K. Martin, University of Toronto) are plotted and compared to the recent trend where $A(V_t)$ is scaling in proportion to minimum-feature-size F. $A(V_t)$ has been improving with F, but significant improvements are not expected beyond the 65 nm technology node, where 350 nm$/F = 5.4$.***

3) NMOS may have less V_t match accuracy than PMOS.
4) At high values of gate drive $V_{gt}(\equiv V_{gs} - V_t)$, the saturation current I_{dsat} matching constant $A(I_{\mathrm{dsat}})$ is dominant over $A(V_t)$.

2) Mitigating Local Systematic Matching Errors: Device lithographic patterns below 100 nm have significant corner rounding and width nonuniformity. OPC of the reticle patterns mitigates corner rounding, pinching, and other local pattern proximity effects when they are within a distance of approximately 3 λ [7]. After OPC, the reticle no longer has a 1-to-1 correspondence with the layout, as is shown in Fig. 3. Optically dense patterns in close proximity to one another tend to cause local pattern expansion or pattern bridging. Optically sparse areas contribute to cause pattern narrowing (pinching or necking). Pinching is especially harmful in performance critical layers such as poly (see Fig. 4). Poly pinching or rounding can contribute to mismatch errors and to significant increases in device leakage, especially at the gate edge. Because leakage is strongly dependent on channel length, excessive leakage resulting from pinching or rounding can have a stronger effect on yield than mismatch errors. Good device matching and leakage control must include the effect of patterning other devices within a local group.

Regular poly patterns in a single direction with a favorable pitch require less OPC, and higher poly density reduces poly RIE etch loading. Regular patterning of poly and active-area layers, as well as extra gate extension that makes device performance less sensitive to layer overlap, are important components of LFD [8].

There are many other factors that affect the final dimensions of the device in silicon. These include

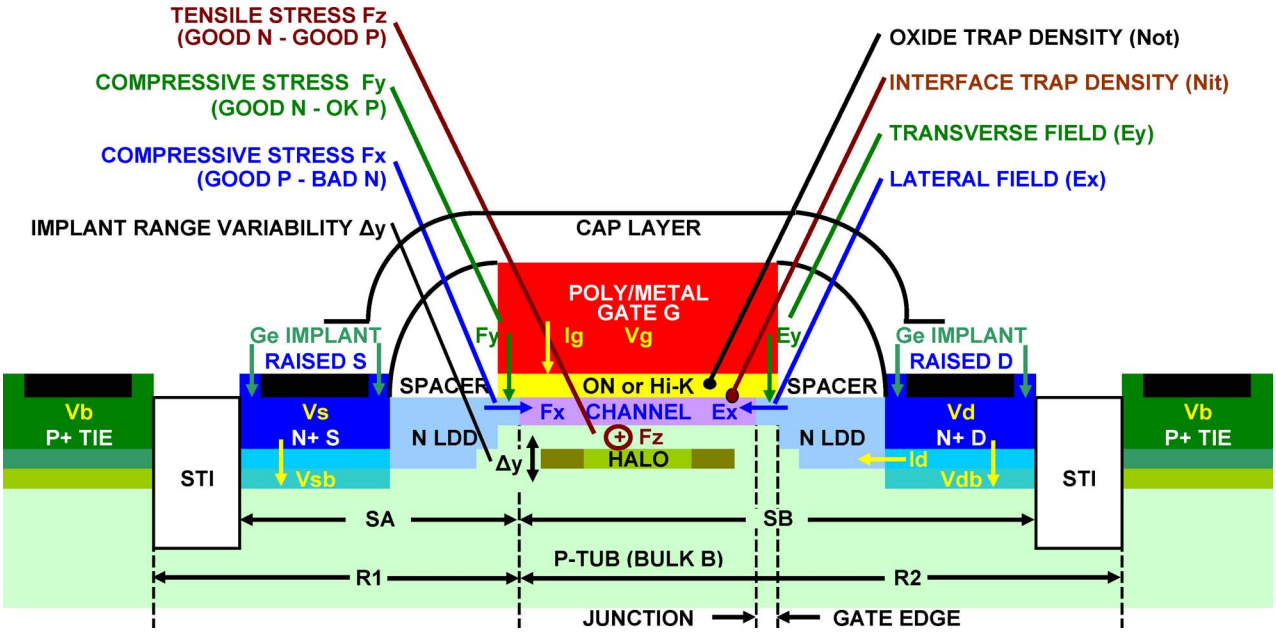

Fig. 2. ***NMOS cross-section. In addition to stress from cap layers and Ge raised source-drain (S-D) implants, device dimensions such as distance from source-channel boundary to nearby STI (SA and SB), proximity and regularity of overlying metal patterns, and short distances to other device patterns within the local*** ($< 2\ \mu$m) ***stress field induce transverse*** (F_y) ***and lateral*** (F_x ***and*** F_z) ***stress components, which affect threshold and mobility. Increasing the distance to P+ ties increases local tub (bulk) resistance components R1 and R2, which isolate the device MOS model substrate node from the device subcircuit symbol*** V_b ***node and degrade HF performance. Hot carrier reliability stress is dependent on the sum of transverse and lateral fields*** E_y ***and*** E_x***. These fields are increased near the drain by increasing source to bulk*** (V_{sb}) ***and drain*** (V_d) ***to gate*** (V_g) ***or source*** (V_s) ***voltages in various combinations. As hot carrier stress increases, damage to channel from interface trap density*** (N_{it}) ***affects threshold and mobility, while gate oxynitirde (ON) or high-dielectric-constant (Hi-K) insulator trap density*** (N_{ot}) ***affects threshold and gate leakage.***

exposure variation (exposure latitude), focus variation (depth of focus), and mask size variation (mask error enhancement factor) [9]. Isoprobability contours for these factors are shown graphically in Fig. 5. While these factors result in device variability that is to some extent layout-dependent, they are still present even if good LFD is practiced. LFD requires slightly more layout area but shortens the time required to produce the layout and requires less postlayout extraction rework to remove lithographic deficiencies flagged by DFR/DFP/DFM tools.

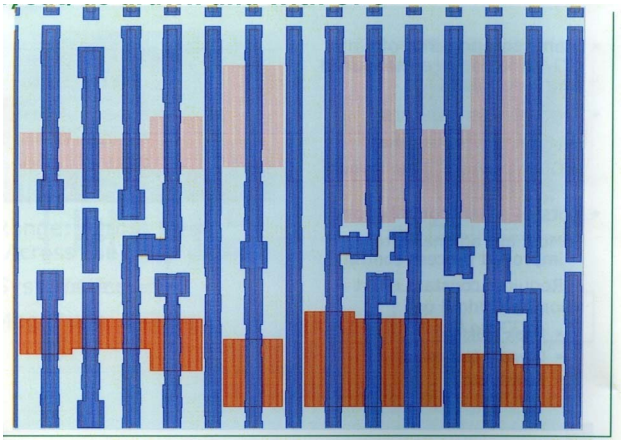

Fig. 3. ***After OPC, the reticle patterning no longer has a 1 : 1 correspondence with the layout [8].***

WPE effects are caused by substrate implant ions being reflected off the well edge-definition photoresist and increase device thresholds when their paths end under the gate. The effect of WPE is mitigated by increasing the distance from the gate to the well edge. When the distance from gate to well edge is $> 2\ \mu$m, the WPE effect is very small. Matching all gate-to-well distances minimizes matching errors, but for distances $< 1\ \mu$m, significant V_t increases are probable. A better strategy is to use extended well patterns. In the amplifier cell shown in Fig. 6, wells are extended to the full X dimension of the cell, and adjacent cells to the left and right extend the well pattern.

The STI effect is a result of device oxide isolation trench producing stress in the channel affecting both V_t and mobility. The effect is inversely proportional to the distances SA and SB from the gate to each trench, as shown in Fig. 2. The effect can be mitigated by providing two gates per device. In this case, the average gate-to-trench distance is approximately equal to the default SPICE model value. WPE and STI effects may cause large current variations in the design of cascode devices.

The value of the distance-dependent resistance from the gate to the substrate (N-well or P-tub) tie affects

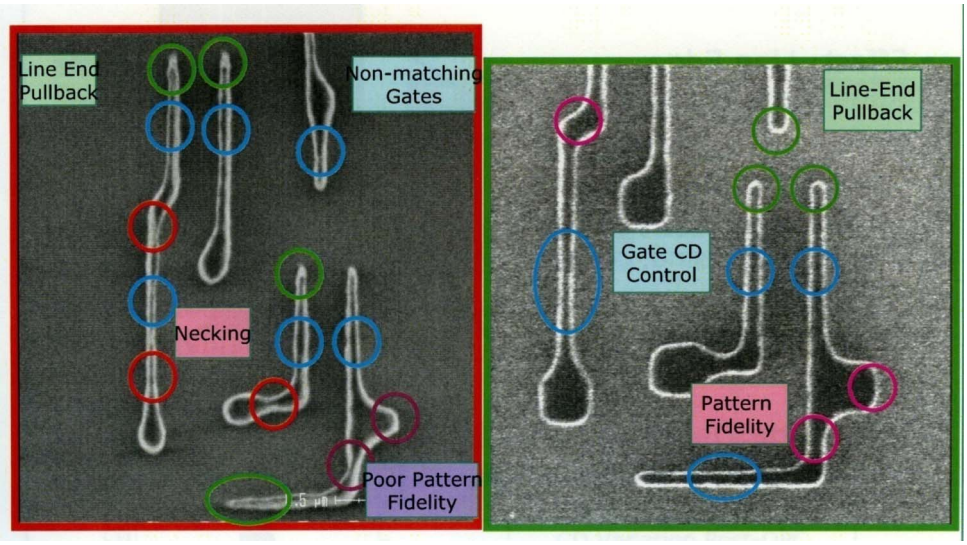

Fig. 4. *Following the application of OPC to the reticle, poly layer patterning achieves higher CD accuracy in the final silicon (right panel) [8].*

performance at high frequencies. RF models include the effect of tie resistance and substantially increase simulation time. In an RF model, a basic SPICE device model is used for a four-terminal MOS transistor, which is imbedded in a physical-layout-dependent RF subcircuit. In this model, the MOS device S/D junction diodes are coupled to the device model bulk node (V_b). V_b is then coupled to the transistor subcircuit bulk terminal node through resistors (R1 and R2 in Fig. 2). If the gate-to-tie distance is minimized, an RF model may not be required to achieve simulation accuracy for nanoscale technology applications in the range below 10 GHz. An example of device layout intended to mitigate intended to mitigate OPC, STI, and gate-to-tie-distance effects is shown in Fig. 6.

3) Mitigating Long-Distance Systematic Matching Errors: It is not surprising that in the nanoscale technology era, classical analog design techniques are being abandoned in favor of digital calibration, adaptation, or signal averaging methods. Low-area component matching is becoming more difficult, while the area and power overhead for digital calibration circuits is decreasing. Open-loop

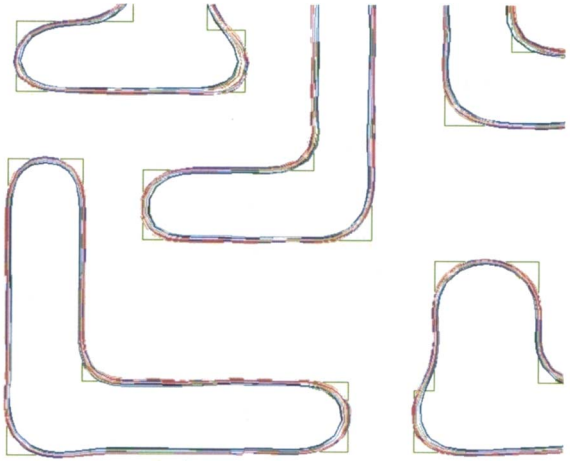

Fig. 5. *Factors that affect the final dimensions of the device in silicon include dose variation (exposure latitude), focus variation (depth of focus), and mask size variation (mask error enhancement factor). Differences between the layout geometry and isoprobability contours for each of these effects are shown [9].*

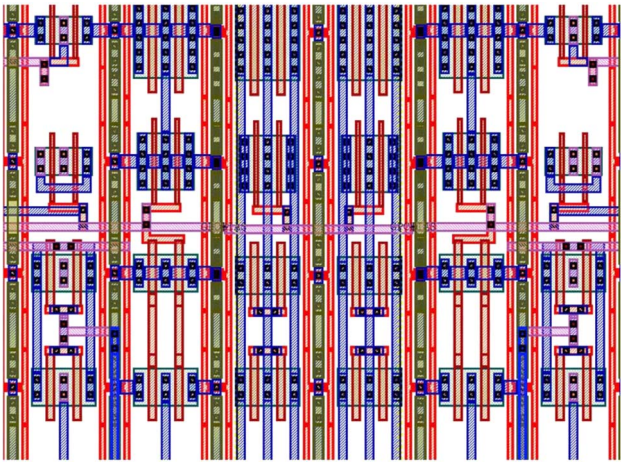

Fig. 6. *A portion of an amplifier cell with regular device pitch in both X and Y directions (upper metal layers removed for clarity). For best HF performance, all devices' substrate ties are placed on either side of two-finger gate patterns. Grounded stripes of poly are interposed between device active area and all substrate ties to minimize the need for reticle compensation (OPC) and also reduce poly etch loading to achieve good CD accuracy.*

amplifiers with resistive loads and digital calibration have significantly reduced power requirements and have increased resolutions of ADCs to 12 bits [10], [11]. CMOS ADC resolutions in the 13–14 bit range now use C-DAC array matching [12] or background calibration techniques [13]. Most calibration alternatives do not eliminate the requirement for some degree of matching, and precise matching is still required for the design of accurate supporting circuits such as bias current mirrors.

Conventional C-DAC precise capacitor weighting networks are usually arranged in a square pattern, with the capacitor selection ordered to keep the average distance from the center as close as possible to zero. This type of ordering, commonly known as common centroiding, assumes that process gradients are substantially uniform (linear) across the array. Unfortunately, capacitor patterns at the edge of the array tend to be underetched as a result of low-density patterns of the same material tens of micrometers outside the array causing heavy RIE loading. Patterns in the center of the array are usually dense, and RIE loading is minimized. It is therefore difficult to reduce first-order (linear) process gradients and the second-order (quadratic) lithographic errors induced by asymmetrical patterns outside a square C-DAC array. Linear insulator thickness process gradients affect poly-poly (PIP) and MIM capacitor electrode spacing. Quadratic lithographic gradients affect PIP, MIM, and finger-capacitor edge pattern uniformity.

A one-dimensional centroiding methodology can be implemented to minimize the mismatch caused by first and higher order gradients along a linear array. Three centroiding concepts can be combined to substantially mitigate first- and second-order long-distance process and lithographic gradients in precision C-DACs to achieve differential and integral linearity in excess of 14 bits without calibration [14]. Nanoscale CMOS gigasample current DACs (I-DACs) have significant second-order g_m variations as a result of silicon stress induced by neighboring lithographic patterns. Second-order centroiding combinations for both precision C-DACs and high-speed I-DACs are required to achieve high linearity without requiring digital calibration.

C. Nanoscale CMOS Reliability Issues

The present trend in analog CMOS having independent IP providers favors more circuit design activities in fabless centers that have only limited reliability and engineering capability. Unfortunately, many of the foundry-reported reliability stress limits are documented in terms of predicting digital, not analog, circuit end of life. The net result is that there are gaps between the required and the available device reliability data and simulation tools. Despite efforts within the industry to narrow what is being called "the reliability gap," it is widening with time [15].

Circuit simulation tools presently lack the capability to predict the effect of several reliability stress effects, including gate insulator time-dependent dielectric breakdown (TDDB), hot carrier injection (HCI), negative bias temperature instability (NBTI), and junction breakdown as a function of device terminal voltages. TCAD tools can accurately predict several of these effects, but they run too slowly to be useful in most circuit simulators.

As operating voltages are scaled down, gaining or losing 100 mV of dynamic range for an analog circuit may be the difference between meeting a design specification easily, or perhaps not. It is therefore important for analog designers to understand the limits of the process they are using. Of greater importance is the understanding of how rapidly the device reliability stress factors vary in the region of operation just inside and outside the process voltage limits. An extra 100 mV of signal swing may be of great help in achieving required dynamic range goals, but it also may result in an unacceptable reduction in product target lifetimes.

While analog designers favor the use of cascode circuit configurations ($V_{sb} > 0$) to mitigate drain-source voltage (V_{ds}) stress effects, most foundry device qualification is based on $V_{sb} = 0$ test conditions. The condition $V_{sb} > 0$ is playing a stronger role in both hot and cold carrier damage effects for highly scaled devices. Unfortunately, the foundries are providing inadequate reliability models and/or qualification data for analog design using the $V_{sb} > 0$ condition, where devices are subjected to the CHISEL mode of HCI stress.

In nanoscale analog CMOS design, there is no good substitute for understanding TDDB, HCI, NBTI, and EM process stress factors. How these stress factors vary with operating conditions, the required foundry qualification test data, and how to alter circuit topologies and device geometries to mitigate their effect are discussed in this section. We will also explore methods for extending device terminal voltage limits under certain conditions beyond foundry-specified voltage limits.

1) Nanoscale Device Reliability: Many aspects of device, circuit parameter, and supply voltage scaling were accurately predicted by Dennard *et al.* in 1974 [16]. Changes in device and circuit performance were indicated in terms of the dimensionless scaling factor κ shown in Table 2. At power supply voltages of 3.3 V, and drawn gate lengths of 350 nm, it was not difficult to obtain high signal-to-noise ratios in the design of high-gain wide-band amplifier stages. Downscaling L significantly increased speed performance in HP technology nodes for both analog and digital circuits. The prediction of constant power density [16] was apparently based on static current. Supply voltage and static current both scale as $1/\kappa$. However, dynamic current scales in direct proportion to κ ($I = f * C * V$) and the corresponding power density scales as κ^2. The κ^2 dependence of power density on scaling has placed the highly temperature-dependent EM reliability factor at the top of the list for most likely early nanoscale chip failure.

Table 2 Scaling Results for Circuit Performance [16]. Device or Circuit Parameter Changes With Scaling Factor κ

Device or Circuit Parameter	Scaling Factor
Device dimension *Tox, L, W*	$1/\kappa$
Doping concentration	κ
Voltage *V*	$1/\kappa$
Current *I*	$1/\kappa$
Capacitance $\varepsilon A/t$	$1/\kappa$
Delay time/circuit *VC/I*	$1/\kappa$
Power dissipation/circuit *VI*	$1/\kappa^2$
Power density *VI/A*	1

As lithographic minimum feature size (F) decreased to 100 nm, ITRS HP target voltages for that node decreased by $1/\kappa$, down to 1.0 V. Reliability-limited device lifetime was long enough to allow supply voltages to be increased by 200 mV for core, but not for I/O devices. Operation above HP target voltages is commonly described by the foundries as the overdrive mode of device operation. Fully complementary, cascoded high-gain stages consumed too much of the OD supply voltage to be practical for wide dynamic range ($>$ 60 dB) applications. Beyond the 100 nm technology node, operation of some analog circuit current sources, input stages, and even output stage drivers, at voltages above VCC_{MAX}, is necessary. The nanoscale-era circuit designer's challenge is how to operate all devices safely within the reliability limits of the process.

In cases where device reliability stress is high, a matter of serious concern is how many devices are being operated under such conditions. As the number of devices operated under high stress increases, the chip lifetime reduction probability also increases. The disadvantage of using a device in a high stress mode must therefore be weighed against the performance advantage gained. The chip reliability lifetime design target must be in accord with the reliability target for the complete system. In cases where several identical chips, such as DRAMs, are used in a single system, the importance of IC target lifetime increases. The statistical analysis of IC long-term reliability is, however, beyond the scope of this paper.

2) Time-Dependent Dielectric Breakdown: TDDB of gate insulating material results from the cumulative effect of insulator trapped charge (N_{ot}) buildup during short-term and long-term high-field stress. High N_{ot}-charge-induced local fields build up within the insulator, causing gate leakage, excess noise, and, eventually, dielectric breakdown [17]. Short-term stress events include RIE-induced gate currents, electrostatic discharge (ESD) currents, and startup conditions that increase gate-to-surface potentials. Long-term stress conditions include steady-state operation, signal overload events, and power-down modes with bias current off and supply voltage on.

A common assumption with regard to gate insulator reliability is that, as long as operating $V_{gs} < VCC_{\mathrm{MAX}}$, TDDB stress will be within acceptable limits. That is not the case where, under startup conditions, V_{gs} limits are inadvertently exceeded for several milliseconds per startup. Precision analog circuits can undergo serious noise and V_t performance degradation as a result of gate-insulator trap (N_{ot}) buildup prior to device TDDB failure. The lifetime for gate insulators is normally defined as t_{50}, the time it takes for 50% of the devices to fail at a high level of gate current. Breakdown charge (Q_{bd}) levels for short-term high-stress events are different from those for long-term low-stress conditions. TDDB for thin films of SiO_2 has been well characterized [18], and maximum gate dielectric electric fields (E_{ox}) for prior processes have been surprisingly uniform at a value of approximately 5 MV/cm.

It is important to note that, for $E_{\mathrm{ox}} > 4.8$ MV/cm, SiO_2 breakdown occurs at a faster rate. Simple models do not take into account the fact that, at high fields, secondary processes create excess damage. Those processes are a result of higher energy carriers' being injected into the oxide. TDDB process qualification is always performed at $E_{\mathrm{ox}} \gg 5$ MV/cm, with T elevated to at least 125 °C (398 K). Substantially higher temperatures can produce accurate test results in shorter times, or in an equivalent time at lower field strength [19].

The log of t_{50} is a linear function of E_{ox} with y-intercept A and slope (field acceleration parameter) B. A simple model for TDDB is

$$\log t_{50} = A - BE_{\mathrm{ox}}. \tag{1}$$

Here, the coefficient A depends on gate material and oxide thickness. It includes a term E_a/k_BT, where E_a is the thermal activation energy of the gate material process and k_B is the Boltzmann constant. For a typical 3.2 nm gate oxide, using values of 1.32 and 1.1 for A at 9.0 and 3.5 nm, respectively [20], the resulting log t_{50} curve is shown in Fig. 7.

Although the curves in Fig. 7 are for SiO_2, and not for a future-process gate insulator, it is still helpful to use them to understand some aspects of long-term versus short-term stress. Manufacturable solutions are currently unlikely for ultrascaled gate dielectrics that use nitrogen or hafnium having more than twice the thickness of SiO_2 equivalents [21]. In Fig. 7, note the following.

1) An increase in gate field E_{ox} by only 10% decreases t_{50} by a factor of ten. However, if TDDB lifetime exceeds expectations by more than an order of magnitude during process qualification, allowing 100 mV more V_{gs} might be acceptable as a specified overdrive (OD) limit for core devices.

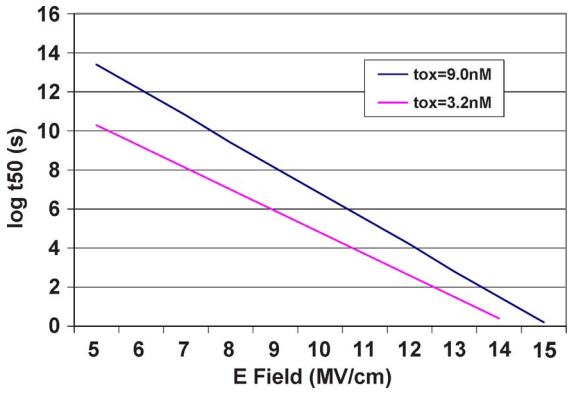

Fig. 7. *Time to 50% breakdown extrapolated to 5 MV/cm from [20] SiO_2 accelerated TDDB data > 10 MV/cm. Brief but frequent startup voltage excursions increasing E values to 10 MV/cm or less could have serious reliability consequences.*

2) For some technology nodes, TDDB dominates reliability lifetime limits, and no OD conditions are permissible, even in short-term standby circuit modes.
3) While ESD events that create voltage excursions up to $3 \times VCC_{\mathrm{MAX}}$ for 1 μs might be acceptable, frequent startup voltage excursions that increase maximum E_{ox} values $2\times$ for several milliseconds have potentially serious reliability consequences.

In order to mitigate the effects of TDDB during startup, the drain circuit must remain sufficiently conductive to allow the source voltage to track the gate voltage within V_{gs} limits. However, during standby with supply voltage on, a high stress TDDB condition exists if the current supplying the device source is stronger than that supplying the drain. Then the source current supply can pull the source potential all the way down to V_{ss}, creating a potential V_{gs} voltage in excess of specified foundry overdrive limits, an EOD condition. Because of this possibility, the analog designer needs to take extreme care in simulating all possible startup and standby conditions. Some circuit simulators are capable of setting flags if an excess V_{gs} condition occurs during startup or standby; most are not.

LP CMOS processes generally offer a thicker core oxide for operation at slightly higher V_{gs} limits and significantly reduced GBW performance. A much thicker second or third oxide device is usually available to satisfy higher voltage I/O requirements. The addition of an LP or second oxide device to a nanoscale HP core technology requires at least one additional thermal cycle for oxide growth and some degradation of the core device performance. Accurate simulation requires the use of models appropriate for the specific oxide combinations present in the process.

3) HCI Effects: Operation of bulk CMOS high dc gain, wide-dynamic-range output stages within the < 1 V HP technology VCC_{MAX} limits at the 45 nm technology node and beyond is a significant design challenge. It is therefore useful to understand both the specified limits of the process and how rapidly reliability stress factors increase as a limit is reached, or exceeded. All reliability stress factors do not impose equal device lifetime limits when supply voltages reach VCC_{MAX}.

Foundry-prescribed OD voltages VCC_{OD} in excess of VCC_{MAX} are usually specified so that no single reliability factor will cause unacceptable reliability lifetime degradation. For example, under certain conditions, HCI stress may, or may not, substantially degrade device lifetime at V_{ds} voltages (VCC_{EOD}) in excess of specified OD limits. Low HCI-stress factors include low drain current, long device L, certain drain voltage waveform shapes, etc. Other limits, such as the V_{gs} limit based on TDDB stress, have fewer low-stress factors. Reliability stress factors do not cause step changes in device lifetimes at the specified process voltage limits.

HCI in the case where $V_{sb} = 0$ is commonly known as the CHE mode of device degradation [22], [23]. In CHE, impact ionization results in carrier pairs generated near the drain end of the channel causing interface trap (N_{it}) and N_{ot} buildup. The buildup of N_{it} and N_{ot} near the gate–channel interface degrades the I–V characteristics and eventually results in device failure.

Cascode circuits where $V_{sb} > 0$ have been used for a long time in analog, voltage multiplier, and current-mode logic digital designs, to overcome V_{ds} limitations [24]. Operation at $V_{sb} > 0$ has been used by EEPROM circuit designers to obtain higher electron programming energies and currents. As V_{sb} is increased, the drain-bulk field increases, and one carrier from the first carrier pair generated by impact ionization near the drain may be accelerated toward the bulk with sufficient energy to generate a secondary carrier pair. One member of that pair is consequently accelerated upward, toward the gate acquiring high energy and capability for creating damage. In an NMOS device, this process is known as channel-initiated secondary electron generation, or CHISEL mode of HCI stress [25]. The process can also occur in PMOS devices at nearly the same voltage stress levels, with equally serious damage consequences [26].

a) HCI stress in CHE mode: In CHE, the buildup of N_{it} is directly proportional to transverse field strength (E_y) and to the ratio of bulk-to-drain current $(I_b/I_d)^n$, where n is slightly lower than 0.5. At low V_g, I_b/I_d is low, but E_y is high and proportional to $V_d - V_g$. Because E_y is high, impact ionization multiplication factors are high. However, very low drain current results in low I_b, even in the presence of a high avalanche multiplication factor. Conversely, for high V_g, $V_d - V_g$ is low, E_y is low, and low impact ionization rates result in reducing I_b to a small value. Most process qualification for HCI damage is usually with $V_{sb} = 0$ at maximum substrate current, where V_{gs} (or $V_d - V_g$) is approximately $V_{ds}/2$. Testing at maximum substrate current is not the condition that corresponds to

the maximum HCI damage rate when V_{sb} is more than a small fraction of V_{ds}. At high V_{sb}, gate current is the best indicator for damage in an NMOS device.

The relative magnitudes of the device terminal voltages determine the location of the hot carrier injection along the channel and have a substantial effect on the type and degree of degradation. With $V_{sb} = 0$, lateral field (E_x) is relatively independent of V_g and generally peaks between the drain junction and the edge of the lightly doped drain region. High transverse (E_y) fields result from $V_d - V_g$. E_x and E_y work together to accelerate carriers in the channel and create HCI. As V_g approaches V_d, E_y is substantially reduced but E_x fields are still high. Substrate current decreases significantly as N_{it} and N_{ot} peak values diminish but widen, and move toward the center of the channel. There, the N_{it} and N_{ot} peak values become more efficient in degrading analog g_m and V_t characteristics. Increasing device length by 25% to 50% over minimum reduces maximum E_x at the drain. However, decreasing maximum $V_d - V_g$ reduces only E_y. While bulk current is reduced to a small value as V_g approaches V_d, the damage rate for analog g_m and V_t may be reduced by only 20% because E_x remains relatively unchanged [27].

Peak V_{ds} foundry voltage limits for HCI stress are commonly based on a digital-design requirements where I_{dsat} is degraded at end-of-life by 10%. However, excess device g_m, noise, and V_t shifts are consequences of increased N_{it} and N_{ot} trap levels prior to device failure. A 20% increase in V_{ds} beyond VCC_{OD} limits can easily reduce the lifetime of an NMOS by a factor of ten because of the exponential dependence of damage rate on V_{ds}. The output of an amplifier or voltage-controlled oscillator (VCO) with an inductive load may have peaks above the supply voltage during normal operation. In that case, the use of peak values of V_{ds} for evaluating HCI stress conditions may be a bit conservative. In circuit designs where an extra 50–100 mV of output peak voltage could provide a valuable increase in dynamic range, it is useful to estimate the effect of the additional peak voltage on the damage rate.

b) Obtaining additional device headroom in CHE mode: HCI damage where $V_{sb} = 0$ progresses at a rate proportional to $e^{V_{ds}/C} * f(t)$, where C is the voltage stress coefficient $B * VCC_{MAX}$. The value of C is obtained from foundry reliability data, and typical values of $B = C/VCC_{MAX}$ range from 0.1 to 0.2 at the 100 nm node. The second term $f(t)$ is a slowly varying nonlinear function of time, such as t^n. Typically n ranges from 0.2 to 0.5.

When an inductive load is present, such as in a high-frequency VCO, the drain terminal voltage can easily exceed the supply voltage. Extra headroom can be obtained when peak V_d values for sinusoidal waveforms are allowed to exceed VCC_{MAX} by a value up to an excess overdrive limit V_{EOD}. If the relationship between stress and $V_d - V_s$ were linear, then the HCI damage rate would be independent of the amplitude of the sinewave. In that case V_{EOD} could exceed VCC_{MAX} by the zero-to-peak amplitude of the sinewave, V_{0-pk}. The stress relationship is unfortunately not linear; however, given a value for the coefficient C, the value of V_{EOD} can be determined.

The damage rate equation can be used to find the voltage difference (ΔV) between the average sinewave voltage and VCC_{MAX}, where the damage rate for the sinewave is the same as it is for dc. Once ΔV is determined, then V_{EOD} is related to VCC_{MAX} by

$$V_{EOD} = VCC_{MAX} - \Delta V + V_{0-pk}. \tag{2}$$

Substituting $VCC_{MAX} - \Delta V + V_{0-pk} \times \sin(\omega t)$ for V_{ds} in the first term of the steady-state damage rate

$$e^{\frac{VCC_{MAX} - \Delta V + V_{0-pk}\sin(\omega t)}{B \cdot VCC_{MAX}}}. \tag{3}$$

Equation (3) is then integrated over one cycle of the sinewave, and the result is equal to the dc damage rate term $e^{V_{ds}/C}$ or $e^{VCC_{MAX}/C}$. When the variable changes $V_{0-pk} = A * VCC_{MAX}$, $C = B * VCC_{MAX}$, and $\Delta V = E * VCC_{MAX}$ are made, the equality between ac and dc damage rate is

$$\int_0^{2\pi} \frac{1}{2\pi} e^{1/B - E/B + (A/B)\sin(\theta)}\, d\theta \equiv e^{1/B}. \tag{4}$$

Cancelling out the $e^{1/B}$ terms, factoring out the $e^{E/B}$ term, and taking the natural log, the last equation reduces to

$$\frac{E}{B} = \ln\left\{\frac{1}{2\pi}\int_0^{2\pi} e^{(A/B)\sin(\theta)}\, d\theta\right\}. \tag{5}$$

The value of E/B can be obtained using numerical integration methods to evaluate the sine integral from zero to 2π, and taking the log of the integration result. In Table 3, E/B as a function of A/B is shown for a practical range of A/B (recall $A/B = V_{0-pk/C}$) with $V_{sb} = 0$. Where $A/B = 2$ $(A/B = 0.2/0.1)$ and $VCC_{MAX} = 1.0$ V, an extra 119 mV of headroom can be gained; then $V_{EOD} = 1.119$ V. For a higher amplitude sinewave, where $A/B = 4$, the gain in headroom is 159 mV. Note that the table value $E/A = \Delta V/V_{0-pk}$. Therefore, $\Delta V = E/A * V_{0-pk}$ and $V_{EOD} = VCC_{MAX} - \Delta V + V_{0-pk}$. As the sinewave amplitude increases, the fraction of amplitude that contributes to increased stress (E/A) increases, but not fast enough to prevent V_{EOD} from increasing beyond VCC_{MAX}.

Table 3 Sinewave Excess Overdrive Factor E/B Is Computed From A/B, Where $A = V_{0-pk}/VCC_{MAX}$ and $B = C/VCC_{MAX}$. Values of ΔV^* and Excess Overdrive Voltage Limit V^*_{EOD} Are Shown for a $VCC_{MAX} = 1.0$ V, But They Can Be Computed for Any Process When VCC_{MAX} and C, the Foundry-Specified Voltage Stress Coefficient, Are Known

A/B (V_{0-pk}/C)	E/B (ΔV/C)	E/A ($\Delta V/V_{0-pk}$)	ΔV^* (V) for $VCC_{MAX} = 1.0$V	$V_{EOD}{}^*$ (V) for $VCC_{MAX} = 1.0$V
1.0	0.236	0.236	0.023	1.077
2.0	0.824	0.412	0.081	1.118
3.0	1.585	0.523	0.157	1.142
4.0	2.425	0.606	0.241	1.158
5.0	3..305	0.661	0.329	1.170

Some pipeline ADC stages and analog sample-and-hold circuits have signal distributions that are substantially uniform over a moderately wide voltage range. For such a stage, A is related to the peak-to-peak amplitude (V_{p-p}), as in the square wave case by $V_{p-pMAX} = 2A * VCC_{MAX}$. In a pipeline AD stage with uniform signal distribution, the output is approximated by a square wave with uniform voltage probability density between $-A$ and $+A$

$$\frac{E}{B} = \ln\left\{\frac{1}{2A}\left[\int_{-A}^{A} e^{(x/B)}\,dx\right\}\right. . \tag{6}$$

Using an analytical solution for the integral yields the following result for E/B:

$$\frac{E}{B} = \ln\left\{\frac{1}{2}\frac{B}{A}\left(e^{A/B} - e^{-A/B}\right)\right\}. \tag{7}$$

For higher amplitudes, such as $A/B = 4$, 208 mV of extra headroom is gained, in the $VCC_{MAX} = 1.0$ V example of Table 4. Excess headroom for the random voltage distribution case is higher than in the sinewave case.

Table 4 Constant-Amplitude Squarewave Excess Overdrive Factor E/B Computed From A/B, Where $A = V_{p-p}/(2VCC_{MAX})$ and $B = C/VCC_{MAX}$. Values of ΔV^* and Excess Overdrive Voltage Limit V^*_{EOD} Are Given for $VCC_{MAX} = 1.0$ V, But Can Be Computed for Any Process When Both E/A and VCC_{MAX} Are Known

A/B (V_{0-pk}/C)	E/B (ΔV/C)	E/A ($\Delta V/V_{0-pk}$)	ΔV^* (V) for $VCC_{MAX} = 1.0$V	$V_{EOD}{}^*$ (V) for $VCC_{MAX} = 1.0$V
1.000	0.161	0.161	0.016	1.084
2.000	0.595	0.298	0.060	1.140
3.000	1.206	0.402	0.121	1.179
4.000	1.920	0.480	0.192	1.208
5.000	2.697	0.539	0.270	1.230

This is because the probability of the waveform's being at any voltage at any time is uniform, while the sinewave spends a larger fraction of time at peak amplitudes.

This general method can be used for various signal distributions and other reliability stress factors with similar exponential dependence on voltage. For an asymmetric digital subscriber line signal-processing case, having typically 2 : 1 peak-to-average ratios, the signal statistics are even more favorable to excess overdrive voltage allowance. Good analog design procedure, where EOD operating modes are planned, requires extensive consultation with device engineers who understand all of the relevant reliability factors, an analysis of how the EOD will affect total chip reliability, backup by process qualification data, and an awareness that operation too far outside the normal HP limits may require consideration of additional damage modes in the device lifetime predictions.

c) HCI stress in CHISEL mode: Series-connected CMOS devices, as in current–source–differential pair and cascode circuit combinations, reduce V_{ds} drops across the individual devices. However, while V_{ds} is reduced, V_{sb} is increased in one or more of the series-connected devices. When $V_{sb} > 0$, the hot carriers have higher energy as a result of high transverse field (E_y) acceleration and cause more HCI (CHISEL mode) damage. High Ey also results in more NBTI damage. Voltage ranges of interest in circuits with series-connected CMOS include $V_{db} > VCC_{MAX}$ and $V_{sb} > 0$ up to $V_{sb} = VCC_{MAX}$. For those ranges, accurate data exist for model parameter extraction, but in many cases, long-term reliability data do not. Such data are vital in overcoming headroom limitations that result from future $1/\kappa$ voltage scaling. Early access to $V_{sb} > 0$ preliminary reliability projections is required for nanoscale analog CMOS circuit design.

When V_{sb} voltages reach half the supply voltage, the damage rate compared to the condition $V_{sb} = 0$ can more than double. When V_{sb} approaches VCC_{MAX}, the damage rate compared to $V_{sb} = 0$ can increase by an order of magnitude or more. At low values of V_{sb}, the damage is concentrated near the drain end of the channel and moves toward the center as V_{sb} increases, resulting in the generated traps' becoming more efficient in degrading V_t, I_{dsat}, and g_m [28], [29].

For NMOS in CHISEL, the HCI damage buildup is proportional to gate current (I_g), not substrate current. The damage buildup rate is proportional to I_g^n, where n is approximately 0.5 [30]. The maximum damage for NMOS in CHISEL is not at maximum substrate current, where V_{gs} is approximately $V_{ds}/2$ and HCI device qualification is usually performed. It occurs where V_{gs} is near V_{ds}. Low values of $V_d - V_g$ do not mitigate high HCI damage in CHISEL mode.

For PMOS in CHISEL, the maximum damage rate does not necessarily correspond to maximum gate current conditions. In a PMOS device, maximum damage can occur at a gate voltage where gate current is near zero. This

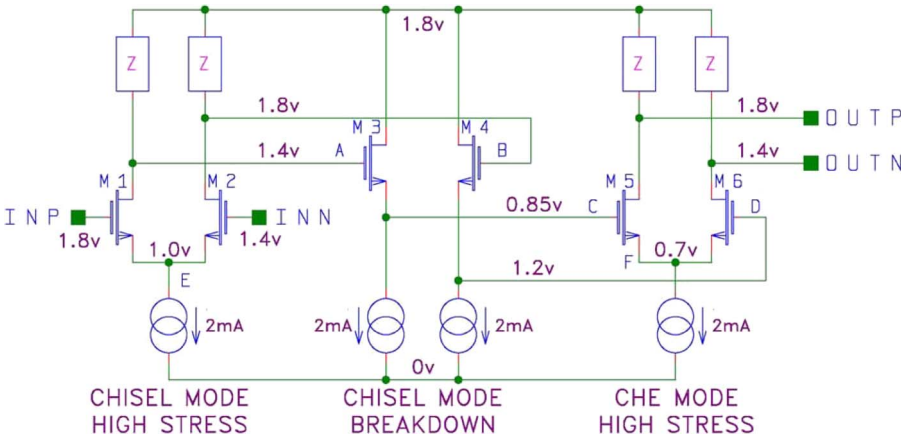

Fig. 8. *Series-device circuit configurations using 100 nm core devices operated at $VCC = 1.8$ V, which could potentially have CHE or CHISEL mode stress with $V_{sb} \gg 0$ and V_{db} approaching 1.8 V.*

has been explained as the result of the current cancellation effect from the two carrier currents of opposite polarity's combining to generate the damage [31]. Damage under these conditions is more closely correlated to the sum of the absolute value of both hole and electron gate currents. Maximum stress conditions may be difficult to determine from external device currents because the hole and electron currents are of opposite sign.

Some series-device circuit configurations using 100 nm core devices operated at $VCC = 1.8$ V, which could have potentially high CHE or CHISEL mode stress with $V_{sb} > 0$, are shown in Fig. 8. As in CHE, slight increases in device length will help mitigate high CHISEL mode stress. However, device operation where V_{db} or V_{gb} exceeds the HP VCC_{MAX} or VCC_{OD} specifications for the process (EOD conditions) must be subjected to careful reliability analysis.

4) Negative Bias Temperature Instability: NBTI occurs in PMOS devices as interface traps (N_{it}) are formed at high gate voltage and at high temperatures. This process does not require the presence of high E_x and the resulting hot carriers. Trap formation creates V_t shifts and g_m degradation. Device qualification is performed at high temperatures, high V_{gs}, and, usually, $V_{sb} = 0$.

NBTI occurs when Si_3-Si-H bonds at the Si–SiO_2 interface are broken by cold holes that originate in the inversion layer, causing hydrogen to be released. The resulting dangling bonds act as interface traps Si-(N_{it}), where (N_{it}) indicates that the SiO_2 molecule is absent. The NBTI damage rate increases as the gate voltage increases. As the nitrogen concentration in the oxide is increased to raise the gate dielectric constant, NBTI degradation also increases [32]. At a given V_{gs} and with increasing V_{sb}, more N_{it} damage is created as Si-O bonds are broken by hot holes. The damage shifts from the drain end toward the center of the channel, and both g_m degradation and V_t shifts are accelerated [33].

A number of alternative geometries for device formation have been proposed for ultrascaled MOS devices beyond the 22 nm technology node. Unfortunately, NBTI effects significantly increase in narrow-width planar, triple-gate, and surround-gate MOSFETs as device dimensions are scaled down. Increasing device width W well beyond minimum dimensions is an analog design option that tends to mitigate NBTI effects in highly scaled devices. This occurs because the hydrogen diffusion length L_D grows with time (as $D_H * t^{1/4}$, where D_H is the hydrogen diffusion constant) and the damage rate increases in proportion to L_D/W [34].

5) Electromigration: EM failures are induced when current through a conductive material causes the formation of small voids. As a result, the material resistance gradually builds up to the point where the increase in resistance is high. At that point, thermal heating causes a rapid acceleration of the process. EM design guidelines generally define conductor end of life to occur at a specific percentage change in resistance, such as 10%. Beyond the 100 nm technology node, the EM problem is considered within the industry to be severe. Analog circuit design dimensioning of HF device width, length, and number of gate fingers must be done concurrently with layout-related EM considerations. Otherwise, a number of layout and/or design changes will be required when the layout is complete and a number of EM rule violations are found using postlayout DFR tools.

HF analog device layouts rely on shared drain structures to improve device GBW. The benefits of reduced parasitic capacitance from shared drain structures have increased in importance as devices and metal interconnects are scaled down. Device layouts that do not share drains, but instead alternate source and drain connections, and suffer from progressively increasing metal height-to-spacing ratios

under scaling. The result is high drain-source parasitic capacitance.

EM qualification testing is normally done at very high elevated temperatures, such as 573 K, using Black's model for the time to 50% failure (t_{50}) [35]

$$t_{50} = Aj^{-n}e^{E_a/k_BT}. \quad (8)$$

Here, A is a constant based on the conductor material and geometry, E_a is the activation energy of the process, and k_B is the Boltzmann constant. The activation energy typically ranges from 0.5 to 1.0, while the exponent n ranges from 2.0 to 1.0 depending on the type of structure (contact, via, or metal) and material. With the present ability to fabricate more than 1 billion devices on a chip, operating temperatures on high-speed ICs are going up and EM current temperature coefficients, which are inversely proportional to temperature, are going down. An increase in temperature from 100 to 125 °C can easily reduce t_{50} by a factor of three. The effects of strong dependence of EM limits on temperature can be mitigated by spreading out circuits that generate high power in small areas.

Maximum EM current limits for long metal lines increase in cases where the line length (L) is short. Bletch explained this effect as back stress in short metal lines countering, to some extent, the effect of forward current stress [36]. However, the mitigating effect of short L on back stress becomes less significant as the dielectric material becomes softer with the Lo-K dielectric material targeted for use in advanced technology nodes [37].

The consistency of analog device current density over several process nodes allows some general predictions to be made regarding the future of copper damascene as a suitable interconnect material for analog HF devices. Consider a two-gate-finger high-GBW NMOS device operating at a moderate current density of 10 μA/***F***, with a typical length of 1.25 ***F***. If the drain is shared, the current density at the drain is 20 μA/***F***, and the intermediate metal can be up to 2 ***F*** in width. Both EM maximum current coefficient (mA/μm) and device intermediate metal width have scaled in proportion to $1/\kappa$ for each of the past several copper-interconnect technology nodes. EM-limited HF-device drain current has therefore been scaling approximately as $1/\kappa^2$.

The maximum device W (in ***F*** units) is plotted for one, two, and three layers of intermediate metal (Mx) in Fig. 9. The constant value of 5.5 ***F*** of the red curve represents the minimum width for a two-contact device that can be constructed at each technology node. One layer of Mx is adequate to carry the current at 65 nm; two layers are required at 45 nm and three at 32 nm. Because maximum width falls below the minimum width required to accommodate two drain contacts beyond the 32 nm node, conductor material or device geometry changes will be required at the 22 nm node.

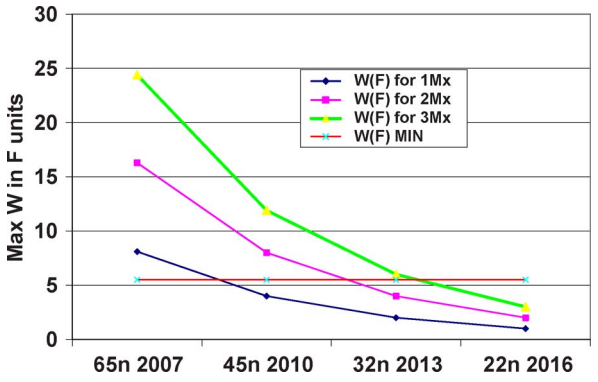

Fig. 9. ***Maximum device W (in F units) for one, two, and three layers of Mx. Calculations assume an industry average EM current limit for short-length wires of 1 mA/μm at the baseline (65 nm) technology node. When maximum device W falls below minimum width required for two drain contacts at W ≈ 5.5 F (W MIN—red line), an additional metal layer is required.***

Interconnect problems may be a more serious threat than devices to the scaling trends predicted by Moore's law. Eventually, dual damascene copper metallization schemes must be replaced by another technology, such as SWCNs [38]. Metal resistance and EM heating effects increase inversely with electron mean-free paths in a conductor. Mean-free paths in copper and similar materials are in the range of a few tens of nanometers. If mean-free paths in excess of 5 μm can be achieved at the 22 nm technology node, SWCN-bundle R-C delay performance will offer a significant improvement over copper for long ($\gg 5$ μm) lines. Many formidable challenges must be overcome before the fabrication of SWCN bundles can be compatible with conventional CMOS processes. These include reducing the high temperatures (> 600 °C) presently required to grow the nanotubes [39]. The low-resistance properties of SWCN graphene bundles are shared with GNRs [40]. Because GNRs are planar, they may be patterned using nanoscale lithography methods.

D. Nanoscale MOS Transistor Performance

Clearly, transistor performance is changing as CMOS technologies are scaled down into the nanoscale regime (see, for example, [41]). Optimizing MOSFETs is challenging due to conflicting device performance requirements for digital and analog circuits. For digital circuits, the I_{on}/I_{off} tradeoff dominates. The challenge lies in reducing I_{off} while minimizing the intrinsic delay and achieving high I_{on}. For analog and RF circuits, cutoff frequency (f_T), intrinsic gain (g_m/g_{ds}), linearity, noise, and device mismatches constitute the performance metric.

In this section, the ways the different electrical parameters of MOSFETs change as the active area of the transistor is scaled down are discussed.

1) Intrinsic Speed: The frequencies of parasitic poles in analog CMOS circuits are often related to the cutoff frequency f_T of the transistors. In this case, the maximum speed of a feedback amplifier is related to f_T since a certain phase-margin is required for stability.

The cutoff frequency is one of the few parameters that improves as channel length is scaled down. The maximum cutoff frequency f_{Tmax} is limited by the saturation velocity of the charge carriers in the channels and is given approximately by

$$f_{\mathrm{Tmax}} = \frac{v_{\mathrm{sat}}}{2\pi L_{\mathrm{eff}}} \tag{9}$$

where v_{sat} is the saturation velocity of the charge carriers and L_{eff} is the effective channel length (see, for example, [42]). In practice, the transistors are operated at lower frequencies due to, for example, parasitic capacitances. Let us denote the practical maximum operating frequency of a single transistor, assuming that there are no interconnects, as the intrinsic speed f_i. To estimate f_i, we use a diode-connected transistor excited by a current source and measure the -3 dB frequency of the voltage across the device. The bandwidth of the small-signal transfer function v_{out}/i_{ac} of this circuit is used as the estimate of f_i. Under certain assumptions, the intrinsic speed increases as CMOS technology is scaled down. To show this, an approximate expression for f_i is derived from the small signal equivalent circuit of the circuit in Fig. 1. The resulting estimate for f_i is

$$f_i = \frac{1}{2\pi}\frac{g_m + g_{ds}}{C_{gs} + C_{db}} \cong \frac{1}{2\pi}\frac{g_m}{C_{gs} + C_{db}} \tag{10}$$

where g_m, g_{ds}, C_{gs}, and C_{db} are the transconductance, channel conductance, gate-source capacitance, and drain-bulk capacitance of the transistor, respectively. This expression indicates that the speed is proportional to the ratio of the transconductance and the sum of the capacitors. This ratio depends on the biasing of the transistor and, in particular, the inversion level of the transistor, which is represented here by a voltage defined as

$$V_{il} \equiv \frac{2I_D}{g_m} \tag{11}$$

where I_D is the transistor drain current.

Because V_{il} is a measure of the inversion level of the transistor channel, the subscript *il* is included. Note from (11) that V_{il} can be easily calculated. The factor two is included to make V_{il} an estimate of the gate-voltage overdrive and the saturation voltage. Note that V_{il} is equal to the simple expression used for the saturation voltage in the square-law model.

In most of the allowed biasing range, f_i increases as V_{il} increases. Hence, to achieve high speed, a high inversion level is needed. As the technology is scaled down, it turns out that the transistor transconductance does not change for a given current density, which can be understood by observing the following fundamental expression of the strong-inversion drain current in the case of a fully velocity-saturated transistor channel (see, for example, [42]):

$$I_D = W c_{\mathrm{ox}}(V_{gs} - V_t) v_{\mathrm{sat}}. \tag{12}$$

Here, W is the width of the transistor gate; c_{ox} is the oxide capacitance per gate area; V_{gs} is the gate-source voltage; and V_t is the threshold voltage. To obtain an expression for the transconductance, the derivative of (12) with the respect to V_{gs} is taken, and the result is

$$g_m = W c_{\mathrm{ox}} v_{\mathrm{sat}}. \tag{13}$$

Here, the assumption has been made that the threshold voltage does not depend on the gate-source voltage. For ideal scaling, $W \propto L$ and $c_{\mathrm{ox}} \propto 1/L$. Hence, (13) shows that the transconductance does not depend on L.

Still, capacitors do have the potential to be made smaller as technologies are scaled down. Under ideal scaling, C_{gs} scales proportional to L. The plot in Fig. 10 shows that C_{gs} is reduced as technology is scaled down, even when transistor widths are kept independent of the technology node. This is because the CMOS foundries have not employed ideal scaling. The oxide capacitance per unit area is currently scaled down more slowly than the gate area.

As expected, the drain-bulk capacitance is also reduced as the technology is scaled down, as shown in Fig. 11. Since transistor capacitances decrease as the technology is scaled down and transconductance does not change, intrinsic

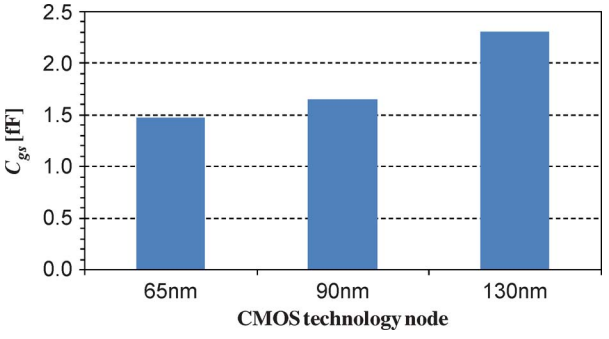

Fig. 10. *Strong inversion gate-source capacitance C_{gs} for three different CMOS technology nodes. Minimum gate lengths and a gate width of 2 μm were used. Calculated using SPICE.*

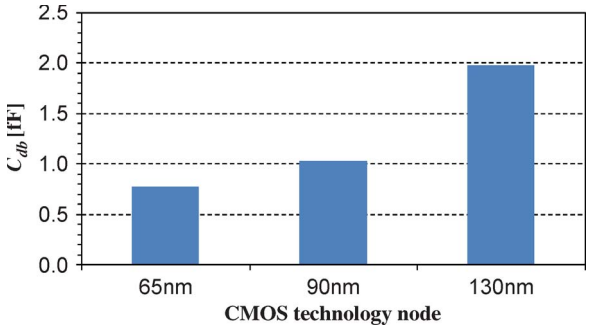

Fig. 11. *Drain-bulk capacitance C_{db} at zero drain-source voltage for three different CMOS technology nodes. Minimum gate lengths and a gate width of 2 μm were used. Calculated using SPICE.*

speed will increase with scaling. This is illustrated in Figs. 12 and 13.

To develop a better sense of the improvement in f_i as technology is scaled down, the relative f_i normalized to the values obtained for the 0.13 μm technology node has been plotted in Fig. 13. At the commonly used inversion level of $V_{DD}/8$, 65 and 90 nm technologies are potentially 2.4 and 1.9 times faster, as compared to the 0.13 μm technology. Note also that the simulations indicate that this speed improvement can be further boosted by an additional increase in V_{il} (if possible).

2) Intrinsic Gain: Degradation of the transistor's intrinsic gain is considered one of the major challenges in relation to the design of high-performance analog circuits in scaled-down technologies. Device engineers encounter a tradeoff between cutoff frequency and intrinsic gain when they are optimizing the device design at a given technology node. As pointed out in [43], maintaining the intrinsic gain across technology nodes is not feasible since

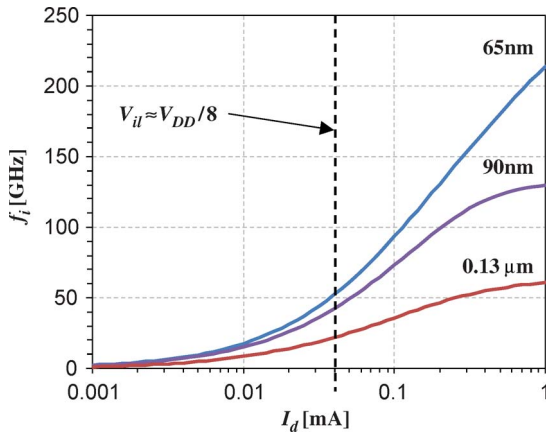

Fig. 12. *SPICE simulations of NMOS transistor intrinsic speed f_i versus drain current for three different CMOS technology nodes. Minimum gate lengths and a gate width of 2 μm were used.*

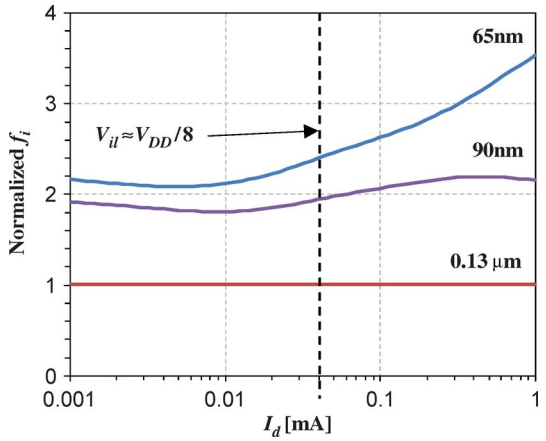

Fig. 13. *SPICE simulations of NMOS normalized f_i versus drain current for three different CMOS technology nodes. Minimum gate lengths and a gate width of 2 μm were used.*

it requires increasing the threshold voltage as gate lengths are scaled down—an increase that, in most cases, is not acceptable. Hence, the situation where the intrinsic gain is reduced for every technology node has to be tolerated. This is illustrated in Fig. 14. Notice that, at an inversion level of $V_{DD}/8$, the intrinsic gain for the 65 nm technology is reduced by almost 80% as compared to the 0.13 μm technology. Notice also that the maximum intrinsic gain for all technologies plotted appears close to the $V_{DD}/8$ inversion level.

3) Gate Leakage Current: Gate leakage current cannot be neglected in the process of designing analog circuits in nanoscale CMOS technologies. In modern state-of-the-art CMOS technologies, the dominating transport mechanism for gate leakage is tunneling through the thin gate oxide (see, for example, [44]–[46]). Gate leakage current is extremely sensitive to the oxide thickness. The introduction of oxynitride gate formulations at the 90 nm technology node allowed target gate capacitance to be achieved with a greater physical thickness, thereby increasing tunneling barrier width [47]. However, physical thickness was still less than at the 130 nm technology node. It is evident from Fig. 15 that, at the 65 nm technology node, the gate leakage current has increased by more than six orders of magnitude as compared to the level for the 0.18 μm technology node. In addition to this oxide thickness dependency, the gate leakage current depends on the gate-source voltage, the gate-drain voltage, and the gate area.

One obvious implication of a nonzero gate leakage current is that the gate terminal now includes a tunnel conductance in parallel with the traditional capacitances. To investigate the implications of the gate leakage current, study the simple circuit in Fig. 16(a), which contains only a voltage source $v_G (= V_{GS} + v_{gs})$ and an NMOS transistor.

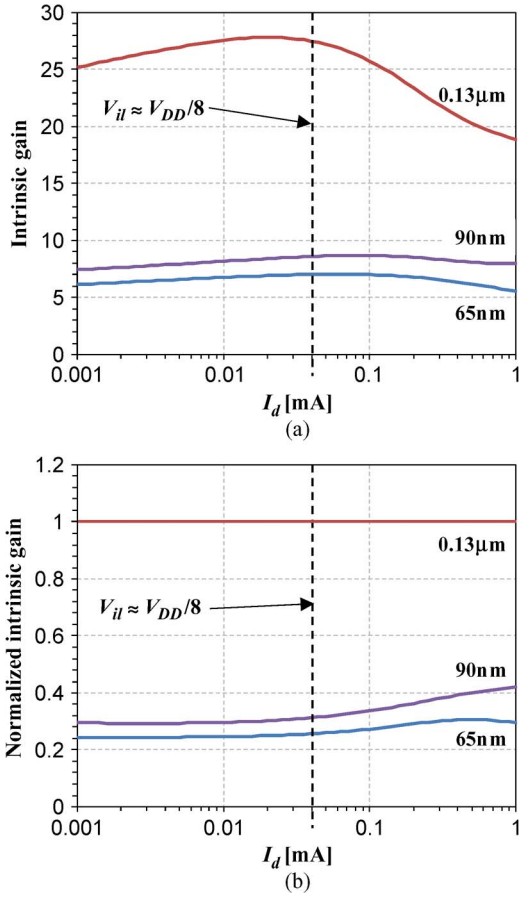

Fig. 14. *SPICE simulations of (a) NMOS transistor intrinsic gain versus drain current for three different CMOS technology nodes and (b) relative intrinsic gain normalized to the 0.13 µm technology node. Minimum gate lengths and a gate width of 2 µm were used.*

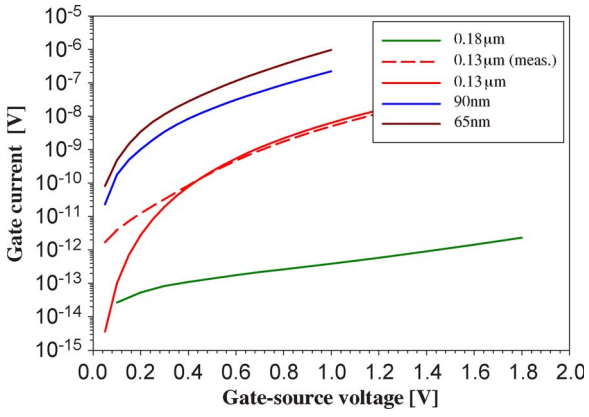

Fig. 15. *Measurements and SPICE simulations of gate leakage current for different CMOS technologies. Gate dimension for 0.18 µm technology: W = 100 µm, L = 0.18 µm. Gate dimensions for all other technologies: W = 100 µm, L = 0.13 µm. Source, drain, and bulk voltages set to 0 V.*

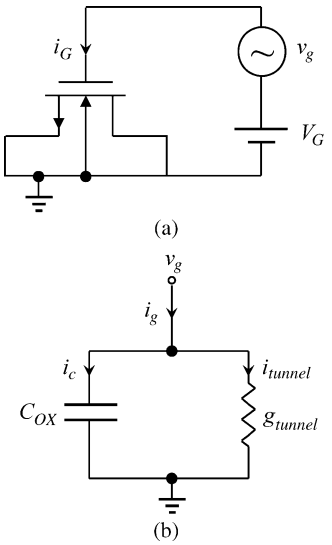

Fig. 16. *A simple circuit (a) for investigation of the impact of gate leakage current containing one NMOS transistor and a voltage source and (b) its small signal equivalent, based on an assumption of low to moderate frequencies and strong inversion.*

Assuming that the NMOS transistor is biased in strong inversion, the small signal operation of the circuit can be approximated by the linear circuit shown in Fig. 16(b). Since i_c depends on frequency while i_{tunnel} does not, there is a characteristic frequency f_{ch} where the two currents have equal magnitude [41]. This frequency is given by

$$f_{\text{ch}} = \frac{g_{\text{tunnel}}}{2\pi C_{\text{OX}}}. \tag{14}$$

Since both the tunnel conductance g_{tunnel} and C_{OX} are proportional to the gate area, f_{ch} does not depend on the gate area. Notice that, for frequencies much larger than f_{ch}, i_c becomes much larger than i_{tunnel}. In these cases, the gate current is mostly capacitive and the gate behaves as a conventional MOSFET gate. To extract f_{ch} from SPICE simulations, the phase of the ac small signal gate current versus frequency is plotted. The frequency at which the phase shift is 45° corresponds to f_{ch}. A plot of the simulated f_{ch} for different CMOS technology nodes is shown in Fig. 17. The figure indicates that, for signal frequencies above 1 MHz, even the gates of 65 nm transistors, which have the highest gate leakage currents, behave like conventional MOSFET gates. In this case, the gate leakage current does not degrade the high-frequency performance of circuits.

4) Nonlinearity: The nonlinearity of a CMOS analog signal-processing circuit depends on the nonlinearity of the transistor drain current and also on some other factors. To study how this nonlinearity changes with technology,

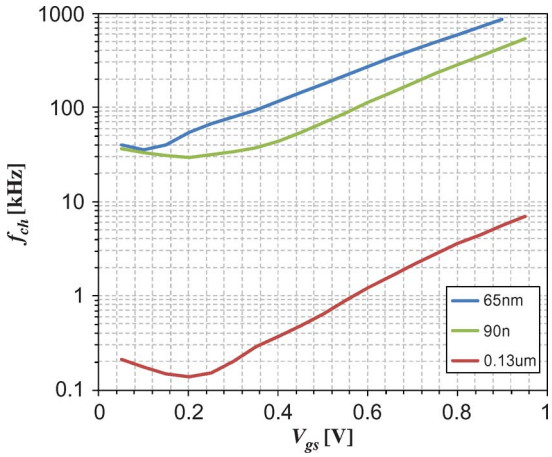

Fig. 17. *Simulated, using SPICE, f_{ch} versus V_{gs} for different CMOS technologies. $V_{DS} = V_{DD}/2$.*

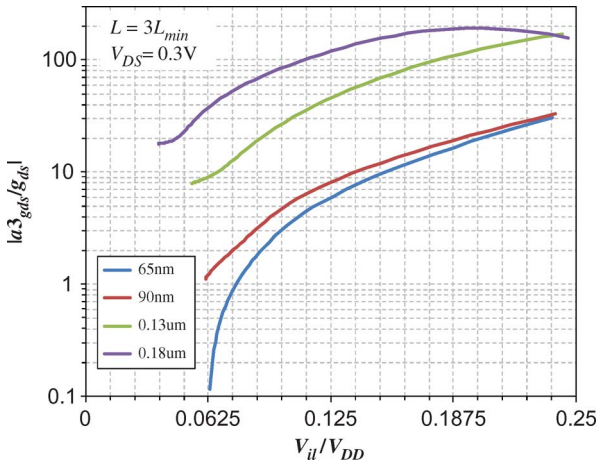

Fig. 18. *Simulated, using SPICE, $a3_{gds}/g_{ds}$ versus inversion level for different CMOS technology nodes.*

the Taylor expansion of the drain current around the bias point is written as

$$\begin{aligned} i_D(v_{GS}, v_{DS}, v_{BS}) &= i_D(V_{GS}, V_{DS}, V_{BS}) + g_m v_{gs} + a2_{g_m} v_{gs}^2 + a3_{g_m} v_{gs}^3 + \cdots \\ &+ g_{ds} v_{ds} + a2_{g_{ds}} v_{ds}^2 + a3_{g_{ds}} v_{ds}^3 \, g_{mb} v_{bs} + \cdots \\ &+ a2_{g_{mb}} v_{bs}^2 + a3_{g_{mb}} v_{bs}^3 + a2_{g_m g_{ds}} v_{gs} v_{ds} + \cdots \\ &+ a3_{2g_m g_{ds}} v_{gs}^2 v_{ds} + a3_{g_m 2g_{ds}} v_{gs} v_{ds}^2 + \cdots \\ &+ a2_{g_m g_{mb}} v_{gs} v_{bs} + a3_{2g_m g_{mb}} v_{gs}^2 v_{bs} + \cdots \\ &+ a3_{g_m 2g_{mb}} v_{gs} v_{bs}^2 + a2_{g_{ds} g_{mb}} v_{ds} v_{bs} + \cdots \\ &+ a3_{2g_{ds} g_{mb}} v_{ds}^2 v_{bs} + a3_{g_{ds} 2g_{mb}} v_{ds} v_{bs}^2 + \cdots \\ &+ a3_{g_m g_{ds} g_{mb}} v_{gs} v_{ds} v_{bs} + \cdots \end{aligned} \tag{15}$$

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

$$a2_{g_m} = \frac{1}{2} \frac{\partial^2 i_D}{\partial v_{GS}^2} \bigg|_{\substack{v_{GS}=V_{GS} \\ v_{DS}=V_{DS} \\ v_{BS}=V_{BS}}} \tag{16}$$

$$a3_{g_{ds}} = \frac{1}{6} \frac{\partial^3 i_D}{\partial v_{DS}^3} \bigg|_{\substack{v_{GS}=V_{GS} \\ v_{DS}=V_{DS} \\ v_{BS}=V_{BS}}} \tag{17}$$

$$a3_{2g_m g_{mb}} = \frac{1}{6} \frac{\partial^3 i_D}{\partial v_{GS}^2 \partial v_{BS}} \bigg|_{\substack{v_{GS}=V_{GS} \\ v_{DS}=V_{DS} \\ v_{BS}=V_{BS}}}. \tag{18}$$

In differential analog circuits, often the third-order terms in (15) dominate the harmonic distortion of the transistor drain current. Therefore, it is relevant to study how the third-order coefficient $a3_{gds}$ varies with technology, relative to the channel conductance g_{ds}. The channel conductance is a complex function of biasing, transistor dimensions, and technology. If an assumption is made that a gate length related to the minimum length of the technology L_{min} is used—for example, $L = 3L_{min}$—and the inversion level is proportional to V_{DD}, simulation results indicate that g_{ds} increases as the technology is scaled down. If we scale V_{DS} at the same rate as V_{DD}, which is probably the most common scenario, simulations indicate that the ratio $a3_{gds}/g_{ds}$ is almost independent of technology at an inversion level of $V_{DD}/8$. However, if a way is found to keep the drain-source voltage of the transistors unchanged, the ratio improves (decreases) as the technology is scaled down, as shown in Fig. 18. One way to take advantage of the reduced nonlinearity at constant V_{DS} when the technology is scaled down is to operate critical parts of the analog circuitry at a supply voltage higher than the nominal supply voltage for the CMOS technology used. More on this can be found in Section II-C.

III. CIRCUIT TRENDS

A. Design Techniques to Minimize Non-Idealities in Nanoscale CMOS

There are a number of circuit design techniques employed to address the major nonidealities of nanoscale processes when they are used in SOC designs. The most significant nonidealities, in order, are: a) poor predictability of process parameters, especially those determining impedances and capacitances, b) device mismatches due to both random variances and systematic and stress-related mismatches, and c) low transistor output impedances that are bias and temperature dependent. First, general circuit design approaches that minimize non-idealities are

discussed. Secondly, system-level approaches and especially the use of digital calibration are addressed. One of the more subtle, but nevertheless important, consequences of nanoscale technologies has to do with the system design methodology. For example, system design and calibration approaches that were too complicated to be considered in traditional approaches now not only become possible but can often be implemented with only a minimal power and area penalty due to the minimal size of nanoscale digital calibration logic. One example of this is the inclusion of skew adjust circuitry for time-interleaved ADCs, which can be included in nanoscale technologies, with the only significant penalty being additional design time. Another example is where the addition of small digital calibration circuits for mismatch allows for significantly smaller analog device sizing, which in turn often reduces power.

1) Circuit Design Techniques: There are a number of circuit design approaches used to minimize nonidealities; our first discussion has to do with current biasing approaches. There are three different popular biasing approaches: constant current, constant voltage, and constant gm.

a) Bias approaches: The constant-current approach, is based on establishing on-chip accurate and temperature-insensitive bias currents. This approach is based on applying an accurate bandgap voltage (obtained from an on-chip generator) across an accurate off-chip resistor; this generates an accurate current, which is copied and sent to various on-chip blocks using parallel current sources and current mirrors. A second alternative is similar but uses an on-chip resistor rather than an off-chip resistor, as in the previous approach. The on-chip resistor (which may actually be one of the resistors in the bandgap circuit) will have large process inaccuracies (±20%) but will still have good matching to other on-chip resistors; the voltages generated across other matched on-chip resistors can have accuracies better than 2%. In one variant of this method, a digitally programmable current mirror is placed between the current generator and the second resistor, resulting in a digitally programmable reference voltage.

A third current bias approach is to use a constant-gm bias current (see, for example, [3] and [48]); in this approach, bias currents are generated that, when used to bias transistors and/or amplifiers, result in predictable and temperature-insensitive transistor transconductances, and, to a lesser degree, bandwidths.

b) Feedback and replica biasing: The use of feedback in ICs to help minimize processing errors is very prevalent and powerful. An example is the voltage reference circuit shown in Fig. 19. In this example, a fully differential op-amp is used to take a single-ended voltage reference, which is often generated using a bandgap voltage, and produce a "floating" differential, low-impedance voltage reference that is often needed by ADCs.

Another example of using feedback is in phase-locked-loop systems, whereby a phase detector, a lead-lag low-pass filter, and a controlled oscillator are configured in a feedback loop to drive the output phase and frequency of the controlled oscillator and to accurately track the phase and frequency of an input reference clock. By adding digital dividers and running the oscillator at a frequency greater than the input reference frequency (which is normally equal to the divided-down frequency), very accurate on-chip clock frequencies are achieved. This is the base of on-chip clock drivers.

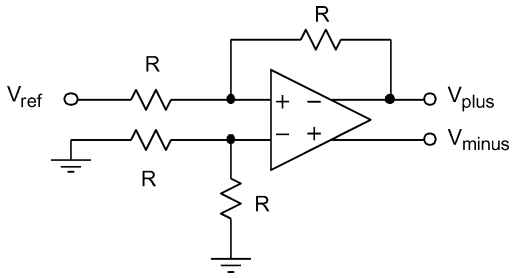

Fig. 19. ***Using feedback with a fully differential op-amp to generate a "floating reference voltage."***

A related example is using a feedback system to stabilize the oscillation frequency of an infinite-Q biquad; a replica of the control signal is then employed to tune other biquads and filters used in the signal path. The control signal automatically adjusts for errors due to both process and temperature variability. This automatic adjustment is based on the assumption that the signal-path biquads are affected in the same way as the infinite-Q biquad in the controlled use.

Many replica bias circuits operate at low frequencies or at dc that allow them to be realized using minimal power. An example of this occurs in realization of high-frequency buffers, such as may be found in high-frequency sample-and-holds. The example shown in Fig. 20 uses complementary n-channel and p-channel source follower buffers in the signal path. Normally, the cascade of these complementary blocks has a dc offset-voltage that occurs due to process and temperature-dependent threshold and mobility differences. The replica bias circuit uses a feedback loop to adjust the bias current of the input n-channel source-follower so the overall offset-voltage is minimized; using the same control voltage to adjust the bias voltage of the matched source-follower in the signal path minimizes the signal-path offset voltage to the level of a good approximation.

The previous examples demonstrate the use of primarily analog techniques used to minimize process-voltage-temperature (PVT) variations; the next section discusses the use of digital calibration to minimize PVT nonidealities of analog circuits.

c) Digital calibration of analog circuits: Digital calibration involves implementing digitally programmable analog components, which can be adjusted using a "digital engine." Digital calibration usually involves adding a

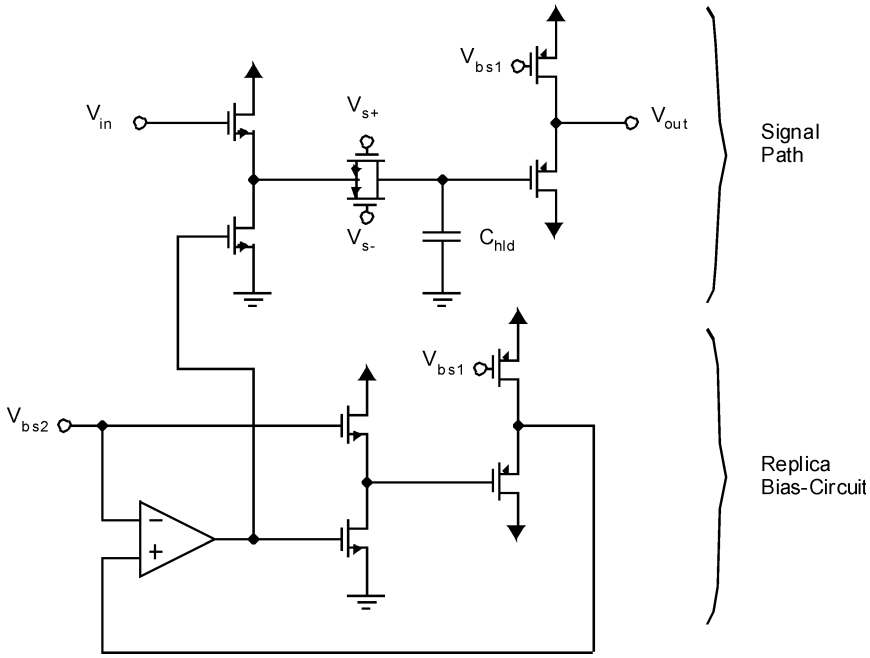

Fig. 20. ***Using a replica-bias circuit to minimize the offset-voltage through two complementary source-followers.***

significant amount of digital circuitry. In nanoscale CMOS, assuming that the digital circuitry is implemented using a digital-synthesis flow, the area of the added digital circuitry is usually insignificant, and it seldom results in a significant increase in total area of the complete analog system (in the chip). In addition, digital calibration is often accompanied by other system changes, for example, the replacement of a large analog loop-filter with a digital loop-filter. These changes can often result in reduced total area.

There are many alternatives for realizing digitally programmable analog components. One choice is to realize programmable capacitors or resistors. Since these are often large and composed of a large number of unit elements, the increase in area required to make them digitally programmable is often minimal. Another popular alternative is to replace a current source with a digitally programmable-current source. A good approach to this replacement [3] is to use perhaps 4 bits of equal thermometer-coded elements for the most significant bits, and then to use binary-coded elements for as many least significant bits as are required to achieve the necessary resolution. In the realization of the digitally programmable analog elements, integral linearity is not usually important; rather, differential linearity and monontonicity are usually of primary importance.

Alternative approaches for digital calibration are calibration at startup or calibration during system operation, either offline or online. Digital calibration at startup is the most popular approach; it is used to minimize processing and especially matching errors. It does not minimize errors due to temperature or power-supply voltages, but these can usually be minimized by other analog circuit techniques. For example, either a feedback or a replica bias loop can be used to minimize the temperature-dependent variations; since this loop does not need the range required to minimize process variations, which are minimized by startup digital calibration, its gain is much smaller. This approach, which helps to minimize noise introduced by the tuning circuits, is very powerful for tuning the frequency of oscillators in phase-locked loops (PLLs).

Calibration during system operation is usually more complicated than calibration at startup, but it has the advantage of minimizing all PVT errors. Calibration can take place either offline or while the system is in use. The use of offline calibration is very application-dependent and only possible when it is known a priori that the system will periodically not be needed for times that are sufficient for calibration to take place. The use of online calibration usually takes two approaches: in the first approach, the calibration takes place using the normal signals that are present during typical operation. A question of concern in this approach is whether the signals span a space large enough so all important parameters can be calibrated. For example, consider using calibration in ADCs. In a flash ADC, where the offset voltage of each comparator is calibrated, the comparators used for signals near the extremes of the input voltage range may never be exercised. However, the normal data signals may be adequate during calibration of the timing skew error between different time-interleaved ADCs [49].

It has often been proposed that special calibration signals be injected into the signal path during the actual operation, either in a way that causes them to be "not

seen" by the system or in a way that results in them being cancelled out after the calibration error detection block, through the use of either a cancelling signal or a filter that does not significantly affect the desired signals (see, for example, [50]). When the former approach is used, the injected signals may often be small and random, and may dither the desired signals with minimal negative effects. Concerns to be addressed in this approach are whether a large enough space is spanned for adequate calibration and whether the introduced noise is really small enough so that system performance does not degrade. In the latter approach, where a calibrating signal is injected and later cancelled after use, there is the important concern about whether the final cancellation is truly adequate. For example, nonlinearities in the system may limit the accuracy of canceling the calibration signal. In some systems, even small residues of noncancelled signals may introduce spectral error tones that are harmful to the system's operation. Although this latter approach has often been considered in the research community, it has not often been used for production ICs.

Two important considerations about any digital calibration scheme are: what is the error or objective function minimized during the adaptive calibration and what is the method used to drive the objective function towards zero? These considerations are, to a large degree, independent or orthogonal. Invariably, in digital calibration approaches, the calibration engine will be implemented completely using digital circuits, and, most often, the engine will be designed using a digital synthesis approach, which helps minimize design time and circuit size. The digital synthesis approach also allows portability and reusability.

The most popular adaptation engine for digital calibration is simply an up–down counter that saturates if extreme values are reached. An alternative approach involves using a linear (but digital) lead-lag filter, assuming that a linear error function is available. A third alternative is to use a successive-approximation approach. All of these approaches have been successfully used. In many adaptation loops, the correlation between an objective error function and the sensitivity of that error function to the parameter being adjusted is driven towards zero. When adaptation speed is crucial, then the adaptation constant can be normalized by the power of the sensitivity filter. This operation requires a division iteration, but it can be implemented iteratively without too much hardware [51]. In all approaches, one must consider the strength of the correlation between the error signal and the parameter being tuned, as compared to uncorrelated noise, which can drive the calibration engine in an arbitrary direction. The larger the noise, the greater the need for filtering in the adaptive loop and, consequently, the loop adaptation speed must be constrained to become slower through the use of a smaller adaptation constant. Another critical consideration is how much latency is inside the loop. As a rough rule of thumb, the time constant of the closed adaptive calibration loop is constrained to be a minimum of ten times greater than the latency around the calibration loop. A large part of the development of any good calibration approach is finding an objective function where the correlation between the objective function and the parameter being adapted is large, the uncorrelated components of the objective function are small, and the signals are not degraded too much by analog offset voltages that occur due to component mismatches.

Once an adequate calibration loop has been designed, then it is worthwhile to minimize the complexity of the solution; a commonly used method is to use just the sign of one or both of these signals instead of correlating the actual error signal with the sensitivity signal [52]. The approaches described have the advantage that a multiplication operation is eliminated; which simplifies the solution and, perhaps more important, makes the digital circuitry faster. The disadvantage of using the signs of the signals rather than the actual signals is that adaptation in the presence of dc offsets may suffer and possibly not even converge, wandering and bias may be larger, and adaptation speed can be considerably slower. Another common simplification for calibration or adaptive loops is to update the parameters being adapted at a lower sample frequency than that used for system signals. This approach is preferable when adaptation speed is not critical and when significant error components are not missed in the process of adapting at a subsampled rate.

d) Regulator and high-voltage approaches: An increasingly popular approach to minimization of errors due to inadequate power supply regulation and/or noisy digital supplies is the use of on-chip regulators. Modern nanoscale technologies almost always have at least two on-chip supply voltages. The most important one is commonly called the core power-supply voltage, which is used for the core digital logic; it is currently often around 1.0 V. In addition, there is invariably a higher power-supply voltage required for the digital interface circuits. Current popular choices for this are 3.3, 2.5, or 1.8 V. Current practice often includes linear regulators, which are designed using the thick oxide transistors and powered by the high supply voltage, in order to generate a clean on-chip power supply voltage. This power supply voltage is used to power critical analog blocks, such as the oscillators in PLLs. Currently, high-frequency on-chip switching regulators (with off-chip inductors) are not often used. An advantage of using a linear on-chip regulator is the increased accuracy of the on-chip voltage being generated (around 2%), which mitigates slow-corner performance degradation. Another advantage of using on-chip regulation is that it allows the use of analog blocks with limited power supply rejection; for example, replacing p-channel active amplifier loads with resistive loads minimizes parasitic capacitances at the outputs of gain stages, and thereby results in faster circuits.

Another practice gaining popularity is to design analog circuits that use core transistors but have power-supply

voltages significantly greater than the core digital supply voltage (such as 1.5 or 1.8 V). This practice can pose a significant reliability issue when not carefully done; normally, the junction-to-substrate voltage can be as large as two times the core supply voltage without breakdown; however, 1) gate-to-source, drain, or any other diffusion voltage differences must be kept within VCC_{OD} limits in steady state and only allowed to exceed those limits for a very short total time on startup to prevent TDDB breakdown; 2) drain-to-source voltage differences must be kept within waveform-dependent VCC_{EOD} limits in the CHE-mode of device operation; and 3) drain-to-source voltages must be restricted to be less than foundry-specified VCC_{MAX} limits in the CHISEL-mode of device operation.

B. ADC Figure of Merit in the Nanoscale Era

In this section, we discuss the evolution of the ADC FOM and how nanoscale CMOS technologies may affect the figure of merit of future ADCs. We briefly discuss a phenomenon we call the "FOM cliff," a technology specific sampling frequency where the FOM worsens rapidly.

If we compare the FOM of different ADCs, we can determine which ADC is "better." But we have to use the right figure of merit. Historically, the Walden figure of merit has been used for ADCs. In 1994, and again in 1999, Walden did a study of ADCs [53], [54]. Data for 150 ADCs was collected, and an empirical FOM was deduced. The figure-of-merit was presented as

$$\mathrm{FOM} = \frac{2^N f_s}{P}. \tag{19}$$

This figure-of-merit has been presented in many forms, but in the recent years it has been common to present it as

$$\mathrm{FOM} = \frac{P}{2^{\mathrm{ENOB}} f_s} \tag{20}$$

where ENOB is the effective number of bits, f_s is the sampling frequency, and P is the power dissipation. With this figure of merit a lower figure is better. The FOM is usually reported in picojoule/conversion step. The state of the art at the time of writing was less than 0.5 pJ/conversion step,[1] but this has surely been reduced since then.

In recent years, the validity of (20) has been questioned [55]. The FOM expression works well when you compare ADCs with the same number of bits, but not when comparing ADCs with different ENOB. It is unfair for high-ENOB ADCs and too lenient on low-ENOB ADCs. The central point of discussion is the 2^{ENOB} factor in (20), this factor should be $2^{2\mathrm{ENOB}}$. The argument is as follows: assume that the power dissipation is dominated by the sampling capacitor, whose value is set by thermal noise power requirements. This would be a thermal noise limited ADC, which is often the case for ADCs with more than 10–12 bits. In most ADCs, the thermal noise power is proportional to kT/C where, in this case, C is equal to the sampling capacitor C_s.

Assume a quantizer step of

$$V_{\mathrm{LSB}} = \frac{V_{\max}}{2^N} = \frac{2V_{pp}}{2^N} = \frac{2\eta_v V_{DD}}{2^N} \tag{21}$$

where $V_{\max}$ is the maximum peak-to-peak differential signal swing. The quantization noise power is therefore given by

$$V_{\mathrm{LSB}}^2 = \frac{4(\eta_v V_{DD})^2}{12 \cdot 2^{2N}} = \frac{(\eta_v V_{DD})^2}{3 \cdot 2^{2N}}. \tag{22}$$

By allowing a thermal noise power equal to 1/4 the quantization noise power, the size of the sampling capacitor can be written as

$$C_s = \gamma \frac{k_B T}{V_{n\_rms}^2} = \frac{12\gamma k_B T}{(\eta_v V_{DD})^2} 2^{2N} \tag{23}$$

where γ is the proportionality constant for the thermal noise. Hence, to increase the resolution by one bit, the size of the sampling capacitor has to be quadrupled.

The power dissipation is usually proportional to the sampling capacitor in a thermal noise limited ADC. As a result, when the size of the sampling capacitor quadruples, the required power dissipation quadruples. If we in (20) keep the FOM and increase the ENOB by one bit, the power dissipation can only double. Hence, a more correct FOM for an ADC that is thermal noise limited is

$$\mathrm{FOM} = \frac{P}{2^{2\mathrm{ENOB}} f_s}. \tag{24}$$

Ideally, all ADCs should be thermal noise limited; this would ensure ADCs with the best energy efficiency. An interesting question is: at what resolution is a converter thermal noise limited? It is difficult to give an exact answer to this question, since it involves knowing the specific architecture of the ADC, but we can give an estimate of the inverse. At what resolution will a converter not be thermal noise limited? Assume that the sampling capacitor is given by (23). Fig. 21 shows a plot of the required size of the sampling capacitor for different power supply voltages. The

[1]ADC data from ISSCC and JSSC: http://www.wulff.no/carsten/doku.php/carsten/electronics/adcfom.

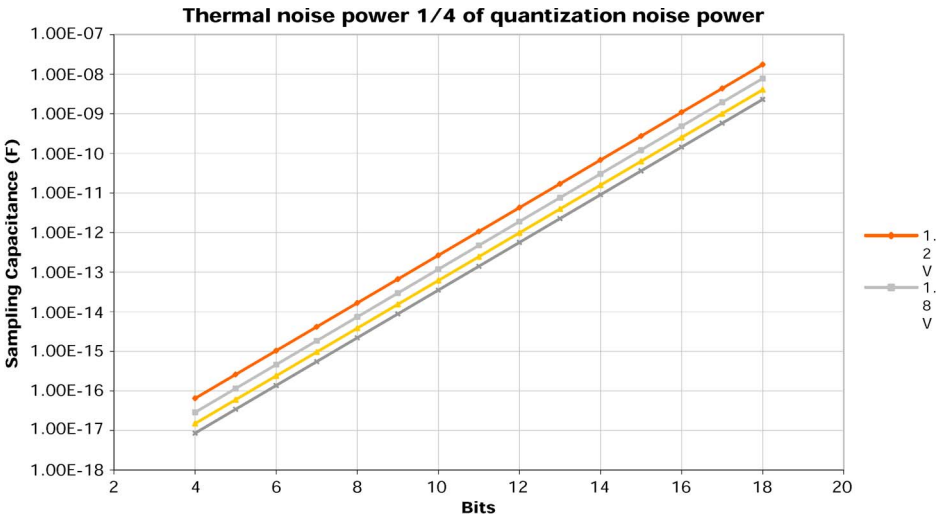

Fig. 21. ***Necessary sampling capacitance as a function of bits.***

number of bits is on the x-axis and the sampling capacitor size is on the y-axis. The other parameters used are $V_{DD} =$ 1.2 V/1.8 V/2.5 V/3.3 V, $\gamma = 1$, $\eta_v = 0.8$, and $T = 353$ K (80 °C). For 1.2 V and 10 bit, a sampling capacitor of 300 fF is needed; for 8 bit, 20 fF; and for 6 bit, 1 fF.

Designing a thermal-noise-limited 8-bit ADC is difficult; the power has to be dominated by the 20 fF sampling capacitor. In current nanoscale CMOS technologies, it is not uncommon to have on the order of 10 fF parasitic capacitance at circuit nodes due to routing. As a consequence, it is unlikely that the sampling capacitor will dominate the power dissipation. A 6-bit thermal-noise-limited ADC at high frequencies ($>$ 100 MHz) will be next to impossible with current processing technology. With a sampling capacitor of 1 fF, and parasitic capacitances on the order of 10 fF, the power dissipation due to the needed sampling capacitor will be an order of magnitude less than the power dissipation due to parasitic capacitances. It remains to be seen if it is at all possible to make a 6 bit, $>$ 100 MHz, thermal-noise-limited ADC in current, and future, CMOS technologies. But as the CMOS technology is scaled down, the parasitic capacitances also scale, which is an advantage for low-resolution converters. Thus it is expected that the FOM for ADCs with less than 8 bit will improve more than the FOM for 12-bit ADCs in the future nanoscale CMOS technologies. In conclusion, we recommend that one use the FOM in (24) when comparing ADCs with different number of bits, but remember that a 0.01 fJ/step 6-bit converter is very hard to design, while a 15-bit converter with the same FOM is commonplace.

1) A Figure of Merit Study: The FOM for lower bit converters is greatly affected by the fact that $<$ 10 bit ADCs are usually not thermal noise limited. Fig. 22 shows the result of a study of the ADC FOM. The ADCs were published in the *Journal of Solid State Circuits;* the latest ADC in the study was from 2005. Power dissipation, sampling frequency, and effective number of bits were collected when the data were available. For the converters where ENOB was not reported, 1 bit was subtracted from the reported resolution. By subtracting one bit, we are confident that a better FOM than the actual FOM is not reported. For very high-resolution converters (above 14 bits), the figure of merit reaches its minimum at about 10^{-17} J/step. Below 14 bits, the converters become less and less dominated by thermal noise, and the figure of merit is worse. For 6-bit ADCs the FOM is almost four orders of magnitude worse than the 14 bit converters, suggesting that there is much to be gained by designing more efficient 6-bit ADCs.

2) The FOM Cliff: Sampling frequency is one of the key measures of any ADC, and usually faster is better. Often we can increase the sampling frequency of a design without degrading the FOM, since power dissipation is proportional to sampling frequency.

The required sampling capacitor does not change with frequency, but the power dissipation does. If the sampling capacitor is much larger than the parasitic capacitances, the power dissipation will be proportional to the frequency. A larger current, however, is needed to drive the same size capacitor at a higher speed; and to keep the gate overdrive constant, the W/L must be increased. A larger transistor area leads to larger parasitic capacitance. Soon the parasitic capacitances are comparable to the sampling capacitor, at which point we waste more and more power for an improvement in speed. For a given CMOS technology and beyond a certain frequency, the FOM will thus drop rapidly. This is what we call the "FOM cliff." If one

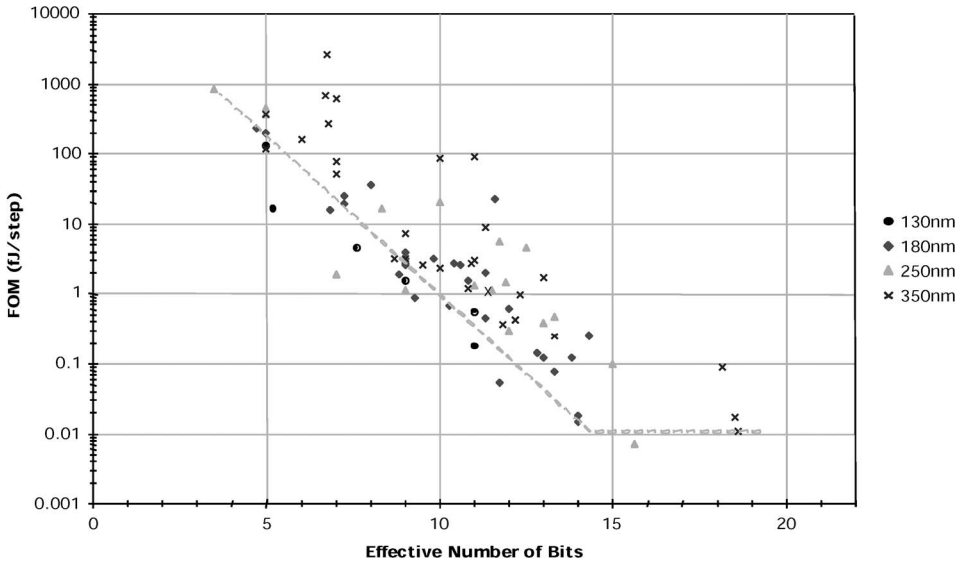

Fig. 22. ***FOM as a function of effective number of bits and technology.***

must operate an ADC beyond the FOM cliff, we advise investigating alternative architectures, such as time-interleaving. Pushing an ADC past the FOM cliff reduces the energy efficiency of the design.

Data from Nyquist ADCs published at ISSCC in the years 2000–2007 were collected. A plot of these data is shown in Fig. 23, with sampling frequency on the *x*-axis and 1/FOM on the *y*-axis for different designs in 90–180 nm CMOS. For 180-nm technology, the FOM cliff is just above 100 MHz. There is small improvement in going to finer technologies, mainly due to reduced parasitic capacitance, but going to a finer technology does not improve the situation much.

Beyond the FOM cliff, other architectures may give a better energy efficiency. One such architecture is time-interleaving, where multiple reduced-rate converters are used in sequence, resulting in a speed increase proportional to the number of parallel converters. Time-interleaved converters have been reported with superior FOM. In [56], eight pipelined ADCs were time-interleaved to make an 11-bit 1-GHz ADC. The parallel pipelined ADCs had a sampling frequency of 125 MHz, which is on the FOM cliff. The ADC is marked in the figure, and it is an order of magnitude better than the others, none of which are time-interleaved ADCs. It should be mentioned that time-interleaved converters have design challenges, like

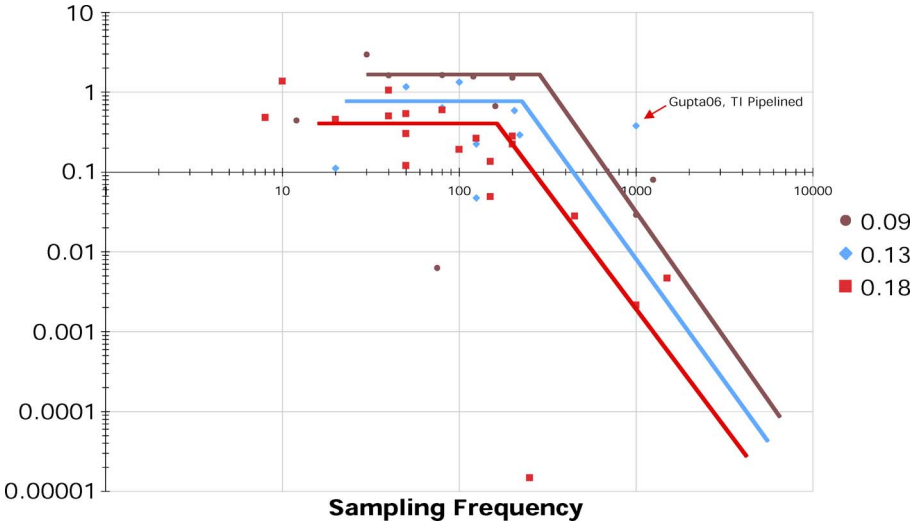

Fig. 23. ***FOM cliff for Nyquist ADCs in ISSCC 2000–2007.***

timing skew, that might exclude their use from certain applications.

IV. CONCLUSION

In the nanoscale era of analog CMOS design, an understanding of the physical factors affecting circuit reliability and performance, as well as the method of mitigating or overcoming them, is becoming increasingly important. The first part of this paper presented factors affecting device matching, including those relating to single devices as well as local and long-distance matching effects. Several reliability effects are discussed, including physical design limitations projected for future down-scaling. The condition $V_{sb} > 0$ is required for cascode circuit configurations. The role of other terminal voltages is discussed as $V_{sb} > 0$ increases both hot and cold carrier damage effects in highly scaled devices. In some cases, it may be helpful to exceed foundry specified drain-source voltage limits by a few hundred millivolts. Models were presented for achieving this, which include the dependence on the shape of the output waveform.

The second part of this paper focused on trends in device characteristics and how they influence the design of nanoscale analog CMOS circuits. A number of circuit design techniques employed to address the major nonidealities of nanoscale CMOS technologies were presented. Achieving high energy efficiency in ICs capable of accommodating 10^8 devices is becoming critically important. This paper also presented a survey of the evolution of FOM for ADCs.

As nanoscale analog CMOS design pushes forward into future technology nodes, there is no doubt that more physical limits and process variables will come into view. Many past fabrication and circuit design "limits" have been pushed out through hard work and innovation. While even more complex device models, design techniques, and methods for mitigating the new limits will be required in emerging areas such as the 50 GHz to subterahertz regime, analog CMOS design should remain alive and well deep into the nanoscale era of CMOS technology. ■

REFERENCES

[1] B. Dennington, "Low power design from technology challenge to great products," in *Proc. IEEE ISLPED '06*, Oct. 2006, p. 213.

[2] L. Lewyn and J. Meindl, "Physical limits of VLSI dynamic random-access memories," *IEEE J. Solid-State Circuits*, vol. SSC-20, pp. 231–241, Feb. 1985.

[3] D. Johns and K. Martin, *Analog Integrated Circuit Design*. New York: Wiley, 1997.

[4] Y.-M. Sheu *et al.*, "Modeling the well-edge proximity effect in highly scaled MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, pp. 2792–2798, Nov. 2006.

[5] C.-Y. Chan, Y.-S. Lin, Y.-C. Huang, S. Hsu, and Y.-Z. Juang, "Impact of STI effect on flicker noise in 0.13-μm RF nmOSFETs," *IEEE Trans. Electron Devices*, vol. 54, pp. 3383–3392, Dec. 2007.

[6] W. Abadeer, "Reliability monitoring and screening issues with ultrathin gate dielectric devices," *IEEE Trans. Device Mater. Rel.*, vol. 1, pp. 60–68, Mar. 2001.

[7] S. Saxena, C. Hess, H. Karbasi *et al.*, "Variation in transistor performance and leakage in nanometer-scale technologies," *IEEE Trans. Electron Devices*, vol. 55, pp. 131–144, Jan. 2008.

[8] R. Klein, "Overview of process variability," in *Proc. ISSCC 2008 Microprocessor Forum F6: Transistor Variability Nanometer-Scale Technol.*, Feb. 7, 2008, pp. A1–A24.

[9] L. Liebmann, "Computational lithography and its impact on design," in *Proc. ISSCC 2008 Microprocessor Forum F6: Transistor Variability Nanometer-Scale Technol.*, Feb. 7, 2008, pp. B1–B52.

[10] C. R. Grace, P. J. Hurst, and S. H. Lewis, "A 12 b 80 MS/s pipelined ADC with bootstrapped digital calibration," in *ISSCC Dig. Tech. Papers*, Feb. 2004, pp. 460–539.

[11] A. Nazemi, C. Grace, L. Lewyn *et al.*, "A 10.3 GS/s 6 bit (5.1 ENOB at Nyquist) time-interleaved/pipelined ADC using open-loop amplifiers and digital calibration in 90 nm CMOS," in *VLSI Symp. Dig. Tech. Papers*, Jun. 2008, pp. 18–19.

[12] S. Ray and B.-S. Song, "A 13b Linear 40 MS/s pipelined ADC with self-configured capacitor matching," in *ISSCC Dig. Tech. Papers*, Feb. 2006, pp. 852–861.

[13] S.-T. Ryu, S. Ray, B.-S. Song *et al.*, "A 14b-linear capacitor self-trimming pipelined ADC," in *ISSCC Dig. Tech. Papers*, Feb. 2004, pp. 460–539.

[14] L. Lewyn and M. Loose, "A 1.5 mW 16b ADC with improved segmentation and centroiding algorithms and litho-friendly physical design (LFD) used in space telescope imaging applications," in *Proc. ACE IEEE CICC'09*, in press.

[15] M. Campbell, V. Gerousis, J. Hogan, J. Kibarian, L. Lanza, W. Ng, D. Pramanik, A. Strojwas, and M. Templeton, "When IC yield missed the target, who is at fault?" in *Proc. ACE IEEE DAC'04*, Jun. 2004, p. 80, panel session.

[16] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SSC-9, pp. 256–268, Oct. 1974.

[17] J. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Trans. Device Mater. Rel.*, vol. 1, pp. 43–58, Mar. 2001.

[18] E. Harrari, "Dielectric breakdown in electrically stressed thin films of thermal SiO2," *J. Appl. Phys.*, vol. 49, pp. 2478–2489, Apr. 1978.

[19] A. Yassine, H. Nariman, M. McBride, M. Uzer, and K. Olasupo, "Time dependent breakdown of ultrathin gate oxide," *IEEE Trans. Electron Devices*, vol. 47, pp. 1416–1420, Jul. 2000.

[20] Y. Chen, J. Suehle, C. Shen, J. Bernstein, C. Messick, and P. Chaparala, "A new technique for determining long-term TDDB acceleration parameters of thin gate oxides," *IEEE Electron Device Lett.*, vol. 19, pp. 219–221, Jul. 1998.

[21] H. Stork, "It's all about scale," *IEEE SSCS Newsletter*, Jan. 2007.

[22] C. Hu, "Lucky electron model of channel hot electron emission," in *IEDM Tech. Dig.*, Dec. 1979, p. 22.

[23] E. Sangiorgi, B. Ricco, and P. Olivio, "Hot electrons and holes in MOSFET's biased below the Si-SiO2 interfacial barrier," *IEEE Electron Device Lett.*, vol. 6, pp. 513–515, Oct. 1985.

[24] A. Abidi, "On the operation of cascode gain stages," *IEEE J. Solid-State Circuits*, vol. 23, pp. 1434–1437, Dec. 1988.

[25] Y. Nakagome, E. Takeda, H. Kume, and S. Asai, "New observation of hot-carrier injection mechanism," *Jpn. J. Appl. Phys.*, vol. 19, p. 85, 1983, suppl. 19–1.

[26] F. Driussi, D. Esseni, L. Selmi, and F. Piazza, "Observation of a new hole gate current component in p+ -poly gate p-channel MOSFETs," in *Proc. ESSDERC Conf. 2000*, 2000, pp. 136–139.

[27] S. Mahapatra, C. D. Parikh, V. R. Rao, C. R. Viswanathan, and J. Vasi, "A comprehensive study of hot-carrier induced interface and oxide trap distributions in MOSFETs using a novel charge pumping technique," *IEEE Trans. Electron Devices*, vol. 47, pp. 171–177, Jan. 2000.

[28] F. Driussi, D. Esseni, L. Selmi, and F. Piazza, "Substrate enhanced degradation of CMOS devices," in *IEDM Tech. Dig.*, 2000, pp. 105–108.

[29] F. Driussi, D. Esseni, L. Selmi, and F. Piazza, "Damage generation and location in n- and p-MOSFETs biased in the substrate-enhanced gate current regime," *IEEE Trans. Electron Devices*, vol. 49, May 2002.

[30] F. Driussi, D. Esseni, and L. Selmi, "On the electrical monitor for device degradation in the CHISEL stress regime," *IEEE Electron Device Lett.*, vol. 24, pp. 357–359, May 2003.

[31] I. C. Chen, S. E. Holland, and C. Hu, "Electron trap generation by recombination of electrons and holes in SiO2," *J. Appl. Phys.*, vol. 61, no. 9, p. 4544, 1987.

[32] N. Jha and V. Rao, "A new oxide trap-assisted NBTI degradation model," *IEEE Electron Device Lett.*, vol. 26, pp. 687–689, Sep. 2005.

[33] S. Mahapatra, D. Saha, D. Varghese, and P. B. Kumar, "On the generation and recovery of interface traps in MOSFETs subjected to NBTI, FN, and HCI stress," *IEEE Trans. Electron Devices*, vol. 53, pp. 1584–1592, Jul. 2006.

[34] H. Kufluoglu and M. A. Alam, "Theory of interface-trap-induced NBTI degradation for reduced cross section MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, pp. 1120–1130, May 2006.

[35] J.-C. Huang and W.-T. Chien, "Some practical concerns on isothermal eletromigration tests," *IEEE Trans. Semicond. Manuf.*, vol. 14, pp. 387–394, Nov. 2001.

[36] I. A. Blech and C. Herring, "Stress generation by electromigration," *Appl. Phys. Lett.*, vol. 29, p. 131, 1976.

[37] C. Hau-Riege, A. Marathe, and V. Pham, "The effect of low-K ILD on the electromigration reliability of Cu interconnects with different line lengths," in *Proc. IEEE 41st Annu. Int. Rel. Phys. Symp. 2003*, Mar. 2003, pp. 173–177.

[38] A. Naeemi, R. Savari, and J. D. Meindl, "Performance comparison between carbon nanotube and copper interconnects for gigascale integration (GSI)," *IEEE Electron Device Lett.*, vol. 26, pp. 84–86, Feb. 2005.

[39] A. Naeemi and J. D. Meindl, "Design and performance modeling for single-walled carbon nanotubes as local, semiglobal, and global interconnects in gigascale integrated systems," *IEEE Trans. Electron Devices*, vol. 54, pp. 26–37, Jan. 2007.

[40] A. Naeemi and J. D. Meindl, "Electron transport modeling for junctions of zigzag and armchair graphene nanoribbons (GNRs)," *IEEE Electron Device Lett.*, vol. 29, pp. 497–499, May 2008.

[41] A. J. Annema, B. Nauta, R. van Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE J. Solid-State Circuits*, vol. 40, pp. 132–143, Jan. 2005.

[42] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. New York: McGraw-Hill, 1999, p. 501.

[43] M. Garg, S. S. Suryagandh, and J. C. S. Woo, "Scaling impact on analog performance of sub-100 nm MOSFETs for mixed mode applications," in *Proc. 33rd Eur. Solid-State Device Res. Conf. (ESSDERC '03)*, 2003, pp. 371–374.

[44] H. Sasaki, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "1.5 nm direct-tunneling gate oxide Si MOSFET's," *IEEE Trans. Electron Devices*, vol. 43, pp. 1233–1242, Aug. 1996.

[45] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in n-MOSFET's with sub-2 nm gate oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 1636–1644, Aug. 2000.

[46] C. Chang-Hoon, N. Ki-Young, Y. Zhiping, and R. W. Dutton, "Impact of gate direct tunneling current on circuit performance: A simulation study," *IEEE Trans. Electron Devices*, vol. 48, pp. 2823–2829, Dec. 2001.

[47] A. Veloso, M. Jurczak, F. Cubaynes, R. Rooyackers, S. Mertens, A. Rothschild, M. Schaekers, A. Al-Shareef, R. Murto, C. Dachs, and G. Badenes, "RPN oxynitride gate dielectrics for 90 nm low power CMOS applications," in *Proc. 32nd Eur. Solid-State Device Res. Conf.*, 2002, pp. 159–162.

[48] B. Razavi, *Design of Analog CMOS Integrated Circuits*. New York: Mc-Graw-Hill, 2001.

[49] A. Haftbaradaran and K. Martin, "Mismatch compensation techniques using random data for time-interleaved A/D converters," in *Proc. IEEE Int. Conf. Circuits Syst. (ISCAS 2006)*, May 2006, pp. 3402–3405.

[50] E. Siragusa and I. Galton, "A digitally enhanced 1.8 V 15b 40 MS/s CMOS pipelined ADC," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2004.

[51] K. Martin, "Power-normalized update-algorithm for adaptive filters—Without divisions," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1782–1786, Nov. 1989.

[52] A. Shoval, D. Johns, and M. Snelgrove, "Comparison of DC offset effects in four adaptive LMS algorithms," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 176–185, Mar. 1995.

[53] R. H. Walden, "Analog-to-digital converter technology comparison," in *IEEE GaAs IC Symp. Tech. Dig.*, Oct. 1994, pp. 217–219.

[54] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, Apr. 1999.

[55] H.-S. Lee and C. G. Sodini, "Analog-to-digital converters: Digitizing the analog world," *Proc. IEEE*, vol. 96, pp. 323–334, Feb. 2008.

[56] S. Gupta, "A 1 GS/s 11 b time-interleaved ADC in 0.13μm CMOS," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, 2006.

ABOUT THE AUTHORS

Lanny L. Lewyn (Life Senior Member, IEEE) received the B.S. Eng. degree (with honors) and the M.S.E.E. degree from the California Institute of Technology, Pasadena, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1984.

His work at Stanford on physical limits of VLSI circuits resulted in publication of the first closed-form solution for the MOS device surface potential that was continuous from weak to strong inversion. Another result was the first solution of the dRAM alpha particle soft error problem using the combination of Hamming error-correcting codes, particle range data, and separation of bits on nonadjacent bit lines. His high-resolution ADC and DAC work relies on litho-friendly design without autocalibration. This work includes an 18-b CMOS DAC IC that was licensed to Toshiba for high-volume production in early 4× OS audio disc players, a 14-b CMOS ADC IC that was the first converter used for DSL in the United States and Europe by industry-pioneers PairGain and Alcatel, and a 1.2 mW 16-b CMOS ADC ×36-array in an image-processing ASIC (SIDECAR) that replaced the main survey camera (ACS) image processor in the May 2009 Hubble space telescope servicing mission (SM-4). That ADC design will be used again in the main IR camera (NIRCAM) of the next-generation space telescope (JWST), scheduled to fly in 2014. His current circuit design work includes overcoming headroom problems resulting from device voltage limitations in nanometer analog CMOS circuits. That work includes high speed (1–10 GHz), high resolution (10–14b) CMOS pipeline ADCs, comparators, amplifiers, low-jitter clock distribution, and the next-generation low-power $>$ 10 GBS SERDES line driver for the SnowBush IP Division of Gennum in 28 nm CMOS technology. His physical design work is focused on overcoming stress, lithography, and reliability limitations in deep submicrometer processes. This work also includes the development and refinement of dimensionless schematic capture and layout techniques to port designs from micrometer technology nodes down through 28 nm. His past positions include Manager, Nuclear Space Instrument Development, NASA-JPL; Director of R&D, Pacemaker Division, American Hospital Supply; Manager, Advanced CMOS Circuit Development, Hughes Solid State Products Division; and Manager, Analog IC Design, Pargain Division, Globespan Inc. He is currently President of Lewyn Consulting Inc. (LCI), Laguna Beach, CA; is a Course Instructor for the Mead Education Group on the subjects of reliability and physical design issues in nanoscale CMOS technologies; and serves on the Technical Advisory Board of the Snowbush IP Division, Gennum Physical Design Center, Aguascalientes, Mexico. He has received 29 U.S. patents in CMOS and bipolar circuits.

Dr. Lewyn was invited to present the keynote paper at IEEE NORCHIP 2009.

Trond Ytterdal (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the Norwegian Institute of Technology in 1990 and 1995, respectively.

He was a Research Associate with the Department of Electrical Engineering, University of Virginia (1995–1996), and as a Research scientist with the Electrical, Computer and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY (1996–1997). From 1997 to 2001, he was a Senior ASIC Designer with Nordic Semiconductor, Trondheim, Norway. In 2001, he became a Professor at the Department of Electronics and Telecommunication, Norwegian University of Science and Technology (NTNU). His present research interests include design of analog integrated circuits, behavioral modeling and simulation of mixed-signal systems, modeling of nanoscale transistors, and novel device structures for application in circuit simulators. He has published more than 130 scientific papers in international journals and conference proceedings. He is a coauthor of *Semiconductor Device Modeling for VLSI* (Englewood Cliffs, NJ: Prentice-Hall, 1993), *Introduction to Device Modeling and Circuit Simulation* (New York: Wiley, 1998), and *Device Modeling for Analog and RF CMOS Circuit Design* (New York: Wiley, 2003) and has contributed to several other books published internationally. He is also a Codeveloper of the circuit simulator AIM-Spice.

Prof. Ytterdal is a member of The Norwegian Academy of Technological Sciences.

Carsten Wulff (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the Department of Electronics and Telecommunication, Norwegian University of Science and Technology (NTNU), in 2002 and 2008, respectively.

During his Ph.D. work at NTNU, he worked on open-loop sigma–delta modulators and analog-to-digital converters in nanoscale CMOS technologies. In 2006–2007, he was a Visiting Researcher with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. He is currently a Research and Development Engineer with Nordic Semiconductor, Trondheim, Norway. His present research interests include analog and mixed-signal CMOS design and design of high-efficiency analog-to-digital converters.

Kenneth Martin (Fellow, IEEE) was a Professor at the University of California, Los Angeles (1980–1991). He received tenure (Associate Professor) in two years and became a full Professor in 1987. In 1985, he founded the Integrated Circuits and Systems Laboratory (ICSL) and major field at UCLA, which was the incubator of many high-tech companies in southern California, including Broadcom. He became an Endowed Professor at the University of Toronto, Toronto, ON, Canada, in 1991 and an Adjunct Professor in 2008. While there, he coauthored (with D. Johns) *Analog Integrated Circuit Design* (New York: Wiley, 1997). He has also coauthored numerous other books and has authored or coauthored well over 100 papers. He is the inventor on numerous patents. In 1998, he cofounded Snowbush Microelectronics, which had grown to 50 employees (including a Mexican Design Center) by 1997, when it was acquired by Gennum Corp. He was Chief Technical Officer with Gennum Corp. for one year after acquisition until leaving to found Granite SemiCom Inc. in early 2009.

Prof. Martin received the Outstanding Young Engineer Award from the IEEE Circuits and Systems Society in 1984. He received the NSF Presidential Young Investigator's Award from 1985 to 1990. He was a corecipient of the Beatrice Winner Award at the 1993 ISSCC and a corecipient of the 1999 IEEE Darlington Best-Paper Award from the IEEE Transactions on Circuits and Systems. He received the 1999 CAS Golden Jubilee Medal from the IEEE Circuits and Systems Society.