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

Kwangok Jeong,[†] Andrew B. Kahng,[†‡] Chul-Hong Park[†] and Hailong Yao[‡]
[†]ECE and [‡]CSE Departments, Univ. of California at San Diego, La Jolla, CA 92093
E-mail: {kjeong,chpark,hailong}@vlsicad.ucsd.edu, abk@ucsd.edu

ABSTRACT

In sub-100nm 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 stepper 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 simultaneously improve both timing yield and leakage power of a given design. We formulate the placement-aware dose map optimization as a quadratic program, and solve it using an efficient quadratic programming solver. In this paper, we mainly focus on the placement-aware dose map optimization problem; in the Appendix, we describe the complementary but less impactful dose map-aware placement optimization, and an efficient cell swapping heuristic. Experimental results are promising: with typical 90nm stepper (ASML Dose Mapper) parameters, we achieve more than 8% improvement in minimum cycle time of the circuit without any leakage power degradation.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: INTEGRATED CIRCUITS—*Design Aids*; J.6 [**Computer Applications**]: COMPUTER-AIDED ENGINEERING

General Terms

Algorithms, Design, Performance

Keywords

Dose Map, Placement, Timing Yield, Leakage Power Reduction

1. 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-100nm 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 (DFM) techniques [1] 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, etc. A statistical leakage minimization method is proposed in [2], which obtains significant improvement in total leakage reduction by simultaneously varying the threshold voltage, gate sizes and gate lengths. [3] 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* [7, 8], allows for optimization of ACLV (Across-Chip Linewidth Variation) and AWLV (Across-Wafer Linewidth Variation)[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* [5], 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., [6]) 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 (causing a 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 (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]

In this paper, 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 our work are as follows.

- A novel method of enhancing circuit performance and parametric yield based on the dose map technology.

[1]ACLV is primarily caused by the mask and scanner, while AWLV is affected by the track and etcher [9].

[2]Optimization of gate CDs according to setup or hold timing (non-)criticality has been used by [3]. What we propose below uses a coarser knob (i.e., the dose map) for design-aware CD control, but has the advantage of not requiring any change to the mask or OPC flows.

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.
DAC 2008, June 8–13, 2008, Anaheim, California, USA.
Copyright 2008 ACM 1-59593-381-6/06/0006 ...$5.00.

- A new design- and equipment-aware dose map optimization (*DMopt*) method that uses dose to modulate gate poly CD across the exposure field, so as to optimize a function of delay and leakage power of the circuit.
- A new dose map-aware placement optimization (*dosePl*) heuristic that considers systematic 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 on a placement-aware timing and leakage optimized dose map. Our paper is organized as follows. Section 2 introduces fundamentals on the DoseMapper concept. Section 3 describes details of the design-aware dose map optimization. Section 4 discusses the overall optimization flow. The experimental results are presented in Section 5.

2. DOSEMAPPER FUNDAMENTALS

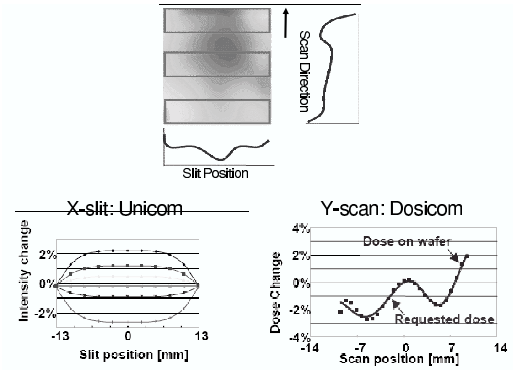

Figure 1: Unicom-XL and Dosicom, which change dose profiles in slit- and scan-directions, respectively. Source: [4].

Figure 1 shows the intrafield DoseMapper concept. In Figure 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 [10] 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 up to 8^{th} order in the dose recipe. Additionally, a maximum gradient constraint of 1% per mm at mask scale in the slit direction is applied in the CD Analyzer to calculate the dose recipe; this limits the correction range for higher-order corrections. Overall, a correction range of $\pm 5\%$ can be obtained with Unicom-XL for the full field size of 26mm 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, $D_{set}(x)$, is used to model parameters for a dose recipe formed of Legendre polynomials (Legendre functions of the first kind) as

$$D_{set}(x) = \sum_{n=1}^{8} L_n P_n \qquad (1)$$

where L_n are Legendre coefficients and P_n are Legendre polynomials. Up to eight Legendre coefficients can be supported. The correction range for the scan direction is $\pm 5\%$ (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).

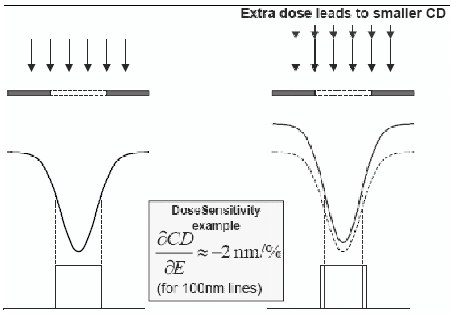

Figure 2: Dose sensitivity: increasing dose (red color) decreases the CD.

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

3. DOSE MAP OPTIMIZATION FOR IMPROVED DELAY AND LEAKAGE

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.*

In the following, for simplicity of exposition we assume that the reticle area taken 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 the 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 $R = |r_{i,j}|_{M \times N}$ 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 improvement. However, G cannot be too small because of current DoseMapper equipment limitations. G can be determined so as to balance between DoseMapper equipment constraints and timing yield improvement.

Circuit Delay Calculation

A typical dose sensitivity D_s at $\leq$ 90nm is -2nm/% [6]; we assume this value below in our experimental evaluations. Gate length changes linearly with dose tuning, i.e, $\Delta L_p = D_s \times d_{i,j}$, where ΔL_p is the gate length change of gate p and $d_{i,j}$ is a percentage value which specifies the relative change of dose in the rectangular grid $r_{i,j}$ wherein the gate is located.

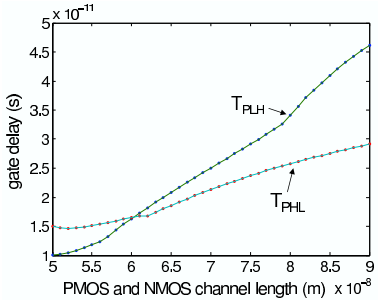

Figure 3: Delay of an inverter vs. gate length.

Figure 3 shows SPICE-calculated transistor delay values as gate lengths are varied in an inverter that is implemented in 70nm technology node. Channel lengths of the PMOS and NMOS devices are equal. T_{PLH} and T_{PHL} represent the low to high propagation delay and the high to low propagation delay, respectively. From Figure 3, the gate delay varies linearly with gate length around the nominal feature size of 70nm. Our background experiments tested Liberty delay model tables of 50 different standard cell masters, and confirmed all the cell masters such an approximate linear relationship for a given pair of input slew and load capacitance.

When the gate length 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 increases linearly as the gate length increases. Since gate length increases linearly when the dose on the gate varies, there is a linear relationship between the change of gate delay and the change of dose on the gate, i.e, $\Delta t_p = t'_p - t_p = A_p \times \Delta L_p = A_p \times D_s \times d(r(p))$. Here, t_p and t'_p are the delay of gate p before and after the percentage dose change $d(r(p))$ in the rectangular grid $r(p)$ where gate p is located, ΔL_p is the change in gate length of gate p, and A_p is a fitted parameter that is 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, a different value of A_p is obtained from processing of Liberty nonlinear delay model tables. Total runtime of this procedure for an entire production standard-cell library is less than an hour on a single processor.

For circuit delay calculation, without loss of generality we consider a combinational circuit with n gates as in [11]. Sequential circuits may be addressed similarly, e.g., by 'unrolling' them, using standard techniques, to 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.

Leakage Power Quadratic Approximation

For simplicity, we do not include dose-dependent change of wire delay in our problem formulation; note that a dose map optimization on the transistor gate layer of the IC will not affect wire pattern, 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.

- **Objective:** minimize $\lambda \times T + \Delta P_{leakage}$
- **Subject to:**

$$L \le d_{i,j} \le U \quad \forall\, i \in [1,M],\ j \in [1,N] \tag{2}$$

$$\begin{cases} |d_{i,j} - d_{i+1,j+1}| \le \delta \ \forall\, i \in [1,M-1],\ j \in [1,N-1] \\ |d_{i,j} - d_{i,j+1}| \le \delta \quad \forall\, i \in [1,M],\ j \in [1,N-1] \\ |d_{i,j} - d_{i+1,j}| \le \delta \quad \forall\, i \in [1,M-1],\ j \in [1,N] \end{cases} \tag{3}$$

$$\begin{cases} a_q \le T & \forall\, q \in fanin(0) \\ a_p + t'_q \le a_q & \forall\, p \in fanin(q) \quad (q = 1, \cdots, n) \\ 0 \le a_{n+1} & \\ t'_p = t_p + A_p \times D_s \times d(r(p)) & \end{cases} \tag{4}$$

In our optimization, we assume that the change of leakage power of a gate is a quadratic function of the gate length changes[4], i.e, $\Delta P(\Delta L_p) = \alpha \times \Delta L_p + \beta \times (\Delta L_p)^2$ for gate p. 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 map to adjust the channel lengths of the gates and thereby reduce a weighted sum of circuit delay and 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.

Equation (2) specifies the correction range on the dose, where L and U are user-specified or equipment-specific parameters for the lower and upper bounds on the dose. Equation (3) specifies a smoothness constraint on the dose, which implies that the doses in neighboring rectangular grids should differ by a bounded amount. Equation (4) denotes the delay constraint when the delays of the gates are scaled during the dose adjustment process. In Equation (4), a_p represents the arrival time at node p, which is the maximum delay from source node $n+1$ to node p; $r(p)$ is the rectangular grid in which gate p is located; and $d(r(p))$ is the change in percentage of dose in the grid $r(p)$. The calculation of the total leakage power of the gates in the circuit is given by Equation (5). Note that the parameters A_p in Equation (4) and α_p and β_p in Equation (5) are all gate-specific, i.e., 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.

$$\Delta P_{leakage} = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{p \in r_{i,j}} \alpha_p \times D_s \times d_{i,j} + \beta_p \times D_s^2 \times d_{i,j}^2 \tag{5}$$

The above problem formulation[5] is a quadratic programming problem, which can be solved using classic quadratic programming methods. In particular, we use CPLEX [12] in the experimental platform described below.

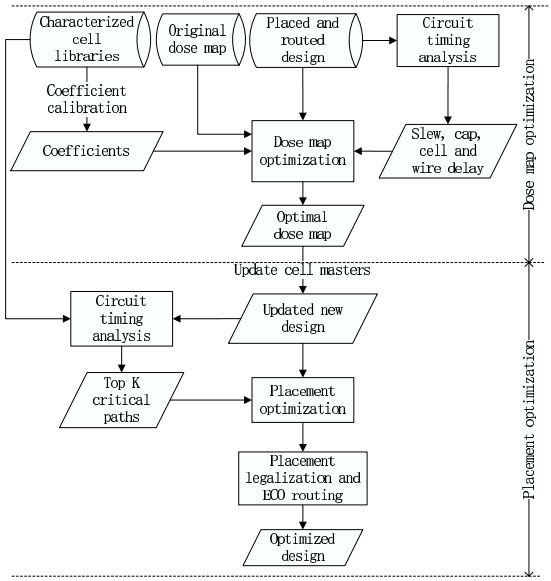

Figure 4: Flow of the timing and leakage power optimization with integrated *DMopt* and *dosePl* (in the Appendix).

4. TIMING AND LEAKAGE POWER OPTIMIZATION FLOW

The Overall Optimization Flow

Figure 4 shows the 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 the original dose map, the characterized standard-cell timing libraries (or, other timing models that comprehend

[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.

[5]The optimization result is feasible for the equipment, as a consequence of the constraints (2) and (3).

the impact of dose on transistor gate length) for different gate lengths, and 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 parasitics data, timing analysis can be performed to generate the input slews and output load capacitances of all the cells. With the input slews and output load capacitances of all the cells, the original dose map, and characterized cell libraries of different gate lengths, the dose map optimization is executed to determine doses that adjust gate lengths of the cells for timing and leakage optimization, subject to dose map constraints. Finally, the optimal dose map is output.

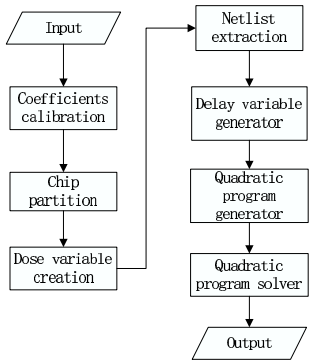

Figure 5: Detailed view of design-aware dose map optimization flow.

According to the optimal dose map, the cell instances in different grids of the dose map will have different gate lengths as well as different cell masters in the characterized cell libraries, i.e., the design will be updated according to the dose map. With the characterized cell libraries, timing analysis is performed on the new design with the updated cell masters to identify the top K (e.g., $K = 10000$) 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.

Summary of the Dose Map Optimization Flow

The dose map optimization in Figure 5 is summarized as follows. The input consists of the original dose map, the characterized cell libraries of different gate lengths, and the input slews and output capacitances of all the cells in the circuit. From the characterized cell libraries of different gate lengths, the coefficients in the linear function of delay and quadratic function of leakage power on gate length 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 field is then partitioned into rectangular grids. For each grid, a variable $d_{i,j}$ represents the amount of dose change in the grid. Maximum circuit delay is captured using variables a_p that represent the arrival time at the output of each cell. When all the variables are obtained, a quadratic programming 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 weighted sum of circuit delay and total leakage power of all the cells. Finally, a quadratic programming solver solves the problem and finds the optimal dose change in each grid based on the original dose map, from which the optimal dose map is calculated.

5. EXPERIMENTAL RESULTS

Table 1: Characteristics of Artisan TSMC 90nm designs.

Design	Chip Size (mm^2)	#Cell Instances	#Nets
AES	0.25	21944	22581
JPEG	1.09	98555	105955

To assess the effectiveness of the proposed dose map optimization algorithm, we first sweep the dose change from −5% to +5% for all the rectangular grids in industrial testcase AES (shown in Table 1) and perform timing analysis using *Synopsys PrimeTime* (version Z-2006.12) [13] and leakage power estimation using *Cadence SOC Encounter* (v06.10) [14]. The timing analysis and leakage power estimation are based on pre-characterized cell libraries with gate length variants. Delay and leakage power results are given in Table 2. "MCT" refers to minimum cycle time and "$P_{leakage}$" refers to the total leakage power of all the cells. Table 2 shows 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 rectangular grids cannot obtain timing yield improvement without leakage power increase. However, our proposed dose map optimization algorithm can obtain substantial timing yield improvement with little or no increase in total leakage power.

The timing and leakage optimization flow is implemented in C++ programming language and tested on industrial testcases as given in Table 1. In the experiments, the smoothness bound δ is set to be 2, and the dose sensitivity D_s is -2nm/%. The parameters A_p, α_p and β_p are calibrated using PrimeTime and SOC Encounter based on the pre-characterized cell libraries. The parameter λ balances between timing and leakage power. Different λ values are tested in our experiments, which show that increasing λ results in better timing improvement but with degraded leakage power optimization.[6] Table 3 shows the dose map optimization results. More than 8% improvement is obtained in minimum cycle time with little or no degradation in total leakage power.

6. 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 of the cells in the circuit, which is useful for optimization of gate delay and gate leakage power. Experimental results show more than 8% improvement in minimum cycle time of the circuit at no cost of leakage power increase. Our ongoing work includes the testing of the proposed dose map-placement co-optimization system on more test cases, especially on larger industrial 65nm designs.

7. REFERENCES

[1] P. Gupta and A. B. Kahng, "Manufacturing-aware physical design", *Proc. IEEE/ACM ICCAD*, 2003, pp. 681-687.

[2] S. Bhardwaj, Y. Cao and S. Vrudhula, "Statistical leakage minimization through joint selection of gate sizes, gate lengths and threshold voltage," *IEEE ASPDAC*, 2006, pp. 953-958.

[3] P. Gupta, A. B. Kahng, P. Sharma and D. Sylvester, "Gate-length biasing for runtime-leakage control," *IEEE Trans. on CAD*, 25(8) (2006), pp. 1475-1485.

[4] `http://wps2a.semi.org/cms/groups/public/documents/membersonly/van_schoot_presentation.pdf`.

[6] λ is the reported in the experimental results as a scaled value. In our studies, λ is used to balance between the delay (ns) and leakage power (nW) of cell instances. A typical λ value is around 200.

Table 2: Delay and leakage values of AES when dose change $d_{i,j}$ is swept from 0% to +5% and from 0% to −5%. The straightforward way of increasing dose cannot obtain delay improvement without the cost of leakage increase.

Dose change	$d_{i,j}$ = 0	$d_{i,j}$ = +0.5	$d_{i,j}$ = +1	$d_{i,j}$ = +1.5	$d_{i,j}$ = +2	$d_{i,j}$ = +2.5	$d_{i,j}$ = +3	$d_{i,j}$ = +3.5	$d_{i,j}$ = +4	$d_{i,j}$ = +4.5	$d_{i,j}$ = +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
$P_{leakage}$ (uW)	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
Dose change	$d_{i,j}$ = 0	$d_{i,j}$ = -0.5	$d_{i,j}$ = -1	$d_{i,j}$ = -1.5	$d_{i,j}$ = -2	$d_{i,j}$ = -2.5	$d_{i,j}$ = -3	$d_{i,j}$ = -3.5	$d_{i,j}$ = -4	$d_{i,j}$ = -4.5	$d_{i,j}$ = -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
$P_{leakage}$ (uW)	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

Table 3: Dose map optimization results with 20×50 rectangular grids and dose correction range ±5%. Significant improvement in delay can be obtained without leakage degradation, or even with leakage reduction.

AES	Nom L_{gate}	λ = 1.0		λ = 1.1		λ = 1.2		λ = 1.3		λ = 1.4		λ = 1.5	
		DMopt	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)
MCT (ns)	1.990	1.825	8.310	1.819	8.620	1.819	8.622	1.815	8.776	1.814	8.839	1.805	9.327
$P_{leakage}$ (uW)	2430.2	2350.8	3.270	2370.4	2.462	2382.2	1.975	2396.7	1.381	2424.9	0.217	2433.6	-0.138
Runtime (s)	–	232.667		239.900		310.786		270.844		127.617		141.948	
JPEG	Nom L_{gate}	λ = 0.6		λ = 0.8		λ = 1.0		λ = 1.2		λ = 1.4		λ = 1.6	
		DMopt	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)
MCT (ns)	2.906	2.776	4.480	2.726	6.172	2.703	6.968	2.687	7.536	2.673	8.007	2.670	8.124
$P_{leakage}$ (uW)	4354.2	3817.1	12.336	4009.5	7.916	4096.4	5.920	4182.1	3.953	4276.0	1.796	4343.2	0.252
Runtime (s)	–	1961.216		1670.229		2108.800		1946.394		1895.096		1835.088	

[5] I. Pollentier, S. Y. Cheng, B. Baudemprez *et al.*, "In-line lithography cluster monitoring and control using integrated scatterometry," in *Proc. SPIE on Data Analysis and Modeling for Process Control*, vol. 5378, 2004, pp. 105–115.

[6] J. B. van Schoot, O. Noordman, P. Vanoppen *et al.*, "CD uniformity improvement by active scanner corrections," *Proc. SPIE Symp. on Optical Microlithography XV*, vol. 4691, 2002, pp. 304–314.

[7] 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," *Proc. SPIE Symp. on Optical Microlithography XVIII*, vol. 5754, 2005, pp. 760–772.

[8] N. Jeewakhan, N. Shamma, S.-J. Choi *et al.*, "Application of dosemapper for 65-nm gate CD control: strategies and results," *Proc. SPIE Symp. on Photomask Technology*, vol. 6349, 2006, p. 63490G.

[9] R. Seltmann, R. Stephan, M. Mazur *et al.*, "ACLV-analysis in production and its impact on product performance," *Proc. SPIE Symp. on Optical Microlithography XVI*, vol. 5040, 2003, pp. 530–540.

[10] http://www.asml.com/

[11] C.-P. Chen, C. C. N. Chu and D. F. Wong, "Fast and exact simultaneous gate and wire sizing by lagrangian relaxation," *IEEE Trans. on CAD* 18(7) (1999), pp. 1014–1025.

[12] http://www.ilog.com/products/cplex/

[13] Synopsys PrimeTime. http://www.synopsys.com/

[14] Cadence SOC Encounter. http://www.cadence.com/

APPENDIX

A. DOSE MAP-AWARE PLACEMENT

A.1 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 section, 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 and swap those pairs for timing yield improvement.

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. 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. Our intuition is that moving a cell within its bounding box has 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.

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.

(1) 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.[7]

(2) HPWL-based (half-perimeter wire length) 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 is below a predefined threshold (e.g., 20% in our experiments reported below) will the cell swap be attempted.

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. The priority for a critical cell during swapping is decided according to the following two factors.

(1) Number of critical paths that pass through the cell. The more critical paths pass through a given cell, the more beneficial it is to swap the cell to a higher-dose region. Higher priorities are assigned to cells on a greater number of critical paths.

(2) 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 Equation (6) where C_l is the crit-

[7] In the experimental results below, this threshold is chosen proportionally to the chip dimension divided by the square root of gate count, which is about 13um for both design AES and JPEG.

Table 4: Experimental results of dose map optimization followed by incremental placement process. The chip is partitioned into 20×50 rectangular grids and the dose correction range is $\pm 5\%$.

AES	Nom L_{gate}	$\lambda = 1.0$				$\lambda = 1.2$				$\lambda = 1.5$			
		DMopt	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)
MCT (ns)	1.990	1.825	8.310	1.811	9.002	1.819	8.622	1.807	9.194	1.805	9.327	1.799	9.617
$P_{leakage}$ (uW)	2430.2	2350.8	3.270	2353.0	3.180	2382.2	1.975	2384.4	1.887	2433.6	-0.138	2435.1	-0.200
Runtime (s)	–	232.667		3.392		310.786		3.317		141.948		4.902	
JPEG	Nom L_{gate}	$\lambda = 0.6$				$\lambda = 1.2$				$\lambda = 1.6$			
		DMopt	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)
MCT (ns)	2.906	2.776	4.480	2.761	4.989	2.687	7.536	2.668	8.179	2.670	8.124	2.647	8.910
$P_{leakage}$ (uW)	4354.2	3817.1	12.336	3817.1	12.336	4182.1	3.953	4182.1	3.953	4343.2	0.252	4343.3	0.251
Runtime (s)	–	1961.216		173.849		1946.394		131.202		1835.088		86.832	

ical paths where $cell_l$ 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. For cells in the same critical path, they are sorted in non-increasing order according to their weights.

$$W(cell_l) = \sum_{cell_l \in C_l} e^{-slack(C_l)} \quad (6)$$

Algorithm 1 *dosePl* cell swapping heuristic for timing yield improvement.

1. Find cells in top K critical paths by golden timing analysis;
2. Compute weights for critical cells as in Equation (6);
3. Sort critical paths in non-decreasing order according to their slacks;
4. **for** $k = 1$ to K **do**
5. Sort the cells in critical path c_l in non-increasing order according to their weights;
6. **for all** cell $cell_l \in$ critical path c_k **do**
7. **if** # swapped cells in path c_k $n(c_k) > \gamma_1$ **then break**; **end if**
8. Compute bounding box b_l of cell $cell_l$ in path c_k;
9. Get the set of rectangular grids R intersected with b_l;
10. Sort the grids in R in non-increasing order according to the dose $d(r)$ in each grid r;
11. Set $flag \leftarrow$ **false**;
12. **for all** $r \in R$ **do**
13. **if** $d(r) < d(r(cell_l))$ **then break**; **end if**
14. Sort the non-critical cells NC in grid r in non-decreasing order by Manhattan distance from $cell_l$;
15. **for all** $cell_m \in NC$ **do**
16. **if** $dis(cell_l, cell_m) > \gamma_2$ **then break**; **end if**
17. **if** $cell_l \in b_m$ and $cell_m \in b_l$ and $\Delta HPWL(cell_l) < \gamma_3$ and $\Delta HPWL(cell_m) < \gamma_3$ **then**
18. Swap $(cell_l, cell_m)$;
19. Update the number of swapped cells $n(c_s)$ for all critical paths c_s such that $cell_l \in c_s$;
20. Set $flag \leftarrow$ **true**;
21. **break**;
22. **end if**
23. **end for**
24. **if** $flag =$ **true then break**; **end if**
25. **end for**
26. **end for**
27. **end for**

Pseudocode of the cell swapping heuristic: The pseudocode of our cell swapping heuristic is presented as Algorithm 1. The cell swapping process is based on the critical paths, which are first sorted in non-increasing order according to their total path delays. Cells of a given path are then swapped. Since it is not necessary to swap all the cells in a critical path to improve the path's timing, the number of cells swapped for each path is recorded and 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 $0.2 \times$ the cell count on the path). The swapping process checks the bounding box constraint, the dose constraint, the distance between candidate swapping pairs, and computes HPWL-based wirelength increase when the pair is swapped. If the candidate pairs pass all the checks, they will be swapped and the corresponding critical paths are updated to record the increased number of swapped pairs. The cell swapping process continues until all critical paths are considered. When the swapping process finishes, the perturbed placement is legalized and routed by a standard placement tool's ECO (Engineering Change Order) placement and routing functionality. After final (ECO) routing, golden timing analysis is performed with updated parasitics to evaluate the circuit delay improvement.

A.2 Experimental Results

The experimental results of dose map-placement co-optimization are given in Table 4. From the results, *DMopt* can strongly improve the timing yield considering the cost in leakage power increase. Cell-swapping based *dosePl* further improves the result up to 9.6% for AES and 8.9% for JPEG.

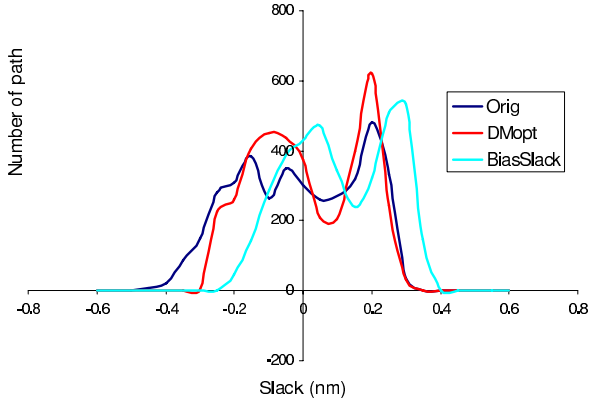

Figure 6: Slack profiles of design AES before *DMopt*, after *DMopt* and the biased design when all the gates in the top 10000 critical paths are enforced using maximum possible exposure dose (+5%).

Figure 6 shows the slack profiles of design AES, including (i) the original design, (ii) the design after dose map optimization ($\lambda = 1.5$, dose correction range is $\pm 5\%$, 20×50 rectangular grids), and (iii) the design 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 Figure 6, the worst slack of the original design is improved significantly by dose map optimization process. But, this lessens opportunity for the following placement process. On the one hand, the difference between the worst slacks of the dose-optimized design and the biased design ("best" design) is quite small (less than 0.05ns); on the other hand, in the dose map-optimized design, the number of critical paths, whose slack values are quite near the worst slack value, is large. To further improve timing, the placement process has to swap many cells to consider all the critical paths with slack values near the worst case. This can introduce many wire detours and make other non-critical paths become critical. In light of the relatively small opportunity left for *dosePl* process, the improvement confirms the effectiveness of the cell swapping based algorithm.[8]

[8]We have also tried to follow the dose map-specific placement ECO with another dose map optimization. However, this did not result in any further improvement.