Improving OPC Quality Via Interactions Within the Design-to-Manufacturing Flow

P. Gupta[a], A.B. Kahng[a,b] and C.-H. Park[b]

[a]Blaze DFM Inc., Sunnyvale
[b]ECE Department, University of California at San Diego

ABSTRACT

Today's design-manufacturing interface lacks essential mechanisms to link disparate disciplines and tool sets. In this paper, we describe three specific mechanisms for improving OPC quality via interactions within the design-to-manufacturing flow. Our studies of these improvements have yielded promising results.

1. INTRODUCTION

Optical lithography has been a key enabler of the aggressive IC technology scaling implicit in Moore's Law. Minimum feature sizes have outpaced the introduction of advanced lithography hardware solutions, so that gate-length and CD tolerances are extremely difficult to achieve. Hence, resolution enhancement techniques (RETs) such as optical proximity correction (OPC), phase shift masks (PSM), and Off-Axis Illumination (OAI) are being pushed ever closer to fundamental resolution limits.[6] RETs, which are imperative during mask data preparation (MDP) today, increase mask cost and should be used judiciously. Existing design-manufacturing interfaces suffer from lack of communication across disciplines and/or tool sets. The result is that both design and manufacturing have limited information about each other, and conservative assumptions must be made on both sides. This leads to sub-optimal performance due to too much guardbanding, and high mask costs and large turnaround time due to over-correction.

We review three techniques that link design and manufacturing for better and cheaper masks.

- *Design-aware optical proximity correction (OPC).* Here we attempt to pass designer's intent to OPC and reduce over-correction. OPC can increase the mask data volume by over 5X; mask cost and turnaround time are proportional to mask data volume. Our technique selectively applies levels of OPC, with higher levels of OPC being applied to devices that are considered critical to circuit performance. We show up to a 34% reduction in mask data volume.

- *Placement for better depth of focus (DOF).* We investigate the feasibility and benefit of minor placement modifications to enhance printability. OAI improves resolution at certain pitches at the expense of others. Pitches where resolution is deteriorated due to OAI, also known as *forbidden pitches*, prevent the correct application of Sub-Resolution Assist Features (SRAFs) and should be avoided. We describe a methodology that perturbs standard-cell placements to reduce the occurrence of forbidden pitches and increase the number of inserted SRAFs.

- *Topography-Aware OPC.* We propose a novel flow and method to drive OPC with a wafer topography map of the layout that is generated by CMP simulation. The wafer topography variations result in local defocus, which we explicitly model in our OPC insertion and verification flows. Our experimental validation uses 90nm foundry libraries and industry-strength OPC and scattering bar recipes. We find that the proposed topography-aware OPC can give up to 90% reduction in edge placement errors at the cost of little increase in mask cost.

The remainder of this paper is organized as follows. In Section 2, we describe our design-aware methodology for OPC effort reduction. Section 3 describes our placement alteration technique to enhance design printability. Interactions between CMP and OPC are explored in Section 4. Section 5 concludes and mentions ongoing work.

Photomask and Next-Generation Lithography Mask Technology XII, edited by Masanori Komuro,
Proceedings of SPIE Vol. 5853 (SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15
doi: 10.1117/12.617372

2. DESIGN-AWARE OPC

In this section we focus on OPC, which is a major contributor to mask costs as well as design turnaround time. More than a 5X increase in data volume and several days of CPU runtime are common side effects of OPC insertion in current designs.[5] OPC affects MDP, defect inspection (and implicitly defect repair), and the mask-writing process itself. Today, variable-shaped electron beam mask writers, in combination with vector scanning*, comprise the dominant approach to high-speed mask writing. In the standard MDP flow, the input GDSII layout data is converted into the mask writer format by *fracturing* into rectangles or trapezoids of different dimensions. With OPC applied during MDP, the number of line edges increases by 4-8X over a non-OPC layout, driving up the resulting GDSII file size as well as fractured data (e.g., MEBES format) volume.[8] Mask writers are hence slowed by the software for e-beam data fracturing and transfer, as well as by the extremely large file sizes involved. Moreover, increases in the fractured layout data volume† lead to disproportionate, super-linear increases in mask writing and inspection time. Compounding these woes is the fact that the total cost to produce low-volume parts is now dominated by mask costs[11] since masks costs cannot be amortized over a large number of shipped products. There is a clear need to reduce the negative implications of OPC on total design cost while maintaining the printability improvements provided by this crucial RET step.

We observe that OPC has traditionally been treated as a purely geometric exercise wherein the OPC insertion tool tries to match every edge as best as it can. As we show in our work, and has been observed by Gupta et al.,[7] such "over-correction" leads to higher mask costs and larger runtimes. A first approach to driving RET explicitly by performance considerations was proposed at DAC-2003 by Gupta et al..[7] Their work proposes selective OPC based on an assumption of several available levels of correction. We describe a design-aware OPC methodology that is demonstrated to be highly implementable within the limitations of current industrial design flows.

2.1. Practical Methodology for Design-Aware OPC

Our flow passes design constraints to the OPC insertion tool in a form that it can understand. As previously mentioned, OPC insertion tools are driven by *edge placement error* (EPE) *tolerances*. Typical model-based OPC techniques break up edges into *edge-fragments* that are then iteratively shifted outward or inward (with respect to the feature boundary) based on simulation results, until the estimated wafer image of each edge-fragment falls within the specified EPE tolerance. EPE (and hence EPE tolerance) is typically signed, with negative EPE corresponding to a decrease in CD (i.e., moving the edge inward with respect to the feature boundary). An example of a layout fragment and its EPE is shown in Figure 1. Mask data volume is heavily dependent on the assigned EPE tolerance that the OPC insertion tool is asked to achieve. For example, Figure 2 shows the change in MEBES file size for a cell with applied OPC as the EPE tolerance is varied. In this particular example, loosened EPE tolerances can reduce data volume by roughly 20% relative to tight control levels.

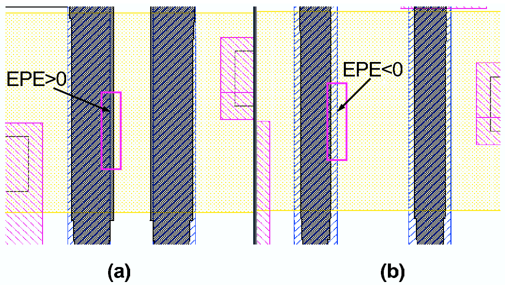

Figure 1. The signed edge placement error (EPE).

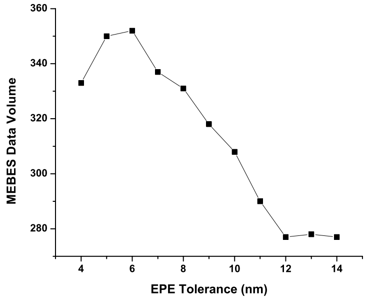

Figure 2. Mask data volume (kB) vs EPE tolerance for a NAND3X4 cell in TSMC 130nm technology.

Since model-based OPC corrects for pattern-dependent CD variation, which is systematic and predictable, we assert that OPC actually determines *nominal timing*, rather than parametric yield as assumed in the work

*Compared to traditional raster scanning, vector scanning allows features to be scaled up or down in size while maintaining sharpness, but the write cost is proportional to feature complexity: the mask pattern must be decomposed into a set of disjoint "shots" or "flashes", each of which takes roughly constant (unit) time.

†E.g., according to the 2003 ITRS,[13] the maximum single-layer MEBES file size increases from 216GB in 90nm to 729GB in 65nm.

of Gupta et al..[7] This allows us to base our OPC insertion methodology on traditional corner-case timing analysis tools instead of (currently non-existent from a commercial standpoint) statistical timing analysis tools. Our methodology adopts a slack budgeting-based approach - as opposed to the sizing-based approach used previously[7] - to determine EPE tolerance values for every feature in the design. For simplicity, our description and experiments reported here are restricted in two ways: (1) we apply selective EPE tolerances in OPC to only gate poly features, and (2) every gate feature in a given cell instance is assumed to have the same EPE tolerance (the approach may be made more fine-grained using the same techniques that we describe). Figure 3 shows our design-aware OPC flow. The quality of results generated by the flow are measured as MEBES data volume of fractured post-OPC insertion layout shapes as well as OPC insertion tool runtime, which can be prohibitive when run at the full-chip level. In the remainder of this section, we describe details of the major steps of Figure 3.

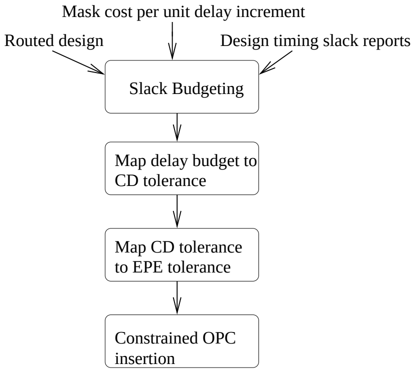
Mask cost per unit delay increment
Routed design
Design timing slack reports
Slack Budgeting
Map delay budget to CD tolerance
Map CD tolerance to EPE tolerance
Constrained OPC insertion

Figure 3. A design-aware OPC flow finds quantified EPE tolerances for layout features and drives OPC with these tolerances.

To map delay budgets found from a linear programming-based formulation to CD tolerances, we require characterization of a standard-cell library with varying gate-lengths. Using such an augmented library, along with input slew and load capacitance values for every cell instance, we can map delay budgets to the corresponding gate lengths. For example, if a particular instance with specified load and input slew rate has a delay budget of 100ps, then we can select the longest gate-length implementation of this gate type that meets this delay. This largest allowable CD will lead to a more easily manufactured gate with less RET effort. CD tolerance of each cell in the design is calculated by subtracting budgeted gate-lengths from nominal gate-lengths.

The next step in our flow maps CD tolerances to signed EPE tolerances. Again, obtaining EPE tolerances is crucial since this is the parameter which OPC insertion tools understand and can exploit. As noted above, in this work we assume positive and negative EPE tolerance to be the same. Since CD is determined by two edges, the worst-case CD tolerance is twice the EPE tolerance.

2.2. Experimental Setup and Results

Now we describe our experiments and the results obtained to validate the design-aware OPC methodology.

Test Cases. We use seven combinational benchmarks drawn from ISCAS85 suite of benchmarks and Opencores.[17] These benchmark circuits are synthesized, placed and routed in a restricted TSMC 0.13 μm library containing 32 cell macros with cell types of BUF, INV, NAND2, NAND3, NAND4, NOR2, NOR3, and NOR4. The test cases are *c432* (337 cells), *c5315* (2093 cells), *c6288* (4523 cells), *c7552* (2775 cells), and *alu128* (12403 cells).

Library Characterization. We assume a total of EPE tolerance levels ranging from ±4nm to ±14nm. Corresponding to each EPE tolerance, the worst case gate-length is $130nm + EPE\_Tolerance$. We map cell delays to EPE tolerance levels by creating multiple .lib files for each of the 10 worst case gate-lengths using circuit simulation. For simplicity, we neglect the dependence of delay on input slew in our analysis but this could easily be added to the framework.

Expected mask cost for each cell type is extracted as a function of EPE tolerance. We run model-based OPC using *Mentor Calibre* on individual cells followed by fracturing to obtain MEBES data volume numbers for each (cell, tolerance) pair. Though the exact corrections applied to a cell will depend somewhat on its placement environment, stand-alone OPC is fairly representative of data volume changes with changing EPE tolerance.

Test	Traditional OPC Flow				Design-aware OPC Flow							
	CD Distribution		OPC	Delay	Budgeting	CD Distribution				OPC	Delay	Norm.
	All Gates (nm)		Runtime	(ns)	Runtime	All Gates (nm)		Critical Gates (nm)		Runtime	(ns)	MEBES
	mean	σ	(s)		(s)	mean	σ	mean	σ	(s)		Volume
alu128	126.1	1.48	51516	3.28	11	131.5	4.93	130.8	2.04	33535	3.28	0.76
c7552	126.2	1.89	7149	1.59	4	132.0	4.77	130.1	1.99	5142	1.59	0.78
c6288	126.0	1.37	12830	5.21	9	131.4	4.45	129.7	1.27	9710	5.21	0.82
c5315	126.1	1.82	4539	1.94	3	131.7	4.70	129.7	1.89	4247	1.94	0.79
c432	126.8	1.57	1020	1.33	1	131.3	3.90	129.9	1.67	737	1.33	0.83

Table 1. Impact of design-aware OPC optimization on Cost and CD. All runtimes are based on a 2.4GHz Xeon machine with 2GB memory running Linux.

Finally, we calculate the sensitivity of mask cost to delay change under the assumption that cost reduction is a linear function of delay increase. This assumption is based on linearity between gate delay and CD as well as the rough linearity shown in Figure 2 between data volume and EPE tolerance. We then build a .lib-like look-up table of correction cost sensitivities (with respect to the tightest EPE tolerance of 4nm).

Design-aware OPC with Calibre. Our OPC flow involves assist-feature insertion followed by model-based OPC. The EPE tolerance is assigned to each gate by the *tagging* command within Calibre. We first separate the entire poly layer into gate poly and field poly components. The field poly tolerance is taken to be ±14nm while gate poly tolerance ranges from ±4nm to ±14nm. We take 1nm as our step size* when applying OPC to obtain very precise correction levels. We set the iteration number to the minimum value beyond which adding mask cost and CD distribution show little sensitivity to OPCs, which is found experimentally. After model-based OPC is applied, we perform 'printimage' simulations in Calibre to obtain the expected as-printed wafer image of the layout. Average gate CD and its standard deviation are extracted from this wafer image. The corrected GDSII is fractured into MEBES using CalibreMDP. The total mask data volume is then determined based on the MEBES file sizes.

Results. We synthesize the benchmark circuits using *Synopsys Design Compiler*. Place and route is performed using *Cadence Silicon Ensemble*. *Synopsys Primetime* is used to output the slack report of the top 500 critical paths as well as the load capacitance for each driving pin. As noted above, STA is run with a modified 134nm (tightest EPE tolerance) library with pin capacitances corresponding to 144nm (loosest EPE tolerance) to remain conservative after slack budgeting. We use *Cplex v8.1*[19] as the mathematical programming solver to solve the budgeting linear program. Since the circuit sizes are fairly small, we use only a single iteration to solve the budgeting problem.

Table 1 compares the runtime and data volume results for design-aware OPC and traditional OPC. The budgeting approach ensures that there is no timing degradation going from the traditional to the design-aware OPC flow. Moreover, unlike sizing, budgeting does not involve iterations with timing analysis. As a result, budgeting runtimes are negligibly small, ranging from 1s to 11s. The important result is the amount of mask cost reduction achieved, whether measured as runtime of model-based OPC or fractured MEBES data volume. The design-aware OPC flow reduces MEBES data volume by 17%-24% which directly translates to substantial mask write time improvements. OPC runtimes are improved by 6%-34% which translates to substantial absolute turnaround time savings. For instance, the design-aware OPC flow saves 5 hours compared to the traditional OPC flow on a small 12,000-gate benchmark.

3. PLACEMENT FOR BETTER DOF

Combinations of RET techniques can provide certain advantages for lithography manufacturing, e.g., OAI and OPC, together with SRAF, achieve enhanced CD control and focus margin at minimum pitch. However, whenever OAI is applied, there will always be (non-minimum) pitches for which the angle of illumination works with the angle of diffraction to produce a bad distribution of diffraction orders in the lens. These pitches are called *forbidden pitches* because of their lower printability, and designers should avoid such pitches in the layout. However, it is very difficult to consider all possible forbidden pitches in the design stage, particularly since the forbidden pitches are dependent on optical conditions which are often tuned in manufacturing. The resulting *forbidden pitch problem* for the manufacturing-critical poly layer must be solved before detailed routing, since routing "locks in" the poly layer layout. At the same time, we wish to address the forbidden pitch problem as late as possible, to avoid extra rework upon modification of the manufacturing process recipe. In this paper, we describe a novel dynamic programming-based algorithm for *AFCorr* (Assist-Feature Correctness), which uses flexibility in detailed placement to avoid forbidden pitches and the manufacturing uncertainty caused by them.

*Step size is the minimum perturbation to an edge that model-based OPC can make. Smaller step sizes lead to better correction accuracy at the cost of runtime.

3.1. Assist Feature Correction Methodology

Modified Design and Evaluation Flow. To account for new geometric constraints arising from SRAF OPC in physical design, we add forbidden pitch extraction and post-placement optimization into the current ASIC design methodology. Figure 4 shows the modified design and evaluation flows in the regime of forbidden pitch restrictions. Of course, we must assume that the library cells themselves have been laid out with awareness of forbidden pitches, and indeed our experiments with commercial libraries confirm that there are no forbidden pitch violations in poly geometries within commercial standard cells. SRAF insertion rules for enhancing DOF margin are determined based on best and worst focus models.* Post-placement optimization generates a new placement which is more conducive to insertion of SRAFs, thus allowing a larger process window to be achieved. The two layouts generated by conventional and assist-correct flow undergo comprehensive SRAF OPC. The amount and impact of the applied RET is a function of the circuit layout. Thus we can evaluate how assist-correct placement impacts circuit performance and printability/manufacturability using measures of SRAF and EPE. The following subsections give more details of forbidden pitch extraction and its design implementation.

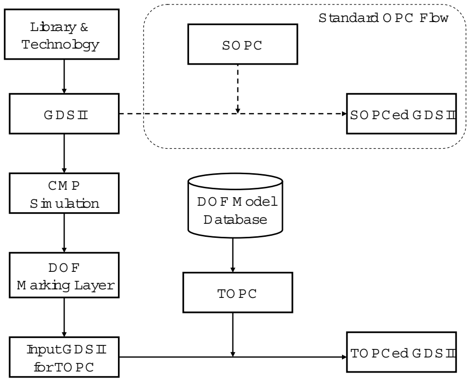

Figure 4. Modified design and evaluation flows: Forbidden pitch extraction and post-placement optimization are added to the traditional ASIC design flow.

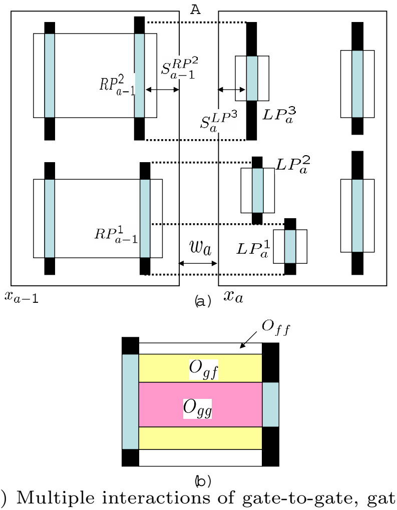

Figure 5. (a) Multiple interactions of gate-to-gate, gate-to-field, and field-to-gate, and (b) overlapped area in the region A of (a).

SRAF and Forbidden Pitch Rules. Lack of space prohibits insertion of a sufficient number of SRAFs, and as a result, patterns violate CD tolerance through defocus. Forbidden pitches are pitch values for which the tolerance of a given target CD is violated. Allowable pitches are all pitches other than forbidden pitches.

Our SRAF insertion rule is initially generated based on the theoretical background given by Shi et al..[12] Positioning of SRAFs is then adjusted based on OPC results. Large CD degradation through-pitch increases pattern bias as model-based OPC is applied, and this requires trimming of the SRAF rule to guarantee better process margin and prevent the SRAFs from printing.† After applying SRAF OPC with a best-focus model, test patterns are simulated with the worst-defocus model. This evaluation yields the forbidden pitches, considering maximum printability and manufacturability. The forbidden pitch rule is determined based on CD tolerance and worst defocus level which can be changed by requirements of device performance and yield. We report that CD tolerance is assumed to be ±10% of minimum line width while the worst defocus level is assumed to be $0.5\mu m$.

Assist Feature Correction. Given a cell C_a, let LP_a and RP_a be the sets of valid poly geometries in the cell which are located closest to left and right outlines of the cell respectively. Only the geometries with length larger than minimum allowable length of SRAF features are considered. Define $s_a^{LP^i}$ to be the space between the left outline of the cell and the i^{th} left border poly geometry. Also assume a set $AF = AF_1, \ldots, AF_m$ of spacings which are "assist-correct". I.e., if the spacing between two gate poly shapes belongs to the set AF, then required number of assist features can be inserted between the two poly geometries. AF_j denotes the j^{th} member of the set of assist-feature correct spacings AF when AF is assumed to be sorted in increasing order. Note that the

*In general, the best focus is shifted from zero to about $0.1\mu m$ due to refraction in the resist. The worst defocus is the maximum allowable defocus corner for manufacturability in a lithography system.

†More complicated approaches to SRAF rule generation may involve co-optimization of model-based OPC and SRAF insertion. We do not address such involved optimizations of OPC, since the focus of our work is OPC-aware design and not OPC itself.

set AF may contain a number of spacings which correspond to varying SRAF widths. Let w_a denote the width of cell C_a and x_a denote its (leftmost) placement coordinate in the given standard cell row (indexed from left to right). In addition, let δ be the cell placement perturbation to adjust the spacing between cells. Then the assist-correct placement perturbation problem is formulated as follows.

$$\text{Minimize} \sum \mid \delta_i \mid$$
$$\delta_{a+1} + x_{a+1} - x_a - \delta_a - w_a + s_{a+1}^{LP^k} + s_a^{RP^g} \in AS$$
$$\text{s.t. } LP^k \text{ and } RP^g \text{ overlap}$$

The objective can be made aware of cells in critical paths by a weighting function. Since the available number of allowable spacings is very small, obtaining a completely assist-correct solution is usually not possible in a fixed cell row width context. Therefore, a more tractable objective is to minimize the expected CD error at a predetermined defocus level. We solve this "continuous" version of the above problem by a dynamic programming approach. The recurrence relation is given below.

$$Cost(1, b) = \mid x_1 - b \mid$$
$$Cost(a, b) = \lambda(a) \mid (x_a - b) \mid +$$
$$Min_{i=x_{a-1}-SRCH}^{x_{a-1}+SRCH}\{Cost(a-1, i) + HCost(a, b, a-1, i)\}$$

Here, $Cost(a, b)$ is the cost of placing cell a at placement site number b. The cells and the placement sites are indexed from left to right in the standard cell row. We restrict the perturbation of any cell to $\pm SRCH$ placement grid points. This is done to contain the delay and runtime overheads of AFCorr placement post-processing. The factor λ decides the relative importance of preserving the initial placement and the final AFCorr benefit achieved, and is a function of the cell instance. In the current implementation λ is directly proportional to the number of critical paths that pass through the given cell instance. $HCost$ corresponds to the printability deterioration in defocus conditions for the vertically-oriented poly geometries closest to the cell boundary; the $HCost$ term depends on the difference between the current nearest-neighbor spacing of the polys and the closest assist-feature correct spacing. The method of computing $HCost$ is shown in Figure 6.

```
HCost(a,b,a-1,i) of Cell $C_a$
Input:
  User-defined weight for overlapping field polys : $c_{ff}$
  User-defined weight for overlapping gate polys : $c_{gg}$
  User-defined weight for overlapping gate and field polys : $c_{gf}$
  Origin x (left) coordinate and length of cell $C_a$ = b
  Origin x (left) coordinate and length of cell $C_{a-1}$ = i
  Width of cell $C_a = w_a$
  Width of cell $C_{a-1} = w_{a-1}$
Output:
  Value of $HCost$
Algorithm:
  Case $a = 1$ : $HCost(1, b) = 0$
  Case $a > 1$ Do
  $N$:= cardinality of the set $RP_{a-1}$
  $M$:= cardinality of the set $LP_a$
    For $(k = 1 ; k = N ; k = k + 1)${
      For $(g = 1 ; g = M ; g = g + 1)$ {
  Let $Hspace(k, g)$ denote the horizontal spacing between $RP_{a-1}^k$
  and $LP_a^g$. $O_{ff}(k, g)$, $O_{fg}(k, g)$ and $O_{gg}(k, g)$ denote
  the field-to-field, field-to-gate and gate-to-gate overlap lengths
  between $RP_{a-1}^k$ and $LP_a^g$.
  $slope(j)$ is the degradation of CD with respect to pitch when
  spacing between two poly geometries is between $AF_j$ and $AF_{j+1}$.
    /* Calculate overlap weight between $RP_a^k$ and $LP_{a-1}^g$ */
          $weight(g, k) = slope(j) \times (Hspace(k, g) - AF_j)$
            $\times(c_{ff}O_{ff}(k, g) + c_{gf}O_{gf}(k, g) + c_{gg}O_{gg}(k, g))$
             s.t.  $AF_{j+1} > Hspace(k, g) \geq AF_j$,
        $Hcost(a, b, a-1, i) += weight(g, k)$
      }
    }
```

Figure 6. $HCost$ calculation.

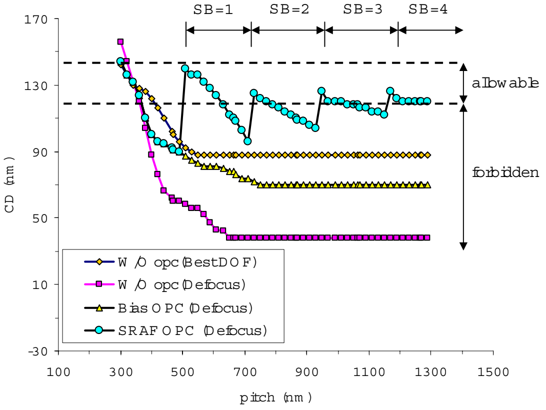

Figure 7. Evaluation of proximity plots in through-pitch: Best focus without OPC, worst defocus without OPC, worst defocus with BIAS OPC, and worst defocus with SRAF OPC.

O_{gg}, O_{ff} and O_{gf} respectively correspond to length of overlapped area in the cases of gate-to-gate, field-to-field, and gate-to-field poly as shown in Figure 5. In addition, c_{gg}, c_{ff}, and c_{gf} are proportionality factors which decide the relative importance of printability for gate and field poly. Typically, gate poly geometries need to be better controlled through process as they directly impact chip performance. Therefore, a typical order is $c_{gg} \geq c_{fg} \geq c_{ff}$. The term *slope*$(j)$ is defined as delta CD difference over delta pitch between AF_j and AF_{j+1}. Thus perturbation cost is a function of *slope*, length and weight of overlapped polys, and space for SRAF insertion. The algorithm takes a legal placement as an input and outputs as a legal placement with better depth of focus properties. The runtime of the algorithm is $O(ncell \times SRCH)$, where $ncell$ is the total number of cells in the design.

3.2. Experiments and Discussion

Experimental Setup. We synthesize *alu*128 benchmark design from *Opencores* in *Artisan TSMC 0.13μm* and *Artisan TSMC 0.09μm* libraries using *Synopsys Design Compiler v2003.06-SP1*. *alu*128 synthesizes to 13279 cells and 8722 cells in 130nm and 90nm technologies respectively. The synthesized netlists are placed with row utilization ranging from 50% to 90% using *Cadence First Encounter v3.3*. All designs are trial routed before running timing analysis. On the lithography side, we use *KLA-Tencor Prolith* to generate models for OPC. *Mentor Graphics Calibre* is used for model-based OPC, SRAF OPC and optical rule checking (ORC). Simulation is performed with wavelength λ =248nm and numerical aperture NA = 0.6 for 130nm, and λ =193nm and NA = 0.75 for 90nm. An annular aperture with σ =0.85/0.65 is used for both processes.

Proximity plots with fixed line width of 0.13μm are illustrated in Figure 7. Exposure dose focuses on the pattern in the minimum pitch of 0.13μm. CD degradation increases through-pitch as the defocus level increases. Patterns in the pitches of over 0.4μm before OPC are outside the allowable tolerance range at the worst defocus of 0.5μm. After BIAS OPC, pitches up to 0.38μm are allowable for CD tolerance while all pitches larger than than 0.38μm should be forbidden. After evaluating SRAF OPC patterns with the worst defocus model, a set of forbidden pitches is obtained as follows: [0.37, 0.509], [0.635, 0.729], [0.82, 0.949], and [1.09, 1.169]. Forbidden pitches still remain after SRAF OPC even though OPC considerably reduces forbidden pitches in comparison to BIAS OPC. SRAF rules are generated based on the criteria mentioned above, with results shown in Table 2. SRAF width is 60nm for 130nm and 40nm for 90nm technology.

	0.13μm Litho.		0.09μm Litho.	
	Pitch($X : \mu m$)	Slope	Pitch($X : \mu m$)	Slope
#SRAF = 0	$0 \leq X < 0.51$	0.28	$0 \leq X < 0.41$	0.162
#SRAF = 1	$0.51 \leq X < 0.73$	0.22	$0.41 \leq X < 0.57$	0.075
#SRAF = 2	$0.73 \leq X < 0.95$	0.105	$0.57 \leq X < 0.73$	0.062
#SRAF = 3	$0.95 \leq X < 1.17$	0.07	$0.73 \leq X < 0.89$	0.050
#SRAF = 4	$1.17 \leq X$	0.02	$0.89 \leq X$	0.012

Table 2. SRAF rule table in 0.13μm and 0.09μm lithography.

Experimental Results. The post-placement optimization is performed based on forbidden pitches and slopes of CD error within them. After AFCorr placement perturbation, we obtain a new placement wherein the coordinates of cells have been adjusted to avoid the forbidden pitches. We use three printability quality metrics. *Forbidden Pitch Count* is the number of border poly geometries estimated as having greater than 10% CD error through-focus. *EPE Count* is the number of edge fragments on border poly geometries having greater than 10% edge placement error at the worst defocus level. This is estimated by ORC. *SB Count* is the total number of scattering bars or SRAFs inserted in the design. A higher number of SRAFs indicates less through-focus variation and hence is desirable. We use $c_{fg} = c_{gg} = c_{ff} = 0.33$, $\lambda(a) = \frac{sitewidth}{10} \times$ number of top 200 critical paths passing through cell a and $SRCH = 5$. All the results have been tabulated in Table 3. Reductions of EPE and forbidden pitch are investigated in each utilization. The increase in total number of SRAFs inserted is also shown in Table 3. Forbidden Pitch Count improves 84%-98% in 130nm and 72%-90% in 90nm. EPE Count enhances 62%-76% in 130nm and 74%-85% in 90nm. In addition, SB Count has the range of improvement 0.1%-6.4% for 130nm and 0%-7.4% for 90nm. Note that these latter numbers are small as they correspond to the entire layout rather than just the border poly geometries.

The number of total SRAFs increases as the utilization[*] decreases, since there is increased white space between cells. The benefit of AFCorr decreases with lower utilization, because there is increased availability of whitespace for SRAF insertion. With additional insertion of SRAFs, there is a small increase in SRAF OPC runtime (< 3.6%) and final data volume (< 3%). The change in estimated post-trial route circuit delay ranges from -7% to +11%, but it should be emphasized that this is a very noisy estimate.

[*] Cell utilization is the percentage of floorplan area used for actual cell placements. Lower utilization implies larger whitespace in the design.

	Utilization (%)	90		80		70		60		50	
	Flow	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr
130nm	# Forbidden	8315	1305	6883	389	4224	121	2347	38	185	3
	# SB	158987	169158	173673	183172	185493	191874	195741	198948	212079	212365
	# EPE	6572	2462	5098	1312	4198	1210	2760	742	216	50
	Runtime (s)	6721	6732	6839	6899	6878	6923	6943	6944	7032	7039
	GDS (MB)	42.9	41.9	41.8	42.3	42.2	42.2	44.9	44.9	45.2	45.4
	Delay (ns)	4.21	4.49	4.547	4.444	4.501	4.372	5.142	4.976	5.051	4.942
90nm	# Forbidden	5171	965	3484	510	1801	323	1130	291	53	10
	# SB	115652	124262	139182	142167	153904	155120	164264	165397	182572	182579
	# EPE	9118	2505	6229	1292	2468	635	2134	600	349	33
	Runtime (s)	4835	5011	5451	5535	5529	5632	5685	5698	5943	5944
	GDS (MB)	41.1	42.3	41.2	43.2	42.2	42.3	42.9	42.8	43.6	43.6
	Delay (ns)	2.478	2.305	2.458	2.602	2.522	2.47	2.867	3.176	3.113	3.046

Table 3. Summary of AFCorr results. Runtime denotes the runtime of SRAF insertion and model-based OPC. The AFCorr perturbation runtime ranges from 2 to 3 minutes for all test cases. GDS size is the post SRAF OPC data volume.

4. WAFER TOPOGRAPHY-AWARE OPC

Depth of focus is the major contributor to lithographic process margin. One of the major causes of focus variation is imperfect planarization of fabrication layers. Presently, OPC (Optical Proximity Correction) methods are oblivious to the predictable nature of focus variation arising from wafer topography. As a result, designers suffer from manufacturing yield loss, as well as loss of design quality through unnecessary guardbanding. The wafer topography variations result in local defocus, which we explicitly model in our OPC insertion and verification flows. Our new TOPC methodology informs OPC insertion by estimated defocus values derived from simulation of the *chemical-mechanical planarization* (CMP) process. After fabrication of a given chip layer, variation in topography creates focus variation in the lithography used to create the next layer of the chip. We use CMP simulation to compute a topographic map over the chip layout; this yields for each layout feature an associated height. The overall TOPC methodology, as distinguished from standard OPC (SOPC), is summarized in Figure 8.

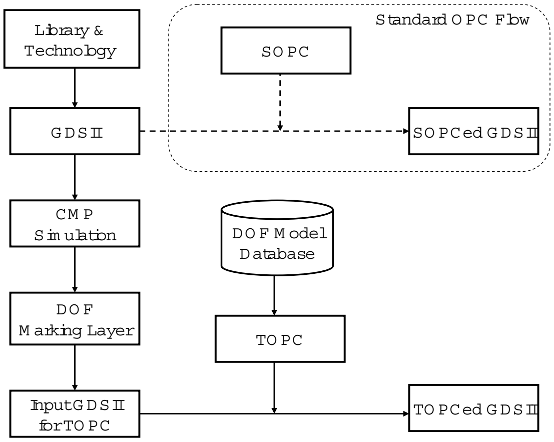

Figure 8. Modified design and evaluation flow: a map of thickness variation from CMP simulation is converted to defocus marking layers and then into GDSII for input to TOPC.

While the CMP simulation yields a continuous topographic map, it is necessary to use only a small number of discrete defocus values when calculating the OPC solution. Thus, the central problem is to assign one of the available defocus values to each layout feature, while reflecting the topographic map as accurately as possible. The details of the entire TOPC flow are given in.[20] We reproduce some results from[20] in the following.

4.1. Experiments and Results

In our experiments, we synthesize the *alu*128 benchmark design from *Opencores* in *Artisan TSMC 0.09μm* libraries using *Synopsys Design Compiler v2003.06-SP1.* The synthesized netlists are placed with row utilization of 90% using *Cadence First Encounter v3.3.* All designs are trial routed before running timing analysis. On the lithography side, we use *Sigma-C SOLID-C* to check CD. *Mentor Graphics Calibre* is used for model-based OPC, SRAF OPC and optical rule checking (ORC). All tests are run on an Intel Xeon 2.4GHz CPU. The calculated maximum thickness variation of Metal 2 is 0.26 μm. We assume maximum DOF variation to be composed of topography variation (50% contribution) and other factors (50%). In our testcase, topography contribution, half of total thickness variation, is $0.13\mu m$. We construct two testcases:

- *CASE I.* Assume the stepper machine focuses on the average of the topography; This is DML1 in our case.
- *CASE II.* Assume the stepper machine focuses on DML2. Therefore, DML0 corresponds to $-0.2\mu m$ defocus.

Assuming that the Bossung plots are symmetrical about 0 focus, metal lines have three different DOF values in CASE I and four different values in CASE II. During TORC, non-topography factors (ΔD), account for 0.13 μm defocus. As a result in TORC, a feature with $0.1\mu m$ thickness value (stepper focusing on $0\mu m$) will have worst case DOF range of $-0.03\mu m$ to $0.23\mu m$.

Table 4 shows results of SOPC and TOPC according to EPE count, which is the number of edge fragments on metal having greater than 10% of CD error. Each OPC'ed metal line is evaluated by TORC with DOF models, i.e., $0.13\mu m$, $0.26\mu m$, and $0.39\mu m$. Specifically, TOPC can reduce EPE count by between 68% and 75% as compared to the standard OPC flow. However, the improvement in process window and potential yield comes at the cost of some increase in data volume and OPC runtime, which is shown in Table 5.

	SOPC	TOPC	Improvement(%)
CASE I	4652	1510	67.5
CASE II	12855	3295	74.3

Table 4. Comparison of EPE counts of SOPC and TOPC

5. CONCLUSIONS

We have presented three techniques that consider design and manufacturing information together to improve mask quality and to make masks cheaper. The first technique, design-aware OPC, proposes a practical means of reducing mask costs and the computational complexity of OPC insertion through performance-driven OPC assignment. In particular, we use edge placement errors to drive OPC insertion tools to correct only as needed to meet timing specifications. Our results on several benchmarks ranging from 300 to 12,000 cells show up to 24% reduction in MEBES data volume, a standard metric for RET complexity. Furthermore, the runtime of the OPC insertion tool is reduced by up to 34%, a critical improvement since running OPC tools at the full-chip level is extremely time-consuming during the physical verification stage of IC design.

The second technique, a novel placement-perturbation approach called AFCorr, is a practical and effective means of achieving assist feature compatibility in physical layouts. AFCorr leads to reduced CD variation and enhanced DOF margin. Our results indicate the following. (1) AFCorr placement perturbation can achieve up to 98% reduction in number of cell border poly geometries having forbidden pitch violations. The corresponding reduction in edge placement error is up to 85%. (2) We achieve up to 7.4% increase in the number of inserted scattering bars in the benchmark design. (3) The increases of data size, OPC runtime and maximum delay overheads of AFCorr are within 3%, 4% and 11% respectively. (4) The runtime of AFCorr placement perturbation is negligible ($\sim$ 3 minutes) compared to the runtime of OPC ($\sim$ 2 hours).

The third technique is a methodology for wafer-topography aware OPC. With an experimental testbed of 90nm foundry libraries, industry OPC recipes, and commercial OPC and ORC software tools, we have confirmed that our technique achieves up to 75% reduction in edge placement errors at worst-case defocus. With dimensions scaling faster than the lithographic process, depth of focus and hence awareness of topographic variation in RET will become increasingly important. Thus, we believe that topography-aware techniques such as ours will be critical for reducing parametric variation - particularly of interconnect performance - in future technology nodes.

	Original	SOPC	SOPC	TOPC	TOPC
			Runtime		Runtime
Testcase	GDS (MB)	GDS (MB)	(min.)	GDS(MB)	(min.)
CASE I	2.3	3.8	35	4.2	43
CASE II	2.3	3.8	35	4.4	45

Table 5. Comparison of OPC runtime and data volume between SOPC and TOPC. Note that SOPC result does not change between CASE I and CASE II.

Our ongoing and future work involves improving the above techniques in terms of quality, runtime and usability. Another area of interest that can be explored is design timing and power validation after traditional “tapeout”. Already, commercial extraction tools try to take into account CMP dependent dishing and erosion effects (though in a very conservative and worst-cased manner) in metal parasitic extraction. A closer to wafer signoff flow is becoming increasingly essential. Errors in dimensions can come from various sources that span OPC, focus/exposure changes in lithography, as well as CMP-based thickness variations. Many of these effects are well-modelable, and hence simulatable, predictable, and compensatable. Hence, simulation results can be passed upstream into the design flow. For instance, shapes generated by optical simulation (e.g. ORC or Optical Rule Check) for devices as well as wires can be passed to an analysis flow to compute timing and power.

REFERENCES

1. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi and V. De, “Parameter Variations and Impact on Circuits and Microarchitecture”, in Proc. IEEE/ACM DAC, 2003, pp. 338-342.
2. E. Bozorgzadeh, S. Ghiasi, A. Takahashi and M. Sarrafzadeh, “Optimal Integer Delay Budgeting on Directed Acyclic Graphs”, in Proc. IEEE/ACM DAC, 2003.
3. Y. Cao, P. Gupta, A. B. Kahng, D. Sylvester and J. Yang, “Design Sensitivities to Variability: Extrapolations and Assessments in Nanometer VLSI”, *Proc. ASIC/SOC* , 2002, pp. 411-415.
4. C. Chen, E. Bozorgzadeh, A. Srivastava and M. Sarrafzadeh, “Budget Management with Applications”, *Algorithmica*, vol 34, No. 3, Jul. 2002, pp. 261-275.
5. P. Gupta, F.-L. Heng and M. Lavin, “Merits of Cellwise Model-Based OPC”, *Proc. SPIE International Symposium on Microlithography*, 2004, to appear.
6. P. Gupta and A. B. Kahng, “Manufacturing-Aware Physical Design”, in Proc. IEEE/ACM ICCAD, Nov. 2003, pp. 681-687.
7. P. Gupta, A. B. Kahng, D. Sylvester and J. Yang, “A Cost-Driven Lithographic Correction Methodology Based on Off-the-Shelf Sizing Tools”, in Proc. IEEE/ACM DAC, Jun. 2003, pp. 16-21.
8. S. Murphy, Dupont Photomask, *SEMATECH: Mask Supply Workshop*, 2001.
9. R. Nair, C.L. Berman, P.S. Hauge and E.J. Yoffa, “Generation of Performance Constraints for Layout”, in *TCAD*, 8(8), 1989, pp. 860-874.
10. R. Rao, A. Srivastava, D. Blaauw and D. Sylvester, “Statistical Estimation of Leakage Current Considering Inter- and Intra-Die Process Variation”, *Proc. ISLPED*, 2003, pp. 84-89.
11. M.L. Rieger, J.P. Mayhew and S. Panchapakesan, “Layout Design Methodologies for Sub-Wavelength Manufacturing”, in Proc. IEEE/ACM DAC, 2001, pp. 85-92.
12. X. Shi, S. Hsu, F. Chen, M. Hsu, R. Socha, and Micea Dusa, “Understanding the Forbidden Pitch Phenomenon and Assist Feature Placement”, Proc. SPIE, Vol. 4689, 2002, pp. 985-996.
13. International Technology Roadmap for Semiconductors, 2003, http://public.itrs.net
14. http://www.synopsys.com/products/mixedsignal/hspice/hspice.html
15. http://www.synopsys.com/products/logic/design_compiler.html
16. http://mentor.com/calibre/datasheets/opc/html/
17. http://www.opencores.org/projects/
18. http://www.kla-tencor.com
19. http://www.ilog.com
20. P. Gupta, A.B. Kahng, C.-H. Park, K. Samadi and X. Xu, “Topography-Aware Optical Proximity Correction for Better DOF margin and CD control”, *Proc. SPIE PMJ*, 2005, to appear.