Analog/Mixed-Signal Design Challenges in 7-nm CMOS and Beyond

Alvin L. S. Loke, Da Yang, Tin Tin Wee, Jonathan L. Holland*, Patrick Isakanian, Kern Rim, Sam Yang, Jacob S. Schneider, Giri Nallapati, Sreeker Dundigal, Hasnain Lakdawala, Behnam Amelifard, Chulkyu Lee, Betty McGovern, Paul S. Holdaway**, Xiaohua Kong, and Burton M. Leary

Qualcomm Technologies Incorporated, 5775 Morehouse Drive, San Diego, CA 92121 USA
*Qualcomm Technologies Incorporated, 8041 Arco Corporate Drive, Raleigh, NC 27617 USA
**Qualcomm Incorporated, 5775 Morehouse Drive, San Diego, CA 92121 USA
aloke@qti.qualcomm.com

***Abstract*—The economics of CMOS scaling remain lucrative with 7-nm mobile SoCs expected to be commercialized in 2018. Driven by careful design/technology co-optimization, modest reduction in fin, gate, and interconnect pitch as well as process innovations continue to offer compelling node-to-node power, performance, area, and cost benefits to advance logic and SRAM to the next foundry node. However, analog/mixed-signal circuits do not fully realize these improvements. They become more cumbersome to design, having worse parasitic resistance and capacitance, stronger layout-dependent effects, and layout growth in some situations. Furthermore, early adopters of these cutting-edge finFET nodes must cope with the complications of design concurrent with technology development for shorter product time-to-market. We provide an overview of the key process technology elements enabling 7 nm and beyond to address analog/mixed-signal design challenges. From this insight, we offer layout guidelines aimed to reduce design vulnerability to technology and model immaturity.**

I. Introduction

The mobile system-on-chip (SoC) remains the main driver for CMOS scaling with the necessary market volume to justify the enormous investments in process and design development. Smartphones with 7-nm SoCs are expected to debut in 2018. Foundries and leading design houses are preparing for continued demand at 5 nm. Although physical feature scaling has slowed down, meticulous design/technology co-optimization (DTCO) continues to squeeze enough overall power, performance, area, and cost (PPAC) improvement to justify another node. However, analog/mixed-signal (AMS) designs such as PLLs, wireline I/O, data converters, regulators, and bandgap references do not fully reap these benefits. SoC technologies are tailored for logic and SRAM as they dictate die area and cost. Consequently, the AMS device palette, comprising logic FETs plus long-channel and I/O varieties, passives, BJTs, and ESD devices, cannot be optimized independently from logic and SRAM. Instead, it is largely derived from existing process capability and inevitably compromised.

This paper presents the challenges faced when AMS designs migrate into 7 nm and beyond. We highlight technology innovations that have enabled digital scaling to explain their impact on AMS design. Early adopters of new technology nodes face the additional burden of designing alongside technology development to reduce product time-to-market [1]. Here, speculative target-based device models are prone to adjustments throughout the design cycle and even after tape-out. With some understanding of the technology, we can implement layout practices that make AMS designs more resilient to model retargeting.

II. Technology Scaling Enablers

A. Fully Depleted Bulk FinFET

The bulk finFET, shown in Fig. 1, was introduced into manufacturing at the 22-nm node [2]. Shortly thereafter, foundries began offering finFETs at 16/14 nm [3], 10 nm [4], and 7 nm [5]. The area-efficient finFET device architecture offers superior short-channel control and is particularly suited for low-power CMOS. Its fully depleted operation and 3-D structure attenuate the coupling of the channel surface potential to the body and drain. The resulting stronger gate control of the FET on-off transition reduces the subthreshold swing *SS* (ΔV_{GS} per decade change in subthreshold current) and drain-induced barrier lowering *DIBL* (V_T reduction per ΔV_{DS}). Drive and leakage currents comparable to a planar structure can be achieved at a much lower V_T and supply voltage (V_{DD}) to reduce dynamic power (Fig. 2). For example, the 22-nm finFET in [2] demonstrates *SS* of 69–72 mV/decade and *DIBL* of 46–50 mV/V at 25 °C, a marked improvement from numbers as high as 100 mV/decade and 200 mV/V from a competitive 32-nm planar technology [6]. Equally important, less DIBL also improves the effective drive current at a given I_{Dsat} for faster CMOS switching [7] as well as the saturation r_{out} for better analog intrinsic gain [8].

With dramatic *SS* and *DIBL* reduction enabled by the finFET, subsequent node-to-node improvements have been modest. Narrower and taller fin profile as well as junction optimization have merely trimmed *SS* and *DIBL* to 65 mV/decade and 35 mV/V respectively at 7 nm [5]. Nonetheless, better fin process control and the use of gate-stack instead of implant-based V_T tuning have been instrumental to reduce device variation, enabling even lower V_{DD} operation. Moving to 5 nm, the finFET is expected to continue to evolve, perhaps incorporating higher channel mobility materials such as Ge and III-V compounds [9]. The manufacturability of gate-all-around and more exotic sub-60-mV/decade device architectures is not yet proven.

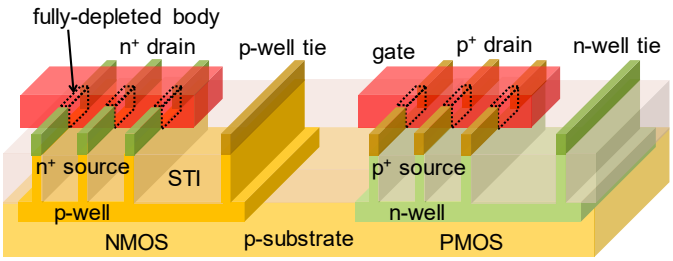

Fig. 1. Fully depleted bulk CMOS finFETs.

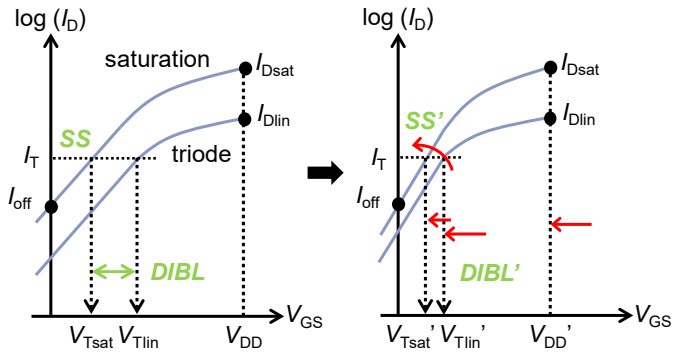

Fig. 2. Enabling lower V_{DD} with reduced *SS* and *DIBL*.

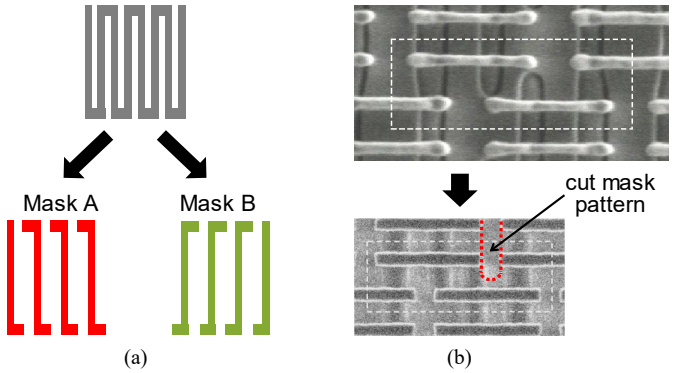

Fig. 4. (a) Pitch splitting and (b) use of cut mask [11].

B. *Lithography and Self-Aligned Patterning*

Critical dimensions (CDs) and pitch continue to shrink with time but at a pace falling short of the historical $1/\sqrt{2}$ node-to-node scaling factor [10]. See Fig. 3. In fact, as early as 28 nm, the technology node name has become a marketing tag for overall PPAC, no longer defined by the minimum gate length (L_{min}). The 193-nm immersion (193i) scanner has remained the industry workhorse since 32 nm. Although 13.5-nm extreme ultraviolet (EUV) is finally available starting at 7 nm, its usage is limited to a few metal layers due to high tool cost and limited source power. Meanwhile, innovations have emerged to overcome the 193i single-exposure pitch limit of ~80 nm for further scaling at the expense of added process complexity and cost.

Pitch splitting or litho-etch-litho-etch (LELE) debuted at 20 nm for printing contacts and the lowest metal levels (Fig. 4(a)) [11]. A pattern of alternate lines (Mask A) is first transferred to a thin hard mask and the remaining unprinted alternate lines are exposed with a second resist pattern (Mask B). A common etch subsequently transfers the combined hard mask and resist patterns to the underlying layer. Pitch splitting adds the complexity of layout coloring or mask decomposition as well as new rules to cope with misalignment tolerance and pattern density balance between the two masks. This technique can be extended to include a third mask (LELELE) to further reduce metal pitch to 48 nm [10].

Cut masks are used to shorten line end-to-end spacing (Fig. 4(b)). Cut patterns slice orthogonally through a pattern of continuous lines devoid of line end corner rounding and pullback artifacts. As the cut ends fall outside the line patterns, their artifacts are not transferred to the final pattern. The cuts can be incorporated as an extra exposure or a hard mask pattern depending on whether the line pattern is positive tone (e.g., gate) or negative tone (e.g., metal trench). Introduced at 45-nm gate patterning for denser SRAM cells [12], cut masks are ubiquitous in fin, gate, and interconnect patterning. The tight cut pitch in 10 nm and below has even necessitated pitch-splitting the cuts.

Overlay control is equally critical as feature size for area scaling and has spawned key self-alignment innovations that are, in principle, insensitive to mask-to-mask misalignment.

Spacer-based patterning [13], also known as self-aligned double patterning (SADP), is routine for fin construction [2]. Shown in Fig. 5, spacers of a uniform width are formed on the sidewalls of a sacrificial mandrel grating that is subsequently removed, leaving behind a sea of spacers at half the mandrel pitch to pattern the underlying material. The pitch of the spacer-based pattern is limited to half of the 193i limit, but the technique can be repeated recursively. In self-aligned quadrature patterning (SAQP), the spacers derived from the original mandrel (Spacer 1) become the mandrels for forming a second set of spacers (Spacer 2) at one quarter the original mandrel pitch. SAQP has enabled fin pitch to scale well below 40 nm [14]. Spacer-based patterning also offers less line width variation than pitch splitting. Unlike a conventional resist-patterned line, the edge roughness along the opposite sides of a spacer-patterned line is correlated because the spacer dielectric deposition process is conformal. This benefit has initiated the use of SADP for gate patterning at 10 nm to reduce device variation at a contacted gate pitch (CGP) of 64 nm [4]. SADP can also be augmented to offer a "wimpy" device with a slightly longer gate commonly used in processor design. Here, spacers are formed with an initial width of $L_{min}+\Delta L_{wimpy}$. With an extra mask, some

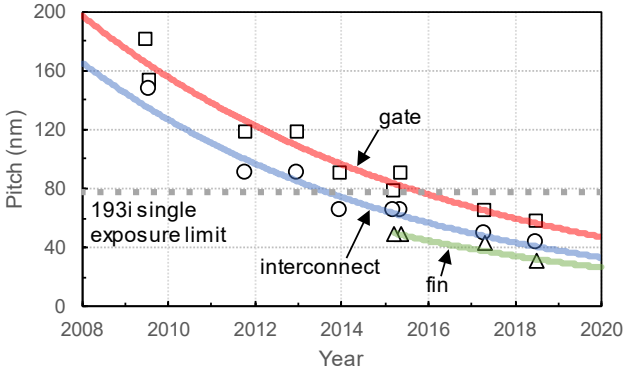

Fig. 3. Foundry scaling of fin, gate, and interconnect pitch [10].

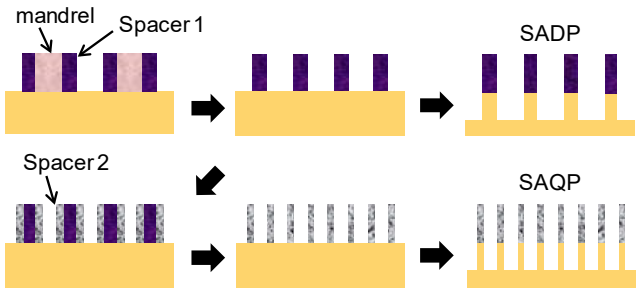

Fig. 5. Spacer-based patterning [13].

spacers are trimmed to a width of L_{min} before both L_{min} and $L_{min}+\Delta L_{wimpy}$ spacers are used to pattern the poly-Si beneath. Yet another mask is required to print much longer devices for analog design. SADP and SAQP have additionally been applied to middle- and back-end-of-line (MEOL and BEOL) metal patterning [14]. Unlike gate patterning, it is the dielectric of a metal layer that is patterned due to the damascene nature of contact and metal/via formation. Here, SADP and SAQP must incorporate a trim mask to add resist patterns that bridge contiguous spacers prior to etching to make wider spaces possible [15]. The mandrel spacing can also be adjusted to offer some flexibility in metal width.

BEOL scaling has also been enabled by the self-aligned via (SAV) [16]. Dual-damascene SAVs are formed by first storing the overlying metal trench pattern in a hard mask. Rectangular via openings that orthogonally cut across the trench width are then patterned. Selective against the trench hard mask, the via etch is bounded by the hard mask opening and can only proceed at the desired intersection of the via and trench patterns.

Introduced at 22 nm [2], self-aligned MEOL contacts have become necessary at 10 nm to support aggressive CGPs. The tight separation between source/drain contacts and the gate is prone to overlay-related contact-to-gate shorts. As such, self-aligned source/drain contacts (SACs) are used. They are formed by capping the gate with an insulator that is compositionally different from the contact dielectric to protect the gate against a misaligned diffusion contact etch [2]. Self-aligned gate contacts (SAGCs) have also been implemented [14]. Denser standard cells can be achieved by eliminating the need to land contacts on the gate outside the active area. SAGCs require the source/drain contacts to be capped with an insulator that is different from both contact and gate cap dielectrics to protect the source/drain contacts against a misaligned gate contact etch. See Fig. 6. Foundries are yet to adopt SAGCs at the 7-nm node.

Scaling will continue to drive down feature pitch and eventually be limited by corner rounding and overlay (components of edge placement error) of the self-aligned cut and contact/via patterns [10]. Migrating to EUV mitigates these issues and resumes the simplicity of single-exposure patterning. However, as features continue to scale, pitch splitting and spacer-based patterning will eventually be applied to EUV.

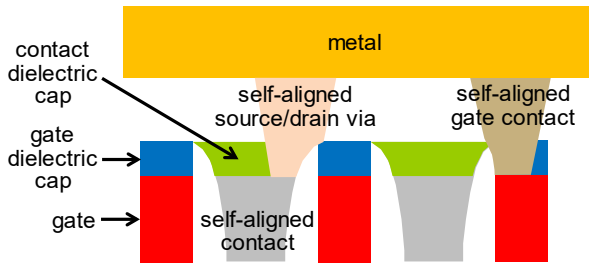

Fig. 6. Gate and contact dielectric caps for self-aligned MEOL contacts [10].

C. High-K Gate Dielectric and Metal Gate (HKMG)

Foundries have adopted HKMG since 28 nm to improve short-channel control and reduce tunneling leakage and gate charge depletion that plagued the poly-Si/SiON stack [17]. See Fig. 7. Increased C_{ox} comes from a high permittivity (HK) HfO_2 dielectric covered by a thin metal gate (MG) layer. The remaining gate volume is filled with a lower resistivity metal to reduce gate resistance. The interfacial oxide, HK, and MG layers are deposited using atomic layer deposition (ALD) for precise thickness and stoichiometric control. V_T is tuned by uniquely adjusting the flatband voltage (V_{FB}) of the HKMG stack for each NMOS and PMOS V_T flavor offered. In 10 nm and beyond, the fin volume surrounded by the gate is so small that V_T adjustment by a few stochastically implanted dopants increases device variation. Consequently, foundries employ schemes that modulate the MG work function (ϕ_M) with different types and thicknesses of materials [4] as well as possibly embed dipoles (fixed charge) in the HK dielectric [18]. A fully depleted device requires less ΔV_G to transition from accumulation to inversion. For this reason, quarter-gap instead of band-edge gate work functions are chosen to achieve the V_T values required for optimum drive and leakage [19].

The MG/HK interface is delicate and prone to thermally aggravated ϕ_M instability. Hence, the gate is formed with a damascene replacement metal gate (RMG) integration that introduces the HKMG layers after the source/drain anneal for better V_T control [11]. In RMG, the patterned poly-Si gate from conventional front-end processing is sacrificial, removed completely after the contact dielectric is polished to expose the gate surface. The resulting gate trench is filled with the HKMG stack, and excess deposition above the trench is polished away. The gate is then recessed and covered with the dielectric cap to withstand the SAC etch. RMG requires postponing the silicide module until after the SAC etch as silicide cannot tolerate the HK post-deposition anneal [9]. Unfortunately, this resequencing constrains silicide to form only at the bottom of the contact opening. To reduce contact resistance to the fins, trench contacts lining the entire device width have become standard.

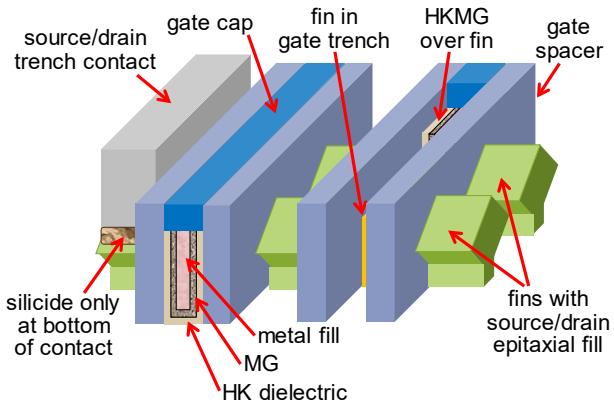

Fig. 7. Cut view of finFET with HKMG.

D. Process-Induced Mechanical Strain

Mechanical stressors have been employed to boost electron and hole mobilities since 90 nm. Because silicon is piezoresistive, as little as 1% lattice strain can increase mobility by several times [20]. Channel stress is introduced by straining surrounding regions with techniques such as source/drain fin epitaxy and gate stress. The desired longitudinal channel strain is

tensile in NMOS and compressive in PMOS. Mobility enhancement remains far more effective for PMOS, and this has resulted in PMOS short-channel drive strength matching and even exceeding that of NMOS [4], [5]. Mobility boost is, however, much less effective in longer channels where the larger fin volume between source and drain succumbs less to the intentional surrounding pressure. Source/drain fin recess and subsequent epitaxial growth with *in situ* doping continue to be critical modules and active opportunities for device improvement.

E. Middle-End-Of-Line (MEOL)

Starting at 20 nm, tighter CGP has resulted in an increasingly complex and far costlier MEOL for Metal-1 to contact the underlying high density of transistors [21]. Previously a single-mask module in 28 nm, the MEOL even at 10 nm requires well over a dozen masks. The finFETs are contacted with independently formed SACs and gate contacts. Furthermore, additional levels of local vias (Via-0, gate via, and source/drain via) and metal (Metal-0) are required to support dense local routing. Each MEOL level requires a very aggressive pitch, e.g., 40 nm for 7-nm Metal-0 [5], which necessitates multiple patterning; four masks per 7-nm MEOL level is not uncommon.

MEOL specification is driven by extensive DTCO study and carefully optimized to balance process complexity (yield risk) and logic/SRAM area reduction. For example, standard cell area can be reduced by special process constructs such as diffusion-to-diffusion jumpers, cross-coupling connections, and single-diffusion breaks (SDBs) [10]. SDBs (Fig. 8) can potentially save 10% logic area by eliminating dummy gate waste. In finFET fabrication, the end of an active area must terminate at a dummy gate spacer for better control of source/drain fin epitaxy. If a narrow shallow trench isolation (STI) cannot be inserted under a single dummy gate, abutted devices cannot terminate on a shared dummy gate and a double diffusion break (DDB) becomes necessary. With advances in aggressive STI oxide fill, SDB became available as early as the 14-nm node.

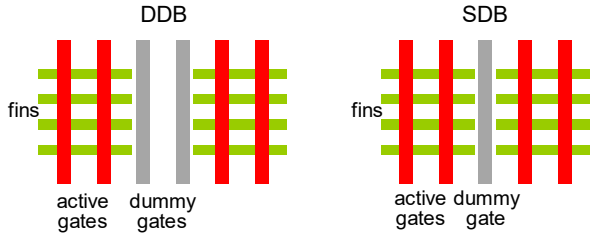

Fig. 8. Layout view of single vs. double diffusion break [22].

III. AMS Device Palette

A. Thin-Oxide Core Transistors

Short-channel FETs continue to offer PPAC advantages to digital CMOS-like AMS circuits such as ring VCOs. Analog circuits also experience some benefits. The 7-nm finFET offers 25–35% reduced V_T variation compared to 16 nm [5]. Sensitive high-speed circuits which require short-channel bandwidth performance, such as SerDes receivers, still require offset correction. The intrinsic gain of short-channel finFETs has not experienced notable node-to-node improvement but is already 3× better than a planar device [8]. Although device width quantization can be challenging for SRAM and some logic designs, its impact on analog designs remains minimal. With voltage headroom constraining gate overdrives to 50 mV or lower, g_m is granular enough at 10–100 µA/V per fin for sufficient design flexibility to attain typical values of 1 mA/V. The already weak body effect in finFET is even weaker in 7 nm with $\Delta V_T < 5$ mV for $|\Delta V_{BS}| = V_{DD}$. This offers some headroom relief as well as eliminates the need for hot n-wells (where the body of a PMOS is tied to its source instead of V_{DD}) and possibly even deep n-wells for RF noise isolation.

Long-channel FETs are still essential in most analog circuits for realizing near ideal current sources. Unfortunately, with HKMG integration, the maximum allowable gate length has been dramatically reduced from ~1 µm in pre-HKMG nodes to as low as 240 nm [23] to limit the extent of RMG polish dishing (Fig. 9) and gate-density-induced mismatch [24]. Due to the limited MG conductivity, gate charge cannot be completely contained inside the thin MG layer and spills into the metal fill. As a result, the effective gate work function is also influenced by the work function and height of the metal fill.

The stacked FET of Fig. 10 has become ubiquitous for building longer channel equivalents. The higher r_{out} is realized through source degeneration of the top device in the stack (in the case of NMOS) operating in saturation; the others operating in triode. Although the intermediate diffusions add layout area, the stacking of shorter channel devices, which benefit most from strain-enhanced mobility, may reduce overall area as fewer fingers are required. Stacking has also been shown to reduce device mismatch [23]. However, it compromises r_{out} at higher frequencies as the intermediate diffusions shunt more ac current to ground (Fig. 10). For example, the simulated r_{out} of a stack of L_{min} devices rolls off at 25% lower frequency than a single 240-nm device with the same low-frequency r_{out}. This can degrade analog metrics such as high-frequency intrinsic gain and common-mode noise rejection.

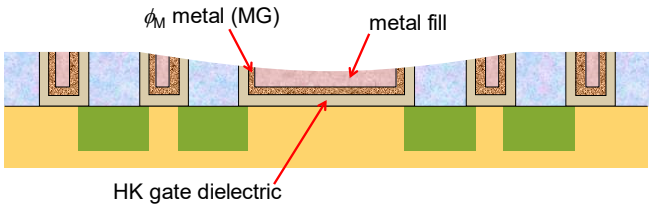

Fig. 9. HKMG long-channel gate dishing from RMG polish [24].

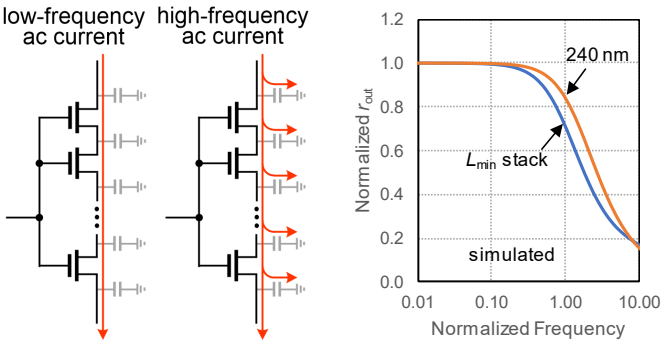

Fig. 10. Stacked FET r_{out} at low and high frequencies.

B. *Thick-Oxide I/O Transistors*

The 7-nm node continues to offer 1.8-V I/O transistors primarily to support general-purpose I/Os (GPIOs) that communicate with peripheral ICs made in cheaper technology nodes. This legacy support comes with nontrivial technology and design challenges. FinFETs with thicker gate dielectrics are increasingly difficult to build because fin pitch scaling requires a more aggressive MG ALD fill capability [25]. It is possible to integrate a second and wider fin pitch to accommodate 1.8-V devices at the expense of a more complex SAQP scheme. On the design side, the scaling of core V_{DD} to as low as 0.5 V in SoC sleep mode complicates the design of voltage level shifters that bridge the wide core and I/O voltage gap. Consequently, there is mounting pressure to lower the I/O FET voltage to 1.5 or 1.2 V. A thinner I/O device gate dielectric also offers PPAC benefits. As a result, memory interfaces like LPDDR4X, which link the SoC to a higher supply DRAM die, are already shifting to lower and more SoC-friendly signaling voltages. With reduced SoC core V_{DD}, a lower GPIO voltage is inevitable, but time is required to steer the entire chipset ecosystem.

C. *Passives (Resistors, Capacitors, and Inductors)*

The precision MEOL thin-film resistor, shown in Fig. 11(a), continues to be used. HKMG integration made the unsilicided poly-Si resistor obsolete at 20 nm. The MEOL resistor is composed of a thin refractory metal compound (e.g., TiN) that is deposited and subtractively etched with a dedicated mask. Built specifically for AMS usage, its integration is decoupled from the finFET. Thinning the resistor film for a higher sheet resistance will increase variation, making resistor area scaling an outstanding impediment to overall AMS area scaling.

Several capacitor options are available. Linear capacitors are still realized as interdigitated metal-oxide-metal (MOM) fingers in the BEOL stack. BEOL pitch scaling has certainly improved the attainable capacitance per unit area. It has, however, also increased the bottom layer parasitic capacitance which degrades the efficiency of ac coupling. A metal-insulator-metal (MIM) plate capacitor placed in the far BEOL may be available. It is primarily used for supply noise decoupling and does not have low enough plate resistance for high-speed AMS design. As additional processing is required, MIM capacitors are only justified in more expensive server ICs, not in mobile SoCs. The accumulation-mode varactor of Fig. 11(b) is also available for noise decoupling and *LC*-VCO tuning. The thick-oxide flavor is preferred for low gate leakage. In a fully depleted device with smaller *SS*, PLL VCO gain (dC_G/dV_G) is higher, but this benefit may require more careful varactor biasing given the narrower V_G window for useful tuning.

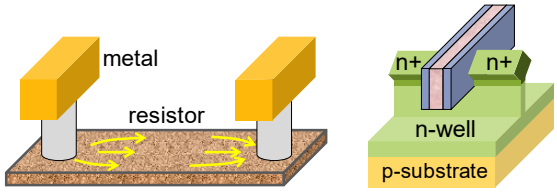

Fig. 11. (a) MEOL thin-film resistor and (b) finFET varactor.

Inductor design is minimally impacted as the thick upper BEOL layers used for building planar spirals remain unscaled to maintain low-droop power distribution. Subtle degradation of inductor *Q* will occur. Each metal level in the increasingly taller interconnect stack under the coil requires more stringent surrounding dummy fill to minimize accumulation of BEOL topography.

D. *PNP-BJT and ESD Diodes*

The finFET equivalents of the PNP-BJT and STI ESD diodes are shown in Fig. 12. The diode-connected PNP-BJT (analog diode) is routinely used for bandgap references and thermal sensors. The ESD diodes find their usual application in I/O signal pads where they shunt potential ESD currents to nearby supply clamps. As fin width is a small fraction of the fin pitch, the vertical well resistance is high despite higher fin doping beneath the device to suppress subsurface punchthrough leakage. The result is higher diode series resistance (R_D) although junction area capacitance is reduced.

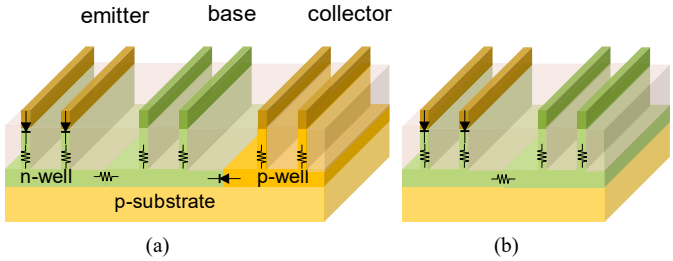

Fig. 12. (a) PNP BJT and (b) STI ESD diodes in finFET technology.

IV. AMS Design Impact

A. *Parasitic Resistance*

Arguably the most challenging aspect of AMS design with finFETs is coping with parasitic device, MEOL, and BEOL resistance; a burden that only gets worse in each new node.

Parasitic device resistance comes from the source/drain, gate, and well. FinFET source/drain resistance is high as currents funnel from trench contacts into the narrow fins through a diffusion with limited silicide. Short-channel gate resistance is also high, even with metal gates. The resistance is highest on top of the fin where the gate is already thin and made even thinner after being recessed for SAC formation. Contacting the gate on both sides of an active area and using groups of fewer fins are common area-bloating remedies to mitigate growing nonquasistatic effects. High R_D in analog and ESD diodes has degraded the diode ideality factor at higher currents (Fig. 13). In bandgap references (Fig. 13), this has forced the use of smaller current ratios (*N*) to generate the PTAT ΔV_{BE} which comes at a price of higher mismatch sensitivity [26]. The use of higher diode current ratios is possible in thermal sensors but requires R_D cancellation techniques. For example, [27] employs two ΔV_{BE} measurements and ratios. Higher well resistance is also responsible for some layout growth due to the higher density of well taps required to prevent latch-up. In addition, latch-up aggressors and victims in I/O pads, already surrounded by double guard rings, must be substantially spaced apart.

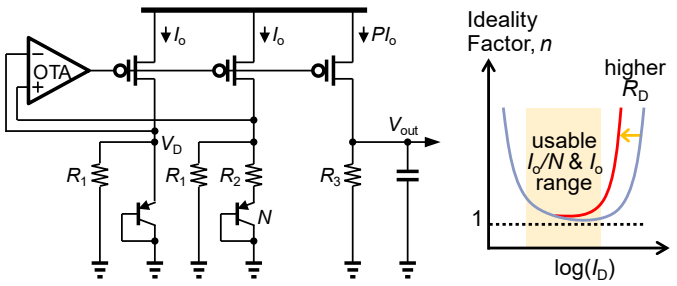

Fig. 13. Low-voltage bandgap reference [26].

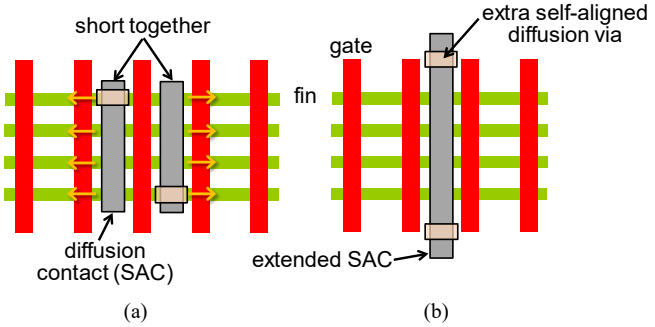

Fig. 14. (a) Double-source layout and (b) SAC extension.

Finer geometries and additional via levels have added substantially more resistance in the MEOL. Dense metallization enforces unidirectional routing which requires supply and signals to traverse through more vias, each with highly resistive contact interfaces. Techniques to reduce contact and via resistance are becoming vitally important, even at the expense of increased capacitance. For example, the double-source layout for multi-fingered devices and SAC extension for more diffusion vias (Fig. 14) are increasingly used to mitigate droop in high-current circuits such as I/O transmitters and clock buffers.

The resistance concern extends into the BEOL, especially at the lowest levels with tightest pitch for logic routability. Metal pitch scaling increases resistance at an exponential rate because the resistive TaN diffusion barrier that clads the copper wire is not scaling with pitch, leaving behind diminishing volume of the lower resistivity copper. For example, reducing the metal pitch from 80 to 48 nm results in a 6× increase in line resistance [22]. Although conductivity at these dimensions may be degraded by surface scattering and quantum confinement, the recent use of alternative barrier-less metal materials, such as cobalt and ruthenium, is showing improvement over copper for line pitch as aggressive as 36 nm [10], [14].

B. *Parasitic Capacitance*

The compact 3-D finFET geometry and corresponding denser interconnects to access the finFETs have heightened parasitic capacitance. In fact, in migrating from planar to finFET CMOS, dynamic power reduction required aggressive V_{DD} scaling to offset the higher capacitance [28]. C_{GS} and C_{GD} are particularly high due to the gate sidewall coupling to the trench SACs and epitaxial source/drain fill between fins, impacting analog design in a variety of ways. For example, in Fig. 15(a), higher C_{GD} (Miller) coupling in a low-dropout (LDO) regulator with a PMOS pass element causes worse high-frequency supply noise rejection. In another example (Fig. 15(b)), higher C_{GS} injects more kickback noise in a single-ended LPDDR receiver, commonly implemented as a PMOS differential amplifier with one input tied to a V_{ref} threshold. Increased C_{GS} and C_{GD} in a varactor also degrades VCO tuning range.

Scaling will continue to increase interconnect capacitance. The current interconnect system incorporates a porous SiO_2 with an already low dielectric constant (K) of 2.6–2.7. Increasing dielectric porosity to further reduce K will exacerbate mechanical integration, package stress, and reliability issues.

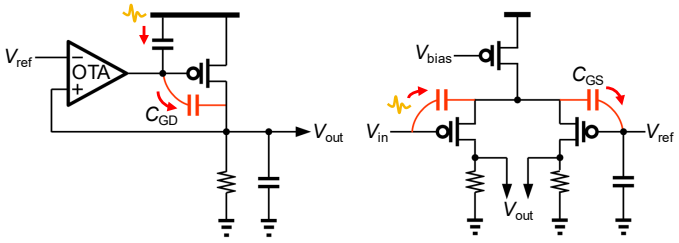

Fig. 15. (a) LDO regulator with PMOS pass element and (b) LPDDR receiver.

C. *Layout-Dependent Effects (LDEs)*

The addition of FET performance elements such as process-induced mechanical strain and HKMG has given rise to local transistor LDEs which have been identified and incorporated into the transistor models and layout extraction. LDEs can be responsible for non-trivial post- vs. pre-layout simulation discrepancies that make design iterative and time-consuming.

Process-induced strain for mobility boost is more aggressive in each new node. Channel stress, especially in short-channel devices, is more readily influenced by its source/drain volume and CGP, and is also perturbed by the stress and proximity of other devices in the same active area (also called OD for oxide definition), surrounding dielectric isolation, and neighboring devices as depicted in Fig. 16 [29], [30]. For instance, the device current per fin depends on the number of contiguous fins and fingers. The finFET 3-D geometry has also given rise to new complex stress interactions such as the gate cut effect shown in Fig. 17(a) [22]. Here, cutting the gate between adjacent groups of fins disrupts the mechanical support provided by a continuous gate and consequently modulates the stress of fins near the cut. SDBs also have a profound impact by severing the continuity of the heavily strained diffusions [22]. This is a manifestation of the well-established LOD (length of OD) effect where channel stress and mobility depend on gate location within an OD as captured by the gate-to-OD-edge distances (SA and SB) of the individual gates [31]. Short ODs are partic-

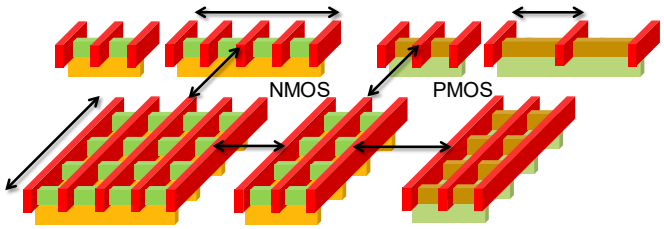

Fig. 16. Layout-dependent effects due to surrounding isolation and devices.

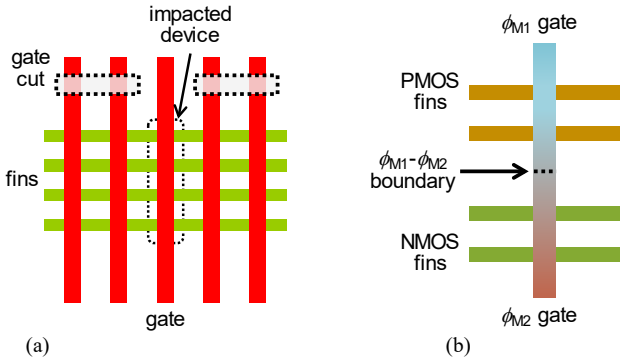

Fig. 17. (a) Gate cut effect and (b) metal boundary effect.

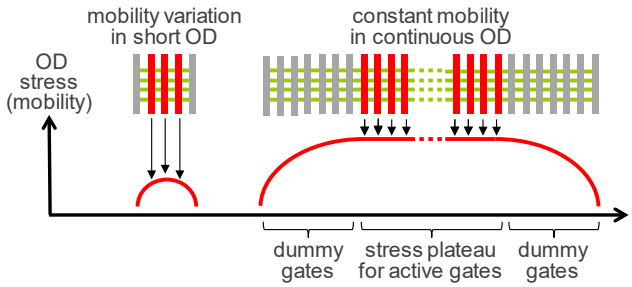

Fig. 18. Continuous OD concept to establish stress plateau for device matching.

ularly prone to stress variation and have motivated the use of continuous OD for better matching. Shown in Fig. 18, contiguous dummy gates on both OD ends build up a stress plateau on which active gates are placed to reduce mobility sensitivity to SA and SB. In a continuous OD, the higher plateau mobility reduces the number of device fingers to attain a given current.

The metal boundary effect (MBE) has been observed with the addition of HKMG [33]. When a single gate is composed of two dissimilar ϕ_M materials for more compact layout (e.g., inverter NMOS and PMOS in Fig. 17(b)), fins near the ϕ_M boundary experience some V_T shift. It is postulated that interdiffusion of the ϕ_M metals is responsible for this phenomenon.

D. Spacer-Based Patterning

Gate and metal SADP have imposed new layout constraints. Some non-minimum metal spacing is no longer allowed. Variation in pattern density introduces spacer deposition and etch loading effects that affect the spacer width. Such effects have been observed in planar CMOS where neighboring long-channel devices would impact short-channel performance through the gate spacer width and corresponding extension resistance.

Gate SADP CD control can be improved by grouping short-channel devices together to avoid mixing the spacer-patterned short-channel devices with the longer channel devices that are patterned with another mask. For example, in a current mirror, auxiliary switches added for enable control are typically short-channel devices placed in a sea of long-channel devices. For channel length consistency, it is preferred to convert the long-channel mirror devices to stacked devices or increase the channel length of the auxiliary devices as depicted in Fig. 19.

In the MEOL and BEOL, layout may need to comply with new minimum perimeter density constraints, in addition to the usual area density rules, to enforce more pattern uniformity in

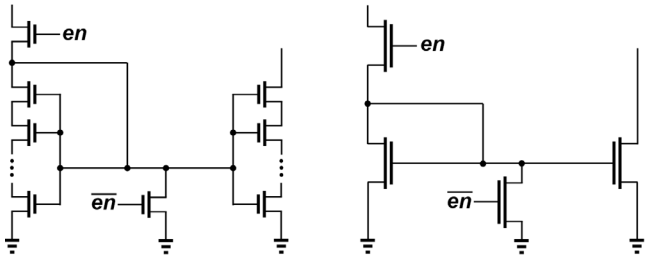

Fig. 19. Current mirror with auxiliary enable (*en*) control in all short- or long-channel devices.

line widths and spaces. For this reason, a set of parallel narrow lines is preferred over a single wider line, providing higher perimeter density at the cost of area.

E. Other Layout Considerations

Tougher layout rules in each new node have been forcing AMS layout styles to increasingly resemble logic gate arrays. Intended to minimize long-range variation and cumulative wafer topography, the number of area density, perimeter density, and corresponding density gradient constraints continue to grow. Contact, via, and even cut mask density need to comply to a manufacturable range. Density rules further extend to larger checking windows and to the cumulative density of multiple contiguous metal levels. As a result, layout closure requires more iterative rework of smaller cells in the layout hierarchy.

AMS floorplanning has also become far more tedious. Significant bloat is incurred to accommodate the increasing use of required dummy gates, well taps, and guard rings. Area growth also results from the necessary transition fill zones inserted between regions of different device types and pattern densities.

V. Concurrent Technology Development

A. Perspectives

SoCs built in the latest nodes are designed concurrently with technology development for faster time-to-market [1]. The profit premium for early market entry is simply too high to wait for technology readiness. As a result, the models provided to early adopters for design enablement are inherently speculative and optimistic. They are based on realistic and nontrivial projections of future technology capability but aggressive enough to entice new product interest.

In this fluid design infrastructure vulnerable to missed projections, designers accept the burden of periodic model and design rule updates, some very late and dramatic, during the design cycle. The situation is more precarious for AMS design. The process is continuously tuned to match short-channel logic and SRAM device targets as its top priorities. Besides logic FETs and the MEOL resistor, the AMS device palette consists entirely of slave devices prone to model adjustments as the fab approaches the logic device targets; long-channel and I/O FETs being frequent victims. The logic and SRAM palettes are not immune either. Each new node is less mature than the previous one when risk production must start with the lead product. With growing mask counts leading to longer wafer lead times, the fab simply has fewer cycles of silicon learning in its development timeline to achieve the aggressive logic and SRAM targets.

B. Reducing AMS Design Exposure

When models and design rules are updated, the rework effort for AMS circuit and layout is typically greater than that for logic and SRAM. Design automation is limited for the broad classes of AMS circuits, each with unique design requirements. However, because AMS subsystems consume far less area than logic and SRAM, and are not as densely compacted, device-level layout can often tolerate incremental layout growth with little to no impact on die cost. We can exploit this trait to incorporate some design resilience to model changes. Table I proposes layout guidelines that address areas where model parameters are not stable during process development due to ongoing optimization and are susceptible to retargeting. Critical process areas include source/drain epitaxy, RMG work function tuning of the device V_T flavors, and contact formation.

TABLE I. LAYOUT GUIDELINES TO REDUCE AMS DESIGN EXPOSURE TO MODEL RETARGETING

Layout Guideline	Benefit
Use continuous OD stress plateau for active device placement	Desensitize devices from stress-related LDEs (process-induced strain, diffusion epitaxy, STI)
Attach dummy devices to OD ends	
Avoid single-diffusion break	
Use only one ϕ_M metal in each gate	Eliminate metal boundary LDE
Avoid using gate as interconnect	Eliminate gate cut LDE
Add contacts on both sides of gate	Reduce impact of gate resistance with ϕ_M metal tuning to adjust V_T
Use groups of fewer fins	
Use redundant SACs (e.g., double source layout)	Reduce impact of SAC interface resistance and diffusion epitaxy
Use redundant diffusion vias with SAC extension	Reduce impact of diffusion via interface and diffusion resistance
Do not unnecessarily push design rules to the limit	Reduce exposure to design rule update to more conservative limit

VI. CONCLUSION

With incessant SoC technology focus on logic and SRAM, AMS design in the remaining CMOS nodes is an increasingly tedious endeavor of managing technology-imposed non-idealities. AMS designers are pressed to comprehend the technology even more than ever to develop ways to overcome these impairments. The impact of parasitics and layout-dependent effects will only get worse in each new node. Contact and via resistances are so high that they are expected to ultimately limit scaling [9]. From a cost perspective, the overall area of AMS subsystems has only been scaling at a modest rate of 0.8–0.9× with each new node, unlike the target entitlement of 0.5× for logic and SRAM area. This disparity is lowering the threshold for migrating to system-in-package alternatives where, for example, AMS subsystems are partitioned to another on-package die that is cheaper and more AMS-friendly. Meanwhile, AMS designers must continue to assimilate the high-performance and low-power benefits of advanced SoC technology to maintain legacy system capabilities as well as enable new ones not possible in earlier nodes. Implementation just involves more perspiration.

REFERENCES

[1] L. Bair, "Process/product interactions in a concurrent design environment," in *IEEE CICC*, 2007.

[2] C. Auth *et al*., "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *IEEE Symp. VLSI Technology*, 2012.

[3] S.-Y. Wu *et al*., "A 16nm finFET CMOS technology for mobile SoC and computing applications," in *IEEE IEDM*, 2013.

[4] H.-J. Cho *et al.*, "Si finFET based 10nm technology with multi Vt gate stack for low power and high performance applications," in *IEEE Symp. VLSI Technology*, 2016.

[5] S.-Y. Wu *et al*., "A 7nm CMOS platform technology featuring 4th generation finFET transistors with a 0.027um^2 high-density 6-T SRAM cell for mobile SoC applications," in *IEEE IEDM*, 2016.

[6] P. Packan *et al.*, "High performance 32nm logic technology featuring 2nd generation high-k + metal gate transistors," in *IEEE IEDM*, 2009.

[7] L. Wei *et al*., "Exploration of device design space to meet circuit speed targeting 22nm and beyond," in *Int. Conf. SSDM*, 2009.

[8] F.-L. Hsueh *et al*., "Analog/RF wonderland: circuit and technology co-optimization in advanced finFET technology," in *IEEE Symp. VLSI Technology*, 2016.

[9] C. Hou, "A smart design paradigm for smart chips," in *IEEE ISSCC*, 2017.

[10] D. Yang, "SoC scaling challenges in the era of the single digit technology nodes," in *Int. Workshop Advanced Patterning Solutions*, 2017.

[11] W. Arnold *et al.*, "Manufacturing challenges in double patterning lithography," in *IEEE ISSM*, 2006.

[12] C. Auth *et al.*, "45nm high-k + metal-gate strain-enhanced transistors," in *IEEE Symp. VLSI Technology*, 2008.

[13] Y.-K. Choi, T.-J. King, and C. Hu, "A spacer patterning technology for nanoscale CMOS," *IEEE Trans. Electron Devices*, vol. 49, no. 3, 2002.

[14] C. Auth *et al.*, "A 10nm high performance and low-power CMOS technology featuring 3rd generation finFET transistors, self-aligned quad patterning, contact over active gate and cobalt local interconnects," in *IEEE IEDM*, 2017.

[15] Y. Woo *et al*., "Design and process technology co-optimization with SADP BEOL in sub-10nm SRAM bitcell," in *IEEE IEDM*, 2015.

[16] R. Brain *et al*., "Low-k interconnect stack with a novel self-aligned via patterning process for 32nm high volume manufacturing," in *IEEE IITC*, 2009.

[17] S. H. Yang *et al*., "28nm metal-gate high-K CMOS SoC technology for high-performance mobile applications," in *IEEE CICC*, 2011.

[18] C. Y. Kang *et al*., "The impact of La-doping on the reliability of low Vth high-k/metal gate nMOSFETs under various gate stress conditions," in *IEEE IEDM*, CA, 2008.

[19] L. Chang *et al*., "Gate length scaling and threshold voltage control of double-gate MOSFETs," in *IEEE IEDM*, 2000.

[20] V. Chan *et al*., "Strain for CMOS performance improvement," in *IEEE CICC*, 2005.

[21] M. Rashed *et al.*, "Innovations in special constructs for standard cell libraries in sub 28nm technologies," in *IEEE IEDM*, 2013.

[22] S. Yang *et al.*, "10 nm high performance mobile SoC design and technology co-developed for performance, power, and area scaling," in *IEEE Symp. VLSI Technology*, 2017.

[23] F.-L. Hsueh, "Device challenges for scaled analog-RF," in *IEEE Symp. VLSI Technology*, Short Course, 2017.

[24] S. Yang *et al.*, "High-performance mobile SoC design and technology co-optimization to mitigate high-K metal gate process variations," in *IEEE Symp. VLSI Technology*, 2014.

[25] A. Wei *et al*., "Challenges of analog and I/O scaling in 10nm SoC technology and beyond," in *IEEE IEDM*, 2014.

[26] H. Banba *et al.*, "A CMOS bandgap reference circuit with sub-1-V operation," *IEEE J. Solid-State Circuits*, vol. 34, no. 5, 1999.

[27] "ADT7461 ±1°C temperature monitor with series resistance cancellation," *ON Semiconductor Pub. No. ADT7461/D*, 2014.

[28] E. Terzioglu, "Design and technology co-optimization for mobile SoCs," in *IEEE ICICDT*, Keynote, 2015.

[29] J. Faricelli, "Layout-dependent proximity effects in deep nanoscale CMOS," in *IEEE CICC*, 2010.

[30] M. Garcia Bardon *et al.*, "Layout-induced stress effects in 14nm & 10nm finFETs and their impact on performance," in *IEEE Symp. VLSI Technology*, 2013.

[31] R.A. Bianchi *et al*., "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *IEEE IEDM*, 2002.

[32] X. Xi *et al*., BSIM4.3.0 MOSFET Model User's Manual, *Regents of Univ. California at Berkeley*, 2003.

[33] M. Hamaguchi *et al.*, "New layout dependency in high-K/metal gate MOSFETs," in *IEEE IEDM*, 2011.