J. Low Power Electron. Appl. **2012**, *2*, 1-29; doi:10.3390/jlpea2010001

OPEN ACCESS

Journal of Low Power Electronics and Applications

ISSN 2079-9268

www.mdpi.com/journal/jlpea/

Review

CMOS Leakage and Power Reduction in Transistors and Circuits: Process and Layout Considerations

Eitan N. Shauly [1,2]

[1] TowerJazz Corporation, Migdal Ha'Emek, 10556, Israel; E-Mail: eitan.shauly@towerjazz.com; Tel.: +972-4-6506570; Fax: +97204-547788

[2] The Department of Materials Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

Received: 12 December 2011; in revised form: 14 January 2012 / Accepted: 16 January 2012 / Published: 27 January 2012

Abstract: Power reduction in CMOS platforms is essential for any application technology. This is a direct result of both lateral scaling—smaller features at higher density, and vertical scaling—shallower junctions and thinner layers. For achieving this power reduction, solutions based on process-device and process-integration improvements, on careful layout modification as well as on circuit design are in use. However, the drawbacks of these solutions, in terms of greater manufacturing complexity (and higher cost) and speed degradation, call for "optimized" solutions. This paper reviews the issues associated with transistor scaling and related solutions for leakage and power reduction in terms of topological design rules and layout optimization for digital and analog transistors. For standard cells and SRAMs cells, leakage aware layout optimization techniques considering transistor configuration, stressors, line-edge-roughness and more are presented. Finally, different techniques for leakage and power reduction at the circuit level are discussed.

Keywords: low leakage; low power; layout optimization; transistor scaling; leakage-related-stressors; design-aware leakage reduction

---

1. Introduction

Transistor scaling that has driven the CMOS technology for the last 45 years increased the transistor density. However, the power consumption and the leakage current of scaled down transistors increase rapidly and thus, some "classical" scaling rules like gate oxide thinning can no longer be maintained. The increased number of transistors per chip and reduction in die size leads to rapid increase in power. Due to the fact that the device clock frequency has increased with each new generation, but the power supply was not scaled down at the same ratio the dynamic power is now the dominant power factor for 65 nm platforms (Table 1, Figure 1). To overcome this problem, in most cases, foundries are offering platforms with several technologies. The different "technologies" refer to the thin-oxide transistors' parameters and operational voltage per the related application (Table 1). In most cases, the SL (Standard Logic for General Purposes) technology will have FEOL (Front-End-of-Line) with thinner gate oxide thickness, lower operation voltage, higher drive currents and lower threshold voltages compared to the Low-Power (LP). Interactive audio and/or video mobile platforms have both dynamic and static high power consumption. To meet these opposite demands, "mixing technology" is also proposed [1,2], with a "triple-gate". We will discuss this solution later on. The BEOL (Back-End-of-Line, metal and dielectric layers) and most of the analog passive components like resistors, junction varactors and thick-oxide MOSFET-varactors are common for all technologies at the same platform.

Table 1. Typical device specifications for 65 nm to 32 nm technologies, for both Standard Logic for General Purposes, Low-Power and High Performances. Data from [1,3–5].

Platform (node)	65 nm						45 nm		40 nm			32 nm		
Technology Application	Standard Logic for General Purposes			Low Power			High Performance LP		LP for mobile/WiFi			General Purpose—High-Performance		
V_{dd} **(V)**	1			1.2			1		1.1			0.9		
Stressors							SMT, SiGe, cSEL		SMT, cSEL			SMT, SiGe, cSEL		
Gate Formation	Poly/SiON						Poly/SiON		Poly/SiON			HK/MG		
Reference	ST, NXP, CEA-LETI Minatec [1]						Fujitsu [3]		TOSHIBA [4]			IBM Alliance [5]		
T_{ox}**_Inv N/P (Å)**	20.5/22.5			26/27.5			19.1/20.5		24.5/25.5			12/14		
V_t **Type**	HVt	SVt	LVt	HVt	SVt	LVt	HVt	SVt	HVt	SVt	LVt	HVt	SVt	LVt
Ion_N/Ion_P (µA/µm)	670/295	830/398	950/450	420/210	610/310	740/390	970/630	1220/765	487/235	715/295	840/370	855/550	1050/650	1250/790
Isub_N/Isub_P (nA/µm)	5/3	51/40	130/130	0.015/0.009	0.36/0.10	5/2.5	10/10	100/100	0.03/0.03	0.4/0.4	6/3	1/1	10/10	100/100
Igate_N/Igate_P (A/cm^2)	8/3			0.02/0.005			30/20		0.2/0.07			0.4/0.2		
Gate Delay RO 1FO (ps/gate)	14	10.5	8.8	25.5	17.5	13.5								

Figure 1. Total power ratio evolution *vs.* platform (node) and technology (application) [2] (left). "SL" refers to "Standard Logic for General Purposes" and "LP" for "Low-Power". IBM leakage values for high-performances and LP technologies for the different platforms [6] (right).

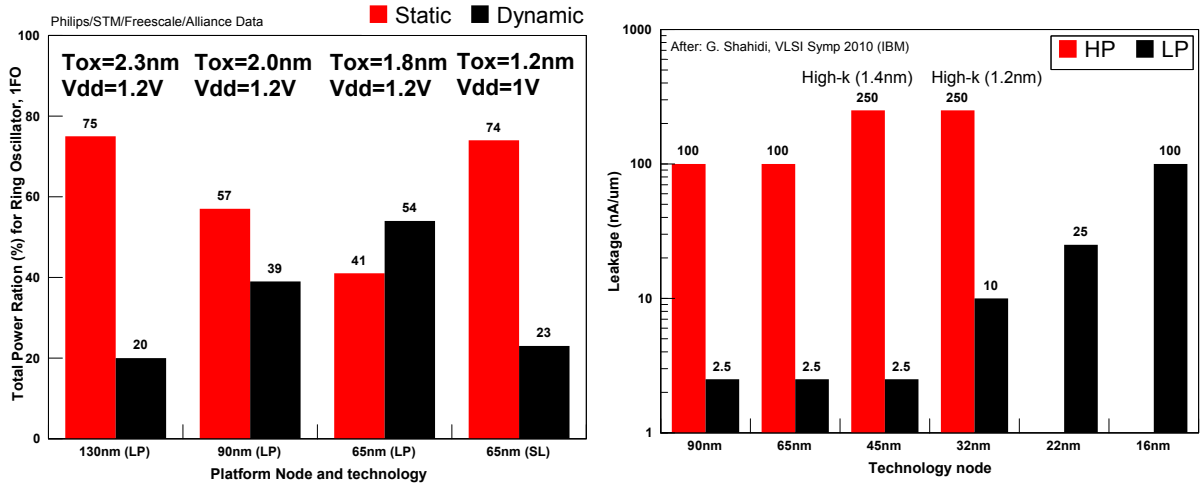


Table 1 contains a short list of benchmark specifications for 65 nm down to 32 nm platforms, including specifications for supply voltage (V_{dd}), gate oxide thickness, drive current (I_{on}) and sub-threshold leakage (I_{sub}). As can be seen, I_{sub} typical values go up by several orders of magnitudes as technology is scaled down where V_{dd} is reduced by only 30%. This is the main challenge that will be discussed in this paper. We will use Table 1 all along this paper, for technologies comparison. It is interesting to compare IBM I_{sub} for HV (High-Speed) and LP technologies, as shown in Figure 1, with the data in Table 1: for 65 nm, the HP leakage is close to the SL/LVt, and LP is close with the LP/LVt. For 45 nm, IBM use a high-*k dielectric*, and the leakage values are lower by a factor of 5 compared to LP/HVt in [4].

There are several sources for power dissipation (P) in digital CMOS circuits [7]:

$$P_{avg} = P_{dynamic} + P_{static} = (P_{short} + P_{switch}) + P_{static} = I_{sc}V_{dd} + \alpha C_L V_{dd}^2 f + I_{leak}V_{dd} \quad (1)$$

The first term P_{short} is the power consumed during gate voltage transient time, that in CMOS technology is only related to the direct path short circuit current (I_{sc}) which flows when both the NMOS and PMOS transistors are simultaneously active, conducting current directly from supply V_{dd} to ground or V_{ss}.

The second term, P_{switch} refers to the dynamic component of switching power due to charging and discharging C_L—is the total loading capacitance, f is the clock frequency and α is the average switching activity factor (typical value for α is 20% for logic blocks in 65 nm technology [8]). Some techniques for P_{switch} reduction are described in the next section.

Imperfect cut-off of the transistor leads to leakage (I_{leak}) and power dissipation (P_{static}) even without any switching activity. With an increasing number of gates both the total capacitance and the channel width are relevant for the leakage increase.

This paper is organized as follows: we will start by generally reviewing the transistor's leakage components related to scaling. Later, we will describe some layout parameters and related design rules

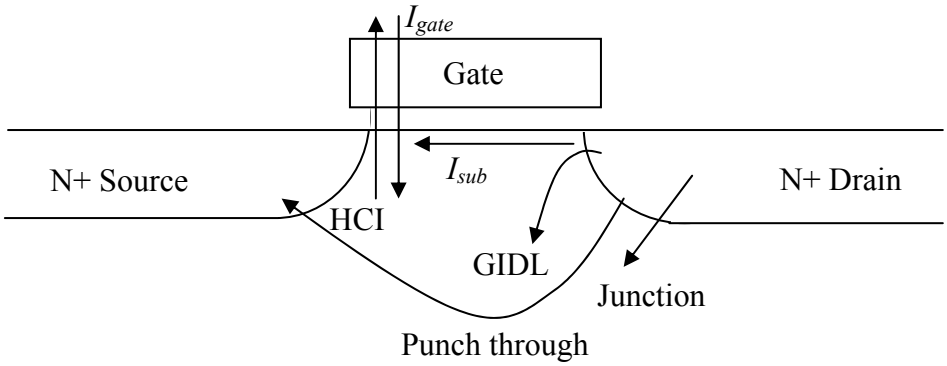
that affect leakage. Some guidelines for layout optimization for power (or leakage) reduction will be given. In Section 3, we describe some operational and layout consideration for power reduction in SRAM. Finally, techniques for power reduction in transistors and circuit level will be discussed for both core and SRAM.

2. Transistors Leakage Components

The main feature of transistors scaling is the reduction in V_{dd}, the threshold voltage (V_t), effective channel length, T_{ox} and doping levels and depth (Table 1). In this section, we will discuss some of the dependency of the transistor leakage components to these parameters.

As analyzed in [7,9,10], the overall leakage currents can be divided into several components (Figure 2), taking place under different bias conditions. At very low gate voltage, a potential difference between source and drain still results in sub-threshold static leakage current, I_{sub}. Among the many parameters, I_{sub} dependence on higher threshold voltage (V_t) and operation temperature is the most significant, reducing I_{sub} in an exponential manner with increasing V_t and decreasing temperature, respectively. Basically, lower channel doping, shorter effective channel length and longer transistor width will reduce V_t and increase I_{sub}. In addition, the body-factor and DIBL (Drain-Induced-Barrier-Lowering) parameters, that depend on the 1D and 2D doping profiles of the V_t adjust halo/pocked and extensions implants will also affect I_{sub}.

Figure 2. Schematic description of the different leakage currents and mechanisms in deep-submicron transistor.



Transistor scaling also means shallower and more abrupt extensions and S/D junctions. Although more abrupt junctions provide improved short channel effect, the rising doping concentrations and the high electric field (>10^6 V/cm) across the reverse-biased p-n junction lead to leakage due to Band-To-Band-Tunneling (BTBT) [9]. Higher gate-to-drain voltage increases the vertical field in the drain depletion layer, and reduces the depletion width at the gate-drain overlap area, resulting in Gate-Induced-Drain-Leakage (GIDL) [9]. For having a good V_t control, and to reduce the I_{sub} leakage, the dopants concentration near the surface are kept high. However, an increase of the drain voltage lowers the potential barrier for the majority carriers at the source side, thus leading in "additional" I_{sub} leakage and the punchthrough.

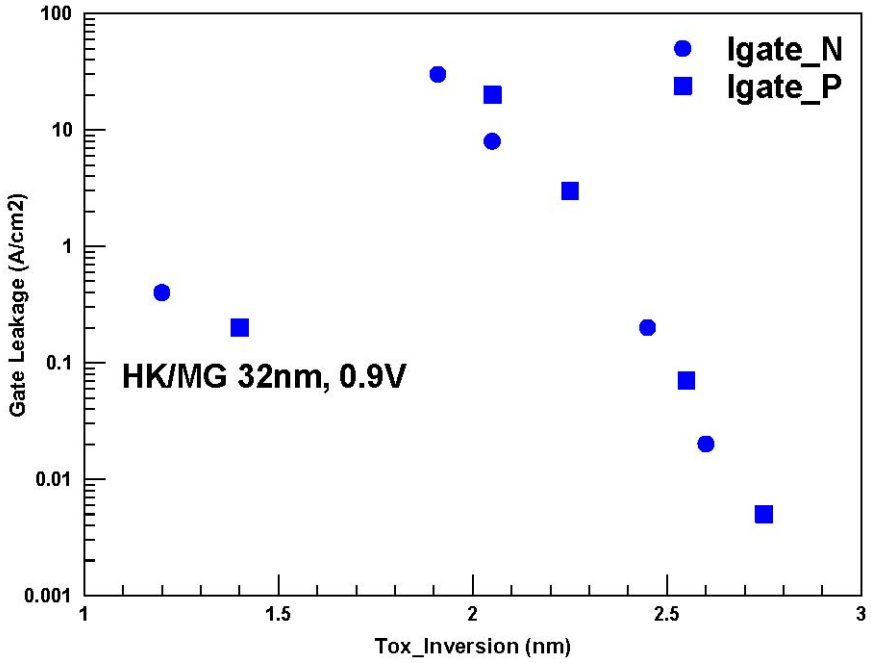
For technology generation of 65 nm and below, due to aggressive gate oxide thickness (T_{ox}) reduction, direct tunneling through the gate oxide leads to gate leakage (I_{gate}), that becomes dominant

over I_{sub}. In [7], the gate leakage is simply approximated using W (transistor width) and K_1 and K_2 that are constants which can be extracted experimentally:

$$I_{gate} = K_1 \cdot W \left(\frac{V_{DD}}{T_{ox}} \right)^2 e^{-K_2 T_{ox}/V_{DD}} \quad (2)$$

Figure 3 describes the gate leakage dependence on the gate oxide thickness. The exponent is much more dominant then the (V_{dd}/T_{ox}) part in the pre-exponent.

Figure 3. Gate leakage *vs.* Gate oxide thickness for Poly/SiON (65 nm to 40 nm platforms) and HK/MG (for 32 nm), based on data from Table 1. For the same effective oxide thickness, the gate leakage is lower by ~ 3 orders of magnitudes comparing to oxynitridization thermal oxide.



For 130 nm, I_{sub}, GIDL and junction leakage, cover ~95% of the overall leakage, and I_{gate} < 5%. For 90 nm, I_{gate} is ~40% and for 65 nm, it is >90%. Note that these percentages refer to leakages at room-temperature. As temperature goes-up, both I_{sub} and the junction leakage become more dominant [2]. Another factor which affects the ratio between the different components is the V_t target: in multi-V_t technology, having for example 3 types of V_t's, the high-V_t (HVt) will have 25% leakage due to I_{gate}, 25% leakage due to diodes and ~50% leakage due to I_{sub}. Regular (or Standard V_t, SVt) will have <5% for I_{gate} and diode and ~90% for I_{sub}. In Low-V_t, I_{sub} is the dominant (>98%) [1]. The 32 nm SL (Standard Logic for General Purposes) foundry technology node is the first one to use *high-k* material that allows reducing I_{gate} while keeping good gate control on the channel. About 3 order of magnitude reduction of I_{gate} can be achieved for the same effective oxide thickness (Figure 3).

In addition to gate current due to tunneling, Hot Carrier Injection (HCI) at the channel pinch-off area leads to impact ionization and leakage injection into the gate oxide.

Another aspect of scaling is the increase of inter-die thermal gradients due to the increase of the local power densities. Higher thermal gradients increase the voltage drop due to increased leakage.

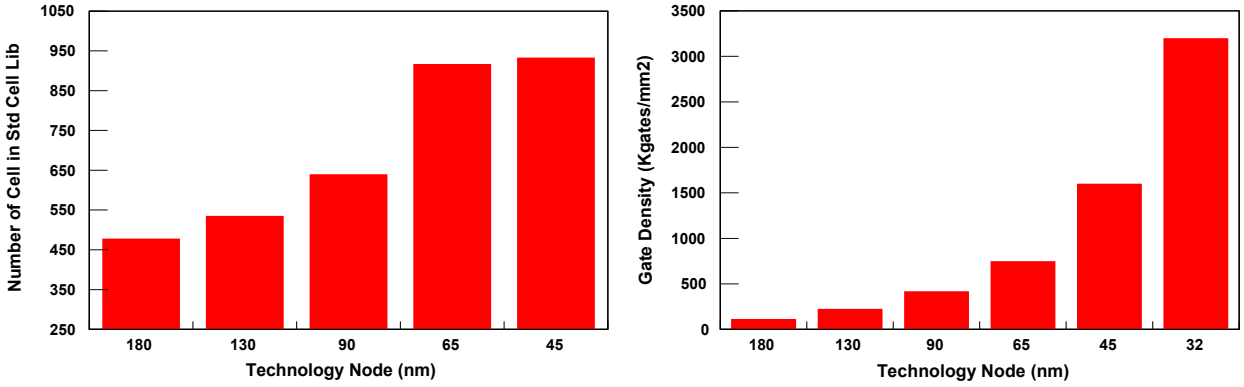
This voltage drop affecting the clock skew. Kawa [11] found voltage drop of 12% and 16% for 30° thermal gradients for 0.18 μm and 0.13 μm technology nodes, respectively.

3. Transistor and Cell Level Leakage Analysis and Optimization

3.1. Topological Design Rules and Layout Optimization

In addition to continued reduction in transistor dimensions along the scaling, also the transistor configuration (or "transistor layout"), as used by standard cells become more and more complex. At this section, we will discuss some of the transistor leakage dependency to the layout "style". Although the number of the different functions supported almost did not change during the years, the number of different cell types has increased in ~×1.2 at every technology node (Figure 4). The gate density is increased by factor of ×2 as required by basic scaling. More cell types and with more demanding design rules increase the challenge to reduce the leakage dependence on layout.

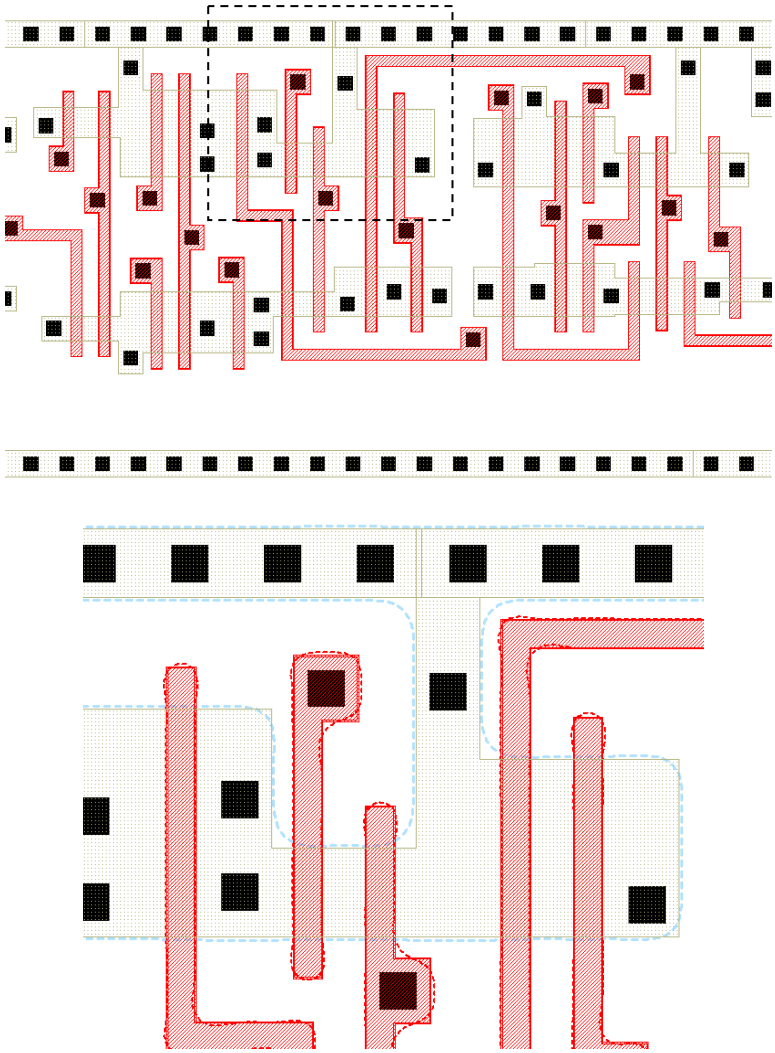
Figure 4. Number of different cells *vs.* technology nodes (left), library density *vs.* technology node (right). Both are for CMOS technology, for logic and Analog applications.



Another aspect for analysis of the complex topography, is the OPC (Optical-Proximity-Correction) implemented by the semiconductor foundry after the design is completed and prior to mask making. Basically, during OPC, small corrections are made to the design by attaching (or removing) small polygons. This OPC procedure takes place for the active area (AA), poly, all the metal layers and in the advanced platforms (≤90 nm), also for contacts and vias.

Figure 5 shows a snapshot from a standard cell library used in mass production, and the "on-silicon" shapes, based on modeling that takes into consideration the OPC and the manufacturing photolithography illumination conditions.

Figure 5. A snapshot from standard cell library, shows the drawn active area and poly complexity. The dash-box marked an area shown at the picture below, with drawn data and the layout after OPC.

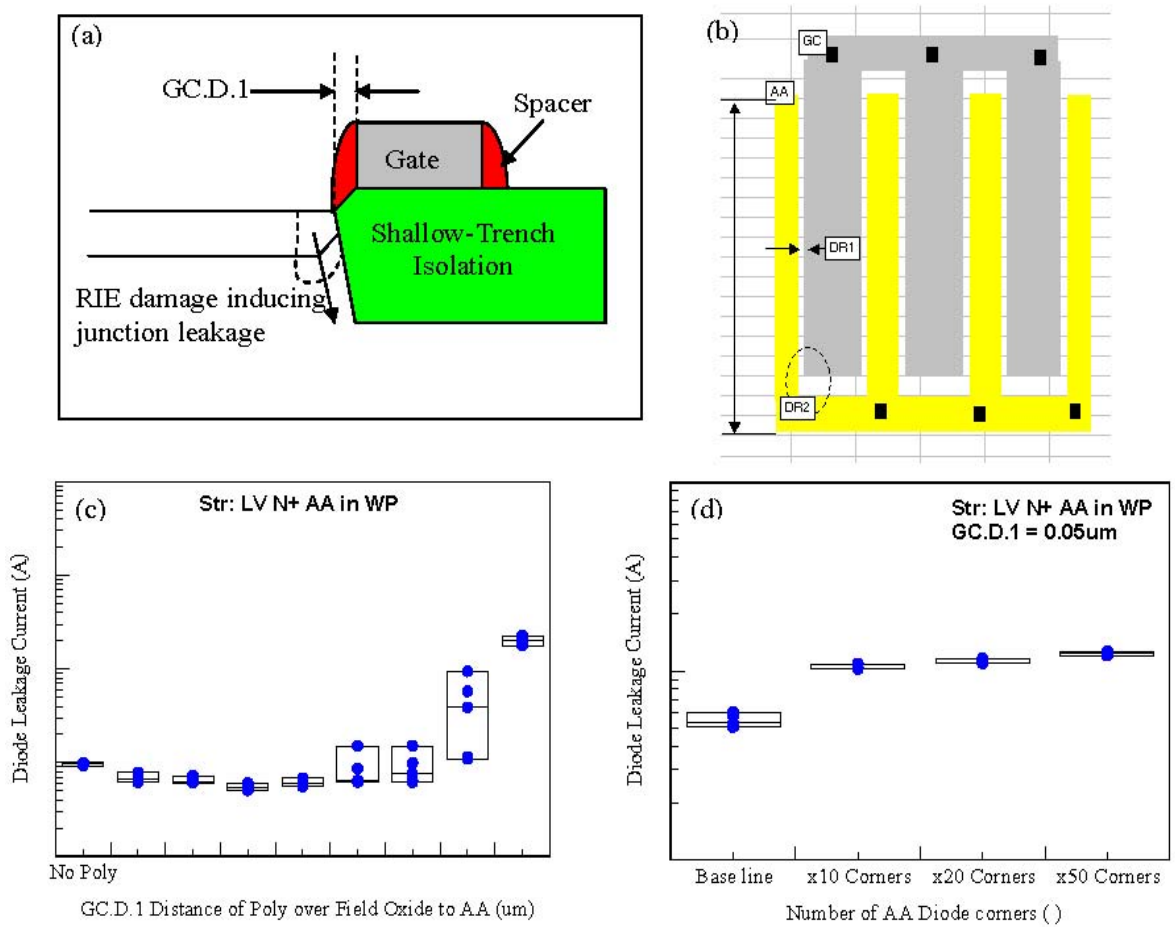


About 20 TDR (Topological Design Rules) are needed, for drawing the cells shown above. Among them, several rules have a direct relation with the transistor leakage, and therefore, should be optimized for low-power design. The analysis below covers some of these layout rules that are listed in Table 2.

Table 2. Main design-rules needed for building transistors in standard-cell library. See Figure 16 for illustration of these rules.

Rule	Design Rule Description	180 nm	130 nm	90 nm	65 nm
GC.D.1	Distance of Poly (over STI) to related AA	0.10	0.07	0.05	0.05
GC.X.2	Extension of poly beyond AA (end-cap)	0.22	0.18	0.16	0.14
AA.D.3	Distance between WN to N+ in WP	0.43	0.34	0.22	0.16
AA.E.3	Enclosure of WN around P+ in WN	0.43	0.34	0.22	0.16
CS.D.1/2	Distance of CS over AA to related Gate	0.16	0.11	0.11	0.09

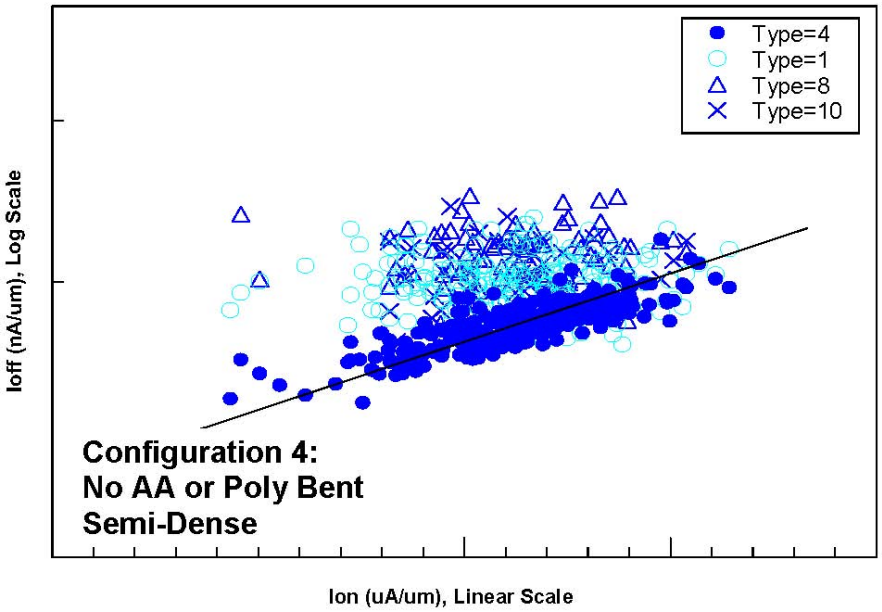
Figure 6. The effect of GC.D.1 (Distance of Poly over STI to related AA) on the junction leakage (**a**) schematic x-section showing the spacer etch damage at the junction corner; (**b**) top-view of the electrical test structure used for checking the junction leakage as a function of GC.D.1; (**c**) junction leakage dependence on GC.D.1 (all with the same number of AA corners); (**d**) junction leakage dependence on the number of corners (all with the same GC.D.1).



GC.D.1 and transistor configuration: Several papers already discussed the effect of the distance between the poly (over STI) to related AA. If the poly is too close and rounded, it may affect the transistor gate length [12], and because of that, it is always recommended (if possible) to have larger distance. In terms of process and leakage interaction, the distance GC.D.1, also affects the exact corner location of the spacer/AA (Figure 6). In case the spacer corner is located too close to the AA/STI boundary, the damage to the silicon substrate during the spacer etch-back can cause junction leakage. The example below (Figure 6), shows the N+/WP junction leakage, as function of GC.D.1, using a dedicated test structure consisting of diffusion comb interdigitated with the poly over STI comb. As can be seen, if the distance is large enough, the leakage is low and almost similar to that of junction w/o poly-near-by. However, for a too short distance, the leakage and the leakage spread both increase. Figure 6 also shows the dependence of the leakage value on the number of diffusion corners. Naturally, the higher the number of corners, the higher the junction leakage (for the same value of GC.D.1). Therefore, for low-power design relaxing GC.D.1 and avoiding using complex transistors with a large amount of AA corners is recommended.

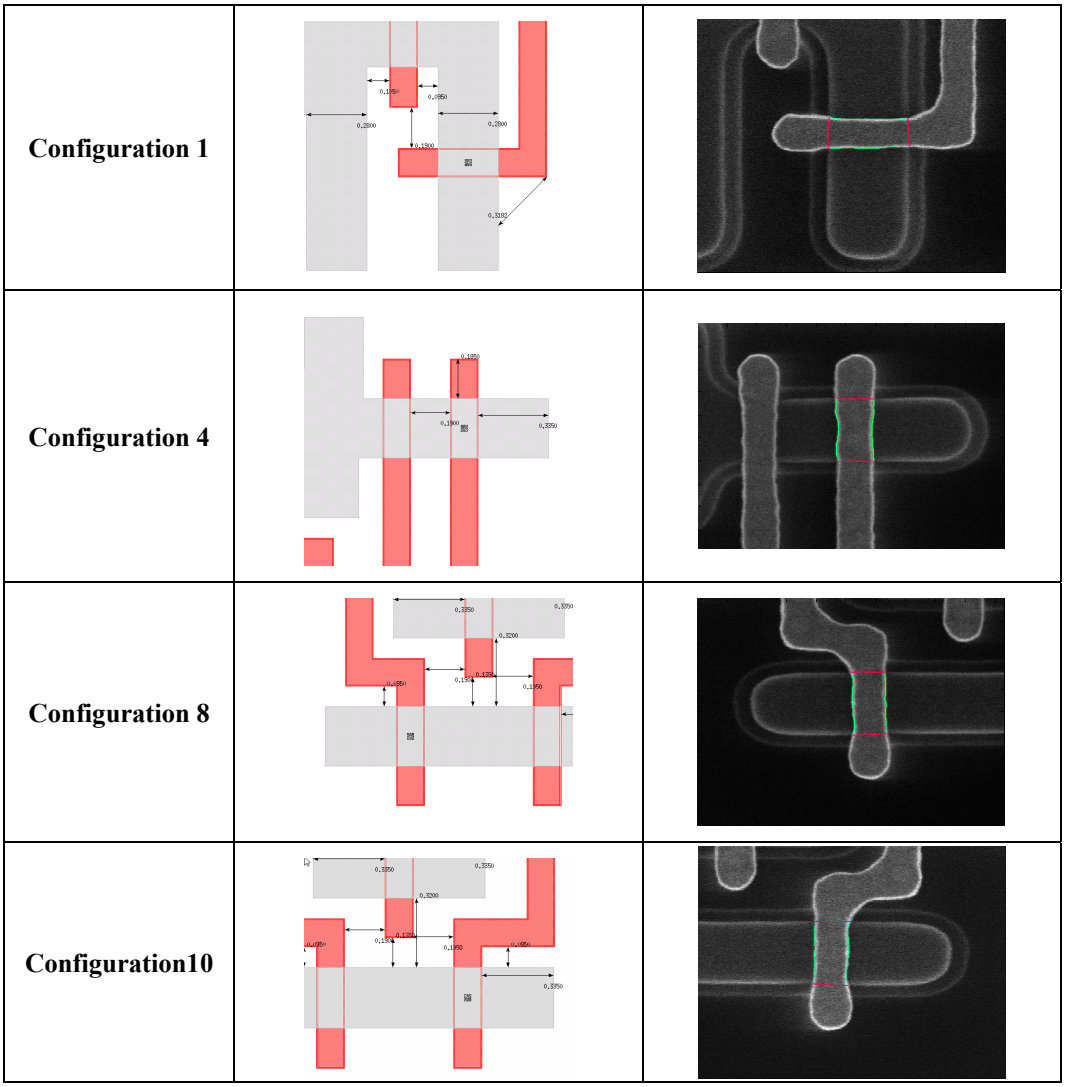
The transistor leakage also depends on the complex AA/Poly configuration. For analysis, a study methodology was developed [13,14] consisting of systematic Edge-Contour-Extraction (ECE) from transistors, taken along the manufacturing line. In general, the SEM (Scanning Electron Microscopy) ECE algorithm is based on CAD (GDS) to SEM pattern recognition, followed by initial and final 2D edge extraction. About ~3000 transistors were measured for the analysis. Device modeling (based on SPICE simulation) was then used, to predict the nominal values as well as the device performance variability of I_{on} and I_{sub}. The SEM analysis was done with measurement steps of 2 nm, so for every transistor gate, the min/max, average and standard deviation of the width and length were measured and calculated. I_{on} was calculated based on W_{avg} and L_{avg}—average width and length of every transistor, respectively. I_{sub} was calculated based on L_{min}, σ_L and W_{avg} were L_{min} and σ_L are the minimum gate length and the related standard deviation of every transistor. More details on this calculation method are given in [15]. The I_{on}/I_{sub} characteristic was used, in order to compare the performance of different transistor configurations. Generally, shorter gate length resulted in higher drive current and higher leakage current. The I_{on}/I_{sub} chart (Figure 7), gives the possibility to characterize configurations that yield lower leakage current for the same drive current. This is the main advantage of using the I_{on}/I_{sub} chart instead of looking on I_{on} or I_{sub} separately for variability analysis.

Figure 7. Transistors leakage current (I_{sub}) *vs.* drive current (I_{on}). Data for site #2 (only) and for 4 different configuration types [14].



Analysis of the different clusters at the I_{on}/I_{sub} chart using Calibre DFM (Design-for-Manufacturing) property (Mentor Graphic) showed that each cluster is related to a different transistor configuration. The most frequent configurations (Figure 8) were configuration 4 (37%), configuration 1 (32%), configuration 8 (11%) and configuration 10 (8%). In the other 12% of transistors, 18 different configurations were defined.

Figure 8. Layout view and SEM micrograph of the most popular transistors configurations analyzed for I_{on}/I_{sub} ratio. The transistor marked with ■ in the center is under detection. Note that configurations 8 and 10 are similar but with different transistors under detections [14].



Configuration 1 consists of a U-shape AA, with isolated poly gate. This configuration is known to have higher AA width variability [15,16]. The transistor under detection at configuration 4 does not have any AA or poly corners close the gate area, and can be referred to as "semi-dense" poly. Configuration 8 and 10, are very similar: both have poly bent at minimum design rule distance to the gate area, and the poly gate can be referred to as isolated. The only difference between these two configurations is the local area that is very similar but not identical. The electrical performances of each one of the configurations are shown in Figure 9. Configuration 4 shows the lowest I_{sub}, with about 20% lower leakage compared to configuration 8 or 10. This "better performances" can be attributed to the lack of AA and poly corners near the transistor gate, as well as to the semi-dense poly line. On the other hand, configurations 8 and 10 show the "worse performances" due to the isolated poly line, L_{min} was narrow and yield high leakage current. In addition, the poly bent and the related OPC, may affect the transistor width (as well as the transistor minimum length), as proposed at the SEM micrograph

(Figure 8). It is clear from the I_{on}/I_{sub} chart, that these two configurations, had the highest (the worse) ratio and therefore, are less recommended to be used for low power or low leakage applications. Configuration 1, also shows bad I_{on}/I_{sub} ratio, correlated with the AA width spread as well as the isolated poly gate as can be clearly seen at the SEM micrograph (Figure 8).

Figure 9. I_{sub} for the 4 different transistors configurations. Data is from central site only [14].

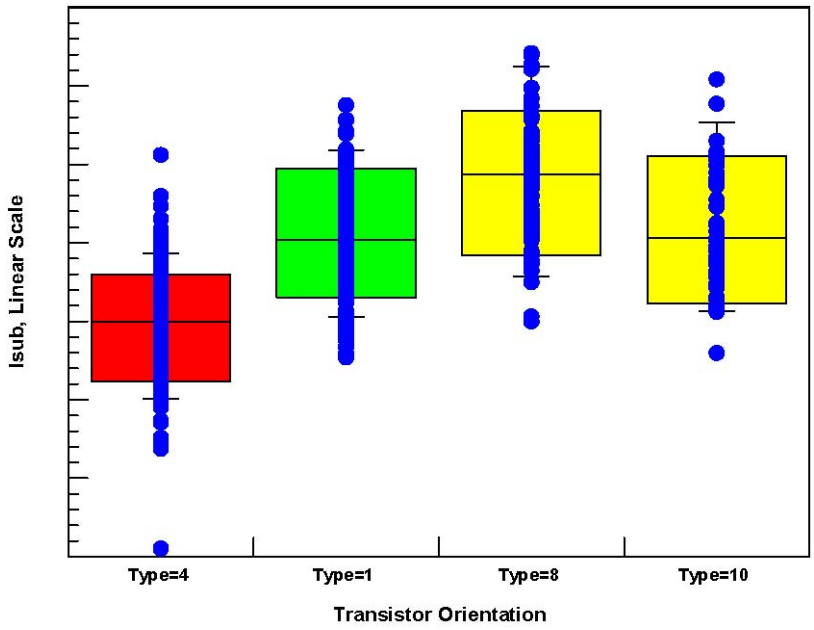
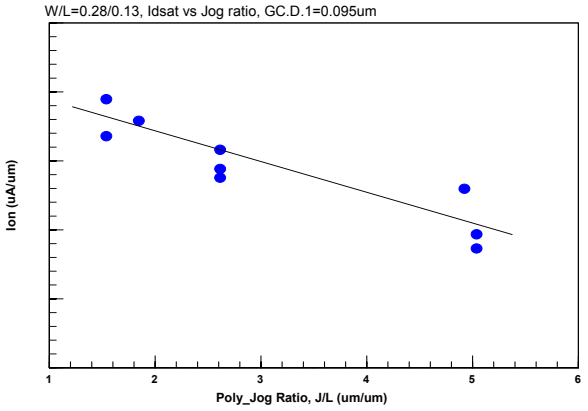
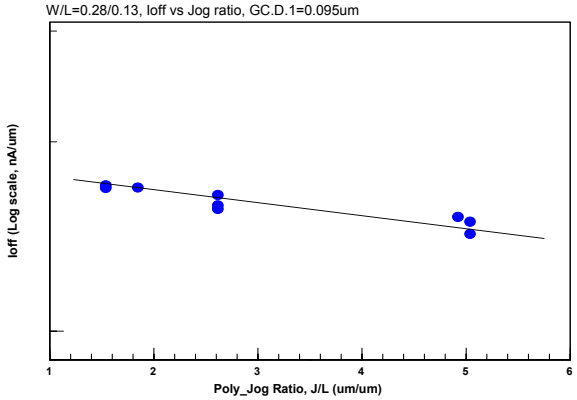
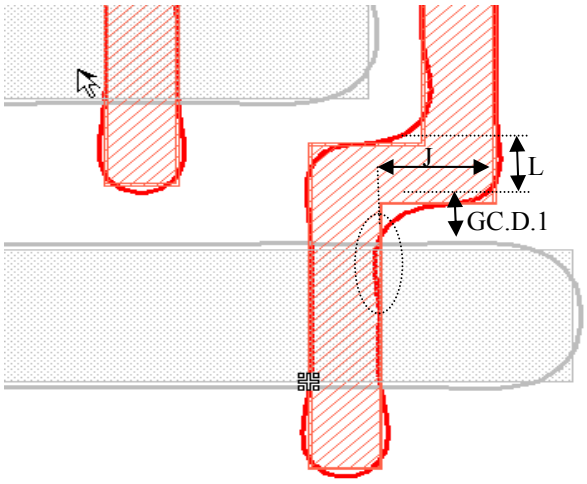


Figure 10. Dependence of drive current (up) and leakage current (below) on poly jog ratio. The layout (right) shows the poly jog ratio definition [13].



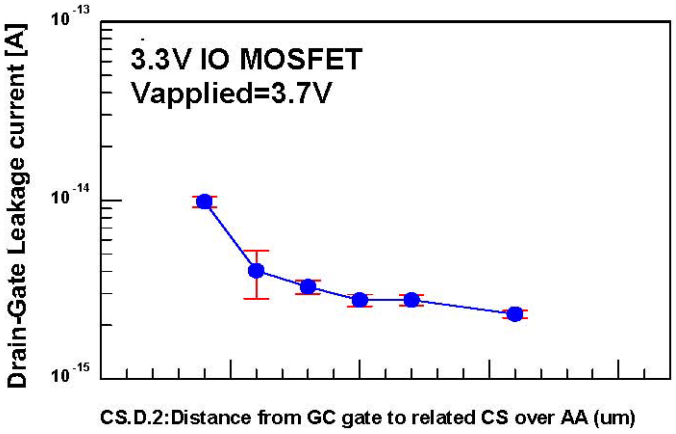




Configurations 8 and 10 were also studied in [14]. A large array of standard cells was OPC treated followed by "silicon simulation", to simulate the optical and etch manufacturing conditions. After the physical parameters were extracted from the "on-silicon" structures, device simulation was performed. Some correlation was found between the I_{on} and I_{sub} values and the ratio of J/L (Figure 10) where J is the length of the parallel poly to the AA and L is the length of the poly line. The parameters L and J determine how close the poly corners are to each other. Close corners will cause the OPC corrections to interfere with each other, causing channel length profile to undershoot in the jog side of the channel. This poly line-width undershoot increases I_{sub} because it is a function of L_{min}.

CS.D.1/2: Contacts too close to transistor gate, may have higher electrical field between the gate and the drain and as a result, may lead to higher I_{sub} (Figure 11). This is the reason that at some cases, CS.D.2 for thick oxide MOSFETs that use V_{dd} of 3.3 V and have larger electrical field between the contact and the gate, use larger distances compared to thin oxides (CS.D.1). In addition, contacts too close to the gate increase the overall gate capacitance, and degrade the transistor switching speed. Process improvement for this rule for leakage reduction result mostly from contact etches profile optimization and some selective OPC. A special test chip was proposed for monitoring CS.D.1 leakage levels [17].

Figure 11. The effect of the distance of CS to gate for 3.3V IO MOSFET.



Comparison among several vendors of standard cells using "ranking methodology" was presented in [18]. The ranking rule was based on fab manufacturing information data regarding the physical and electrical sensitivity of the structure to the design rule type and value. The overall design score was calculated using the ranking rule and its "weight". Table 3 below (taken from [18]), shows the results for 4 different Std Cells libraries from 3 different vendors. Vendor C yielded the highest score for both GC.D.1 and CS.D.1. Vendor B2 received the lowest score for these two parameters.

Table 3. Design score for GC.D.1 and CS.D.1 design rules for 4 different IP blocks, of 3 different vendors [18].

Rule	Rule Description	Rule Weight	A	B1	B2	C
RGC.D.1	Distance between Poly (over STI) to AA Edge	8	86.10%	82.80%	88.80%	98.40%
RCS.D.1	Distance between CS to Poly gate	7	83.70%	84.20%	84.20%	91.50%

High LER (Line-Edge-Roughness) and higher LWR (Line-Width-Roughness), also degrade the transistor leakage current. If we assume that the poly is composed of N segments in series, having a length l_i, so the overall I_{sub} of the transistor at V_{gs} close to 0 V will be [19]:

$$I_{sub} \propto \sum\nolimits_N \exp\left(\exp\left(-l_i / l\right)\right) \tag{3}$$

where l is constant. The first term at the summation will be L_{min} (the minimum gate length at the specific transistor). This segment will have much higher leakage than the other terms because of the exponential dependence. This can also explain the decision to use $(L_{min} + \sigma_L)$ for the I_{sub} calculation [15]. The natural conclusion from this is that higher transistor variability means also higher power consumption. Kim *et al.* [20], found that for poly gates having widths of 80 nm~90 nm, increasing *LWR* from <7.1 nm to 14~21 nm, increased I_{sub} by 1.5~2 orders of magnitude.

LER is a strong function of the image conditions. At poly layer photolithography, the poly is "dark" and the poly space is "clear" or "bright". In order to improve image fidelity and reduce variability, the transition from bright-to-dark needs to be steep. For reducing LWR, it is recommended to have a fixed (and optimal) space between poly the gates: to the near transistor or to dummy transistor. The size of this optimal space is set by the image conditions used by the technology—the wavelength, the numerical aperture, the illumination conditions as well as the photo-resist conditions like thickness and viscosity. Standard cells libraries used the minimum poly width of the technology for almost all gates. However, the position of the different transistors over the AA can not be fixed due to contact located in between for some cases. In addition, if the library supports multi-V_t, so the distance between different types of transistors should be maintained. This "fix space" is the base for using regular and gridded design with restrictive design rules (RDR) that introduced in 45 nm and below platforms.

Ban and Pan [21] proposed an algorithm for LER-aware poly optimization in order to minimize the leakage related LER, by setting an optimal space. The procedure placed poly gates at the best locations and introduced dummy poly to eliminate boundary conditions. As an example, 6 cells simulated with 32 nm technology conditions, showed leakage reduction by up to 47% (average 40%). In another work [22], Ban *at al*., presented a layout optimization based on comprehensive sensitivity metric which seamlessly incorporate proximity effects and process variations. Based on that information, standard cell layout optimization (poly gate and AA layout adjustments) is taking place, to minimize the delay at nominal and corner conditions. Using 45 nm Std Cell library, they demonstrated a leakage reduction of 7~91% at that corner.

3.2. Leakage Reduction in Transistor Level—CMOS and SRAM

Threshold voltage reduction is the simplest way to overdrive the transistors, and reduce propagation delay. However, V_t reduction means an exponential increase of I_{sub}. (Figure 12). By using a very high V_t values for non-critical paths, the leakage can be reduced by 2~3 orders of magnitude. Figure 13 shows the V_t scaling from 0.25 μm platform down to 65 nm for Standard (or regular) V_t. As can be seen, V_t values were no longer reduced beyond 90 nm. For Low Power, V_t higher by 100 mV~200 mV was used.

Figure 12. Typical values of I_{sub} *vs.* threshold voltage for nMOS, 65 nm for GP and for LP, based on data from [23], of MOTOROLA, STMicroelectronics, Philips and CEA-LETI and data from TOSHIBA [24]. For pMOS devices (not seen), V_t values were adjusted for having the same leakage levels.

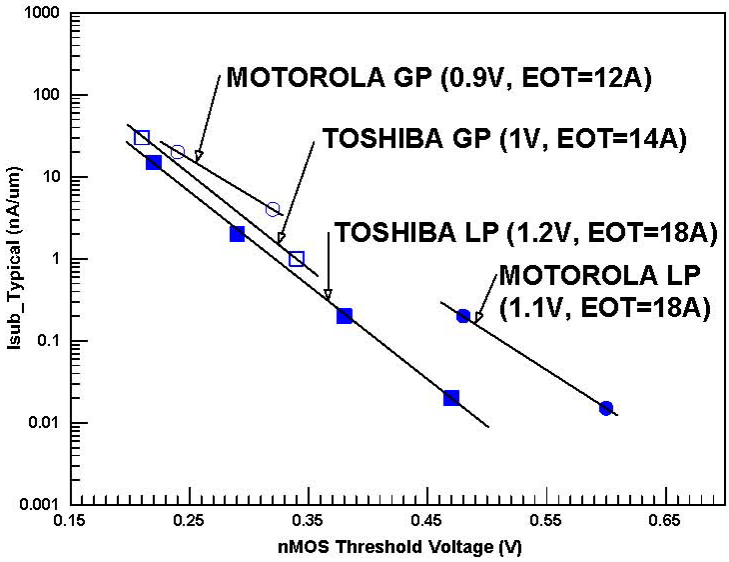
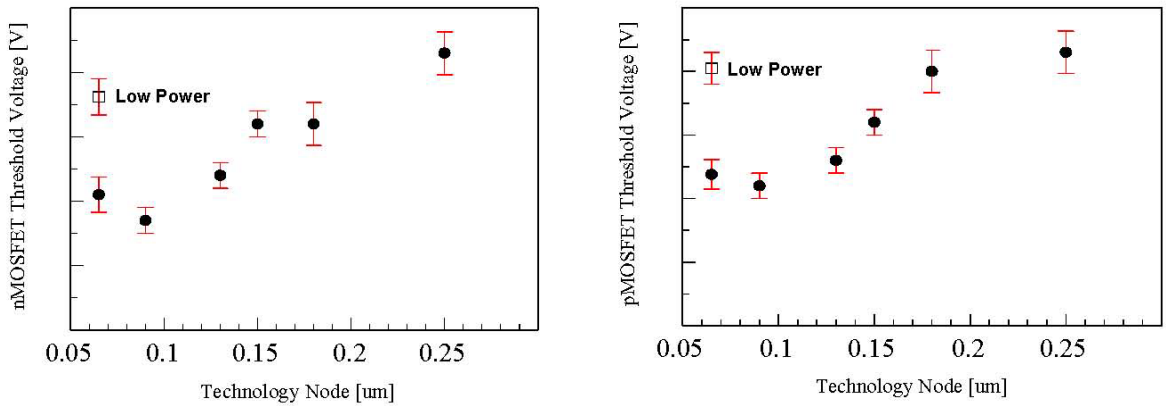


Figure 13. Typical values of thin oxide threshold voltage standard V_t (SVt) transistors, nMOS (left) and pMOS (right), for Standard Logic for General Purposes and for LP (Low Power) applications.



Basically, V_t change can be done by doping adjustments (of the channel and/or the SDE—Source-Drain-Extensions), adjustment of the gate oxide effective thickness and/or the work-function difference of the gate electrode (in the case of metal-gates) or by body biasing. In multiple-V_t, also known as "Multiple Threshold Voltage CMOS" (MTCMOS, or dual-V_t CMOS or DVTCMOS), two (or more) types of transistors are fabricated: Standard (SVt) or Regular-V_t and high-V_t (HVt). In most cases, this is done by an additional two V_t implant masks or two SDE implant masks. In high-density standard cells, this technique can be limited for "mixing" standard cell libraries, due to the layout design-rules related to the other layers. For SRAMs, the mask data preparation done by the foundry assign the relevant HVt implant masks also to the SRAM array. In case the design is without HVt, the dedicated VNS (Vt implant for nMOS SRAM) mask described above (or another VPS mask) are used. Another way to adjust the V_t is by the V_t roll-off behavior. However, modern CMOS devices use high

doping levels of halo implants, in order to reduce the V_t roll-offs. In addition, in case a larger L is used, the gate capacitance will also be increased.

"Mixed" technology refers to the case of having simultaneously the Standard Logic transistors and the Low-Power transistors, having a different gate oxide thickness. In the example of the 65 nm platform described in Table 1, the two technologies can be used separately or "mixed" together [1,2]. These combinations have a triple-gate oxide and it is not manufacturing friendly due to an additional mask penalty, between +1 mask for the gate oxide process only, and up to +8 masks for the case V_t and SDE implants also need to be separated. In addition, the complexity of having an oxide-strip at a very small window, the additional thermal budget, and the fact that there are two different gates with a close thickness target are problematic due to the oxidation kinetics [25]. However, it was successfully developed for 28 nm technology, having gate oxides thickness of 16A for Low Power Standby (1.1 V) and 13.5 A for Low Power (0.8 V) [26]. In summary, this combination is one of the ways to reduce the overall circuit leakage, but it introduces many process challenges and has a high cost penalty.

It is known that the back bias (body bias, or reverse body biasing - RBB) can modify the transistor V_t. However, higher body bias increases GIDL, degrades V_t variability, and in multi-V_t transistors induces different body-biasing sensitivity that depends on V_t [1]. In this case, closely located transistors can not share the same N or P wells and because of that, triple wells are needed. Such wells have high area penalty due to additional layout design rules. It is important to note that, while the reverse bias increases V_t, it also increases the junction current and decreases the junction capacitance. In [27], a novel technique to minimize the standby leakage was proposed. In order to overcome the performance's degradation using RBB due to increase in GIDL, DIBL and BTBT currents, H-J Jeon *at al.* [27], proposed a standby leakage power reduction technique, based on optimal body bias voltage. This voltage was determined by the ratio of I_{sub} and the band-to-band tunneling current (I_{BTBT}). For circuit implementation, they proposed a control system that includes monitoring circuit, current comparator and charge pump. The leakage monitoring circuit input both I_{sub} and I_{BTBT} into the current comparator that increase or decrease the body voltage applied to the chip core by the charge pump. Implementation of this technique to 32 nm MOSFET technology ISCAS85 benchmark circuits yield 400~1500× leakage reduction. Yasuda *et al*. [28] succeeded in reducing the sensitivity of the body-biasing to threshold voltage by careful channel and gate engineering—they increased the channel contour doping (by adjustment of the punch through implant dose and energy) and shifted the channel from the surface (buried channel). Taking advantage of the V_t shift by the work-function modulation of the Hf-based gate dielectric, the peak concentration of the channel impurity profile was positioned in a deeper channel region, away from the surface, and without lowering the V_t.

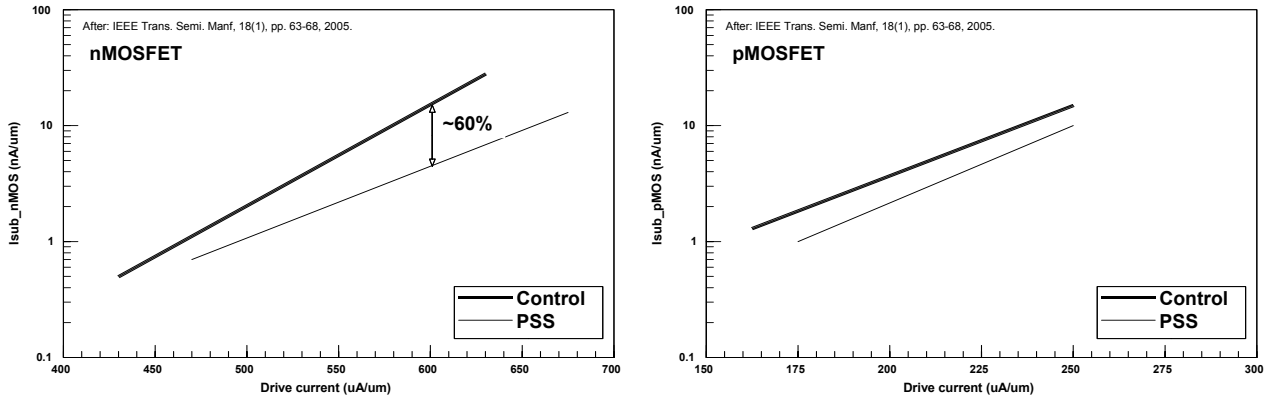
The main drawback of reducing the leakage by increasing the channel doping for V_t, is the reduction in the transistors currents due to a lower overdrive, which leads to degradation in delay time, that for an inverter is given by:

$$\tau_d \propto \frac{C_L V_{dd}}{\mu \cdot (\varepsilon_{ox}/T_{ox}) \cdot (W_{eff}/L_{eff}) \cdot (V_{dd} - V_t)^v} \tag{4}$$

where v is a fitting constant (that is correlated to the velocity saturation index), μ, ε_{ox}, W_{eff} and L_{eff} are the channel mobility, the gate oxide dielectric constant, the effective width and length of the transistors, respectively.

The 90 nm technology was the first node in which performance enhancement was done using stressors [29]. These stressors can be STI [30], Stress Memorization layers [31], nitride located under D1 that was also used as Contact Etch-Stop Layer (cSEL) [32] and eSiGe (elevated SiGe) [33]. Stress induced by the salicided Source-Drain active area can also improve performance [34]. Basically, stress induced into the channel can improve or degraded the carrier's mobility, and as a result change the transistors currents. The level of improvement (or degradation) depends on the level of the stress induced, the type (tensile or compressive) as well as on the direction of the strain induced into the silicon. For example, compressive stress induced by STI along the x-axis (along the channel length), will improve the drive current for pMOS transistors. However, the compressive stress at the y-axis (along the channel width) will degrade the drive current for the same pMOS transistor, and because of that, higher tensile stress resulting from AA salicidation at the y-axis will improve the current. In the work reported in [35], all stress techniques listed above were used in a 90 nm platform. The overall currents improvement was up to 15%. It consisted of improvement due to cSEL (~7%), from STI (~7%) and from salicidation (~5%). Naturally, the improvement of all stress components is not "cumulative".

Figure 14. I_{sub}/I_{on} curves for nMOS (left) and pMOS (right) transistors, with and without stresses induced. PSS is "Process Strained Si" that includes: cESL, STI and silicided layer. Data is from [35].

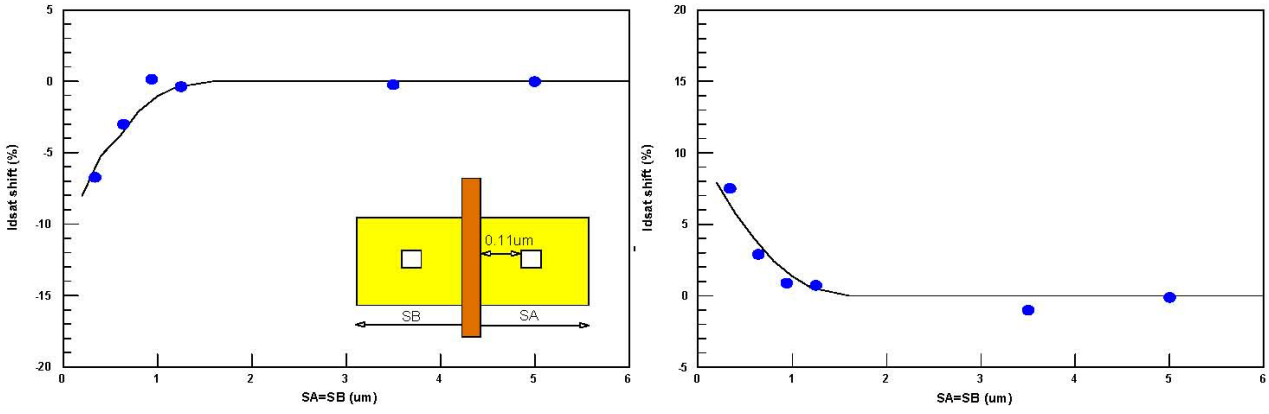


The main advantage of the different mobility enhancement techniques is the increase in drive current without leakage degradation (Figure 14). Based on that, by a careful layout modification, the leakage current can be improved by keeping other parameters in place. As a basic example, assume a transistor with a specific gate length that yields drive current and leakage based on the I_{on}/I_{sub} charts. Increase of the gate length, will reduce the leakage and the drive current. However, by using stressors, we can re-set the drive current back to place, while still having this low leakage levels. In the example of Figure 14, the leakage for the nMOS can be reduced by ~60% while keeping the same drive current. In addition to this example, the I_{sub} reduction is observed to taper off quickly with longer gate length. It is important to note, that stress can also *increase* the leakages related to junctions. Wang *et al*. [36] studied the effect of mechanical uni-axial stress on junctions fabricated in 65 nm technology. They found, that for junction in nMOS, where the BTBT is the major component of the junction leakage, the dependence on stress (generated by tensile cSEL) is weak, and even using high-stress layers (thicker,

as for 45 nm technology), the junction leakage degradation was <7%. For nMOS, higher tensile stress reduced the leakage. However, for junction in the pMOS, where the stress was generated by both compressive cSEL and eSiGe elevated S/D, due to the fact that the leakage mechanism was based on both BTBT and generation current, high stress would degrade the leakage by up to 25%. These facts should be taken into consideration for future technologies. Examples for stress affecting layout modification are:

- Expending the AA (Source and Drain) edges beyond gate (GC.X.1), for different STI induce stress (Figure 15). The stress range and magnitude are up to GC.X.1 = 1.3 μm, and <10%, respectively;
- Re-placement of contacts with distance to gate (CS.D.1). This is because contacts "punching" of the cSEL layer, and release some of the stress. For this reason, also re-set of the Source and Drain contact pitch may improve performance [37];
- Poly space between transistor fingers (without contacts) [38]. This is because smaller space also means narrow cSEL layer, or narrow eSiGe stressor trench, or both. In [3], up to 7.8% degradation was seen for different poly spaces;
- Location of tensile/compressive nitride cSEL boundary layer over STI (Figure 16) and separating nMOSFET and pMOSFET [38].

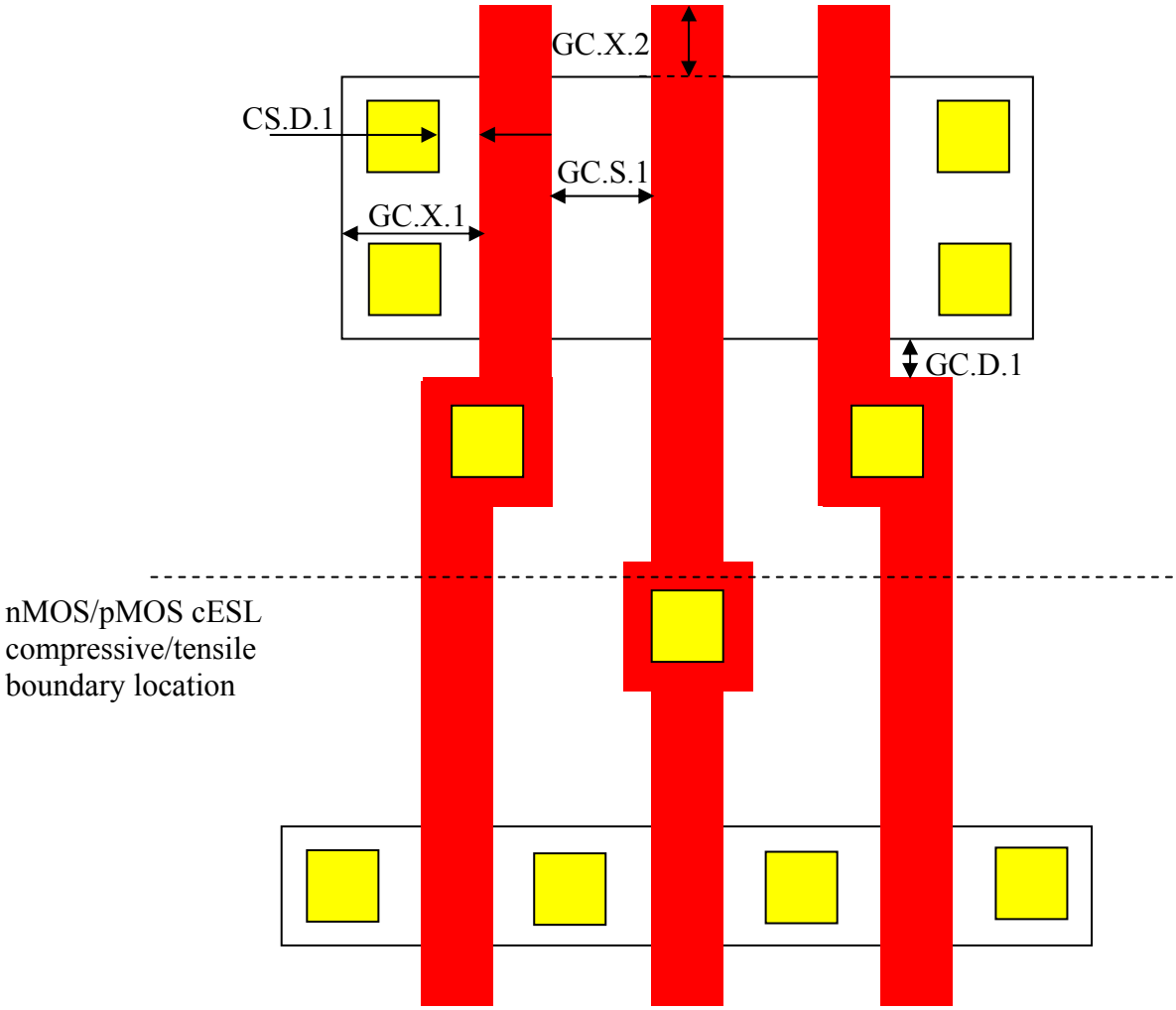
Figure 15. The effect of AA extension beyond gate (GC.X.1) LOD stress and on drive current for (left) nMOSFET, (right) pMOSFET. For both devices, W = 1.2 μm, L = 0.13 μm, Single finger, and SA = SB, where SA and SB are the extensions from left and right of the transistor, respectively.



The first research work to tackle timing closure for standard cell by layout modifications using active area depended mobility of strained silicon was made by [38]. In there work, GC.S.1 was adjusted, to modify the stress induced by eSiGe stressor. Later, Joshi *et al.* [39], developed a methodology for stress-aware layout optimization, with a constraint that the cell area will not change, have similar switching delays or less, and lower leakage. Because dual-V_t (HVt, LVt) was available, the algorithm also "assigned" the optimal V_t type per case, together with the layout optimization. This approach was used successfully for the 65 nm technology design having the following parameters: V_{dd} = 1 V, nMOS_HVt = 334 mV, pMOS_HVt = −391 mV, nMOS_LVt = 243 mV, pMOS_LVt = −280 mV. The I_{on} and I_{sub} ratio for LVt/HVt was ×1.24/×16 for the nMOS and

×1.32/×29 for the pMOS. The stress-aware layout modification included changes in GC.X.1, CS.D.1 and CS pitch, and setting of the location of tensile/compressive nitride cSEL layer located over STI (Figure 17). Comparison was also made between using dual-V_t with single thin-oxide thickness only, and using dual-V_t with stress-aware layout modification. Analysis showed that for the same delay time, up to 34% reduction of leakage was obtained. For the same leakage values, up 10% delay time reduction was achieved using this methodology.

Figure 16. Typical 3 input NOR gate with some of different Topological Design Rules.



The results of Table 4 [39] clearly show, that the combined approach improved significantly the leakage power while keeping the same delay time. Improvement in critical delay time for iso-leakage was also seen while comparing to dual-V_t only approach. Maximum leakage improvement was 38.5% and with average value of 23.8%.

Figure 17. Leakage power versus delay tradeoff curve for the circuit c7552, which includes 1993 gates, for dual-V_t with and without stress-aware layout modification [39].

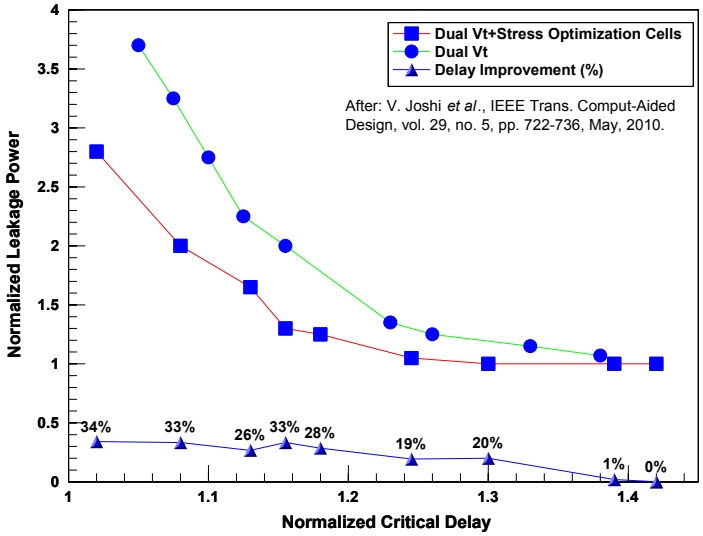


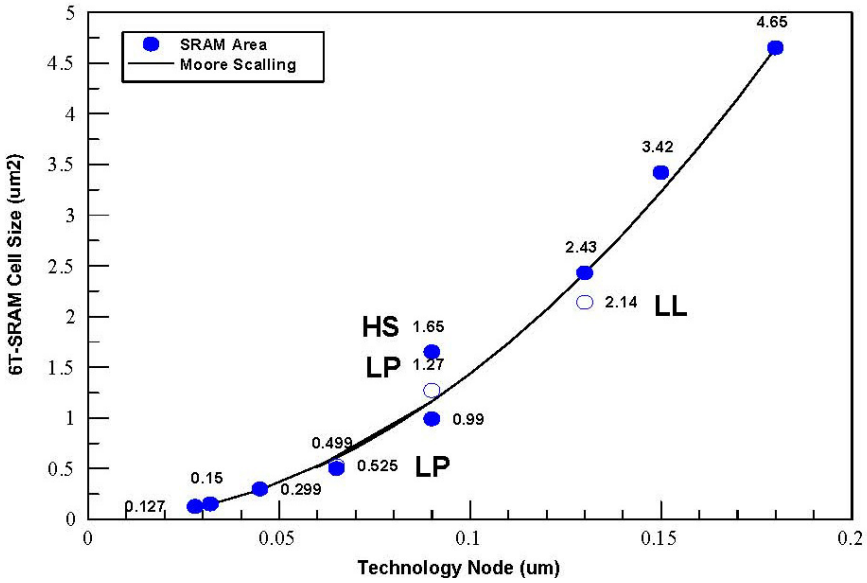
Table 4. Improvement in leakage and delay, comparing Dual-V_t (HVt/LVt) approach to Dual-V_t with stress-aware layout optimization, based on data from [39]. 12 different circuits were used, having number of gates from 166 and up to 37,560.

12 Circuit	Number of Gates	Comparison for Iso-Delay Against Only Dual-V_t Assignment				Comparison for Iso-Leakage Against Only Dual-V_t Assignment			
		Stress + V_t based assignment		Only stress based assignment		Stress + V_t based assignment		Only stress based assignment	
		Improvement in Leakage	Area Overhead	Improvement in Leakage	Area Overhead	Improvement in Delay	Area Overhead	Improvement in Delay	Area Overhead
Minimum		14.70%	0.10%	4.70%	0.20%	4.60%	0.20%	2.30%	0.20%
Median		22.45%	0.30%	5.10%	0.35%	5.10%	0.30%	2.95%	0.35%
Maximum		38.50%	0.90%	12.00%	0.90%	5.80%	0.90%	3.60%	1.00%

4. Low Power Consideration for SRAM

Technology scaling decreased the overall SRAM area by factor of ×2 (or more) for each generation (Figure 18). The 0.13 µm platform was the first in which two bit-cells were used by foundries for high volume manufacturing: 2.43 µm^2, that is a direct shrink from 0.18 µm, and 2.14 µm^2, for high-density low-leakage application. Down to 80 nm, a 6-T (six transistors) SRAM Bit cell of type A to D was used [40]. The 65 nm foundry technology [41], introduced a new layout configuration, that did not have any AA or Poly corners that could be rounded as explained above. This "thin" vertical height also reduced the bit-line loads and improved noise immunity. For 45 nm or 32 nm technologies, the straight poly lines could also be supported with line-cut double-patterning [42].

Figure 18. 6-T SRAM Bit-Cell area trend, used by pure-player foundries. The data refers to SRAM used in Standard Logic for General Purpose technology, unless indicated differently: HS = High-Speed, LP = Low power and LL = Low Leakage.



The total leakage in SRAM is roughly expressed as [1]:

$$2 \cdot I_{sub\_latch} + I_{sub\_PU} + I_{gate\_latch} + I_{gate\_PU} \tag{5}$$

were I_{sub_latch} and I_{sub_PU} are the subthershold current for the nMOS latch and the pMOS Pull-Up transistor, respectively. I_{gate_latch} and I_{gate_PU} are the gate currents for nMOS latch and pMOS Pull-Up, respectively.

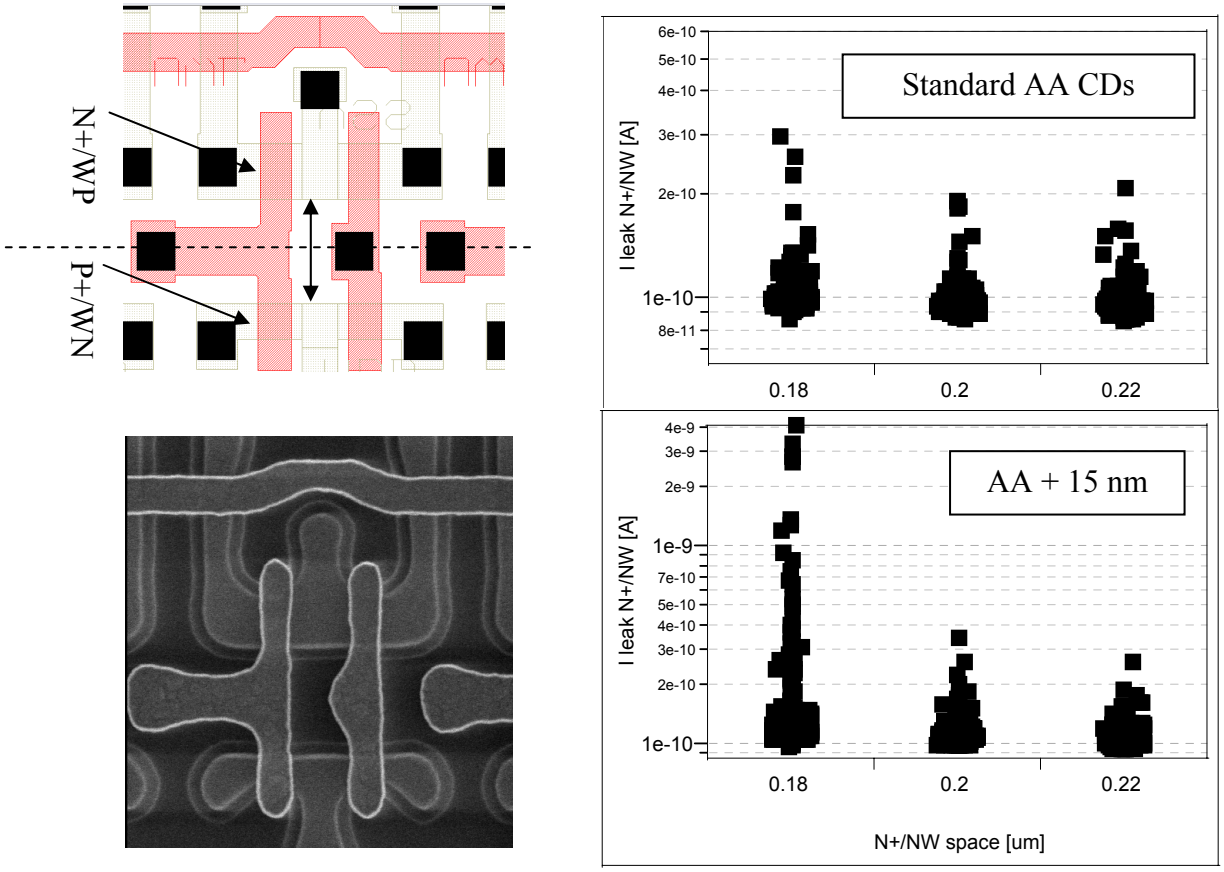
In order to reduce off-state leakage, in many cases the SRAM array has higher V_t. This is most important for the nMOS pull-down (PD) and costs an additional dedicated VNS (dedicated V_t implant for the SRAM nMOS) mask. Note, that this mask also increases the nMOS Path-Gate (PG) threshold voltage. The penalty is that both the write delay and the read delay increase. In some cases, another additional mask is used in order to increase the V_t of the pMOS Pull-Up (PU) transistors results in reduction in V_{dd}-to-ground leakage, but with a penalty of write delay [43]. The higher V_t results also in improved static noise margin (SNM) in the cell, which allows reduced β ratio (or cell ratio), that is defined as β = (width/length of nMOS PD)/(width/length of nMOS WL). The reduction of β improves the cell read current [9].

A major contributor of leakage for SRAMs is the gate-to-channel leakage of the PD nMOS transistors in the "ON" state. An increase of the gate oxide thickness can reduce this leakage (that has an exponential behavior, see Equation (2) above). However, the gate thickness is set by the logic transistors (both nMOS and pMOS). Solutions like "multiple" gate oxide thicknesses ([43], called MoxCMOS in [9], dual-T_{ox} CMOS or DTOCMOS in [44]) were also proposed. For advanced technologies, which use high-k gate oxide materials, reduced gate leakages for the same effective gate oxide thickness are achieved (See Figure 3 and Table 1). Yasuda *et al*. [28] reported that by replacement of the SiON gate material for HfSiON, where both have the same effective oxide thickness, the gate leakage components in (3) become negligible, the total stand-by power consumption is reduced by a factor of 5. In addition, Yang *et al*. [45] reported, that for 32 nm Low

Power technology, the adoption of a gate-first Hf-based high-*k* process, improved V_t mismatch by 50% (comparing to 45 nm technology), due to thicker gate oxide that provided better channel control. V_t mismatch improvement reduced SRAM soft fail rate.

One of the SRAM scaling parameters refers to space reduction between the nMOS AA to the pMOS AA, and it affects all A-D types of 6-T "tall" SRAMs [40]. This space is composed of: AA.D.3+AA.E.3, where AA.D.3 is the distance between WN to N+ in WP and AA.E.3 is the enclosure of WN around P+ in WN (see Figure 19). Based on Table 2, the values for these two rules are scaled down by a factor of ~0.7. However, the limiting factor for nMOS-pMOS AA space reduction is the punchthrough. Figure 19 shows a typical layout and SEM top-view micrograph of 6-T SRAM type D. Assuming that the distance is 2 × 0.22 = 0.44 μm, leakage measurements for the standard photolithography conditions show an increase of the leakage value when this distance is reduced by 2 × 0.04 μm. For stability testing, a process window having larger AA by 0.015 μm, reduced the minimum space between diffusions to 2 × (0.18 − 0.015) = 0.33 μm. As seen from Figure 19, the leakage goes up. For more scalability, both N-Well and P-Well tub profiles as well as the STI depth and slope need to be optimized.

Figure 19. 6-T SRAM (2.14 μm^2, type-D) Bit-Cell layout and SEM Top-View taken after Poly etch (left), N+/PW to P+/WN leakage as function of space, at two process conditions (right).

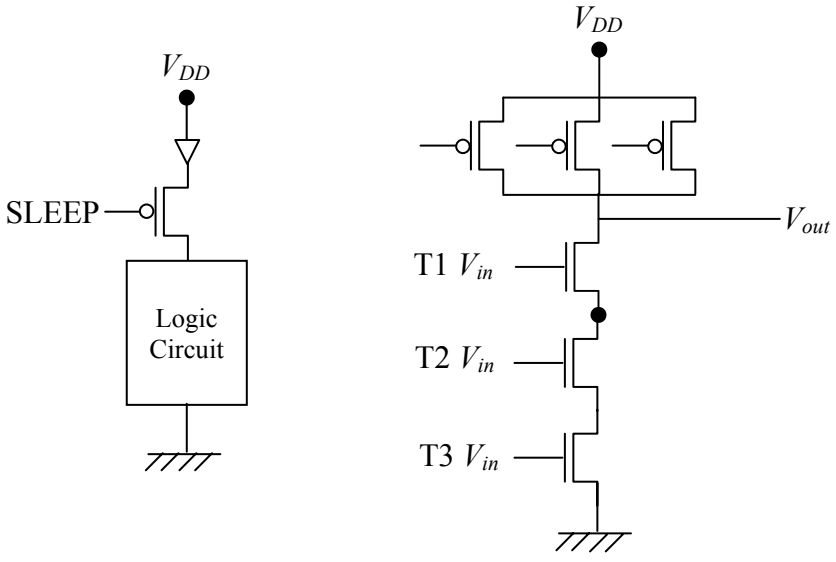


The area reduction of SRAM requires more aggressive design rules than those allowed by the platform design kit. For AA, poly and contacts, this "violations" are mostly related to enclosure of AA and poly around contacts, CS.D.1, Poly-end-caps (GC.X.2) and as explained above the—distance N+/PW to P+/WN. All these "violations", demand a careful OPC treatment. In most cases, foundries use a dedicated OPC treatment for the SRAM array. For an SRAM bit-cells having area of 2.14 μm^2 and used in 0.13 μm platform LL (low leakage) technology, an overall typical cell current of 5 pA/cell (Max < 10 pA/cell) was achieved using a dedicated OPC.

5. Circuit Level Techniques for Power and Leakage Reduction

This paper focuses on device and process level power reduction techniques and therefore, circuit level solutions will be covered mostly from design-rules point of view. Power reduction techniques at the circuit level are listed in [46]. For mobile applications, where the product is in a standby mode most of the time, the most effective way is to cut the leakage by switching off the inactive circuits. The basic method is to insert a power switch in series between a digital circuit block and its supply line (Figure 20). When entering the sleep mode, the gate of this power supply switch transistor is raised above V_{dd}, to decrease I_{sub}, which depends exponentially on V_{gs} (gate-to-source voltage). The drawback of this V_{dd} "boosting" is that I_{gate} also increase exponentially (2), and the gate oxide may wear out. Naturally, an "optimal" V_{dd} should be applied. In [47], a circuit that automatically biases the power switch gate transistor to its minimal leakage point, and efficiently compensated for temperature and corner variations was presented. If dual-V_t or triple gate oxides are used, the power switch transistor will have the HVt and thick oxide. Another way is to use reverse biasing (with SVt or LVt), to obtain lower leakage and reasonable performance, as explained above.

Figure 20. Schematic of the power switch transistor used to cut the supply into the logic circuit in a sleep mode (left). Stack and sub-stack of a NAND3 (right). It is recommended that the same transistor type should be used (pMOS at this case) in the parallel structure.



In this section, a short review of circuit solution for digital design will be presented. After, we will focus on design optimization to reduce leakage of large SRAM array.

For CMOS: I_{sub} flowing through a stack of series-connected transistors is reduced when more than one transistor in the stack is turned off. For example, the leakage of a two-transistor stack is one order of magnitude less the leakage of a single transistor. This effect is known as the stacking effect [9] or self-service biasing [48]. Leakage reduction takes place because the voltage level of the intermediate node (between the two transistors) is positive. This leads to a negative V_{gs} and to a negative V_{bs} (body-to-source potential) and also to reduction in V_{ds} (drain-to-source voltage). All these yield lower I_{sub}. For example, in 3 input NAND gates in stack, that were simulated using 65 nm technology with 17 Å gate oxide thickness, turning-off 1 transistor reduced I_{sub} by 23% (by 7% for 2 turned-off and ~4% for all 3 turned-off) [49].

Sill *et al.* [44], performed a simulation analysis for selecting the best transistor type, using both dual-V_t and dual-T_{ox} (DVTCMOS and DTOCMOS). From the results, they extracted two design rules for transistors stacks:

- The Delay rule—within mixed stack, the L-V_t Transistor (with low V_t doping and thin oxide), has to be placed as close as possible to the gate output to achieve best results for the time delay;
- The leakage rule—within mixed stack, the H-V_t Transistor (having high V_t doping and thick oxide), has to be placed at the end of the stack (away from the output) to achieve best leakage result.

Using these recommended rules, a library of ten standard gates in 65 nm technology was created. The example below shows different possible realizations for NAND3 (Figure 20), with the relative leakage and results of performance (Table 5) for the case where all transistors are made with LVt and Low-T_{ox}. Delay improvement of 6% with the same leakage value (compare #3 and #2) was achieved by placing the H-V_t transistor in the center (T2), and allowed the L-V_t transistor to be close to the output (T1), as defined by the gate delay rule. Leakage improvement of 20% with the same gate delay time (compare #4 and #3) was achieved by placing the two H-V_t far from the output (T2, T3), as defined by the leakage rule. More details on static leakage reduction through simultaneous V_t, T_{ox} and transistors' state assignment can be found in [50].

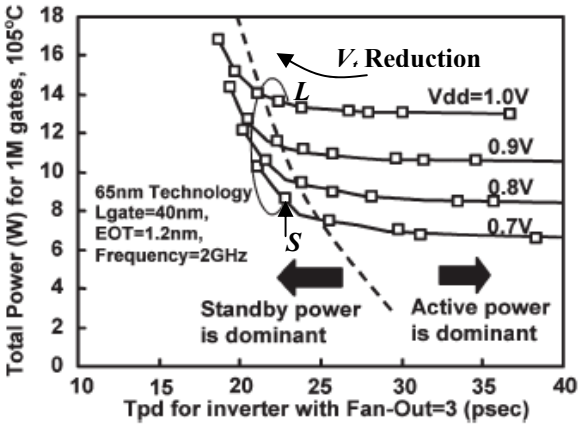
Table 5. Comparison of possible mixed-gates realization, based on data for NAND3 from [44]. "H" means high H-V_t transistor (having high V_t doping and thick oxide), "L" means low L-V_t transistor (having low V_t doping and thin oxide). Refer to Figure 20 for nMOS transistors locations.

Realization Combination #	pMOS Transistors	nMOS Transistors	Relative Leakage	Relative Gate Delay
1	LLL	T1 = L, T2 = L, T3 = L	100%	100
2		T1 = L, T2 = H, T3 = L	50%	118%
3	HHH	T1 = H, T2 = L, T3 = L	50%	124%
4		T1 = L, T2 = H, T3 = H	30%	126%

Supply voltage reduction is also an effective method for switching power reduction due to the quadratically dependence (1). Following the basic scaling rules, the supply voltage should be reduced by a factor κ in order to maintain a constant electric field. However, although V_{dd} reduction yields lower dynamic power consumption it also degrades the circuit performances that cannot be

compensated by V_t reduction. Morifji *et al*. [8], analyzed the dependence of the total power consumed by 1 M gates at 105 °C, on delay time. A 65 nm platform was used, and the gate delay time was calculated by CV/I for inverter with FO = 3. The total power was estimated with clock frequency of 2GHz and switching activity of 20% (1). The implicit variables were the V_{dd} and the V_t. For high-speed demands (Figure 21), V_t should be reduced, and cause the standby power to increase. The dotted line is the boundary where the dominant power changes from being mostly an active power to being mostly a standby power, depending on the operation frequency and the switching activity. Based on that, it is proposed [8], that in SoC (System-on-Chip) composed of different circuits—each circuit may have an optimized V_{dd} (and V_t target) per need. For example, in Logic Core or clocks with 100% duty that seeks for high speed and high activity V_{dd} (and V_t) will scale down aggressively (see point "S" in Figure 21). On the other hand, logic with low frequency or low activity will have higher V_{dd} and higher V_t (see point "L" in Figure 21).

Figure 21. Estimated total power consumption for 1 M gates at 105° as a function of delay of the FO3 inverters, simulated using 65 nm technology [8]. "*S*" is the working point with low V_t and low V_{dd} that provides high speed and *L* is the working point for high V_t and high V_{dd} to minimized leakage.



For SRAM Cell: SRAM cell stability can be observed using its eye (or "butterfly") property where its size is the Static-Noise-Margin (NSM). Basically, SNM degraded for lower V_{dd}, lower V_t or lower β. However, in the case of memory array (where many cells are connected together on a single bit-line), lower V_t will increase the leakage current. When the leakage current becomes comparable to the cell current (that is reduced due to lower V_{dd}), the array will fail. Therefore, both small leakage of the transfer gate and large cell currents are required. This can be partially achieved by longer gate length of the PG transistor and wider width for PD transistor. An optimal $V_t/V4$ combination can be found after setting the β [8]. Dual-power solutions [51] show power reduction by 20%~40% [26].

In [52], two design techniques that reduce the static power dissipation due to I_{gate} and I_{sub} reduction were presented. The first one (titled "PP-SRAM" in [7]) is based on replacing the nMOS path-transistors with pMOS and re-set the transistors widths and V_t levels. This new configuration showed 26% reduction in gate leakage current, 37% in power dissipation and 15% improvement in SN. However, the cell area increased by 16.5%. The second configuration (titled "IWL-VC SRAM" in [7]) is based on improvement of the dynamic voltage scaling method, titled NC-SRAM in [53]. Basically,

this method uses two nMOS path transistors (NC1 and NC2), which provides different ground levels and reduces the gate leakage by ~50% and ~57% power dissipation. In [7], a 3rd pass transistor is added to reduce the gate voltage of the path-gate (Word-Line) transistor yielding another 16% leakage reduction. For both the NC-SRAM and the IWL-VC SRAM since only 2 or 3 transistors are added per row, the area penalty is negligible. More design techniques for SRAM power reduction can be found in [46,54].

6. Summary and Conclusions

The rapid reduction in transistors dimensions results in increase leakages and power dissipation. This demands efforts in several aspects. At the transistor level, the increased leakages have different origins, and therefore reduction requires careful new process integration including novel materials. Another aspect is the leakage dependence on the layout that gives the possibility to reduce leakages by clever layout optimization. Some layout-aware procedures including automated tools for leakage reduction were proposed. Finally, some circuit-based solution linked to layout design rules was described.

The presented analysis revealed correlation between leakages and transistor configurations. Guidelines for leakage reduction based on the use of different stressors, the dependence of leakage on LER, *etc*. were specified. For SRAMs, different circuit level techniques, like multi-V_t, Multi-T_{ox}, body bias adjustment, and power-switching were discussed as possible approaches for leakage reduction.

Acknowledgment

The author wishes to thank Magnet Office in the Israeli Ministry of Industry for providing partial financial support of this work. Many thanks to Yakov Roizin, Amram Eshel and Eng. Israel Rotstein for fruitful discussions, and to O. Menadeva and S. Levi from Applied Materials, Israel for the design aware experiments reviewed in this paper.

References

1. Skotnicki, T.; Fenouillet-Beranger, C.; Gallon, C.; Boeuf, F.; Monfray, S.; Payet, F.; Pouydebasque, A.; Szczap, M.; Farcy, A.; Arnaud, F.; *et al.* Innovative materials, devices, and CMOS technologies for low-power mobile multimedia. *IEEE Trans. Electron Device* **2008**, *55*, 96–130.
2. Tavel, B.; Duriez, B.; Gwoziecki, R.; Basso, M.T.; Julien, C.; Ortolland, C.; Laplanche, Y.; Fox, R.; Sabouret, E.; Detcheverry, C.; *et al.* 65 nm LP/GP Mix Low Cost Platform for Multi-Media Wireless and Consumer Applications. In *Proceedings of the 35th European Solid-State Device Research Conference (ESSDERC 2005)*, Grenoble, France, 12–16 September 2005; Volume 50, pp. 573–578.
3. Miyashita, T.; Ikeda, K.; Kim, Y.S.; Yamamoto, T.; Sambonsugi, Y.; Ochimizu, H.; Sakoda, T.; Okuno, M.; Minakata, H.; Ohta, H.; *et al.* High-Performance and Low-Power Bulk Logic Platform Utilizing FET Specific Multiple-Stressors with Highly Enhanced Strain and Full-Porous Low-k Interconnects for 45-nm CMOS Technology. In *Proceedings of the IEEE International Electron Devices Meeting, (IEDM 2007)*, Washington, DC, USA, 10–12 December 2007; pp. 251–254.

4. Watanabe, R.; Oishi, A.; Sanuki, T.; Kimijima, H.; Okamoto, K.; Fujita, S.; Fukui, H.; Yoshida, K.; Otani, H.; Morifuji, E.; *et al.* A Low Power 40 nm CMOS Technology Featuring Extremely High Density of Logic (2100 kGate/mm^2) and SRAM (0.195 μm^2) for Wide Range of Mobile Applications with Wireless System. In *Proceedings of the IEEE International Electron Devices Meeting, (IEDM 2008)*, San Francisco, CA, USA, 15–17 December 2008; pp. 641–644.
5. Arnaud, F.; Liu, J.; Lee, Y.-M.; Lim, K.-Y.; Kohler, S.; Chen, J.; Moon, B.-K.; Lai, C.-W.; Lipinski, M.; Sang, L.; *et al.* 32 nm General Purpose Bulk CMOS Technology for High Performance Applications at Low Voltage. In *Proceedings of the IEEE International Electron Devices Meeting, (IEDM 2008)*, San Francisco, CA, USA, 15–17 December 2008; pp. 633–636.
6. Shahidi, G.G. Design-Technology Interaction for Post-32 nm Node CMOS Technology. In *Proceedings of the 2010 Symposium on VLSI Technology (VLSIT)*, Honolulu, HI, USA, 15–17 June 2010; pp. 143–144.
7. Helms, D.; Schmidt, E.; Nebel, W. Leakage in CMOS Circuits—An Introduction. In *Integrated Circuit and System Design. Power and Timing Modeling, Optimization and Simulation, 14th International Workshop (PATMOS 2004)*; Springer: Berlin, Germany, 2004; pp. 17–35.
8. Morifuji, E.; Yoshida, T.; Kanda, M.; Matsuda, S.; Yamada, S.; Matsuoka, F. Supply and Threshold-Voltage Trends for Scaled Logic and SRAM MOSFETs. *IEEE Trans. Electron Device* **2006**, *53*, 1427–1432.
9. Roy, K.; Mukhopadhyay, S.; Mahmoodi-Meimand, H. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. *Proc. IEEE* **2003**, *91*, 305–327.
10. Mann, R.W.; Abadeer, W.W.; Breitwisch, M.J.; Bula, O.; Brown, J.S.; Colwill, B.C.; Cottrell, P.E.; Crocco, W.G., Jr.; Furkay, S.S.; Hauser, M.J.; *et al.* Ultralow-power SRAM technology. *IBM J. Res. Dev.* **2003**, *471*, 553–563.
11. Kawa, J. Low power and power management for CMOS—An EDA perspective. *IEEE Trans. Electron Device* **2008**, *55*, 186–196.
12. Wong, B.P.; Mittal, A.; Cao, Y.; Starr, G. *Nano-CMOS Circuit and Physical Design*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
13. Shauly, E.N.; Parag, A.; Krispil, U.; Rotstein, I. Device performances analysis of standard-cells transistors using silicon simulation and build-in device simulation. *Proc. SPIE* **2010**, doi:10.1117/12.845622.
14. Shauly, E.; Parag, A.; Khmaisy, H.; Krispil, U.; Adan, O.; Levi, S.; Latinski, S.; Schwarzband, I.; Rotstein, I. Standard cell electrical and physical variability analysis based on automatic physical measurement for design-for-manufacturing purposes. *Proc. SPIE* **2011**, doi:10.1117/12.881841.
15. Shauly, E.; Drori, R.; Cohen-Yasour, M.; Rotstein, I.; Peltinov, R.; Bartov, A.; Latinski, S.; Siany, A.; Geshesl, M. Accurate device simulations through CD-SEM-based edge-contour extraction. *Proc. SPIE* **2008**, doi:10.1117/12.772648.
16. Wang, P.-H.; Lee, B.; Han, G.; Rouse, R.; Hurat, P.; Verghese, N. Addressing Parametric Impact of Systematic Pattern Variations in Digital IC Design. In *Proceedings of the IEEE Custom Integrated Circuits Conference (CICC '07)*, San Jose, CA, USA, 16–19 September 2007; pp. 587–590.
17. King, M.-C.; Chin, A. New test structure to monitor contact-to-poly leakage in sub-90 nm CMOS technologies. *IEEE Trans. Semi. Manf.* **2008**, *21*, 244–247.

18. Vaserman, Y.; Shauly, E.N. Design ranking and analysis methodology for standard cells and full-chip physical optimization. *Proc. SPIE* **2009**, doi: 10.1117/12.812972.
19. Singhal, R.; Balijepalli, A.; Subramaniam, A.; Liu, F.; Nassif, S.; Cao, Y.; Singhal, R. Modeling and Analysis of Non-Rectangular Gate for Post-Lithography Circuit Simulation. In *Proceedings of the 44th ACM/IEEE Design Automation Conference (DAC '07)*, San Diego, CA, 4–8 June 2007; pp. 823–828.
20. Kim, H.-W.; Lee, J.-Y.; Shin, J.; Woo, S.-G.; Cho, H.-K.; Moon, J.-T. Experimental investigation of the impact of LWR on sub-100-nm device performance. *IEEE Trans. Electron Device* **2004**, *51*, 1984–1988.
21. Ban, Y.; Pan, D.Z. Modeling of Layout Aware Line-Edge Roughness and Poly Optimization for Leakage Minimization. accepted for publication. *IEEE Trans. Emerg. Sel. Top. Circuits Syst.* **2011**, *1*, 1–10.
22. Ban, Y.; Sundareswaran, S.; Pan, D.Z. Total Sensitivity Based on DFM Optimization of Standard Libray Cells. In *Proceedings of the ISPD '10 Proceedings of the 19th International Symposium on Physical Design*, New York, NY, USA, 14–17 March 2010; pp. 113–120.
23. Arnaud, F.; Boeuf, F.; Salvetti, F.; Lenoble, D.; Wacquant, F.; Regnier, C.; Morin, P.; Emonet, N.; Denis, E.; Oberlin, J.C.; *et al.* A Functional 0.69 μm^2 Embedded 6T-SRAM bit cell for 65 nm CMOS platform. In *Proceedings of the 2003 Symposium on VLSI Technology*, Kyoto, Japan, 10–12 June 2003; pp. 65–66.
24. Utsumi, K.; Morifuji, E.; Kanda, M.; Aota, S.; Yoshida, T.; Honda, K.; Matsubara, Y.; Yamada, S.; Matsuoka, F. A 65 nm Low Power CMOS Platform with 0.495 μm^2 SRAM for Digital Processing and Mobile Applications. In *Proceedings of the 2005 Symposium on VLSI Technology*, Washington DC, 14–16 June 2005; pp. 216–217.
25. Lin, Q.; Ma, M.; Vo, T.; Fan, J.; Wu, X.; Li, R.; Li, X.-Y. Design-for-manufacturing for Multigate oxide CMOS process. *IEEE Trans. Semi. Manf.* **2008**, *21*, 41–45.
26. Wu, S.-Y.; Liaw, J.J.; Lin, C.Y.; Chiang, M.C.; Yang, C.K.; Cheng, J.Y.; Tsai, M.H.; Liu, M.Y.; Wu, P.H.; Chang, C.H.; *et al.* A Highly Manufacturable 28 nm CMOS Low Power Platform Technology with Fully Functional 64 Mb SRAM Using Dual/Tripe Gate Oxide Process. In *Proceedings of the 2009 Symposium on VLSI Technology*, Honolulu, HI, USA, 16–18 June 2009; pp. 210–211.
27. Jeon, H.-J.; Kim, Y.-B.; Choi, M. Standby leakage power reduction technique for nanoscale CMOS VLSI systems. *IEEE Trans. Instrum. Meas.* **2010**, *59*, 1127–1133.
28. Yasuda, Y.; Akiyama, Y.; Yamagata, Y.; Goto, Y.; Imai, K. Design methodology of body-biasing scheme for low power system LSI with multi-V_{th} transistors. *IEEE Trans. Electron Device* **2007**, *54*, 2946–2952.
29. Wong, B.; Zach, F.; Moroz, V.; Mittal, A.; Starr, G.; Kahng, A. *Nano-CMOS Design for Manufacturability*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
30. Tilke, A.; Stapelmann, C.; Eller, M.; Bach, K.-H.; Hampp, R.; Lindsay, R.; Conti, R.; Wille, W.; Jaiswal, R.; Galiano, M.; *et al.* Shallow trench isolation for the 45-nm CMOS node and geometry dependence of STI stress on CMOS device performance. *IEEE Trans. Semi. Manf.* **2007**, *20*, 59–67.

31. Eiho, T.; Sanuki, E.; Morifuji, T.; Iwamoto, G.; Sudo, K.; Fukasaku, K.; Ota, T.; Sawada, O.; Fuji, H.; Nii, M.; *et al.* Management of Power and Performance with Stress Memorization Technique for 45 nm CMOS. In *Proceedings of the 2007 IEEE Symposium on VLSI Technology*, Kyoto, Japan, 12–14 June 2007; pp. 218–219.
32. Lee, K.; Kang, C.; Yoo, O; Young, C.; Bersuker, G.; Park, H.; Lee, J.; Hwang, H.; Lee, B.; Lee, H.-D.; *et al.* A Comparative Study of Reliability and Performance of Strain Engineering using CESL Stressor and Mechanical Strain. In *Proceedings of the IEEE International Reliability Physics Symposium (IRPS 2008)*, Phoenix, AZ, USA, 27 April–1 May 2008; pp. 306–309.
33. Ota, K.; Sanuki, T.; Yahashi, K.; Miyanami, Y.; Matsuo, K.; Idebuchi, J.; Moriya, M.; Nakayama, K.; Yamaguchi, R.; Tanaka, H. Scalable eSiGe S/D technology with less layout dependence for 45-nm generation. In *Proceedings of the 2006 Symposium on VLSI Technology*, Honolulu, HI, USA, 13–15 June 2006, pp. 64–65.
34. Luo, Y.; Nayak, D.K. Enhancement of CMOS performance by process-induced stress. *IEEE Trans. Semi. Manf.* **2005**, *18*, 63–68.
35. Ge, C.-H.; Lin, C.-C.; Ko, C.-H.; Huang, C.-C.; Huang, Y.-C.; Chan, B.-W.; Perng, B.-C.; Sheu, C.-C.; Tsai, P.-Y.; Yao, L.-G.; *et al.* Process-Strained Si (PSS) CMOS Technology Featuring 3D Strain Engineering. In *Proceedings of the IEEE International Electron Devices Meeting (IEDM '03)*, Washington, DC, USA, 8–10 December 2003; pp. 371–374.
36. Wang, T.-J.; Ko, C.-H.; Chang, C.-J.; Wu, S.-L.; Kuan, T.-M.; Lee, W.-C. The effects of mechanical uniaxial stress on junction leakage in nanoscale CMOSFETs. *IEEE Trans. Electron Device* **2008**, *55*, 572–577.
37. Ban, Y.; Pan, D.Z. Compact Modeling and Robust Layout Optimization for Contacts in Deep Sub-wavelength Lithography. In *Proceedings of the Design Automation Conference (DAC)*, Anaheim, CA, USA, 13–18 July 2010.
38. Chakraborty, A.; Shi, S.X.; Pan, D.Z. Layout Level Timing Optimization by Leveraging Active Area Dependent Mobility of Strained-Silicon Devices. In *Proceedings of the Design, Automation & Test Europe (DATE)*, Munich, Germany, 10–14 March 2008.
39. Joshi, V.; Cline, B.; Sylvester, D.; Blaauw, D.; Agarwal, K. Mechanical stress aware optimization for leakage power reduction. *IEEE Trans. Comput. Aided Des.* **2010**, *29*, 722–736.
40. Venkatraman, R.; Castagnetti, R.; Kobozeva, O.; Duan, F.L.; Kamath, A.; Sabbagh, S.T.; Vilchis-Cruz, M.A.; Jhy Liaw, J.; You, J.-C.; Ramesh, S. The design, analysis, and development of highly manufacturable 6-T SRAM bitcells for SoC applications. *IEEE Trans. Electron Device* **2005**, *52*, 218–226.
41. Zhang, K.; Bhattacharya, U.; Chen, Z.; Hamzaoglu, F.; Murray, D.; Vallepalli, N.; Wang, Y.; Zheng, B.; Bohr, M. SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction. *IEEE J. Solid State Circ.* **2005**, *40*, 895–901.
42. Smayling, M.; Axelrad, V. Simulation-Based Lithography Optimization for Logic Circuits at 22 nm and Below. In *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices, SISPAD '09*, San Diego, CA, USA, 9–11 September 2009; pp. 1–4.
43. Amelifrad, B.; Fallah, F.; Pedram, M. Leakage minimization of SRAM cells in a dual-V_t and dual-T_{ox} technology. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2008**, *16*, 851–860.

44. Sill, F.; You, J.; Timmermann, D. Design of Mixed Gates for Leakage Reduction. In *Proceedings of the 17th ACM Great Lakes Symposium on VLSI*, New York, NY, USA, 11–13 March 2007.
45. Yang, H.S.; Wong, R.; Hasumi, R.; Gao, Y.; Kim, N.S.; Lee, D.H.; Badrudduza, S.; Nair, D.; Ostermayr, M.; Kang, H.; *et al.* Scaling of 32 nm Low Power SRAM with High-k Metal Gate. In *Proceedings of the IEEE International Electron Devices Meeting (IEDM 2008)*, San Francisco, CA, USA, 15–17 December 2008; pp. 233–236.
46. Narendra, S.G. Challenges, Design Choices in nanoscale CMOS. *ACM J. Emerg. Technol. Comput. Syst.* **2005**, *1*, 7–49.
47. Valentian, A.; Beigne, E. Automatic gate biasing of an SCCMOS power switch achieving maximum leakage reduction and lowering leakage current variability. *IEEE J. Solid State Circ.* **2008**, *43*, 1688–1698.
48. Paul, A.C.; Agarwal, A.; Roy, K. Low-power design techniques for scaled technologies. *Integration* **2006**, *39*, 64–89.
49. Rahman, H.; Chakrabarti, C. A leakage estimation and reduction technique for scaled CMOS logic circuits considering gate-leakage. In *Proceedings of the International Symposium on Circuits and Systems*, Vancouver, Canada, 23–26 May 2004; pp. 297–300.
50. Lee, D.; Zhai, B.; Blaauw, D.; Sylvester, D. *Ultra Low-Power Electronics and Design*; Macii, E., Ed.; Kluwer Academic Publishers: New York, NY, USA, 2004.
51. Chang, M.-C.; Chang, C.-S.; Chao, C.-P.; Goto, K.-I.; Ieong, M.; Lu, L.-C.; Diaz, C.H. Transistor- and circuit-design optimization for low-power CMOS. *IEEE Trans. Electron Device* **2008**, *55*, 84–95.
52. Razavipiur, G.; Afazali-Kusha, A.; Pedram, M. Design and analysis of two low-power SRAM cell structures. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2009**, *17*, 1551–1555.
53. Elakkumanan, P.; Narasimhan, A.; Sridhar, R. NC-SRAM—A low-leakage memory circuit for ultra deep submicron designs. In *Proceedings of the IEEE International SOC (Systems-on-Chip) Conference*, Rochester, NY, USA, 17–20 September 2003; pp. 3–6.
54. Chuang, C.-T.; Mukhopadhyay, S.; Kim, J.-J.; Kim, K.; Rao, R. High-performance SRAM in nanoscale CMOS: Design challenges and techniques. In *Proceedings of the IEEE International Workshop on Memory Technology, Design and Testing*, Taipei, Taiwan, 3–5 December 2007; pp. 4–12.