Shaping Gate Channels for Improved Devices

Puneet Gupta[1] (puneet@ee.ucla.edu), Andrew B. Kahng[2] (abk@cs.ucsd.edu),
Youngmin Kim[4] (kimyz@eecs.umich.edu), Saumil Shah[3] (saumil@blaze-dfm.com),
Dennis Sylvester[4] (dennis@eecs.umich.edu)

[1]EE Dept., University of California, Los Angeles
[2]ECE and CSE Dept., University of California, San Diego
[3]Blaze DFM Inc. Sunnyvale, CA
[4]EECS Dept., University of Michigan, Ann Arbor

Abstract

With the increased need for low power applications, designers are being forced to employ circuit optimization methods that make tradeoffs between performance and power. In this paper, we propose a novel transistor-level optimization method. Instead of drawing the transistor channel as a perfect rectangle, this method involves reshaping the channel to create an optimized device that is superior in both delay and leakage to the original device. The method exploits the unequal drive and leakage current distributions across the transistor channel to find an optimal non-rectangular shape for the channel. In this work we apply this technique to circuit-level leakage reduction. By replacing every transistor in a circuit with its optimally shaped counterpart, we achieve 5% savings in leakage on average for a set of benchmark circuits, with no delay penalty. This improvement is achieved without any additional circuit optimization iterations, and is well suited to fit into existing design flows.

1. Introduction

While continued technology scaling plays an important role in improved circuit performance, the associated increase in power density has led to the rise of power-limited yield loss. The proportion of total power contributed by subthreshold leakage has been shown to increase with every technology node [1]. Over the years, several circuit optimization techniques to meet timing and power constraints have been proposed at the standard cell level [2,3,4] and at the transistor level [5].

Cell and transistor-level optimization techniques involve assigning different characteristics such as threshold voltage and gate length to different cells or devices based on slack availability. Circuit optimization is typically a two-phase process. The first step is that of library optimization, in which multiple variants of a standard cell are created with different power-performance tradeoffs. This can be done either by changing the characteristics of the entire cell simultaneously or changing those of every transistor in the cell. The second step is the design optimization step, wherein an optimizer assigns these variants to different instances in a circuit in order to meet simultaneous timing and power constraints.

A characteristic of all these methods is that locally there is a trade-off between leakage and delay, and the burden is on the circuit optimizer to intelligently assign these different variants to all cells in a circuit. Optimization of large circuits in feasible time is a difficult problem and has been the focus of much research ([6,7] are just a few examples). For a standard cell based methodology, the number of variants required for every cell could be very large since different variants are suited to different slack characteristics. An enormous library size places further emphasis on an efficient optimization algorithm. We also note that, for high-performance applications, designs typically have tight delay constraints. Under tight constraints, there is often not enough slack to be used for appreciable power reduction. As a consequence, optimization results are significantly affected by optimizer quality, tightness of constraint and circuit complexity.

In this work, we propose a novel channel optimization technique in which timing and leakage optimization is performed purely at the transistor level. From an existing transistor, we create a new *dominant* device that possesses better delay, leakage and input capacitance characteristics than the original transistor. Depending on the application, we can focus the optimization towards either timing or leakage, but in every case the optimized transistor is dominant. Once we create an optimized transistor for every length and width combination in the library, we replace each transistor in every standard cell with its optimized counterpart. This creates a new set of standard cells that are superior in all characteristics to the existing cells. Since there is no tradeoff between delay and leakage involved at the standard-cell level, the circuit optimization step simply involves substituting every standard cell in the circuit

Design for Manufacturability through Design-Process Integration II, edited by Vivek K. Singh,
Michael L. Rieger, Proc. of SPIE Vol. 6925, 69250I, (2008) · 0277-786X/08/$18 · doi: 10.1117/12.772889

with the dominant cell, without any slack or leakage analysis. The dominance property ensures that the circuit generated by substitution has better timing and leakage than the original circuit. There is no dependence on complex large-scale optimizers in this method. The efficiency of this method is, therefore, independent of circuit complexity.

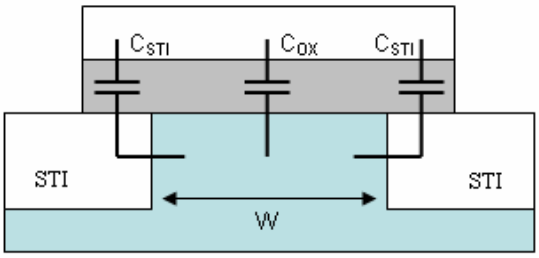


Figure 1. Cross section of a MOS device with Shallow Trench Isolation Technology

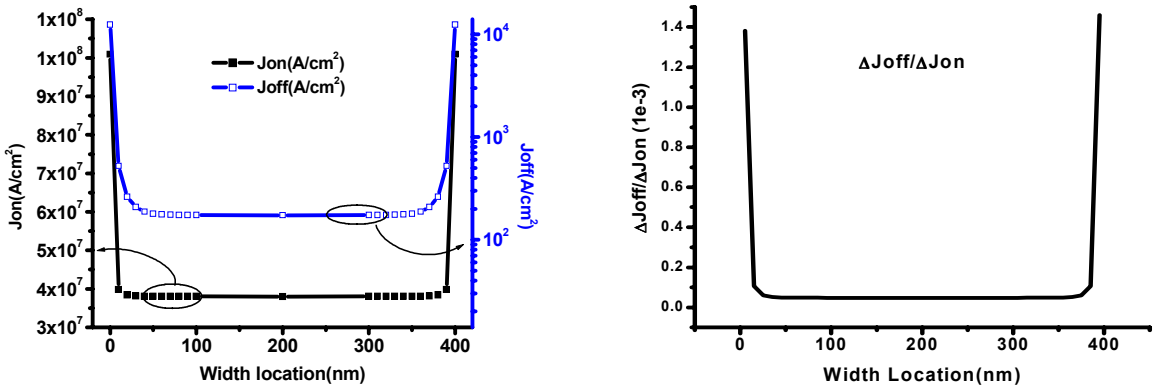
The motivation behind this method is provided by the "edge-effect" in the transistor channel characteristics. References [8-11] discuss this effect, which manifests as an unequal distribution of drive and leakage current densities across the width of the channel. In the Berkeley Short-Channel IGFET Model (BSIM) [12], this effect is modeled using the narrow-width component of the threshold voltage model [10]. The observation that the threshold voltage of modern MOSFETs is much lower near the edges than the center of the channel is central to our method. Of prime interest is the fact that the *relative* values of the drive (Ion) and leakage (Ioff) currents change across the channel. By selectively changing the properties of different portions of the channel, we can improve either total Ioff or Ion, without negatively affecting the other.

Ion and Ioff are dependent on a number of factors, such as dopant concentration, oxide thickness and gate-length. It is difficult to assign different implants and oxides to different parts of the same device, hence we focus on adjusting the gate length. We selectively increase the length at some points in the channel, while decreasing the length at other locations which effectively changes the shape of the device from a perfect rectangle to a non-rectangular shape. We refer to this process as intra-gate biasing (IGB). This non-rectangular transistor can be created such that it is superior in all characteristics to the original device.

In this paper we describe this channel-shaping methodology and its application to circuit optimization. The rest of the paper is organized as follows. The edge-effect phenomenon is explained in detail in Section 2. Section 3 briefly describes how to fit a mathematical model that can capture this effect in the absence of a 3D TCAD setup. Section 4 discusses the feasibility of this method from a fabrication perspective. Section 5 explains the algorithm that we use to generate an optimal shape for a transistor of a given width and length. Section 6 details our experimental setup. The results are described in Section 7 and Section 8 concludes the paper.

2. Edge effect: physical explanation

This section reviews the effects that lead to an unequal distribution of drive and leakage current densities depending on distance from the device edge. The model we use is discussed in greater detail in [8-11]. For completeness, we review some of that discussion here. Figure 1 shows the cross section of a MOS device with shallow trench isolation (STI) technology. To compensate for the line-end shortening effect due to defocus, the gate polysilicon is required to extend over the diffusion area to some extent. The extension over the STI region is termed as the line-end extension. The figure shows parasitic fringing capacitances (C_{STI}) between gate, STI sidewall, and active area, where there is a polysilicon gate extension over the isolation. The fringing capacitance leads to higher capacitive coupling and hence higher surface potential near the edges. This in turn leads to reduced threshold voltage (Vth) near the edges. This effect is exacerbated by dopant scattering due to STI edges [8], and the well proximity effect (WPE) [9]. These effects are pronounced near the device edges and roll off sharply as we move towards the center of the device. This lowering of Vth near the edges causes the on and off current densities to be much higher near the edges than at the center. Although this effect appears in both Ioff and Ion, the leakage current is an exponential function of threshold voltage, whereas the drive current has a sub-quadratic (typically of degree 1.3-1.5 [13]) dependence.



(a) (b)

Figure 2. (a) Drive and leakage current densities (Jon and Joff) and (b) ΔJoff/ΔJon for a small change in gate-length as a function of location along channel.

Figure 2 (a) shows the on and off current densities as a function of position, based on 3D TCAD simulation. The TCAD setup is calibrated to closely match an industrial 90nm SPICE setup. The data shown here is for a 400nm wide NMOS device. PMOS devices tend to show similar characteristics. We notice that Jon increases only by 3x going from the center to the edges, whereas the off current density increases by 30x. Clearly, the Joff/Jon ratio is much higher near the edges than near the center. We also observe the change in Joff and Jon as we change the length of the device. Figure 2 (b) shows a plot of ΔJoff/ΔJon as a function of location along the channel, for a 4nm change in length across the entire gate. This graph suggests that increasing the gate-length closer to the edges, provides large leakage savings for a given delay overhead. At the same time, decreasing gate length near the center can provide delay improvement for a relatively small leakage overhead. This analysis provides the basis behind our intra-gate biasing methodology.

3. Mathematical model in absence of a TCAD setup

To be able to use this shaping methodology in an industrial optimization flow, we require the ability to compute Ion and Ioff for the optimized gates. Current circuit analysis tools can only handle perfect rectangular gates and cannot analyze non-rectilinear geometries. In [11], the authors describe a method to model non-rectangular gates.

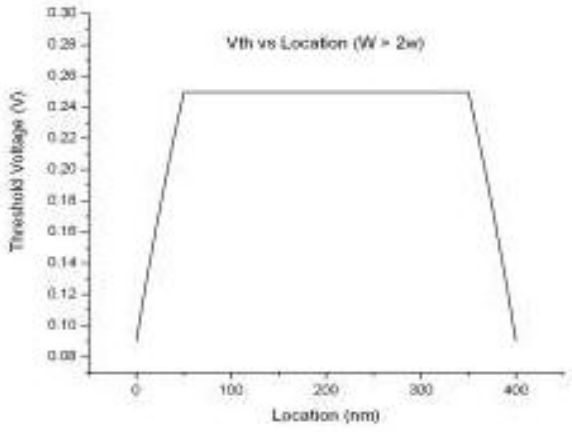


Figure 3. Vth as a function of location

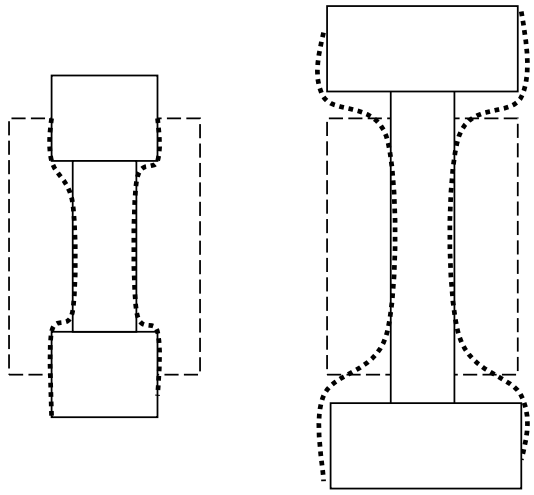
The Jon and Joff data obtained from TCAD simulation as shown in Figure 2 can also be generated using a model as described in [11]. The method develops a mathematical formula for Vth as a function of position along the channel width. Figure 3 shows plots of Vth vs. location for a particular device. Using this Vth model, we can calculate the drive and leakage current densities along the entire channel width. The development of this model is purely on the basis of SPICE or measurement-based current vs. device width curves, which are readily available for any

technology. This model is useful when a realistic TCAD setup is not available for the technology. This model can be used in our flow to perform optimization as well as to map the optimized gates to equivalent rectangular gates.

4. Manufacturing feasibility and impact

The unusual shape of the resulting optimized channel raises some concerns about printability of such gate shapes. Arbitrary shapes in general can be difficult to print. For instance, non-Manhattan edges may need to be "stairstepped" to be put on a mask. Therefore, we avoid non-Manhattan shapes in the channel. Given this, there are two ways to print non-rectangular channel shapes:

1. *Active Shape Perturbation*: As the name suggests, in this case the channel layout is explicitly drawn as intended. It is assumed that the downstream RET processes will be able to reproduce the shape faithfully on the wafer. With advancements in OPC, this is not an unreasonable expectation. The channel shape here can be well-controlled but it may require minor design-rule waivers and a smaller layout grid resolution.
2. *Passive Shape Perturbation*: In this case, the channel is not explicitly drawn in a non-rectilinear fashion. The shape change is induced by changing the layout of field polysilicon geometries adjoining the gate. The corner rounding as a result of lithographic imaging effectively produces a perturbed channel shape. This may be the only feasible option in certain cases where a bent gate shape is prohibited by a technology. The disadvantage of such a perturbation scheme is inadequate control of the channel shape.



(a) **(b)**

Figure 4. (a) Active shape perturbation example with sample lithographic printed image contour shown in dotted line. (b) An example of passive shape perturbation. Note that in (b) the drawn gate channel is a perfect rectangle.

Figure 4 shows examples of active and passive shape perturbation. In either case, the number of corners is likely to increase in the layout (more so in active shape perturbation). Corners and small jogs can be difficult to OPC. In the channel shape solutions that we propose, we restrict ourselves to introducing at most four extra jogs in the channel shape (i.e., the channel looks like a "dumbbell"). Moreover, in this work we consider active shape perturbation and assume that the channel is printed as drawn. In a future work, the impact of wafer contours (as a result of imperfect printing) and passive perturbation can also be taken into account.

5. Optimal shaping solution

To perform channel shaping targeting improved total device Ioff, Ion and capacitance, we again refer to the current density and sensitivity curves in Figure 2. Figure 2 (b) shows the sensitivity of leakage and drive current to small changes in gate-length at different points along the channel. The plot shows that the sensitivity near the edges is up to 10X greater than the sensitivity near the center of the device. It is clear that increasing the gate length near

the edges will lead to a large decrease in leakage along with some decrease in drive current. This reduction in drive current can be compensated for by reducing gate length near the center of the device. Although this compensation is achieved at the cost of some increased leakage, we can expect that this small increase in leakage will not offset the gains provided by the large edge-gate length. Here, we assume that the technology allows slightly sub-minimal gate-lengths, which is not unreasonable [5].

The channel shaping process, including the form of the final optimized shape, is shown in Figure 5. The original gate length is *l* and width is *w*. The optimized gate has center length $l_1 < l$, edge length $l_2 > l$, and a width of the l_2 region of *w'*. The total width of the gate remains *w* to maintain layout compatibility with the original cell. The equivalent rectangular gate shown at the right in the figure is a "leakage-equivalent" geometry with *leq>l*.

In the optimization process, any shape that has either lower Ion or higher capacitance than nominal is an infeasible solution. We use a first-order approximation for the input capacitance, assuming it to be directly proportional to the total area of the gate. Of all feasible solutions, the one with the lowest total leakage current is taken as the optimal solution.

The optimal shape is found by the process of exhaustive enumeration with pruning of provably sub-optimal or infeasible solutions. In our library, we find that the maximum width for a single device is 1um. Devices larger than that are split into multiple fingers, where we can apply this technique independently to different fingers.

We constrain l_1 and l_2 to be on a 2nm grid. For *w',* a very fine grid raises manufacturability concerns, whereas a coarse grid reduces available optimization space. In this setup, we choose a 10nm grid for *w'*. l_1 is always less than *l*, l_2 is always greater than *l*, and *w'* has to be less than half the device width. For our technology with a drawn nominal gate-length of 100nm, the ranges for l_1, l_2 and *w'* are 90-98nm, 102-110nm and 10-490nm, respectively. This gives us a total of 1225 combinations per device. Since every combination is simply a sum of current densities, runtime is acceptable even if all combinations are investigated. However, we observe that some combinations are provably sub-optimal or infeasible and can be omitted from the enumeration process. For example, consider a fixed value of l_1 and l_2. In this case if a particular value of w' has lower Ion or higher area (capacitance) than the nominal case, it is useless to increase w' further since this only serves to reduce Ion and increase area further. This would amount to pushing the solution further into the infeasible region. As another example, in the case of a fixed value of l_1 and *w'*, if a particular value of l_2 has higher Ion and lower area than nominal, we do not need to reduce l_2 further since this will increase Ion and Ioff, and reduce capacitance further. In other words, all solutions with l_2 below this value will be feasible, but suboptimal. These and other observations allow us to prune a considerable number of combinations from the total evaluation space.

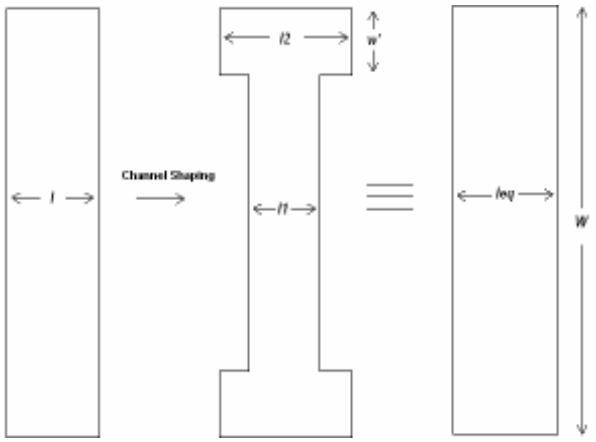


Figure 5. Channel shaping and equivalent rectangular gate.

Once an optimal shape for the gate has been identified, it is mapped onto an equivalent rectangular gate to allow SPICE-based characterization of the modified standard cells. It is important to note that the lengths of the equivalent gates are different for Ioff and Ion. Specifically, the equivalent *l* for Ion is lower or equal to the length of the original gate, whereas the equivalent Ioff length is larger. New standard cells are created using these equivalent gate lengths. Separate sets of cells are created for delay and leakage equivalence.

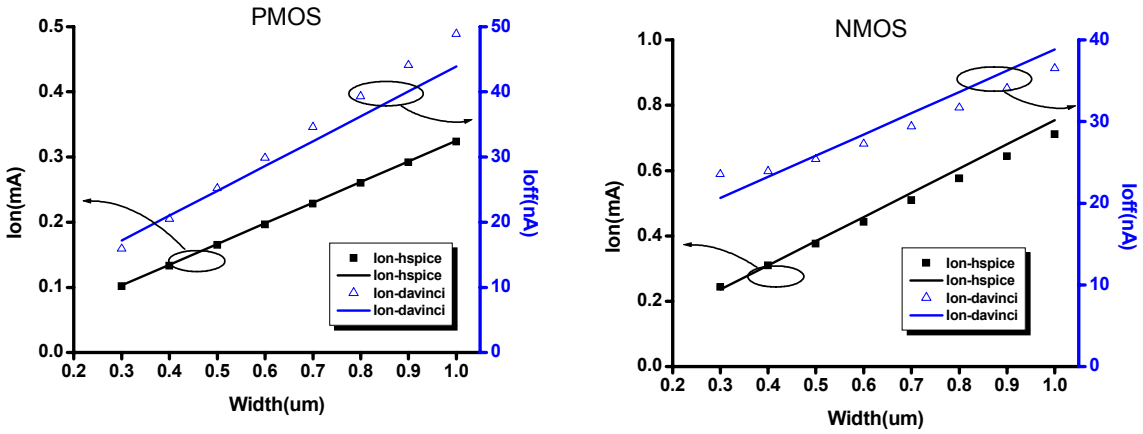


Figure 6. Matching accuracy between SPICE and TCAD

6. Experimental Setup

For our experimental setup, we use an industrial 90nm technology with BSIM4.3 SPICE[12] models and a closely matched TCAD setup. The matching accuracy is shown in Figure 6. The TCAD setup is shown to track the SPICE data accurately over the entire range of widths. The standard cell library under consideration consists of a subset of the total available cells. We first extract a list of all width and length combinations from the SPICE netlists. The channel shaping step is repeated for every such combination in the library, and equivalent lengths are generated. Nominally all devices in a library have the same length, and we need to perform shaping only for all width values.

Variants of each of the standard cells are created by substituting the lengths of the original devices in each cell with the equivalent lengths. We note that this step is just for the characterization process, as cell characterization tools only accept rectangular gates.

Two distinct variants are created for each cell, for both Ion and Ioff equivalence and characterized using a commercial software. In addition to delay and leakage, we also compute the equivalent input capacitance of every cell. In the library, we simply scale the input capacitance by the ratio of the areas of the optimized and original devices. Finally, we merge the delay, leakage and input capacitance values to create an optimized library correctly reflecting the newly shaped devices.

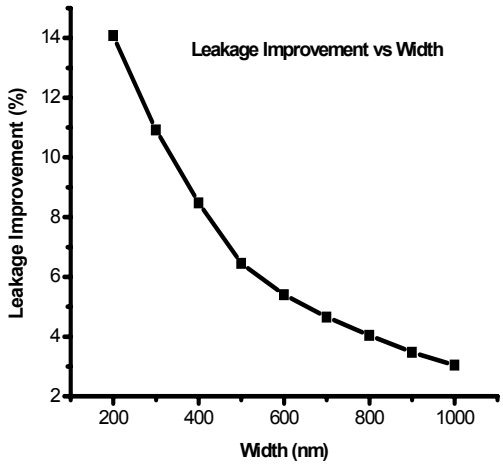


Figure 7. Leakage improvement by channel shaping as a function of NMOS device width

7. Results

This section discusses in detail the leakage improvements obtained at the transistor and circuit levels by our channel shaping method.

Figure 7 shows the leakage of IGB optimized gates of different widths with respect to a rectangular NMOS device with the same delay. PMOS characteristics are similar. We see that the improvements decrease as the width increases, from approximately 14% for very narrow devices to 3% for wide devices. This arises due to the sharp roll-off of the edge effect. For large width devices, the edge regions occupy a lesser portion of the total width. Since only a small portion of the total width can be biased positively, and the remaining width has to be negatively biased by at least 2nm to conform to our manufacturing grid, we find that a larger portion of the leakage savings is lost. This phenomenon is a consequence of the 'dumbbell' shape that we restrict ourselves to in the interests of manufacturability.

Table 1 shows the results of applying IGB to various ISCAS [15] and MCNC [16] benchmarks. We find that we obtain 4.9% average leakage savings over all benchmarks. It is worthwhile to note that our method does not suffer from any design or optimization time overhead. Also, this method is orthogonal to, and therefore is *additive* with, any other optimization procedure that has already been applied to the circuit. As mentioned before, this is possible since the method is applicable irrespective of the available slack. We also note that the savings are independent of circuit topology. The leakage improvement is only dependent on the usage statistics of different standard cells in each circuit. Thus this method has very good scalability.

Table 1. Leakage Savings for Benchmark Circuits

Circuit Name	Orig. Delay (ns)	Opt. Delay (ns)	Orig. Leakage (uW)	Opt. Leakage (uW)	% Imp.
C432	1.87	1.87	9.60	9.11	5.1
C1908	2.24	2.24	11.98	11.38	5.0
C2670	1.55	1.55	18.62	17.68	5.0
C3540	2.84	2.84	4.44	4.22	4.9
C5315	1.96	1.95	31.93	30.46	4.6
C6288	5.62	5.61	39.66	38.38	3.2
C7552	3.19	3.19	36.78	35.08	4.6
i2	0.86	0.86	13.55	12.80	5.5
i3	0.45	0.45	6.07	5.74	5.4
i4	0.58	0.58	5.46	5.21	4.6
i5	0.52	0.52	9.22	8.77	4.9
i6	0.59	0.59	10.72	10.23	4.8
i7	0.72	0.72	14.22	13.50	5.1
i8	1.01	1.01	25.19	23.98	4.8
i9	1.37	1.37	16.28	15.40	5.4
i10	2.27	2.27	55.82	53.06	4.9

8. Conclusions and future work

To our knowledge, this paper presents the first ever transistor optimization technique to exploit unequal distribution of drive and leakage current across a transistor channel. Our intra-gate length biasing methodology selectively biases the channel length for portions of the device. Using this technique, we are able to modify existing devices to achieve lower leakage, delay and input capacitance. We are able to create devices with up to 14% lower leakage than the unoptimized devices, with the same delay and capacitance. Using these optimized devices, we create an optimized standard cell library with every cell consisting of the optimally shaped devices. Using these new standard cells in circuits, we are able to achieve up to 5.5% average leakage reduction on benchmark circuits with zero delay penalty. The method achieves leakage reduction without any reliance on large-scale circuit optimizers and the improvements are independent of the available circuit slack.

It is also interesting to note that this method fits very well into an existing design flow. It can be inserted into the end of the flow, before the GDS is sent to the foundry. This method can be applied after all other circuit

optimizations have been performed and timing closure has been achieved. This is possible because our method does not change cell footprints, nor does it affect circuit timing.

As technologies scale further and device widths become smaller, we expect our method to provide considerably larger leakage improvements. Ongoing work on this method is in the following areas.

1. Verification of manufacturability and optimization results based on silicon fabrication of the shaped devices.
2. Investigation of the manufacturability of various shapes, in addition to the current 'dumbbell' shape, allowing for improved leakage optimization.
3. Timing optimization by changing the focus of the dominant device generation method from maximum leakage reduction to maximum timing reduction.
4. Generation of a large number of leakage and timing optimized devices for use with a circuit optimization algorithm for enhanced leakage reduction.
5. Investigation of benefits of a non-rectilinear diffusion shape which forms a slope under the channel region due to L or T-shape diffusion layouts [17].

10. References

[1] S. Narendra, *et. al*., "Leakage Issues in IC Design: Trends, Estimation and Avoidance", *Tutorial, ICCAD,* 2003.

[2] V. Sundarajan and K. Parhi, "Low Power Synthesis of Dual Threshold Voltage CMOS VLSI circuits," *ISLPED,* 1999, pp. 139-144.

[3] L. Wei, *et al*, "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits", *ACM/IEEE Design Automation Conference*, 1998, pp. 489-494.

[4] P. Gupta, *et al*, "Selective gate-length biasing for cost-effective runtime leakage control", *ACM/IEEE Design Automation Conference*, 2004, pp. 327-330.

[5] P. Gupta, A.B. Kahng and S. Shah, ``Standard Cell Library Optimization for Leakage Reduction", *Proc. IEEE/ACM DAC*, 2006.

[6] S. Sirichotiyakul, *et al*., "Duet: An Accurate Leakage Estimation and Optimization Tool for Dual-Vt Circuits", *IEEE Transactions on VLSI Systems*, April 2002, pp. 79-90.

[7] C. Chen, *et. al*., "Fast and Effective Gate-Sizing with Multiple-Vt Assignment using Generalized Lagrangian Relaxation", *Proc. Asia South Pacific - Design Automation Conference,* 2005, pp. 381-386.

[8] C. Pacha, *et al*., "Impact of STI-Induced Stress, Inverse Narrow Width Effect, and Statistical Vth Variations on Leakage Current in 120nm CMOS", *Solid-State Device Research conference, 2004, Proceedings of the 34th European*, Sept. 2004, pp. 397- 400.

[9] I. Polishchuk, N. Mathur, C. Sandstrom, P. Manos, O Pohland, "CMOS Vt-control improvement through implant lateral scatter elimination," *Semiconductor Manufacturing, 2005, IEEE International Symposium on*, pp. 193- 196.

[10] K.K-L. Hsueh, J. J. Sanchez, T. A. Demassa, and L. A. Akers, "Inverse-Narrow-Width Effects and Small-Geometry MOSFET Threshold Voltage Model", *IEEE Transactions on Electron Devices*, v. 35, No. 3, March 1988, pp. 325-338.

[11] P. Gupta, A.B. Kahng, Y. Kim, S. Shah, and D. Sylvester ``Modeling of non-uniform device geometries for post-lithography circuit analysis", *Proc. SPIE Microlithography*, 2006.

[12] BSIM4.3.0 User's Manual, 2003.

[13] T. Sakurai and A.R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas", *IEEE Journal of Solid-State Circuits, April 1990*, pp. 584-594.

[15] F. Brglez and H. Fujiwara, "A neutral netlist of 10 combinational benchmark circuits and a target translator in Fortran," *Proc. ISCAS*, 1989 pp. 695-698.

[16] http://www.cbl.ncsu.edu.

[17] P. Gupta, A.B. Kahng, Y. Kim, S. Shah and D. Sylvester, ``Investigation of Diffusion Rounding for Post-Lithography Analysis", *Proc. IEEE/ACM ASPDAC* , 2008.