A Framework for Systematic Evaluation and Exploration of Design Rules

Rani S. Ghaida and Puneet Gupta, Electrical Engineering Department, UCLA, {rani,puneet}@ee.ucla.edu

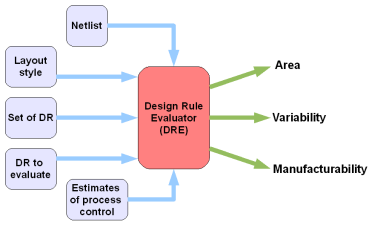


Figure 1: Overview-diagram of DRE framework.

***Abstract*—This paper offers a novel framework for early and systematic evaluation of design rules and layout styles in terms of major layout characteristics of area, manufacturability, and variability. Due to the focus on co-exploration in early stages of technology development, we use first order models of variability and manufacturability (instead of relying on accurate simulation) and layout topology/congestion-based area estimates (instead of explicit and slow layout generation). The framework is used to efficiently co-evaluate several debatable rules (evaluation for a 104-cell library takes 24 minutes). Results show that: a) diffusion-rounding mainly from diffusion power-straps is a dominant source of variability, b) fixed gate-pitch implementation has significant cell-area overhead compared to 1D-poly implementation (18%), and c) 1D-poly restriction, which improves manufacturability and variability, have insignificant area overhead compared to 2D-poly (<1%). In addition, we explore M1-pitch and gate-spacing rules using our evaluation framework. This exploration yields almost identical values as those of a commercial 65nm process, which serves as a validation for our approach.**

I. Introduction

Design rules have been the primary contract between technology and design. While current approaches for defining design rules are largely unsystematic and empirical in nature [1–3], this paper proposes the first framework to explore area-manufacturability-variability tradeoffs of design rules systematically and in a quantitative manner. Rather than fine-tuning DRs, our goal is to make early decisions *before* exact process and design technologies are known. At this stage, accurate evaluation methods and models are unlikely to be available and the return on investment of using them is fairly low. As a result, we use simple but justified approximations for manufacturability and variability unlike [4, 5] that rely on layout generation or perturbation. Since design rule space is very large, we further use fast layout topology generation methods to estimate area as opposed to full-blown layout generation. The accuracy of the former is surprisingly good and allows for explicit "layout style" guidelines as we show later in this paper.

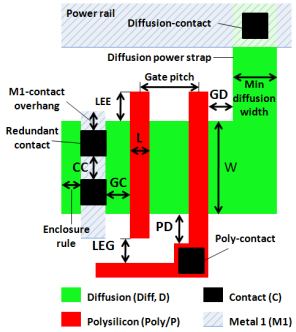
Figure 1 shows the structure of the DR Evaluator (DRE). Given SPICE netlists of cells (possibly scaled down from a previous technology generation), layout style and preferences (e.g., redundant contacts), design rules and their values (see Figure 2), and estimates of process control (e.g., overlay error distribution), only the values of DRs to be evaluated are modified while all other rules remain unchanged. This modified set of DRs is then used to estimate the layout and determine major metrics of area, manufacturability, and variability[1].

II. Area Estimation

The number of design rules is growing tremendously and design rule manuals (DRM) are becoming unmanageable as we move toward smaller feature sizes [6, 7]. As a result, our framework was designed for *fast* evaluation necessary to enable DR exploration/optimization.

This section describes the methods for fast layout topology generation and congestion estimation used in DRE framework.

[1]For brevity and continuity, we skip certain details about the methods and algorithms that were used and do not consider effects of DRs on other circuit characteristics such as delay, power, and reliability along with designability.



DR	Value[nm]
L	$\geq$ 50
M1 1/2 pitch	$\geq$ 65
Gate pitch	$\geq$ 190
Min diff width	$\geq$ 90
LEE	$\geq$ 55
LEG	$\geq$ 75
GC	$\geq$ 35
CC	$\geq$ 75
GD, PD	$\geq$ 50
M1-C overhang	$\geq$ 35
Diff enclosure	$\geq$ 5
Poly enclosure	$\geq$ 5

Figure 2: Illustration of major DRs, their notations and values in FreePDK 45nm process.

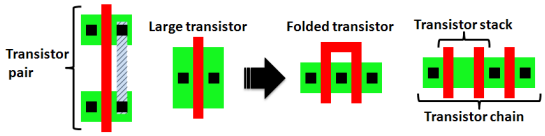


Figure 3: Techniques and notations used in layout topology generation.

A. Layout Topology Generation

Major transistor placement techniques used for layout-area reduction are highlighted in Figure 3. Figure 4 outlines the flow of transistor placement used for layout estimation and describes the algorithms used at each step. We illustrate the application of these steps on a standard cell in Figure 5.

B. Routing Estimation

Rather than performing the time-consuming step of actual routing, DRE estimates routing to approximate the wire length and models congestion and its effect on layout area.

Source/drain (S/D) contacts connected to power supply are located as close as possible to the power rail while meeting DR requirements. All other S/D contacts are located near p/n interface to reduce the length of wires connecting p-to-n type transistors. Figure 6 shows details on transistor interconnections.

C. M1-Congestion Estimation

Occupied track-length in a particular routing-direction is determined as the sum of wire length, line-end spacing DR between wire segments, and track-length blocked by wires in the orthogonal direction. At intersections of tracks in orthogonal directions (excluding the ones needed to form connections), a track-length equal to the minimum line-width plus spacing DRs is blocked.

Cell-area is increased if M1 track-congestion (defined as the ratio of occupied to available track-length) is larger than a certain threshold. This threshold depends on the intra-cell routing efficiency and empty space required on M1 to access the cell I/O pins. Furthermore, routing efficiency is a function of the proportion of non-preferred direction wire length to total wire length. To capture these effects, we model track-congestion threshold as follows:

$$C_{threshold} = \alpha + \left|\frac{U_x - U_y}{U_x + U_y}\right| \times \beta - \gamma, \qquad (1)$$

where U_x and U_y are the track utilization in x and y directions, which excludes track-blockage from the orthogonal direction wiring. α and β parameters are a function of intra-cell routing efficiency and γ is a function of empty space left for inter-cell

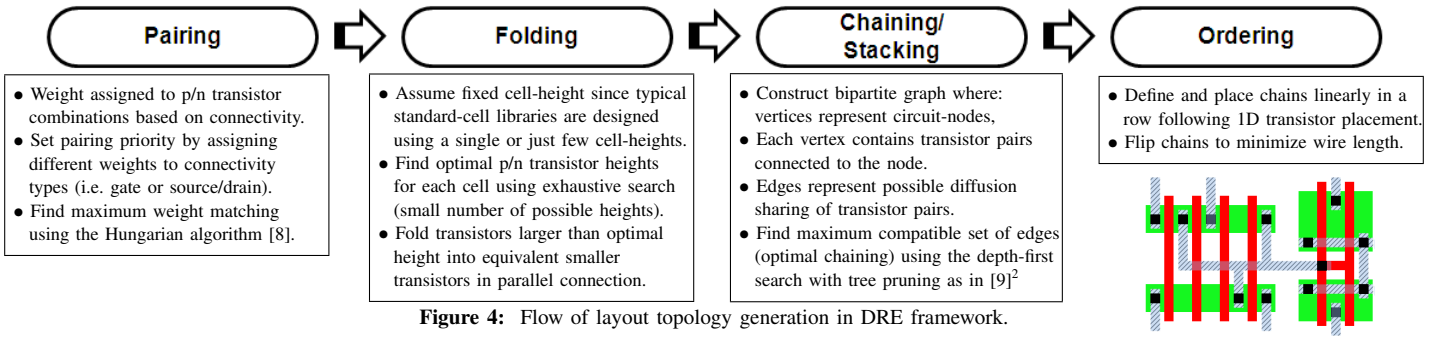


Figure 4: Flow of layout topology generation in DRE framework.

Figure 6: Single-trunk Steiner tree routing: S/D-to-gate interconnections on M1 and poly layers and S/D-to-S/D interconnections on M1 only.

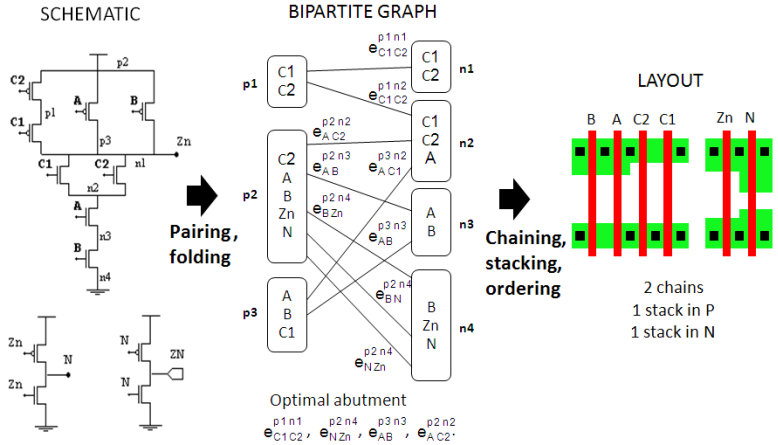


Figure 5: Example that illustrates our layout topology generation for 4-input OAI standard-cell.

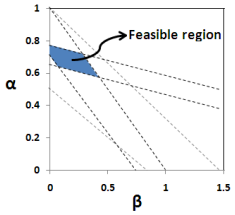


Figure 7: Every single cell implementation adds one lower and one upper bound line. Lower bound lines are defined by cell congestion. Upper bound lines are defined by cell congestion *after* area increase due to congestion (equal to 1 if area was not increased). α and β are approximated by coordinates of the feasible region's geometric centroid.

router to access cell-I/O pins. The values of all these parameters are specific to the router. Figure 7 depicts one method to extract α and β parameters either from trial routes of few cells or from cells of a previous generation library.

Layout estimation accuracy is validated in Figure 8. The source of imperfect area estimation is from using different layout styles[3].

III. Manufacturability

Our manufacturability index for evaluating DRs is the probability of survival (POS) from three major sources of failure[4]: a) contact-defectivity (a.k.a. contact-hole failure); b) overlay error (i.e. misalignment between layers) coupled with lithographic line-end shortening (a.k.a. pull-back); c) random particle defects.

POS associated with contact-hole failure is equal to the number of non-redundant contacts in the layout times contact-hole failure rate. In case of contact-redundancy, duplicated contacts are assumed to always yield since the probability for *two* contacts connected to the *same* pin to fail is negligible.

Overlay vector components in x and y directions are described by a normal distribution with zero mean and process-specific 3σ estimate. We compute POS from overlay causing: failure to connect between contact and poly/M1/diffusion, gate-to-contact short defect, and always-on device caused by poly-to-diffusion overlay error. Connection failure at contacts occurs when the area of overlap with top/bottom connecting layers is smaller than a certain threshold-value. Thus, we consider overlay in both x and y directions in this analysis. In gate-related failure analysis, overlay in just one direction is considered since gates are presumably unidirectional. Moreover, we assume all layers are aligned to a reference alignment mark on substrate[5] and overlay between different layers and the reference layer to be independent[6]. The overall POS from overlay is then calculated as the product of POS from independent overlay errors. If overlay is assumed to be completely a die-to-die variation, then POS of the die is p (equal to POS of the most overlay-critical spot in layout). On the other extreme, if overlay is completely random within-die variation, then POS of the die is p^n, where n is the total number of critical spots in the design. Reality is closer to the former situation (since field and wafer level components dominate intra-field components [12]), which is our assumption in this paper.

Critical area analysis is performed for open and short defects at M1/poly/contact layers and short defects between gates and diffusion-contacts. For fast analysis, we use the virtual artwork approach proposed in [13]. Poly and contact layers are represented by strips separated by spacing-DRs; whereas for M1 layer, this separation corresponds to the spacing that makes the wires as far apart as possible. The virtual artwork representation allows quick calculation of critical area as a function of defect size by applying a closed-form equation model. Probability of failure from random particles is then inferred from average critical area for all defect sizes and average defect density following standard methods [14] and using the defect size distribution model of [15, 16]. Overall POS (i.e. complement of probability of failure) from all sources is then calculated as the product of POS from individual sources.

IV. Variability

In sub-wavelength lithography regime, three sources of printing imperfection causing gate-dimension variation are dominant: a) diffusion and poly corner-rounding; b) line-end tapering under overlay error and line-end pull-back; c) CD variability. Figure 9 depicts these imperfections and locations at which they affect variability. The contribution of each source to gate length and width variations (ΔW and ΔL) is modeled independently. The total change in drive current is set as our variability index

[2]The algorithm is made faster by limiting the number of iterations (to six). This has negligible effect on the quality of results since the optimal solution is among the first few examined solutions in almost every case. Moreover, the algorithm is modified to favor solutions with more transistor stacks over regular chaining or vice versa.

[3]In particular, Nangate-cells are generated with preference for transistor chaining over pairing; whereas, our method enforces pairing of transistors with same gate-signal. The latter is better approach for modern 1D-poly and fixed gate-pitch implementations.

[4]More involved models of lithography induced failures are part of our ongoing work.

[5]This can be modified to conform with the process alignment strategy.

[6]In reality overlay of different layers with the reference layer have some degree of correlation. This can be dealt with by reducing the amount of overlay (i.e. use smaller 3σ for overlay distribution).

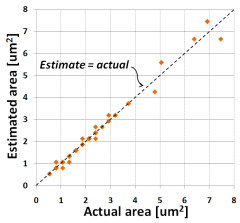


Figure 8: Accurate area estimation for the entire Nangate Open Cell Library [10] (104 cells) with 2.4% average error. Runtime of evaluation procedure is 24 minutes real time on a 2GHz clock speed and 2MB cache processor.

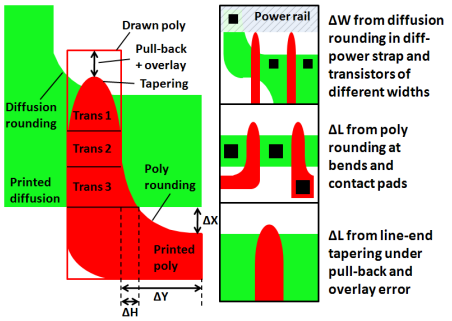


Figure 9: Slicing model, rounding model parameters, and the sources of gate length and width variability considered in DRE framework.

Table I: PROCESS CONTROL PARAMETERS WITH VALUES PROJECTED BY ITRS [11]. CRITICAL M1/POLY/CONTACT WIDTHS CORRESPOND TO TYPICAL MINIMUM ACCEPTABLE WIDTH FOR THE DEFECT NOT TO CAUSE A FAILURE.

Parameter	45nm process	65nm process
Avg defect density $[faults/m^2]$	1395	1757
Critical defect size [nm]	34	45
Max defect size [nm]	250	250
Fab cleanliness parameter	3	3
Contact-holes rate [ppm]	0.00004	0.00004
Overlay (3σ) [nm]	13	15
Line-end pull-back (mean) [nm]	10	14
Gate CDU (3σ) [nm]	2.6	3.3
Critical M1 line-width [nm]	10	15
Critical poly line-width [nm]	15	20
Critical contact-width [nm]	10	15

for evaluating and comparing DRs and is calculated using the following equation:

$$\Delta(\frac{W}{L}) = \frac{\sum_{allgates} \left|\Delta(\frac{W}{L})_i\right|}{(\frac{W_{tot}}{L})_{ideal}}, \quad (2)$$

where i represents the source of variability[7].

Since the resulting ΔW and ΔL are not across the entire gate, we quantify their contribution to $\Delta(\frac{W}{L})$ by modeling devices as parallel slices of transistors[8].

The rounding-shape is a function of corner dimensions and is modeled empirically to give $< 0.8nm$ error with measured data from printed-image simulations on a fairly wide range of practical corner-dimensions, i.e. ΔX and ΔY, depicted in Figure 9, ranging from $30 \rightarrow 70nm$ and $10 \rightarrow 200nm$ respectively.

Line-end tapered shape and gate-length at the transistor-edge are described using the model offered in [21][9] while accounting for line-end pull-back (mean value) and overlay errors (from distribution). Line-ends are assumed to extend beyond the gate as far as possible unless minimum line-end extension (LEE) rule is enforced by the user.

CD uniformity (CDU) is not directly affected by any DR, but it is well believed that layout-regularity (such as fixed gate-pitch) can considerably improve CDU [22–24].

After determining all $\Delta(\frac{W}{L})$ terms from different sources, we compute the absolute sum of all terms for the entire layout with the intention of highlighting actual gate variability. Finally, the drive current variability index is calculated using Equation 2.

V. Experimental Setup and Results

In this section, we evaluate and analyze major debatable DRs, compare DR sets from standard and low power processes, and explore M1-pitch and two gate-spacing related DRs collectively.

A. Testing Setup

We use four benchmark designs from [25] synthesized using Nangate 45nm Open Cell Library (scaled for testing with 65nm process). Designs correspond to two processing cores and two video controllers and differ by the number of cell-instances ($4358 \rightarrow 43156$) and the number of unique cell-types ($54 \rightarrow 94$).

Experiments were performed using 45nm open-source FreePDK process and 65nm process from a commercial vendor. Estimates of process control parameters associated with each process are summarized in Table I. CDU value in the table is for 2D-poly patterning. For fixed pitch 1D-poly, we use CDU 3σ improvement factor of 47% over 2D-poly reported by IBM in [23] and assume half the improvement is from unidirectional patterning.

α and β parameters of the congestion threshold model (Equation 1) are extracted from Nangate cells using the method discussed in Section II-C and γ parameter of the model is set to zero (i.e. no extra space requirement for I/O pin-access).

Since the area of the benchmark designs is relatively small, we normalize POS values to a $100mm^2$ chip-area. We determine for the base case in each experiment the number of design copies that can fit in $10 \times 10mm$ chip size with 80% cell-area utilization and find the corresponding number of contacts and critical area.

B. Evaluation of DRs and Layout Styles

Results of DR evaluations are a strong function of the base set of rules, layout styles, library architecture, and design type and, hence, they are *not generalizable*.

Three configurations of poly-patterning styles were investigated: a) no poly-routing, i.e. 1D-poly, b) limited poly-routing, and c) non-restricted poly-routing, i.e. 2D-poly. In case of 1D-poly configuration, poly is used only to connect dual gates (i.e. gates of same transistor-pair). In case of limited poly-routing, it is also used to connect adjacent gates in the same p or n network. In case of 2D-poly, it is used to perform all gate interconnections unless blocked by previous routing or diffusion power-straps.

Figure 10 shows area, manufacturability, and variability trade-offs associated with 1D/2D-poly, fixed gate-pitch, and diffusion/M1 power-straps rules for 7/9/11-track cell-heights on 45nm process. Important observations and interpretation of results are brought forward next.

1) 2D vs. 1D-poly: 1D-poly leads to much less variability compared to 2D-poly at negligible area overhead, which has two reasons. First, cells have fairly simple poly-patterning as a result of pairing transistors with same gate-signal. Second, gate-alignment requirement for 1D-poly induces negligible area overhead in FreePDK, which uses the same rule for minimum and contacted gate-pitch. Limited poly-routing and 2D-poly results are almost identical and, thus, allowing U-shape and W-shape poly-patterns with RET complications might not bring real benefits.

2) Multiple vs. fixed gate-pitch: Fixed gate-pitch implementation has large area overhead compared to 1D-poly implementation if used in combination with diffusion power-straps; however, this overhead is much smaller if fixed gate-pitch is combined with M1 power-straps. The reason for this discrepancy is that, according to FreePDK DRs, diffusion power-straps require larger separation between gates than the minimum gate spacing-rule and, hence, gate-spacing needs to increase further for gates to fall on the proper pitch. On the other hand, for the case of M1 power-straps, gate-separation is constant and only isolated-gate spacing is increased.

3) Diffusion vs. M1 power-strap: Diffusion power-straps results in much larger variability than in the case of M1 power-strap[10]. The reason for this large effect of diffusion rounding is the fact that cells are packed in the horizontal direction to minimize cell-width and minimum DRs are used. In contrast, poly-rounding and line-end tapering effects are much less important because

[7]We realize that this estimate is approximate as effects from different sources can interfere. Nevertheless, it is a good indicator of worst-case variability and process control requirement.

[8]More accurate slicing models of [17–20] can also be embedded in the framework if they are available.

[9]$L_i = 2a\left(1 - |\frac{h_i - k}{b}|^n\right)^{\frac{1}{n}}$, where l_i is the gate-length at i location in the line-end extension, h_i is the distance from i to gate-edge, a is half the nominal gate-length, b is the line-end extension, and k and n parameters describe the taper-shape. In our experiments, we use $k = 0$ and $n = 3$.

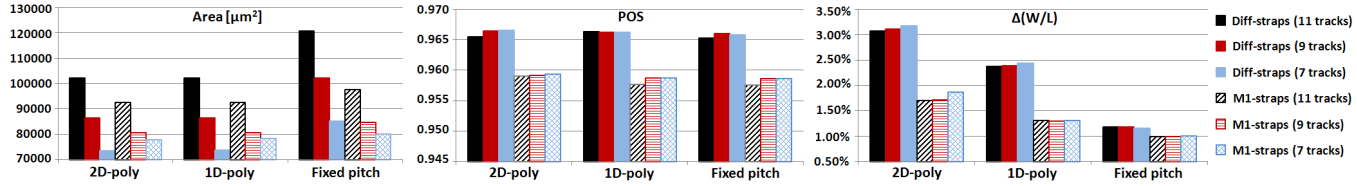


Figure 10: Evaluation of restrictive DRs on 45nm FreePDK process for 7/9/11-track cell-heights.

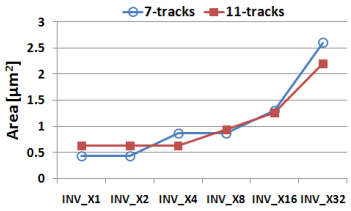


Figure 11: Increasing area with increasing transistor-width for 7/11-track cell-height.

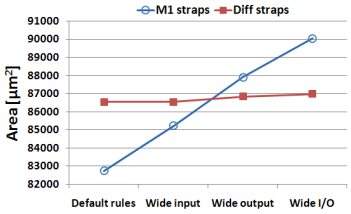


Figure 12: Cell-area for increasing pin-access requirements.

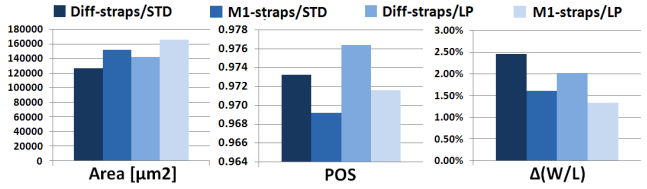


Figure 13: Restrictive DR study for commercial standard and low power 65nm-processes.

cells are normally relaxed in the vertical direction (cell-height being fixed). Furthermore, an area overhead is associated with diffusion power-strap style for some cell-height cases for the reason highlighted in Section V-B2. However for a small cell-height (7 tracks), diffusion power-straps lead to area improvement because they reduce M1-congestion, which affects cell-area seriously in this case. Besides, diffusion power-straps have some manufacturability benefits. Gate-to-contact shorts are reduced and contact redundancy for power connections is implemented at no cost since these contacts are placed on the power-rail in this case.

4) Cell-height: Results show a minor effect of cell-height decision on variability. This is because affected poly rounding and line-end tapering are second-order sources of variability as discussed earlier. The smallest cell-area of the benchmark designs is achieved with 7-track cell-height. However, this is not true for all cells as Figure 11 shows.

5) Pin-access requirement: In this study, we consider the requirements of double-sized input M1-pads, double-sized output M1-wires, and the combination of both. Results reported in Figure 12 show that pin-access requirement has negligible effect on cell-area with diffusion power-straps, but considerable effect with M1 power-straps.

C. Comparison of Different Processes and DR Exploration

Figure 13 compares DRs of a standard and a low power 65nm process from the same commercial vendor with diffusion/M1 power-strap style and 1D-poly patterning. Contrary to the previous study with 45nm process, diffusion power-straps lead to smaller cell-area in this case. This large area improvement is due to *reduced* M1-congestion and gate-spacing at power connections according to design rule values of the 65nm commercial process.

The framework is used for exploration of M1-pitch for the cases of diffusion and M1 power-straps and 1D-poly. In Figure 14, we plot normalized area and POS as a function of M1-pitch. In this exploration, we fix cell-height to 9 tracks of the first horizontal metal layer (M3) that has the same pitch as M1. Area increases linearly with M1-pitch due to increasing M1-congestion and increasing cell-height with increasing track-pitch. POS is almost unaffected by M1-pitch decision[11]. Reducing the pitch makes the layout susceptible to small defect sizes; however, it also leads to a smaller layout-area, which affects critical area in an opposite way.

In another experiment, we co-optimize gate-to-diffusion (GD) and gate-to-contact (GC) rules in 65nm process. We perform the study for all benchmark designs and use diffusion power-straps and 1D-poly patterning styles. Results are depicted in Figure 15

[10]It is important to note that area-impact of diffusion/M1 power-straps style depend strongly on design rule values as Figure 13 would show in a later study that different DRs yield completely opposite results.

[11]This is partly due to the absence of an elaborate 2D printability model in our framework.

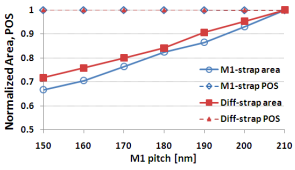


Figure 14: M1-pitch exploration for diffusion/M1 power-straps style.

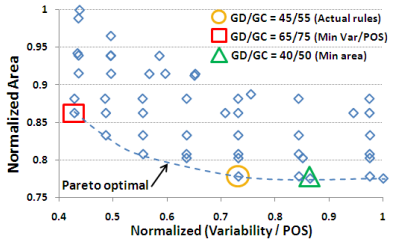


Figure 15: Co-exploration of GC/GD rules (see figure 2) for 65nm commercial process.

where each data point represents a GD/GC value. The solution corresponding to process GD/GC actual values falls exactly on the Pareto optimal frontier. Although quite simplistic, this example provides compelling evidence of our evaluation metrics fidelity and validates our approach.

VI. Conclusions and Future Work

In this paper, we illustrated potential applications of our framework for collective DR evaluation and exploration as well as comparison of DRs from different processes. To the best of our knowledge, this is the first work that includes all area/manufacturability/variability metrics in the evaluation. Nevertheless, this is just the first step and our ongoing work pursues the following directions: a) addressing design rule effects on other layout and circuit characteristics including performance, power, reliability, and some notion of designability; b) introducing a 2D printability model (not based on field simulation), for example, derived from [26, 27]; c) extrapolating DR evaluation to the chip level and include intermediate and global metal/via layers; d) studying interactions and tradeoffs of variability and area, as in [28] for example.

References

[1] L. Capodieci *et al.*, in *Proc. IEEE/ACM DAC'04*, 2004, pp. 311–316.
[2] V. Dai *et al.*, in *Proc. SPIE*, vol. 7275, 2009, p. 727517.
[3] S. Chang *et al.*, in *Proc. SPIE*, vol. 7275, 2009, p. 72750D.
[4] A. R. Subramaniam *et al.*, in *Proc. ASP-DAC'08*, 2008, pp. 474–479.
[5] S. Kobayashi *et al.*, in *Proc. SPIE*, vol. 7028, 2008, p. 70280O.
[6] L. Liebmann *et al.*, in *Proc. SPIE*, vol. 7275, 2009, p. 72750A.
[7] J. Yang *et al.*, in *Proc. SPIE*, vol. 6156, 2006, p. 61560A.
[8] J. P. S. Kuhn, *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.
[9] C. Hwang *et al.*, *IEEE Trans. on CAD*, vol. 9, no. 7, pp. 781–786, 1990.
[10] [Online]. Available: http://www.si2.org/openeda.si2.org/projects/nangatelib
[11] International Technology Roadmap for Semiconductors, Report 2007.
[12] B. Eichelberger *et al.*, in *Proc. SPIE*, vol. 6924, 2008, p. 69244C.
[13] W. Maly, *IEEE Trans. on CAD*, vol. CAD-3, no. 3, 1985.
[14] I. Koren *et al.*, in *Proc. of the IEEE*, vol. 86, no. 9, 1998, pp. 1819–1837.
[15] d. G. J. P., in *Proc. IEEE/ACM SLIP'01*, 2001, pp. 135–163.
[16] C. H. Stapper, *IBM J. of R&D*, vol. CAD-3, no. 3, pp. 549–557, 1983.
[17] P. Gupta *et al.*, in *Proc. ASP-DAC'08*, 2008, pp. 480–485.
[18] R. Singhal *et al.*, in *Proc. of DAC'07*, 2007, pp. 823–828.
[19] P. Gupta *et al.*, in *Proc. SPIE*, vol. 6156, 2006, p. 61560T.
[20] S. X. Shi *et al.*, in *Proc. IEEE/ACM ICCAD'06*, 2006.
[21] P. Gupta *et al.*, in *Proc. SPIE*, vol. 7028, 2008, p. 70283A.
[22] M. C. Smayling *et al.*, in *Proc. SPIE*, vol. 6925, 2008, p. 69250B.
[23] L. W. Liebmann *et al.*, in *Proc. SPIE*, vol. 5379, 2004, pp. 20–29.
[24] L. Pileggi *et al.*, in *Proc. IEEE/ACM Design Automation Conference DAC'03*, 2003.
[25] [Online]. Available: http://www.opencores.org/
[26] A. B. Kahng *et al.*, in *Proc. SPIE*, vol. 6349, 2006, p. 63490H.
[27] M. Cho *et al.*, in *Proc. Design Automation Conference DAC'08*, 2008, pp. 504–509.
[28] K. Jeong *et al.*, in *Proc. ISQED'08*, 2008, pp. 790–797.