Dose Map and Placement Co-Optimization for Improved Timing Yield and Leakage Power

Kwangok Jeong, *Student Member, IEEE,* Andrew B. Kahng, *Fellow, IEEE,* Chul-Hong Park, *Member, IEEE,* and Hailong Yao, *Member, IEEE*

***Abstract*—In sub-100 nm CMOS processes, delay and leakage power reduction continue to be among the most critical design concerns. We propose to exploit the recent availability of fine-grain exposure dose control in the step-and-scan tool to achieve both design-time (placement) and manufacturing-time (yield-aware dose mapping) optimizations of timing yield and leakage power. Our placement and dose map co-optimization can improve both timing yield and leakage power of a given design. We formulate the placement-aware dose map optimization as quadratic and quadratic constraint programs which are solved using efficient quadratic program solvers. In this paper, we mainly focus on the placement-aware dose map optimization problem; in the Appendix, we describe a complementary but less impactful dose map-aware placement optimization based on an efficient cell swapping heuristic. Experimental results show noticeable improvements in minimum cycle time without leakage power increase, or in leakage power reduction without degradation of circuit performance.**

***Index Terms*—Dose map, leakage power reduction, placement, timing yield.**

I. Introduction

Continued scaling of feature sizes in integrated circuits (ICs) drives improvements of integration complexity and device speed with each successive technology node. In sub-100 nm process nodes, manufacturing variations are the primary sources of design performance variability and parametric yield loss. To minimize the impact of manufacturing variations on performance variability, the manufacturing process itself can be improved, and/or designs can be made robust to variations. Improvements to the manufacturing process require, most prominently, advanced techniques in reticle enhancement, mask making, and optical lithographic equipment—all of which increase the manufacturing cost and subsequently the design cost. As a result, so-called design for manufacturability techniques [2] have received great attention within the electronic design and electronic design automation communities.

Critical dimension (CD) variation is a dominant factor in the variation of delay and leakage current of transistor gates in integrated circuits. With advanced manufacturing processes, CD variation is worsening due to a variety of systematic variation sources at both within-die and reticle or wafer-scale; the latter sources include radial bias of spin-on photoresist thickness, etcher bias, reticle bending, uniformity of wafer starting materials, and so on [17]. A statistical leakage minimization method is proposed in [3], which obtains significant improvement in total leakage reduction by simultaneously varying the threshold voltage, gate sizes and gate lengths. Gupta *et al.* [4] proposed to apply gate-length (CD) biasing only on the devices in non-critical paths for leakage power control without negative effects on timing.

A recent technology from ASML, called *DoseMapper* [8], [9], allows for minimization of across-chip linewidth variation (ACLV) and across-wafer linewidth variation (AWLV)[1] using an exposure dose (or, simply, dose) correction scheme. DoseMapper in the ASML tool parlance exercises two degrees of control, *Unicom-XL* and *Dosicom* [6], which respectively change dose profiles along the lens slit and the scan directions of the step-and-scan exposure tool.

Today, the DoseMapper technique is used solely (albeit very effectively, e.g., [7]) to reduce ACLV or AWLV metrics for a given integrated circuit during the manufacturing process. However, to achieve optimum device performance (e.g., clock frequency) or parametric yield (e.g., total chip leakage power), not all transistor gate CD values should necessarily be the same. For devices on setup timing-critical paths in a given design, a larger than nominal dose on poly layer (causing a

Manuscript received June 3, 2009, revised November 10, 2009. Date of current version June 18, 2010. A preliminary version of this paper appeared in [1]. Extensions beyond [1] include discussion and experiments on granularities of the partitioned exposure field for dose tuning, new problem formulations for dose map optimization either for improved leakage under timing constraint or for improved circuit performance under leakage constraint, dose map optimization on poly and active layers simultaneously, and improved cell swapping strategy for further timing improvement after dose map optimization. We also test our algorithms using a set of new 65 nm testcases besides the 90 nm testcases used in [1]. This paper was recommended by Associate Editor, J. Hu.

K. Jeong is with the Department of Electrical and Computer Engineering, University of California at San Diego (UCSD), La Jolla, CA 92093 USA (e-mail: kjeong@vlsicad.ucsd.edu).

A. B. Kahng is with the Department of Computer Science and Engineering and the Department of Electrical and Computer Engineering, University of California at San Diego (UCSD), La Jolla, CA 92093 USA (e-mail: abk@ucsd.edu).

C.-H. Park is with the Semiconductor Research and Development Center, Samsung Electronics, Seoul 137-857, Korea (e-mail: chul.h.park@samsung.com).

H. Yao was with the Department of Computer Science and Engineering, University of California at San Diego (UCSD), La Jolla, CA 92093 USA. He is now with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: hailongyao@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2010.2048397

[1] ACLV is primarily caused by the mask and scanner, while AWLV is affected by the track and etcher [10].

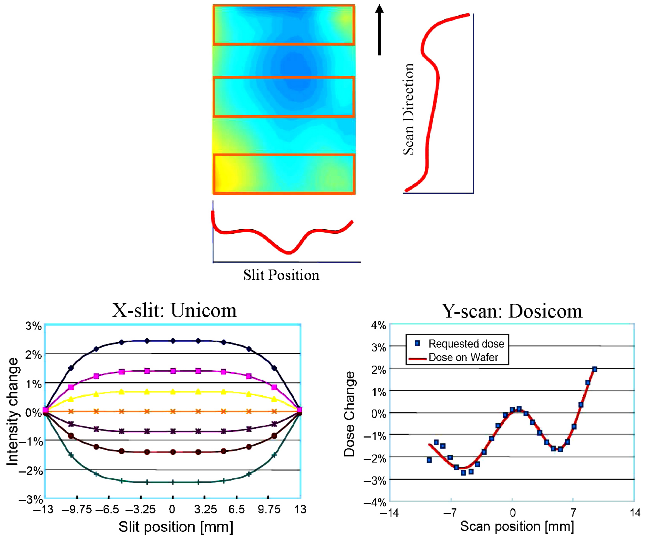


Fig. 1. Unicom-XL and Dosicom, which change dose profiles in slit and scan-directions, respectively. Source: [5].

smaller than nominal gate CD) will be desirable, since this creates a faster-switching transistor. On the other hand, for devices that are on hold timing-critical paths, or in general that are not setup-critical, a smaller than nominal dose on poly layer (causing a larger than nominal gate CD) will be desirable, since this creates a less leaky (although slower-switching) transistor. What has been missing, up to now, is any connection of such "design awareness"—that is, the knowledge of which transistors in the integrated-circuit product are setup or hold timing-critical—with the calculation of the DoseMapper solution.[2] The Zeiss/Pixer Critical Dimension Control (CDC) technology [16] also enables adaptivity in the manufacturing flow to meet the required CD specifications. The CDC technology modifies the local mask transmissivity (which translates into local CD changes on the wafer during the lithography process) without removing the pellicle, thus allowing for tool installations either at the mask manufacturing site, or at the fab line. In this paper, we focus on the DoseMapper technology for tuning of transistor gate dimensions.

We propose a novel method to enhance timing yield as well as reduce leakage power by combined dose map and placement optimizations. The contributions of this paper are as follows.

1) A novel method of enhancing circuit performance and parametric yield based on the dose map technology.
2) A new design-aware and equipment-aware dose map optimization (*DMopt*) method that uses dose to modulate gate CD (for poly layer) and/or gate width (for active layer) across the exposure field, so as to either optimize timing under the constraint of leakage power or reduce leakage power under the constraint of timing.
3) A new dose map-aware placement optimization (*dosePl*) heuristic that considers systematic gate CD changes at different areas within a given dose map, and seeks to optimize circuit timing yield by selectively re-placing critical and near-critical cell instances based on golden extraction and timing analysis results.

Note that two distinct optimizations are possible, i.e., the placement-aware dose map optimization (*DMopt*) and the dose map-aware placement optimization (*dosePl*). This paper mainly focuses on *DMopt*. However, *dosePl* (in the Appendix) is also attempted. This paper is organized as follows. Section II introduces fundamentals of the DoseMapper concept and the dose map optimization problems for improved timing yield and for reduced total leakage power. Section III provides details of our problem formulations for design-aware dose map optimization. Section IV discusses the overall optimization flow, and experimental results are presented in Section V. In Section VI, the conclusion is drawn and further research directions are presented.

II. Preliminaries of Dose Map Optimization Problem

A. *DoseMapper Fundamentals*

Fig. 1 shows the intrafield DoseMapper concept. In Fig. 1, the slit exposure correction is performed by Unicom XL. The actuator is a variable-profile gray filter inserted in the light path. The default filter has a second-order (quadratic) profile, and ASML [11] recommends use of a quadratic slit profile to model data in the slit direction. It is also possible to obtain a customized profile: lithography systems with Unicom XL (e.g., the XT:1700i machine) support a slit profile represented by polynomials of up to the 6^{th} order in the dose recipe. Overall, a correction range of ±5% can be obtained with Unicom-XL for the full field size of 26 mm in the X-direction.

Scan exposure correction is realized by means of Dosicom, which changes the dose profile along the scan direction. The dose generally varies only gradually during scanning, but the dose profile can contain higher-order corrections depending on the exposure settings. The dose set, $D_{\text{set}}(y)$, is used to model parameters for a dose recipe formed of Legendre polynomials (Legendre functions of the first kind) as

$$D_{\text{set}}(y) = \sum_{n=1}^{8} L_n P_n(y) \tag{1}$$

where y is a floating variable ($|y| \le 1$) related to the scan position, L_n are Legendre coefficients, and $P_n(y)$ are Legendre polynomials of variable y. Up to eight Legendre coefficients can be supported. The correction range for the scan direction is ±5% (10% full range) from the nominal energy of the laser. When the requested X-slit and Y-scan profiles are sent to the lithography system, they are converted to system actuator settings (one Unicom-XL shift for all fields, and a dose offset and pulse energy profile per field).

Dose sensitivity is the relation between dose and critical dimension, measured as CD (nm) per percentage (%) change in dose. Increasing dose decreases CD as shown in Fig. 2, i.e., the dose sensitivity has negative value. To calculate the dose sensitivity [$\triangle CD/\triangle E$, (nm/%)], a focus-exposure matrix (FEM) must be exposed on a product wafer for each product

[2] Optimization of gate CDs according to setup or hold timing (non-)criticality has been used by [4]. What we propose below uses a coarser knob (i.e., the dose map) for design-aware gate CD control, but has the advantage of not requiring any change to the mask or OPC flows.

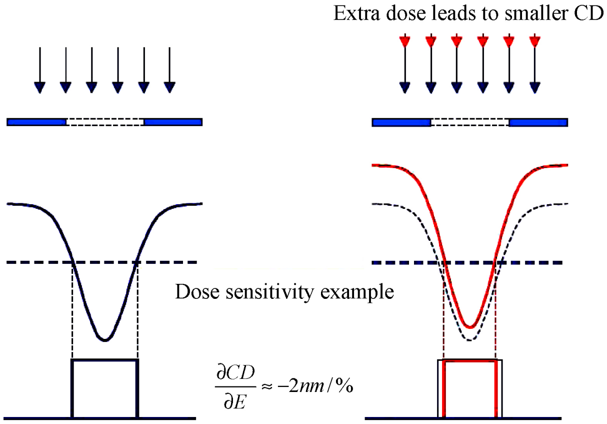

Extra dose leads to smaller CD
Dose sensitivity example
$\frac{\partial CD}{\partial E} \approx -2nm/\%$

Fig. 2. Dose sensitivity: Increasing dose (red color) decreases the CD. Source: [11].

layer using standard production settings [e.g., reticle (6% attPSM), resist, and illumination settings].

B. Dose Map Optimization Problem

The design-aware dose map problem, for the objective of timing yield and leakage power, can be stated as follows. *Given placement P with timing analysis results, determine the dose map to improve timing yield as well as reduce total device leakage.* Specifically, there are two dose map optimization problems with different objectives: one is to minimize total leakage power under a clock period upper bound constraint, and the other is to minimize clock period under a leakage power upper bound constraint. These two dose map optimizations may be formulated using quadratic program (QP) and quadratically constrained program (QCP) (see Section III), and solved using efficient commercial solvers [13].

In the following, for simplicity of exposition we assume that the reticle area taken up by a single copy of the integrated circuit is the same as the area of the exposure field. In practice, the exposure field will contain one or more copies of the integrated circuit(s) being manufactured. It is simple to extend our proposed algorithms to the case where the exposure field contains multiple copies of the integrated circuit(s) being manufactured: smoothness or gradient constraints are scaled, and multiple copies of the dose map solution are tiled horizontally and vertically.

For the dose map optimization problem, we partition the exposure field into a set of rectangular grids $R = |r_{i,j}|_{M\times N}$ on both active and poly layers, where the (uniform) width and height of rectangular grid $r_{i,j}$ are both less than or equal to a user-specified parameter G. G controls the granularity of the dissected rectangular grids: a smaller value of G corresponds to a larger number of rectangular grids, along with a more precisely specified new dose map and better timing yield and/or leakage power improvement. However, G cannot be set too small, due to the DoseMapper equipment limitations. In general, G can be determined so as to balance between DoseMapper equipment constraints and timing yield and/or leakage power improvement, and different values of G may be used for different layers. In the discussion below, we assume that the same G values are used for both active and poly layers.

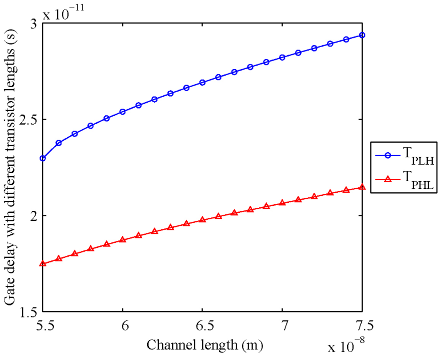


Fig. 3. Delay of an inverter versus gate length.

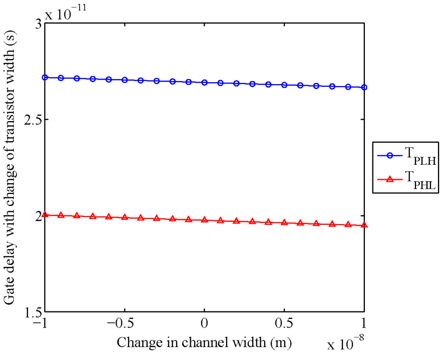


Fig. 4. Delay of an inverter versus change in gate width.

Dose map optimization using different granularities of the partitioned rectangular grids is tested and discussed in Section V.

In this paper, we mainly focus on dose map optimization on the poly layer, i.e., for modulation of gate length. We have also tried dose map optimization on both the active and poly layers to simultaneously modulate gate width and length when optimizing timing and leakage power. For consistency of exposition, in the following sections we state the circuit delay and leakage power estimation equations, as well as our problem formulations, considering both gate width and gate length variations.

C. Circuit Delay and Leakage Power Calculation

We assume the dose sensitivity D_s to have the typical value of -2 nm/% [7] in our experimental evaluations. Gate length and gate width change linearly with dose tuning, i.e., $\Delta L_p = D_s \times d^P_{i,j}(p)$ and $\Delta W_p = D_s \times d^A_{i,j}(p)$, where ΔL_p is the change in gate length of gate p, ΔW_p is the change in gate width of gate p, and $d^P_{i,j}(p)$ and $d^A_{i,j}(p)$ are percentage values which specify the relative changes of dose for poly and active layers in the rectangular grid $r_{i,j}$ wherein gate p is located.

Fig. 3 shows SPICE-calculated delay values as gate lengths are varied in an inverter that is implemented in 65 nm technology with equal channel lengths of the PMOS and NMOS devices. Fig. 4 shows SPICE-calculated inverter delay values as gate widths of the PMOS and NMOS devices are changed by the same delta value. In Figs. 3 and 4, T_{PLH} and T_{PHL} represent the low to high propagation delay and the high to low propagation delay, respectively. From the two figures, the gate delay varies linearly with both gate length and gate width around the nominal feature size (i.e., 65 nm) and the original transistor widths. Our background experiments tested Liberty delay model tables of 36 different 65 nm standard cell masters, and confirmed in all cell masters such an approximate linear relationship at each pair of input slew and load capacitance values. Similar studies at 90 nm were conducted in [1].

When gate length and/or gate width changes in a small range, the effects of the change on other topologically adjacent gates are typically small.[3] Hence, we assume that the gate delay decreases linearly as the gate width increases, and increases linearly as the gate length increases. Since gate length (width) changes linearly when the dose on the gate for poly (active) layer varies, there is a linear relationship between the change of gate delay and the change of exposure dose on the gate for both poly and active layers, i.e., $\Delta t_p = t'_p - t_p = A_p \times \Delta L_p + B_p \times \Delta W_p = A_p \times D_s \times d^P_{i,j}(p) + B_p \times D_s \times d^A_{i,j}(p)$. Here, t_p and t'_p are the delay values of gate p before and after the percentage dose changes $d^P_{i,j}(p)$ and $d^A_{i,j}(p)$ on poly and active layers in the rectangular grid $r_{i,j}$ where gate p is located, ΔL_p and ΔW_p are the changes in gate length and gate width of gate p, and A_p and B_p are fitted parameters that are dependent on input slew and load capacitance of each gate. In other words, for each distinct standard cell, and for each combination of input slew and load capacitance, different values of A_p and B_p are obtained from processing of Liberty nonlinear delay model tables. Total runtime of this procedure for the 65 nm production standard-cell library (36 combinational cells and nine sequential cells) is less than 1 min on a single processor using our Liberty processing and curve fitting tool. The fitted parameters can also be used to compute the change in gate delay when only the dose on poly layer changes (i.e., only for gate length modulation), in which case the dose change on active layer $[d^A_{i,j}(p)]$ is 0.

For circuit delay calculation, without loss of generality we consider a combinational circuit with n gates as in [12]. Sequential circuits may be addressed similarly, e.g., by "unrolling" them, using standard techniques, into combinational circuits that traverse from primary inputs and sequential cell outputs, to sequential cell inputs and primary outputs. For a given combinational circuit, we add to the corresponding circuit graph one fictitious source node, which connects to all primary inputs, and one fictitious sink node, which connects from all primary outputs. Nodes are indexed by a reverse topological ordering of the circuit graph, with the source and sink nodes indexed as $n + 1$ and 0, respectively.

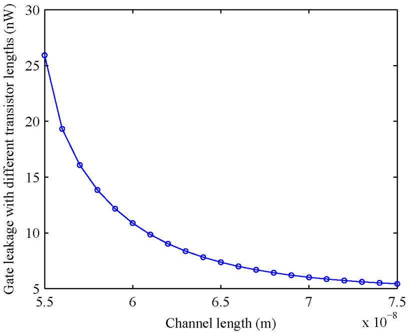


Fig. 5. Average leakage of a 1X inverter (INVX1) versus gate length (VDD = +1.0-V, temperature = +25 °C, process = TT).

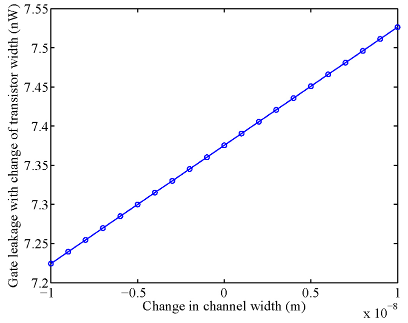


Fig. 6. Average leakage of a 1X inverter (INVX1) versus the change in gate width (VDD = +1.0-V, temperature = +25 °C, process = TT).

Fig. 5 shows SPICE-calculated average transistor leakage values with simulation condition (VDD = +1.0-V, temperature = +25 °C, process = TT) as gate lengths are varied in a minimum-size inverter that is implemented in 65 nm technology, where channel lengths of the PMOS and NMOS devices are equal. Fig. 6 shows SPICE-calculated average transistor leakage values with the same inverter and simulation condition, as all channel widths of the PMOS and NMOS devices are changed by the same delta value. From Figs. 5 and 6, the gate leakage varies exponentially with gate length and linearly with the change in gate width around the nominal feature size (i.e., 65 nm) and the original transistor widths. We also performed background experiments on Liberty leakage values of 36 different standard cell masters, and confirmed these exponential (linear) relationships between gate leakage and gate length (width). Similar analyses for the 90 nm technology node can be found in [1]. In our optimization, we assume that the change of leakage power of a gate is a quadratic function of the change in gate length[4] and a linear function of the change in gate width, i.e., $\Delta Leakage(\Delta L_p, \Delta W_p) = \alpha_p \times (\Delta L_p)^2 + \beta_p \times \Delta L_p + \gamma_p \times \Delta W_p$ for gate p. The calculation of the change in total leakage power of the gates in the circuit is given by (2). Note that the parameters α_p, β_p, and γ_p are gate-specific, i.e., different values of the parameters are used for different types of gates. Similar to the computation of gate delay, the fitted parameters can also be used to compute the change in leakage power when only the dose on poly layer changes, in which case there is no dose change on active layer

[3] We recognize that off-path loading, slew propagation, and crosstalk timing windows can all change, and will be eventually accounted for precisely by golden signoff analysis. However, we assume in our optimization framework—as is fairly standard in the sizing literature—that these effects are negligible, and we validate our results with golden signoff analysis.

[4] We recognize that leakage power is exponential in gate length. We use a quadratic approximation to facilitate the problem formulation and solution method.

(i.e., ΔW_p is 0)

$$\begin{aligned}\Delta Leakage \ &= \ \textstyle\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{p\in r_{i,j}} \alpha_p \times D_s^2 \times d_{i,j}^P(p)^2 \\ &\quad +\beta_p \times D_s \times d_{i,j}^P(p) + \gamma_p \times D_s \times d_{i,j}^A(p).\end{aligned} \tag{2}$$

III. Problem Formulation of Dose Map Optimization for Improved Delay and Leakage

For simplicity, we do not include dose-dependent change of wire delay in our problem formulation; note that a dose map optimization on the poly and active layers will not affect wire layout patterns, and thus will not affect golden wire parasitics. In our implementation, wire delay is obtained from golden static timing analysis reports and added in between gates. Assume that the original dose in the chip area is uniform. The goal of the design-aware dose map optimization (*DMopt*) is to tune the dose maps on poly and active layers simultaneously to adjust the channel lengths and widths of the gates and thereby optimize circuit delay and/or total leakage power, subject to upper and lower bounds on delta dose values per grid, and a dose map smoothness bound to reflect the fact that exposure dose must change gradually between adjacent grids. In the following problem formulations, we use delta leakage instead of total leakage power to facilitate the computation. By minimizing/constraining delta leakage, i.e., the change in total leakage power, the total leakage power will be minimized/constrained. For computing delta leakage power, three fitted parameters (i.e., α_p, β_p, and γ_p) are needed as in (2). However, for computing the total leakage power, four fitted parameters are needed (i.e., a constant item is needed besides coefficients α_p, β_p, and γ_p) because we assume a quadratic relation between the change in leakage power and the change in doses on active and poly layers. Since delta leakage is enough for the following problem formulations, we use delta leakage rather than total leakage power to avoid the constant item in the estimation.

A. Design-Aware Dose Map Optimization on Poly Layer

The design-aware dose map optimization for poly layer can be formulated into two different problems, i.e., the quadratic program and the quadratic constraint program based on different types of constraints (i.e., either linear or quadratic).

1) *Dose Map Optimization for Improved Leakage Under Timing Constraint:* The optimization problem on poly layer is formulated as a quadratic program as follows.

- *Objective:* Minimize $\Delta Leakage$.
- *Subject to:*

$$L \le d_{i,j}^P \le U \quad \forall\ i \in [1, M] \quad j \in [1, N] \tag{3}$$

$$\begin{cases}|d_{i,j}^P - d_{i+1,j+1}^P| \le \delta\ \forall\ i \in [1, M-1] \quad j \in [1, N-1] \\ |d_{i,j}^P - d_{i,j+1}^P| \le \delta \quad \forall\ i \in [1, M] \quad j \in [1, N-1] \\ |d_{i,j}^P - d_{i+1,j}^P| \le \delta \quad \forall\ i \in [1, M-1] \quad j \in [1, N]\end{cases} \tag{4}$$

$$\begin{cases}a_q \le T & \forall\ q \in fanin(0) \\ a_r + t'_q \le a_q & \forall\ r \in fanin(q) \qquad (q = 1, \cdots, n) \\ 0 \le a_{n+1} \\ t'_p = t_p + A_p \times D_s \times d_{i,j}^P(p) & \qquad (p = 1, \cdots, n)\end{cases} \tag{5}$$

$$T \le \tau_L. \tag{6}$$

Equation (3) specifies the correction ranges on the dose for the poly layer, where L and U are user-specified or equipment-specific lower and upper bounds on the dose. Equation (4) specifies smoothness constraints on the dose for the poly layer, i.e., that the doses in neighboring rectangular grids should differ by a bounded amount.[5] Equation (5) specifies the delay constraint when the delays of the gates are scaled during the dose adjustment process. In (5), a_p represents the arrival time at node p, which is the maximum delay from source node $n+1$ to node p; $d_{i,j}^P(p)$ is the change in percentage of dose in rectangular grid $r_{i,j}$ on the poly layer in which gate p is located. The parameter A_p is gate-specific, and different values of the parameters are used for different types of gates as well as for gates of the same type that have different input slews and load capacitances. Equation (6) captures the user-specified upper bound (i.e., τ_L) on the delay of the longest path in the circuit. The calculation of the change in total leakage power of the gates $\Delta Leakage$ in the circuit is given by (2), where only poly layer related leakage [i.e., $\alpha_p \times D_s^2 \times d_{i,j}^P(p)^2 + \beta_p \times D_s \times d_{i,j}^P(p)$] is computed. Since the constraints are linear and the objective is quadratic, this gives an instance of quadratic program.

2) *Dose Map Optimization for Improved Timing Under Leakage Constraint:* The optimization problem on poly layer is formulated as a quadratically constrained program as follows.

- *Objective:* Minimize *T*.
- *Subject to:* Equations (3), (4), (5), and

$$\Delta Leakage \le \xi_L. \tag{7}$$

Equations (3), (4), and (5) are as discussed in the previous problem formulation. Equation (7) specifies the constraint on the change in the total leakage power of all cell instances, where ξ_L is a user-specified parameter for the constraint. Since the constraint in (7) is quadratic and the objective is linear, this yields an instance of quadratically constrained program.

B. Design-Aware Dose Map Optimization on Both Poly and Active Layers

1) *Dose Map Optimization for Improved Leakage Under Timing Constraint:* The optimization problem on both poly and active layers is formulated as a quadratic program as follows.

- *Objective:* Minimize $\Delta Leakage$.

[5] As stated in Section II-A, the dose generally varies gradually. To reflect the gradual property of dose profiles, the smoothness constraint is specified.

- *Subject to:* Equations (3), (4), and

$$L \leq d^A_{i,j} \leq U \quad \forall\, i \in [1, M] \quad j \in [1, N] \tag{8}$$

$$\begin{cases} |d^A_{i,j} - d^A_{i+1,j+1}| \leq \delta \;\forall\, i \in [1, M-1] & j \in [1, N-1] \\ |d^A_{i,j} - d^A_{i,j+1}| \leq \delta \quad \forall\, i \in [1, M] & j \in [1, N-1] \\ |d^A_{i,j} - d^A_{i+1,j}| \leq \delta \quad \forall\, i \in [1, M-1] & j \in [1, N] \end{cases} \tag{9}$$

$$\begin{cases} a_q \leq T & \forall\, q \in fanin(0) \\ a_r + t'_q \leq a_q & \forall\, r \in fanin(q) \qquad (q = 1, \cdots, n) \\ 0 \leq a_{n+1} & \\ t'_p = t_p + A_p \times D_s \times d^P_{i,j}(p) & \\ \qquad + B_p \times D_s \times d^A_{i,j}(p) & \qquad\qquad (p = 1, \cdots, n) \end{cases} \tag{10}$$

$$T \leq \tau_{WL}. \tag{11}$$

Similar to (4) for the poly layer, (8) specify the correction ranges on the dose for the active layer. Equation (9) specifies smoothness constraints on the dose for the active layer, and (10) specifies the delay constraint when the delays of the gates are scaled during the dose adjustment process on both poly and active layers. a_p and $d^P_{i,j}(p)$ are defined as in (5), and $d^A_{i,j}(p)$ is the change in percentage of dose in grid $r_{i,j}$ on the active layer wherein gate p is located. The parameter B_p is gate-specific, similar to A_p used in (5). Equation (11) specifies the constraint on the delay of the longest path in the circuit, where τ_{WL} is a user-specified parameter for the constraint. The calculation of the change in total leakage power of the gates $\Delta Leakage$ in the circuit is given by (2) which considers the impact of both gate length and gate width variations on leakage power.

2) *Dose Map Optimization for Improved Timing Under Leakage Constraint:* The optimization problem on both poly and active layers is formulated as a quadratically constrained program as follows.

- *Objective:* Minimize *T.*
- *Subject to:* Equations (3), (4), (8), (9), (10), and

$$\Delta Leakage \leq \xi_{WL}. \tag{12}$$

Equations (3), (4), (8), (9), and (10) are as discussed in previous problem formulations. Equation (12) specifies the constraint on the change in the total leakage power of all cell instances, where ξ_{WL} is a user-specified parameter for the constraint. Again, since the constraint in (12) is quadratic and the objective is linear, we have an instance of quadratically constrained program.

The above problem formulations[6] are either quadratic program or quadratically constrained program, which can be solved using classic quadratic programming methods. In particular, we use CPLEX [13] in the experimental platform described below.

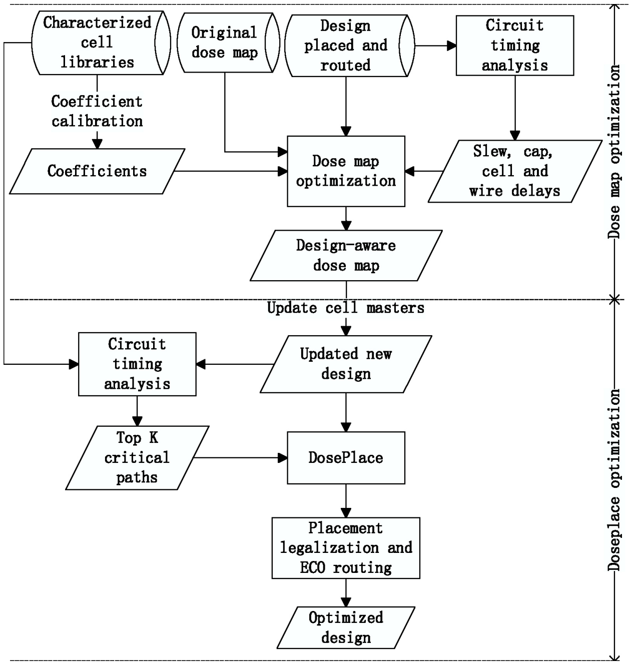


Fig. 7. Flow of the timing and leakage power optimization with integrated *DMopt* and *dosePl* (in the Appendix).

IV. Timing and Leakage Power Optimization Flow

A. Overall Optimization Flow

Fig. 7 shows our whole flow integrating *DMopt* together with *dosePl* (discussed in the Appendix) for timing and leakage optimization. Note that the timing and leakage optimization flow is carried out after V_{th} and V_{dd} assignment processes. For the timing and leakage related dose map optimization problem, the input consists of: (1) the original dose maps (i.e., those calculated to minimize ACLV and AWLV metrics, based on in-line metrology) for both poly and active layers; (2) the characterized standard-cell timing libraries (or, other timing models that comprehend the impact of dose on transistor gate lengths and widths) for different gate lengths and gate widths; and (3) the circuit with placement and routing information. By "placement and routing information," we also include implicit information that is necessary for timing and power analyses, e.g., extracted wiring parasitics. With the nominal gate-length cell timing and power libraries, and the circuit itself with its placement, routing and parasitic data, timing analysis can be performed to generate the input slews and output load capacitances of all the cell instances. With the input slews and output load capacitances of all the cell instances, the original dose maps, and characterized cell libraries of different gate lengths and gate widths, the dose map optimization is executed to determine doses that adjust gate lengths and gate widths of the cells for timing and leakage optimization, subject to dose map constraints. Finally, the optimized design-aware dose maps on both layers are generated.

According to the optimized design-aware dose maps on both poly and active layers, the cell instances in different grids of the dose maps will have different gate lengths and widths as well as different cell masters in the characterized

[6] The optimization result is feasible for the equipment, as a consequence of the constraints (3), (4), (8), and (9).

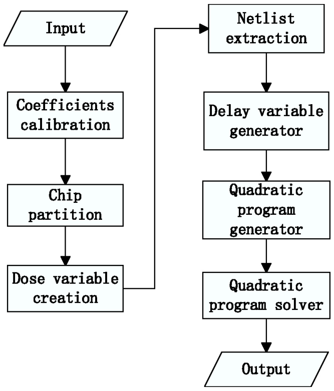


Fig. 8. Detailed view of design-aware dose map optimization flow.

cell libraries.[7] Thus, the design's netlist representation must be updated according to the dose maps. Using the characterized cell libraries, timing analysis is performed on the new design with the updated cell masters to identify the top-K (e.g., $K = 10,000$) critical paths for the complementary *dosePl* (see the Appendix) process to optimize. The *dosePl* process is based on a cell swapping strategy, which may introduce an illegal placement result. Therefore, a legalization process is invoked to legalize the swapped cells. ECO routing is then executed for the affected wires to refine the design with optimized timing yield.

B. Summary of the Dose Map Optimization Flow

The dose map optimization in Fig. 8 is summarized as follows. The input consists of the original dose maps on both layers, the characterized cell libraries of different gate lengths and widths, and the input slews and output capacitances of all the cells in the circuit. From the characterized 65 nm cell libraries of different gate lengths and widths and 90 nm cell libraries of different gate lengths,[8] the coefficients in the linear function of delay and quadratic function of leakage power are calibrated. Note that when gate delay calculation in the cell libraries adopts a lookup table method, where the entries are indexed by input slews and output capacitances, the coefficients of the delay functions may be calibrated for each entry in each delay table. Then, according to the input slew and output capacitance values that were obtained for each cell in the previous step, the coefficients associated with the nearest entry (or, entries with interpolation) in the table will be applied to calculate the delay of the cell.

The exposure fields on both poly and active layers are then partitioned into rectangular grids. For each grid on poly (active) layer, a variable $d^P_{i,j}$ ($d^A_{i,j}$) represents the amount of dose change in the grid. Maximum circuit delay is captured using variable a_p that represents the arrival time at the output of cell p. When all the variables are obtained, a quadratic program (resp. quadratically constrained program) problem instance is generated by introducing the dose map correction range constraints, dose map smoothness constraints, and the delay constraints, as well as the objective of minimizing the total leakage power of all the cells under timing constraint (resp. minimizing the timing of the circuit under leakage constraint). Finally, a quadratic program (resp. quadratically constrained program) solver finds the optimal dose change in each grid based on the original dose maps; this yields our optimal design-aware dose maps.

TABLE I

Characteristics of Artisan TSMC 65 nm and 90 nm Designs

Design	Chip Size (mm^2)	#Cell Instances	#Nets
AES-65	0.058	16 187	16 450
JPEG-65	0.268	68 286	68 311
AES-90	0.25	21 944	22 581
JPEG-90	1.09	98 555	105 955

V. Experimental Results

To assess the effectiveness of the proposed dose map optimization algorithms, we first sweep the dose change on poly layer from −5% to +5% for all the cell instances in the 65 nm design AES-65 and the 90 nm design AES-90 (shown in Table I) and perform timing analysis using *Synopsys PrimeTime* (version Z-2006.12) [14] and leakage power estimation using *Cadence SoC Encounter* (v07.10) [15]. The timing analysis and leakage power estimation are based on pre-characterized 65 nm and 90 nm cell libraries with gate length and gate width variants. Delay and leakage power results are given in Tables II and III, where "MCT" refers to minimum cycle time and "Leakage" refers to the total leakage power of all the cells. From the tables, in the extreme cases the dose changes on poly layer correspond to the maximum timing yield improvement or leakage power reduction (i.e., $d^P_{i,j} = +5$ and $d^P_{i,j} = -5$). The results show that timing yield improvement can be obtained at the cost of leakage power increase, whereas leakage power reduction can be obtained at the cost of timing yield degradation. Uniform dose change in all the cell instances cannot obtain timing yield improvement without leakage power increase. However, our proposed dose map optimization algorithms can obtain substantial timing yield improvement without increase in total leakage power, as well as leakage power reduction without degradation in timing yield.

The timing and leakage optimization flow is implemented in C++ and tested on industrial testcases as given in Table I. In Table I, there are two different classes of testcases. AES-65 and JPEG-65 are 65 nm designs, and AES-90 and JPEG-90 are 90 nm designs. In the experiments, the dose sensitivity D_s is −2 nm/%. The parameters A_p, B_p, α_p, β_p, and γ_p are calibrated using PrimeTime and SoC Encounter based on the pre-characterized cell timing and leakage libraries. Since different libraries (i.e., 90 nm and 65 nm) are used for different designs, two sets of parameters are calibrated from

[7] When the gate lengths and widths are computed from the optimized dose maps, it is possible that the computed values do not exactly match the available drive strengths of the cell masters in the characterized cell libraries. Thus, a rounding step is needed to snap the computed gate lengths and widths to the cell masters with nearest drive strengths.

[8] We focus on the dose map optimization methods for 65 nm testcases. However, 90 nm testcases are also used in dose map optimization for gate length modulation to provide extra supporting experimental data.

TABLE II

DELAY AND LEAKAGE VALUES OF 65 NM DESIGN AES-65 WHEN DOSE CHANGE $d^P_{i,j}$ IS SWEPT FROM 0% TO −5% AND FROM 0% TO +5% ON POLY LAYER

Dose change	$d^P_{i,j} = 0$	$d^P_{i,j} = -0.5$	$d^P_{i,j} = -1$	$d^P_{i,j} = -1.5$	$d^P_{i,j} = -2$	$d^P_{i,j} = -2.5$	$d^P_{i,j} = -3$	$d^P_{i,j} = -3.5$	$d^P_{i,j} = -4$	$d^P_{i,j} = -4.5$	$d^P_{i,j} = -5$
MCT (ns)	1.638	1.658	1.677	1.695	1.715	1.730	1.750	1.766	1.786	1.806	1.824
imp. (%)	–	−1.22	−2.38	−3.48	−4.70	−5.62	−6.84	−7.81	−9.04	−10.26	−11.36
Leakage (μW)	448.0	421.0	397.7	375.9	356.9	339.7	324.8	311.7	299.9	289.3	279.6
imp. (%)	–	6.03	11.23	16.09	20.33	24.17	27.50	30.42	33.06	35.42	37.59
Dose change	$d^P_{i,j} = 0$	$d^P_{i,j} = +0.5$	$d^P_{i,j} = +1$	$d^P_{i,j} = +1.5$	$d^P_{i,j} = +2$	$d^P_{i,j} = +2.5$	$d^P_{i,j} = +3$	$d^P_{i,j} = +3.5$	$d^P_{i,j} = +4$	$d^P_{i,j} = +4.5$	$d^P_{i,j} = +5$
MCT (ns)	1.638	1.622	1.601	1.578	1.557	1.537	1.517	1.497	1.474	1.452	1.427
imp. (%)	–	0.98	2.26	3.66	4.95	6.17	7.39	8.61	10.01	11.36	12.88
Leakage (μW)	448.0	478.0	513.4	552.8	600.4	655.2	722.2	800.1	893.5	1008.6	1142.2
imp. (%)	–	−6.70	−14.60	−23.39	−34.02	−46.25	−61.21	−78.59	−99.44	−125.13	−154.96

Straightforward way of increasing dose cannot obtain delay improvement without incurring leakage increase.

TABLE III

DELAY AND LEAKAGE VALUES OF 90 NM DESIGN AES-90 WHEN DOSE CHANGE $d^P_{i,j}$ IS SWEPT FROM 0% TO −5% AND FROM 0% TO +5% ON POLY LAYER

Dose change	$d^P_{i,j} = 0$	$d^P_{i,j} = -0.5$	$d^P_{i,j} = -1$	$d^P_{i,j} = -1.5$	$d^P_{i,j} = -2$	$d^P_{i,j} = -2.5$	$d^P_{i,j} = -3$	$d^P_{i,j} = -3.5$	$d^P_{i,j} = -4$	$d^P_{i,j} = -4.5$	$d^P_{i,j} = -5$
MCT (ns)	1.990	2.011	2.031	2.057	2.078	2.093	2.115	2.135	2.155	2.172	2.188
imp. (%)	–	−1.031	−2.076	−3.359	−4.401	−5.155	−6.296	−7.257	−8.283	−9.142	−9.949
P_{leakage} (μW)	2430.214	2324.525	2225.130	2135.234	2054.458	1980.457	1914.474	1850.809	1796.545	1746.507	1699.788
imp. (%)	–	4.349	8.439	12.138	15.462	18.507	21.222	23.842	26.075	28.134	30.056
Dose change	$d^P_{i,j} = 0$	$d^P_{i,j} = +0.5$	$d^P_{i,j} = +1$	$d^P_{i,j} = +1.5$	$d^P_{i,j} = +2$	$d^P_{i,j} = +2.5$	$d^P_{i,j} = +3$	$d^P_{i,j} = +3.5$	$d^P_{i,j} = +4$	$d^P_{i,j} = +4.5$	$d^P_{i,j} = +5$
MCT (ns)	1.990	1.971	1.950	1.932	1.905	1.893	1.868	1.845	1.818	1.791	1.758
imp. (%)	–	0.964	2.029	2.915	4.257	4.906	6.161	7.302	8.652	10.012	11.661
P_{leakage} (μW)	2430.214	2546.756	2678.096	2824.598	2994.978	3180.969	3404.057	3654.222	3939.749	4253.778	4619.039
imp. (%)	–	−4.796	−10.200	−16.228	−23.239	−30.893	−40.072	−50.366	−62.115	−75.037	−90.067

Straightforward way of increasing dose cannot obtain delay improvement without incurring leakage increase.

TABLE IV

RESULTS OF DOSE MAP OPTIMIZATION ON POLY LAYER, I.E., GATE LENGTH (L_{gate}) MODULATION WITH SMOOTHNESS BOUND $\delta = 2$ AND DOSE CORRECTION RANGE ±5%

AES-65	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				30 × 30 μm^2 grids			
	L_{gate}	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)
MCT (ns)	1.638	1.631	0.44	1.607	1.89	1.632	0.35	1.626	0.71	1.637	0.07	1.637	0.07
Leakage (μW)	448.0	409.7	8.54	441.3	1.49	434.3	3.05	445.4	0.57	447.9	0.01	447.1	0.19
Runtime (s)	–	72	–	108	–	18	–	335	–	9	–	46	–
JPEG-65	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				30 × 30 μm^2 grids			
	L_{gate}	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)
MCT (ns)	2.179	2.174	0.25	2.081	4.52	2.178	0.04	2.102	3.54	2.172	0.31	2.159	0.91
Leakage (μW)	2915.5	2312.7	20.67	2922.3	−0.23	2480.9	14.91	2913.4	0.07	2843.1	2.48	2909.8	0.19
Runtime (s)	–	490	–	891	–	292	–	558	–	61	–	929	–
AES-90	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				50 × 50 μm^2 grids			
	L_{gate}	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)
MCT (ns)	1.990	1.975	0.75	1.861	6.47	1.981	0.44	1.872	5.91	1.989	0.05	1.927	3.19
Leakage (μW)	2430.2	1823.2	24.98	2386.1	1.82	1901.6	21.75	2370.2	2.47	2172.4	10.61	2406.0	1.00
Runtime (s)	–	176	–	227	–	85	–	145	–	16	–	92	–
JPEG-90	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				50 × 50 μm^2 grids			
	L_{gate}	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)	*QP*	imp. (%)	*QCP*	imp. (%)
MCT (ns)	2.906	2.894	0.41	2.667	8.23	2.901	0.16	2.689	7.45	2.887	0.65	2.757	5.11
Leakage (μW)	4354.2	3422.5	21.40	4244.4	2.52	3453.6	20.68	4273.5	1.85	3822.2	12.22	4308.3	1.06
Runtime (s)	–	2157	–	3644	–	1194	–	2068	–	243	–	2545	–

TABLE V

RESULTS OF DOSE MAP OPTIMIZATION ON BOTH POLY AND ACTIVE LAYERS USING QUADRATICALLY CONSTRAINED PROGRAM FOR IMPROVED TIMING, I.E., GATE LENGTH (L_{gate}) AND GATE WIDTH (W_{gate}) MODULATION, WITH SMOOTHNESS BOUND $\delta = 2$ AND DOSE CORRECTION RANGE ±5%

AES-65	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				30 × 30 μm^2 grids			
	L_{gate}&W_{gate}	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)
MCT (ns)	1.638	1.601	1.89	1.586	3.17	1.626	0.10	1.610	1.71	1.647	0.07	1.630	0.48
Leakage (μW)	448.0	441.3	1.49	447.0	0.22	445.4	0.57	447.7	0.06	447.9	0.01	446.5	0.32
Runtime (s)	–	108	–	179	–	335	–	548	–	46	–	141	–
JPEG-65	Nom	5 × 5 μm^2 grids				10 × 10 μm^2 grids				30 × 30 μm^2 grids			
	L_{gate}&W_{gate}	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)
MCT (ns)	2.179	2.081	4.52	2.090	4.10	2.102	3.54	2.093	3.93	2.159	0.91	2.153	1.21
Leakage (μW)	2915.5	2922.3	−0.23	2922.0	−0.22	2913.4	0.07	2915.5	0.00	2909.8	0.19	2907.9	0.26
Runtime (s)	–	891	–	1561	–	558	–	747	–	929	–	3184	–

TABLE VI

RESULTS OF DOSE MAP OPTIMIZATION ON BOTH POLY AND ACTIVE LAYERS USING QUADRATIC PROGRAM FOR IMPROVED LEAKAGE POWER, I.E., GATE LENGTH (L_{gate}) AND GATE WIDTH (W_{gate}) MODULATION, WITH SMOOTHNESS BOUND δ = 2 AND DOSE CORRECTION RANGE ±5%

AES-65	Nom	$5 \times 5\,\mu\text{m}^2$ grids				$10 \times 10\,\mu\text{m}^2$ grids				$30 \times 30\,\mu\text{m}^2$ grids			
	L_{gate}&W_{gate}	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)
MCT (ns)	1.638	1.631	0.44	1.635	0.18	1.632	0.35	1.636	0.10	1.637	0.07	1.631	0.45
Leakage (μW)	448.0	409.7	8.54	383.8	14.33	434.3	3.05	409.2	8.64	447.9	0.01	444.9.0	0.69
Runtime (s)	–	72	–	110	–	18	–	44	–	9	–	13	–
JPEG-65	Nom	$5 \times 5\,\mu\text{m}^2$ grids				$10 \times 10\,\mu\text{m}^2$ grids				$30 \times 30\,\mu\text{m}^2$ grids			
	L_{gate}&W_{gate}	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)	L_{gate}	imp. (%)	Both	imp. (%)
MCT (ns)	2.179	2.174	0.25	2.177	0.09	2.178	0.04	2.177	0.11	2.172	0.31	2.179	0.01
Leakage (μW)	2915.5	2312.7	20.67	2301.1	21.07	2480.8	14.91	2434.0	16.52	2843.1	2.48	2763.2	5.22
Runtime (s)	–	490	–	1232	–	292	–	531	–	61	–	93	–

the different libraries and used in the dose map optimization for the corresponding testcases. Table IV shows the dose map optimization results on poly layer. In Table IV, QP refers to the quadratic program for improved total leakage under timing constraint, and QCP refers to the quadratically constrained program for improved timing under leakage constraint. Different sizes of rectangular grids are used in the dose map optimization, i.e., $5 \times 5\,\mu\text{m}^2$, $10 \times 10\,\mu\text{m}^2$, and either $30 \times 30\,\mu\text{m}^2$ (for 65 nm cases) or $50 \times 50\,\mu\text{m}^2$ (for 90 nm cases). The dose smoothness bound is $\delta = 2$,[9] and the dose correction range is ±5%. From the results, the finer the rectangular grids, the greater the improvement in the timing of the circuit or in the total leakage power. We observe different optimization quality between 90 nm testcases (AES-90 and JPEG-90) and 65 nm testcases (AES-65 and JPEG-65). Average leakage reduction for 90 nm testcases under timing constraints for $5 \times 5\,\mu\text{m}^2$ grids is 23.2% but that of 65 nm testcases shows 14.6%. Average MCT reduction for 90 nm testcases under leakage constraints for $5 \times 5\,\mu\text{m}^2$ grids is more than 7.4%, but that of 65 nm testcases shows 3.4%. There are two reasons for the above optimization discrepancy between 90 nm and 65 nm designs. The first reason is that $5 \times 5\,\mu\text{m}^2$ grids have different granularities for the different designs. From Table I, the average number of cell instances in a grid of $5 \times 5\,\mu\text{m}^2$ is 2.2 for the 90 nm testcases and 6.3 for the 65 nm testcases. As discussed above, the finer the rectangular grids, i.e., the fewer cell instances in one grid, the better the optimization quality. The smaller average number of cell instances per grid for the 90 nm testcases permits larger improvements. The second reason is the difference in timing criticality (slack distribution) of the testcases before optimization. Table VII shows the timing criticality of each testcase as the number of critical paths within a specific range of timing. More paths in the 65 nm testcases have delay values near the MCT, which makes it difficult for the dose map optimization to remove all those paths to improve timing. However, in the 90 nm testcases, the number of such critical paths is small, making it easier for the dose map optimization to improve timing. For these reasons, more substantial leakage and timing improvements are observed for the 90 nm testcases.

Table V shows the dose map optimization results using the quadratically constrained program for improved timing on both poly and active layers for 65 nm designs. From the results, slightly better timing improvement is obtained using simultaneous modulation in both gate length and gate width. Table VI shows the dose map optimization results using the quadratic program for improved leakage power on both poly and active layers. Again, only the 65 nm designs are tested. From the results, slightly better leakage improvement is obtained using simultaneous modulation of gate length and gate width than only using gate length modulation. The maximum change in gate width is 10 nm according to the dose sensitivity −2 nm/% and dose correction range ±5%, which is relatively small compared with the transistor widths of cells in the 65 nm standard cell library (the minimum transistor width in 65 nm cells is around 200 nm, while the maximum width is more than 650 nm). As a result, there is only slight impact of gate width modulation on the cell's delay and leakage, and the related timing and/or leakage improvements are not significant.

TABLE VII

PERCENTAGE OF CRITICAL TIMING PATHS IN TESTCASES

Design	95–100% MCT (%)	90–100% MCT (%)	80–100% MCT (%)
AES-65	16.54	28.98	41.98
JPEG-65	4.80	9.89	30.23
AES-90	0.91	4.54	22.84
JPEG-90	0.12	0.35	3.92

In one case (JPEG-65 with $5 \times 5\,\mu\text{m}^2$ grids in Table V), the dose map optimization using simultaneous gate width and gate length modulation obtains slightly worse results than using only gate length modulation. We attribute this to the use of more fitted parameters (i.e., B_p and γ_p for gate width related delay and leakage) in estimation of cell delay and leakage, which can introduce more estimation errors. From the Liberty delay model tables of 36 different 65 nm standard cell masters, for all the arcs (i.e., rise and fall) with all the slew/load combinations, we perform curve fitting for cell delay versus gate length using the least square method. When only gate length changes, 21 different characterized libraries are needed corresponding to the 21 different dose values for poly layer in Table II. In this case, the maximum sum of squares of the residuals for all the fitted curves is 0.0005. When both gate length and gate width change, total 441 (i.e., 21 × 21) characterized libraries are needed, which is a combination of 21 different dose values for the poly layer (i.e., the change in gate length) and 21 different dose values for the active layer

[9] Different smoothness bounds in different directions, i.e., slit and scan directions (see Section II), may be specified, respectively. Here, we use an average example value for both directions.

(i.e., the change in gate width). In this case, the maximum sum of squares of the residuals for all the fitted curves is 0.0101, which is much larger than 0.0005. The increased error in curve fitting may be caused by the increased number of variables (i.e., gate width) and the increased number of characterized libraries.

From the results in Tables II and III, we recognize that smaller dose change results in smaller timing improvement, e.g., Table II shows $d_{i,j}^P = +1$ corresponding to 2.26% timing improvement versus $d_{i,j}^P = +5$ corresponding to 12.88% improvement. Therefore, tighter smoothness bounds (i.e., $\delta < 2$) will result in smaller timing improvement by enforcing smaller available dose changes within each rectangular grid. By testing different sizes of the rectangular grids, the smoothness bounds are also elaborated, i.e., the effective smoothness bound of a given smoothness value is different for different rectangular grids. For example, the effective smoothness bound of smoothness value $\delta = 2$ over $50 \times 50\,\mu\text{m}^2$ grids (i.e., 2%/50 μm) is tighter than that over $10 \times 10\,\mu\text{m}^2$ grids (i.e., 2%/10 μm). As mentioned in the Introduction, the Zeiss/Pixer CDC technology also enables adaptivity in the manufacturing flow to meet the required CD specifications. We note that our methods can be used for any emerging technology that enables the fine-grain tuning of CD (i.e., along with relaxed effective smoothness bound) during manufacturing. Moreover, the sizes of our testcases (Table I) are very small, with the largest area (90 nm JPEG-90) being only a little over 1mm^2. For designs of larger sizes, we anticipate that our methods will obtain better timing and leakage improvements.

VI. Conclusion

We have proposed a novel method to improve the timing yield of the circuit as well as reduce total leakage power, using design-aware dose map and dose map-aware placement optimization. We focus mainly on the placement-aware dose map optimization. The dose map-aware placement optimization is also attempted on a placement-aware timing and leakage optimized dose map. The proposed method is based on the fact that the exposure dose in the exposure field can change the gate/transistor lengths and widths of the cells in the circuit, which is useful for optimization of gate delay and gate leakage power. Our ongoing work includes extension of the dose map optimization methodology to minimize the delay variation of different chips across the wafer or the exposure field.

Appendix
Dose Map-Aware Placement

A. Cell-Swapping Based Optimization

After a placement-specific dose map has been calculated, it is natural to ask whether a dose map-specific placement can further improve the result. In this Appendix, we describe a simple cell swapping-based dose map-aware placement (*dosePl*) optimization. The *dosePl* problem can be stated as follows. Given the original placement result and a timing and leakage-aware dose map, determine cell pairs to swap for timing yield improvement.

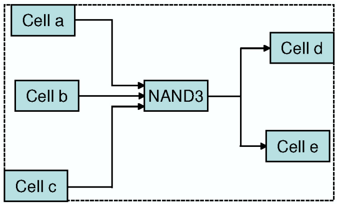


Fig. 9. Bounding box of a 3-input NAND3 cell: moving the cell within its bounding box has lower likelihood of increasing total wire length.

The basic idea behind the cell swapping-based optimization method is to swap cells on timing-critical paths (referred to as *critical cells* hereafter) to high-dose regions and non-critical cells to low-dose regions, to further enhance the circuit performance under leakage constraint.

We define the *bounding box of a cell* as the bounding box of all the cell's fanin cells and all of its fanout cells, as well as the cell itself. Fig. 9 shows the bounding box of a (NAND3) cell, denoted by the dashed line. Our intuition is that moving a cell within its bounding box has a lower likelihood of increasing total wire length or timing delay than moving it outside the bounding box. Thus, we seek pairs of cells $cell_l$ with bounding box b_l and $cell_m$ with bounding box b_m in different dose regions, such that $cell_l$ is in b_m and $cell_m$ is in b_l. With this restriction, we filter out candidate cell swaps that are too disruptive to wirelength and timing.

(1) *Additional Heuristics to Avoid Wirelength Increase:* When two cells satisfy the condition that they are located in each other's bounding boxes, it is still possible for total wirelength to increase. We thus adopt the following heuristics to further filter out unpromising cell pairs.

- Distance between the two cells to be swapped When the distance between two cells is very large, the impact of cell swapping on total wirelength is potentially large. Therefore, we avoid considering swaps of cells that are far apart.[10]
- Half-perimeter Wire Length (HPWL)-based wire length comparison We may also filter cell swaps by computing updated HPWL-based wirelength estimates; only if the estimated wirelength increase for all incident nets (e.g., the four nets incident to the "NAND" cell in Fig. 9) is below a predefined threshold (e.g., 20% in our experiments reported below) will the cell swap be attempted.

(2) *On the Number of Swaps and Cell Priority:* For a given critical path, several cell swaps may suffice to reduce the path delay, and further cell swapping will introduce unnecessary wirelength and leakage increase. So, an upper bound on the number of cells swapped for each critical path is specified in our heuristic's implementation (e.g., one cell per critical path in the experiments below). The priority for a critical cell during swapping is decided according to the following two factors.

- Number of critical paths that pass through the cell The more critical paths that pass through a given cell, the more

[10] In the experimental results below, this threshold is chosen proportionally to the gate pitch, which is computed as the chip dimension divided by the square root of gate count in the chip.

beneficial it is to swap the cell to a higher-dose region. Higher priorities are assigned to cells that are on a greater number of critical paths.

- Slack of critical paths The larger the total path delay (= smaller slack) of a given critical path, the more important it is to swap cells on the path to achieve cell delay improvement. Therefore, higher priority is assigned to cells on paths with greater timing criticality.

  Based on the above two heuristic factors, critical cells are assigned weights as calculated in (13) where C_l is the set of critical paths on which $cell_l$ is located. In our implementation, cells are processed path by path (obtained from golden timing analysis), in order from most timing-critical to least critical. Therefore, cells on more-critical paths always have higher priorities than cells on less-critical paths. Cells in the same critical path are sorted in non-increasing order of weights that are computed as

$$W(cell_l) = \sum_{cell_l \in C_l} e^{-slack(C_l)}. \qquad (13)$$

(3) *On the Leakage Power Increase:* When $cell_l$ and $cell_m$ are to be swapped, the increase in their combined leakage power $\Delta Leak_{l,m}$ may be estimated beforehand. If $\Delta Leak_{l,m}$ is less than a given fraction γ_4 (e.g., 10%) of the original leakage power $Leak_{l,m}$ of the two cells, they will be swapped. Otherwise, no swapping will be performed, so as to avoid large leakage increase. Because one cell is swapped to a higher dose region (i.e., leakage increases) and the other one is swapped to a lower dose region (i.e., leakage decreases), it is not always the case that cell swapping will result in leakage power increase.

(4) *Pseudocode of the Cell Swapping Heuristic:* The pseudocode of one round of our cell swapping heuristic is given as Algorithm 1. In each round of cell swapping, a maximum of γ_5 swaps are allowed (e.g., one swap for each round of cell swapping in our experiments). The cell swapping process is based on the critical paths, which are first sorted in non-increasing order according to their slacks. Cells of a given path are then swapped. Since it is not necessary to swap all the cells in a critical path to improve its timing, the swapping process for a path is terminated when the number of swapped cells reaches a user-defined parameter γ_1 (in our experiments, up to one cell is swapped on each path). For a given candidate swapping pair, the swapping process checks the bounding box constraint, the dose constraint, and distance, then computes HPWL-based wirelength increase and leakage increase when the pair is swapped. If a candidate pair passes all the checks, its cells are swapped and we update the number of swapped cells for the corresponding critical paths. The cell swapping process continues until all critical paths are processed or the number of swaps reaches γ_5. When one round of the swapping process finishes, the perturbed placement is legalized and routed by an engineering change order (ECO) placement and routing process. After final (ECO) routing, golden timing analysis is performed with updated parasitics to evaluate the circuit delay improvement. If the circuit delay is improved, the swapping is accepted. Otherwise, the swapped cell instances are rolled back to their previous cell masters and another round of cell swapping is performed with those swapped cells marked as fixed (i.e., those cells cannot be swapped again in the following cell swapping process). The total number of rounds of cell swapping in our experiments is 10. A larger number of swapping rounds can obtain better timing improvement. However, the improvement cannot be guaranteed because only swapping cells on minimum-slack critical paths can improve timing yield, but those cells may not be swappable due to the swapping constraints, e.g., distance, leakage increase, etc.

Algorithm 1 *dosePl*: one round of cell swapping heuristic for timing yield improvement.

```
Find cells in top $K$ critical paths by golden timing analysis;
Compute weights for critical cells as in (13);
Sort critical paths in non-decreasing order according to their slacks;
Set $numSwaps \leftarrow 0$;
for $k = 1$ to $K$ do
    Sort the cells in critical path $c_k$ in non-increasing order according to their weights;
    for all cell $cell_l \in$ critical path $c_k$ do
        if # swapped cells in path $c_k$ $n(c_k) > \gamma_1$ then break; end if
        Compute bounding box $b_l$ of cell $cell_l$ in path $c_k$;
        Compute the set $R$ of rectangular grids that intersect with $b_l$;
        Sort the grids $r \in R$ in non-increasing order according to their doses $d(r)$;
        Set $flag \leftarrow$ false;
        for all $r \in R$ do
            if $d(r) < d(cell_l)$ then break; end if // $d(cell_l)$ is the dose on $cell_l$
            Sort the non-critical cells $NC$ in grid $r$ in non-decreasing order by
            Manhattan distance from $cell_l$;
            for all $cell_m \in NC$ do
                if $dist(cell_l, cell_m) > \gamma_2$ then break; end if
                if $cell_l \in b_m$ and $cell_m \in b_l$ and $\Delta HPWL(cell_l) < \gamma_3$ and
                $\Delta HPWL(cell_m) < \gamma_3$ and $\Delta Leak_{l,m} < Leak_{l,m} \cdot \gamma_4$ then
                    Swap $(cell_l, cell_m)$;
                    Update the number of swapped cells $n(c_s)$ for all critical paths $c_s$
                    such that $cell_l \in c_s$;
                    Set $flag \leftarrow$ true;
                    $numSwaps$ ++;
                    if $numSwaps \geq \gamma_5$ then return; end if
                    break;
                end if
            end for
            if $flag$ = true then break; end if
        end for
    end for
end for
```

TABLE VIII

EXPERIMENTAL RESULTS OF DOSE MAP OPTIMIZATION ON POLY LAYER USING QUADRATICALLY CONSTRAINED PROGRAM FOR IMPROVED TIMING, FOLLOWED BY INCREMENTAL PLACEMENT PROCESS

Testcase	AES-65			JPEG-65		
	Nom L_{gate}	*QCP*	*dosePl*	Nom L_{gate}	*QCP*	*dosePl*
MCT (ns)	1.638	1.607	1.601	2.179	2.081	1.847
Leakage (μW)	448.0	441.4	441.4	2915.5	2922.3	2922.3
Runtime (s)	–	108	40	–	924	149

The chip is partitioned into rectangular grids of size 5 × 5 μm^2, the dose correction range is ±5%, and dose smoothness bound is $\delta = 2$.

B. Experimental Results

The experimental results of dose map-placement co-optimization are given in Table VIII. From the results, *DMopt* (*QCP*) first improves the timing yield under leakage power constraint. Cell-swapping based *dosePl* further improves the results.

Fig. 10 shows four slack profiles of design AES-65, including: (1) the original design (*Orig*); (2) the design (*DMOpt*) after dose map optimization on poly layer for improved timing

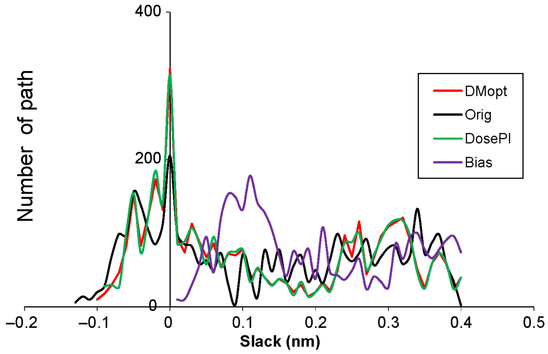


Fig. 10. Slack profiles of design AES-65 before *DMopt*, after *DMopt*, after *dosePl*, and the biased design when all the gates in the top 10 000 critical paths receive maximum possible exposure dose (+5%).

(dose correction range is ±5%, smoothness bound is $\delta = 2$, and rectangular grids are of sizes $5 \times 5\,\mu\text{m}^2$); (3) the design (*dosePl*) after placement optimization, and (4) the design (*Bias*) when all the gates in the top 10000 critical paths are enforced using maximum possible dose (i.e., +5% on the original dose). The purpose of enforcing the maximum possible exposure dose on the critical gates is to find out the optimization headroom left after *DMopt* process. From Fig. 10, the worst slack of the original design is first improved by dose map optimization process, and then further improved by placement optimization process. However, it is difficult for the dose map and placement optimization to improve the slacks for all the paths on the "hill" around the critical slack value of 0 ns. Besides, as shown by Table II, though there is seeming headroom left between the optimized design and the "optimal" (*Bias*) design, it is impossible to reach the "optimal" design without dramatically increasing the total leakage power.[11]

References

[1] K. Jeong, A. B. Kahng, C.-H. Park, and H. Yao, "Dose map and placement co-optimization for timing yield enhancement and leakage power reduction," in *Proc. IEEE/Assoc. Comput. Mach. Design Autom. Conf.*, 2008, pp. 516–521.

[2] P. Gupta and A. B. Kahng, "Manufacturing-aware physical design," in *Proc. IEEE/Assoc. Comput. Mach. Int. Conf. Comput.-Aided Design*, 2003, pp. 681–687.

[3] S. Bhardwaj, Y. Cao, and S. Vrudhula, "Statistical leakage minimization through joint selection of gate sizes, gate lengths and threshold voltage," in *Proc. Asia South Pacific Design Autom. Conf.*, 2006, pp. 953–958.

[4] P. Gupta, A. B. Kahng, P. Sharma, and D. Sylvester, "Gate-length biasing for runtime-leakage control," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 8. 2006, pp. 1475–1485.

[5] [Online]. Available: http://wps2a.semi.org/cms/groups/public/documents/membersonly/van_schoot_presentation.pdf

[6] I. Pollentier, S. Y. Cheng, B. Baudemprez, *et al.*, "In-line lithography cluster monitoring and control using integrated scatterometry," in *Proc. Int. Soc. Opt. Engineers Data Anal. Modeling Process Control*, vol. 5378. 2004, pp. 105–115.

[7] J. B. van Schoot, O. Noordman, P. Vanoppen, *et al.*, "CD uniformity improvement by active scanner corrections," in *Proc. Int. Soc. Opt. Engineers Symp. Opt. Microlithography*, vol. 4691. 2002, pp. 304–314.

[8] G. Zhang, M. Terry, S. O'Brien, *et al.*, "65nm node gate pattern using attenuated phase shift mask with off-axis illumination and sub-resolution assist features," in *Proc. Int. Soc. Opt. Engineers Symp. Opt. Microlithography*, vol. 5754. 2005, pp. 760–772.

[9] N. Jeewakhan, N. Shamma, S.-J. Choi, *et al.*, "Application of dosemapper for 65-nm gate CD control: Strategies and results," in *Proc. Int. Soc. Opt. Engineers Symp. Photomask Technol.*, vol. 6349. 2006, pp. 63490G-1–63490G-11.

[10] R. Seltmann, R. Stephan, M. Mazur, *et al.*, "ACLV-analysis in production and its impact on product performance," in *Proc. Int. Soc. Opt. Engineers Symp. Opt. Microlithography*, vol. 5040. 2003, pp. 530–540.

[11] [Online]. Available: http://www.asml.com

[12] C.-P. Chen, C. C. N. Chu, and D. F. Wong, "Fast and exact simultaneous gate and wire sizing by Lagrangian relaxation," *IEEE Trans. Comput.-Aided Design Integr. Circuits System.*, vol. 18, no. 7. 1999, pp. 1014–1025.

[13] *ILOG CPLEX* [Online]. Available: http://www.ilog.com/products/cplex

[14] *Synopsys PrimeTime* [Online]. Available: http://www.synopsys.com

[15] *Cadence SOC Encounter* [Online]. Available: http://www.cadence.com

[16] [Online]. Available: http://www.smt.zeiss.com

[17] A. B. Kahng, "Key directions and a roadmap for electrical design for manufacturability," in *Proc. Eur. Solid-State Circuits Conf.*, 2007, pp. 83–88.

Kwangok Jeong (S'07) received the B.S. and M.S. degrees in electrical engineering from Hanyang University, Seoul, Korea, in 1997 and 1999, respectively. Since 2006, he has been pursuing the Ph.D. degree from the VLSI CAD Laboratory, University of California, San Diego.

He was with the Computer-Aided Engineering Team, Samsung Electronics, Seoul, Korea, from 1999 to 2006. His current research interests include physical design and very large scale integration design-manufacturing interface.

Mr. Jeong received three Best Paper Awards at the Samsung Technical Conference, in 2001, 2002, and 2004, respectively, during his work with Samsung Electronics, and the Honorable Outstanding Researcher Award, in 2005.

Andrew B. Kahng (F'10) received the A.B. degree in applied mathematics (physics) from Harvard College, Cambridge, MA, and the M.S. and Ph.D. degrees in computer science from the University of California at San Diego (UCSD), La Jolla.

He was an Assistant Professor in 1989, an Associate Professor in 1994, and a Full Professor in 1998, with the Department of Computer Science, University of California, Los Angeles. In 2001, he was a Professor with the Department of Computer Science and Engineering (CSE) and the Department of Electrical and Computer Engineering, UCSD, La Jolla. In 2004, he co-founded Blaze DFM, Inc., Sunnyvale, CA, and was the CTO of the company until resuming his duties with UCSD in 2006. He has published over 300 journal and conference papers. Since 1997, his research in integrated circuit design for manufacturability has pioneered methods for the automated phase-shift mask layout, variability-aware analyses and optimizations, chemical–mechanical polishing fill synthesis, and parametric yield-driven, cost-driven methodologies for chip implementation.

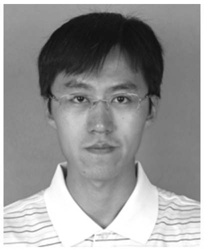
Dr. Kahng was an Associate Chair with the UCSD CSE Department from 2003 to 2004. He was the Founding General Chair of the 1997 Association for Computing Machinery (ACM)/IEEE International Symposium on Physical Design, a Co-Founder of the ACM Workshop on System-Level Interconnect Prediction, and defined the physical design roadmap as a Member of the Design Tools and Test Technology Working Group (TWG) in 1997, 1998, and 1999 renewals of the International Technology Roadmap for Semiconductors. From 2000 to 2003, he was the Chair of both the U.S. Design Technology Working Group and the Design International Technology Working Group, and he continues to serve as a Co-Chair of the Design Information Technology Working Group. He has been an Executive Committee Member of the MARCO Gigascale Systems Research Center since its inception in 1998. He has received the National Science Foundation Research Initiation and Young Investigator Awards, three Best Paper Nominations, and six Best Paper Awards.

[11] We have also tried to follow the dose map-specific placement ECO with a second dose map optimization, i.e., applying the *DMopt* and *dosePl* optimizations in alternation. However, this did not result in any further improvement.

Chul-Hong Park (S'03–M'10) received the B.S. and M.S. degrees in mathematics from Kyung Hee University, Seoul, Korea, and the Ph.D. degree in electrical and computer engineering from the University of California at San Diego, La Jolla.

He is currently a Principal Engineer with the Semiconductor Research and Development Center, Samsung Electronics, Seoul, Korea. He currently leads the Integration of Design and Technology Team which has developed computational design rules and computational lithography solutions. He has published over 30 journal and conference papers, and is the holder of seven patents. His current research interests include nanometer physical design, design for manufacturing, integration of transistor/layout/circuit, and design/lithography for emerging technologies.

Dr. Park received the 1998 Honorable Outstanding Researcher Award and two Best Paper Awards at Samsung Electronics. He also received the 2009 Semiconductor Research Corporation Inventor Recognition Award.



Hailong Yao (S'07–M'09) received the B.S. degree in computer science and technology from Tianjin University, Tianjin, China, in 2002, and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2007.

From 2007 to 2009, he was a Post-Doctoral Research Scholar with the Department of Computer Science and Engineering, University of California at San Diego, La Jolla. He has been an Assistant Professor with the Department of Computer Science and Technology, Tsinghua University, since 2009. His current research interests include very large scale integration physical design and design-manufacturing interface.

Dr. Yao received two International Conference on Computer Aided Design Best Paper Nominations in 2006 and 2008, respectively.