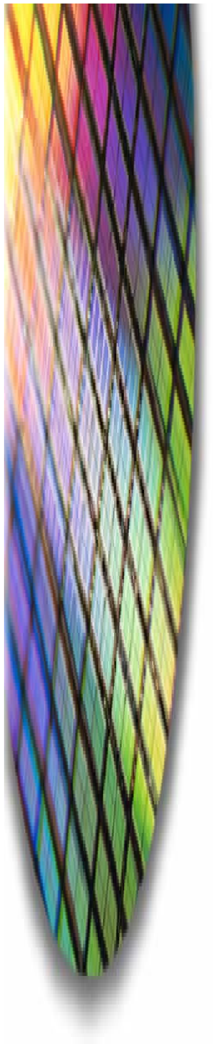


Stress Aware Active Area Sizing, Gate Sizing and Repeater Insertion

Ashutosh Chakraborty
ashutosh@cerc.utexas.edu

David Z. Pan
dpan@ece.utexas.edu

ECE Department,
University of Texas at Austin

Outline

- Intro. to source/drain (S/D) SiGe technology
- Active Area (AA) aware Delay Model
- AA aware Optimal Repeater Insertion (ORI)
- Concurrent AA and Gate Sizing
- Conclusions

Outline

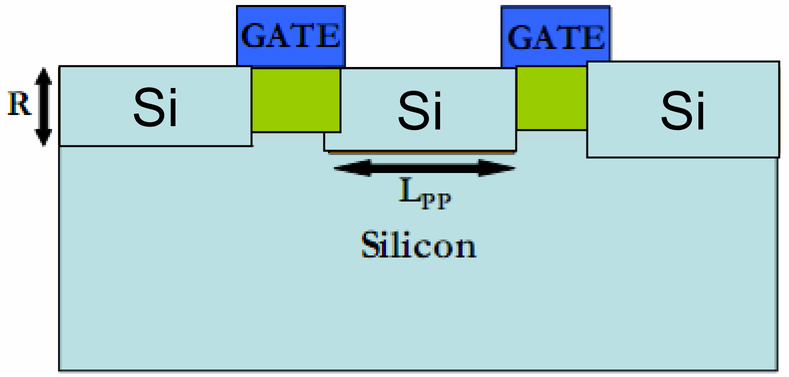
- **Intro. to source/drain (S/D) SiGe technology**
- SiGe Active Area (AA) aware Delay Model
- AA aware Optimal Repeater Insertion (ORI)
- Concurrent AA and Gate Sizing
- Conclusions

Stress/Strain Basics

- Squeezing lattice produces compressive stress
- Pulling lattice apart produces tensile stress
- In direction of charge carrier flow,
  - **Compressive stress improves PMOS performance**
  - Tensile stress improves NMOS performance
  - Larger stress means more performance benefit.

Basics of S/D SiGe Technology

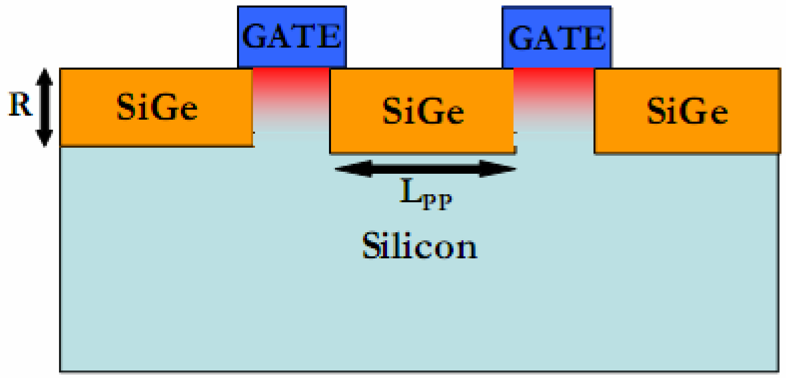
- SiGe instead of Si in S/D regions.
- Imparts compressive strain
- **Increase PMOS speed**



- Popular? Feasible?
  - Yes. Now routinely used by processor manufacturers.
  - AMD using it for 45nm, plans also for 32nm [RTP ’08]
  - Intel used it for 65nm and 45nm [IEDM ‘07]
  - Sony has used it [VLSI Symp ‘08]
  - Manufacturing cost up by only 4% [VLSI Symp 08]

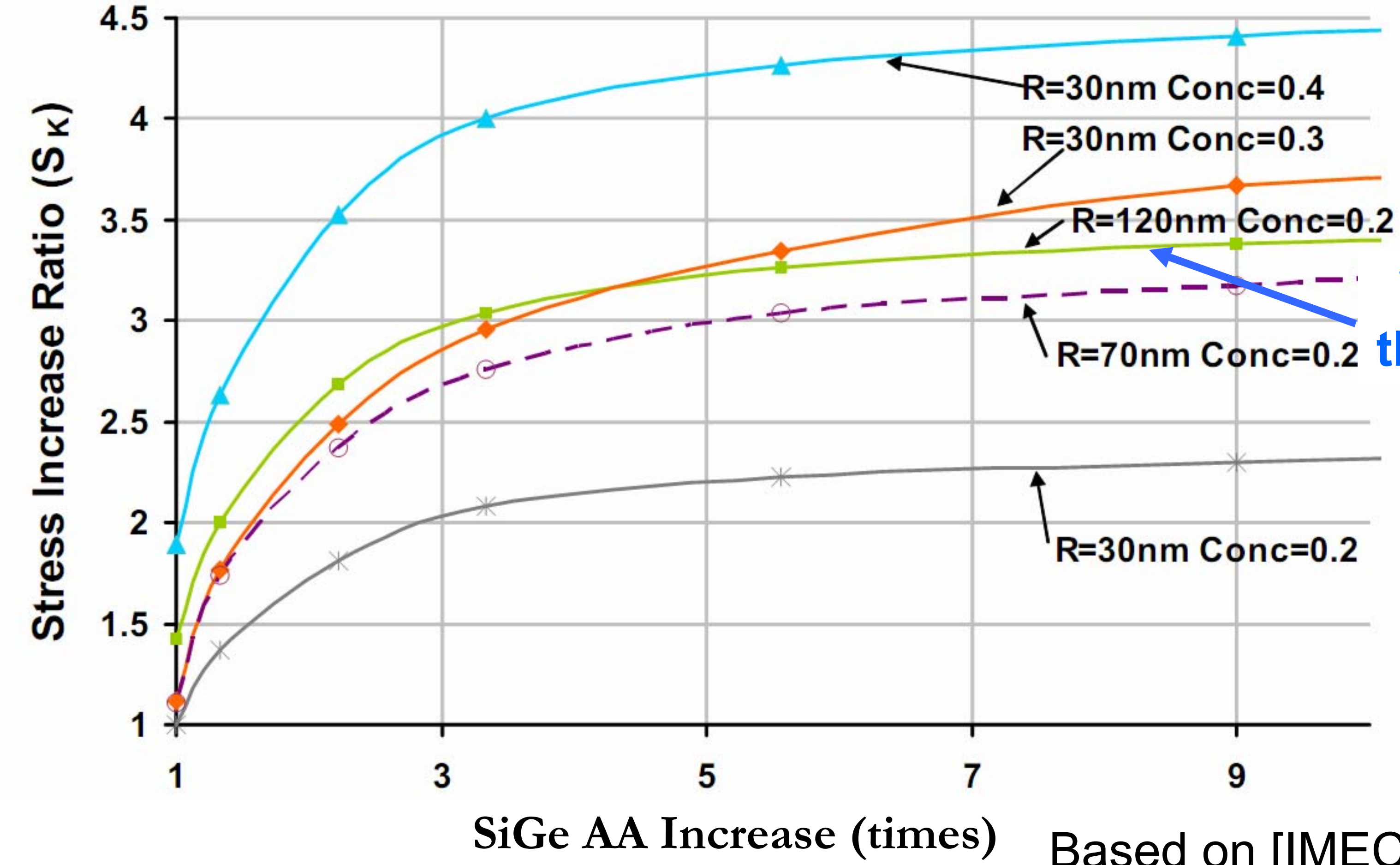
S/D SiGe Aware active area (AA) sizing

- Factor affecting this mobility enhancement:
  - Active area dimension (Lpp)
  - Concentration of Ge
  - Recess depth R



- Layout designer can control active area (AA) size
  - Traditionally, trained to minimize it
  - However, with S/D SiGe, increasing AA helps!

Stress as function of SiGe AA size

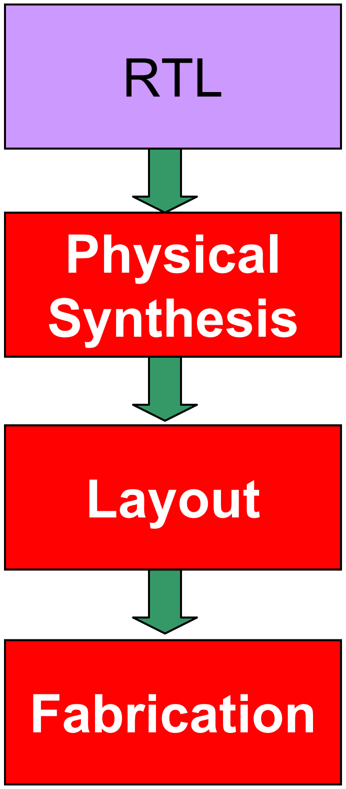


We used this setting

Based on [IMEC, 2006]

Previous Works

- Modeling SiGe AA increase impact
  - Eneman, VLSI Symp '05
  - Simoen, Trans Elect. Dev. '08
  - Applied Materials '07 report
- SiGe AA aware layout optimization
  - Chakraborty, DATE 08
  - Joshi, ISPD 08
  - Joshi, DAC 08
- SiGe AA aware physical synthesis
  - None existing.
  - This work targets this void.



Motivational Example

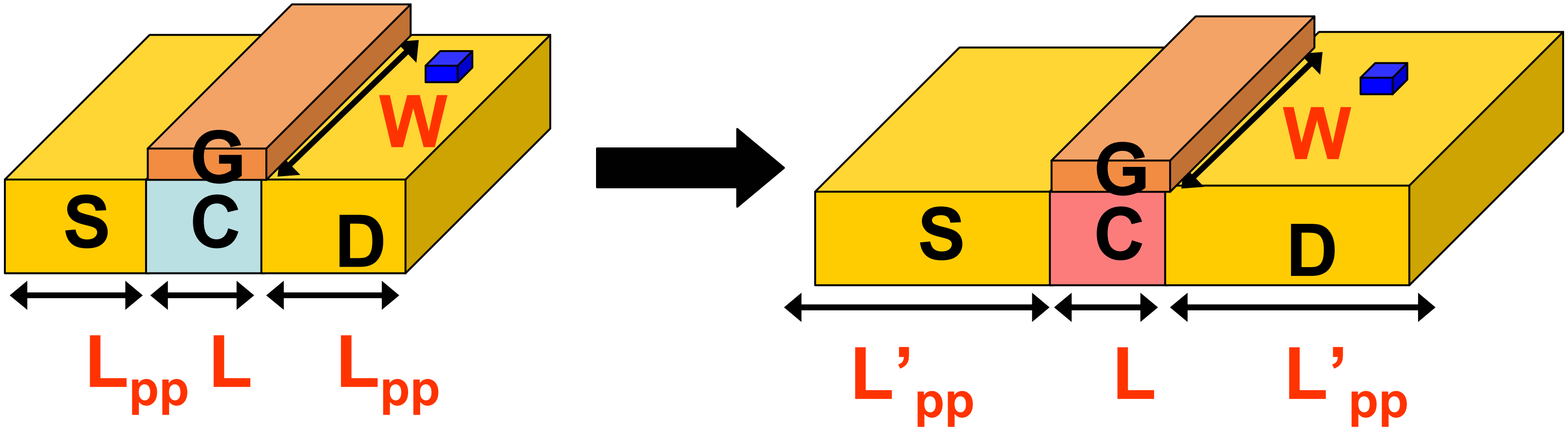
- You have a product without S/D SiGe (1 GHz)
- Soon will use S/D SiGe. (magically get 1.5 GHz)
- Is change required at physical synthesis stage?
  - Gate sizing algorithms
  - Repeater insertion algorithms
  - Buffer planning tools
- Yes. Must change these to exploit fully
  - Approximately 10% lesser module delay
  - Approximately 10% lesser global interconnect delay
  - Can get 1.65 GHz!

Outline

- ♦ Intro. to source/drain (S/D) SiGe technology
- ♦ **SiGe Active Area (AA) aware Delay Model**
- ♦ AA aware Optimal Repeater Insertion (ORI)
- ♦ Concurrent AA and Gate Sizing
- ♦ Conclusions

Cell Delay Model Derivation

- Analyze cell layout to obtain RC switch model
  - Consider AA aware PMOS resistance values
  - Consider increased self-loading capacitances



- Compute new fall and rise time
- Average fall and rise delay to get cell delay

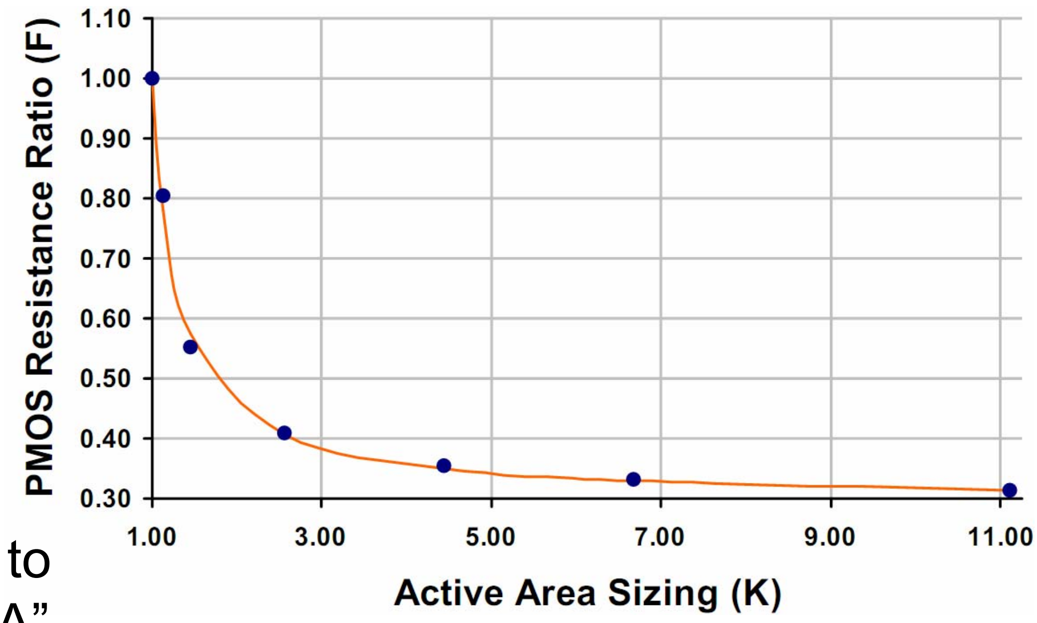
PMOS Resistance Decrease

- Stress ⇔ Mobility (μ) ⇔ R_{ON} (ON resistance)

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

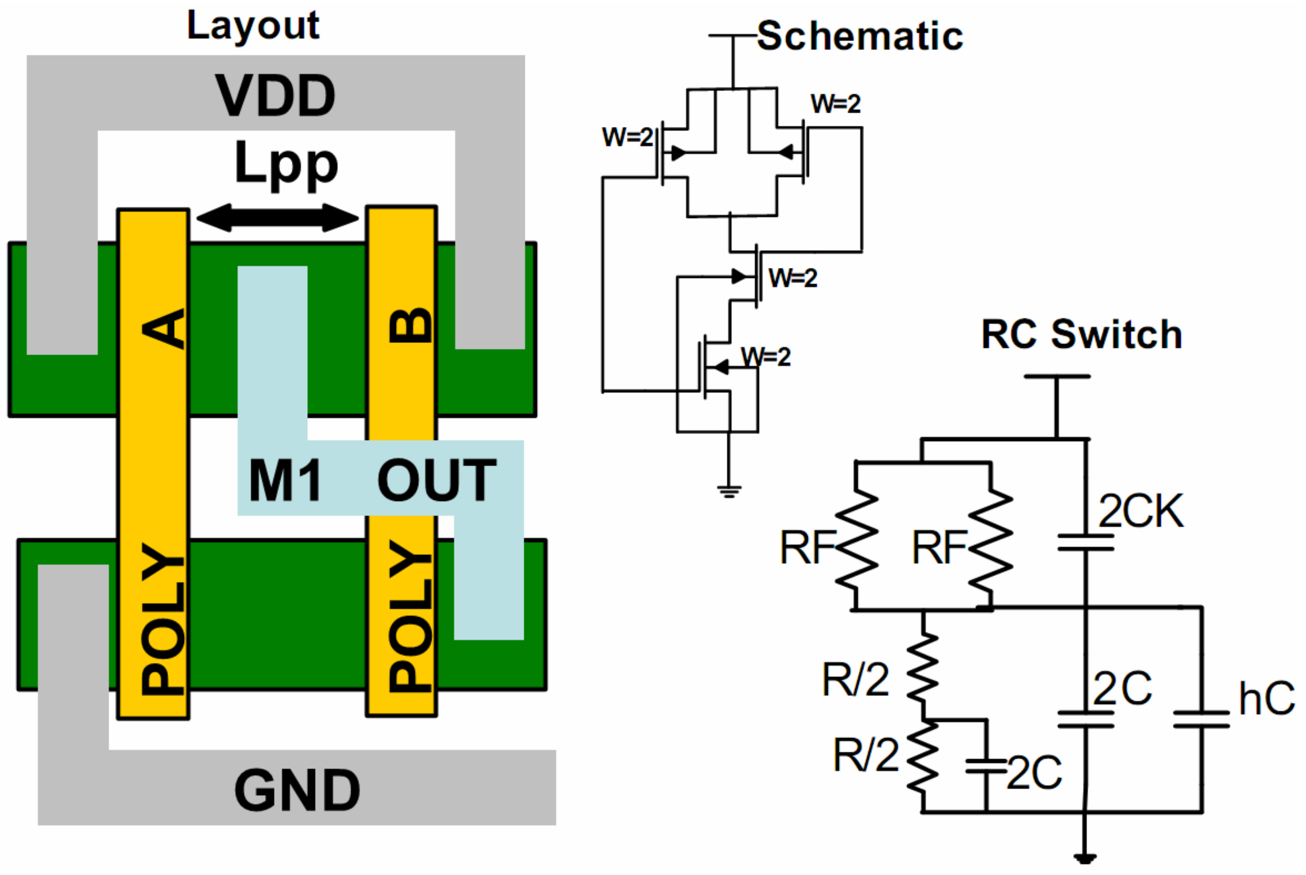
By curve fitting

$$F = \frac{K}{A \times K - A + 1}$$



Relates PMOS R_{ON} decrease to SiGe AA increase. Value of “A” depends on Ge conc and recess depth. For our settings, A = 3.4.

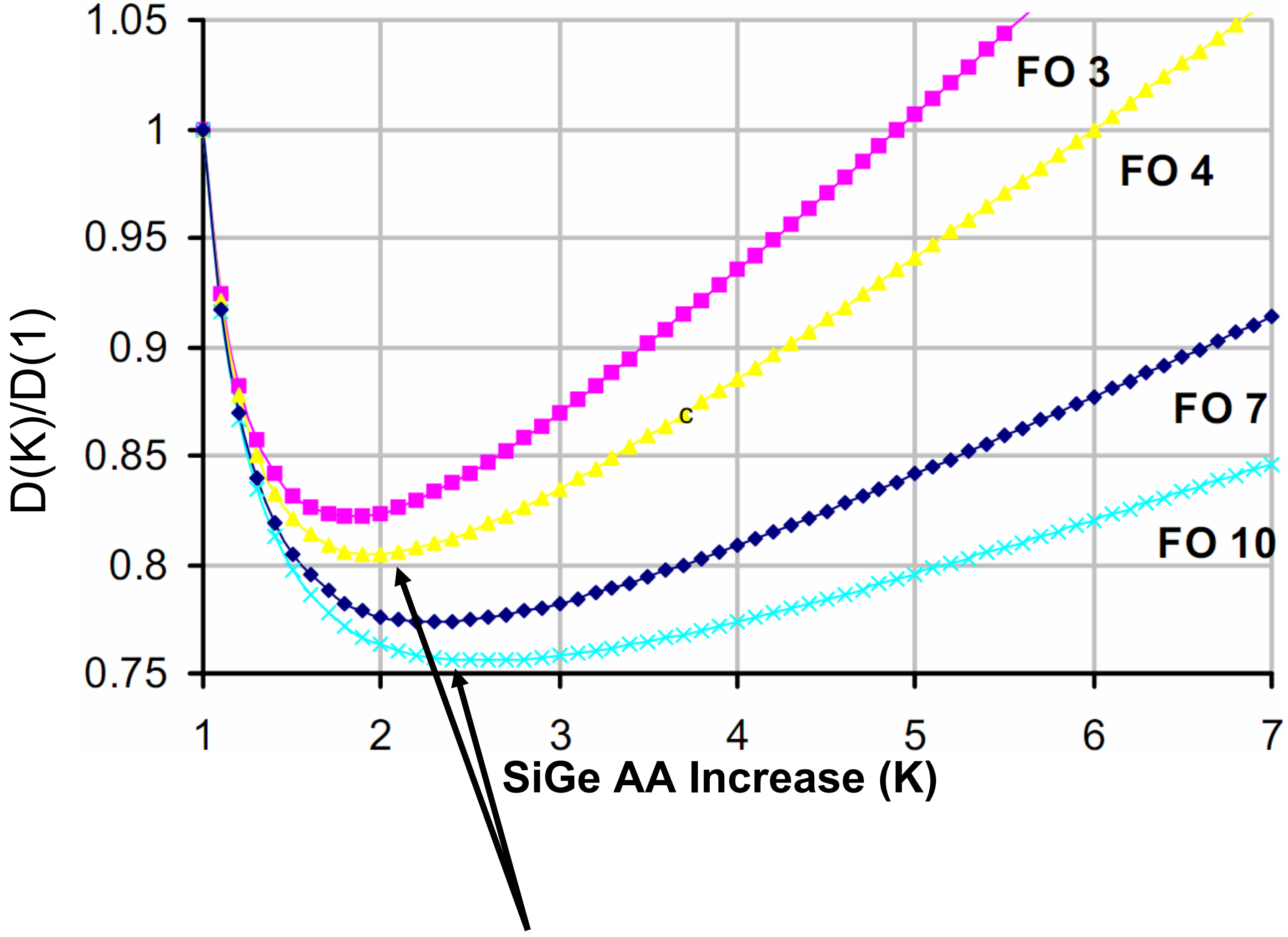
Example [NAND Gate]



D(K) = RC ((F+1)(K+0.5*h+1) + 0.5)

Characteristic Delay Equation

NAND Cell Delay (different fan-outs)



For FO-4, 2X Lpp => 20% delay decrease

For FO-10, 2.5X Lpp => 25% delay decrease

Delay Decrease for other Gates

Gate	Characteristic Delay Equation	ΔD @ FO4
INV	RC (F+1)(K+0.5*h+0.5)	-17.9%
2 input NAND	RC ((F+1)(K+0.5*h+1) + 0.5)	-21.9%
3 input NAND	RC ((F+1)(2K+0.5*h+1.5) + 1.5)	-16.2%
2 input NOR	RC ((F+1)(2K+0.5*h+0.5) + FK)	-16.4%
3 input NOR	RC ((F+1)(3K+0.5*h+1) + 3FK)	-16.1%

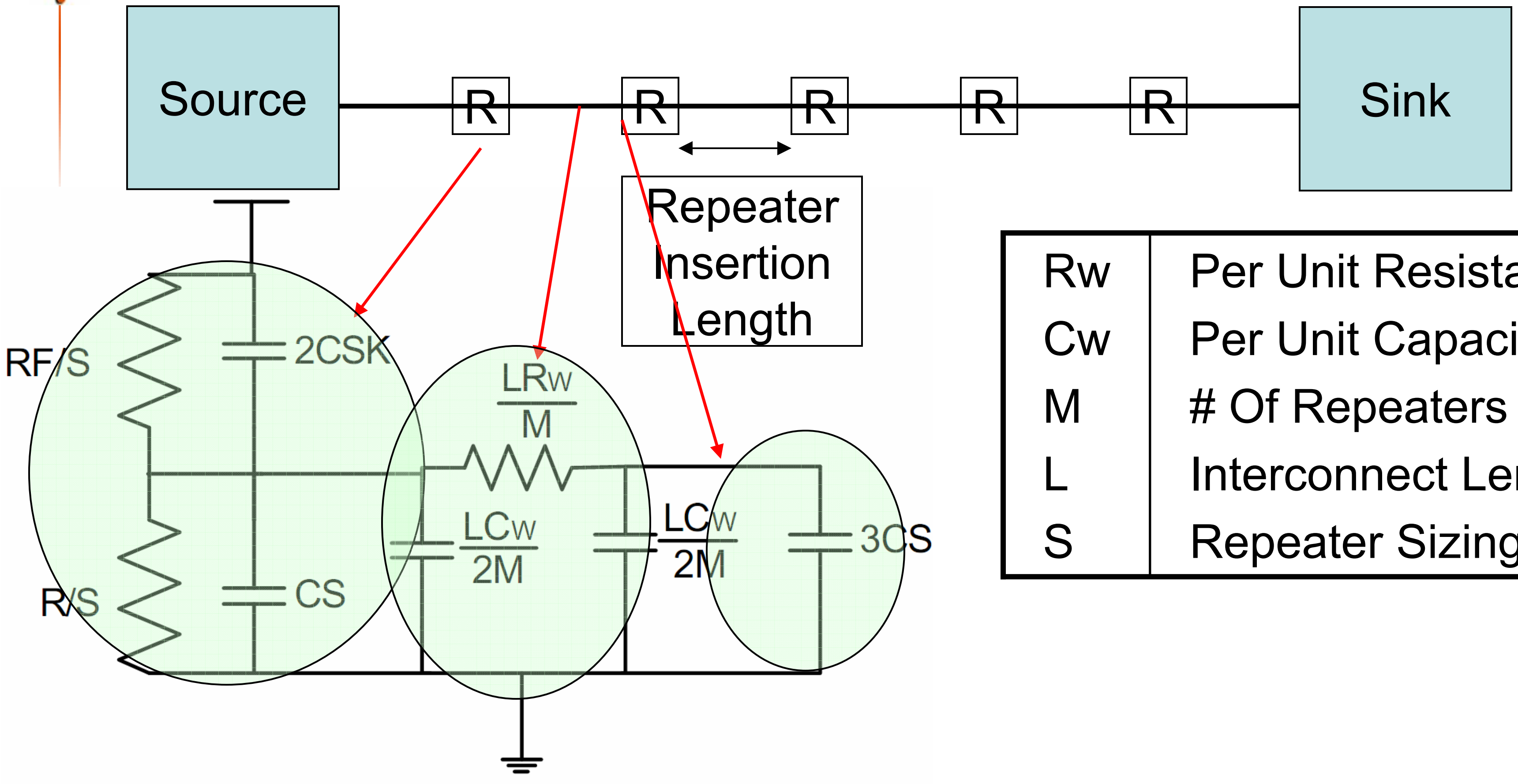
Outline

- Intro. to source/drain (S/D) SiGe technology
- Cell Delay Model Derivation
- **AA aware Optimal Repeater Insertion (ORI)**
- Concurrent AA and Gate Sizing
- Conclusions

Optimal Repeater Insertion (basics)

- Target: Minimize delay through interconnects.
- Divide a long interconnect into several parts. A repeater is inserted to drive each of these.
- AA sizing aware ORI: Apart from gate size, number of repeaters, also determine optimal AA size of the repeater cell.

AA Sizing Aware Repeater Insertion

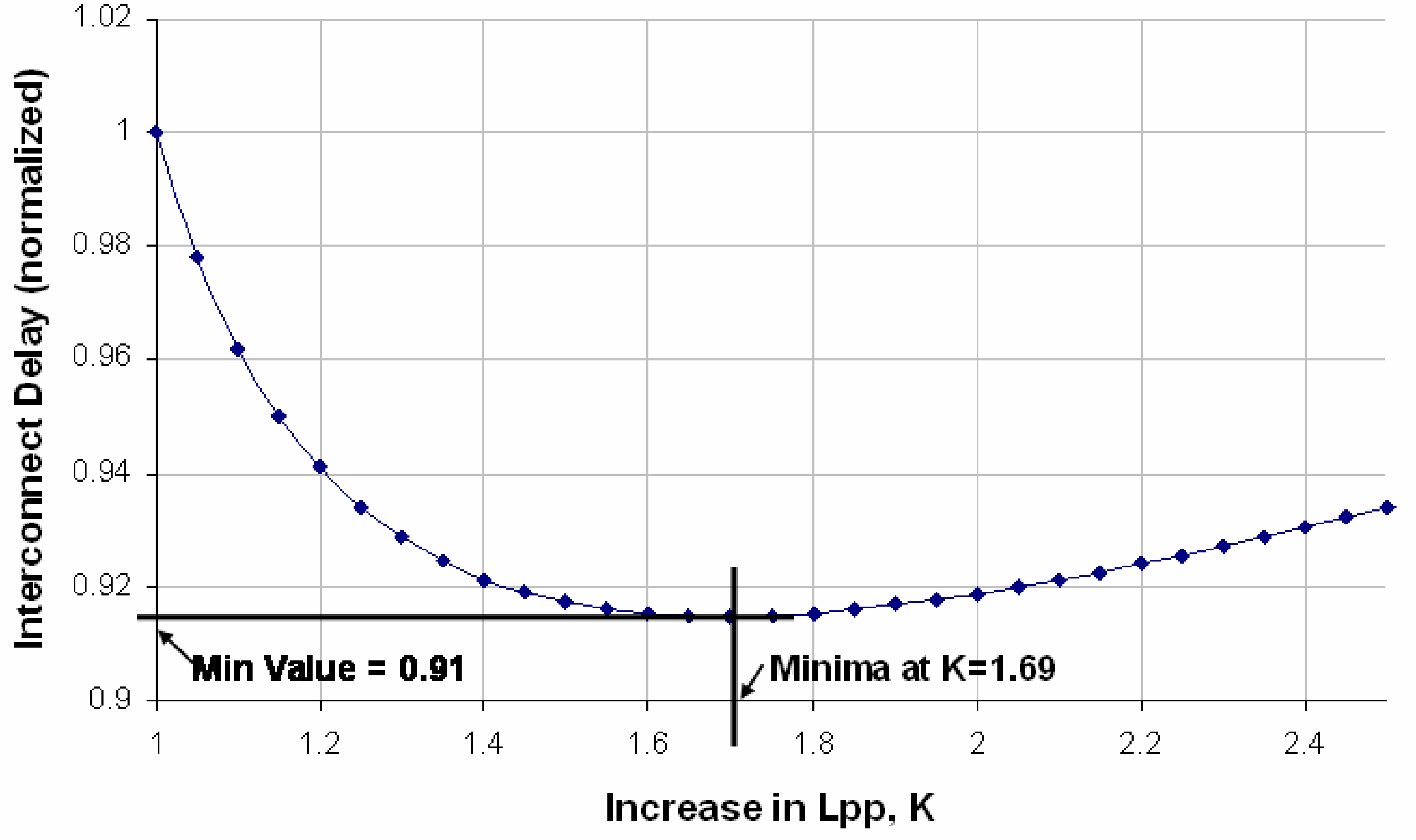


Rw	Per Unit Resistance
Cw	Per Unit Capacitance
M	# Of Repeaters
L	Interconnect Length
S	Repeater Sizing

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

Minimum Interconnect Delay

- Minimizing the delay equation analytically…



Results [ORI for Performance]

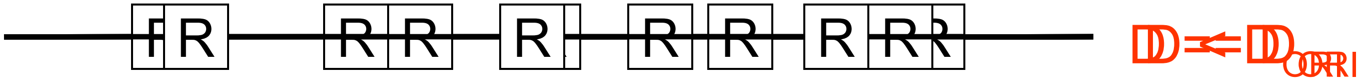
Metric	Traditional ORI	AA Sizing + ORI
Delay	D	0.91*D (=> -9%)
# Repeaters	M	1.04*M (=> +4%)
Gate Size	S	0.87*S (=> -13%)
AA Size	1	1.7 (=> +70%)
Total Power	P_{total}	1.1*P_{total} (=>+10%)

Thus, 9% better delay than the “optimal” repeater insertion solution without SiGe AA size change.

What if the aim is not to maximize performance? i.e. iso-delay case (compared to traditional ORI)

Results [Reducing # of interconnects]

- Reduce no. of repeaters until AA sizing aware sub-optimal repeater insertion delay is same as traditional post ORI delay.



- 45% reduction in number of repeaters!
  - Very interesting for layout level timing closure stability

Outline

- Intro. to source/drain (S/D) SiGe technology
- Cell Delay Model Derivation
- AA aware Optimal Repeater Insertion (ORI)
- **Concurrent AA and Gate Sizing**
- Conclusions

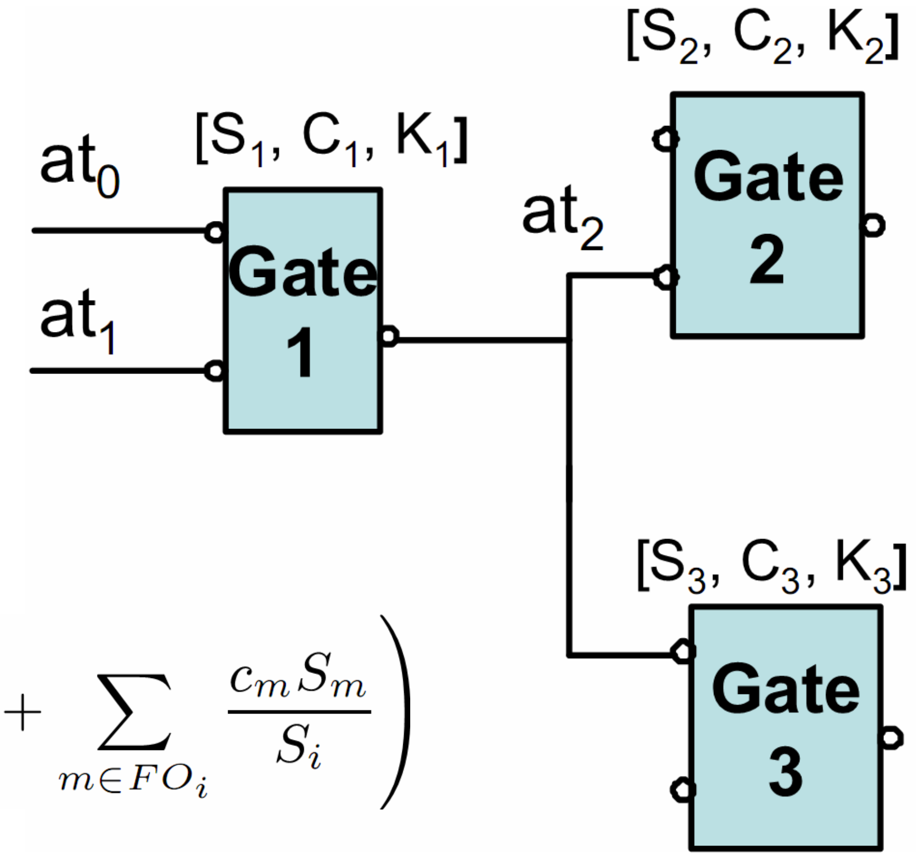
CGAS: Concurrent Gate and AA Sizing

- Target: Minimize a convex objective
  - Delay through the module, or
  - Power under delay budget, or other.

- Determine gate size of each cell and its active area sizing.

CGAS: Formulation

- Let tuple {S, C, K} represent gate size, input pin capacitance, and active area sizing for a gate.



- Delay of gate i:

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

CGAS: Formulation

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

- Are all constraints convex?
  - All e
  - Firs paper) **as long as** fitting parameter A is >= 1

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

Results [CGAS on IWLS benchmarks]

Bench	Num Gates	Delay GS	Delay CGAS	%Perf Imprv	% Δ Cap.
C6288	3316	1320	1175	11.0	3.2
C880	502	340	309	9.0	0.4
frg1	149	178	159	10.7	0.2
k2	1163	323	295	8.7	0.5
C7552	2581	734	687	6.4	0.4
large	481	262	236	9.8	0.3
vda	628	222	199	10.2	0.6
des	3759	270	233	13.9	1.3
C5315	2007	449	400	11.0	1.5
Average:				10.1	0.9

Note: All delay values are multiples of RC

More than 10% reduction in delay over Gate Sizing (GS)

Less than 1% capacitance increase on average

Outline

- Intro. to source/drain (S/D) SiGe technology
- Cell Delay Model Derivation
- AA aware Optimal Repeater Insertion (ORI)
- Concurrent AA and Gate Sizing
- **Conclusions**

Conclusions

- When moving to S/D SiGe, physical synthesis must be revisited to extract maximum benefit.

- Proposed SiGe AA sizing aware RC model with cap increase and PMOS R_{ON} decrease.

- For long global interconnects, with SiGe AA sizing of repeaters, delay reduced further by 9%.

Conclusions (contd...)

- Or reduce repeater count by 45%. Break cycle: timing analysis ⇔ buffering ⇔layout legalization

- Concurrent gate and SiGe AA sizing (CGAS) proposed and proven as a convex problem.

- For module delay reduction, CGAS reduces delay by 10% over non-AA aware sizing.

References

- “Scalability of the SiGe S/D technology for the 45-nm technology node and beyond,” in *IEEE Transactions on Electron Devices*, July 2006.
- 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
- S. Boyd et al. , *Convex Optimization*. Cambridge Univ. Press, March 2004.

Questions?

Notation

- In the rest of the work:

Increasing the active area of a gate by K times reduces its PMOS's resistance by F times. These are related by the formula

$$F = \frac{K}{A \times K - A + 1}$$

A = 1 : PMOS resistance independent of K

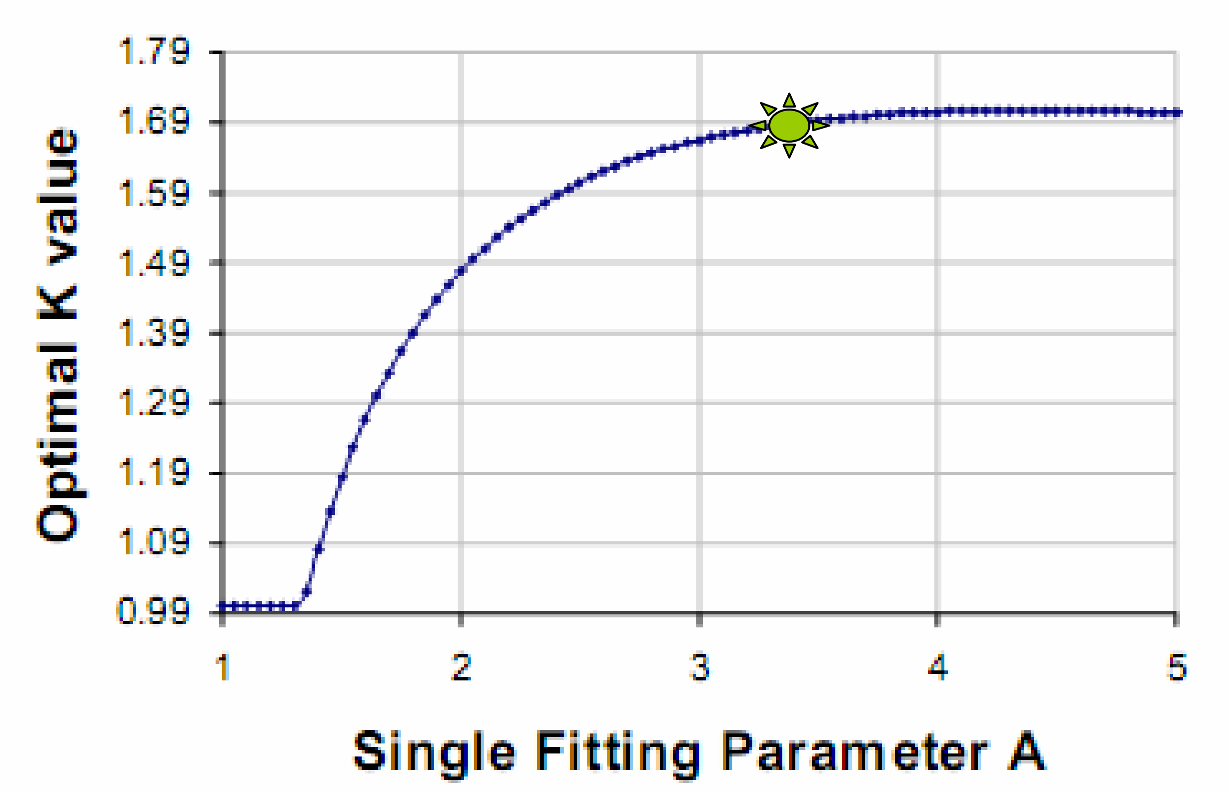
A < 1 : PMOS resistance increases with higher K

A > 1 : PMOS resistance reduces with higher K

In our curve fit, A = 3.4

Optimal K Value

- Depends on fabrication technology
  - i.e. on single fitting parameter A



Flow Used for CGAS

Benchmark

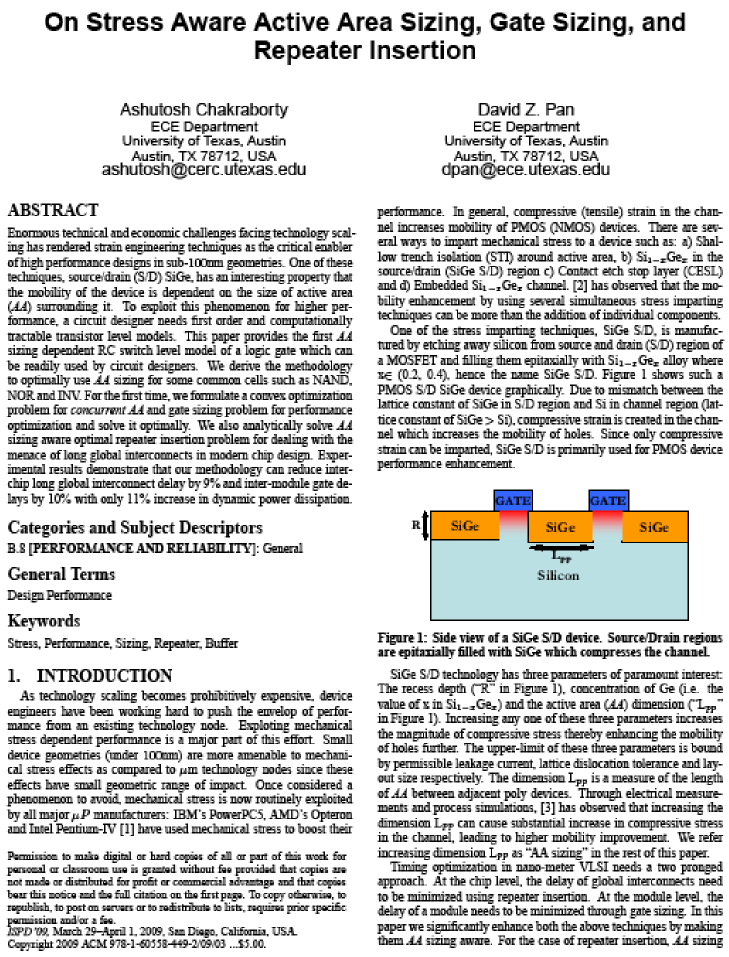
Optmz and Tech map in SIS

2-nand 2-nor inv decomposition used

C++ tool writes out the constraints

MOSEK + AMPL solvers report results

Link to the paper



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.

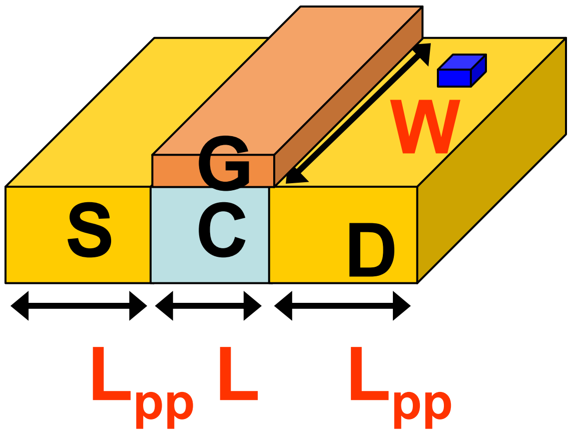
One of the stress imparting techniques, SiGe S/D, is manufactured by etching away silicon from source and drain (S/D) region of a MOSFET and filling them epitaxially with $Si_{1-x}Ge_x$ alloy where x$\in$ (0.2, 0.4), hence the name SiGe S/D. Figure 1 shows such a PMOS S/D SiGe device graphically. Due 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.

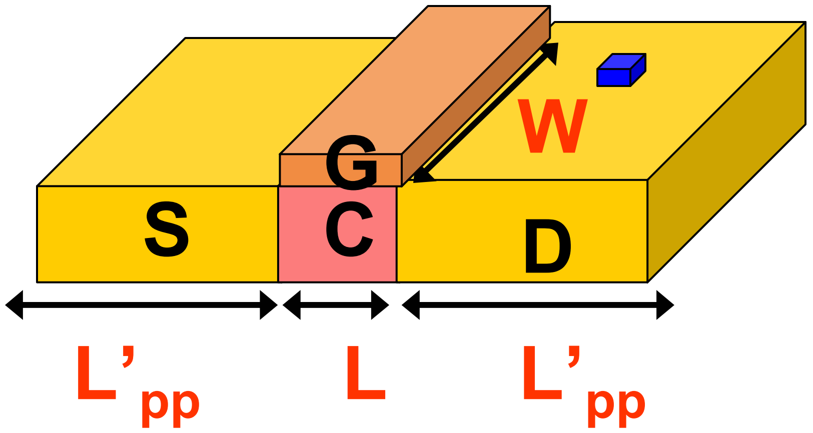
Figure 1: Side view of a SiGe S/D device. Source/Drain regions are epitaxially filled with SiGe which compresses the channel.

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.

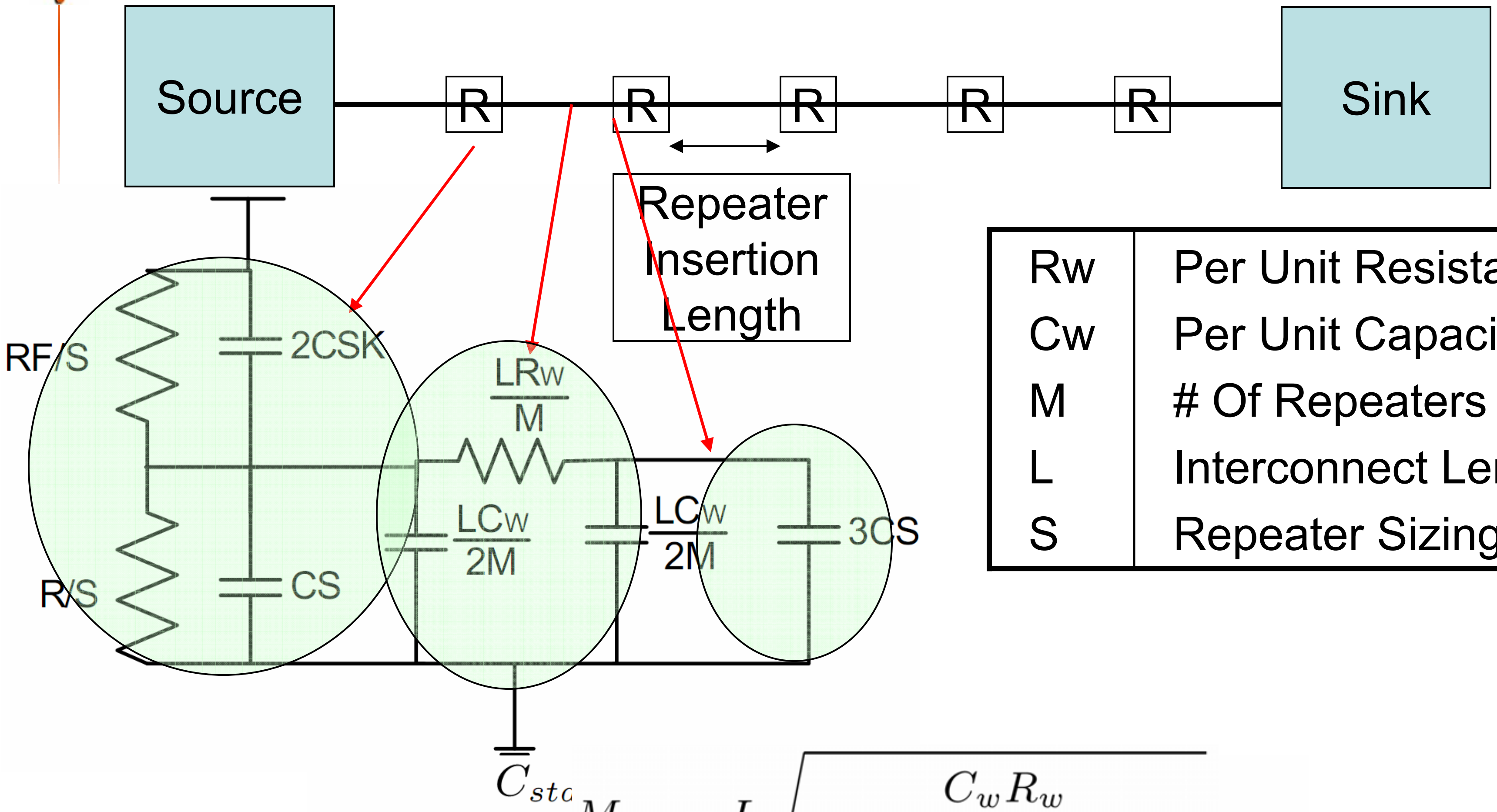
Timing optimization in nano-meter VLSI needs a two pronged approach. At the chip level, the delay of global interconnects need to be minimized using repeater insertion. At the module level, the delay of a module needs to be minimized through gate sizing. In this paper we significantly enhance both the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


G
S
C
D
W
L_{pp}
L
L_{pp}


G
S
C
D
W
L'_{pp}
L
L'_{pp}

AA Sizing Aware Repeater Insertion



Repeater Insertion Length

Rw	Per Unit Resistance
Cw	Per Unit Capacitance
M	# Of Repeaters
L	Interconnect Length
S	Repeater Sizing

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

$$D_{tot} = \sqrt{2L^2 RC R_w C_w} \times \sqrt{1+F} \times \left(\sqrt{3} + \sqrt{2+K}\right)$$

$$D_{tot} = M \times D_{stage}$$