Self-Compensating Design for Focus Variation

ABSTRACT

Process variations have become a bottleneck for predictable and high-yielding IC design and fabrication. Linewidth variation (ΔL) due to defocus in a chip is largely systematic after the layout is completed, i.e., dense lines "smile" through focus while isolated (iso) lines "frown". In this paper, we propose a design flow that allows explicit compensation of focus variation, either within a cell (*self-compensated cells*) or across cells in a critical path (*self-compensated design*). Assuming that *iso* and *dense* variants are available for each library cell, we achieve designs that are more robust to focus variation. Design with a self-compensated cell library incurs ~11-12% area penalty while compensating for focus variation. Across-cell optimization with a mix of *dense* and *iso* cell variants incurs ~6-8% area overhead compared to the original cell library, while meeting timing constraints across a large range of focus variation (from 0 to 0.4um). A combination of original and iso cells provides an even better self-compensating design option, with only 1% area overhead. Circuit delay distributions are tighter with self-compensated cells and self-compensated design than with a conventional design methodology.

1. INTRODUCTION

Within-die process variation has become one of the most important considerations in IC manufacturing, particularly as lithography moves into the deeply subwavelength regime. Variation can occur at the fabrication stage (intrinsic variation) or during circuit operation (dynamic variation) [1]. There are two major components to intrinsic variation: random and systematic [1],[2],[5]. Because of the strong layout dependency of the systematic component, estimation of systematic variation is impossible until layout information is available. Due to numerous variation sources and their interactions, systematic variation is difficult to predict and often treated as random.

Effective channel length (L_{eff}) variation is one of the clearest determinants of IC performance [3]. Prohibitive increases in the cost of process control necessitate relaxed control of L_{eff} from a manufacturing perspective, shifting the focus to more proactive management of L_{eff} variation from a design perspective. Across chip linewidth variation (ACLV) control is critical to the timing and functionality of a design [4]. Various RETs (Resolution Enhancement Techniques) such as SRAF (Sub-Resolution Assistant Feature), OPC (Optical Proximity Correction), and PSM (Phase Shifting Mask) are commonly used to achieve this in current design-to-manufacturing flows [14],[15].

One of the major sources of L_{eff} variation is focus. Such focus variations can occur, for example, due to changes in wafer flatness or lens imperfections. Traditional corner-case timing analysis flows are very pessimistic in worst-casing focus impact on critical dimensions. This is because layout pitch and focus have very systematic interactions, as shown by so-called Bossung plots (e.g., Figure 1). Recent work [6] notes that comprehending systematic through-pitch and through-focus variations in design can reduce timing uncertainty by up to 30%.

2. COMPENSATING FOCUS VARIATION

Systematic variation can be mitigated to some extent by performing OPC and inserting assist features, but it cannot completely be removed for various reasons (modeling errors, algorithmic inaccuracies, etc.). The remaining linewidth variation due to layout is significant even after the use of complex RET techniques, with isolated and dense lines retaining opposite behavior under varying defocus [6]. Thus, there is a possibility of compensating for systematic variation in the design itself. This compensation can be achieved in two ways:

- *Self-compensated cell layout.* This is a correct-by-construction methodology that relies on within-cell compensation of focus variation. For example, variation can be compensated in series-connected NMOS, if one device becomes thinner (thus, faster) under defocus, and the other device becomes fatter (thus, slower). This can be achieved by making one device "iso" and the other device "dense".
- *Self-compensated physical design.* This refers to compensation across cells (e.g., in a critical path). Consider two cells G1 and G2 that lie on the critical path G1 → G2. Focus variation, if not corrected by applying expensive RETs, can cause variation in delay of the critical path and potential timing failures or parametric yield loss. However, if G1 is explicitly made "iso" while G2 is made "dense", then focus variation can be compensated. Assuming that iso and dense versions of library cells are available, designs that are robust to focus variation become possible.

In this paper we compare and contrast the two approaches put forth above. For example we seek to compare the area overheads of self-compensated libraries vs. across-cell optimizations. We also study a hybrid flow that augments the original library (e.g., with insertion of iso-variant instances) to achieve design robustness. Section 3 describes the construction of a cell library that consists of each version of cells, and Section 4 describes self-compensating design. We present experimental results in Section 5, and Section 6 provides conclusions.

3. ISO/DENSE/SELF-COMPENSATED CELLS

3.1 CD Measurement

To analyze iso/dense/self-compensated behavior with defocus, we use a five-line pattern and sweep the space between the three center line from 180nm to 480nm. SRAF (scattering bar) insertion and OPC are performed on these patterns using Calibre [7]. The average linewidth of the center line is then measured for each pattern.

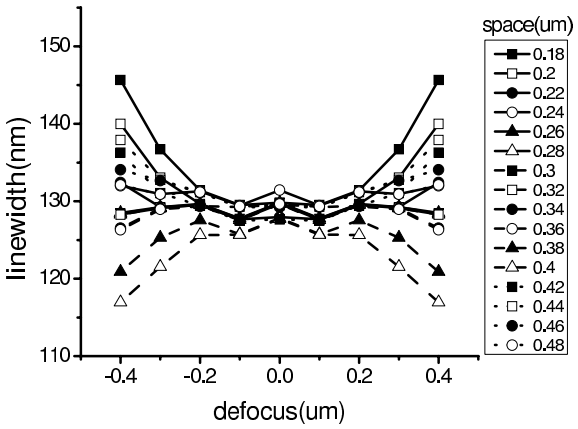

Figure 1. Linewidth variation with defocus level (nominal linewidth = 130nm).

Figure 1 shows the variation in this critical dimension (CD) for different space values at nine different defocus values (-0.4um to +0.4um). In our study 0.0um indicates in-focus conditions and 0.4um is the worst-case defocus level. The figure shows distinct space ranges where the patterns behave as iso, dense or self-compensated.

Based on Figure 1, we generate a look-up table (LUT) using the function *CD = f(LS, RS, F)*, where LS is the left space, RS is the right space, and F is defocus. This allows us to obtain the exact degree to which specific patterns act isolated, dense, or self-compensated, and also to predict CD given defocus and spacings. The tolerance of the self-compensated devices is set at 4nm since the 3σ for the gate CD control is 4nm in 130nm technology [8]. Thus, if linewidths are 4nm larger than nominal at 0.4um defocus, we assume those patterns are "dense"; similarly, if linewidths are 4nm smaller than nominal, we classify the patterns as "iso". Finally, if the CD variation is less than 4nm at 0.4um defocus, we consider the pattern "self-compensated". The first scattering bar insertion point is at a spacing of 420nm, therefore, the "most-iso" pattern has a spacing of roughly 400nm. At 420nm spacing and above, the pattern reverts to "dense" behavior as a result of scattering bar insertion. At the "most-dense" spacing (i.e., 180nm on each side), the linewidth increases 13% from nominal and in the "most-iso" case (i.e., 400nm on each side), the linewidth decreases 11% from nominal at the 0.4um defocus point.

The optimal scattering bar placement and width depend on numerous factors such as wavelength (λ), numerical aperture (NA), illumination type, and others [9]. Reference [10] provides equations for optimal size and placement (defined as SRAF to main pattern spacing) of scattering bars, which are (0.2 ~ 0.25)*(λ/NA) and (0.55 ~ 0.75)*(λ/NA) respectively. We use an optical model of 248nm wavelength and annular illumination with NA=0.7 lens. Table 1 shows the parameters used in Mentor Calibre to generate the linewidth variation with focus level.

3.2 Edge Devices

Special consideration is required for *edge devices*, i.e., devices that are closest to the cell boundary. For example, since there is only one poly line for NMOS and PMOS in an INVX1 (inverter size 1) layout, these are both edge devices. We identify two different cases of edge devices: Case 1 has no neighboring devices on either side (e.g., INVX1), while Case 2 has no neighboring device on exactly one side (e.g., left-most or right-most devices in cells except INVX1 and INVX2 which have no fingers). To investigate the edge effect in Case 1, we first sweep the spacing from 180nm to 1um symmetrically on both sides. For Case 2, we fix one side at 180nm or 380nm since most edge devices in TSMC 130nm standard cells have one of these two spaces on one side (i.e., without contact or with contact between poly lines). The spacing on the other side is swept up to 2um. Figures 2 and 3 show linewidth vs. spacing in both Case 1 and 2.

Table 1. Parameters used in CalibreWB

Parameters	Values
λ (wavelength)	248nm
NA	0.7
Illumination type	annular
Scattering bar width	60nm
Scattering bar placement	180nm
Linewidth (nominal)	130nm

As can be seen from the graph in Figure 2, linewidth is insensitive to focus after two SBs are inserted on each side of the poly line.

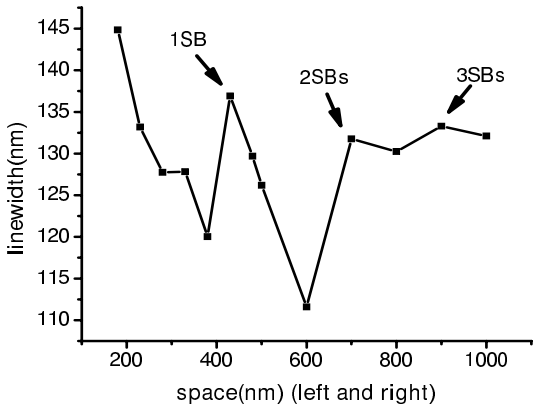

Figure 2. Linewidth variation at 0.4um defocus in Case 1. The arrows indicate scattering bar (SB) insertion points.

Figure 3 shows the Case 2 edge effect. When two adjacent poly lines are 1.2um apart (i.e., 2 SBs are inserted at each side), the linewidth does not vary much even if the spacing becomes larger. Since the distance from edge devices to the cell boundary for all cells is over 600nm in this technology (making the distance of two neighboring poly lines more than 1.2um), we assume that all edge devices in Case 2 follow the behavior seen in Figure 3.

3.3 Area Estimation

The LUT that is constructed from Figure 1 gives CD and also the spacing between poly lines of iso/dense/self-compensated cells. Layout area is then estimated to quantify the area penalty and to extract the parasitic capacitance (i.e. AD, AS, PD, and PS) for these three versions of cells. We estimate the area of these layouts by modifying the space between each device based on the original TSMC013um standard cell layouts.

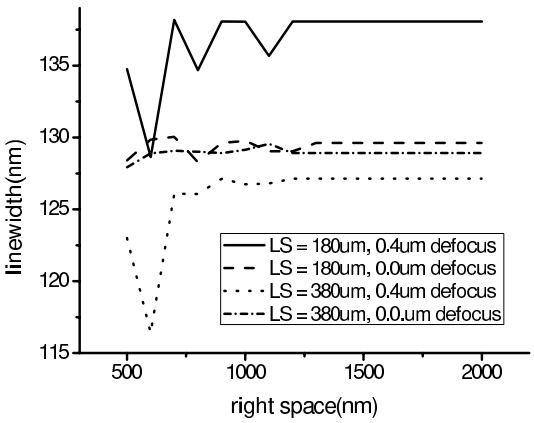

Figure 3. Linewidth with spacing from 0.5um to 2um at 0.0um and 0.4um defocus in Case 2.

Figure 4 shows a representative layout of a large 2-input NAND gate in 130nm technology. Iso/dense/self-compensated devices co-exist in the original cell. Changing inter-device spacings allows us to generate iso/dense/self-compensated versions of cells. Figure 5 shows the average area increase from the original layout for these variants of the base cells. The area increase of iso versions of cells is 17% since they require more space to make all devices appear isolated. The area of self-compensated cells increases about 10% and there is a 3% average area increase in the dense versions.[1]

3.4 Library Generation

Using the layout estimation and LUT that is based on Figure 1, we build SPICE netlists for all the different versions of cells. We consider 21 frequently used cells (INV: x1, x2, x6, x8, x12, NAND2(3) and NOR2(3): x1, x2, x4, x6). The LUT provides the exact isoness/denseness/self-compensatedness of devices at various spacings and two defocus points (0.0 and 0.4um). The CDs of edge devices follow the observations described in Section 3.2. CD is measured at 0.0 and 0.4um defocus value. Therefore we arrive at 4 different libraries (original, iso, dense, and self-compensated libraries) that are characterized at both zero-focus and 0.4um defocus. HSPICE (using Autochar [12]) is employed to generate detailed .lib models.

4. SELF-COMPENSATING DESIGN

4.1 Self-Compensated Cells

Within a cell, self-compensated devices are constructed by modifying the cell layout to have explicitly iso and dense spacings. Starting from the original TSMC 0.13um standard-cell layouts, we generate new versions of cells that are "self-compensated" by adjusting the spacing between poly lines. As can be seen from Figure 1, the self-compensating space range is 260nm to 340nm on each side. However, in our study self-compensation is achieved

[1] Spacing is sometimes increased such that additional scattering bars can be inserted, making devices behave dense. Thus, the dense cells exhibit small area increases over the original library.

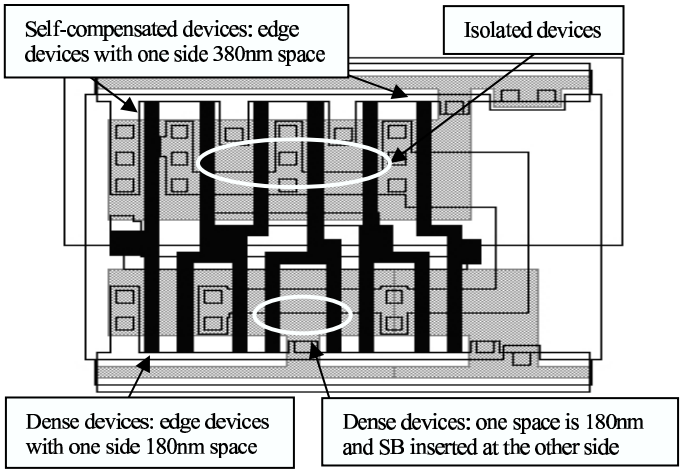

Figure 4. Sample cell layout (NAND2X6) showing isolated, dense, and self-compensated devices.

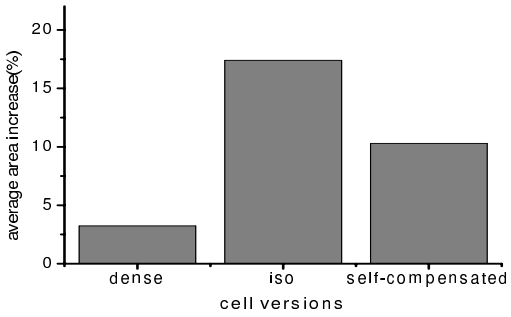

Figure 5. Average area increase for the 3 variants of base cells compared to the original layouts.

by having iso spacing on one side and dense spacing on the other. This allows for smaller cell layout areas with the same linewidth-focus behavior vs. using self-compensated spacings on both sides.

4.2 Optimization (Self-Compensated Physical Design)

4.2.1 Dense with Iso Design

As noted in Section 2, designs that are more robust to focus variation are possible given the availability of self-compensated cells. Another option is to generate optimized circuits using both *dense* and *iso* cells to meet timing at all focus points. This problem of iso-dense self-compensating physical design can be solved as a sizing problem. Since dense cells are slower (at worst-case focus) and smaller while iso cells are faster and bigger, we start with the circuit initially synthesized with dense cells, then swap in iso versions to meet timing at the worst-case defocus level.

Initially, synthesis with the "dense" library results in the slowest timing with small area. The optimization of delay versus area is implemented using a sensitivity-based approach to minimize area penalty while instantiating "iso" counterparts of "dense" cells in the circuit to meet timing constraints. In our experiments, the required time at the primary outputs is set to be 1% higher than the worst-case delay with the original library at 0.0 defocus.[2] The

[2] Because the original library is a mixture of iso, dense, and self-compensated devices, at 0.0um defocus level some cells show better

sensitivity of all gates with respect to a change from "dense" to "iso" variants can be defined as [16]:

$$Sensitivity = \frac{1}{\Delta A + K_1} \sum_{arcs} \frac{\Delta D}{slack_{arc} - S_{\min} + K_2} \quad (1)$$

where ΔA is the change in area and ΔD is the change in delay due to swapping "dense" with "iso". S_{min} is the worst slack in the circuit when synthesized using the "dense" library, and the arcs consist of all rise and fall transitions from each input to output of the gate. The term $slack_{arc}$ is the difference between arrival and required times of the timing arc, and K_1 and K_2 are small positive numbers to ensure stability of the equation. Pseudocode for the first phase of our optimization process is as follows:

```
While worst_slack is negative
{
Calculate sensitivities of all gates in the circuit
Sort sensitivities in non-increasing order
Swap the "dense" version with "iso" cell based on the
order of sensitivities
Calculate new_delay of circuit
Update Worst_slack
}
```

As the pseudocode indicates, we first sort sensitivities in non-increasing order. The gate with maximum sensitivity is then swapped with its corresponding "iso" version. Incremental timing analysis updates the *worst_slack* value and new sensitivities are then calculated if the timing is not met. Figure 6 shows an example of the swapping. Since all gates are dense at first, the design may not meet timing at worst-case defocus. Changing from dense to iso will compensate for the focus along critical paths. The optimization iterates this swapping until timing constraints are met.

Even after the above optimization procedure (which ensures timing correctness at both best and worst focus conditions), the circuit may not meet timing constraints at intermediate values of focus since delay variation with focus may be highly non-linear and even non-monotone. Thus, the timing constraint should be checked across defocus levels. If the extreme defocus point is out of the permissible focus range or the maximum delay is less than the required time, no more steps are needed. However, if the maximum-delay defocus point is within the permissible focus range, a post-processing step is required to globally meet the timing constraint. At the extreme defocus point, we can apply the same sensitivity-based optimization process shown above to ensure that the optimized circuit meets timing throughout the expected defocus range.

4.2.2 Original + Iso Design

The original library is a mixture of iso, dense, and self-compensated devices and the delay variation at 0.4um defocus of original cells is (from Table 2) 2.5% to 4.1%. As a result, taking a design implemented with the original library and then optimizing it through the introduction of iso cells provides another option to reducing through-focus variability with small area overhead. This approach is particularly suitable for near-term technologies that already have libraries available – in this case users could choose to supplement this pre-existing library rather than creating two new (iso and dense) libraries. This latter option, corresponding to the strategy in Section 4.2.1 is more applicable in exploratory technologies such as 45nm today for which library design has not yet commenced. Starting from circuits synthesized with the original library at 0.0um defocus, "iso" cells are instantiated to meet timing at the worst case defocus level (i.e., 0.4um). To this end, the same sensitivity-based optimization process described in Section 4.2.1 can be applied.

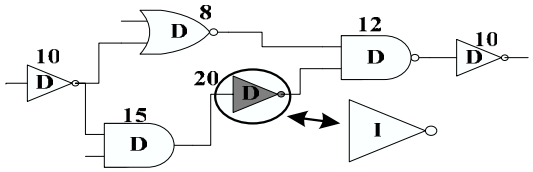

Figure 6. Illustration of optimization process (D denotes "dense" and I denotes "iso" cell; numbers are example sensitivities of gates to swapping to "iso" counterparts)

4.2.3 Original + Reduced-Iso Design

In Section 4.2.2, an iso library containing all base cells is used to improve circuits synthesized with the original (focus-unaware) library. If characterization or library generation effort is a concern, then the frequency count of iso gates actually used by the approach of Section 4.2.2 enables us to filter the iso cells needed for effective optimization. In this way we find it is possible to achieve the timing constraint across all circuits studied with only a subset of all iso cells. In a small experiment, we take the 11 most commonly used cells (INV: x1, x6, x12; NAND2(3) and NOR2(3): x1, x6) out of 21 base cells and apply the same two-phase optimization process as described earlier.

5. RESULTS

To quantify delay variation with defocus across the iso/dense/self-compensated libraries and using our optimization approaches, timing libraries for three different variants of each cell are generated as described in Section 3. ISCAS85 benchmark circuits are then synthesized at three different timing constraints (i.e. minimum, 10% and 20% slower than minimum timing) using Synopsys Design Compiler [13],[17].

The worst-case path delay of selected ISCAS85 benchmarks with iso/dense/self-compensated libraries at 0.0 and 0.4um defocus level is shown at Figure 7. Average delay variations of all benchmarks for the circuits synthesized with 3 different timing constraints are shown in Table 2. As can be seen in Figure 7 and Table 2, the dense cells at 0.4um defocus give 15% slower timing than the original library, and iso cells at 0.4um defocus give 5.6% to 9% faster timing than the original library at 0.0um defocus. The reduction of delay with the iso version at 0.4um defocus is slightly smaller than we might expect from the linewidth versus focus curves of Figure 1 because the iso cells have larger parasitics (and hence larger input capacitances) which degrades the speed. The self-compensated cells in which the devices are modified to tolerate the defocus variation by canceling the iso-ness and dense-ness of the patterns shows minimum variation (less than 1%) at 0.4um defocus (in Figure 7, the lines representing original at 0.0um and self-compensated at 0.4um defocus overlap each other).

timing in the original library than in their iso counterparts (this is not true at 0.4um defocus).

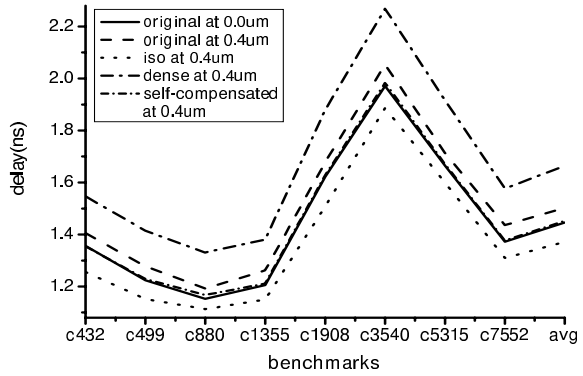

Figure 7. Delay variation of ISCAS85 benchmarks with various libraries at 0.0um and 0.4um defocus

Table 2. Average delay variation across benchmarks at 0.4um defocus with various libraries at 3 timing constraints.

Cell versions	Average Δ delay (%)		
	Min_delay	10% slower	20% slower
original	2.5	4.1	4.1
iso	-8.9	-5.6	-6.2
dense	14.8	14.9	15.2
self-compensated	0.5	0.5	1.0

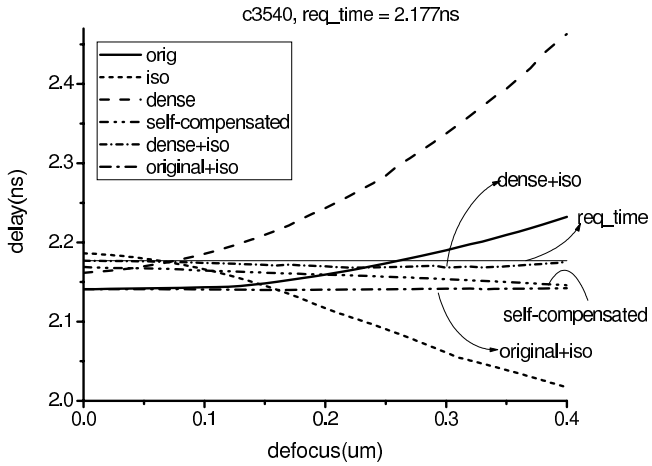

Figure 8. Delay vs. defocus showing the effective compensation in self-compensating design options.

Looking more closely at how the various optimization choices affect timing across the entire defocus range of interest, Figure 8 shows the delay variation for c3540 from 0.0um to 0.4um defocus. As can be seen clearly from the curves, self-compensated cells, dense with iso optimization, and original with iso optimization all satisfy the timing requirement throughout the defocus range. The latter two approaches each benefit from the post-processing step that examines circuit delay at intermediate focus conditions – without this step, timing cannot be guaranteed through the defocus range. The post-processing step typically makes only 5-10 additional cell swaps to ensure timing, so little additional area penalty is incurred.

Figure 9 shows the area penalty of the self-compensating design options, Note that all design options with the exceptions of iso (as described in footnote 2) and dense are able to meet the appropriate timing constraint (also, the original library itself is unable to meet the timing constraint across focus conditions). We observe a 10-12% area penalty for a self-compensated cell based design compared to a 6-9% area penalty for the self-compensating design approach that uses a combination of dense and iso cell variants. Furthermore, the original + iso compensation scheme results in less than 1% area overhead while meeting timing.

Table 3 shows the gate distribution after the full optimization procedure. In the dense and iso optimization of Section 4.2.1, approximately 32% of dense gates must be replaced to satisfy the timing constraint across defocus levels. However, in the original + iso option (Section 4.2.2), only 11% of the original instances needed to be replaced with their iso counterparts. These results clearly explain the very small area penalty seen in the original + iso optimization approach; only a small amount of the larger iso variants must be included.

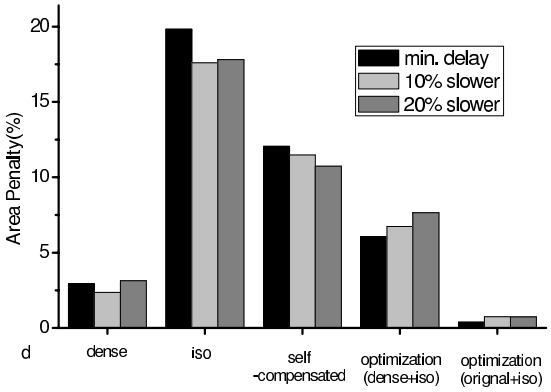

Figure 9. Average area increase from original layout with different design options.

Table 3. Gate distribution after optimization in both cases

benchmarks	Dense + iso		Original + iso	
	dense(%)	iso(%)	original(%)	iso(%)
c432	0.59	0.41	0.88	0.12
c499	0.58	0.42	0.90	0.10
c800	0.71	0.29	0.88	0.12
c1355	0.61	0.39	0.87	0.13
c1908	0.61	0.39	0.88	0.12
c2670	0.58	0.42	0.78	0.22
c3540	0.68	0.32	0.91	0.09
c5315	0.79	0.21	0.93	0.07
c7552	0.71	0.29	0.92	0.08
average	0.68	0.32	0.89	0.11

5.1 Distribution of Focus and Delay

Monte-Carlo simulation with 1000 trials is applied to investigate the impact of defocus variation on delay distribution. A normal distribution of focus with mean = 0.0um and 3σ = 0.4um is assumed. The delay of circuits using the original, iso, dense, self-compensated, and optimization approaches of Section 4.2 are calculated. Figure 10 shows 1000 Monte-Carlo simulation results for the c3540 circuit. Self-compensated, dense and iso, and original with iso library options meet timing requirement at all randomly chosen defocus points and exhibit very delay distributions relative to the original library as well as iso and dense alone. In particular, the two optimization strategies suggested in Sections 4.2.1 and 4.2.2 demonstrate appreciably tighter distributions than the self-compensated cell-based approach.

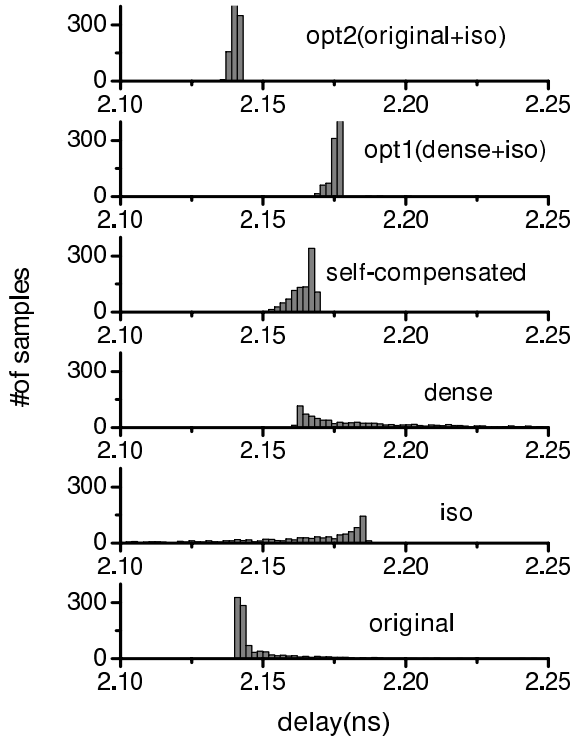

Figure 10. Stacked histogram showing the distribution of delay considering Monte Carlo sampled defocus conditions for c3540 (req_time = 2.177ns)

6. CONCLUSION AND FUTURE WORK

A novel design technique to compensate for lithographic focus variation is proposed in this paper. Given self-compensated cells, which modify devices to cancel expected focus variation, designs that are much more robust to focus variation become possible. Compensated design for focus variation is achievable with small area penalties, assuming that dense and iso counterpart cells are also available. Since the original standard cells are a mixture of iso, dense, and self-compensated devices, we can also choose to add dedicated iso cells to order to meet timing at worst case defocus conditions. We observe that with dense and iso library options we can achieve a compensated design with 6-9% area overhead (compared to 10-12% in a self-compensated library based design) and by supplementing the original library with isolated variants, through-focus timing can be guaranteed with only 1% area penalty.

Our results are based on a 130nm technology – compensation in more advanced technologies such as 65nm is worth investigating as the impact will be even greater. Also, our current sensitivity-based approach is used to optimize delay vs. area. While iso cells become faster under defocus conditions they also exhibit greatly enhanced leakage under defocus conditions since leakage grows exponentially when L_{eff} becomes less than its nominal value. Therefore, joint optimization in the delay/area/leakage space is another compelling area of study.

7. REFERENCES

[1] Y. Cao, *et al.*, "Design Sensitivities to Variability: Extrapolations and Assessments in Nanometer VLSI", *Proc. ASIC/SOC,* 2002, pp. 411-415.

[2] S. R. Nassif, "Design for Variability in DSM Technologies", *Proc. ISQED* 2000, pp. 451-454.

[3] S. R. Nassif, "Within-Chip Variability Analysis", *Proc. IEDM,* 1998, pp. 283-286.

[4] M. Orshansky, L. Milor, P. Chen, K. Keutzer, C. Hu, "Impact of Systematic Spatial Intra-Chip Gate Length Variability on Performance of High-Speed Digital Circuits", *ICCAD*, 2000, pp. 62-67.

[5] P. Gupta and A. B. Kahng, "Manufacturing-Aware Physical Design", *ICCAD,* 2003, pp. 681–687.

[6] P. Gupta and H. Fook-Luen, "Toward a systematic-variation aware timing methodology," *Proc. DAC*, 2004, pp. 321-326.

[7] Calibre version 2004.1_7.33, *http://www.mentor.com.*

[8] "International Technology Roadmap for Semiconductors 2003," *http://public.itrs.net/Files/2003ITRS/Home2003.htm.*

[9] T. Yorick, *et al*, "ArF imaging with off-axis illumination and subresolution assist bars: a compromise between mask constraints and lithographic process constraints," *Proc. SPIE*, 2002, vol. 4691, pp. 1522-1529.

[10] A. J. Lori, T. R. Michael, D. Jason, and J. Christiane, "Effect of scattering bar assist features in 193-nm lithography," *Proc. SPIE*, 2002, vol. 4691, pp. 861-870.

[11] HSPICE version 2004.03, *http://www.hspice.com.*

[12] Autochar - Automates the characterization of digital circuits, *http://directory.fsf.org/design/cad/autochar.html.*

[13] Design Compiler version V-2003.12, *http://www.synopsys.com.*

[14] A. B. Kahng and Y. C. Pati, "Subwavelength Lithography and its Potential Impact on Design and EDA", *Proc. DAC*, 1999, pp. 799-804.

[15] L. W. Liebmann, S. M. Mansfield, A. K. Wong, M. A. Lavin, W. C. Leipold, T.G. Dunham, "TCAD Development for Lithography Resolution Enhancement", *IBM J. RES. & DEV*, vol. 45, no. 5, 2001.

[16] S. Sirichotiyakul, *et al.*, "Stand-by power minimization through simultaneous threshold voltage selection and circuit sizing" *Proc. DAC*, 1999, pp. 436-441.

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