Optimization of FinFET-based circuits using a dual gate pitch technique

Sravan K. Marella[1], Amit Ranjan Trivedi[2], Saibal Mukhopadhyay[2], and Sachin S. Sapatnekar[1]
[1]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332

Abstract— **Source/drain stressors in FinFET-based circuits lose their effectiveness at smaller contacted gate pitches. To improve circuit performance, a dual gate pitch technique is proposed in this work, where standard cells with double the gate pitch are selectively used on the gates of the circuit critical paths, at minimal area and power costs. A stress-aware library characterization is performed for FinFET-based standard cells by obtaining stress distributions using finite element simulations on a subset of structures. The stresses are then employed to create look-up tables for mobility multipliers and threshold voltage shifts, for subsequent performance characterization of FinFET-based standard cells. Finally, a circuit delay optimizer is applied using the dual gate pitch approach and is compared with an alternative gate sizing approach. Using a combination of gate sizing and the dual gate pitch approach, it is shown that the average power delay product improves by 12.9% and 15.9% in 14nm and 10nm technologies, respectively.**

I. Introduction

Modern lithography improves printability and reduces critical dimension (CD) variations by requiring transistor gates in a standard cell to lie on a regular grid [1]. To achieve high density, the contacted gate pitch (i.e., the minimum allowable distance between the centers of two adjacent transistor gates with a contact in between) is typically set to be uniform. In successive technology generations, this parameter is reduced to achieve higher integration densities.

This notion of a constant pitch significantly impacts the performance of FinFETs [2] that are used in advanced technologies to offer stronger control over short-channel effects and provide higher on:off current ratios as compared to conventional planar transistors. As in planar transistors, FinFET performance can be greatly enhanced using strain engineering [3] by placing stressors in the fin, in the source/drain region between the transistor gates. However, strain engineering faces two difficulties in FinFET technologies:

- *Reduced stressor volume:* Reduced gate pitches imply that the volume of stressor in the source/drain region is constrained, limiting the effectiveness of strain engineering [4].
- *Fin edge effects:* Source/drain-induced stresses relax along the free edges at the end of the fin [5], resulting in lower stresses and lower mobilities for transistors closer to the fin edge.

To overcome the *reduced stressor volume*, i.e., the dependence of source/drain stressor volume with contacted gate pitch, techniques such as densified STI [6], or metal-gate-induced stress help to incorporate additional stress over and above source/drains stressors. Other methods include a lattice-mismatched strain relaxed buffer that may be grown below the active fin, but this is better suited for Ge-based fins and is impractical for silicon-based channels [4]. To address *fin edge effects*, alternative layout topologies have been proposed, using fewer fins and moving multi-fingered transistors toward the center of the fins [5]. The effectiveness of this technique is limited to multi-finger gates with very short fins; it is inapplicable to minimum-sized standard cells; therefore it does not provide significant improvements for many standard cells in a library. Alternatively, multiple dummy gates may be added [7] at the ends of a fin (i.e., more than the single dummy gate that is normally used), thus moving the stress-relaxed end of the fin away from the functional transistors; however, these dummy gates incur significant area overheads. Furthermore, we show that the improvements due to additional dummy gates diminish as the number of active gates increases. On average, a standard cell in Nangate 15nm library contains 6 transistor gates: altering the layout topology and adding dummy gates provides improvements only for transistors near the edge of the fin, but changing the contacted gate pitch can provide significant improvements in strain for all transistors.

This work proposes using standard cells with double the minimum contacted gate pitch on *selected gates* that lie on critical paths, in order to improve the worst-case delay of a circuit. Doubling the gate pitch increases the source/drain stressor volume and provides greater mobility and threshold voltage improvements, but incurs about twice the area (standard cell width), increased parasitic diffusion capacitance, and some increase in the leakage power. However, it will be shown that the improvements in mobility and threshold voltage outweigh the disadvantages when used selectively on the critical paths to optimize circuit delay. Since only a few selected gates are modified to double gate pitch, the layout impact is not large, and the area impact is further mitigated by the white space that is available in typical row-based placements due to incomplete row utilization.

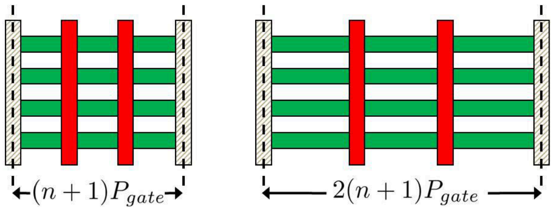

$\leftarrow(n+1)P_{gate}\rightarrow$
$\longleftarrow 2(n+1)P_{gate} \longrightarrow$

Fig. 1. Pull-up/pull-down transistors with nominal and double the gate pitch.

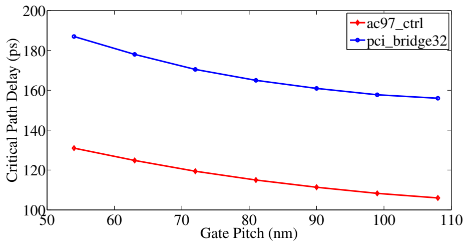


Fig. 2. Change in critical path delay with gate pitch for two benchmarks.

To illustrate the idea, Fig. 1 shows four-fin, two-gate structures with 1× (nominal) and 2× contacted gate pitch; these may represent a pull-up or pull-down network of a two-input standard cell with a single dummy gate (in gray) at each end of the fin. An increase in the contacted gate pitch increases the length of the green source/grain region between the gates, where the stressors lie, and applies additional stress, enhancing performance. This reduces the critical path delay, and its trend as a function of a uniform gate pitch (applied to every cell in the layout), is illustrated in Fig. 2 for 14nm FinFET-based implementations of the ac97_ctrl and pci_bridge32 circuits.[1]

This work is supported in part by NSF CCF-1421606 and SRC 2009-TJ-2234.

[1] Here, the pitch is increased from the 1× (54nm) value to 2× (108nm) in 9nm steps to illustrate the trend, but only 54nm and 108nm are legal values.

Our approach accounts for layout dependency [5] by characterizing stress in the underlying layout and translating its impact on SPICE transistor model parameters. Our contributions are:

- We determine the stress on each transistor using finite element method (FEM) based characterizations for a subset of layout structures. The corresponding mobility multipliers and threshold voltage shifts are stored as a look-up table.
- The look-up tables are employed to build and characterize a standard cell library with two versions of each cell, one with the standard gate pitch and one with twice the pitch.
- We apply the notion of dual gate pitches to optimize benchmark circuits, comparing our approach with conventional gate sizing, where selected gates on the critical paths are up-sized to improve worst-case path delay

The paper is organized as follows. Section II introduces the FinFET structural and layout parameters along with the stressors used in this work. Next, Section III elaborates on the finite element simulation methodology employed to simulate the process-induced stress. This is followed by a discussion of the analytical techniques used to obtain mobility multipliers and threshold voltage shifts, stored as a look-up-table, for standard cell characterization in Section IV. Finally, in Section V, we use a sensitivity-based algorithm to improve the worst-case delay of 14nm/10nm benchmark circuits and compare our method with a similar gate sizing approach.

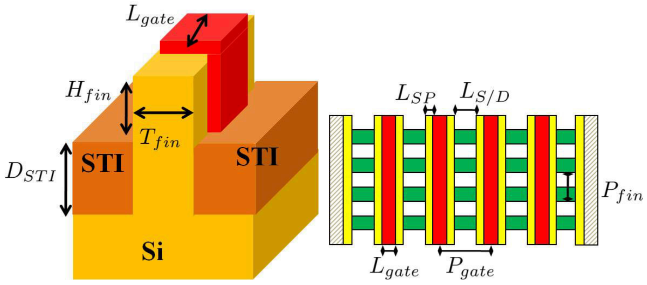
II. FinFET Parameters and Stressors

The magnitude of engineered stress depends upon the FinFET geometry and layout parameters. This section describes the FinFET structure and the intentional stressors considered in this work.

A. FinFET structure and layout

FinFET transistors belong to the family of three-dimensional multi-gate transistors, with the gate wrapping around the channel on three sides. The structure is characterized by two sidewalls and a top surface. If a hard mask exists on the top surface, it is treated as a double-gate transistor, else it acts as a triple-gate transistor. We consider triple-gate transistor structures in this work, but the concepts are applicable to double-gate FinFETs too.

A representative FinFET structure is shown in Fig. 3(a). The FinFET is characterized by fin height, H_{fin}, fin thickness, T_{fin}, and gate length, L_{gate}. The electrical width, W_{fin}, for a triple-gate structure is determined as $W_{fin} = 2 \times H_{fin} + T_{fin}$. Often, multiple fins are used to improve the drive current and to reduce variability of a given transistor. For a multi-fin device with N_{fin} fins, the total electrical width is given as $N_{fin} \times W_{fin}$, i.e., this can be increased in quantized integer steps. The fin is partially surrounded by recessed shallow trench (STI) made up of SiO_2. We consider a Hi-K metal gate technology, where the Hi-K gate oxide is made up of HfSiO, while the metal gate is made up of TiN metal.



(a) (b)

Fig. 3. (a) Basic FinFET structure (b) Layout of a 4-fin-4-gate cell with dummy poly (dashed grey) at the ends.

In Fig. 3(b), the layout top-view of a four-fin four-transistor cell is shown. The gate is flanked by a dielectric low-k spacer (yellow regions) of thickness L_{SP} that reduces the gate-to-source/drain capacitance. The terms P_{gate} and P_{fin} represent the gate pitch and fin pitch, respectively. The length, $L_{S/D}$, of the source/drain region can be derived from the primary parameters as:

$$L_{S/D} = P_{gate} - L_{gate} - 2 \times L_{SP} \tag{1}$$

FinFET-based standard cells are flanked by a single dummy gates (shaded grey) at the end of the fin, as shown in Fig. 3(b). Thus, for a given gate pitch P_{gate}, the width of a standard cell with n active transistors, in the pull-up or pull-down network, is given as an integer multiple of the gate pitch as $(n+1)P_{gate}$. The FinFET structural and layout parameters used in this work are given in Table I.

TABLE I

FinFET parameters

	L_{gate}	H_{fin}	T_{fin}	L_{SP}	P_{fin}	P_{gate}
14nm	18nm	30nm	10nm	10nm	48nm	54nm
10nm	14nm	30nm	8nm	9nm	40nm	48nm

CMOS integrated circuits use logic gates typically with one to four independent inputs, and the number of NMOS/PMOS transistors in a minimum-sized gate is identical to the number of inputs. For logic gates with higher drive strengths, fingered layouts are used. Here, we consider logic gates of strengths 1×, 2×, and 4×, and for inverters or buffers (typically used to drive large loads), we also consider 8×, 16×, and 32× standard cells. Therefore, in this paper, the number of active NMOS/PMOS transistors in a gate take values $NumTran \in \{1, 2, 3, 4, 6, 8, 12, 16, 32\}$.

B. Intentional stressors

Intentional stress can be engineered into transistor channels to boost mobilities and hence circuit performance [3]. Positive (negative) valued stress is termed as tensile (compressive). For PMOS (NMOS) transistor type a compressive (tensile) stress along the channel direction improves the hole (electron) mobilities. The following state-of-the-art strain engineering techniques are considered in this work:

- **Source/drain stress**: Lattice-mismatched SiGe (SiC) alloy is grown epitaxially in source/drain regions to generate compressive (tensile) stress for PMOS (NMOS) transistors.
- **Initial STI stress**: Although regular STI is recessed below the channel and has minor impact [3], densified STI can develop initial stresses in the range of GPa [6].
- **Initial gate stress**: The metal gate can be incorporated with initial stresses that relax to induce stress in the channel. An initial tensile (compressive) stress in the gate creates compressive (tensile) stress in the channel [3].

We also assume the presence of one dummy gate at the edge of the fin. This generates some compressive stress at the edge of the layout, instead of the stress relaxation that is seen in its absence.

III. FinFET Stress Modeling and Characterization

Post manufacturing, the lattice-mismatched stress in the source/drain regions, together with the initial stresses in the STI and the metal gate relax and induce stress in the FinFET channel. This section discusses the finite element method (FEM) based stress modeling methodology that we develop for obtaining stress distributions in the transistor channel in a standard cell layout.

In general, finite element simulations must be performed for all the standard cells in the layout. However, recognizing structural similarities between the standard cells, we build a set of stress primitives. For instance, the fin structure for logic gates INV_X2, NAND2_X1, and NOR2_X1 consists of the same number (two gates) of pull-up and pull-down transistors, and they differ only in their electrical connectivity. We ignore the stress due to the contacts whose contribution is negligible compared to other stressors [4]. Specifically, we perform stress simulations to characterize fins with $n \in NumTran$ gates and a dummy gate on each end, where $NumTran$ is as defined in Section II-A.

A. Stress modeling

The mechanical stress in a system is determined by its three normal stress components, σ_{ii},σ_{jj}, σ_{kk}, and three shearing stress components τ_{ij}, τ_{jk}, and τ_{ki}. Here i, j, k denote the three mutually perpendicular directions of a coordinate system. In integrated circuits, a suitable coordinate system is along the Miller index directions [110], [$\bar{1}$10], [001] represented by x',y', and z'. It can be noted that z' direction corresponds to z-axis direction of the Cartesian coordinate system, while $x' - y'$ are obtained by rotating Cartesian $x - y$ axes by 45^o. Since mobility variations in FinFET devices are primarily due to the $\sigma_{x'x'}$ component along the channel direction x', we concern our discussion to $\sigma_{x'x'}$ component alone. Finite element simulations are performed for various FinFET layout geometries using ABAQUS [8] with the dimensions in Table I for each of the $n \in NumTran$ gate structures for the 1× (nominal) and 2× gate pitches. The stresses in each transistor region are obtained by numerically averaging the tensor components along fin width, fin height and channel length as:

$$\overline{\sigma'} = \frac{1}{L_{gate}} \frac{1}{H_{fin}} \frac{1}{T_{fin}} \int \sigma' \mathbf{dx'dy'dz} \qquad (2)$$

The Young's modulus (denoted by E) in GPa for the materials Si, SiO$_2$, TiN, and HfSiO are: 162, 71.7, 640, and 110, respectively. The corresponding Poisson's ratio (denoted by ν) for the materials Si, SiO$_2$, TiN, and HfSiO are: 0.28, 0.16, 0.25, and 0.2.

B. Simulation of stress relaxation

The magnitude of the initial stress in lattice-mismatched source/drain regions depends upon the mole fraction of the impurity (Ge or C) in the epitaxially grown alloy materials. For a Ge concentration of x% and C concentration of y%, the corresponding alloy materials are represented as Si$_{1-x}$Ge$_x$ and Si$_{1-y}$C$_y$, respectively. The lattice constants of Si, Ge, and C are 0.546nm, 0.566nm, and 0.347nm, respectively. The lattice constants of the alloy materials are obtained by Vegard's law, which gives the resultant lattice constant as a linear combination of individual lattice constants. Clearly, the lattice constant of Si$_{1-x}$Ge$_x$ (Si$_{1-y}$C$_y$) is greater (smaller) than the lattice constant of Si. In this work, we choose a Ge (C) concentration 50% (2%). Thus, when the corresponding alloy materials are epitaxially grown in the source/drain regions, SiGe has an initial compressive stress, while SiC is under a tensile stress in the neighbouring PMOS and NMOS channels, respectively. Moreover, the stress thus developed is isotropic in nature. The initial stress in the source/drain regions is computed as [9]:

$$S_{ii}^{S/D} = \frac{E_{Si}}{1 - 2\nu_{Si}} \left(\frac{a_{Si} - a_D}{a_{Si}} \right) d$$

Here, $S_{ii}^{S/D}$ for $i \in x', y', z$ denotes the initial stress component. The terms E_{Si} and ν_{Si} represent the Young's modulus and Poisson's ratio of silicon. The terms in the braces correspond to the lattice-mismatched strain. The terms a_{Si} and a_D correspond to the lattice constants of silicon and the impurity $D \in \{Ge, C\}$, respectively. The term d denotes the impurity concentration and equals 50% for Ge, and 2% for C.

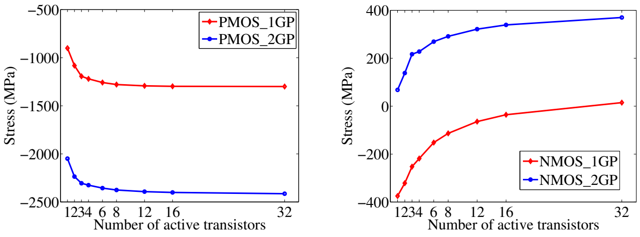
In addition, the corresponding initial isotropic stresses in STI and metal gate are assigned values of $S_{ii}^{STI} = -1$GPa and $S_{ii}^{gate} = +1$GPa for $i \in \{x', y', z\}$.

The gate-last approach is captured by a two-stage simulation:

- The system is first simulated with initial stresses of $S_{ii}^{S/D}$ and S_{ii}^{STI} in the source/drain and STI regions. The gate is absent in this step to simulate the replacement gate process. The averaged stress tensor in the transistor channel is denoted by $\overline{\sigma'}_{(S/D,STI)}$.
- Next, it is simulated with an initial stress of S_{ii}^{gate} in the gate region to simulate the gate-last approach to obtain the corresponding channel-averaged stress tensor, $\overline{\sigma'}_{Gate}$.

The total stress in each individual transistor channels is obtained by a linear superposition of the components of two tensors as:

$$\overline{\sigma'}_{\mathbf{Total}} = \overline{\sigma'}_{\mathbf{(S/D,STI)}} + \overline{\sigma'}_{\mathbf{Gate}} \qquad (3)$$



(a) (b)

Fig. 4. Average $\sigma_{x'x'}$ channel stress among all the transistors due to intentional stress in (a) PMOS and (b) NMOS transistors. Here 1GP and 2GP correspond to 54nm and 108nm, respectively.

For the rest of the discussion, the stress tensor components denote the channel averaged stress distributions. To characterize the effect of increased gate pitch, we observe the $\sigma_{x'x'}$ component along the channel length, averaged among all the transistors in a fin. Fig. 4 plots the average stress in structures with $n \in NumTran$ gates, separately for PMOS and NMOS transistors. We observe that:

- The layout dependency is evident from the different magnitudes of stress based on the number of gates in the layout, as also observed in [5]. Channel stress becomes more compressive for PMOS transistors as the number of gates increases. For NMOS, at the nominal 54nm gate pitch, the stress becomes less compressive, while at 2× gate pitch, it becomes more tensile.
- From Fig. 4(a), the $\sigma_{x'x'}$ component in the PMOS transistors becomes more compressive as gate pitch doubles. On the other hand, from Fig. 4(b) for the nominal (54nm) NMOS gate pitch case, $\sigma_{x'x'}$ is compressive for a smaller number of transistors and tends to be tensile as the number of active gates increase. Furthermore, the $\sigma_{x'x'}$ component is tensile with the double gate pitch, indicating that the SiC source/drain stress dominates.

IV. STRESS-AWARE STANDARD CELL CHARACTERIZATION

Having characterized the stress in a fin, we now focus on the impact of stress on electrical parameter variations in specific standard cells. In this section, we will present analytical mobility and threshold voltage variation models based on analytical piezoresistivity and deformation potential theory, respectively. These models are used to populate look-up tables that determine the mobility multipliers and threshold voltage shifts for each transistor within a standard cell [10], which are fed to HSPICE simulations for library characterization.

A. Obtaining mobility multipliers and threshold voltage shifts

Mobility variations: According to piezoresistivity theory, the changes in transistor mobility can be expressed as a function of applied stress. A complete mathematical model is presented in [11]. Here, we assume, transistor channels are oriented along the [110] axis direction. For each transistor k in a standard cell with $n \in NumTran$ transistors, we obtain the mobility variation as:

$$\frac{\Delta\mu^k}{\mu} = \pi'_{11}\sigma^k_{x'x'} + \pi'_{12}\sigma^k_{y'y'} + \pi_{12}\sigma^k_{z'z'} \qquad (4)$$

Here, μ denotes the carrier mobility and $\Delta\mu^k$ the change in mobility in the k^{th} transistor. The terms π'_{11} and π'_{12} are the piezoresistivity coefficients in the primed coordinate system, and π_{12} is a piezoresistivity coefficient in the unprimed coordinate system since the z-axis remains constant the translated coordinate system. The stress components $\sigma^k_{x'x'}$, $\sigma^k_{y'y'}$, and $\sigma^k_{z'z'}$ are the channel averaged normal stress components in the k^{th} transistor obtained from FEM simulations outlined in Section III.

The electrostatics in a FinFET transistor differ from bulk technology and so are their piezoresistivity coefficients. The piezoresistivity values for FinFET-based transistors are given in Table II [12].

TABLE II
FinFET Piezoresitivity Coeffs. in (100) Si

	π'_{11} (Pa^{-1})	π'_{12} (Pa^{-1})	π_{12} (Pa^{-1})
NMOS	452×10^{-12}	256×10^{-12}	-576×10^{-12}
PMOS	-450×10^{-12}	238×10^{-12}	101×10^{-12}

During SPICE-level simulations, the corresponding mobility multipliers in the transistor k is given by $1 + \frac{\Delta\mu^k}{\mu}$ [10].

Threshold voltage variations: Applied mechanical stress also causes shifts and splits in electronic band potentials and results in reduction in energy band gap, and thus a reduction in threshold voltage. It has been shown in [13] that epitaxial uniaxial strain results in reduction in energy band gap and hence in threshold voltage. Deformation potential theory relates changes in conduction and valence band energy levels to strains in the crystallographic coordinate system. Thus, stress components from finite element method in the primed coordinate system need to be transformed into strains in crystallographic Cartesian coordinate system. This can be accomplished by familiar stress-strain relations (Hooke's Law) and axis transformations in [14]. The resultant changes in energy band potentials in the k^{th} transistor of a standard cell, with $n \in NumTran$ transistors, are given as:

$$\begin{aligned}\Delta E^{(i)}_{C_{(k)}} &= \Xi_d \left(\epsilon^k_{xx} + \epsilon^k_{yy} + \epsilon^k_{zz}\right) + \Xi_u \epsilon^k_{ii}, i \in \{x', y', z'\} \\ \Delta E^{(hh,lh)}_{V_{(k)}} &= a \left(\epsilon^k_{xx} + \epsilon^k_{yy} + \epsilon^k_{zz}\right) \qquad (5) \\ &\pm \sqrt{\frac{b^2}{4}(\epsilon^k_{xx} + \epsilon^k_{yy} - 2\epsilon^k_{zz})^2 + \frac{3b^2}{4}(\epsilon^k_{xx} - \epsilon^k_{yy})^2 + d^2\left(\epsilon^k_{xy}\right)^2}\end{aligned}$$

Here, $\Delta E^{(i)}_{C_{(k)}}$ is the change in the conduction band potential energy in the carrier band i for the k^{th} transistor. The term $E^{hh}_{V_{(k)}}$ ($E^{lh}_{V_{(k)}}$) denotes the heavy-hole [light-hole] valence band potential of the k^{th} transistor, with a corresponding usage of the positive [negative] sign in the expression. The terms ϵ^k_{xx}, ϵ^k_{yy}, ϵ^k_{zz}, ϵ^k_{yz}, ϵ^k_{zx}, and ϵ^k_{xy} denote the six channel-averaged strain components of the k^{th} transistor in the Cartesian coordinate system. The coefficient terms Ξ_d and a are the hydrostatic deformation potential constants and the terms Ξ_u, b, and d are the shear splitting deformation potential constants. The corresponding values of the constants Ξ_d, Ξ_u, a, b, and d in eV are [15]: 1.13, 9.16, 2.46, -2.35, -5.08.

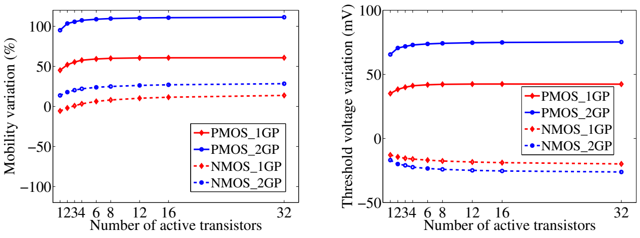
The threshold voltage of PMOS/NMOS transistors can in turn be expressed in terms of changes in conduction band and valence band potentials. In this work, the changes in electronic band potentials are due to the source/drain, STI, and gate stressors. The corresponding threshold voltage shifts in the k^{th} transistor for a structure with $n \in NumTran$ transistors is given by:

$$\begin{aligned} q\Delta V^k_{thp} &= (m-1)\Delta E_{C_{(k)}} - m\Delta E_{V_{(k)}} \\ q\Delta V^k_{thn} &= -m\Delta E_{C_{(k)}} + (m-1)\Delta E_{V_{(k)}} \qquad (6)\end{aligned}$$

where ΔV^k_{thp} and ΔV^k_{thn} are the threshold voltage shifts in PMOS and NMOS threshold voltages, respectively, q is the electron charge, and m (= 1.3 – 1.4) is the body-effect coefficient. $\Delta E_{C_{(k)}}$ is the minimum of the changes in conduction band potentials, $\Delta E^i_{C_{(k)}}$ and $\Delta E_{V_{(k)}}$ are the maximum of the changes in valence band potentials, $\Delta E^{hh}_{V_{(k)}}$ and $\Delta E^{lh}_{V_{(k)}}$ of the k^{th} transistor in a standard cell.

Comparison: The Fig. 5 shows the mobility variations obtained by using Equation (4), and the threshold voltage shifts obtained using Equation (6) for nominal and twice the gate pitch in 14nm technology (P_{gate} = 54nm). From the figures we can deduce the following:

- From Fig. 5(a), we can see that the magnitudes of PMOS and NMOS mobility improvements are higher with double the gate pitch, consistent with observations in Fig. 4. However, with 2× the gate pitch, the relative improvements in PMOS transistors is greater than NMOS transistors and can be explained by the relative magnitudes of stress components.
- From Fig. 5(b), we can observe that the stress-induced threshold voltage shifts in PMOS (NMOS) are positive (negative) valued indicating reduction in threshold voltages. Moreover, when gate pitch is doubled, PMOS and NMOS have increased threshold voltage shifts. This contributes to delay improvements and increase in leakage power with 2× gate pitch. Similar to mobility variations, the PMOS transistors experience higher magnitudes of threshold voltage shifts compared to NMOS transistors at double the gate pitch.



(a) (b)

Fig. 5. (a) Average mobility (b) Average threshold voltage variations over all transistors in PMOS and NMOS FinFETs. Here 1GP and 2GP correspond to 54nm and 108nm, respectively.

B. Library characterization

The standard cell characterization takes the underlying layout into consideration. For a given library of standard cells and their corresponding layouts, the following steps are performed considering a nominal gate pitch and twice its value:

- We obtain stress distributions for different structures with $n \in NumTran$ gates for nominal and double the gate pitch. The stress tensor components are averaged along the channel using Equation (2). We obtain the total stresses simulating gate-last approach using Equation (3) in Section III-B.
- We obtain mobility variation and threshold voltage shifts by applying the piezoresistivity model in Equation (4), and deformation potential theory formulation in Equations (5), (6). The electrical variations are stored in a look-up table as corresponding mobility multipliers and as threshold voltage shifts.
- We apply, during standard cell characterization, based on the number of active transistors in the layout, the look-up table entries by performing HSPICE circuit simulations.
- For delay, we characterize our standard cell library for different supply voltages, load capacitances, and input slopes. For leakage power, we characterize our library for different supply voltages and static input conditions.

We apply this standard cell characterization approach for a 14nm and 10nm PTM [16] technologies in conjunction with BSIM-CMG [17] FinFET transistor models. To allow standard cells with twice the gate pitch to be used on selected gates of the critical paths, we characterize two sets of libraries – one with nominal gate pitch, which we refer to as Library_1GP, and another with twice the gate pitch which is referred to as Library_2GP. Thus it takes twice the time to characterize both sets of libraries, but this is a one-time effort.

V. Results

In Sections III and IV, we have seen that using standard cells with twice the nominal gate pitch improves the magnitudes of engineered mobility and threshold voltage shifts. In this section, we show the circuit delay improvements that can be obtained by using standard cells with twice the gate pitch. We compare our technique with conventional gate sizing approach and we show that a combination of the two results in superior improvements.

A. Timing optimization framework

We begin with a placed circuit netlist with nominal gate pitch, and apply optimization techniques to improve the delay by replacing selected standard cells with twice the gate pitch or by using a higher strength variant (sizing) of the standard cell. For this we chose a TILOS [18] based circuit optimization framework. We find the best delay achievable with our optimization within the given placement area. Typical standard cell rows have enough white space to accommodate the higher strength variants or the double gate pitch variants; for example, our benchmarks show row utilizations ranging from 35% to 80%. The timing optimization is outlined as follows:

1) Find the current most critical path in the design.
2) For each gate on the current critical path, compute the change in the critical path delay, ΔD, and leakage power, ΔL, obtained by either upsizing the gate or by choosing a corresponding gate with twice the gate pitch.
3) Find the gate with the best gain $G = \Delta D/\Delta L$, and replace it with corresponding higher strength variant (sizing) or with a corresponding standard cell with twice the gate pitch.
4) Go to step 1 till convergence criteria is met.

The procedure converges when no possible upsizing/double gate pitch standard cells are found, or if the circuit area exceeds a bound. We compare three strategies for circuit optimization:

- *Only gate sizing (OPT_X):* The cells are replaced by an upsized variant, e.g., INV_X1 may be replaced with INV_X2 (Fig. 6). The cells are chosen from Library_1GP alone.
- *Only double gate pitch (OPT_2GP):* The cells are replaced are corresponding cells with twice the gate pitch, e.g., INV_X1 may be replaced with INV_X1_2GP (Fig. 6). The cells are chosen from Library_2GP alone.
- *Combined optimization (OPT_Comb):* While selecting the cell with best gain, we consider both sizing and double gate pitch options, and chose the cell with a higher gain. Cells can be chosen from either Library_1GP or Library_2GP.

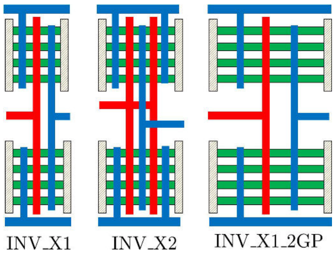


Fig. 6. Layouts of INV_X1, INV_X2 (sizing), INV_X1_2GP with twice the gate pitch. For a gate pitch of 54nm, the corresponding standard cell widths are: 108nm, 162nm, and 216nm.

Gate selection: We now show a example of gate choices during optimization. Table III shows the delay and average leakage power of a set of NAND2 standard cells in the Library_1GP, Library_ExDummy, and Library_2GP. The standard cells in Library_1GP and Library_2GP have nominal and twice the gate pitch, respectively as discussed in Section IV-B. The gates in Library_ExDummy have an additional pair of dummy gates so that the active transistors do not experience fin-edge effects. The column n denotes the number of active transistors in the fin. We can see that a higher strength variant within the same library provides both superior PMOS rise and NMOS fall delays, while the corresponding cell with twice the gate pitch provides better PMOS rise-delay improvements compared to NMOS fall-delay improvements. This is due to the relatively smaller mobility and threshold voltage improvements in NMOS transistors shown in Fig. 5 in Section IV-A. The standard cell leakage power is expected to increase with increased width (higher strength) and with greater threshold voltage shifts (twice the gate pitch). However, doubling the gate pitch incurs comparatively smaller magnitude of leakage power compared to upsizing a gate. Further, it can be seen that the best rise delay improvement from the 2GP case is significantly better than that of the 1GP case. For completeness, we compare the corresponding gates in the Library_ExDummy. We can observe that the delay improvements, obtained by adding additional dummy gates, diminish with the number of active transistors.

TABLE III
DELAY AND LEAKAGE POWER OF 14NM NAND2 CELLS.

Gate	n	Library_1GP		Library_ExDummy		Library_2GP	
		Rise/Fall (ps)	Leakage (nW)	Rise/Fall (ps)	Leakage (nW)	Rise/Fall (ps)	Leakage (nW)
NAND2_X1	2	11.4/19.4	19.9	11.2/16.9	20.5	7.1/18.7	32.2
NAND2_X2	4	7.6/13.1	40.2	6.7/12.3	40.8	3.9/12.7	64.7
NAND2_X4	8	6.1/10.7	80.8	6/10.6	81	3.2/9.2	130.3

B. Circuit-level optimization with dual gate pitches

We apply our techniques to a set of IWLS05 [19] benchmarks described in Table IV. The column denoted #G refers number of number of gates in the corresponding circuit. We use CAPO [20] for circuit placement and PTM SPICE models. Our inputs are:

- Characterized standard cell libraries with nominal (Library_1GP) and double gate pitch (Library_2GP) for 14nm/10nm technology.
- An initial placed netlist with nominal gate pitch cells. We treat this as our reference and term it as the "Nominal" case. Note that the layouts of 14nm and 10nm are different.

We run static timing analysis and compute the dynamic and static leakage power by propagating signal probabilities. We compare the delay, total power, and the power-delay product of the nominal and timing-optimized circuits, where the power-delay product multiplies the total power and the worst-case path delay of the circuit, and is a measure of the energy consumption per clock cycle.

Table IV shows the results of three optimizations, OPT_X, OPT_2GP, and OPT_Comb in a 14nm and 10nm technology. The columns D_0, P_0, and E_0 denote the worst-case critical path delay, total power (sum of dynamic and static leakage power), and the power-delay product of the nominal circuit without optimizations in 14nm technology. We present the changes in the circuit metrics with reference to the nominal case. The columns ΔD_i, ΔP_i, and ΔE_i under 14nm technology indicate the changes in delay, total power, and the power-delay product using optimization i, where $i = 1, 2, 3$ refer to OPT_X, OPT_2GP, and OPT_Comb, respectively. For 10nm technology, the columns D_4 and E_4 denote the nominal delay and power-delay product, respectively. The relative changes in delay (power-delay product) for OPT_X, OPT_2GP, and OPT_Comb are given in columns ΔD_5, ΔD_6, and ΔD_7 (ΔE_5, ΔE_6, and ΔE_7), respectively. Negative (positive) changes indicate improvements (degradations). It can be seen that the total power increases for all the optimizations. This is because, a higher strength variant has greater transistor width, while using cells with twice the gate pitch incurs higher magnitudes of threshold voltage shifts as seen in Fig. 5. Both these effects contribute to increased leakage power of the circuit. Finally, we report the average improvements (degradations) in delay and power-delay product (total power) for all optimizers, over all circuits.

From Table IV, the changes in circuit metrics using the three optimization techniques in 14nm and 10nm technology are summarized as follows:

- *OPT_X*: For 14nm technology, the delay improvements range from 0% to -20.6%, the total power degrades by 0.01% to 2.1%, and the power-delay product changes by -19.0% to 0.01%. We can observe that for the benchmark mem_ctrl in this work, the critical path delay did not change but the total power

TABLE IV

CIRCUIT OPTIMIZATION RESULTS FOR 14NM/10NM TECHNOLOGY

Circuit	#G	14nm Technology												10nm Technology							
		Nominal			OPT_X			OPT_2GP			OPT_Comb			Nominal		OPT_X		OPT_2GP		OPT_Comb	
		D_0	P_0	E_0	ΔD_1	ΔP_1	ΔE_1	ΔD_2	ΔP_2	ΔE_2	ΔD_3	ΔP_3	ΔE_3	D_4	E_4	ΔD_5	ΔE_5	ΔD_6	ΔE_6	ΔD_7	ΔE_7
	(×1K)	(ps)	(μW)	(fJ)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(ps)	(fJ)	(%)	(%)	(%)	(%)	(%)	(%)
ac97_ctrl	9.5	131	845	111	-20.6%	2.0%	-19.0%	-18.3%	1.31%	-17.2%	-22.1%	2.8%	-20.0%	103	94	-19.4%	-18.2%	-28.2%	-26.3%	-29.1%	-26.5%
aes_core	11.9	141	700	99	-9.9%	1.2%	-8.9%	-11.3%	1.34%	-10.2%	-18.4%	3.0%	-16.0%	101	77	-7.9%	-7.0%	-6.9%	-6.1%	-14.9%	-12.1%
des	4.6	264	463	122	-3.0%	0.3%	-2.8%	-8.7%	0.27%	-8.5%	-9.8%	1.3%	-8.7%	206	103	-11.7%	-11.1%	-15.5%	-14.9%	-15.5%	-14.3%
ethernet	28.0	238	1704	406	-6.7%	0.3%	-6.5%	-8.0%	0.33%	-7.7%	-11.3%	0.6%	-10.8%	206	373	-8.3%	-7.9%	-3.9%	-3.8%	-6.8%	-6.6%
i2c	1.0	134	87	12	-18.7%	2.1%	-16.9%	-11.9%	0.98%	-11.1%	-24.6%	6.4%	-19.8%	120	11	-32.5%	-30.6%	-36.7%	-34.0%	-41.7%	-37.2%
mem_ctrl	8.9	253	707	179	0.0%	0.01%	0.01%	-6.7%	0.34%	-6.4%	-9.5%	1.2%	-8.4%	203	153	-6.4%	-6.1%	-5.4%	-5.2%	-10.8%	-9.8%
pci_bridge32	10.0	187	727	136	-9.6%	0.7%	-9.0%	-13.4%	1.10%	-12.4%	-17.1%	1.7%	-15.7%	149	117	-14.1%	-13.2%	-14.8%	-13.9%	-19.5%	-18.1%
spi	3.1	259	262	68	-15.4%	0.8%	-14.7%	-5.0%	0.18%	-4.9%	-17.4%	1.6%	-16.1%	213	60	-6.1%	-5.8%	-6.6%	-6.3%	-19.2%	-17.9%
systemcdes	2.7	208	275	57	-2.4%	0.2%	-2.2%	-6.3%	0.37%	-5.9%	-9.1%	0.8%	-8.4%	166	49	0.0%	0.02%	-13.9%	-12.9%	-9.0%	-8.6%
usb_funct	11.2	192	749	144	-1.6%	0.1%	-1.4%	-10.4%	0.46%	-10.0%	-5.7%	0.5%	-5.3%	146	117	-13.7%	-12.9%	-8.2%	-7.8%	-8.9%	-8.4%
Average					-8.8%	0.8%	-8.1%	-10.0%	0.7%	-9.4%	-14.5%	2.0%	-12.9%	Average		-12.0%	-11.3%	-14.0%	-13.1%	-17.5%	-15.9%

increases by 0.01% (however, the critical paths before and after optimization are different); for the remaining circuits, we can observe improvements in delay and the power-delay product. The average improvements in delay and power-delay product are -8.8% and -8.1%, respectively. For 10nm technology, the ranges of (average) delay and power-delay product changes are: -32.5% to 0% (-12%), and -30.6% to 0.02% (-11.3%), respectively.

- *OPT_2GP*: For 14nm technology, the delay and power-delay improvements range from -5% to -18.3% and -4.9% to -17.2%, respectively, for a total power overhead ranging from 0.18% to 1.34%. The corresponding average improvements in delay and power-delay product are -10% and -9.4%. For 10nm technology, the ranges of (average) delay and power-delay product improvements are: -3.9% to -36.7% (-14%), and -3.8% to -34% (-13.1%), respectively.
- *OPT_Comb*: For 14nm technology, the changes in delay, total power, and power-delay product range are: -5.7% to -24.6%, 0.5% to 6.4%, and -5.3% to -20%, respectively. The average delay and power-delay product improvements are -14.5% and -12.9%, respectively. For 10nm technology, the corresponding ranges (average) of delay and power-delay product changes are: -6.8% to -41.7% (-17.5%), and -6.6% to -37.2% (-15.9%).

From the changes in circuit metrics, we can observe that the performance of the dual gate pitch technique (OPT_2GP) is superior to the only sizing approach (OPT_X) in most of the circuits except ac97_ctrl, i2c, and spi (ethernet, mem_ctrl and usb_funct) in 14nm (10nm) circuits. The relatively smaller delay improvements due to OPT_2GP approach in these circuits, is due to the smaller NMOS fall delays improvements compared to PMOS rise delays as discussed in Section V-A. On the other hand the use of sizing can improve both the rise/fall delays of a given gate on the critical path. This shows that it is worth exploring the possibility of using a combination of both the techniques as demonstrated by the combined approach. In fact, on an average, the OPT_Comb optimization approach provides better delay and power-delay product improvements. In addition, it was observed that the OPT_Comb approach predominantly chooses corresponding cells from Library_2GP (2× gate pitch) over higher strength variants in LIbrary_1GP owing to their superior rise delay improvements at a considerably smaller leakage overhead (refer Section V-A).

VI. CONCLUSIONS

This work demonstrates a dual gate pitch technique to improve the source/drain stressor effectiveness in FinFET-based circuits. In this approach, selected gates on the critical path are replaced with corresponding gates with twice the gate pitch. The stress distributions in the FinFETs are obtained through FEM simulations, and subsequently used to generate look-up tables for mobility multipliers and threshold voltage shifts at SPICE level. A sensitivity-based circuit optimization is employed to optimize circuit delays using sizing, twice the gate pitch, and a combination of both the techniques. It has been shown that the power-delay product of FinFET-based circuits can be improved by performing a concurrent sizing and dual gate pitch optimization.

REFERENCES

[1] T. Jhaveri *et al.*, "Co-optimization of circuits, layout and lithography for predictive technology scaling beyond gratings," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, pp. 509–527, Apr. 2010.

[2] S. Tang *et al.*, "FinFET-a quasi-planar double-gate MOSFET," in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 118–119, February 2001.

[3] A. Nainani *et al.*, "Is strain engineering scalable in FinFET era?: Teaching the old dog some new tricks," in *IEEE International Electronic Devices Meeting*, pp. 18.3.1–18.3.4, December 2012.

[4] G. Eneman *et al.*, "Stress simulations for optimal mobility group IV p- and nMOS FinFETs for the 14 nm node and beyond," in *IEEE International Electronic Devices Meeting*, pp. 6.5.1–6.5.4, December 2012.

[5] M. Bardon *et al.*, "Layout-induced stress effects in 14nm & 10nm FinFETs and their impact on performance," in *Symposium on VLSI Technology*, pp. T114–T115, June 2013.

[6] T. Baldauf *et al.*, "Strained isolation oxide as novel overall stress element for Tri-gate transistors of 22nm CMOS and beyond," in *International Semiconductor Conference Dresden-Grenoble*, pp. 61–63, September 2012.

[7] S. Mujumdar and S. Datta, "Layout-dependent strain optimization for p-channel trigate transistors," *IEEE Transactions on Electron Devices*, vol. 59, pp. 72–78, January 2012.

[8] "ABAQUS CAE Online Documentation." available at `http://www.sharcnet.ca/Software/Abaqus/6.11.2/index.html`.

[9] N. Xu *et al.*, "Effectiveness of stressors in aggressively scaled FinFETs," *IEEE Transactions on Electron Devices*, vol. 59, pp. 1592–1598, June 2012.

[10] M. Bardon and V.Moroz. private email communication, 2015.

[11] D. A. Bittle *et al.*, "Piezoresistive stress sensors for structural analysis of electronic packages," *Journal of Electronic Packaging*, vol. 113, no. 3, pp. 203–215, 1991.

[12] M. Saitoh *et al.*, "Three-dimensional stress engineering in FinFETs for mobility/on-current enhancement and gate current reduction," in *Symposium on VLSI Technology*, pp. 18–19, June 2008.

[13] S. D. dos Santos *et al.*, "Impact of selective epitaxial growth and uniaxial/biaxial strain on DIBL effect using triple gate FinFETs," *Journal of Integrated Circuits and Systems*, vol. 5, pp. 154–159, March 2010.

[14] J. Barber, *Elasticity*. New York City, USA: Springer, 2010.

[15] J.-S. Lim, S. E. Thompson, and J. G. Fossum, "Comparison of threshold-voltage shifts for uniaxial and biaxial tensile-stressed n-MOSFETs," *IEEE Electron Device Letters*, vol. 25, pp. 731–733, November 2004.

[16] "Predictive Technology Model." available at `http://www.eas.asu.edu/$\sim$ptm`.

[17] "Berkeley Short-channel IGFET Model." available at `http://www-device.eecs.berkeley.edu/bsim`.

[18] J. P. Fishburn and A. E. Dunlop, "TILOS: A posynomial programming approach to transistor sizing," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 326–328, 1985.

[19] "IWLS 2005 Benchmarks." available at `http://www.iwls.org/iwls2005/benchmarks.html`.

[20] A. E. Caldwell, A. B.Kahng., and I. L. Markov, "Can recursive bisection alone produce routable, placements?," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 477–482, 2000.