Total Sensitivity Based DFM Optimization of Standard Library Cells

Yongchan Ban, Savithri Sundareswaran*, and David Z. Pan
Dept. of ECE, The University of Texas at Austin, Austin, TX USA
*Freescale Semiconductor, Austin, TX USA
ycban@cerc.utexas.edu, Savithri.Sundareswaran@freescale.com,
dpan@ece.utexas.edu

ABSTRACT

Standard cells are fundamental circuit building blocks designed at very early design stages. Nanometer standard cells are prone to lithography proximity and process variations. How to design robust cells under variations plays a crucial role in the overall circuit performance and yield. In this paper we propose a comprehensive sensitivity metric which seamlessly incorporates effects from device criticality, lithographic proximity, and process variations. We develop first-order models to compute these sensitivities, and perform robust layout optimization by minimizing the total delay sensitivity to reduce the delay variation on the nominal process condition and by minimizing the performance gap between the fastest and the slowest delay corners to reduce the leakage current on the process corner. The results on industrial $45nm$ node standard cells show up to 76% improvement in non-rectangular delay variation under nominal process condition, 24% reduction in the delay difference between the fastest and slowest process corners, and up to 90% reduction in leakage current at the fastest process corner.

Categories and Subject Descriptors: B.7.2 [Hardware, Integrated Circuit]: Design Aids

General Terms: Algorithms, Design, Performance.

Keywords: VLSI, Lithography, Sensitivity, Optimization, DFM.

1. INTRODUCTION

As integrated circuit (IC) process nodes continue to shrink down to $45nm$ and below, the variations of designs are increasing. Among many variation issues, lithography induced non-ideal printability and the underlying process variations are the most fundamental ones which directly impact yield and performance [1, 2]. Despite advances in resolution enhancement techniques (RET) such as optical proximity correction (OPC), phase shifting mask (PSM), off-axis illumination (OAI), lithographic variation continues to be a challenge [3,4]. The control of the gate length variation is critical to the nanometer ICs [1]. For example, 10% gate length variation may cause over 25% timing degradation, and up to 10x leakage in a $45nm$ node CMOS inverter. This requires new models and methods to mitigate the gate length variations.

Standard cells are pervasively used in digital designs as basic circuit blocks. Since a large amount of identical cells will be used repeatedly, any small changes to reduce gate length variation in standard cells can result in significant improvements at the design level. Moreover, since standard cells are the bridge between process and design, all sources of variations of the target process should be taken into account in the cell design [5, 6].

In current industrial flows, DFM optimization is usually performed either through restricted design rules or by identifying opportunities in the standard cell layout to enforce as many recommended rules as practically feasible. It shall be noted that variations still exist even with restricted design rules, e.g., single poly directions and single poly pitch [7].It is largely caused by irregular surrounding patterns, e.g., poly-contact pad to active layer, poly routing line to active, active power rail to poly line, poly end-cap, and so on [8, 9], which can cause different *non-rectangular* gates. The situation becomes even more cumbersome when process variations, such as dosage and focus are taken into consideration. The rule-based approach which is binary in nature will not be able to capture the continuous parametric yield (e.g., timing/leakage) under process variations. Moreover, current DFM optimization usually treats poly/active polygons of every device equally. It shall be noted that different transistors have inherently different delay sensitivities to the same amount of gate length variation. As a general principle, we should ensure that highly sensitive devices be given higher priority during layout optimization while less sensitive devices can allow relatively larger amount of gate length variations.

In this paper, we propose a total sensitivity driven DFM optimization for standard cell performance robustness (i.e., timing/power variations). We first systematically introduce the *total sensitivity metric* and show how to compute them. Then we incorporate the models in our standard cell layout optimization formulations. The objective of the proposed optimizations is to enhance standard cell layouts for improved parametric yield and reduced variations with minimal or no penalty on nominal delay, leakage and area. The major contributions of this paper include the following:

- We propose a comprehensive set of sensitivity metrics for robust cell layout. It consists of the transistor criticality due to device criticality, the non-rectangular

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISPD'10, March 14–17, 2010, San Francisco, California, USA.
Copyright 2010 ACM 978-1-60558-920-6/10/03 ...$10.00.

gate impact due to lithography printability, and process-variations (e.g., due to dosage and defocus).

- We develop analytical models for these sensitivities, and put them together in a seamless manner to form the total sensitivity metric by capturing the correlations of nominal lithography and process variation sensitivities.
- The total sensitivity metric is built into a cell layout optimization engine to minimize the performance gap between process corners. We focus on the best position and spacing of the target poly and active layouts given area constraints of standard cell. The cell layout optimization is formulated into a convex optimization problem for poly layout and a linear optimization for active diffusion layout, respectively; they can be solved efficiently in a global optimal manner.

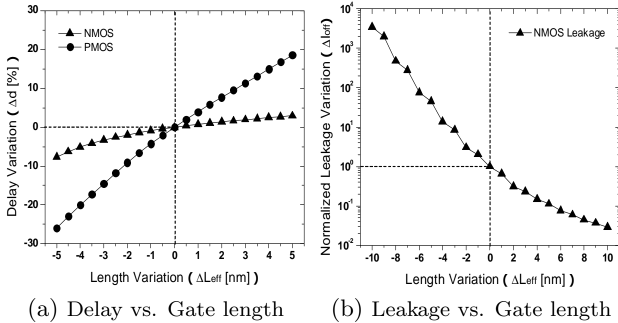
The rest of the paper is organized as follows. Section 2 describes the timing impact of gate length variation and the lithography induced variation. Section 3 presents the total delay sensitivity metric and how to compute it. Section 4 proposes the layout optimization formulation and algorithm using the total sensitivity metric. Experimental results are discussed in Section 5, followed by conclusions in Section 6.

2. PRELIMINARY

2.1 Impact of Gate Length Variation

The most direct impact of systematic gate length variation is the resulting variation of CMOS gate delay and leakage. Figure 1 shows the % delay variation (a) and the % leakage current variation (b) according to the gate length variation in the 45nm node CMOS inverter. In our experiments of the 45nm patterning on a silicon wafer, the gate length variation was up to 10% of the nominal gate length which makes pull-up timing transition delay of un-skewed PMOS decreased over 25% as shown in Figure 1(a). The leakage current variation due to the gate length variation is much more bigger than that of the saturation current or the delay variation. The 10% gate length decrease causes more than tenfold in the leakage current as shown in Figure 1(b). This means that the small improvement to reduce gate length variation can result in significant decrease of the delay and leakage current in a standard cell.

In sub-45nm node standard cell, the gate length variation is still huge (as much as 10% of the nominal gate length) for a semiconductor manufacturing in spite of applying the



(a) Delay vs. Gate length (b) Leakage vs. Gate length

Figure 1: Delay and leakage current variation with gate length variation

strong RET technique like as OPC, an immersion lithography, an off-axis illumination process. This illustrates that all efforts to mitigate these lithography proximity at the final OPC stage are not enough due to restricted design flexibility. Although the lithographic gate length and width variation is a function of the neighboring environment of a cell in a full-chip layout [10], in a sub-45nm mode design industrial standard cells usually have an auxiliary pattern which shields poly patterns near the cell from the proximity effect of neighboring cells. Even though the proximity due to the neighboring cells can not be ignored, our goal in this paper is (1) to mitigate the essential systematic error sources of an intra cell from the lithography proximity and (2) to minimize the overall layout-dependent and process variation impacts by making the layout less sensitive to them and (3) to minimize the delay and leakage variation impacts by introducing topological circuit sensitivity and layout induced sensitivity.

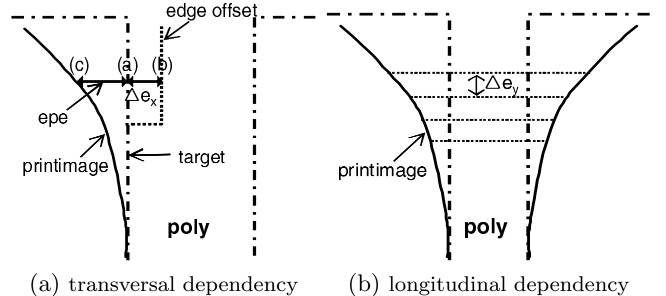
2.2 Lithography Induced Gate Variation

Lithography is an important process step that causes layout (or device geometry) variations. Lithography process is defined by a set of defocus and exposure levels. For nominal defocus and exposure levels, printing of small geometries results in loss of image quality. This results in distorted non-rectangular shapes of the geometries in each layer. Each device in a cell is defined by several mask layers including poly, active, contacts etc. The lithography step impacts all these layers and the rectangular drawn geometries are generally printed as non-rectangular shapes, which depends on the neighborhood geometries in that layer. We term the sensitivity of gate length and gate width variations as layout proximity induced sensitivity.

The lithography induced variation for a nominal process condition can be divided into two components: a transversal (ΔL_x) and a longitudinal (ΔL_y) directional variation of the gate layout as shown in Figure 2. The transversal gate length variation (ΔL_x) results from the spatial frequency of the layout, and it is regarded as the edge placement error (EPE) based on the target layout. The EPE at a given site i (epe_i) is a complex function of mask, then the total gate length is the sum of the target gate length and EPE components. The first-order Taylor series expansion could be used to determine epe_i as a function of edge offset ($\Delta e_{x,i}$) as follows [11]:

$$\Delta L_{x,i} = \frac{\partial L_{x,i}}{\partial e_{x,i}} \Delta e_{x,i}\bigg|_L + \frac{\partial L_{x,i}}{\partial e_{x,i}} \Delta e_{x,i}\bigg|_R . \quad (1)$$

The term ($\frac{\partial L_{x,i}}{\partial e_{x,i}}$) is computed numerically from the Hop-



(a) transversal dependency (b) longitudinal dependency

Figure 2: Lithography induced gate length variation

kins partial coherence equations, and the L and R denote the left and right edge, respectively. Figure 2 (a) shows the definition of each variable of Eq. (1). The point **(c)** is on the printed image resulted in the proximity effect of the point **(a)** on the target layout. To compensate the EPE error, we should pull the edge **(a)** back up to the point **(b)**. Here the difference between **(a)** and **(b)** is the edge offset ($\Delta e_{x,i}$) at a given site i. The longitudinal gate length variation (ΔL_y) is changed through the gate width due to different conduction along the width of the non-rectangular gate region, and it is related with the transversal gate length variation (ΔL_x). To model the longitudinal component, we use a non-rectangular gate model [12].The basic idea is to convert a non-rectangular transistor into several rectangular slices(Δe_y) such that the non-rectangular gate shape is modeled as a single equivalent rectangular transistor with an effective gate length.

3. MODELING OF TOTAL SENSITIVITY

The delay of a standard library cell depends on the circuit topology, the device size, and the layout geometry in the cell. Due to process variations, each device exhibits a certain variation within the cell, meanwhile the device geometry exhibits certain variations due to layout proximity despite a given nominal defocus and exposure levels. Consequently, the total (or effective) delay sensitivity of a cell to process variations should model both (a) the delay sensitivity due to device geometry variations and (b) each device geometry/layout variations due to layout proximity and process variations.

3.1 Circuit Induced Device Criticality

Each cell is characterized for delay sensitivity to gate length and gate width variations using a first-order sensitivity. The variation in each device within the cell results in variation in the delay. Let ΔL_i represent a variation (either gate length or width) in the ith device in a cell with N devices. Then the delay sensitivity, Δd_i for each delay arc, α due to ΔL_i is given as

$$\Delta d_i^\alpha = \frac{\partial d^\alpha}{\partial L_i} \cdot \Delta L_i \tag{2}$$

where the first-order sensitivity $\frac{\partial d^\alpha}{\partial L_i}$ represents the sensitivity contribution for the delay arc, α of the ith device to the cell's delay sensitivity. Considering ΔL_i to be Gaussian distribution, the cell's delay sensitivity due to all devices can be represented as

$$\Delta d^\alpha = \sum_i \frac{\partial d^\alpha}{\partial L_i} \cdot \Delta L_i \tag{3}$$

The delay variation is different from the input delay arcs. Some devices have significant impact on falling arcs while the other devices have significant impact on rising arcs. Thus, to understand the contribution of each device with respect to the cell's total performance, all delay arcs need to be considered together. Consequently, we define a total delay sensitivity index, Ψ as weighted sum of delay-sensitivities due to all delay arcs in a cell. The total sensitivity index for a cell is given as follows:

$$\Psi = \sum_\alpha w^\alpha \cdot \Delta d^\alpha = \sum_\alpha w^\alpha \cdot \sum_i \frac{\partial d^\alpha}{\partial L_i} \cdot \Delta L_i \tag{4}$$

By accumulating all the components of sensitivity due to each device, the Eq. (4) can be rewritten as follows:

$$\Psi = \sum_i \sum_\alpha w^\alpha \cdot \frac{\partial d^\alpha}{\partial L_i} \cdot \Delta L_i = \sum_i \sigma_i \cdot \Delta L_i \tag{5}$$

where, $\sigma_i = \sum_\alpha w^\alpha \frac{\partial d^\alpha}{\partial L_i}$. The total sensitivity index, Ψ now represents a single cell level metric. And, σ_i represents the total weighted sensitivity of the device variation, ΔL_i considering all delay arcs within the cell. That is, σ_i is the contribution of ith device to the total sensitivity index of the cell. We term σ_i as the *device criticality induced sensitivity.* The devices within the cell can be ranked based on their sensitivity contributions to the cell's delay sensitivity. During the layout optimization procedure an additional filter may be added to choose the most sensitive devices first.

3.2 Nominal Lithography Induced Sensitivity

As shown in Section 2.2, lithography induced variation can be classified into two components in timing analysis for non-rectangular gate layout. Given a set of N slices for the gate, with each slice width $W_j = \Delta e_y$, the total current variation could be calculated by summing the current variation per unit width. Note that this current is a function of both the transversal ΔL_x and the longitudinal Δe_y components of given sliced transistor j.

$$\Delta I_{total} = \textstyle\sum_{j=1}^{N} w_j \cdot f\left(\Delta L_{x,j}, \Delta e_y\right) \tag{6}$$

where w_j is the weighting factor which considers the narrow width effect [13].

The longitudinal gate length variation (ΔL_y) is calculated by converting the total current into the gate length which is a function of the layout proximity component (the transversal variation (ΔL_x)) and the device performance component (the current weighting factor (w)). To combine the nominal lithography induced sensitivity and the process induced sensitivity, we define the local layout sensitivity of the each slice with a single metric as follows:

$$\Delta L_j = \Delta L_{y,j} = \frac{\partial L_{y,j}}{\partial e_{y,j}} \cdot \Delta e_{y,j} \tag{7}$$

where $\frac{\partial L_{y,j}}{\partial e_{y,j}}$ is a function of $\Delta L_{x,j}$ and w_j, and we call $\frac{\partial L_{y,j}}{\partial e_{y,j}}$ the *lithography induced sensitivity* for given sliced transistor on nominal process condition.

3.3 Process Variation Induced Sensitivity

This paper reports that the process variation is highly related with the nominal gate length variation due to lithography. To set a systematic sensitivity metric, we first simplify the process variables and then combine the process induced variation into the nominal lithography variation. There are a large number of potential process errors in lithography process, and all variables in lithography either act like dose (linear error), e.g. temperature, photo-resist thickness, and MEEF (mask error enhancement factor), or act like focus (2nd order error), e.g. aberration etc [14]. In the lithographic process, dose and focus errors are the dominant sources of the systematic errors. Using a linear formulation for dose variation parameter, Δp_e and second order formulation for focus variation parameter, Δp_f, the gate length variation can be represented as:

$$\Delta L = \frac{\partial L}{\partial p_e} \Delta p_e + \frac{\partial^2 L}{\partial p_f^2} \Delta p_f^2 \tag{8}$$

where Δp_e is dose error and Δp_f is focus error.

Above equation can be rewritten by using the percentage variation of dose and focus levels for normalization:

$$\Delta L = \frac{\partial L}{\partial \ln p_e} \cdot \%\Delta p_e + \frac{\partial^2 L}{\partial p_f^2} \Delta p_f^2 \quad (9)$$

Note that the effect of focus and exposure levels are actually correlated each other. Thus we should consider both focus and dose error simultaneously and combine them in a single matrix form. Assuming that a CD distribution with focus variation is symmetrical, for a given focus level, Δp_f, the gate length variation is just a function of dose error and can be simplified to a first order form as follows:

$$\begin{aligned} \Delta L &= \frac{\partial L}{\partial \ln p_e}|_{F_0} \left[1 + \alpha \cdot (\Delta p_f)^2\right] \cdot \%\Delta p_e \\ &= \frac{\partial L}{\partial \ln p_e}|_{\Delta p(f)} \cdot \%\Delta p_e \end{aligned} \quad (10)$$

where F_0 is the nominal focus and α is a lithography process-specific constant which is related to wavelength, layout pitch, and refractive index of a material between lens and wafer [14].

In this paper, we set the focus error (Δp_f) around $50nm$ for $45nm$ node device in order to acquire more than $0.1um$ depth-of-focus (DOF) margin. Given focus error, gate length has a different sensitivity from dose variation. Thus, we term $\frac{\partial L}{\partial \ln p_e}|_{\Delta p_f}$ as the *process induced sensitivity* given focus error.

3.4 Total Delay Sensitivity

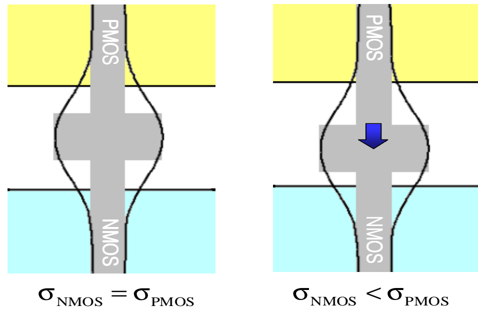
Finally, we can represent a total delay sensitivity by combining the device criticality for all timing arcs from Eq. (5), the local proximity induced sensitivity of the nominal condition from Eq. (7) and the process variation induced sensitivity given focus and dose variation from Eq. (10). As shown in Figure 4 (a), the process induced variation linearly adds on the nominal lithography variation. That means the gate length variation due to process is changed on the basis of the gate length due to the layout proximity. Thus, given a set of N slices for a given transistor i, the total sensitivity is given as

$$\begin{aligned} \Psi_i &= \sigma_i \cdot \sum_{j=1}^{N} \left[\frac{\partial L_{i,j}}{\partial e_{y,j}} \Delta e_{y,j} + \frac{\partial L_{i,j}}{\partial \ln p_e}|_{\Delta p_f} \cdot \pm\%\Delta p_e\right] \\ &= \sigma_i \cdot (\gamma_i + \eta_i)\,|_{\Delta e_y, \pm\%\Delta p_e} \end{aligned} \quad (11)$$

where σ_i is the device criticality induced sensitivity due to unit variation in the geometries, γ_i is the layout induced sensitivity due to local layout proximity at the nominal dose and focus levels, and η_i is the process induced sensitivity due to the process window in which the focus variation is included so that it is a linear formula from a variable, % dose corner variation ($\pm\%\Delta p_e$). If the delay sensitivity of NMOS (σ_p) is smaller than that of PMOS (σ_n) (= PMOS is more critical than NMOS) in a simple inverter, the distance between the PMOS active layout and the poly-contact pad should be larger than that of the NMOS in order to mitigate the delay variation of PMOS by the nominal lithography and process induced sensitivity as shown in Figure 3.

γ_i can have positive or negative value at a given sliced transistor. $\gamma_i > 0$ means the gate length is increased from the nominal, and $\gamma_i < 0$ represents the gate length is decreased. η_i has a positive and negative corner values at the center of γ by lithography dose variation at a given focus tolerance. Δe_y and $\%\Delta p_e$ are user specific variables which means we can designate the sliced width along the gate and the % dose at given focus error in our lithography process. Once we define Δe_y and $\pm\%\Delta p_e$ in the lithography process, we can define the corner values of delay sensitivity ($\pm\Delta d_i$) from the nominal in a sliced transistor. By minimizing the difference between the fastest corner ($-\Delta d_i$) and the slowest corner ($+\Delta d_i$), we can optimize the original layout.

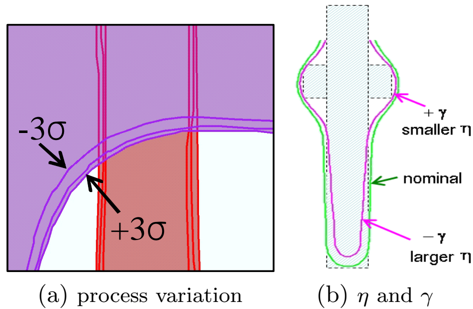
Note that this paper reports that the process induced layout sensitivity (η) is highly correlated with the proximity induced layout sensitivity (γ). As the value of γ goes to the negative direction, the band gap between the best and the worst case corner becomes larger as shown in Figure 4 (b). It means that once the proximity induced sensitivity is calculated, we can estimate the process induced sensitivity. As shown in Figure 5, η is highly dependent on γ that correlates more than 95% in the positive and negative process corners in our experiments. Since a device having a negative γ value has bigger process corner value which causes much more leakage current due to the much deviation of the fastest process corner from the target, we should mitigate the γ variation in order to minimize the performance gap between the fastest process corner and the slowest process corner. By minimizing the performance gap of between the fastest and slowest process corners, we can obtain the process-robust layout.



$\sigma_{NMOS} = \sigma_{PMOS}$ $\sigma_{NMOS} < \sigma_{PMOS}$

(a) same σ_i (b) different σ_i

Figure 3: The topological delay sensitivity σ_i



(a) process variation (b) η and γ

Figure 4: Definition of η and relation with γ

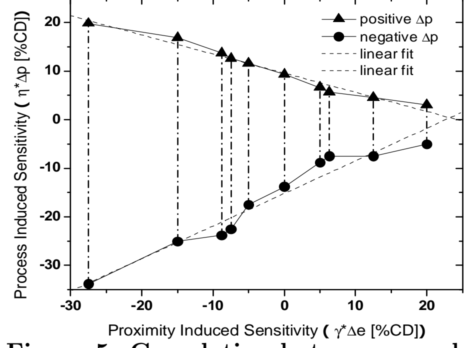


Figure 5: Correlation between γ and η

4. TSDFM: TOTAL SENSITIVITY BASED DFM OPTIMIZATION

4.1 Conventional Cell Optimization Approach

Once the standard cell height and width based on required drive strength are fixed, the cell synthesis and layout optimizations are performed. In the current flow, the standard cell layout optimization is performed by identifying opportunities to enforce as many recommended rules as practically feasible. Since any trivial gate length variation are corrected in the final OPC stage, the critical polygons which are difficult to be optimized in the final OPC stage are just applied to those rules. Those critical polygons on layout are usually poly corner to active area and active corner to poly area. The poly corner comes from poly-to-contact pad, poly routing line, and so on, meanwhile the active corner is caused by active-to-power connection, different skewed or sized devices, and so on.

While implementing design rules, the layout optimization is done targeting poly/active polygons of every device in the layout, without regard to the relative criticality of the devices to variations. Let us revisit the goal of layout optimization for standard cells - the basic objective is to improve parametric yield or reducing systematic variability in cell delay. If there are few devices in the cell that do not exhibit any significant contribution to the systematic delay variations, then any optimization effort on these devices will not help in improving the effective parametric yield. Moreover, in a current recommended rules, since it is difficult for poly and active layout to reflect all proximity rules and process variation rules, the limited information of the systematic variability are just considered. Consequently, there are three issues with the current layout optimization approach for standard cells:

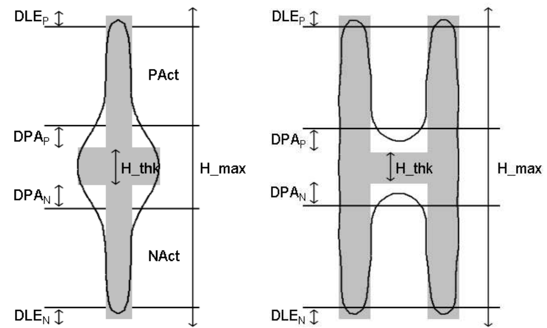
- The design rules are applied to all devices and all layers without any criticality (or sensitivity to variations) information.
- There is no good mechanism to quantify the improvement due to optimization of the standard cells in terms of its performance.
- It is difficult to quantify the impact of systematic lithography proximity effect and its process variation.

In the proposed approach, we use the fact that all devices in a cell are not equally critical and so our model-based approach can take into account the criticality metric of a device. The device criticality as well as the lithography process variability is applied to the cell layout optimization by introducing the proposed total sensitivity.

4.2 Proposed Formulation and Algorithm

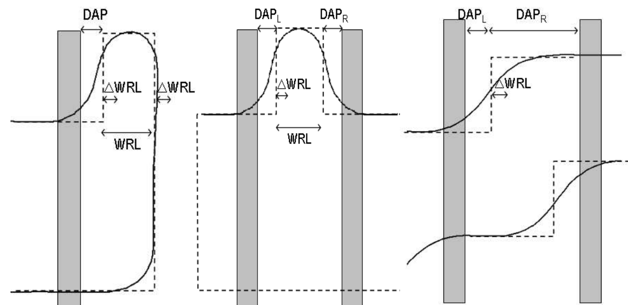
Like as a current optimization of standard cells, we set the optimization variables as the distance of poly corner to active area and active corner to poly area. The representative variables for our model-based approaches are as follows:

- Distance of poly corner to active (DPA): poly corner makes the effective gate length be changed as shown in Figure 6. By providing enough margin, this local sensitivity could be reduced.
- Distance of active corner to poly (DAP): active corner rounding is one of variation sources which results in a slight increase of source-drain current as shown in Figure 7. To prevent gate length from increasing variation, this distance should be increased.
- Distance of poly line-end (DLE): poly line-end is one of lithographic process sensitive areas. It makes the circuit delay decrease, but the leakage current may exponentially increase in this region [9]. By compensating a negative proximity induced sensitivity, such leakage performance degradation can be reduced.



(a) poly contact induced (b) poly routing line induced

Figure 6: Representatives of poly variation



(a) shift active (b) shift/cut active (c) cut active

Figure 7: Representatives of active variation

DPA results in the local gate length variation due to the proximity of gate itself, meanwhile DAP usually causes the local gate width variation due to the active diffusion rounding without any variation on gate line. In a CMOS standard cell, a MOS device is usually connected with other devices by sharing a poly routing line or a poly-metal contact pad. For a example of a CMOS inverter, a pair of p-type and n-type transistor is connected each other at the center of poly-contact pad as shown in Figure 6(a).

4.2.1 Poly Layer Optimization

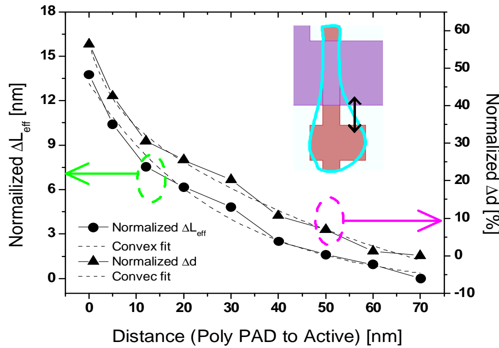
Let us first focus on the poly layout optimization as shown in Figure 8. Let D be the set of geometrically coupled MOS devices (indexed by i), $S(i)$ be the set of slices (j) in i, and $S = \bigcup_{i \in D} S(i)$. The optimization for DPA_i and $DLE_i, \forall i \in D$ can be done as shown in Figure 8 where the objective is to find DPA_i and $DLE_i, \forall i \in D$ which minimize the worst/largest variation among all the devices in D. H_{white} is the space of both actives of the coupled devices, e.g., $H_{white} = DPA_p + H_{thk} + DPA_n$ in Figure 6. H_{thk} is the height of poly-contact or the poly routing line and DLE_{max} comes from the allowable cell height. Since the cell height and width are fixed, we just optimize the layout within a specific area. Therefore, there is no area penalty in our optimization.

$$
\begin{array}{llr}
\textbf{min :} & max \quad \{v_i | \forall i \in D\} & \\
\textbf{s.t :} & & \\
\textbf{(a)} & v_i = (|\Delta d_{i.max}| + |\Delta d_{i.min}|) & \forall i \in D \\
\textbf{(b)} & \Delta d_{i.max} \geq \sigma_i \sum_{j \in S(i)} (\gamma_{ij} + |\eta_{ij}|)|_{\Delta e_y, \%\Delta p} & \forall i \in D \\
\textbf{(c)} & \Delta d_{i.min} \leq \sigma_i \sum_{j \in S(i)} (\gamma_{ij} - |\eta_{ij}|)|_{\Delta e_y, \%\Delta p} & \forall i \in D \\
\textbf{(d)} & \gamma_{ij} \geq a \cdot \sqrt{DPA_i} + b \cdot DLE_i + c & \forall j \in S \\
\textbf{(e)} & \eta_{ij} \geq d \cdot \Delta p_i \cdot \gamma_{ij} + e & \forall j \in S \\
\textbf{(f)} & DPA_{min} \leq DPA_i \leq DPA_{max} & \forall i \in D \\
\textbf{(g)} & DLE_{min} \leq DPA_i \leq DLE_{max} & \forall i \in D \\
\textbf{(h)} & \sum_{i \in D} DPA_i = H_{white} - H_{thk} & \\
\textbf{(i)} & \sum_{i \in D} DLE_i = H_{total} - P_{active} - N_{active} &
\end{array}
$$

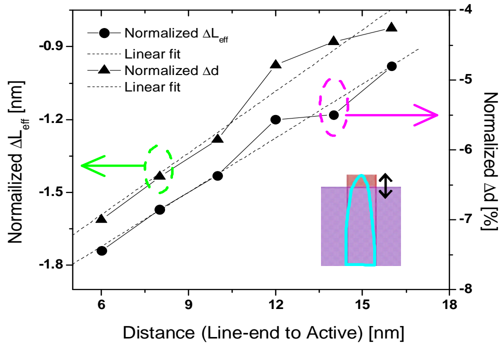
Figure 8: Convex optimization for DPA and DLE

The constraint **(a)** is to compute the performance gap between two delay corners (fastest and slowest) which can be computed as in the constraints **(b,c)** where σ_i is the device criticality induced sensitivity, γ_i is the layout induced sensitivity, and η_i is the process induced sensitivity as we defined in Eq. (11). The γ_{ij} for the j-th slice of $i \in D$ can be obtained from Figure 9, which results in the constraint **(d)**. There are lots of layout shapes of poly corner to active in standard cells. Thus we chose a polygon shape which causes the most severe lithography proximity for an upper bound case, guaranteeing that the gate length variation and its delay impact is never underestimated in our cell library.

In Figure 9, the gate length variation (left Y-axis) and the normalized delay sensitivity (right Y-axis) are reported in terms of the distance between poly corner and active layout in Figure 9(a) which shows an exponential or negative second root trend for both gate length and delay sensitivity.



(a) Poly corner to active (DPA)



(b) Poly line-end to active (DLE)

Figure 9: Delay with poly layout variation

To make a convex function, we approximate the delay trend with a convex equation. Figure 9(b) which has a positive linear trend shows the results with the distance of poly line-end to active. In a similar way, η_{ij} can be described as a function of both γ_{ij} and Δp_i as shown in the constraint **(e)**.

The constraints **(f,g,h,i)** are to satisfy the technology and DRC requirements. **(a,b,c,d)**, and **(e)** are process-dependent parameters extracted from Figure 9. Since $a < 0$ for any process technology, the constraint **(d)** is convex, which enables to solve the formulation in Figure 8 in polynomial time [15]. Also, due to the convexity, we can obtain the globally optimal DPA_i and $DLE_i, \forall i \in D$ which will reduce the largest delay variation among all the devices optimally. By minimizing the total delay sensitivity and by reducing the gap between the fastest and the slowest delay corner, we can achieve the delay reduction on the nominal process condition and the leakage reduction on the process corners.

4.2.2 *Active Layer Optimization*

In a similar fashion, active layout could be optimized by preventing active from corner rounding. Since active layout is much bigger than poly layout in our standard cell, the process induced variation of active is not much sensitive compared to poly layout. Nonetheless, we also optimize active layout because active rounding is one of sources changing the performance estimation of standard cells [16]. There are two kinds of active optimization, e.g., cutting and shifting, to mitigate active corner rounding as shown in Figure 7. Shifting is happened when the DRC margin is enough on the opposite active side from the poly line in Figure 7(a) whereas we cut the active layout when there is some margin on the power rail of active (b) or on the detoured active (c).

The optimization for $DAP_i, \forall i \in D$ can be done as shown in Fig. 10 where the objective is to find $DAP_i, \forall i \in D$ which minimize the amplitude of gate proximity induced variation (γ) among all the devices in D. W_{white} is the space of both

$$
\begin{array}{llr}
\textbf{min :} & \{|\Delta d_i| | \forall i \in D\} & \\
\textbf{s.t :} & & \\
\textbf{(a)} & \Delta d_i \geq \sigma_i \sum_{j \in S(i)} \gamma_{ij}|_{\Delta e_y} & \forall i \in D \\
\textbf{(b)} & \gamma_{ij} \geq a \cdot DAP_i + b & \forall j \in S \\
\textbf{(c)} & DAP_{min} \leq DAP_i \leq DPA_{max} & \forall i \in D \\
\textbf{(d)} & WRL_{min} \leq WRL & \forall i \in D \\
\textbf{(e)} & \sum_{i \in D} DAP_i = W_{white} - WRL &
\end{array}
$$

Figure 10: Linear optimization for DAP

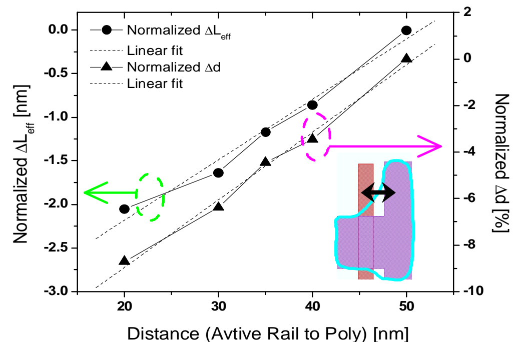


Figure 11: Delay with active layout variation

poly lines of the coupled devices, e.g., $W_{white} = DAP_L + WRL + DAP_R$ in Figure 7. WRL is the width of active layout which can be cut in an allowable DRC margin.

The constraint **(a)** is to compute the delay variation caused by the active layout proximity which can be computed as in the constraints **(b)**. The upper bound EPE variation, γ_{ij} for the j-th slice of $i \in D$ can be obtained from Figure 11, which results in the constraint **(b)**. Figure 11 shows the gate length variation (left Y-axis) and the normalized delay sensitivity (right Y-axis) with the distance between active corner and poly layout. Since the constraint **(b)** is linear, we can solve the formulation in Figure 8 in polynomial time which leads to the globally optimal $DAP_i, \forall i \in D$ to reduce the largest delay variation among all the devices optimally.

4.3 TSDFM Flow

Figure 12 illustrates our TSDFM flow. The flow is divided into three main steps:

1. Calculation of the topological sensitivity (σ): Cell characterization for delay sensitivities is first performed. Then, the devices are ranked for their criticality within a cell based on the sensitivities for all delay arcs.

2. Calculation of the layout sensitivity (γ, η): Based on the non-rectangular shape in the poly and diffusion layers, we define the local layout proximity on nominal lithography condition and the process induced sensitivity given focus and dose condition.

3. Layout optimization: Give all sensitivities, we optimize poly and active layout successively using a convex optimization and a linear programming with the DRC and area constraints.

We check the whole devices in a cell until the total sensitivity of a device has the minimum value for its all timing arc. All sequences are automated with Tcl and Perl script languages.

5. EXPERIMENTAL RESULTS

We implemented TSDFM in Tcl and Perl script language and tested with the industrial 45nm ASIC designs. We used Calibre-WB from Mentor Graphics for model based OPC. The timing analysis and characterization were done by H-Spice circuit simulator from Synopsys. In order to model and solve the convex/linear formulation, we used AMPL /MOSEK 5.0 [17, 18]. All the sequences are implemented and automated in our cell characterization flow. Figure 13 shows the print-image results of the poly and active optimized layout. The black lines are the original layout, and the accompanying graded lines are the optimized layout.

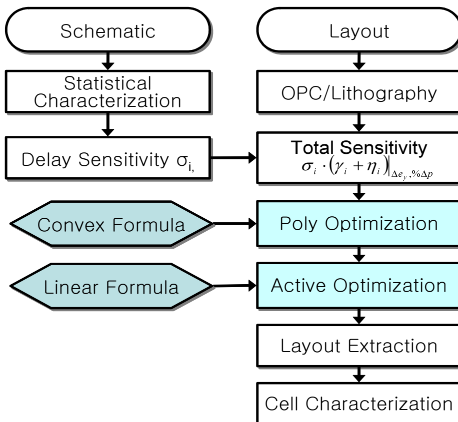
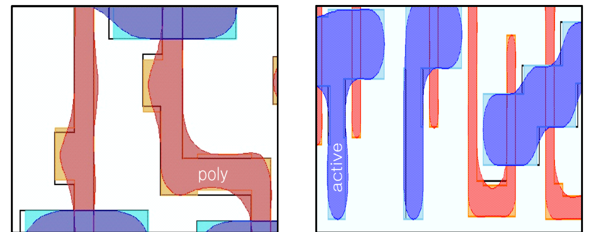


Figure 12: TSDFM driven cell optimization flow

We applied our optimization to the entire 45nm standard cell library. Table 1 shows the delay variation from the nominal delay, and we selected one representative pin without loss of generality. To report our results, we compared our results with a conventional restricted design rule (RDR) approach. We first compared the impact of circuit criticality on delay and measured the % delay difference from the target delay in the column CKT. The lithography simulation was done at the nominal lithography process condition. The results show up to 76% improvement in delay variation. The average improvement of delay variation in entire cells are 43.43%. This implies that the topological sensitivity should be considered when we optimize the cell layout.

In the column PV in Table 1, we compared the impact of process robustness on delay and measured the % delay difference between the slowest (thickest) process corner and the fastest (thinnest) process corner. Thus, each transistor has the same criticality for delay in this case. Since our total sensitivity metric uses an approach to minimize the performance gap between the slowest and the fasted process corner, the results show up to 16% improvement in spite of not considering the device criticality. When the total sensitivity is applied in the column CKT & PV, we can reduce the performance gap due to the process corners as much as 24%. Note that we optimized the cell layout given the cell area constraint, thus there is any area penalty. This means that we could expect more improvement in minimizing the performance variation if we had more area margin of a cell.

For leakage power, we measured the local maximal leakage current which is extracted at the region of a device, e.g. line-end, in which the gate length is the smallest at the fastest process corner. The result of Table 2 shows that the local maximum leakage in a device is decreased up to 91.9% in a cell and as much as 57.5% on average in the D-type Flip-Flop cell. Note that despite the small improvement of gate length variation, we can see the huge amount of improvement on leakage current as I mentioned in Section 2 and Figure 1(b). Another point we should note is that the conventional approach (CONV) has more leakage current than that of the proposed approach (TSDFM) in spite of applying the same OPC and lithography model. This is because the conventional approach cares for the total gate length variation of a whole gate transistor whereas our approach considers the local proximity and process variation effects. This results show that our total sensitivity driven layout optimization is capable of reducing both delay and leakage current given the design constraints.



(a) poly optimization (b) active optimization

Figure 13: Print-image view of optimized layouts

Table 1: Reduction of Delay Variation

Cell	Δdelay with CKT[a]			Δdelay with PV[b]			Δdelay with CKT&PV[c]		
	CONV[d]	TSDFM[e]	Improve	CONV[d]	TSDFM[e]	Improve	CONV[d]	TSDFM[e]	Improve
AOI12X12	2.95%	1.29%	56.20%	19.91%	18.11%	9.04%	19.91%	15.23%	23.51%
CBI4I1X2	1.84%	0.52%	71.63%	18.32%	15.91%	13.19%	18.32%	13.94%	23.90%
FA1X5	3.02%	1.66%	44.98%	22.13%	20.08%	9.27%	22.13%	17.82%	19.50%
HA1X5	5.01%	3.66%	26.99%	22.33%	20.96%	6.14%	22.33%	18.84%	15.64%
MUX21X5	2.78%	1.71%	38.39%	20.34%	17.94%	11.82%	20.34%	15.95%	21.60%
OA12X10	2.68%	1.65%	38.42%	19.03%	18.91%	0.62%	19.03%	18.18%	4.47%
OAI12X10	1.86%	0.51%	72.71%	16.55%	15.88%	4.07%	16.55%	14.16%	14.42%
PAO2X2	4.65%	3.38%	27.18%	21.10%	20.40%	3.31%	21.10%	19.86%	5.87%
XNOR2X2	2.27%	0.93%	58.86%	23.02%	19.36%	15.90%	23.02%	18.59%	19.27%
XOR2X2	3.71%	1.61%	56.63%	18.33%	17.87%	2.48%	18.33%	16.76%	8.55%

[a] The impact of circuit criticality on delay. The lithography simulation is done at the nominal process condition.
[b] The impact of process robustness on delay. Each transistor has the same criticality for delay.
[c] The impact of process robustness on delay. Each transistor has different criticality for delay.
[d] A conventional optimization (RDR) approach.
[e] The total sensitivity driven optimization approach.

Table 2: Reduction of Leakage Current

Position[a]	ΔL	Leakage	Inc[b]	ΔL	Leakage	Inc[b]	Improve
	CONV[c]			TSDFM			(%)
P1	-2.26	2.28E-08	8.73	-1.27	5.40E-09	1.30	85.12
P2	-1.28	5.43E-09	1.31	-0.94	4.61E-09	0.96	26.45
P3	-1.83	6.74E-09	1.87	-1.19	5.20E-09	1.21	35.13
P4	-2.90	3.08E-08	12.10	-1.08	4.94E-09	1.10	90.91
P5	-1.43	5.78E-09	1.46	-1.33	5.54E-09	1.36	7.07
P6	-1.86	6.80E-09	1.89	-0.54	3.63E-09	0.55	71.12
P7	-2.76	2.91E-08	11.40	-1.18	5.17E-09	1.20	89.46
P8	-2.79	2.94E-08	11.52	-2.54	2.63E-08	10.21	11.37

[a] we measured local maximal leakage of D-type Flip-Flop.
[b] Inc is a leakage increment which is a multiple of the nominal leakage current.
[c] A conventional optimization (RDR) approach.

6. CONCLUSION

We have proposed a novel layout optimization approach in standard cell library to minimize the delay sensitivity due to the gate length variation caused by the layout proximity and lithographic process variation at 45nm and below. Our approach practically and effectively improves the circuit performance and hence yield; it has been implemented using a TCL script language, MOSEK convex optimization and linear programming solver. Experimental results with a industrial cell library show that our model-based layout optimization approach can highly decrease the delay and the leakage variation by minimizing the total delay sensitivity and by reducing the gap between the fastest and the slowest delay corner in given layout constraints. In this paper, we focused on a method of intra cell robustness, and we plan to research on the impact of inter cell proximity (e.g. neighboring effect) and the sensitivity aware routing and placement algorithms.

7. ACKNOWLEDGMENTS

This work is supported in part by SRC, NSF CAREER Award, and equipment donations from Intel.

8. REFERENCES

[1] M. Orshansky, L. Milor, P. Chen, K. Keutzer, and C. Hu. Impact of Spatial Intrachip Gate Length Variability on the Performance of High-Speed Digital Circuits. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 21(5):544–553, 2002.

[2] A. Subramaniam, R. Singhal, C. Wang, and Y. Cao. Design rule optimization of regular layout for leakage reduction in nanoscale design. In *Proc. Asia and South Pacific Design Automation Conf.*, Jan 2008.

[3] L. W. Liebmann. Resolution enhancement techniques in optical lithography: It's not just a mask problem. In *Proc. SPIE 4409.*, pages 23–32, September 2001.

[4] M. Cho, K. Yuan, Yongchan Ban, and D. Pan. ELIAD: Efficient Lithography Aware Detailed Router with Compact Post-OPC Printability Prediction. In *Proc. Design Automation Conf.*, Jun 2008.

[5] M. Cote and P. Hurat. Layout Printability Optimization using a Silicon Simulation Methodology. In *Proc. Int. Symp. on Quality Electronic Design*, 2004.

[6] D. Pawlowski, L. Deng, and M. Wong. Fast and accurate opc for standard-cell layouts. In *Proc. Asia and South Pacific Design Automation Conf.*, Jan 2007.

[7] V. Kheterpal, T. Hersan, V. Rovner, D. Motiani, Y. Takagawa, L. Pileggi, and A. Strojwas. Design methodology for ic manufacturability based on regular logic-bricks. In *Proc. Design Automation Conf.*, Jun 2005.

[8] J. Wang, A. Wong, and E. Lam. Performance optimization for gridded-layout standard cells. In *Proc. SPIE 5567*, 2004.

[9] A. Subramaniam, R. Singhal, C. Wang, and Yu Cao. Design Rule Optimization of Regular layout for Leakage Reduction in Nanoscale Design. In *Proc. Asia and South Pacific Design Automation Conf.*, Jan 2008.

[10] A. Kahng, S. Muddu, and C. Park. Auxiliary Pattern-Based OPC for Better Printability, Timing and Leakage Control. *SPIE J. Microlithography, Microfabrication and Microsystems*, 7(1), 2008.

[11] N. Cobb, A. Zakhor, and E. Miloslavsky. Mathematical and CAD Framework for Proximity Correction. In *Proc. SPIE 2726*, 1996.

[12] R. Singhal, A. Balijepalli, A. Subramaniam, F. Liu, S. Nassif, and Y. Cao. Modeling and analysis of non-rectangular gate for post-lithography circuit simulation. In *Proc. Design Automation Conf.*, Jun 2007.

[13] K. Tsai, M. You, Y. Lu, and P. Ng. A new method to improve accuracy of leakage current estimation for transistors with non-rectangular gates due to sub-wavelength lithography effects. In *Proc. Int. Conf. on Computer Aided Design*, Nov 2008.

[14] C. Mack. *Fundamental Principles of Optical Lithography.* Wiley, 2007.

[15] S.P. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003.

[16] P. Gupta, A. Kahng, Y. Kim, S. Shah, and D. Sylvester. Investigation of diffusion rounding for post-lithography analysis. In *Proc. Asia and South Pacific Design Automation Conf.*, Jan 2008.

[17] http://www.ampl.com.

[18] http://www.mosek.com.