Leakage Power Reduction Using Stress-Enhanced Layouts

Vivek Joshi, Brian Cline, Dennis Sylvester, David Blaauw, Kanak Agarwal*
University of Michigan, Ann Arbor, MI. email: {vivekj, btcline, dennis, blaauw}@eecs.umich.edu
*IBM Research, Austin, TX. email: kba@us.ibm.com

Abstract – In recent years, process-induced mechanical stress has emerged as a useful manufacturing technique that enhances carrier transport and increases drive currents. This improvement in current has helped to compensate the decline of device scaling factors in parameters such as t_{ox}, V_{th}, and V_{dd}. In this work, we propose stress as a means to achieve optimal power-performance trade-off by combining stress-based, performance-enhanced standard cell assignment with dual-V_{th} assignment. We study how stress-induced performance enhancements are affected by layout properties and improve standard cell layouts so that performance gains are maximized. We then develop a circuit-level, block-based, stress-enhanced optimization algorithm that includes all layout-dependent sources of mechanical stress. By combining the two performance enhancement techniques (stress-based and dual-V_{th}) for a set of benchmark circuits, we find that our stress-aware optimization, decreases leakage by ~24% on average, for iso-delay, when compared to dual-V_{th} assignment. Similarly, for iso-leakage, our optimization algorithm reduces delay on average by 5%. In both cases, the proposed method only incurs a small area penalty (< 0.5%).

Categories and Subject Descriptors: B.8.0 Performance and Reliability (General)

General Terms: Performance

Keywords: Stress, mobility, layout, leakage, performance.

1. INTRODUCTION

Maintaining integrated circuit (IC) performance and reliability in modern-day semiconductor processes, while continuing aggressive process scaling, is becoming increasingly difficult because of fundamental scaling limitations. Device parameters like oxide thickness (t_{ox}), threshold voltage (V_{th}), and supply voltage (V_{dd}) can no longer be scaled as aggressively as gate length (L) without significantly degrading reliability and exponentially increasing leakage current. Furthermore, as MOSFET's scaled below 100nm, process engineers sought to battle the mobility degradation caused by larger effective electric fields. To ameliorate mobility degradation and the subsequent drive current reduction, several process techniques have been developed which induce mechanical stress in a device's channel. By introducing additional mechanical stress in the channel of a device, one can enhance carrier mobility and achieve higher drive currents. Mobility enhancement has emerged as an attractive alternative to voltage and oxide thickness scaling because it can obtain similar device performance improvement, with reduced effects on reliability and leakage.

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. Since changes in mobility directly influence the drive current, higher carrier mobility improves transistor performance. However, increased mobility not only improves the drain current in the saturation regime of MOSFET operation, but it also increases the subthreshold leakage 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 [3, 4]. It is the relationship of drain current and mobility, in fact, that makes mobility enhancement an interesting alternative to other power/delay optimization techniques. One of the most popular techniques that has been researched considerably in both academia and industry is the dual-V_{th} optimization scheme [5, 6]. This scheme typically uses gate sizing and 2 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 [7].

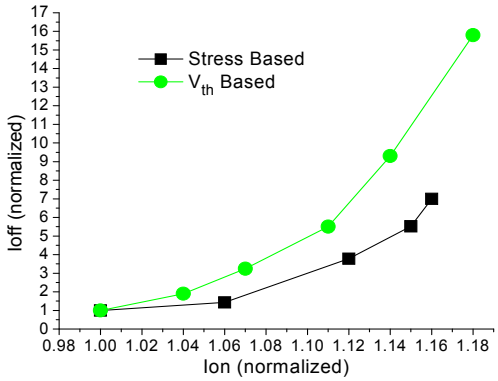
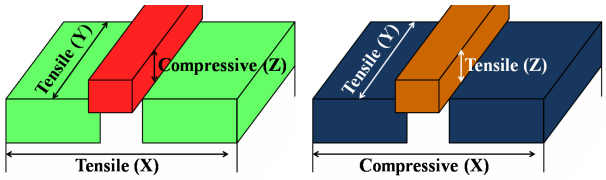


Figure 1. I_{off} versus I_{on} curves for V_{th} assignment and stress-based performance enhancement for a 65nm PMOS device.

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 $I_{D,sub}$), but the approach has only coarse granularity in its selection. Mobility enhancement induced by mechanical stress, however, is layout dependent and provides much finer delay-versus-leakage control (discussed in more depth in Section 2). 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. To illustrate the strengths of both approaches, Figure 1 shows I_{off} vs. I_{on} for an isolated PMOS device across typical values of stress and V_{th}. From this plot, it is clear that stress-based performance enhancement has a better I_{off} vs. I_{on} trade-off while V_{th} has a larger range for both I_{on} and I_{off}. For a 12% improvement in I_{on}, the leakage for the V_{th} case is nearly twice as large as that for the stress-based improvement, and the difference is amplified at higher values of improvement. However, for a given layout, the maximum stress-based enhancement that can be achieved by altering the layout parameters is typically ~20% (for I_{on}).

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) [8, 9] or limited their analysis to only include the PMOS sources of mechanical stress [10, 11]. In [12], the authors explore the effects of a number of different mechanical stress sources and suggest ways to improve standard cell layouts. However, to our knowledge this paper is the first

DAC 2008, June 8–13, 2008, Anaheim, California, USA



X = Longitudinal, Y = Lateral, Z = Si-Depth

(a) NMOS Device **(b) PMOS Device**

Figure 2. Desired stress types for NMOS and PMOS devices.

work to propose a circuit-level, block-based, stress-enhanced optimization scheme that includes all layout-dependent sources of stress.

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 develop general layout rules that serve as guidelines for optimizing transistor performance. Next, these guidelines are used on an industrial 65nm CMOS library, and a performance enhanced version of each standard cell is created. 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}. Experimental results show that we can obtain a 12% performance enhancement 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 standard cells in our library, we find that leakage is reduced by ~2X on average as compared to the V_{th} counterpart. Overall, by combining the two performance enhancement techniques (stress-based and dual-V_{th}) for a few benchmark circuits, we find that our stress-aware optimization, for iso-delay, decreases leakage on average by ~24% 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%. In both cases, our proposed method only incurs a small area penalty (< 0.5%).

The rest of the paper is organized as follows. Background information on mechanical stress in Silicon is included in Section 2. Section 3 presents a study on the layout dependence of stress-based performance enhancement, develops simple guidelines for improving layouts, and discusses the actual improvements seen in an industrial 65nm library. We then propose our stress-aware optimization in Section 4 and present experimental results in Section 5, comparing our approaches with the dual-V_{th} technique. Section 6 concludes the paper.

2. BACKGROUND

As stated previously, mechanical stress in Silicon leads to band splitting and alters the effective mass, which results in carrier mobility changes [1, 2]. Induced stress in the channel can be either tensile or compressive, and NMOS and PMOS devices prefer different stress types in the X, Y, and Z dimensions (shown in Figure 2) [13]. If the correct type of stress is applied in a particular dimension, carrier mobility will increase and the device will consequently produce higher drive current. Mechanical stress, itself, is actually caused by two types of mismatch between material properties. The first, thermal mismatch, is caused by differences in the thermal expansion coefficient of two materials. The second, lattice mismatch, is created when two materials have different lattice constants.

For one of the latest 65nm CMOS technologies, there are five dominant sources of stress (four of which are illustrated in Figure 3): Shallow Trench Isolation (STI), tensile nitride, compressive nitride, Stress Memorization Technique (SMT), and embedded SiGe [14]. The first source, which impacts channel stress in both NMOS and PMOS devices, is Shallow Trench Isolation (STI). STI creates compressive stress longitudinally and laterally (X and Y dimensions) due to thermal mismatch. Of the four sources shown in Figure 3, STI is the only source that is not inherently used to enhance transistor speed. The second and third sources of stress are the tensile and compressive nitride liners, present in only NMOS or PMOS devices, respectively. These liners are permanently deposited over the transistors and they transfer mechanical stress to the channel through the active area and the polysilicon gate [15]. 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 compressively stressed liners into a single, high performance CMOS flow, called the Dual Stress Liner approach. 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. In addition to the permanent tensile liner shown in Figure 3, the fourth stress source, called SMT, is also used to increase the stress in n-Type MOSFET's [16]. 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 annealing and is "memorized" during the re-crystallization of the active area and gate polysilicon. The final source of stress is used to enhance only PMOS devices, and involves expitaxially growing SiGe in cavities that have been etched into the source/drain (S/D) regions of the transistor. Lattice mismatch between the Si and embedded SiGe creates a large compressive stress in the PMOS channel, which results in significant hole mobility improvement. In the embedded SiGe process, NMOS devices are protected by a capping layer that prevents Si recess and SiGe epitaxial growth.

By closely examining these sources of mechanical stress, the layout dependency of the amount of stress transferred to the channel (and, consequently, the drive current enhancement) becomes apparent. For example, the quantity of epitaxial SiGe (and, hence, the stress) that can be grown depends on the length of the S/D regions. Longer diffusion regions will not only increase the stress due to SiGe, but will also move the STI farther away from the channel, reducing its impact on total channel stress. Thus, when mechanical stress is introduced within a device, the drive current is no longer solely dependent upon gate length and width and can be affected both by the layout of the particular transistor, as well as its neighbors. This means that the performance of two transistors with identical gate lengths and widths can differ significantly, depending on their layouts. The layout dependency of mechanical stress is studied further in Section 3.

3. STRESS-AWARE LAYOUT OPTIMIZATION FOR PERFORMANCE ENHANCEMENT

As mentioned in Section 1 and illustrated in Figure 1, mobility enhancement through induced mechanical stress is a viable alternative to V_{th} modification. In this section, we begin by exploring how mechanical stress in CMOS technologies is influenced by layout parameters (i.e., source/drain extension, distance to STI, etc.). From the intuition developed in this study, we then propose standard cell layout guidelines for maximizing stress-induced performance gains. Finally, we implement these guidelines in an industrial 65nm standard cell library and characterize the performance enhanced gates.

3.1. Stress Effects in CMOS Devices

In order to study the layout dependence of stress-based performance enhancement, we used the Davinci 3D TCAD tool [18], which has an extensive set of stress related features. Additionally, we followed the layout design rules from an industrial 65nm CMOS library and device fabrication was simulated in Tsuprem4 [19] (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 stress were found to be consistent with previously published 65nm tech-

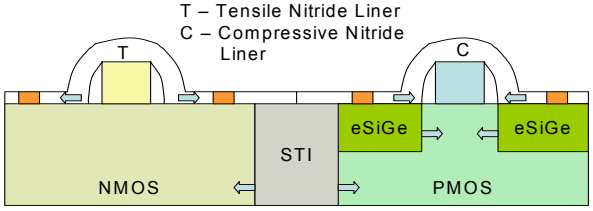


Figure 3. Sources of stress for NMOS and PMOS devices.

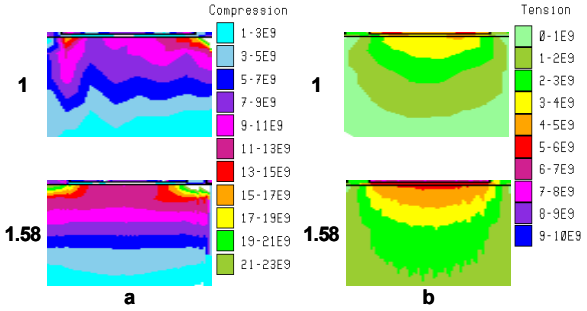


Figure 4. Longitudinal stress component S_{xx} (in Pascals) for normalized $L_{s/d}$ of 1 and 1.58 for a) PMOS b) NMOS.

nology data [14]. 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, embedded SiGe source/drain, 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. SMT, therefore, does not depend on layout properties and can be accurately treated as a uniform increase in NMOS drive current, independent of layout.

For an isolated PMOS device, we increase the active area length ($L_{s/d}$) and examine the corresponding changes in drive current. Increasing active area length has a number of effects: 1) it increases the amount of SiGe, 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; 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.

Figure 4 shows the longitudinal stress (S_{xx}) in isolated PMOS and NMOS devices for two normalized $L_{s/d}$ values of 1 and 1.58. The $L_{s/d}$ values have been normalized to the minimum possible S/D length for a region that contains a contact, in accordance with the layout design rules of the industrial library. Figure 5 shows the drive current, I_{on}, and leakage current, I_{off}, plotted against the S/D length, $L_{s/d}$, for both an NMOS and a PMOS device. Results show that for a 12% and 5% performance increase, leakage current only increases by 3.8X and 1.5X in the PMOS and NMOS device, respectively. This I_{on} versus I_{off} tradeoff, as we have already shown in Section 1, is much better when compared to the enhancement technique where V_{th} is reduced. As mentioned previously, the performance enhancement is also sensitive to contact placement. The experimental results show that about 65% of stress is transferred through the gate and the rest is transferred through the active area. Moving the contacts away from the channel accounts for ~20% of the PMOS drive current improvement and ~80% of the NMOS improvement.

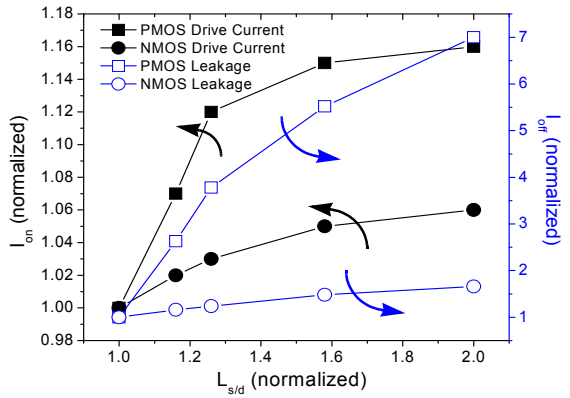


Figure 5. I_{off} and I_{on} versus $L_{s/d}$ curves for stress-based performance enhancement in isolated PMOS, NMOS devices.

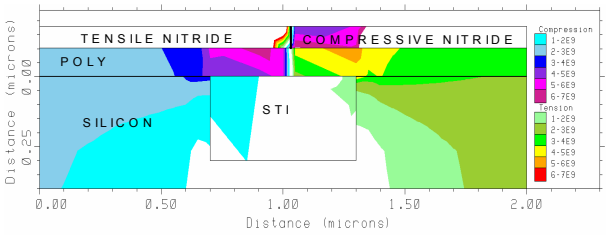


Figure 6. Stress at the Compressive/Tensile Nitride Interface.

3.2. Layout Guidelines for Including Stress

Based on the intuition developed in the previous section, some simple guidelines can be formulated to optimize a given layout for stress-induced performance enhancement. These guidelines should focus on the layout parameters that a designer can modify. 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) are the three main layout parameters, apart from W and L, available to a designer which can be used to modify stress and change current. Using these three parameters, we propose the following layout guidelines for improving the performance of devices under stress.

Guideline 1: ***Increase the active area width to fill up the entire cell.***
This guideline is most readily applied to a gate with NMOS or PMOS stacks that do not use the full width of a cell. The layout, in this case, does not require contacts between intermediate nodes, so their spacing can be significantly tighter than nodes that contain contacts (due to technology design rules). In the absence of stressors, compact layouts are preferred in order to minimize source/drain parasitic capacitance. However, in the presence of stress, the larger capacitances caused by increased active area are dominated by the resulting increase in drive current.

Guideline 2: ***Move the contacts away from gate polysilicon.***
Moving the contacts away from the channel allows more longitudinal stress to be transferred by the nitride layer. In the case of an isolated transistor, move the contact as close to the active edge as design rules permit. For contacts between two gates, place them halfway for similar performance or move the contact closer to the non-critical transistor so that the critical device is stressed more. Moving contacts away from the gate will result in a small increase in the source/drain resistance (< 5Ω), which is outweighed by the resulting gain in drive current.

Guideline 3: ***Laterally, move PMOS/NMOS closer/farther from the nitride interface, respectively.***
Figure 2 shows that the desired lateral stress for both NMOS and PMOS is tensile. However, Figure 6 shows that the lateral stress behavior near the interface of the two nitrides is compressive under the tensile nitride (NMOS side) and tensile under the compressive nitride (PMOS side). Therefore, if possible, it would be beneficial to move the PMOS active area into this region of tensile stress and the NMOS away from the region of compressive stress.

3.3. Layout Guideline Application

As stated previously, we applied the layout guidelines proposed in Section 3.2 to an industrial 65nm standard cell library. The original library contains CMOS logic gates that were designed using present-day layout methodologies that do not consider stress. Hence, most of the cell layouts do not use the full active area available in the cell (especially for series devices), nor do they account for device placement within the cell. Therefore, we can easily apply Guidelines 1, 2, and 3 to the majority of the standard cells in the library without increasing standard cell area. An example of one of the stress-enhanced cells is shown in Figure 7. This particular logic gate is a 3-input NOR gate; our stress-enhanced version is shown in Figure 7a while the original version in

Table 1. Stress-aware layout optimization effects in 65nm standard cells (relative to original, compact layout).

Cell Name	Drive Current Increase		Leakage Increase		Output cap. increase (w/ F04 load)
	NMOS	PMOS	NMOS	PMOS	
Iso-area INV	3%	1.5%	1.2X	1.1X	0%
Incr-area INV	6%	13%	1.9X	3.9X	2.4%
2-input NAND	4.5%	1.5%	1.5X	1.1X	1.3%
3-input NAND	7%	1.5%	2.0X	1.1X	1.9%
2-input NOR	3%	7.5%	1.2X	2.2X	1.9%
3-input NOR	3%	13.5%	1.2X	4.0X	2.7%

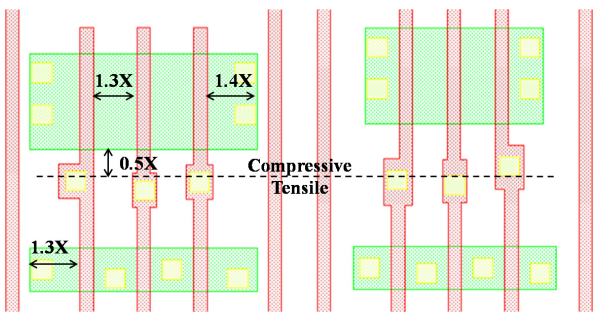
shown in Figure 7b. From this figure, the impact of using our guidelines is clearly seen. After following Guideline 1, the middle S/D lengths for the stacked PMOS devices have increased by 1.3X, while the edge $L_{s/d}$ values have increased by 1.4X and 1.3X for the PMOS and NMOS devices, respectively. One feature of Guideline 1 is the fact that it allows more space to follow Guideline 2. Finally, we were able to laterally move all of the PMOS devices as close to the nitride interface as the design rules allowed, decreasing the distance by 0.5X. After applying the guidelines to the 3-input NOR standard cell, PMOS and NMOS drive current increased by 13.5% and 3%, respectively, while leakage current increased by 4.0X and 1.2X, respectively. On the other hand, when the original cell was used with reduced V_{th} (so that the drive current matched the stress-enhanced case), leakage increased by 9.2X for the PMOS devices and 1.3X for the NMOS devices. Meanwhile, the percentage increase in output capacitance for the stress-optimized layout (with the gate loaded by one FO4) was only 2.7%.

In this same manner, we applied all of the guidelines from Section 3.2 to ~25 standard cells from the 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 apply Guideline 1. Therefore, we chose to create slightly larger stress-enhanced inverters (with ~20% increase in area 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 2).

The stress-enhanced standard cell library is comprised of different sized inverters (iso-area and increased area versions) as well as two- and three-input NAND and NOR gates of varying strengths. Table 1 summarizes the drive current improvements and leakage increase along with the percentage increase in output capacitance (with the gate loaded by one FO4) for the layout optimized versions of the lowest strength gates of each type. When reducing V_{th} in the original cells to match the improvement of the stress-enhanced versions, ~2X larger leakage is observed.

4. Optimization Methodology

Stress-based performance enhancement provides a better leakage versus performance trade-off as compared to V_{th} assignment. However, when the standard cell area is fixed (i.e., the stress-enhanced version occupies the same amount of area as the original version), we can only obtain limited drive current improvement (< 10%) through stress-aware layout optimization. Therefore, we combine the two approaches to simultaneously achieve a larger range of current improvement and more fine-grained control over the performance enhancement (and, consequently, the increase in leakage). Figure 8 shows the leakage and switching delays for various combinations of V_{th} and stress-based optimization for a 3-input NOR gate (L – low, H – high). Low (L) stress optimization corresponds to a standard cell in the library that has not been optimized for stress enhancement (by applying the layout guidelines), while high (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} or low-V_{th}. However, not all gates assigned to low-V_{th} need such a large performance improvement. 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. Alternatively, it also provides a higher performance option when combined with low-V_{th} to further reduce delay.



(a) Stress-Enhanced (b) Original

Figure 7. 3-input NOR Stress Optimization.

For simultaneous V_{th}/stress optimization level selection and sizing optimization, we use an iterative approach similar to [5] 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).

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, and the gates with the highest merit are selected and set to the next higher performance level (lower V_{th} or higher stress optimization level). The performance level is increased from high-V_{th} and no stress optimization to low-V_{th} and high stress optimization in order of increasing leakage. The merit function is shown in (1):

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

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.

However, 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

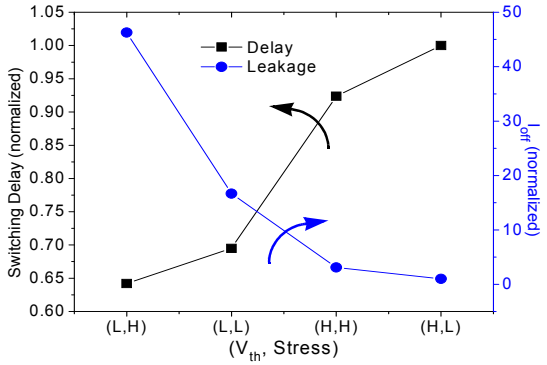


Figure 8. Leakage and switching delays for various combinations of V_{th} and stress-based optimization for 3-input NOR gate.

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 any gate sharing a timing path with a modified gate, as well as the modified gate itself. 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. For the purpose of resizing, we use a delay sensitivity based sizing optimization algorithm [20]. The next section discusses the experimental results obtained when applying this optimization algorithm to benchmark circuits.

5. Experimental Results

The algorithm described in Section 4 was implemented in C and tested on ISCAS85 benchmark circuits and two DSP circuit implementations ("Viterbi1" and "Viterbi2"), that vary in size from 166 to 34082 gates. The circuits were synthesized using an industrial 65nm CMOS technology with a nominal V_{dd} of 1V, $V_{th0,n} = 290\text{mV}$ and $V_{th0,p} = -300\text{mV}$ for the high-V_{th} devices, and $V_{th0,n} = 140\text{mV}$ and $V_{th0,p} = -110\text{mV}$ for the low-V_{th} devices. All of the standard cells (both the original and the stress-enhanced versions) in our library were characterized (using SPICE) 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 source/drain resistance) for each cell were captured during the SPICE-based 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.

Figure 9 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. This is clearly seen in Figure 9 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 [21] for quantifying the trade-off 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 [22]. 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 trade-off curve ($\eta = 2$) in Figure 9. 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.

Figure 10 shows how the percentage improvement (of our combined method over dual-V_{th}) in leakage power and critical delay, and 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 for 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 2 summarizes the improvements seen in two comparisons: 1) combined stress-enhancement and dual-V_{th} versus only dual-V_{th}, and 2) stress-enhancement versus dual-V_{th}. The first two columns state the

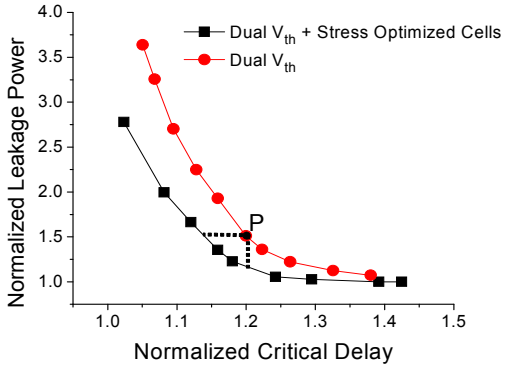


Figure 9. Leakage power versus delay trade-off curve for the circuit c7552 for dual-V_{th} and our approach.

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 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 3.3).

The results clearly show that our combined approach significantly improves the leakage for iso-delay, and also improves critical delay for iso-leakage, when compared to dual-V_{th} based assignment. We get up to 38.5% (23.9% on average) improvement in leakage for iso-delay, and up to 5.8% (5% 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 11 circuits. Even when performing the one-to-one comparison of stress-enhancement versus dual-V_{th}, we achieve up to 7.4% (5.9% on average) improvement in the leakage for iso-delay, and up to 3.6% (3% on average) improvement in delay for iso-leakage. This comparison shows that for this framework and technology, stress-enhancement outperforms dual-V_{th} both in leakage optimization as well as delay optimization. By using stress-enhancement alone, we can eliminate the extra masks and processing steps required for dual-V_{th} designs, thereby reducing 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 trade-off between area and performance, which will increase the performance gap between stress-enhancement and dual-V_{th}.

Figure 11 shows the percentage of gates assigned to low-V_{th} for the dual-V_{th} assignment, and the combined stress enhancement and 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

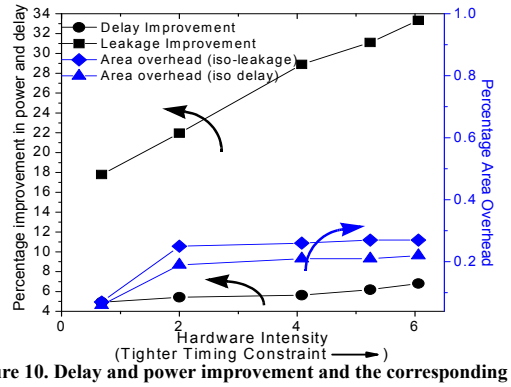


Figure 10. Delay and power improvement and the corresponding area overheads plotted against hardware intensity.

Table 2. 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%
Average		**23.9%**	**0.4%**	**5.9%**	**0.5%**	**5.0%**	**0.5%**	**3.0%**	**0.5%**

expected, for the combined approach, 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 layout optimized 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 about 35% lower than the number for dual-V_{th} assignment.

6. CONCLUSION

In this paper, we studied the dependence of drive current improvement on layout parameters like source/drain length and contact placement, and developed guidelines to improve the layout and maximize performance. We used the guidelines to optimize the layouts of standard cells from a 65nm industrial library. Next, we combined the assignment of these stress-optimized cells with V_{th} assignment in order to optimally trade-off leakage power and performance. The new approach is compared with the traditional dual-V_{th} based assignment technique. Results show that we improve the leakage current by 23.9% on average for identical delay, and improve the critical delay by 5% on average for identical leakage, with a very small area overhead (< 0.5%).

REFERENCES

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

[2] K. Mistry et al., "Delaying Forever: Uniaxial Strained Silicon Transistors in a 90nm CMOS Technology," in *Proc. VLSI Technol. Symp. Tech. Dig.*, pp. 50-51, 2005.

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

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

[5] 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.

[6] 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, 1998.

[7] 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.

[8] R. A. Bianchi et al., "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *Proc. IEDM*, pp. 117-120, 2002.

[9] A. Kahng et al., "Exploiting STI Stress for Performance," in *Proc. ICCAD*, November 2007.

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

[11] M. V. Dunga et al., "Modeling Advanced FET Technology in a Compact Model," in *IEEE Trans. on Elect. Dev.*, Vol. 53, No. 9, pp. 1971-1978, September 2006.

[12] V. Joshi et al., "Stress Aware Layout Optimization," in *Proc. ISPD 2008*, to appear April 2008.

[13] V. Chan et al., "Strain for PMOS performance Improvement," in *Proc. CICC*, pp. 667-674, 2005.

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

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

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

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

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

[19] Manual, Synopsys TSUPREM4, Version 2007.03.

[20] 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.

[21] V. Zyuban et al., "Unified Methodology for Resolving Power–Performance Tradeoffs at the Microarchitectural and Circuit Levels," in *Proc. ISLPED*, pp. 166-171, August 2002.

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

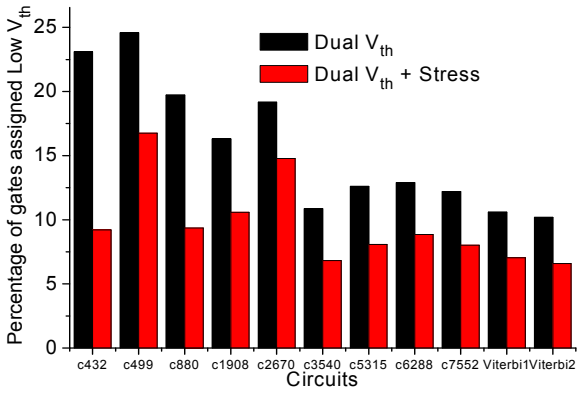


Figure 11. Percentage of gates assigned to low-V_{th} for dual-V_{th} and the combined dual-V_{th} and stress based approach.