Mechanical Stress Aware Optimization for Leakage Power Reduction

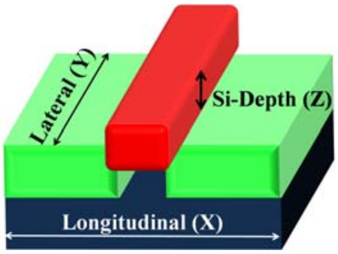
Vivek Joshi, *Student Member, IEEE*, Brian Cline, *Student Member, IEEE*, Dennis Sylvester, *Senior Member, IEEE*, David Blaauw, *Senior Member, IEEE*, Kanak Agarwal, *Member, IEEE*

***Abstract*—Process-induced mechanical stress is used to enhance carrier transport and achieve higher drive currents in current CMOS technologies. This paper explores how to fully exploit the layout dependency of stress enhancement and proposes a circuit-level, block-based, stress-enhanced optimization algorithm that uses stress-optimized layouts in conjunction with dual-V_{th} assignment to achieve optimal power-performance tradeoffs. We begin by studying how channel stress and drive current depend on layout parameters such as active area length and contact placement, while considering all layout-dependent sources of mechanical stress in a 65nm industrial process. We then investigate the three main layout properties that impact mechanical stress in this process and discuss how to improve stress-based performance enhancement in standard cell libraries. While varying the stress-altering layout properties of a number of standard cells in a 65nm industrial library, we show that "dual-*Stress*" standard cell layouts (analogous to "dual-V_{th}") can be designed to achieve drive current differences up to ~14% while incurring less than half the leakage penalty of dual-V_{th}. Therefore, when the flexibility of "dual-Stress" assignment is combined with dual-V_{th} assignment (within the proposed joint optimization framework), simulation results for a set of benchmark circuits show that leakage is reduced by ~24% on average, for iso-delay, when compared to dual-V_{th} assignment. Since mobility enhancement does not incur the exponential leakage penalty associated with V_{th} assignment, our optimization technique is ideal for leakage power reduction. However, our framework can also be used to achieve higher performance circuits for iso-leakage and our joint optimization framework can be used to reduce delay on average by ~5%. In both cases, the proposed method only incurs a small area penalty (<0.5%).**

***Index Terms*—Stress, mobility, layout, leakage, performance.**

I. Introduction and Overview

As INDUSTRY strives to extend Moore's law through aggressive process scaling, significant challenges arise. Maintaining performance and reliability while facing fundamental scaling limitations is a major challenge. We can no longer scale certain device parameters such as gate oxide thickness (t_{ox}), threshold voltage (V_{th}), and supply voltage (V_{DD}) as aggressively as gate length (L) without significantly degrading reliability and exponentially increasing leakage current. Additionally, as MOSFETs continue to scale below 100nm, higher effective fields cause mobility degradation, leading to decreasing drive currents. In order to battle mobility degradation and achieve higher drive currents, modern-day fabrication processes use special means to induce mechanical stress in MOSFETs, which enhances carrier mobility. Mobility enhancement has emerged as an attractive complement to device scaling because it can achieve similar device performance improvements with reduced effects on reliability and leakage.



	NMOS	PMOS
X	Tensile	Compressive
Y	Tensile	Tensile
Z	Compressive	Tensile

Fig. 1. Desired stress types for NMOS and PMOS devices [3].

Mechanical stress in silicon breaks crystal symmetry and removes the 2-fold and 6-fold degeneracy of the valence and conduction bands, respectively [1,2]. This leads to changes in the band scattering rates and/or the carrier effective mass, which in turn affects carrier mobility. Mechanical stress induced in a CMOS channel can be either tensile or compressive. As illustrated in Fig. 1, NMOS and PMOS devices have different desired stress types (compressive or tensile) in the longitudinal, lateral, and Si-depth (vertical) dimensions. By providing the correct type of stress for a device (in one or more dimensions), we can achieve higher drain currents. However, since carrier mobility affects the drain current in all MOSFET operation regimes, increased carrier mobility not only increases saturation current, it also increases subthreshold current. Specifically, short-channel MOSFET saturation drain current, $I_{D,sat}$, has a sub-linear dependence on mobility, μ_0, while the subthreshold drain current ($I_{D,sub}$) dependence on mobility is linear [4,5]. These two relationships between drain current and mobility make mobility enhancement an interesting alternative to other power/delay optimization techniques.

One of the most popular power/delay optimization techniques that has been researched considerably in both academia and industry is the dual-V_{th} optimization scheme [6,7]. This technique typically uses gate sizing and two choices of threshold voltage to optimize a given circuit for some metric (usually delay or power). Since $I_{D,sat}$ and $I_{D,sub}$ are super-linearly and exponentially dependent on V_{th}, respectively, V_{th} can potentially be a powerful optimization parameter. However, since incorporating different threshold voltages adds significant design and process complexity, practical implementations typically restrict the number of threshold voltages to ~2 [8].

One of the main disadvantages of using a dual-V_{th} scheme is, coincidentally, also one of its strengths: each gate in the design can either be high-performance or low-leakage. Dual-V_{th} provides for a wide range of performances (due to the super-linear and exponential dependencies of $I_{D,sat}$ and $I_{D,sub}$ on V_{th}, respectively), but the approach has only coarse granularity in its selection. Mobility enhancement induced by mechanical stress, however, is layout dependent and can therefore provide much finer delay-versus-leakage control without adding to process complexity/cost. This granularity, coupled with the fact that leakage is only linearly dependent on mobility, makes stress-induced mobility enhancement an interesting research topic that can either be directly compared to dual-V_{th} assignment, or used concurrently to provide additional gains in either leakage or delay. Since the leakage penalty incurred by mobility enhancement is significantly less than V_{th} assignment, we focus on leakage reduction in this work. However, for completeness, we also show that our joint optimization framework can be used to reduce circuit delay for iso-leakage.

To date, there has been limited research on the layout dependence of stress-based current improvement. Most of the published work has focused on the effects of Shallow Trench Isolation (STI) [9-12] or limited their analysis to only include the PMOS sources of mechanical stress [13-16]. Reference [17] studies variability in CMOS circuits for a low power 45nm test chip featuring STI and tensile nitride liner as sources of stress (NMOS only). One key result is that NMOS devices show 5% higher performance as source/drain diffusion lengths are increased by 75%, which is qualitatively similar to our results for a process with added stress sources for both PMOS and NMOS. In the last few years, researchers have begun exploring layout optimization techniques involving stress. In [10], the authors presented an active-layer fill insertion technique which optimized circuit delay by exploiting STI stress. However, in the 65nm industrial technology used in this research, we discovered that the STI stress contribution was <10% of the total channel stress, making STI optimization less effective. The first optimization scheme developed to exploit the source/drain length dependency was published in [18], which described a timing closure technique that utilized stress enhanced versions of standard cells to improve path delays. While the authors in [18] do report average delay savings of ~5%, they do not disclose the additional leakage power consumed, nor do they discuss possible leakage versus delay tradeoffs.

This work differs from previously published research in that it incorporates all of the layout dependent sources of stress and, consequently, exploits a larger number of layout properties that affect stress (e.g., source/drain lengths, contact placement, distance from STI, etc.). Additionally, unlike [18], our optimization algorithm is not a one-sided approach that only optimizes delay. The proposed optimization accounts for the tradeoff between leakage and delay and it achieves the largest improvement in leakage power (delay) for identical delay (leakage power). Thus, to our knowledge, this paper is the first work to use stress-enhanced standard cells in a new, circuit-level, block-based, joint optimization framework that improves either leakage power consumption for iso-delay-performance or circuit delay for iso-leakage-power-consumption.

In this paper, we begin by addressing the layout dependency of stress-based performance enhancement. We perform a comprehensive study in order to determine how various layout parameters affect device stress, and then analyze their impact on device performance. From this study we then extract the main layout properties that impact mechanical stress in our industrial, 65nm process. Next, these layout properties allow us to create "high-Stress" and "low-Stress" versions of a subset of standard cells from an industrial 65nm CMOS library (analogous to "low-V_{th}" and "high-V_{th}" cells in a dual-V_{th} library). Finally, we propose a stress-aware optimization algorithm and generate two comparisons: 1) stress-based performance enhancement versus dual-V_{th} assignment, and 2) combined stress-based enhancement with dual-V_{th} versus only dual-V_{th}.

By applying layout-based mobility enhancement, experimental results show that we can obtain a 12% performance increase for PMOS devices (up to about 20%), while only increasing the leakage current by ~3.8X. For NMOS devices, we can achieve a drive current improvement of about 5% while increasing the leakage current by only 1.4X. For the stress-enhanced standard cells in our library, we find that leakage is reduced by ~2X on average when compared to an equivalent V_{th}-modified cell (where V_{th} is changed arbitrarily to match the stress-enhanced delay). Overall, by combining the two performance enhancement techniques (stress-based and dual-V_{th}) for a few benchmark circuits, we find that the combined approach, for iso-delay, decreases leakage on average by 23.8% when compared to dual-V_{th} assignment. Similarly, if we use our optimization algorithm and match leakage (iso-leakage), delay reduces on average by 5.1%. In both cases, our proposed method only incurs a small area penalty (<0.5%).

The rest of the paper is organized as follows. Background for this work is discussed in Section II. Section III presents a study on the layout dependence of stress-based performance enhancement, while Section IV outlines stress-dependent layout properties for our 65nm technology. Results obtained by modifying these properties in 65nm industrial standard cells is discussed in Section V. Section VI includes details on the proposed optimization methodology. The experimental setup and results for the optimization algorithm are reported in Section VII, and Section VIII concludes the paper.

II. Background

This section discusses the two main topics that are the foundation of this work: the sources of mechanical stress (and their dependency on layout properties) and how mobility and V_{th} affect drain current.

A. Mechanical Stress Sources and their Layout Dependence

Mechanical stress in silicon can be generated by either thermal mismatch or lattice mismatch. Thermal mismatch stress is caused by differences in the thermal expansion coefficient, while lattice mismatch stress is caused by differences in lattice constants. Fig. 2 shows the major sources of stress for one of the

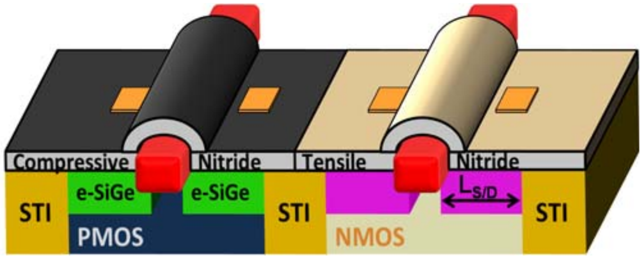


Fig. 2. Sources of stress for NMOS and PMOS devices.

latest 65nm CMOS technologies [19]. The sources are Shallow Trench Isolation (STI), embedded SiGe (only in PMOS devices), tensile/compressive nitride liners (in NMOS/PMOS devices, respectively), and the Stress Memorization Technique (SMT).

Shallow Trench Isolation (STI): STI creates compressive stress longitudinally and laterally due to thermal mismatch [10,12-14] and volume expansion [14]. From Fig. 1, it is apparent that this compressive stress degrades the electron mobility in NMOS devices (in both the longitudinal and lateral directions) [20] and degrades hole mobility in PMOS devices in the lateral direction. However, STI stress that is induced longitudinally (e.g., at the left and right boundaries of standard cells) actually improves hole mobility in PMOS devices.

Embedded SiGe (eSiGe): For PMOS transistors, an eSiGe process is implemented where SiGe is epitaxially grown in cavities that have been etched into the source/drain (S/D) areas [21]. Lattice mismatch between Si and SiGe creates a large compressive stress in the PMOS channel, resulting in significant hole mobility improvement.

Dual-stress Nitride Liners: As shown in Fig. 2, mechanical stress can also be transferred to the channel through the active area and polysilicon gate by depositing a permanent stressed liner over the device [22]. Tensile liners improve electron mobility in NMOS devices, while compressive liners improve hole mobility in PMOS devices. The latest high performance process nodes have simultaneously incorporated both tensile and compressive stressed liners into a single, high performance CMOS flow, called the Dual-Stress Liner technique. In this process, a highly tensile Si_3N_4 liner is uniformly deposited over the entire wafer. The film is then patterned and etched from the PMOS regions. Next, a highly compressive Si_3N_4 liner is deposited, patterned and etched from the NMOS regions.

Stress Memorization Technique (SMT): In addition to the permanent tensile liner shown in Fig. 2, the Stress Memorization Technique (SMT) is also used to increase the stress in n-type MOSFETs [23]. In this technique, a stressed dielectric layer is deposited over all of the NMOS regions, thermally annealed, and then completely removed. The stress effect is transferred from the dielectric layer to the channel during the anneal and is “memorized” during the re-crystallization of the active area and gate polysilicon.

A closer examination of these stress sources shows that the amount of stress transferred to the channel, and, consequently, the drive current enhancement, has a strong dependence on certain layout properties. The amount of eSiGe (and, hence, the stress), for example, depends upon the length of the active area. Longer active area also means that the STI will be pushed further away from the channel, which will lower its effect on the total channel stress. Therefore, the drive current of a transistor depends not only upon the gate length and width (*L* and *W*), but also upon the exact layout of the individual transistor and its neighboring transistors. This means that the performance of two transistors with identical gate lengths and widths can actually differ significantly, depending on their layouts.

Beginning in Section III, we study the layout dependence of stress-based performance enhancement for different device configurations and identify simple layout properties in our 65nm process that allow us to maximize the performance gains due to stress. The idea is to determine the key layout parameters that a layout designer can change to affect transistor performance. Since we are interested in optimizing the layout, uniform techniques such as SMT can be ignored because SMT involves a uniform film deposition, anneal and removal over all of the NMOS regions, which leads to a uniform shift in NMOS drive current that is relatively independent of layout [24].

B. *Drain Current Dependence on Stress and V_{th}*

Modifying carrier mobility directly affects the amount of current that flows between the source and drain terminals of a transistor. Increased carrier mobility increases the drain current, I_D, in all regimes of MOSFET operation, which improves transistor performance (in terms of delay) but increases leakage power. In order to study the delay-versus-leakage tradeoffs involved in stress enhancement, we examine the saturation and subthreshold current equations in order to determine their dependency on carrier mobility. This also allows us to compare mobility enhancement to other performance enhancement techniques, such as V_{th} reduction. Equations (1) and (2) below give the expressions for drain current when the transistor is operating in the saturation and subthreshold regimes, respectively [4,5].

$$I_{D,\,sat} = \frac{\mu_0}{[1 + U_0(V_{GS} - V_T)]} \cdot \frac{C_{ox}}{2aV} \cdot \frac{W}{L_{eff}} \cdot (V_{GS} - V_T)^2$$

$$V = \frac{1 + v_c + \sqrt{1 + 2v_c}}{2} \qquad v_c = U_1((V_{GS} - V_T)/a) \tag{1}$$

$$I_{D,\,sub} = A \cdot e^{\frac{1}{nv_T} \cdot (V_G - V_S - V_{th0} - \gamma' V_S + \eta V_{DS})} \cdot (1 - e^{(-V_{DS})/v_T})$$

$$A = \mu_0 C_{ox} \frac{W}{L_{eff}} v_T^2 e^{1.8} e^{-\frac{\Delta V_{th}}{\eta v_T}} \tag{2}$$

From (1) and (2), it is evident that the saturation drain current ($I_{D,sat}$) has a sub-linear dependence on mobility, μ_0 (due to the vertical field mobility degradation coefficient, U_0) while the subthreshold drain current ($I_{D,sub}$) dependence on μ_0 is linear. The drain current dependence on V_{th}, however, is almost linear in saturation, but is exponential in the subthreshold regime. Therefore, if we obtain identical saturation current improvement using two separate enhancement techniques: 1) stress-based mobility enhancement, and 2) V_{th} reduction, then the corresponding increase in leakage current for the reduced-V_{th} case will be much higher (due to the exponential dependence of $I_{D,sub}$ on V_{th}). Consequently, the reduced increase in leakage current makes mobility enhancement a more attractive option than its V_{th} counterpart.

The benefits of using mobility enhancement over V_{th} reduction is illustrated in Fig. 3, which shows the normalized I_{on} versus I_{off} curves for stress-based and V_{th}-based performance enhancement for an isolated, 65nm PMOS device. The device has three sources of stress: STI, a compressive nitride liner, and eSiGe source/drain regions. Stress is varied by changing the active area length, while the n-channel doping is changed to vary V_{th}. The curves clearly show that the tradeoff is better for stress variation. For a 12% improvement in I_{on}, the leakage for

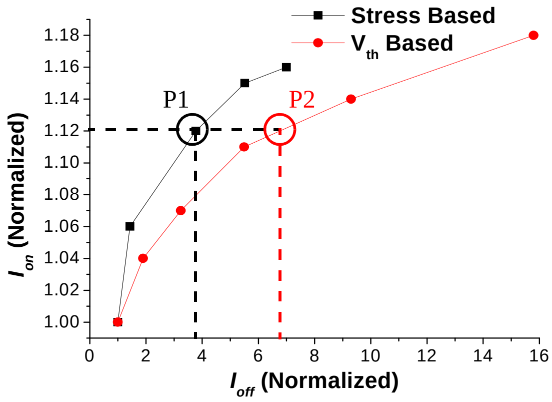


Fig. 3. I_{on} vs. I_{off} for V_{th}- & stress-based enhancement in a 65nm PMOS.

the V_{th} case is nearly twice as large as that for the stress-based improvement (shown in Fig. 3 as points P1 and P2), and the difference is only amplified for higher values of improvement. Also, stress-based improvement allows for more fine-grain improvement control than V_{th} assignment, given that only two or three V_{th} values are typically allowed. Therefore, a designer would prefer to achieve performance improvements through stress-enhancement whenever possible, due to the reduced leakage penalty and increased granularity. The superiority of the stress-based performance improvement technique makes it an appealing option for further investigation. Thus, the next two sections study the layout dependence of stress, and identify the primary layout properties that can be modified so that stress-induced enhancements are maximized.

III. Layout dependence of stress-based enhancement

In order to study the layout dependence of stress-based performance enhancement, we used the Davinci 3D TCAD tool [25], which has an extensive set of stress-related features. Additionally, we followed the layout rules from an industrial 65nm CMOS technology and the device fabrication was simulated in Tsuprem4 [26] (in order to capture the process-induced stress). The stress values were then imported into Davinci, which simulated the device and solved for the stress-based mobility enhancement equations. The resulting values for drive current and leakage were verified against experimental test chip data, which was consistent with previously published 65nm technology data for minimum sized NMOS and PMOS devices [19]. Furthermore, the simulated values of stress were in close agreement with previously reported data for PMOS channel stress while considering all of the layout dependent sources of stress [21]. Due to the absence of any previously published data on the layout dependence of stress or drive-current (due to stress), measured test chip results were used to quantify the impact of layout diversity on device performance. The fabrication process used for this test-chip employs all the known stress enhancement techniques. The hardware data was used to verify the accuracy of our TCAD setup, and the TCAD-based simulation results were found to be in close agreement with the measured data. Our consistency with these fabricated measurements can be attributed to the fact that we model all of the layout dependent sources of stress in the industrial 65nm technology. For a PMOS device, the sources of stress that are layout dependent include the compressive nitride liner, eSiGe, and STI. The NMOS sources, on the other hand, only include the tensile nitride liner and STI. We have ignored the Stress Memorization Technique (SMT) in our simulations, since it involves a uniform deposition and eventual removal of a dielectric layer over all NMOS devices (as discussed previously in Section II-A). SMT, therefore, does not depend on layout properties and can be accurately treated as a uniform increase in NMOS drive current, independent of layout [24].

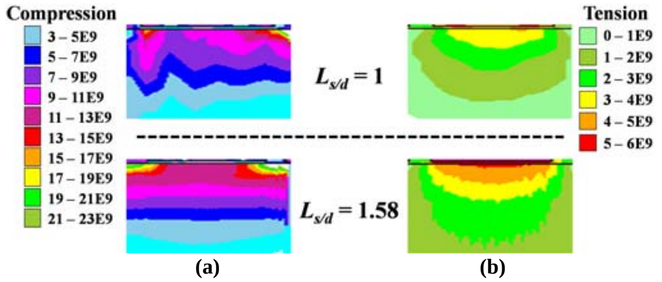


Fig. 4. Longitudinal stress component S_{xx} (in Pascals) for normalized $L_{S/D}$ of 1 and 1.58 for (a) PMOS (b) NMOS.

Previously, Fig. 2 showed the 3D cross-section of an isolated PMOS device surrounded by STI. For the device shown, we increase the active area length ($L_{S/D}$) and examine the corresponding changes in drive current.[1] Increasing active area length has a number of effects: 1) it increases the amount of eSiGe, causing more stress to be transferred to the channel; 2) it increases the distance between the channel and the STI, decreasing the effect STI has on channel stress; and 3) it allows more nitride over the active area. The nitride layer actually transfers stress in two ways – vertically through the gate and longitudinally through the active area. Since active contacts create openings in the nitride layer, the longitudinal component of nitride stress can be increased by moving the contacts away from the channel. Similarly, a source/drain region that does not have any contacts (or has a smaller number of contacts) will have higher channel stress than one that has a high contact density.

Fig. 4a shows the longitudinal stress (S_{xx}) in the same isolated PMOS device for two normalized $L_{S/D}$ values of 1 and 1.58 (the values are normalized to the length of a minimum-sized, contacted S/D region). Fig. 5 shows the PMOS drive current,

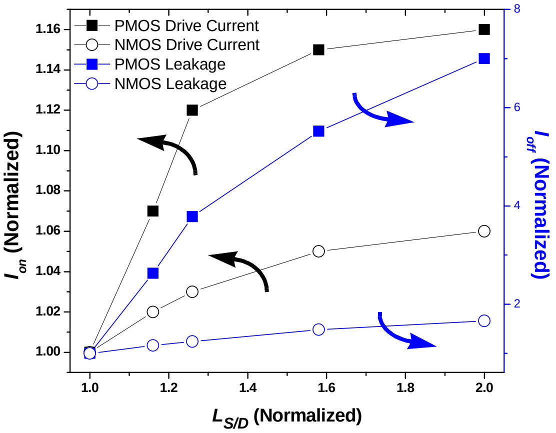


Fig. 5. I_{off} and I_{on} vs. $L_{S/D}$ curves for stress-based performance enhancement in isolated PMOS and NMOS devices.

[1] The authors would like to note that in this work, $L_{S/D}$ is equivalent to both the $L_{S/D}$ and $L_{p/p}$ used in previous works (such as [18]). Thus, in the remainder of the paper, $L_{S/D}$ can refer to any longitudinal S/D dimension.

I_{on}, and leakage current, I_{off}, plotted against $L_{S/D}$, while Fig. 6 shows the normalized PMOS longitudinal stress plotted against $L_{S/D}$. Results show that for a 12% performance increase, leakage current only increases by 3.78X. This I_{on} versus I_{off} tradeoff is much better than the tradeoff produced by the alternative, V_{th}-based enhancement technique, as predicted in Section II-B. Additionally, Fig. 5 shows the saturation point for extending $L_{S/D}$. Increasing the S/D length beyond 1.58 (normalized) yields minimal performance gains, even when active area length and leakage current are increased substantially. Finally, the performance enhancement is also sensitive to contact placement. Moving the contacts away from the channel accounts for nearly 2.6% of the drive current improvement and a device with a non-contacted drain (typically seen in series devices) has ~4% higher performance.

Unlike its PMOS counterpart, NMOS device performance is actually degraded by STI since STI induces compressive stress in the channel. Thus, increasing NMOS $L_{S/D}$ not only pushes away the compressive STI, but it also allows for more contact separation from the channel. Fig. 4b shows the longitudinal stress in an isolated NMOS device for normalized $L_{S/D}$ values of 1 and 1.58. In addition to PMOS I_{on} and I_{off}, Fig. 5 also shows NMOS I_{on} and I_{off} while Fig. 6 shows its normalized longitudinal stress versus $L_{S/D}$. For NMOS devices, a 5% performance gain can be achieved for a 1.48X increase in leakage current. NMOS devices also have the same (normalized) upperbound for $L_{S/D}$ extension as their PMOS counterparts, 1.58. Beyond this value, the area and leakage current penalties do not warrant the minimal gains in I_{on}. The increase in performance in NMOS devices, however, is limited by the fact that we are only increasing the nitride's longitudinal stress through the active area (about 35% of the total stress due to the nitride liner), and pushing away the STI (which has a relatively smaller contribution to the overall channel stress). Experimental results show that almost 80% of the total NMOS improvement is due to moving the contacts and a device with a non-contacted drain has ~2% higher performance.

Next, we studied transistor performance in denser layouts. Fig. 7 shows the channel stress and the corresponding layout view for three PMOS transistors in a 3-input NAND gate. The device in the center (device 2) has higher stress than the two corner transistors because it is surrounded by more eSiGe (its own S/D regions as well as its neighbors' S/D regions). This difference in stress is reflected in their drive current performance, and simulations show that the drive currents for the center and edge devices differ by 8.2%. Furthermore, if there were five devices side-by-side instead of three, the difference would increase to 14.8%. This means that the drive current of a transistor is not only layout-dependent, but it is also *location-dependent*. Similar experiments for NMOS devices show differences of 7.4% and 12.2% for the case of three and five side-by-side transistors, respectively.

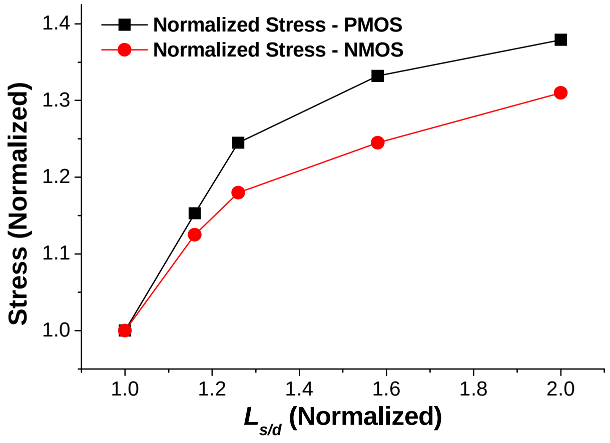


Fig. 6. Longitudinal Stress vs. $L_{S/D}$ for isolated PMOS and NMOS devices.

IV. Layout properties that impact mechanical stress and performance

Based on the intuition developed in the previous section, we now identify 3 simple layout properties in our 65nm technology that can be used to optimize a given layout for stress-induced performance enhancement. Once the properties are presented, the end of this section discusses one other important stress effect: the position-dependency of stress-induced performance enhancement. When mechanical stress is present in MOSFETs, matching W and L does not guarantee similar transistor performance even when neglecting process variation. Apart from W and L, the drive current is also affected by the layout parameters that influence stress: active area length, placement and number of contacts, and device context (i.e., whether the device is surrounded by other transistors or isolated by STI on one or both sides). In this paper, we have already discussed the first two parameters in great detail, while the third parameter (device context) has only been briefly mentioned (at the end of Section III). However, since the device context or position of a transistor within a layout also affects performance, it must be accounted for by the designer, so this phenomenon is discussed in more detail at the end of the section.

Upon finishing the layout dependency study in Section III, we determined that in our 65nm industrial process, the following 3 properties had the largest impact on improving performance (without modifying existing cell boundaries).

Layout Property #1: ***Active Area or Source/Drain Lengths***

Using the length of a transistor's source or drain regions (or, equivalently, changing the amount of active/diffusion area) to modify stress-enhancement is well known technique and has been studied in a number of works [13,16-18]. Increasing the active area moves the STI regions away from the channels and increases the amount of eSiGe in PMOS devices. Moving the STI farther from the channel improves the performance of NMOS devices since STI exerts a compressive

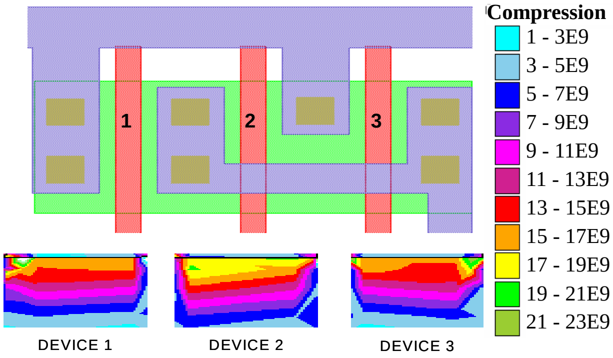


Fig. 7. PMOS devices for a 3-input NAND gate and the corresponding channel stress distribution (in Pa).

stress in the longitudinal direction, which degrades the NMOS electron mobility. For PMOS devices, on the other hand, compressive STI stress is actually beneficial and improves hole mobility. However, increasing the active area for PMOS devices still results in higher stress due to the relatively small contribution of STI compared to the other sources of stress. Measurements show that the stress due to STI represents <10% of the total channel stress. Therefore, the increase in eSiGe and its resulting contribution to PMOS channel stress dominates the stress due to STI and provides a significant increase in hole mobility.

Increasing the active area can most readily be accomplished in a compact pull-up or pull-down network (often containing an NMOS or PMOS stack) that does not use the full width of a cell (Fig. 8 shows the scope for increasing the active area of a PMOS stack in a 3-input NOR gate). In the case of stacked transistors, the layout does not require contacts between intermediate nodes. Thus, their spacing can be significantly tighter because nodes that contain contacts need larger spacing to satisfy the technology's design rules. In the absence of stressors, it is best to minimize the active area in order to reduce the capacitance. However, in the presence of stressors, increasing active area length also results in higher stress in the channel (and, hence, higher drive current), in addition to increasing the source/drain capacitances. In a given CMOS layout, increased S/D capacitance for transistors closer to the output will directly affect the output capacitance, while transistors closer to the V_{DD} and V_{SS} rails will have a smaller effect. Hence, this layout property should be increased in cells with larger output loads, so that the change in capacitance is a small fraction of the total output capacitance. The authors would like to note that the mechanical stress dependence on active area can also be exploited to create high performance versions of standard cells which incur some area penalty, but are assigned optimally within a design.

Layout Property #2: *Contact Placement*

Moving the contacts away from the channel allows more stress to be transferred by the nitride layer. For isolated devices, pulling the contacts as far away from the gate polysilicon as the design rules permit maximizes the stress-enhancement. Contacts between two gates, on the other hand, can either be placed midway for identical performance enhancement of both transistors, or placed closer to the non-critical transistor (increasing stress in the critical device). Moving the contacts away will also result in a small increase in the source/drain resistance, but, in our 65nm study, this increase was typically less than 5Ω (based on sheet resistance calculations for the maximum S/D displacement obtained while creating the stress-aware optimized library), and the resulting gain in drive current outweighed the increase. The maximum S/D contact displacement observed was 60nm.

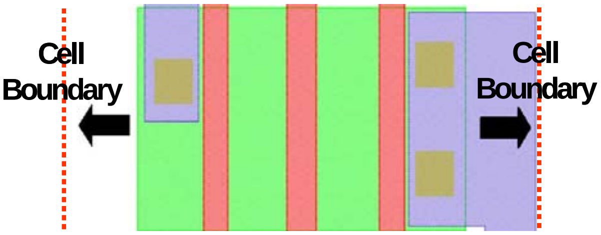
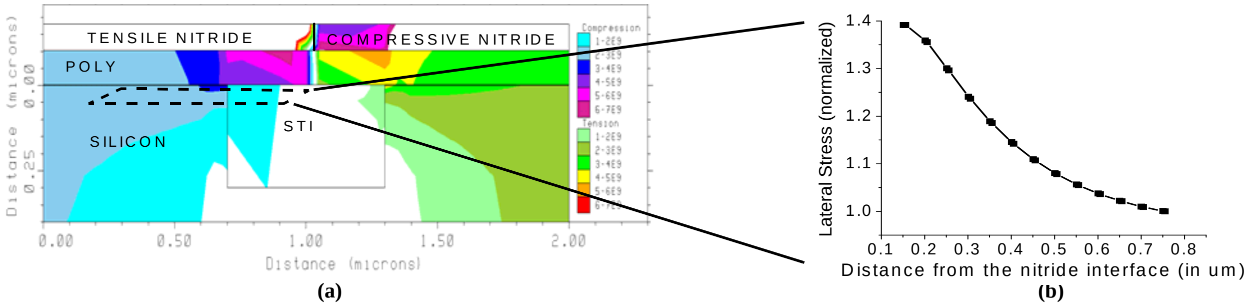


Fig. 8. Application of Layout Property #1 to PMOS stack in 3-input NOR.

Layout Property #3: *Lateral Active Area Placement*

From Fig. 1, we know that the desired stress in the lateral direction is tensile for both NMOS and PMOS devices. Fig. 9a shows the lateral stress behavior near the interface of the two nitride layers (cross-section across the poly going from PMOS to NMOS over STI). Fig. 9b shows the plot of normalized lateral stress (normalized to the stress value at the point farthest from the nitride liner interface) at a depth of 1nm below the Si surface versus the distance from the tensile/compressive liner interface, under the tensile nitride layer. The behavior is interesting in the sense that there is a region of compressive stress under the tensile nitride (the NMOS side) and there is a region of tensile stress under the compressive nitride (the PMOS side). This behavior follows from the physics involved behind the stress-inducing process step. At the compressive/tensile nitride liner interface, each nitride layer exerts an equal and opposite force on the other nitride layer, which imposes the opposite type of stress under the adjacent layer. Therefore, if possible, it is beneficial to move the PMOS active area into this region of tensile stress and the NMOS away from the region of compressive stress. The space for this movement is most readily available when the transistor widths are small but the cell pitch (lateral size) is large (due to pitch uniformity across standard cells). This combination of properties, for example, is common in minimum sized, simple gates (e.g., minimum size inverters, buffers, or 2-input NAND/NOR's).

It should be noted that the lateral active area placement will slightly alter the V_{th} of the shifted devices, due to well edge proximity effects [27-29]. However, since the amount of lateral shift applied to the 65nm standard cells was <0.205μm for the NMOS cells and <0.12μm for the PMOS cells, the corresponding shift in V_{th} was found to be <0.32mV (in both



(a)

(b)

Fig. 9. Stress (in Pascals) at nitride interface for NMOS and PMOS: (a) 2D view across lateral STI (b) Behavior under tensile nitride at channel depth.

HSPICE and TCAD simulations, independently) for all devices.[2] Since this V_{th} shift is relatively small, the reported results described in the remainder of the paper do not include the well edge proximity change induced by Layout Property #3. However, if this shift in threshold voltage becomes appreciable in future processes, our experimental setup can easily be modified to include a well edge proximity model, such as the ones described in [28,29], which will capture the corresponding change in V_{th}.

Apart from these three layout properties, a designer must also be aware of how the channel stress is affected by the position of a device within the layout. Stress in the channel of a device depends not only upon its S/D lengths and contact placement, but also upon its surroundings. As we have shown in the previous section, devices that share their source/drain regions with other transistors have significantly higher stress (and hence drive current enhancement) than those at the edges of an active region (which are therefore bordered by STI), even for identical $L_{S/D}$ and contact placement. This difference in stress can be attributed to the effects of STI, as well as the fact that stressors for a device also affect its neighbors.

Ignoring the position-dependence of stress could lead to a number of design issues. First of all, the location of a transistor could result in an unexpected increase in drive current, resulting in smaller delay and possible hold-time violations, as some gates might be faster than expected. Secondly, the position-dependent current offset could modify the noise margins of a circuit. Hence, for circuits that are sensitive to noise margins (e.g., SRAM cells, Sense Amplifiers, etc.), these deviations must be accounted for either during the design phase (for example, by guardbanding against position-dependent offsets), or during the layout phase (e.g., by modifying the $L_{S/D}$'s to cancel the offsets). Finally, in certain circuits, if the strength of a transistor (in terms of drive current) is increased beyond the expected value, it could cause a substantial drop in performance. A detailed example of context-sensitive design is included in Section V. All in all, designers need to be aware of the effect that position has on performance, especially if pin-to-pin delay, noise margins, or transistor strength are essential to a particular design.

There are three main ways that a designer could capture the position dependence of stress within a particular design: fabrication, TCAD simulation, and electrical circuit simulation. The first solution, fabrication, is an expensive and time consuming endeavor, especially during the early stages of a process's lifetime. The second alternative – using TCAD tools to simulate the position dependence of stress – can be costly in terms of runtime, and convergence becomes extremely difficult when simulating more than 10 devices at once. The final solution, electrical circuit simulation (e.g., HSPICE simulation), promises to be the most efficient in terms of both cost and runtime. Unfortunately, to our knowledge, there has been little research dedicated towards electrical models that capture the layout dependence of stress. Furthermore, of the few that have been published (such as [15]), none have been implemented within an electrical circuit model (e.g., BSIM). The problems associated with each of these solutions make modeling the position dependence of stress an important and interesting research topic that remains largely unexplored.

V. Modifying 65nm standard cell layouts

This section discusses the effectiveness of modifying the layout properties from Section IV in standard cells from an industrial 65nm CMOS technology library. For a given layout, as shown in Section III, a basic tradeoff always exists between the source/drain length, $L_{S/D}$, and the improvement in drive current. By exploiting this tradeoff, we can make faster, but leakier, versions of the standard cells with varying area increments and assign them intelligently to the critical paths in order to optimize performance. The performance enhanced versions all use a combination of the three properties discussed in Section IV: increased $L_{S/D}$, larger poly-to-contact spacing, and stress-aware lateral placement.

For example, Fig. 10a shows the layout for a 3-input NOR gate. It consists of three PMOS transistors in series (a 3-PMOS stack) and three NMOS transistors in parallel. This means that the source and drain of each NMOS is connected to the ground and the output, respectively, necessitating contacts at each node. The PMOS stack on the other hand, only needs one contact to V_{DD} (at the source of the leftmost PMOS) and one contact to the output (at the drain of the rightmost PMOS). Using the classical layout methodology (where stress is ignored and capacitance is minimized), we can shrink the non-contacted S/D regions to lower the parasitic PMOS capacitance. As shown in Fig. 10a (labeled "G1"), the PMOS region has the capability of increasing the source/drain lengths (Layout Property #1) by ~22% without affecting the overall cell area. While increasing the source/drain lengths, we simultaneously shift the contacts away from the gates (Layout Property #2), maximizing performance enhancement. If we increase the active area uniformly for all transistors, drive current improves by ~12% for each PMOS device. Also, there is lateral room to move the NMOS and PMOS active area and exploit the stress dependence of Layout Property #3 (labeled "G3" in Fig. 10a). This leads to further improvements of about 3% and 1.5% for NMOS and PMOS devices, respectively. Therefore, for the 3-input NOR gate, we observe overall improvements in drive current of ~13.5% for

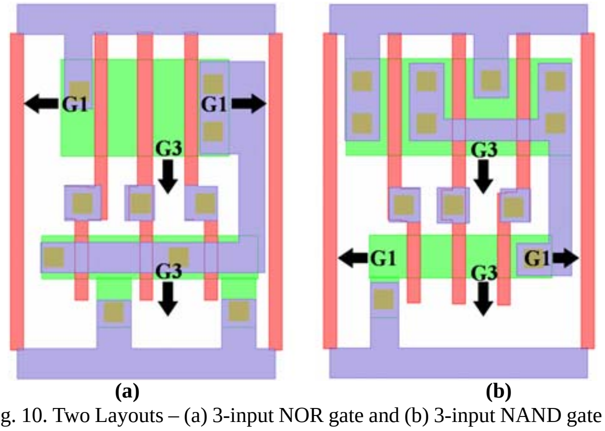


Fig. 10. Two Layouts – (a) 3-input NOR gate and (b) 3-input NAND gate – showing the scope for layout-based stress improvement.

[2] HSPICE well-edge proximity was captured during Calibre PEX parasitic extraction, and then fed into our industrial BSIM models to calculate the effect on V_{th}. Note that the 0.32mV shift reported can be viewed as the *shift* in ΔV_{th} (the change in V_{th} due to well proximity), not total ΔV_{th} itself.

TABLE I

PERCENTAGE CONTRIBUTION OF LAYOUT PROPERTIES 1–3 TO THE OVERALL DRIVE CURRENT IMPROVEMENT FOR PMOS/NMOS STACKS

	Property 1	Property 2	Property 3
NOR3 PMOS	69.6%	19.3%	11.1%
NAND3 NMOS	20.1%	37.8%	42.1%
NOR2 PMOS	53.3%	26.6%	20.1%
NAND2 NMOS	10.1%	27.2%	62.7%

PMOS devices and ~3% for NMOS devices. Similarly, by modifying Layout Properties 1–3 in a 2-input NOR gate, we can achieve drive current improvements of 7.5% and 3% for the PMOS and NMOS devices, respectively.

Similarly, Fig. 10b shows the layout for a 3-input NAND gate. Instead of a PMOS stack, there is an NMOS stack in the NAND gate, so there is a potential to increase the NMOS active area length without affecting the cell area. While altering Layout Properties 1 and 2, we obtain an improvement of ~4% for each of the NMOS drive currents. Also, there is space for moving the active areas to exploit the mobility dependence of Layout Property #3. This leads to further improvements in NMOS and PMOS devices of ~3% and ~1.5%, respectively. Overall, we can achieve a ~7% NMOS performance enhancement and a ~1.5% PMOS performance enhancement. Similarly, by modifying Layout Properties 1–3 of a 2-input NAND, we can obtain drive current improvements of 4.5% and 1.5% for the NMOS and the PMOS devices, respectively. Scope for such layout-based improvements is found in most of the standard cells in our library.

Table I shows the percentage contribution of each layout property to the total drive current improvement achieved for PMOS and NMOS stacks in 2- and 3-input NOR and NAND gates, respectively. The relative contribution of the properties varies between the four cases. This is due to the presence of eSiGe in PMOS which is a major contributor to the overall stress in the channel. As a result, for PMOS devices, altering Layout Property #1 (increasing the active area) results in the maximum improvement as compared to the improvement achieved by modifying the other two properties. However, in the case of NMOS devices, increasing active area results in pushing away the STI, whose contribution to the overall channel stress is relatively smaller. The longitudinal stress due to nitride is increased upon the alteration of Layout Property #2, and Layout Properties 2–3 are the major contributors to the drive current improvement in NMOS devices.

Table II summarizes the results of changing Layout Properties 1–3 in a few standard cells. It reports the percentage drive current improvement, leakage current increase, and the percentage increase in the output capacitance (assuming an FO4 output loading). It also reports the leakage current increase for identical drive current improvements through V_{th} reduction. Comparing the leakage current increase for stress-aware layout optimization to V_{th} reduction re-establishes the superiority of the stress-aware layout optimization. For a 3-input NOR gate, the PMOS leakage current increased by 4X when the layout was optimized to exploit stress dependencies, while the corresponding increase for the V_{th} reduction case was 9.2X. The increase in NMOS leakage for a 3-input NAND gate was found to be 2X for stress-based layout optimization, and 2.4X for the case of V_{th} reduction. Application of Layout Property #1 increased the S/D capacitance since $L_{S/D}$ was increased, but, as shown in Table II, this increase was very small (<3% if we assume an FO4 output loading).

In this same manner, we modified the layout properties from Section IV in ~25 standard cells in a 65nm industrial library, creating a stress-enhanced version of each cell. For the majority of standard cells, the stress-enhanced versions are the same area as the original cells, thus, there is no area penalty. However, since there are no series/stacked devices in inverter layouts, there is negligible space to modify Layout Property #1. The capacitance increase for the “Iso Area INV” is 0% as reported in Table II, because there is only space for the application of Layout Property #3, which does not affect capacitance. Therefore, we decided to create a second, slightly larger, stress-enhanced version of each inverter cell (with ~20% area increase per cell) that achieved larger drive currents (13% increase for PMOS and 6% increase for NMOS). Since the inverters, however, only make up a small subset of our standard cell library, the overall impact on circuit area is <0.5% (as shown later in Table IV). The final stress-enhanced standard cell library is comprised of different sized inverters (iso-area and increased-area versions) as well as 2- and 3-input NAND and NOR gates of varying strengths.

As mentioned in Section III, the position of a device within a layout also affects its stress, and, therefore, its drive current. This position-dependent drive current enhancement can significantly hurt the performance of some circuits. This fact was verified using the circuit shown in Fig. 11, which contains the schematic and partial layout of a basic domino implementation of a 2-input OR gate. Keeper device P2 is a weak PMOS that is used to hold the high state at node N during the evaluation period of the clock, so that N is not discharged by the NMOS leakage currents. The keeper, P2, should be sized large enough to replace the NMOS leakage current and sustain a high voltage at N, but, at the same time, it should be small enough so that the pull-down network can discharge N quickly to minimize the short-circuit current.

Fig. 11 shows two possible layout scenarios for the three PMOS transistors. In one case P2 is located between P1 and P3, while in the other case P1 is in the middle. As shown in Section III, for the two scenarios the drive current for P2 differs by ~8%. This means that the first scenario has higher drive current for keeper P2 than the expected value. As the keeper fights against the pull-down stage, there is a performance loss. HSPICE simulations show that the time taken to discharge node N increases by ~12%. This performance loss can worsen for more aggressively sized cases. For these HSPICE simulations, we approxi-

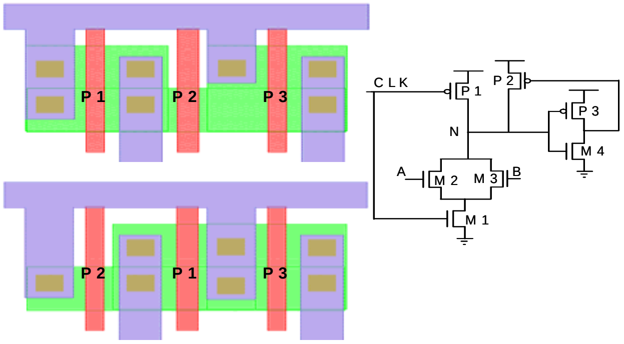


Fig. 11. Basic Domino gate and two possible layouts for the PMOS devices.

TABLE II

SUMMARY OF STRESS-AWARE LAYOUT OPTIMIZATION DRIVE CURRENT IMPROVEMENT AND TRADEOFFS IN 65NM STANDARD CELLS

Cell Name	Percentage drive current improvement by layout optimization		Increase in leakage current by layout optimization		Increase in leakage current for identical drive current improvement by V_{th} reduction		Percentage increase in output capacitance with a FO4 output loading
	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS	
3-input NOR	3%	13.5%	1.22X	4.02X	1.31X	9.20X	2.74%
2-input NOR	3%	7.5%	1.22X	2.24X	1.31X	3.52X	1.92%
3-input NAND	7%	1.5%	1.98X	1.10X	2.36X	1.53X	1.85%
2-input NAND	4.5%	1.5%	1.45X	1.10X	1.68X	1.53X	1.30%
Iso Area INV	3%	1.5%	1.21X	1.10X	1.31X	1.53X	0%
Incr. Area INV	6%	13%	1.86X	3.88X	2.22X	7.04X	2.40%

mated the drive current increase due to stress by changing the relevant mobility numbers in the transistor models.

VI. OPTIMIZATION METHODOLOGY

Stress-based performance enhancement provides a better leakage versus performance tradeoff than V_{th} assignment (as discussed previously in Section II-B). However, when the standard cell area is fixed (i.e., the stress-enhanced version occupies the same/slightly higher amount of area as the original version), we can only obtain limited average drive current improvement through stress-aware layout optimization (<10%). Therefore, we combine stress-optimized assignment with dual-V_{th} assignment to simultaneously achieve a larger range of current improvement and more fine-grained control over the performance enhancement (and, consequently, the increase in leakage). Fig. 12 shows the leakage and switching delays for various combinations of V_{th} and stress-based optimization for a 3-input NOR gate. Low stress (L_{stress}) optimization corresponds to a standard cell in the library that has not been optimized for stress enhancement (by altering the layout properties), while high stress (H_{stress}) optimization corresponds to the layout optimized version of the standard cell. For the dual-V_{th} approach, a gate has only two options to choose from, high-V_{th} (H_{Vth}) or low-V_{th} (L_{Vth}). Introducing stress-based, layout-optimized cells provides an additional reduced leakage option (when performed on a high-V_{th} cell) for gates that require moderate improvements in performance, thereby saving leakage power. Additionally, it also provides a higher performance option when combined with low-V_{th} to further reduce delay.

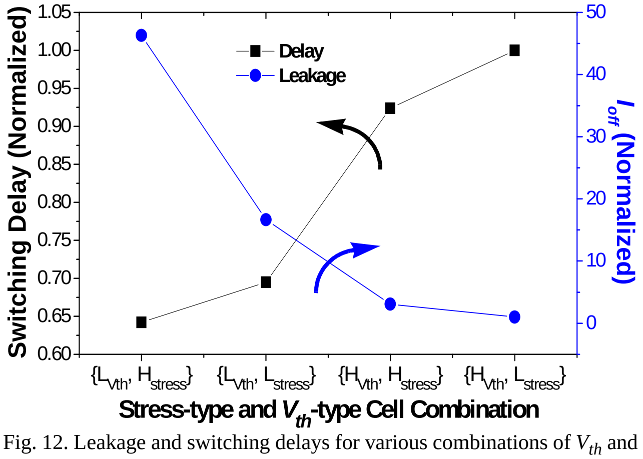


Fig. 12. Leakage and switching delays for various combinations of V_{th} and stress-based optimization for 3-input NOR gate.

For simultaneous V_{th}/stress optimization level selection and sizing optimization, we use an iterative approach similar to [6] that can be divided into two main parts:

1. A certain number of gates in each iteration are assigned to the low-V_{th} or high stress optimization level.
2. The circuit is then rebalanced by reducing the size of the affected gates and other gates are re-sized to compensate for the area reduction (the objective is iso-area).

Initially, all gates are set to their $\{H_{Vth}, L_{stress}\}$ version, to maximize leakage savings. Then, in each iteration, a merit function is evaluated for all gates in a circuit. This merit function rates the increase in total leakage with respect to the performance gain of the circuit. Gates with the highest merit are selected first and set to the next highest performance level. The performance levels for our library are shown in the x-axis of Fig. 12, and, from left to right, are ordered from highest performance (and leakage) to lowest performance (and leakage). This order holds for all standard cells in our library. The merit function is shown below in (3):

$$\text{Merit}(G) = \frac{\Delta I_{off}(G)}{\Delta D(G)}$$
$$\text{where } \Delta D(G) = \sum_{\text{arcs}}^{\alpha} \Delta d_{\alpha}(G) \cdot \frac{1}{k + Slack_{min} - Slack_{\alpha}} \tag{3}$$

Here, $\Delta d_{\alpha}(G)$ is the impact that increased gate performance has on a particular timing arc, α; k is a small negative number; and $Slack_{min}$ is the worst slack seen in the circuit. This weighting function takes the value $1/k$ for timing arcs on the critical paths, and approaches zero for less critical timing arcs.

Once the merit function is evaluated, a circuit's gate sizes are no longer optimal since one or more gates have been assigned to a higher performance level. The resulting decrease in delay creates excess area which can be recovered from the now oversized gates. By shifting this excess area to undersized regions, we can improve performance without increasing area (or only increasing it by a small amount). The candidates for reduction include the modified gate itself along with any gates sharing a timing path with the modified gate. Because modifying a gate has a greater effect on nearby gates, we can identify a modified gate's core of influence to a predetermined logic depth based on the distance of gates (sharing a timing arc with the modified gate) from the changed gate. This depth was experimentally determined to be three levels of logic [6]. For the purpose of resizing, we use a delay-sensitivity-based sizing optimization algorithm [30]. The pseudo code for a given value of target critical delay (T_T) is shown below. Note that Lines 2 and 3 merely provide one set of initial values for T_C and T_N such that the conditions of the while loop are satisfied in the first iteration.

Algorithm 1 STRESS_OPT(T_T) // T_T = Target Delay

```
1:  Set all cells in netlist to {$H_{Vth}$,$L_{stress}$} version
2:  Run Initial STA and baseline sizing
3:  $T_N = T_T + 1$ // $T_N$ = new critical path (CP) delay
4:  $T_C = T_N + \gamma + 1$ // $T_C$ = current CP delay
5:  // $\gamma$ = small constant, checks for >minimal changes in $T_C$
6:  while ( ($T_N > T_T$) and (($T_C$ - $T_N$) > $\gamma$) )
7:      $T_C = T_N$
8:      Evaluate Merit(G) for all gates, G // see (3)
9:      Move gates with highest Merit(G) to next highest
        performance level
10:     Rebalance circuit through sizing
11:     Update STA, find new critical delay, $T_N$
12: end while
```

The next section discusses the experimental results obtained when applying this optimization algorithm to benchmark circuits.

VII. Experimental Setup and Results

The following section describes the library characterization used within our experimental setup, as well as the results obtained from using the proposed optimization scheme on a number of benchmark circuits.

A. *Library Characterization*

To implement our optimization methodology, we first had to characterize our stress-enhanced standard cell library and determine the decrease/increase in propagation-delay/leakage-power, respectively, that the standard cells achieved while exploiting the layout dependencies of stress. The characterization flow is illustrated in Fig. 13 and captures the relative change in propagation delay and leakage power, as compared to the "unstressed" version of a particular standard cell. While characterizing one standard cell, we simulated both the stress-enhanced version and its unstressed counterpart in Tsuprem4 and DaVinci, as discussed in Section III. From these simulations, we were able to calculate the relative increase in I_{on} and I_{off} (referred to as $\Delta I_{on}(X)$ and $\Delta I_{off}(X)$, respectively) for each device, X, within the standard cell. These $\Delta I_{on}(X)$ and $\Delta I_{off}(X)$ values for every PMOS and NMOS device (in every standard cell in our library) were then input directly into the optimization engine. Within the optimization algorithm, $\Delta I_{on}(X)$ is translated to decreasing propagation delay by using an inverse relationship fit: $\Delta d_{\alpha}(X) \propto \frac{1}{\Delta I_{on}(X)}$. Finally, these values, $\Delta d_{\alpha}(X)$ and $\Delta I_{off}(X)$, are used directly in the merit function described in (3).

In order to examine the effect that neighboring cells had on the channel stress of a device, we conducted a simple experiment where the value of I_{on} for a minimum-sized inverter in isolation was compared to the same minimum-sized inverter which had inverters as neighbors on both sides (representing a more "dense" context). We chose the min-sized inverter because of all of the standard cells, it was the most sensitive to changes in context. For the stress-enhanced inverter cell, we observed a 0.8% higher I_{on} and a 2.0% higher I_{off} in the case where neighboring cells were included. However, the corresponding gains in I_{on} and I_{off} (ΔI_{on} and ΔI_{off}) for the stress-enhanced version (compared to the unoptimized version) decreased by <0.1% and <1%, respectively, while considering neighbors. Since the I_{on}/I_{off} gains achieved for stress-enhanced layouts showed little sensitivity to changes in context and because circuit level TCAD simulations were not possible (due to runtime and convergence issues), we used the library characterization of isolated cells to drive the circuit-level analysis in this paper. In the proposed circuit-level optimization (discussed in Section VI), critical cells are iteratively exchanged with their stress-enhanced (or dual-V_{th}) counterparts. While considering the optimization of one particular cell within one iteration, only the type of enhancement is modified. All other parameters like neighborhood, size, and cell type (NAND, NOR, etc.) are held constant.

B. *Experimental Results*

The algorithm described in Section VI was implemented in C and tested on ISCAS85 benchmark circuits, two DSP circuit

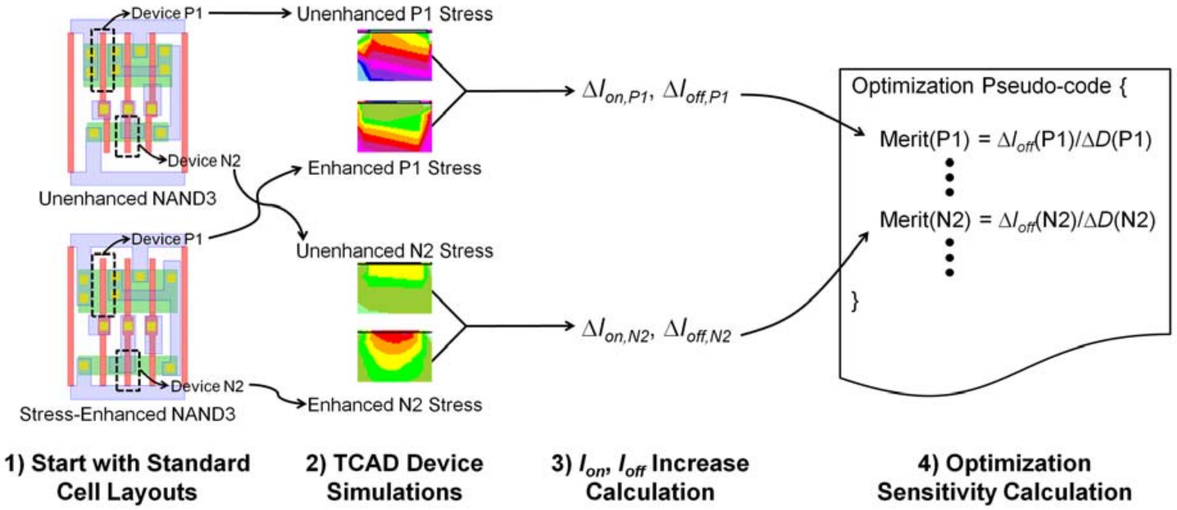

Device P1
Unenhanced P1 Stress
$\Delta I_{on,P1}$, $\Delta I_{off,P1}$
Optimization Pseudo-code {
Merit(P1) = ΔI_{off}(P1)/ΔD(P1)
Merit(N2) = ΔI_{off}(N2)/ΔD(N2)
}
Device N2
Enhanced P1 Stress
Unenhanced NAND3
Device P1
Unenhanced N2 Stress
$\Delta I_{on,N2}$, $\Delta I_{off,N2}$
Device N2
Enhanced N2 Stress
Stress-Enhanced NAND3
1) Start with Standard Cell Layouts
2) TCAD Device Simulations
3) I_{on}, I_{off} Increase Calculation
4) Optimization Sensitivity Calculation

Fig. 13. Stress-enhanced library characterization for stress-aware optimization.

implementations (“Viterbi1” and “Viterbi2”), and a USB 2.0 controller implementation. The benchmarks vary in size from 166 to 37560 gates. The circuits were synthesized using an industrial 65nm CMOS technology with the following specifications:[3]

- $V_{DD,nominal}$ = 1V
- HVT, NMOS V_{th} = 334mV
- HVT, PMOS V_{th} = -391mV
- LVT, NMOS V_{th} = 243mV
- LVT, PMOS V_{th} = -280mV

The resulting spread in I_{on} and I_{off} (between HVT and LVT) was 1.24X/1.32X and 16X/29X, respectively, for NMOS/PMOS transistors. All of the standard cells (both the original and the stress-enhanced versions) in our library were characterized (using HSPICE) at both the high- and low-V_{th} values. The layout-dependent characteristics (e.g., rise/fall delay, rise/fall power, etc.) and parasitics (such as junction capacitance and S/D resistance) for each cell were captured during the HSPICE characterization. All of the improvements discussed in this section use a dual-V_{th} optimization (using simultaneous V_{th} selection and gate sizing) as the basis for comparison.

Fig. 14 shows the leakage power versus critical delay curves for the two techniques: dual-V_{th} assignment and dual-V_{th} assignment combined with stress-aware layout optimization, for one of the larger circuits, c7552. As mentioned earlier, combining stress-based layout optimization with V_{th} assignment provides a better range and more fine-grained control of performance enhancement as compared to the dual-V_{th} based assignment (see Table III for the cell combinations used in each optimization scheme). This is clearly seen in Fig. 14 while comparing both the critical delay for the two techniques at the same value of leakage (iso-leakage), as well as the leakage power at the same value of critical delay (iso-delay). The key metric that we use in our comparisons is known as hardware intensity (η), which was proposed in [31] for quantifying the tradeoff between power and delay of a design. A hardware intensity of *x* means that a 1% decrease in delay leads to an *x*% increase in power. The hardware intensity for the majority of blocks in a microprocessor design is between 2 and 3 [32]. Thus, for a fair evaluation of the proposed approach, we present results for points on the power-delay curve that correspond to a hardware intensity value between 2 and 3. One such point is shown as “P” in the leakage-power-delay tradeoff curve (η = 2) in Fig. 14. For the circuit, c7552, our proposed optimization results in 22% lower leakage power for iso-delay, and 5.4% lower delay for iso-leakage, when compared to dual-V_{th} based assignment at point P.

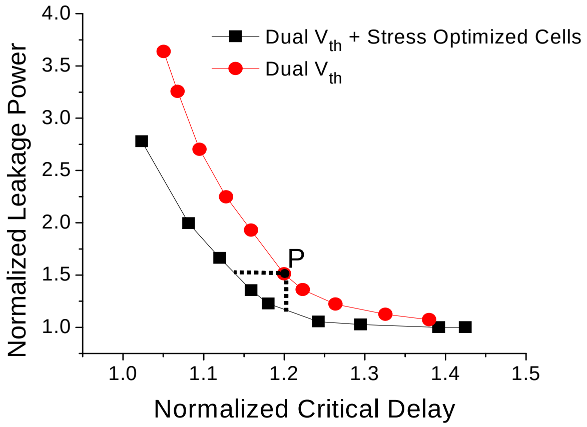


Fig. 14. Leakage power versus delay tradeoff curve for the circuit c7552 for dual-V_{th} and proposed approach.

TABLE III
STRESS AND V_{TH} COMBINATIONS

		Cell Combinations
(1)	Combined stress-enhancement and dual-V_{th}	{L_{Vth}, H_{stress}}, {L_{Vth}, L_{stress}}, {H_{Vth}, H_{stress}}, {H_{Vth}, L_{stress}}
(2)	Only dual-V_{th}	{L_{Vth}, L_{stress}}, {H_{Vth}, L_{stress}}
(3)	Only stress-enhancement	{H_{Vth}, H_{stress}}, {H_{Vth}, L_{stress}}

Fig. 15 shows how the percentage improvement (of our combined method over dual-V_{th}) in leakage power and critical delay, as well as the corresponding area overhead varies with hardware intensity for c7552. Percentage improvement in leakage power increases with increasing hardware intensity because the leakage-power-delay curves for our approach and dual-V_{th} assignment move further apart as delay decreases (or hardware intensity increases). The improvement in critical delay also increases with increasing hardware intensity. The area overhead, however, shows an initial increase as more gates require higher performance, but then becomes fairly constant at higher values of hardware intensity. For the remainder of this section, we report power and delay improvement numbers for points on the leakage-power-delay curves that correspond to a hardware intensity of 2.

Table IV summarizes the improvements seen in two comparisons: 1) combined stress-enhancement and dual-V_{th} (which uses the cell combinations shown in (1) in Table III) versus only dual-V_{th} (see (2) in Table III), and 2) stress-enhancement (see (3) in Table III) versus only dual-V_{th}. The first two columns state the name of the test circuit and its size. The next four columns report the percentage improvement in leakage over the dual-V_{th} case and the corresponding area overhead for iso-delay (for both

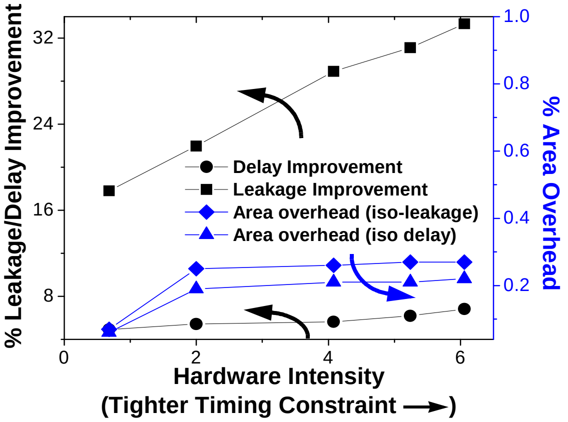


Fig. 15. Delay and power improvement and the corresponding area overhead plotted against hardware intensity.

[3] Reported V_{th} values were obtained using the industry standard “constant current method” [33], where V_{th} is determined by extracting V_{GS} at the point where $|I_{DS}| = 100\text{nA} \cdot \frac{W}{L}$ (with $V_{DS} = V_{DD,nominal}$).

TABLE IV

IMPROVEMENT IN LEAKAGE AND DELAY AS COMPARED TO DUAL-V_{TH} BASED ASSIGNMENT

Circuit	Number of gates	Comparison for iso-delay against only dual-V_{th} assignment				Comparison for iso-leakage against only dual-V_{th} assignment			
		Stress + V_{th} based assignment		Only Stress based assignment		Stress + V_{th} 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
c432	166	38.5%	0.3%	5.4%	0.5%	5.0%	0.5%	3.6%	0.6%
c499	962	20.4%	0.9%	5.1%	0.9%	4.6%	0.9%	3.4%	1.0%
c880	390	33.7%	0.1%	12%	0.2%	5.8%	0.3%	2.3%	0.3%
c1908	432	22.5%	0.6%	7.4%	0.7%	4.7%	0.9%	3.0%	0.9%
c2670	964	14.7%	0.1%	5.1%	0.2%	5.2%	0.3%	3.6%	0.3%
c3540	962	23.9%	0.2%	4.7%	0.3%	4.7%	0.3%	2.5%	0.3%
c5315	1750	22.9%	0.2%	4.9%	0.3%	4.9%	0.2%	2.6%	0.2%
c6288	2470	20.1%	0.9%	5.9%	0.9%	4.6%	0.9%	3.0%	0.9%
c7552	1993	22.0%	0.3%	4.8%	0.2%	5.4%	0.2%	3.1%	0.3%
Viterbi1	14503	21.5%	0.3%	4.9%	0.4%	5.3%	0.3%	2.9%	0.5%
Viterbi2	34082	22.6%	0.3%	5.1%	0.4%	5.2%	0.2%	2.7%	0.4%
USB	37560	22.4%	0.3%	5.2%	0.3%	5.2%	0.4%	2.8%	0.3%
Average		**23.8%**	**0.4%**	**5.9%**	**0.4%**	**5.1%**	**0.5%**	**3.0%**	**0.5%**

comparisons). The last four columns show the percentage improvement in critical delay and the corresponding area overhead for iso-leakage-power (for both comparisons). The small value of area overhead occurs because of the increased area variants of the layout-optimized inverter cells (mentioned in Section V).

The results clearly show that our combined approach significantly improves the leakage power for iso-delay, and also improves critical delay for iso-leakage, when compared to dual-V_{th} based assignment. We get up to a 38.5% (23.8% on average) improvement in leakage for iso-delay, and up to a 5.8% (5.1% on average) improvement in delay for iso-leakage. The area overhead is very small for both the cases – less than 0.5% on average across all 12 circuits. It is worth noting that while our delay improvements are similar to those published in [18], our proposed technique provides the 5.1% delay improvement (on average) for *iso-leakage*.

As mentioned previously, Table IV also includes a one-to-one comparison of stress-enhancement versus dual-V_{th}, where stress-enhancement achieves up to a 7.4% (5.9% on average) improvement in leakage for iso-delay, and up to a 3.6% (3% on average) improvement in delay for iso-leakage (compared to dual-V_{th}). The discrepancy between the leakage improvement of the combined approach (stress + dual-V_{th}) versus dual-V_{th} (23.8% on average) compared to only stress-enhancement versus dual-V_{th} (5.9% on average) arises because the point on the stress-enhancement leakage/delay curve where hardware intensity equals 2 ($\eta = 2$) occurs at a larger delay (e.g., a point to the right of P in Fig. 14). This is explained by the fact that stress-enhancement alone can only achieve $<1/2$ of the performance enhancement of dual-V_{th}. Thus, the leakage comparison between stress-enhancement and dual-V_{th} occurs in the region of leakage-versus-delay where stress does not have as large of an advantage over dual-V_{th} (note the smaller gap between the two curves in Fig. 14 as you move towards larger delays). However, at the new comparison point, for this framework and technology, stress-enhancement still outperforms dual-V_{th} both in leakage optimization as well as delay optimization. This is noteworthy because using stress-enhancement by itself eliminates the extra masks and processing steps required by dual-V_{th} designs, which reduces process complexity and cost. Furthermore, the stress-enhancement versus dual-V_{th} improvement numbers are limited by the fact that we require small or no area overhead for the redesigned standard cells. Using more advanced techniques, we could further improve the stress-enhanced tradeoff between area and performance, which will increase the performance gap between stress-enhancement and dual-V_{th}.

Fig. 16 shows the percentage of gates assigned to low-V_{th} for the dual-V_{th} assignment, as well as the combined "stress enhancement + dual-V_{th}" approach. These numbers are reported for iso-delay points on the leakage-delay curves corresponding to a hardware intensity of 2. As expected, for the combined approach, a lesser number of gates are assigned to low-V_{th} as compared to dual-V_{th} assignment. This is because for the dual-V_{th} assignment, not all gates assigned to low-V_{th} need such a large performance improvement. Combining stress-optimized

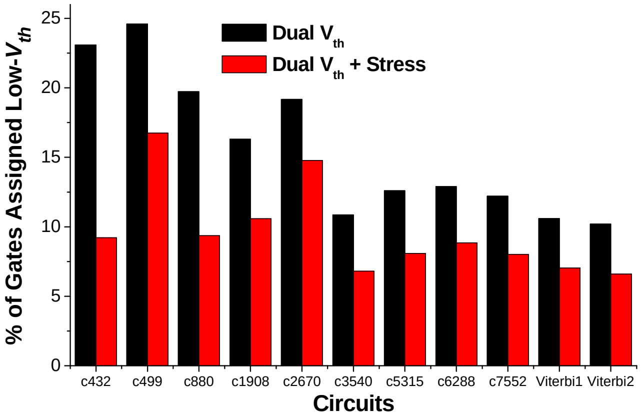


Fig. 16. Percentage of gates assigned to low-V_{th} for dual-V_{th} and the combined dual-V_{th} and stress based approach.

cell assignment with dual-V_{th} assignment provides an additional lower leakage option for the cells that require moderate improvements. This reduces the number of cells that are assigned to low-V_{th}, which, in turn, results in lower leakage current. Typically, the number of gates assigned to low-V_{th} for the combined approach is ~35% lower than the number for dual-V_{th} assignment.

To further investigate the tradeoff that exists between leakage power savings and area overhead, we performed another experiment using a richer library comprised of higher area, stress-enhanced versions of all the cells. The area overhead for the higher area versions was ~20% per cell, and every cell in the richer library had three variants: an original unoptimized version; an iso-area, stress-enhanced version; and an increased area, stress-enhanced version. The richer library provided more intermediate, low-leakage options (in addition to the low-V_{th} cell) for gates requiring moderate improvements. By providing these intermediate performance alternatives, the overall leakage power (for iso-delay) is further reduced as compared to dual-V_{th} assignment. Fig. 17 shows the comparison between the "stress-enhancement + dual-V_{th} assignment" optimization for the richer library and the original, stress-optimized library (with increased area versions for inverters only). It plots the leakage power improvement (for iso-delay) and the corresponding area overhead obtained by using the richer library (compared to the original stress-enhanced library) for six of the larger circuits. On average, using the richer library further improved the leakage power (at iso-delay) by ~12% for an area overhead of ~1% over joint assignment using the original library. This experiment shows that there is scope for further improvement using the richer library. However, the richer library also incurs a higher characterization cost due to the large number of variants for each cell. One approach to minimize this cost would be to only create multiple versions of cells that are used most often (typically the smaller gates such as inverters, NAND's, NOR's, etc.).

VIII. Conclusions

In this paper, we explored the modification of standard cell layouts in order to optimize the stress-based performance enhancement, and proposed a block-based optimization algorithm that combined stress-enhancement with dual-V_{th} assignment to achieve performance gains in leakage or delay. We studied the dependence of drive current improvement on layout parameters like source/drain length and contact placement, and found that the performance of any given layout could be enhanced by increasing the active area length. Based on our observations, we exploited a set of layout properties which maximized the performance improvement of a standard cell without increasing area. When these properties were modified in standard cells from a 65nm industrial library, PMOS and NMOS drive currents attained an average performance enhancement of 6% and 4.4%, respectively, without increasing the cell area. The corresponding average increase in leakage was found to be 2.2X and 1.5X for PMOS and NMOS devices, respectively. Next, we combined the assignment of these stress-optimized cells with V_{th} assignment in order to optimally tradeoff leakage power and performance. When compared to the traditional dual-V_{th} based assignment technique, the new approach reduced leakage current by 23.8% on average for identical delay, and improved critical delay by 5.1% on average for identical leakage, with a very small area overhead (<0.5%).

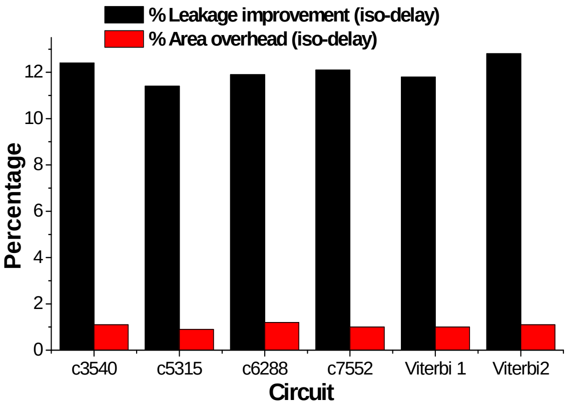


Fig. 17. Delay and power improvement and the corresponding area overhead for the richer library over the original library.

References

[1] F. Andrieu et al., "Experimental and Comparative Investigation of Low and High Field Transport in Substrate- and Process-Induced Strained Nanoscale MOSFETs," *Proc. VLSI Technol. Symp. Tech. Dig.*, pp. 176-177, June 2005.

[2] K. Mistry et al., "Delaying Forever: Uniaxial Strained Silicon Transistors in a 90nm CMOS Technology," *Proc. VLSI Technol. Symp. Tech. Dig.*, pp. 50-51, June 2004.

[3] V. Chan et al., "Strain for CMOS performance Improvement," *Proc. CICC*, pp. 667-674, Sept. 2005.

[4] S. Wolf, "Silicon Processing for the VLSI Era," Lattice Press, 1995.

[5] A. Chandrakasan et al., "Design of High-Performance Microprocessor Circuits," IEEE press, 2001.

[6] S. Sirichotiyakul et al., "Duet: An Accurate Leakage Estimation and Optimization Tool for Dual-V_t Circuits," *IEEE Trans. on VLSI Systems*, Vol. 10, No. 2, pp. 79-90, April 2002.

[7] L. Wei et al., "Design and optimization of low voltage high performance dual threshold CMOS circuits," in *Proc. 35th Design Automation Conference*, pp. 489-494, June 1998.

[8] D. Sylvester and A. Srivastava, "Computer-Aided Design for Low-Power Robust Computing in Nanoscale CMOS,", in *Proc. of the IEEE*, Vol. 95, pp. 507-529, March 2007.

[9] R. A. Bianchi, G. Bouche, and O. Roux-dit-Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *Proc. of IEDM*, pp. 117-120, 2002.

[10]A. Kahng, P. Sharma, and R.O. Topaloglu, "Chip Optimization Through STI-Stress-Aware Placement Perturbations and Fill Insertion," in *IEEE Trans. on CAD*, Vol. 72, pp. 1241-1252, July 2008.

[11] K. Su et al., "A Scaleable Model for STI Mechanical Stress Effect on Layout Dependence of MOS Electrical Characteristics," in *Proc. of CICC*, pp. 245-248, Sept. 2003.

[12]K. Yamada, et al., "Layout-Aware Compact Model of MOSFET Characteristics Variations Induced by STI Stress," in *IEICE Trans. on Elect*, Vol. E91-C, No. 7, pp. 1142-1150, July 2008.

[13]V. Moroz et al., "The Impact of Layout on Stress-Enhanced Transistor Performance," in *Proc. SISPAD*, pp. 143-146, Sept. 2005.

[14]Y.M. Sheu et al., "Modeling Mechanical Stress Effect on Dopant Diffusion in Scaled MOSFETs," in *IEEE Trans. on Electron Devices*, Vol. 52, pp. 30-38, Jan. 2005.

[15] M. V. Dunga et al., "Modeling Advanced FET Technology in a Compact Model," in *IEEE Trans. on Elect. Dev.*, Vol. 53, pp. 1971-1978, Sept. 2006.

[16] G. Eneman et al., "Layout Impact on the Performance of a Locally Strained PMOSFET," in *Proc. of Symp. on VLSI Technology*, pp. 22-23, June 2005.

[17]L. T. Pang et al., "Measurement and Analysis of Variability in 45 nm Strained-Si CMOS Technology," in *IEEE Journal of Solid-State Circuits*, Vol. 44, pp. 2233-2243, Aug. 2009.

[18]A. Chakraborty, S. Shi, and D. Pan, "Layout Level Timing Optimization by Leveraging Active Area Dependent Mobility of Strained-Silicon Devices," in *Proc. of DATE*, pp. 849-855, March 2008.

[19]W. H. Lee et al., “High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-K BEOL,” in *Proc. IEDM*, Dec. 2005.

[20] G. Scott et al., “NMOS Drive Current Reduction Caused by Transistor Layout and Trench Isolation Induced Stress,” in *Proc. of IEDM*, pp. 827-830, 1999.

[21]Z. Luo et al., “Design of high performance PFETs with strained si channel and laser anneal,” in *Proc. of IEDM*, pp. 489-492, Dec. 2005.

[22]H. S. Yang et al., “Dual stress liner for high performance sub-45nm gate length SOI CMOS manufacturing,” in *Proc. of IEDM*, pp. 1075-1077, Dec. 2004.

[23]K. Ota et al., “Novel locally strained channel technique for high performance 55nm CMOS,” in *Proc. of IEDM*, pp. 27-30, 2002.

[24]A. Eiho et al.,”Management of Power and Performance with Stress Memorization Technique for 45nm CMOS,” in Proc. IEEE Symposium on VLSI Technology, pp. 218-219, June 2007.

[25]Manual, Davinci 3D TCAD, Version 2005.10.

[26]Manual, Synopsys TSUPREM4, Version 2007.03.

[27]T. B. Hook et al, “Lateral Ion Implant Straggle and Mask Proximity Effect,” in *IEEE Trans. on Electron Devices*, Vol.50, pp.1946-1951, Sept. 2003.

[28]Y.M. Sheu et al., “Modeling the Well-Edge Proximity Effect in Highly Scaled MOSFETs,” in *IEEE Trans. on Electron Devices*, Vol. 53, pp. 2792-2798, Nov. 2006.

[29]Manual, BSIM4 Spice Model, Version 4.6.1, pp. 115-116.

[30]A. Dharchoudhury et al., “Transistor-level sizing and timing verification of domino circuits in the powerPC™ microprocessor,” in *Proc. ICCD*, pp. 143-148, Oct. 1997.

[31]V. Zyuban et al., “Unified Methodology for Resolving Power-Performance Tradeoffs at the Microarchitectural and Circuit Levels,” in *Proc. ISLPED*, pp. 166-171, Aug. 2002.

[32]S. Burns et al., “Comparative Analysis of Conventional and Statistical Design Techniques,” in *Proc. 44th Design Automation Conference*, pp 238-243, June 2007.

[33]T. Hori, “Gate Dielectrics and MOS ULSI,” New York: Springer-Verlag, 1997.

[34]V. Joshi et al., “Stress Aware Layout Optimization,” in *Proc. ISPD 2008*, pp. 168-174, April 2008.

[35]V. Joshi et al., “Leakage Power Reduction Using Stress-Enhanced Layouts,” in *Proc. 45th Design Automation Conference*, pp. 912-917, June 2008.

Vivek Joshi received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India in 2006. He is currently working toward his Ph.D in electrical engineering at University of Michigan, Ann Arbor.

He was an intern at IBM Austin Research Lab from Oct2008 - July2009, and at Mentor Graphics, Wilsonsonville, during the summer of 2008. His research interests include design manufacturing interface, variation tolerant circuits and systems, and modeling and simulation of systematic sources of variation.

Brian Cline received the B.S. degree in electrical engineering from the University of Texas at Austin in 2004. He received the M.S. degree in electrical engineering from the University of Michigan in 2006, where he is currently working toward his Ph.D.

His research interests include low-power circuit design, variation-aware CAD tool development, and VLSI design optimization for high-performance and low-power designs.

Dennis Sylvester (S '95, M '00, SM '04) received a Ph.D. in electricalengineering from the University of California, Berkeley where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS department.

He is an Associate Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. He previously held research staff positions in the Advanced Technology Group of Synopsys, Mountain View, CA, Hewlett-Packard Laboratories in Palo Alto, CA, and a visiting professorship in Electrical and Computer Engineering at the National University of Singapore. He has published over 250 articles along with one book and several book chapters in his field of research, which includes low-power circuit design and design automation techniques, design-for-manufacturability, and interconnect modeling. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas.

Dr. Sylvester received an NSF CAREER award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and numerous best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He is currently an Associate Editor for IEEE Transactions on CAD and previously served as Associate Editor for IEEE Transactions on VLSI Systems. He is a member of ACM and Eta Kappa Nu.

David Blaauw received his B.S. in Physics and Computer Science fromDuke University in 1986, and his Ph.D. in Computer Science from the University of Illinois, Urbana, in 1991. Until August 2001, he worked for Motorola, Inc. in Austin, TX, were he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan where he is a Professor. His work has focussed on VLSI design with particular emphasis on ultra low power and high performance design. He was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.

Kanak Agarwal (S’01, M’05) received the B.E. degree in electrical engineering from the Birla Institute of Technology and Science, Pilani, India, in 2000 and the M.S. and Ph.D. degrees in electrical engineering from University of Michigan, Ann Arbor, in 2003 and 2004, respectively. His dissertation work focused on various nanometer-design issues such as leakage-power estimation and minimization, impact of process variation on performance and power, and on-chip interconnect modeling. Currently, he is with the IBM Research Lab, Austin, TX. His current research is focused on high-speed and low-power circuit design, and statistical characterization and modeling of process variations. He is also interested in exploring novel device structures and circuit families that can advance the life of scaling beyond conventional CMOS..