On Stress Aware Active Area Sizing, Gate Sizing, and Repeater Insertion

Ashutosh Chakraborty
ECE Department
University of Texas, Austin
Austin, TX 78712, USA
ashutosh@cerc.utexas.edu

David Z. Pan
ECE Department
University of Texas, Austin
Austin, TX 78712, USA
dpan@ece.utexas.edu

ABSTRACT

Enormous technical and economic challenges facing technology scaling has rendered strain engineering techniques as the critical enabler of high performance designs in sub-100nm geometries. One of these techniques, source/drain (S/D) SiGe, has an interesting property that the mobility of the device is dependent on the size of active area (*AA*) surrounding it. To exploit this phenomenon for higher performance, a circuit designer needs first order and computationally tractable transistor level models. This paper provides the first *AA* sizing dependent RC switch level model of a logic gate which can be readily used by circuit designers. We derive the methodology to optimally use *AA* sizing for some common cells such as NAND, NOR and INV. For the first time, we formulate a convex optimization problem for *concurrent AA* and gate sizing problem for performance optimization and solve it optimally. We also analytically solve *AA* sizing aware optimal repeater insertion problem for dealing with the menace of long global interconnects in modern chip design. Experimental results demonstrate that our methodology can reduce inter-chip long global interconnect delay by 9% and inter-module gate delays by 10% with only 11% increase in dynamic power dissipation.

Categories and Subject Descriptors

B.8 [**PERFORMANCE AND RELIABILITY**]: General

General Terms

Design Performance

Keywords

Stress, Performance, Sizing, Repeater, Buffer

1. INTRODUCTION

As technology scaling becomes prohibitively expensive, device engineers have been working hard to push the envelop of performance from an existing technology node. Exploting mechanical stress dependent performance is a major part of this effort. Small device geometries (under 100nm) are more amenable to mechanical stress effects as compared to μm technology nodes since these effects have small geometric range of impact. Once considered a phenomenon to avoid, mechanical stress is now routinely exploited by all major μP manufacturers: IBM's PowerPC5, AMD's Opteron and Intel Pentium-IV [1] have used mechanical stress to boost their performance. In general, compressive (tensile) strain in the channel increases mobility of PMOS (NMOS) devices. There are several ways to impart mechanical stress to a device such as: a) Shallow trench isolation (STI) around active area, b) $Si_{1-x}Ge_x$ in the source/drain (SiGe S/D) region c) Contact etch stop layer (CESL) and d) Embedded $Si_{1-x}Ge_x$ channel. [2] has observed that the mobility enhancement by using several simultaneous stress imparting techniques can be more than the addition of individual components.

One of the stress imparting techniques, SiGe S/D, is manufactured by etching away silicon from source and drain (S/D) region of a MOSFET and filling them epitaxially with $Si_{1-x}Ge_x$ alloy where x$\in$ (0.2, 0.4), hence the name SiGe S/D. Figure 1 shows such a PMOS S/D SiGe device graphically. Due to mismatch between the lattice constant of SiGe in S/D region and Si in channel region (lattice constant of SiGe $>$ Si), compressive strain is created in the channel which increases the mobility of holes. Since only compressive strain can be imparted, SiGe S/D is primarily used for PMOS device performance enhancement.

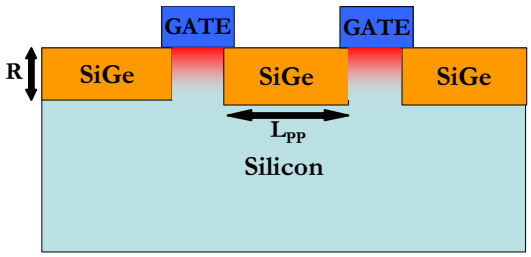

Figure 1: Side view of a SiGe S/D device. Source/Drain regions are epitaxially filled with SiGe which compresses the channel.

SiGe S/D technology has three parameters of paramount interest: The recess depth ("R" in Figure 1), concentration of Ge (i.e. the value of x in $Si_{1-x}Ge_x$) and the active area (*AA*) dimension ("L_{pp}" in Figure 1). Increasing any one of these three parameters increases the magnitude of compressive stress thereby enhancing the mobility of holes further. The upper-limit of these three parameters is bound by permissible leakage current, lattice dislocation tolerance and layout size respectively. The dimension L_{pp} is a measure of the length of *AA* between adjacent poly devices. Through electrical measurements and process simulations, [3] has observed that increasing the dimension L_{pp} can cause substantial increase in compressive stress in the channel, leading to higher mobility improvement. We refer increasing dimension L_{pp} as "AA sizing" in the rest of this paper.

Timing optimization in nano-meter VLSI needs a two pronged approach. At the chip level, the delay of global interconnects need to be minimized using repeater insertion. At the module level, the delay of a module needs to be minimized through gate sizing. In this paper we significantly enhance both the above techniques by making them *AA* sizing aware. For the case of repeater insertion, *AA* sizing

ISPD'09, March 29–April 1, 2009, San Diego, California, USA.

colludes mathematically into the analytical solution leading to 9% further decrease in global interconnect delay as compared to optimal solution without *AA* sizing. Similarly, performing simultaneous gate and *AA* sizing preserves the convexity of the resultant optimization problem which can be solved optimally to achieve more than 10% further reduction in delay through a module compared to gate sizing without *AA* sizing. Such impressive reduction in cycle time of the design comes simply by exploiting stress by layout modification and can be performed by designer.

Rest of this paper is organized as follows: Section 2 presents previous work and our primary contributions. *AA* sizing aware cell delay model is derived in Section 3. In section 4 we solve *AA* sizing aware repeater insertion and simultaneous *AA* and gate sizing showing impressive results. Discussions about various possibilities with *AA* sizing methodology and conclusions derived from this work are presented in Section 5.

2. PREVIOUS WORK AND OUR CONTRIBUTIONS

There has been a plethora of work [4][5][6][2] on the use of various techniques for imparting mechanical stress to NMOS and PMOS devices. A recent work [3] reported the layout sensitivity of mobility of S/D SiGe devices. Essentially, larger SiGe source/drain regions means bigger stressor next to the channel which increases the stress value in the channel, making the device faster. Based on this, [1] proposed using *AA* size modulation for post-routing timing improvement by utilizing whitespace in the design. [1] assumed a simplistic linear delay model for reduction in cell delay as a function of *AA* increase. In addition, they do not consider the capacitance increase while sizing *AA*. As we will show in Section 3, due to the impact of growing *AA* capacitance, the timing improvement is not linear. Recently, [7] touched upon the issue of *AA* sizing for timing improvement but restricted it to only the devices at the boundary of a standard cell. Additionally, [7] deals with only individual gates without embedding them in a real design flow.

Primary Contributions

- We develop the first systematic parameterized model for resistance and capacitance change due to active area sizing. Using these, we build parameterized delay models for basic gates like INV, NAND, NOR etc.
- Guided by our simple yet reasonably accurate analytical model, we enhanced two classic interconnect and gate timing optimization techniques - repeater insertion and gate sizing - by performing them simultaneously with *AA* sizing methodology. Our results show an impressive 10% decrease in interconnect and gate delay.

To the best of our knowledge, our work is the first attempt to abstract out the advantages of strained devices from the process simulation realm systematically into the gate/circuit level and coupling it mathematically with timing optimization techniques which circuit designers use routinely.

3. AA SIZING AWARE CELL DELAY MODEL

ITRS roadmap [8] predicts the contact dimension of $56nm$ for the $45nm$ technology node. After accounting for spacer dimensions, we estimate that the typical L_{pp} distance in the $45nm$ process to be around 90-$100nm$. This estimation also matches the L_{pp} dimension mentioned in [3]. In the rest of the paper, we use the default L_{pp} distance as $90nm$ for the $45nm$ technology node[1]. Consider a nominal PMOS device with $90nm$ L_{pp}. Now, let the *AA* (i.e. L_{pp}) of this device be increased to K times its original value (i.e. to KL_{pp}). Under such circumstances, let S_K denote the multiplicative increase in the mechanical stress due to increase in active area dimensions. Based on the electrical measurements and simulations data in [3], we plot the increase in mobility, S_K, vs the *AA* sizing factor K in Figure 2.

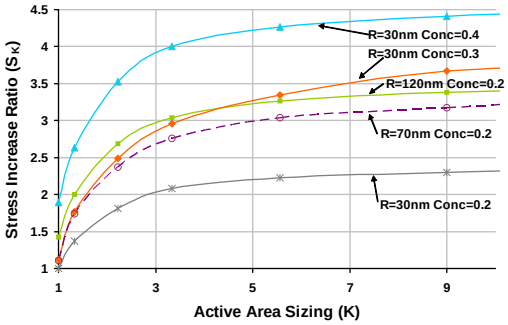

Figure 2: Channel stress vs L_{pp} curve. Stress values normalized to the value at nominal L_{pp} of 90nm.

The different curves above are shown for various recess depths ('R' in figure) and various fractional concentration of Ge in SiGe ('Conc' in figure). From Figure 2 we note that increasing recess depth and Ge concentration increases the stress in the channel and that the stress is much more sensitive to Ge concentration than to recess depth. Qualitatively, one can notice the sharp increase in stress as the L_{pp} is sized upto 3 times its original value (i.e. K=3). The increase in stress is marginal once the value of K goes beyond 5. It is widely accepted that the mobility of the PMOS device is directly proportional ([2] even observed super-linear dependence) to the uniaxial compressive stress[2] in the channel. Therefore the change in mechanical stress, S_K can be mapped on to an equal change in the mobility. From basic device knowledge, we know that the equivalent resistance of a device is given as follows [10].

$$R_{ON} = \frac{V_{DD}}{\mu C_{OX} W \left(V_{GS} - V_T - 0.5 V_{DSAT}\right)} \quad (1)$$

Eqn 1 implies that the equivalent ON resistance of a device is inversely proportional to its mobility. Hence increasing the mobility (by sizing up L_{pp}) decreases the equivalent resistance of the device. This equivalent resistance can be used for switch level RC analysis of the circuit. Let the resistance of the PMOS device scale down by a multiplicative factor F when the L_{pp} distance is increased by K times. Owing to inverse relationship, F is simply the inverse of S_K of Figure 2.

Using the tool GNUplot, we performed curve fitting between F and L_{pp} for the data points corresponding to various Ge concentration and recess depths in Figure 2. For each of these cases, the data point fit with high fidelity ($\geq$ 99.9%) to the functional form

$$F = \frac{K}{A \times K + B}$$

where A and B are constants. Putting the boundary condition that K=1 (which means no increase in *AA* size) would correspond to F=1 (which means no decrease in series PMOS resistance), we get the relationship B=1−A giving

$$F = \frac{K}{A \times K - A + 1} \quad (2)$$

From above, we note that A is the only parameter relating F and K. The value of A will depend on the Ge concentration, recess depth

[1]If for a particular fab this value is different, all that changes in this paper is the base value to which we normalize all stress values

[2]SiGe in S/D region injects uniaxial (only along the channel) stress. STI which surrounds the device area injects bi-axial stress. For bi-axial stress, mobility vs stress curve is in general not linear [9].

and other process parameters. We will refer to the parameter A as *Single Fitting Parameter* (SFP) here on. A=1 would mean PMOS resistance is independent of *AA* sizing. A <1 implies *increase* in PMOS resistance with increasing *AA* size. A >1 implies PMOS resistance *decreases* with increasing *AA* size. Of course, in our case A >1 and we will use this property to prove the convexity of concurrent Gate and *AA* sizing later in the paper.

For a common value of Ge concentration such as 20% [2] and recess depth of 120nm, Figure 3 shows the data points and the fitted curve. The value of *SFP* is found to be 3.4 by curve fitting. For a different SiGe recipe, this constant may be different but it wont affect the general trends of our results. This leads to the final expression as

$$F = \frac{K}{3.4K - 2.4} \tag{3}$$

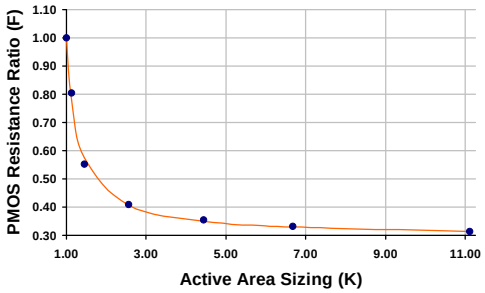

Figure 3: Curve fit for decrease in PMOS series resistance vs increase in L_{pp} for 20% Ge and recess depth 120nm

The above model allows predicting reduction in the series resistance of the PMOS transistor as a function of the increase in length of *AA* (i.e. L_{pp}) around it. We will use the model represented in Eqn 2 for theoretical derivation and that in Eqn 3 when numerical values are required.

With the above understanding and model, we will develop *AA* sizing aware cell delay models. For the sake of brevity, we will take an example of 2-input NAND gate but at the end of this section we will provide the final results for other gates such as 3-input NAND, 2-input NOR, 3-input NOR and inverter gate. Consider the toy layout of a 2-input NAND gate shown in left half of Figure 4 corresponding to the transistor level schematic in top-right. Let symbol R and

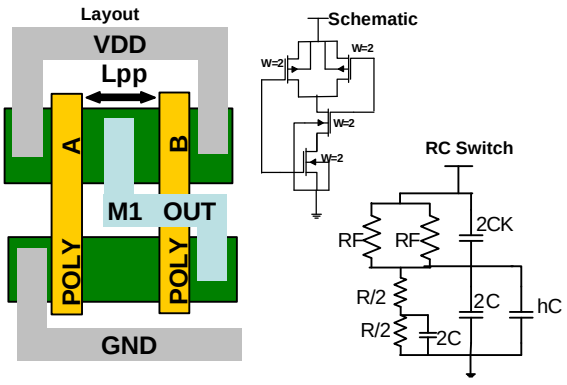

Figure 4: Schematic, Layout and RC Switch models for 2-input NAND Gate. Note the impact of *AA* sizing in switch model

C denote the resistance and capacitance of a unit width NMOS device. Assuming that the original L_{pp} of the PMOS (as shown in the figure) is $90nm$ and it is increased to K times leading to F times decrease in series PMOS resistance, the RC switch level model of the NAND gate driving a fanout load of hC can be represented as the bottom-right diagram in Figure 4. Note that the resistance of the PMOS has been scaled (down) F times ($F \leq 1$) due to *AA* sizing and the diffusion capacitance of the PMOS has been scaled (up) by K ($K \geq 1$) times to account for larger *AA*.

Using the notation that the delay of a cell is the average of the fall and rise times [10], we obtain the *AA* sizing aware 2-input NAND gate delay as

$$D(K) = RC\,(1+F)\,(1+K+0.5h) + 0.5RC \tag{4}$$

Replacing F in terms of *SFP* from Eqn 2 transforms above into

$$D(K) = RC\left(\frac{(A+1)\,K - A + 1}{AK - A + 1}\,(1+K+0.5h) + 0.5\right) \tag{5}$$

The nominal delay of the 2-input NAND cell without any *AA* sizing can be obtained by substituting $K = 1$ in Eqn 5 i.e. it it equal to D(1). Define $\Delta D(K)$ as the ratio $D(K)/D(1)$. For a given value of *SFP* and fanout load h, $\Delta D(K)$ remains only a function of the *AA* sizing. Consider the case of $A = 3.4$ which corresponds to the fit derived in Eqn 3. Under this scenario, we plot the value of $\Delta D(K)$ vs K in Figure 5 for various capacitive load. The term "FO4" refers to fanout of 4 and so on.

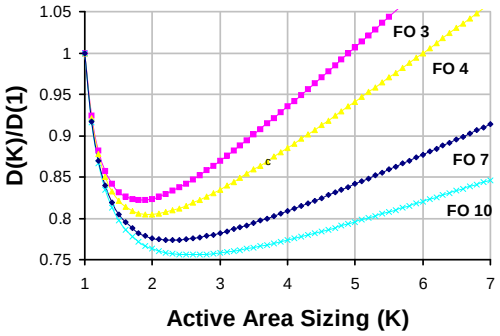

Figure 5: Delay decrease (normalized to unstretched cell) vs the extent of *AA* sizing. Beyond a certain limit, *AA* capacitance's increase overcomes PMOS resistance decrease.

From Figure 5, we observe that starting from $K = 1$ (i.e. L_{pp} = 90nm), the delay decreases monotonically until a particular value of K depending upon the fanout load. Let's call this point K_{opt}. Higher is the capacitive load on the cell, more is the value of K_{opt} meaning that highly loaded cells require larger *AA* size. The delay of the 2-input NAND gate at their optimal delay point is approximately 25% lesser for a fanout of 10 loading and 20% lesser for a fanout of 4 loading. Lets call this delay as D_{opt}. Sizing up the *AA* beyond after corresponding K_{opt} does not improve the delay because the increase in *AA* capacitance outweighs the benefit of reduced PMOS resistance (see Figure 4). In fact, if the *AA* is sized beyond a certain limit, the delay of the gate becomes even bigger than the gate without *AA* increase. Lets call this point as K_{max} and as an example, K_{max}=6 for 2-input NAND gate with fanout of 4 loading in Figure 5. Our analysis can be easily applied to other basic and more complex gates by analyzing its layout and making the equivalent RC switch level network. Based on the method above, we computed K_{opt}, D_{opt} and K_{max} for frequently used gates (NAND, NOR etc) and are shown in Table 1.

Let us call Eqn 4 as the *characteristic delay equation* for a 2-input NAND gate. On similar lines, the characteristic delay equations of other common gates (after derivation following the steps above) are presented in Table 2.

We observe that for the gates in Table 1, the delay of the gate can be reduced by around 17% on an average for fanout of 4 loading and by 23% for fanout of 10 loaded gates. Such a significant decrease in delay of the cell can be very useful for high performance designs. We

	Fanout 4			Fanout 10		
Name	K_{opt}	D_{opt}	K_{max}	K_{opt}	D_{opt}	K_{max}
INV	1.78	0.82	4.63	2.31	0.76	9.54
2-NAND	1.97	0.80	6.00	2.57	0.75	12.45
3-NAND	1.72	0.84	4.22	2.20	0.78	8.31
2-NOR	1.66	0.83	3.79	2.11	0.78	7.46
3-NOR	1.62	0.84	3.55	2.03	0.78	6.66
AVERAGE	1.75	0.83	4.43	2.24	0.77	8.88

Table 1: Optimal *AA* sizing, optimal delay (normalized to cell without *AA* sizing) and maximum sizing for common gates

Name	Equation
INV	$RC\left(F+1\right)\left(K+0.5h+0.5\right)$
2-NAND	$RC\left(\left(F+1\right)\left(K+0.5h+1\right)+0.5\right)$
3-NAND	$RC\left(\left(F+1\right)\left(2K+0.5h+1.5\right)+1.5\right)$
2-NOR	$RC\left(\left(F+1\right)\left(2K+0.5h+0.5\right)+FK\right)$
3-NOR	$RC\left(\left(F+1\right)\left(3K+0.5h+1\right)+3FK\right)$

Table 2: *AA* Sizing aware Elmore Delay Equations for some common gates

also infer that each cell's L_{pp} needs to be sized up by approximately 75% to 125% of its original value for obtaining the best performance.

4. TIMING OPTIMIZATION BY *AA* SIZING

In this section, after introducing basics of gate sizing and repeater insertion, we will enhance these techniques by performing them aware of *AA* Sizing. As the results will show, these techniques when applied simultaneously with *AA* sizing can push the circuit performance further by 10%. Optimal repeater insertion (ORI) and gate sizing (GS) are the two key techniques routinely used by designers for both microprocessor and ASIC design methodology to perform timing and noise closure. Therefore, combining them seamlessly with *AA* sizing is very beneficial from the practical point of view.

4.1 Basics of ORI and GS

In the face of increasing integration and technology scaling, interconnects have been rendered longer and with higher impedance in each generation. ORI problem entails dividing the long interconnect into optimal number of parts and inserting a repeater of optimal gate size so as to reduce total delay over the interconnect. ORI reduces the delay of a global interconnect of length L from a L^2 relationship to linear relationship. This partially alleviates the problem of long global interconnect. Under assumption of same type and structure of repeaters, ORI problem can be solved analytically [10] by obtaining an expression of total delay through the interconnect in terms of the variables such as number of repeaters, gate size of each repeater. This expression can then be minimized w.r.t. the variables to obtain the best configuration of repeaters.

GS problem targets finding the optimal gate size of individual gates in a combinational module so as to minimize an objective such as delay or power under a delay constraint. Consider the channel formed under the gate region (shown as rectangular for simplicity) in Figure 6. The channel, source, drain and gate regions marked as C, S, D and G respectively. The target is to flow current from point S in source region to point D in drain region. The resistance offered to the flow of this current, R_{ch}, is given by

$$R_{ch} = \frac{\rho \times L}{W \times T} \tag{6}$$

Gate sizing and *AA* sizing both aim to reduce the R_{ch} to increase the performance of the device. However, there are several major differences in these techniques as highlighted below:

1. Gate sizing entails modulating the dimension W which reduces R_{ch}. On the other hand, *AA* sizing modifies the dimension L_{AA} which engenders as reduced ρ.

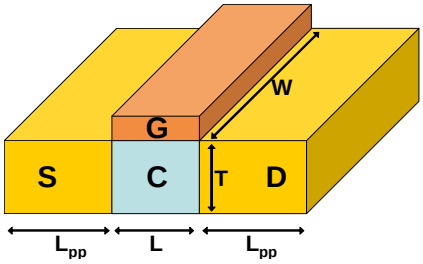

Figure 6: Physical structure of a transistor

2. Gate sizing changes the capacitive load seen by the driver of the current gate. Thus changing the size of a gate has ripple effect on the gate in its fanin cone. On the contrary, *AA* sizing changes the capacitance of the source/drain region which is a *local* effect in the sense that the gates in the fanout and fanin cone do not get affected by increasing L_{pp}.

3. Due to the fixed cell height constraint (dimension W in Figure 6) in a standard cell type of environment, it may not be feasible to perform gate sizing (changing W) in a continuous way. On the other hand, *AA* sizing is more amenable/flexible because of there is no such hard constraint on *AA* size. [3]

4.2 Concurrent *AA* Sizing and ORI

Consider an interconnect schematically represented in the top part in Figure 7 with L, R_w, C_w as its the total length, per unit length resistance and the per unit length capacitance respectively. Assume that M repeaters are inserted in it, i.e. each repeater drives an interconnect of length L/M. Our aim is to achieve fastest delay[4] from the driver to the sink of the interconnect. If S and K represent the gate and *AA* sizing factors for each of these repeaters, its RC switch level model is as shown in the lower part of Figure 7. The resistance reduction factor F is shown in the expression of series resistance of the PMOS transistor. The interconnect segment is modeled as a π network.

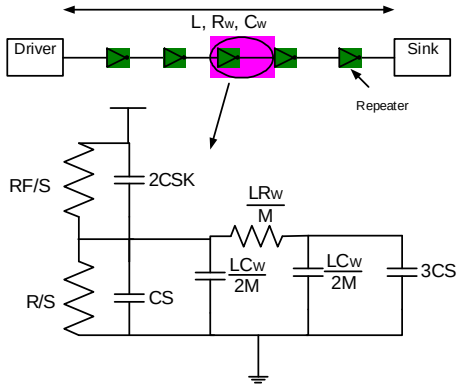

Figure 7: Schematic of a long interconnect with repeaters inserted (top). Switch level RC model for the repeater considering π model of interconnect (bottom)

The switched capacitance and Elmore delay of *one stage* as well as the delay of the whole inverter chain, D_{tot} can be respectively computed from Figure 7 as

[3]We note that to reduce cell library size, both gate sizing and *AA* sizing would need to be discretized, but gate sizing is still *more* discontinuous due to hard constraint on cell height

[4]For objective other than delay minimization, similar approach can be followed by using our switch level RC model

$$C_{stage} = 2CSK + 4CS + LC_w/M \quad (7)$$

$$D_{stage} = \frac{(F+1)\,RC_{stage}}{2S} + \frac{R_w L}{M}\left(\frac{LC_w}{2M} + 3CS\right) \quad (8)$$

$$D_{tot} = M \times D_{stage} \quad (9)$$

4.2.1 Optimizing Delay

To minimize D_{tot}, its partial derivatives with respect to gate size, number of stages and *AA* sizing should be 0. Solving for $\frac{\partial D_{tot}}{\partial S} = \frac{\partial D_{tot}}{\partial M} = 0$ and after some algebra, we obtained the following results.

$$M_{opt} = L\sqrt{\frac{C_w R_w}{2RC(F+1)(K+2)}} \quad (10)$$

$$S_{opt} = \sqrt{\frac{RC_w(1+F)}{6R_w C}} \quad (11)$$

where M_{opt} and S_{opt} represent the optimal number of repeaters and sizing of each of them. As the expressions of S_{opt} and M_{opt} are independent of each other, they can be independently set to their optimal value. Apart from the fixed circuit parameters (i.e. R, C, R_w, C_w and L), M_{opt} and S_{opt} depend on the value of K (the amount by which the *AA* is sized up)[5]. This is the major difference between traditional ORI and ORI considering the freedom of *AA* sizing.

As the *AA* is made bigger (i.e. K increases), the value of series resistance multiplicative factor F decreases (see Eqn 3). Thus, the value of M_{opt} is affected by two reverse trends: $K+2$ term increases whereas $1+F$ term decreases. On the other hand, the trend of S_{opt} is straightforward: as the *AA* is made bigger, $1+F$ term decreases thus decreasing the optimal gate sizing value S_{opt}. The reduction of S_{opt} with increase in *AA* sizing is intuitive because as K increases, more and more performance increase comes from *AA* increase thus requiring increasingly smaller gate sizes. Figure 8 shows the value of M_{opt} and S_{opt} as a function of K normalized to their value without any *AA* sizing up i.e. at K=1.

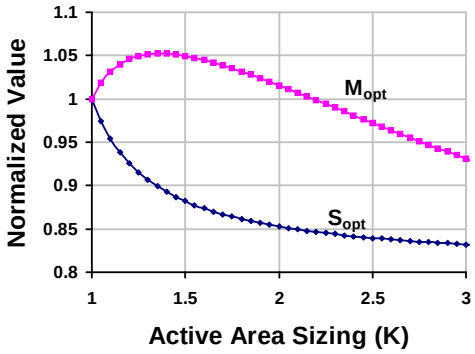

Figure 8: Optimal repeater size (S_{opt}) and number of repeaters (M_{opt}) as a function of *AA* sizing (K).

From Figure 8 we observe that for a gate with slightly larger *AA* (say 25%, corresponding to K=1.25) the number of repeaters for optimal delay is more ($\sim$ 5%) than its counterpart without *AA* sizing. On substituting the optimal number of stages (M_{opt}) from Eqn 10 and optimal inverter size (S_{opt}) from Eqn 11 into Eqn 9, we obtain D_{tot} as

$$D_{tot} = \sqrt{2L^2RCR_wC_w} \times \sqrt{1+F} \times \left(\sqrt{3} + \sqrt{2+K}\right) \quad (12)$$

At this stage, the only unknown independent variable in the above equation is K whose optimal value can be found by solving for $\frac{\partial D_{tot}}{\partial K} = 0$. Using the generic expression for F in terms of K from Eqn 2 into Eqn 12 results in an unwieldy ninth order expression for K which does not provide any insight. Therefore, we computed the value of K_{opt} for a range of numerical values of the *SFP* A. Qualitatively, higher is the value of the *SFP*, sharper is the decrease in series resistance of the PMOS, thus larger is the value of *AA* sizing before the increase in *AA* capacitance starts overcoming the benefit of reduced series resistance. This trend is clear in Figure 9 which shows the value of K_{opt} as a function of the *SFP* using Eqn 3. As

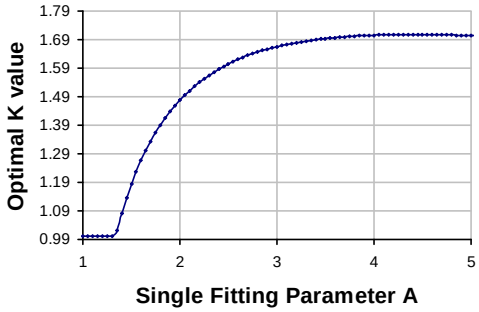

Figure 9: Dependence of optimal *AA* sizing factor K_{opt} on the *SFP* A. Larger A can support larger *AA* increase.

expected, the optimum value of *AA* size, K_{opt} increases as the *SFP* increases. Further, from Figure 9 we observe that once the value of *SFP* is more than a certain value (approximately 3.5 in Figure 9), the dependence of K_{opt} on *SFP* becomes very weak. For very small values of *SFP* ($A \leq 1.3$), the reduction in PMOS resistance is much lesser than the increase in capacitive loading, thus the optimal value of K is equal to 1 i.e. the cell should *not* be made with larger *AA* at all. As long as the functional fit of the form Eqn 2 holds true, the optimal extent of *AA* sizing, K_{opt} can be read off Figure 9.

At this stage, we are ready to solve the *AA* sizing aware ORI problem completely. The result of this problem includes: a) *AA* sizing of each repeater, b) gate sizing of each repeater, c) number of repeaters, d) resultant delay of the interconnect, e) power consumption. For the ease of understanding the impact *AA* awareness had on repeater insertion problem we will present the results normalized to the case of ORI without any *AA* sizing.

- For the case of numerical fit obtained in Eqn 3 the value of K_{opt} can be seen from the Figure 9 as

  $$K_{opt}|_{A=3.4} = 1.69 \quad (13)$$

  This means that the active area of each inverter cell should be increased to 1.69 times its original value for achieving shortest delay. This value of K is used to get other results.

- Using K=1.69 into Eqn 10, the optimal number of repeaters, M_{opt}, is 4% more than its value without any *AA* increase. Placement tools should consider this while whitespace allocation.

- Eqn 11 results in a value of S_{opt} which is 13% *smaller* than its nominal value without *AA* size increase. This implies that inverters with smaller width, but larger *AA* are required for least delay.

- In Figure 10 we plot the fastest delay through the interconnect (normalized to least delay without *AA* sizing) as a function of sizing. At $K = 1.69$, the delay through the long interconnect is reduced by approximately 9%.

- The dynamic power for *each* repeater is given as $C_L V_{DD}^2 f$ where C_L is the switched capacitance as given in Eqn 7 and f, the frequency, can be expressed as $\frac{1}{D_{rptr}}$ where D_{rptr} is

[5]Since the value of F depends only on K as per Equation 2

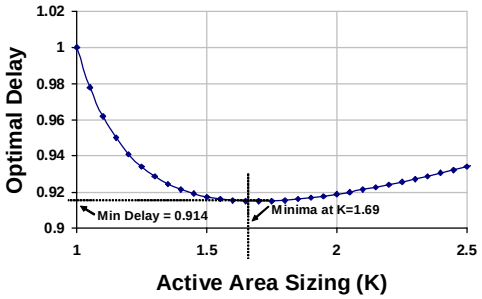

Figure 10: Normalized delay through the repeater chain vs the *AA* sizing of each repeater

the delay through the repeater. The ratio of dynamic power in the complete repeater chain incurred due to increased *AA* size can be calculated as

$$\frac{P_{stretch}}{P_{nominal}} = \frac{[M \times 1/D_{rptr} \times C_L]_{K=1.69}}{[M \times 1/D_{rptr} \times C_L]_{K=1.00}} \quad (14)$$

Substituting Eqn 11 in Eqn 7 and using it along with Eqn 10, Eqn 12 and after some algebra, the above equation condenses into an elegant simple relation below

$$\frac{P_{stretch}}{P_{nominal}} = \frac{\left(1/\sqrt{1+F}\right)|_{K=1.69}}{\left(1/\sqrt{1+F}\right)|_{K=1.00}} = 1.15 \quad (15)$$

Thus, the total dynamic power dissipation of the repeater chain increases by 15% due to *AA* sizing. This increase comes due to three components: a) increased number of repeaters (4%), b) increased *AA* capacitance (2%) and c) increased frequency of operation(9%).

Summary: Table 3 summarizes the results of *AA* sizing aware ORI for minimizing delay of long interconnect. The minimum achievable delay, optimal *AA* size, number of repeaters, repeater sizing, minimum achievable delay and dynamic power dissipation is reported where each of these is normalized to the case without the flexibility of exploiting *AA* sizing. These parameters uniquely define the solution of ORI.

D_{tot}	AA Sizing	Gate Sizing	# Rptrs	Power
0.91	1.69	0.87	1.04	1.15

Table 3: Solution of Active Area Sizing aware ORI Problem. (normalized w.r.t. the values without increasing *AA* size)

4.2.2 Minimizing Repeater Number

AA aware repeater insertion can be used to reduce the repeaters inserted in the iso-delay case w.r.t. repeater insertion without *AA* sizing. Let us assume that the delay through an interconnect after performing ORI alone (i.e. without *AA* sizing) is D_{ORI}. If ORI and *AA* sizing was performed concurrent as in the previous section, let the resultant delay be D_{ORI+DS} (note that $D_{ORI+DS} < D_{ORI}$) . We want to find out by how much can we reduce the number of repeaters in the second case so that the sub-optimal delay D_{ORI+DS} becomes just equal to D_{ORI}. This scenario is of importance for the global interconnects which are not the most critical ones. Since repeaters are power hungry elements and their insertion cause substantial difficulty in physical synthesis, reducing their number is advantageous.

Solving ORI without *AA* size increase (by substituting K=1 in Eqn 10 and Eqn 11 and putting these values back in Eqn 8) yields the result

$$D_{ORI} = 4L\sqrt{3RCR_wC_w} \quad (16)$$

Now, consider the case of ORI concurrently with *AA* sizing. We can put S=S_{opt} and F corresponding to K=1.69 but instead of assigning M=M_{opt}, let M=δM_{opt} where $\delta \in (0, 1]$. In such a case, it can be shown that

$$D_{ORI+DS} = L\sqrt{RCR_wC_w} \times \left(3 + \frac{2}{3\delta} + \frac{2\delta}{3}\right) \quad (17)$$

Summary: Solving Eqn 16 and Eqn 17, we obtain $\delta = 0.55$. This means that by using concurrent ORI and *AA* sizing, we can save nearly half (45%) of repeaters without sacrificing any performance w.r.t. ORI without *AA* size increase. This is a very interesting result and can have substantial impact due to aforementioned benefits of reducing repeater count.

4.3 Concurrent Gate and *AA* Sizing

Gate sizing under Elmore delay is a well under-stood problem and can be solved using convex solvers by exploiting the property of a class of function called *posynomials*, that allows conversion of the optimization problem into a convex programming problem with a straightforward transformation [11]. A larger gate is more effective at driving big capacitive loads at the cost of presenting higher load to its input gate. Based on the understanding of comparison between gate and *AA* sizing in Section 4.1, an obvious question that springs up is about the possibility of combining these two techniques for faster circuits. In section 4.2, we analytically combined ORI and *AA* sizing. That problem was amenable to analytical solution due to the fact that all repeaters were similar to each other. In a general circuit, there can be several different cells, multiple fanin and fanout gates etc and trying to analytically solve for minimum delay is impossible. We thus formulate the Concurrent Gate and Active area Sizing (CGAS) problem as

4.3.1 CGAS Formulation

CGAS: Given a multilevel circuit to be fabricated with SiGe S/D technology, perform concurrent gate and *AA* sizing (CGAS) to minimize an arbitrary convex[6] cost function

Consider Gate 1 in Figure 11 with tuple {S1, C1, K1} which represents its gate sizing, gate capacitance per input pin and *AA* sizing respectively. Gate 1 drives Gate 2 and 3 associated with their tuples

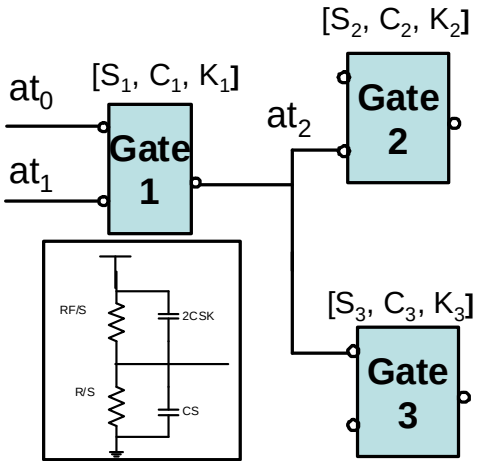

Figure 11: A part of logic circuit under consideration. Gate 1 drives Gate 2 and 3.

{S2, C2, K2} and {S3, C3, K3}. For the sake of clarity, consider

[6]As we will prove later, all the constraints of CGAS are convex therefore having a convex cost function guarantees the problem to be convex and thus can be solved optimally.

for now that Gate 1 is an inverter whose *AA* sizing aware RC switch level model is given in the inset in Figure 11. $a1$ through $a4$ are the signal arrival times at the different locations in the figure. The delay of Gate 1 can be written as

$$D_1 = RC\frac{(1+F_1)}{2}\left(1+2K_1+\frac{(S_2C_2+S_3C_3))}{S_1}\right) \quad (18)$$

where F_1 depends on K_1 as in Eqn 2. Note that S_2C_2 + S_3C_3 can also be written as $\sum_{m\in FO_1} C_mS_m$, where FO_i respresents the fanout of gate i. Using Figure 11, we can write the following equations for the arrival times.

$$at_0 + D_1 \leq at_2$$
$$at_1 + D_1 \leq at_2$$

where D_1 is given in Eqn 18. Now consider the case when Gate 1 is *not* an inverter but some other logic gate. Under such a scenario we can write a generic expression (see particular examples in Table 2) of the delay of gate *i* by generalizing Eqn 18 as

$$D_i = RC\frac{(1+F_i)}{2}\left(a+bK_i+\sum_{m\in FO_i}\frac{c_mS_m}{S_i}\right) \quad (19)$$

where a, b and various $c_m \in \mathbf{R}^+$.

Let us now consider an unoptimized multi-level circuit in which I, O, L represent the set of input pins, output pins and internal logic gates respectively. For each gate i, let at_i, S_i, K_i, FI_i, FO_i denote its arrival time at its output, gate sizing, *AA* sizing factor, fanin gates and fanout gates respectively. *CGAS* can now be written as a mathematical optimization program as

$$\begin{aligned} &Minimize: Delay \\ &Subject\ To: \\ &\quad at_j + D_i \leq at_i \quad \forall i \in L, \forall j \in FI_i \\ &\quad at_i = 0 \quad \forall i \in \{I\} \\ &\quad Delay > at_i \quad \forall i \in \{O\} \\ &\quad at_i > 0 \quad \forall i \in \{L \cup O\} \\ &\quad S_i, K_i > 1 \quad \forall i \in \{L \cup O\} \end{aligned} \quad (20)$$

The dummy variable $Delay$ represents the largest of the arrival times among various output pins of the circuit in the third set of constraints. Therefore, minimizing $Delay$ is equivalent to making the circuit as fast as possible. Though the above formulation is targeted for fastest possible circuit implementation, we note that any other traditional objectives can be handled easily too. For example, for optimizing power consumption, we can perform *AA* minimization under a delay constraint by adding one constraint for the required arrival time of the circuit.

4.3.2 Convexity of CGAS

The objective function as well as all constraints except the first set of constraints in the *CGAS* are convex by observation. We focus on the first set of constraints and show that it is posynomial. Using Eqn 19 for the value of D_i and Eqn 2 for the value of F, each of the first set of delay constraints can be written as

$$\left(1+\frac{K_i}{AK_i-A+1}\right)\times\left(a_i+b_iK_i+\sum_{m\in FO_i}\frac{c_mS_m}{S_i}\right) \leq \frac{2\,(at_i-at_j)}{RC} \quad (21)$$

such that A, a, b, c_m, S_m and $S_i \in \mathbf{R}^+$. Substituting AK_i-A+1=T_i, RHS by Δat and S_m/S_i by S_{mi}, the above can be written as

$$\left(1+\frac{1}{A}+T_i\left(1-\frac{1}{A}\right)\right)\times\left(a+\frac{bT_i}{a}+b\left(1-\frac{1}{a}\right)+\sum_{m\in FO_i}c_mS_{mi}\right) \leq \Delta at \quad (22)$$

On cross multiplying, the *LHS* of the above equation can be separated in terms T_i and various S_{mi} as

$$MNT_i^2+(NL+PM)\,T_i+MT_i\sum_{m\in FO_i}c_mS_{mi}+L\sum_{m\in FO_i}c_mS_{mi} \quad (23)$$

where, L=1 + $\frac{1}{A}$, M=1 − $\frac{1}{A}$, N=$\frac{b}{a}$ and P=$a + b\left(1-\frac{1}{a}\right)$. L, N and M (since A>1 for Eqn 3) $\in \mathbf{R}^+$ by inspection. Using this, the coefficients for T_i^2, $S_{mi}T_i$ and $S_{mi} \in \mathbf{R}^+$. To prove that the above expression is posynomial we need to prove that the coefficient of T_i $\in \mathbf{R}^+$. Since the value of a can be less than 1 (e.g. INV and NOR gate in Table 2) this cannot be done simply by inspection. Thus, we simplify the coefficient of T_i

$$\begin{aligned} NL+PM &= \frac{b}{a}\left(1+\frac{1}{A}\right)+\left(1-\frac{1}{A}\right)\left(a+b\left(1-\frac{1}{a}\right)\right) \\ &= \frac{1}{A}\,(A-1)\,(a+b)+\frac{2b}{a} \end{aligned}$$

From the above form (since A>1), it is clear that the coefficient of T_i also $\in \mathbf{R}^+$ and thus each constraint in *CGAS* is the form of a posynomial. Under a elementary variable transformation using exponential functions, this constraint can be mapped into a convex constraint [11]. Therefore problem *CGAS* is convex and can be solved with existing convex program solvers optimally.

4.3.3 Power Considerations

Though increasing *AA* size reduces the PMOS resistance, it overall increases the *AA* capacitance which impacts the dynamic power dissipation of the design. To quantify the impact of our technique CGAS has on dynamic power (w.r.t. to technique GS), we computed the total switched capacitance for gate sizing, C_{GS} and concurrent gate and *AA* sizing , C_{CGAS}, as

$$C_{CGAS} = \sum_{i\in L}\left(A_i+B_ik_i+\sum_{m\in fo_i}\frac{C_ms_m}{s_i}\right) \quad (24)$$

C_{GS} can by found evaluating $C_{CGAS}|_{\forall k_i=1.00}$. If CGAS can result in a circuit which runs T times faster than the circuit realized with conventional GS, the increase in dynamic power consumption of the new circuit would then be given as

$$\Delta P_{dyn} = \frac{T \times C_{CGAS}}{C_{GS}} \quad (25)$$

Note that since the underlying canonical circuit and logic structure remains the same for the two differently sized circuit realizations, they will have exactly same switching factors for each node. Therefore, we can safely ignore the impact of scaling each capacitance value by corresponding switching factor. Given the solution of of *GS* and *CGAS*, it is possible to find out the value of ΔP_{dyn}. We followed a 2-step procedure to calculate C_{CGAS} and C_{GS} . In the first step, *CGAS* program of Eqn 20 was solved to achieve minimum possible delay. Let this minimum delay be D_{min}^{CGAS}. In the second step, the objective function in Eqn 20 was modified to minimize C_{CGAS} under the constraint that the timing of the resultant circuit is at most D_{min}^{CGAS}. This way, the circuit realization with least switching capacitance subject to timing constraints was obtained. In the same fashion, the minimum dynamic power for circuit realization obtained by traditional GS was also found out. The obtained capacitance values for these two cases determine ΔP_{dyn}.

4.3.4 Experimental Setup and Results

We solved the *CGAS* problem on a variety of benchmarks from IWLS 1991 LGSynth suite [12]. These benchmarks are implementations of a soup of logic blocks and we believe that such mixture of benchmarks can counter structural bias present in them. The original *blif* format of the benchmarks were optimized and technology mapped using SIS program [13]. For the sake of simplicity we restricted our library to a set of 3 gates: 2-input NAND, 2-input NOR and an Inverter. The characteristic delay equations of these gates are in Table 2. We note that typically a cell library can have several other functionally unique cells and a typical industrial application of our technique will require designer to extend Table 2 accordingly for cells such as XOR, AOI with similar results. A C++ program was used to write out the objective and constraints for the tool AMPL [14] which was coupled with the back-end solver MOSEK [15] that solves the convex program using interior point optimization method. Table 4 shows the comparison of efficacy of Gate Sizing (GS) alone and concurrent Gate and Active area Sizing (CGAS) for timing optimization. Column "# Gates" shows the total number of gates in each benchmark. The timing achieved with the respective techniques is shown in Columns "Delay" with the delay enhancement of *CGAS* over *GS* shown in column "Improv". Column "ΔCap" and "ΔP_{dyn}" show the increase in switching capacitance and dynamic power of solution of *CGAS* over *GS*.

Design	Gates	Delay (times RC)		Imprv	ΔCap	ΔP_{dyn}
		GS	CGAS	%	%	%
C6288	3316	1320.65	1175.15	11.01	3.20	14.67
C880	502	340.83	309.19	9.03	0.38	9.44
frg1	149	178.44	159.34	10.70	0.19	10.82
k2	1163	323.06	295.13	8.65	0.49	9.19
C7552	2581	734.07	687.49	6.40	0.43	6.85
large	481	262.51	236.90	9.75	0.30	10.08
vda	628	222.38	199.76	10.17	0.57	10.80
des	3759	270.87	233.22	13.89	1.34	15.43
C5315	2007	449.92	400.20	11.05	1.54	12.78
Avg.				10.08	0.94	11.17

Table 4: Performance Improvement and Dynamic power increase for various benchmarks

The results in Table 4 are very encouraging. We observe that *CGAS* can push the envelop of performance by more than 10% over and above the optimal delay value achievable by traditional gate sizing. For high performance designs, this is very attractive. A 10.08% decrease in design cycle time corresponds to more than 11.2% increase in the chip operating frequency. On comparing the increase in switched capacitance of each benchmark, we found that the capacitance increase is very modest at under 1%. This number is very much dependent on the structure of the benchmark and lies in the wide range of 0.2% to 3.2% for different benchmarks. Results show that the dynamic power increase due to *CGAS* is around 11% on an average. Out of it, nearly 10% increase is simply because the design can switch at a higher frequency of operation. The rest 1% increase is due to capacitance increase for the larger *AA* region. It should be noted that in the absence of *CGAS*, it is impossible to improve performance beyond the optimal solution of traditional gate sizing (without changing V_{DD} or reducing V_{th}). For each of the benchmarks in Table 4, the convex solver took less than a minute to attain the optimal solution.

5. DISCUSSIONS AND CONCLUSIONS

AA sizing for SiGe S/D type devices opens up an exciting dimension for optimizing their performance. Gates with high fanout loads are most benefitted by sizing up *AA* - this could be useful as the interconnect capacitance increases in future technology nodes. Enlarging active area can have mixed impact on the recommendation of using unidirectional poly in layout. For cells whose n-well is larger than p-well (like NAND gate), increasing the active area of PMOS can cause longer wrong way poly routing. On the other hand, for cells which have *smaller* n-well than p-well (like NOR gates), active area sizing of PMOS can actually bring the layout close to unidirectional case. In our results, we observed dynamic power increase of the order of 11%. We believe that for high performance design applications, it is a fair price to pay for getting 10% decrease in design cycle time. Use of *AA* sizing alters footprints of the cells and the design flow should be modified to account for it.c

In this paper, we performed the first gate level systematic study of exploitation of active area (*AA*) dependent mobility of strained CMOS devices. A simple yet accurate empirical model was proposed which fits the observed silicon data very well. The concept of optimal *AA* sizing was introduced and demonstrated using simple gates which showed potential of decreasing delay by 17% for fanout-of-4 load and 23% for fanout-of-10 load. To handle long global interconnect, optimal repeater insertion methodology was significantly enhanced by making it *AA* sizing aware. We derived analytical solutions for this problem for the first time leading to more than 9% decrease in global interconnect delay over and above the solution of traditional repeater insertion. For large scale multi level circuits, the powerful technique of gate sizing was combined with concurrent *AA* sizing for better timing optimization. We proved the convexity of the resultant formulation and solved it on a set of benchmarks to achieve promising timing improvements of more than 10% over the traditional gate sizing results, for only 11% increase in dynamic power.

Future Work: In most of our analysis, discussion on *leakage power* has been conspicuously missing because of the void in proper understanding of impact of strained silicon's layout geometry on leakage power. For the same ON current, the OFF current (leakage current) of a S/D SiGe device can be upto three orders of magnitude [3] *smaller* than that of traditional Si device. However, due to bandgap splitting in SiGe, the leakage of *AA* of a given size is larger for SiGe as compared to Si. Our future work would be to characterize the impact of S/D type SiGe device's layout on the leakage current using device and process simulations.

6. REFERENCES

[1] A. Chakraborty *et al.*, "Layout level timing optimization by leveraging active area dependent mobility of strained-silicon devices," in *DATE*, pp. 849–855, March 2008.

[2] L. Washington *et al.*, "pMOSFET with 200% mobility enhancement induced by multiple stressors," *Electron Device Letters, IEEE*, vol. 27, no. 6, pp. 511–513, June 2006.

[3] G. Eneman *et al.*, "Scalability of the $si_{1-x}ge_x$ source/drain technology for the 45-nm technology node and beyond," in *IEEE Transactions on Electron Devices*, vol. 53, July 2006.

[4] Y.-C. Yeo *et al.*, "Enhanced performance in sub-100 nm CMOSFETs using strained epitaxial silicon-germanium," *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, pp. 753–756, 2000.

[5] W.-S. Liao *et al.*, "Pmos hole mobility enhancement through SiGe conductive channel and highly compressive ild-$hboxSiN_x$ stressing layer," *Electron Device Letters, IEEE*, vol. 29, no. 1, pp. 86–88, Jan. 2008.

[6] H. Yang *et al.*, "Dual stress liner for high performance sub-45nm gate length SOI CMOS manufacturing," *IEDM Technical Digest*, pp. 1075–1077, Dec. 2004.

[7] V. Joshi *et al.*, "Stress aware layout optimization," in *ISPD 2008*, pp. 168–174.

[8] *Lithography Report, pp 17*. http://www.itrs.net/Links/2007ITRS/2007_Chapters.

[9] G. Sun *et al.*, "Hole Mobility in Silicon Inversion Layers: Stress and Surface Orientation," *Journal of Applied Physics*, vol. 102, no. 8, p. 084501, 2007.

[10] J. M. Rabaey *et al.*, *Digital Integrated Circuits*. Prentice Hall Press, 2nd ed., 2003.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[12] *IWLS 1991 LGSynth Benchmarks*. http://www.cbl.ncsu.edu/benchmarks/LGSynth91.

[13] E. Sentovich *et al.*, "Sis: A system for sequential circuit synthesis," Tech. Rep. UCB/ERL M92/41, EECS Department, University of California, Berkeley, 1992.

[14] R. Fourer *et al.*, *The AMPL Book*. http://www.ampl.com, 2nd ed., 2002.

[15] *MOSEK Optimization Tool Manual*. http://www.mosek.com.