MANUFACTURABILITY AWARE DESIGN

by

Jie Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2007

Doctoral Comittee:

Professor Dennis M. Sylvester, Chair

Professor David Blaauw

Professor Kensall D. Wise

Luigi Capodieci, Advanced Micro Devices

ACKNOWLEDGEMENTS

My sincere gratitude to Professor Dennis Sylvester for providing me the opportunity to study at the University of Michigan. Without his vision, encouragement and continuous support, this dissertation was impossible. Professor Sylvester was always around to listen and to give advice. He not only taught me how to ask questions and express my ideas, also taught me to be a careful and meticulous researcher. He always provided timely responses for proof reading and commenting my papers and chapters, extremely patient even after I had made many revisions.

A special thank goes to Dr. Luigi Capodieci, who has always been very supportive and responsible for helping me complete the challenging research that lies behind this dissertation. He was always there listening to my ideas and giving me feedbacks and guidelines that help me go through my problems at many critical moments.

I'm also deeply indebted to the rest of my thesis committee: Professor David Blaauw, Professor Ken Wise, for their insightful comments, hard questions, constructive suggestions and numerous encouragement.

During the course of this work, I was part of the GSRC and SRC program. I would like to thank the sponsors for their continuous support of the projects.

A warm hug to my research fellows for their very valuable inputs and contributions for various projects: Professor Yu (Kevin) Cao, Professor Puneet Gupta, and Professor Andrew Kahng. Thanks also go to my colleagues and friends at VLSI Design / Automation lab and UM for sharing both research and life experiences:

Kanak Agarwal, Himanshu Kaul, Harmander Deogun, Matthew Guthaus, Ashish Srivastava, Youngmin Kim, Aseem Agarwal, Donwoo Lee, Feng Gao, Feng Wang, Hui Zhang, Yunqing Chen, Yong Lei, Jun Cheng, Li Ding, Yan Lan and Cheng Peng. They have made my life at UM very pleasant and memorable:

Last, but not least, my utmost thanks to my family: my parents Yinxiang Yang and Xiumei Zhang, for giving me life in the first place, for educating me with aspects from both arts and sciences, for teaching me how to be a good person and for their unconditional support and encouragement to pursue my interests. Their patient love enabled me to complete this work.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Litho Process and Challenges

Since the invention of the integrated circuit (IC) in 1958, Moore's law has described the unprecedented march of innovation as chip features have become exponentially smaller and the number of transistors exponentially larger. The trend has continued for half a century and is not expected to stop for a decade at least and perhaps much longer. Today the most advanced circuits contain several hundred million components on an area no larger than a fingernail. While the invention and innovation of process steps has been the fundamental enabler of Moore's law's continuity, photolithography has been pushed the most in recent technology generations. It is worthwhile to mention the famous Rayleigh equation that governs optical lithography:

$$R = k_1 \times \lambda / NA \tag{1.1}$$

In Equation 1.1, R is the minimum feature width that we want to resolve. k_1 is a process dependent adjustment factor. λ is the wavelength of light that is used to pattern an integrated circuit layout onto silicon. The opening angle of a lens that is used to project a mask or reticle is defined by the numerical aperture (NA)

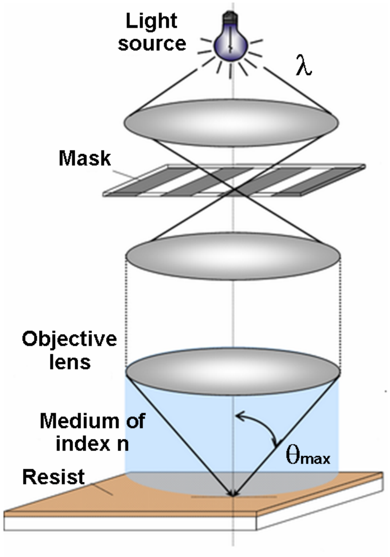


Figure 1.1: $NA = n \times sin(\theta_{max})$, where n is the refractive index of the imaging medium. When the imaging medium is air, n=1.0.

(see Figure 1.1 for detailed definitions). For earlier process technologies, the ratio of target resolution to optical resolution, measured at λ/n, has remained above, or at least at, unity. The wavelength of light has been changed at least four times, ranging from Hg-arc lamps to excimer laser-based systems, and is now operating in DUV (Deep-UV) at 193 nm. As indicated in Figure 1.2, starting at the 350 nm generation, engineers face the challenges of dealing with features printed at less than λ, with k_1 factors approaching 0.3 or even below - where k_1=0.25 is the theoretical lower limit for single exposure optical lithography. In this type of *sub-wavelength lithography*, the projecting light passing through a diffraction limited system results in severe distortion of patterns printed on silicon compared to those created in the semiconductor physical design process. As illustrated in Figure 1.3, distortion effects impact pattern fidelity and edge placement on silicon, and in the worst case can even eliminate patterns entirely due to contrast zones. Figure 1.4 shows the improved pattern fidelity with a commonly used resolution enhancement technologies (RETs)

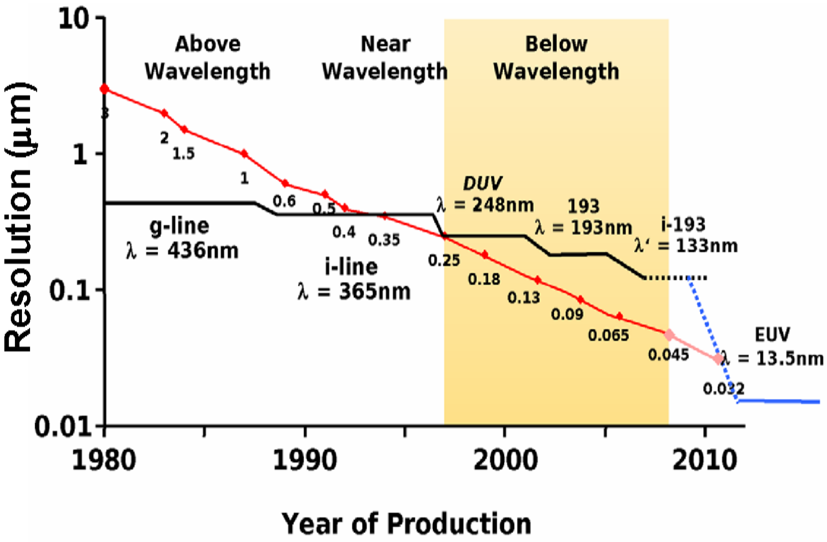


Figure 1.2: Roadmap of wavelength vs. feature size.

- OPC (optical proximity correction) [1].

Besides the difficulties of maintaining feature fidelity, future technology nodes are expected to see increased process variation and decreased predictability of nanometer-scale circuit performance [2]. Despite the relaxation of some 3σ tolerances, there are no known solutions for a number of near-term variability control requirements according to the ITRS (International Technology Roadmap for Semiconductors) [2]. Moreover, observation of key markets that drive the semiconductor industry reveals the potentially large impact of variability on the value of semiconductor products. Semiconductor enterprises must be cognizant of the different risks and ROI (Return on Investment) opportunities from, e.g., an extra increment of T_{ox} or L_{eff} CD (Critical Dimension) control (see Table 1.1), versus new design for value technologies, versus revised performance targets for products, etc. All of these have brought about the need for correction techniques to enhance resolution and avoid unacceptably high circuit and critical path performance variation [3]. Resolution enhancement techniques (RETs) that address three degrees of freedom in lithography, *aperture*,

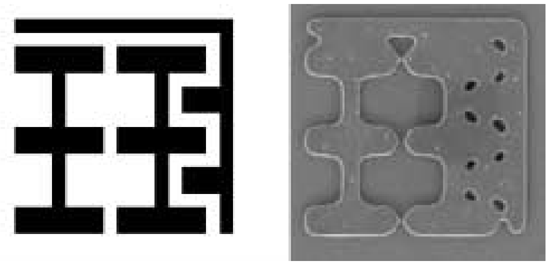


Figure 1.3: Ideal layout (left) can be significantly distarted after wafer processing and etching (right) [1].

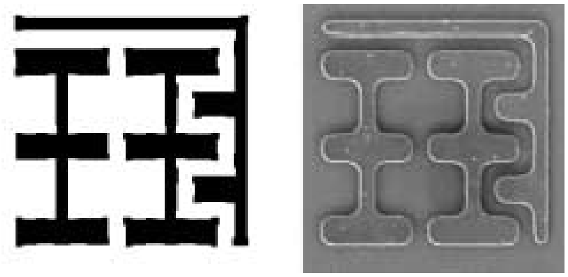


Figure 1.4: Layout modified with OPC to "pre-compensate" for process distortions [1].

phase, and/or *pattern uniformity*, are increasingly adopted in nanometer-scale design (i.e., 130 nm processes and beyond) with respect to not only the number of mask levels incorporating RETs but also the variety of techniques applied.

Due to the technological challenges of controllably printing very small features, the non-recurring engineering (NRE) and turn-around time (TAT) costs of correction (optical proximity correction (OPC), phase-shifting, dummy features) are very high in terms of design time and mask yield/verification. Many costs (yield, mask writing time, data volume, etc.) are directly proportional to the complexity of the shapes needed on the masks (e.g., Figure 1.5) shows the mask data volume for a 45 nm design is projected to be 33X larger than for a 180 nm design). Mask writing time has increased from just a few days to over a month due to RET complexity [5]. This brings up an important relationship between design and lithography costs, namely

Table 1.1: The ITRS requirement of gate dimension variation control is becoming more stringent as the technology scales [2]. MPU, micro-processor unit.

Year	2005	2007	2010	2013
Technology Node	90 nm	65 nm	45 nm	32 nm
MPU gate length	32 nm	25 nm	18 nm	13 nm
MPU Gate CD 3σ	3.3 nm	2.6 nm	1.9 nm	1.3 nm

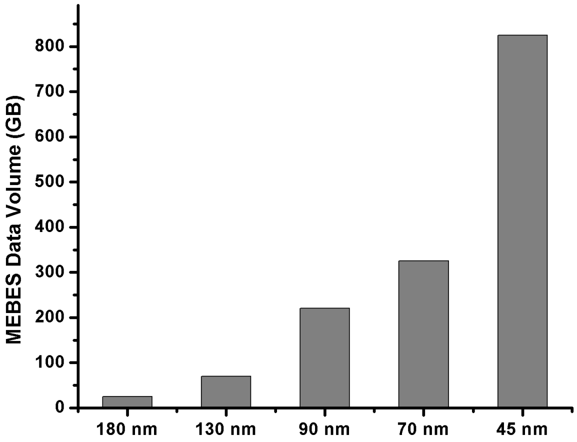


Figure 1.5: Mask data volume as the technology scales [4].

that the total cost to produce low-volume parts (such as most ASIC designs) is dominated by mask costs [6].

1.2 Various RETs

Figure 1.6 showed four basic properties: wavelength, amplitude, phase, and direction. RET (resolution enhancement technology) is wavefront engineering to enhance

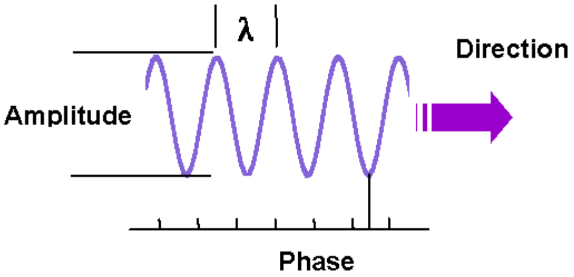


Figure 1.6: Properties of a wave: wavelength, amplitude, phase, and direction.

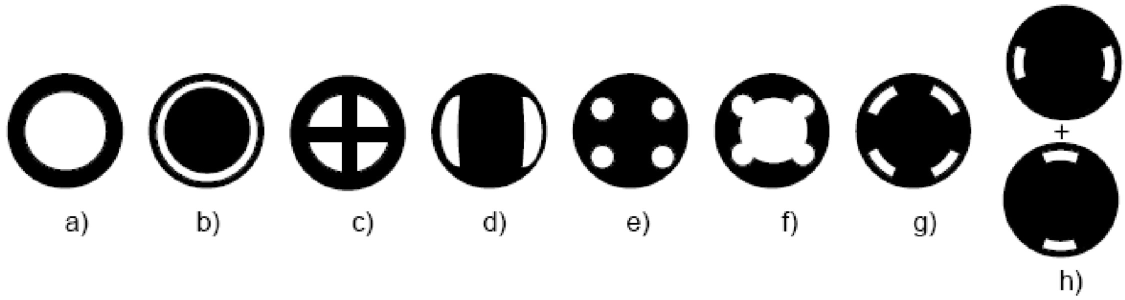


Figure 1.7: Idealized pupil maps for various a) conventional and off-axis illumination schemes: b) annular; c) “fourfold” source; d) “separated” source; e) quadrupole; f) “CAUEST”; g) quasar; and h) dipole.

lithography by controlling these levers.

- *Wavelength.* In lithography, wavelength is set by the choice of light source, and it has been significantly shifted from 365 nm to 193 nm to achieve smaller resolutions since the 80’s. However, if λ has to be further reduced to go beyond the limits of resolution, light is increasingly absorbed by practically any material so that it is impossible to build a refractive optics for small wavelength. It is expected that with immersion tools operating for instance at 193 nm, a resolution of 50 nm and below can be achieved.

- *Direction.* Different illumination types (see Figure 1.7) are used to change the direction of the wavefront to achieve the ultimate resolution of k_1=0.25. These techniques are referred to as OAI (off-axis illumination) and the principles are illustrated in Figure 1.8. As the resolution on the mask becomes small, the diffracted angle resulting from the first-order light becomes so large that it is beyond the acceptance of the exposure lens making the image contrast zero. OAI improves the contrast of the image by transmitting more of the diffracted orders through the lens. The angle of OAI is a function of feature pitch, which leads to a well known phenomenon that the lithographic benefit afforded by OAI erodes as feature pitches vary from the one that the illumination angle has been optimized for. To prevent this loss of process window, dummy features (or sub-resolution assist features (SRAFs)) are added to the layout to lithographically emulate the primary pitch as mentioned in the next paragraph.

- *Amplitude.* Amplitude control comes from changing the shapes of the geometry openings. This technique is known as OPC (optical proximity correction), which include two types: rule-based OPC and model-based OPC. Rule-based OPC applies corrections to the mask based on a predetermined set of rules. Originally only rule-based OPC was needed, e.g., iso-dense biasing, line end extensions or serif additions, and SRAFs insertion (see Figure 1.9). It worked well for simple and moderate needs. From 130 nm technology onwards, rule-based OPC has given way to model-based OPC which uses process simulation to determine corrections to the masks. These corrections, generated in accordance with the results of these simulations, generally provide more accuracy and leads to higher yield at the expense of higher cost and longer run time. Most of the time, a hybrid flow of rule-based OPC followed by model-based OPC is adopted for best performance (illustrated in Figure 1.10).

- *Phase.* Phase shifting masks (PSMs) [8] are used to create interference fringes on the wafer that boost contrast, enabling extremely small features. It has been demonstrated for a variety of applications, and has been used to manufacture commercial devices. Many techniques for phase shifting and phase assignment algorithms have been demonstrated [9]. This technique, however, often results in conflicting phase assignment on layout regions, which can potentially be

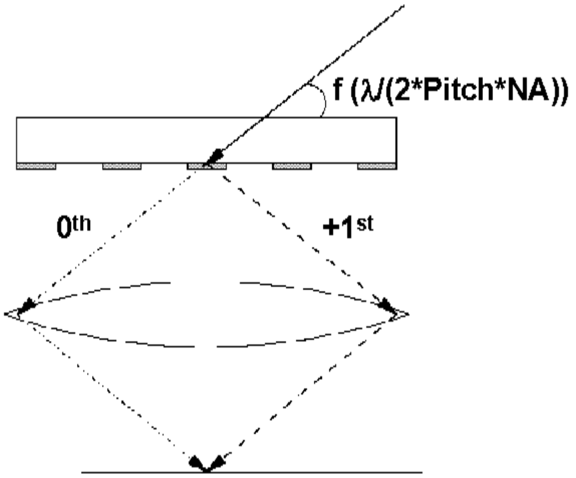


Figure 1.8: Principle of OAI [7].

resolved by enforcing radically restricted design rules (RDRs).

1.3 Modeling Process Variation in Design

Although various RETs have been deployed to enhance feature resolution and improve chip yield, these modifications, significantly affected chip performance. There have been efforts to consider process variation of parameters such as gate CD, oxide thickness, metal width and thickness, temperature, voltage, etc., during circuit performance analysis at the design stage but this requires proper modeling of the variabilities [10; 11; 12]. The most common approach to modeling variability, typically aimed at speed/frequency prediction, is based on "worst case scenarios" (corner cases). This approach assumes all parameters are independent and hence yields overly pessimistic simulations, making the design unnecessarily difficult [13; 14]. Other approaches seek to provide more accurate variability modeling by considering its sources in more detail. Treated statistically, process variations are modeled using a probabilistic framework with effort to accurately model correlations [14; 15]. More than 50% of L_{gate} variation is due to systematic sources [16], which can be modeled accurately once the physical layout is completed. In fact, assumptions about the L_{gate} distribution in Monte Carlo simulations and statistical timing analysis in general, could be made more rigorous by considering realistic systematic contributions (the majority of which arises due to proximity effects) to the overall process variation [51; 17; 18]. With measured data, [51] show a significant systematic intra-chip variability of L_{gate} leading to large circuit path delay variation. That work estimates the location-dependent L_{gate} variation by classifying the layout patterns within 5 categories; in actuality there are many more relevant scenarios, complicating the approach. References [17; 18] use aerial image process simulations to account for systematic L_{gate} variations; however, these simulations are limited to fixed layout patterns and cannot be expanded to full-chip timing analysis. In [19] the authors propose a systematic variation-aware static timing methodology using library-based OPC. However, at the full-chip level OPC features applied to a given library cell will not be identical across all instances and depend largely on neighboring geometries. This approach therefore loses the advantage of model-based OPC, particularly considering those cells with small sizes (in which OPC applied throughout the cell can easily be impacted by neighboring patterns) and high frequency of occurrence on critical paths (e.g., inverters, NANDs) sacrificing the accuracy claimed in their static timing analysis. It is necessary to investigate the impact of modifications on post-signoff designs (especially OPC) given that new steps emerging in advance technologies will notably change performance specs and therefore should be considered in the pre-signoff design flow.

Although RETs have historically been strictly a post-layout procedure, they now need to become part of a cohesive design flow in which libraries and layouts are optimized directly based on conflicts discovered by the RET tool [20]. This “trickle-

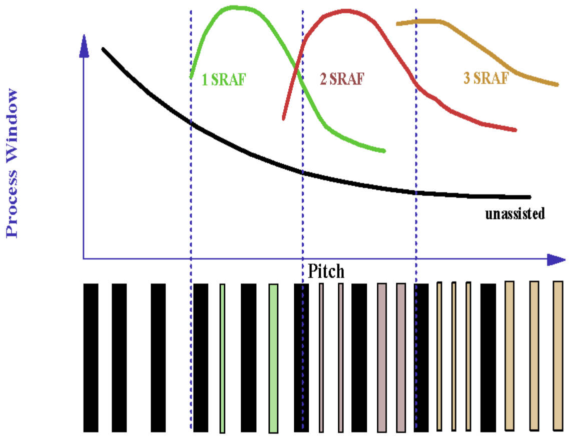


Figure 1.9: SRAFs help maintain process windows for certain pitch ranges.

down" effect of RETs towards the design process also yields more conservative design rules, particularly for the critical polysilicon layer. In particular, the ability to print very tight pitches as well as print a wide range of pitches in a given layer is very difficult for subwavelength lithographic systems. As a result, there is a trend towards limiting the range of allowed pitches in the polysilicon layer [21]. This type of restricted design rule (RDR) seeks to enforce a particular style of layout that is known to be highly manufacturable. As with any design rule, it is a tradeoff between manufacturability and performance, where performance can be measured as layout density, delay, power, etc. By nature, these RDRs seek to push the tradeoff more in favor of the manufacturing side, sacrificing performance in the process. Despite the move towards RDRs, there has been no comprehensive and systematic study of their expected impact on manufacturability and performance. On the other hand, RDRs may over constrain certain design layout and sacrifice the benefits achievable by technology scaling. The use of flexible design rules (FDRs - this approach starts with a RDR-compliant layout and relaxes the rules to recover area in the printability sweet spots, as well as tightens the rules in the hotspot regions) should be considered when RDRs alone are not sufficient to guarantee high yield or waste too much chip

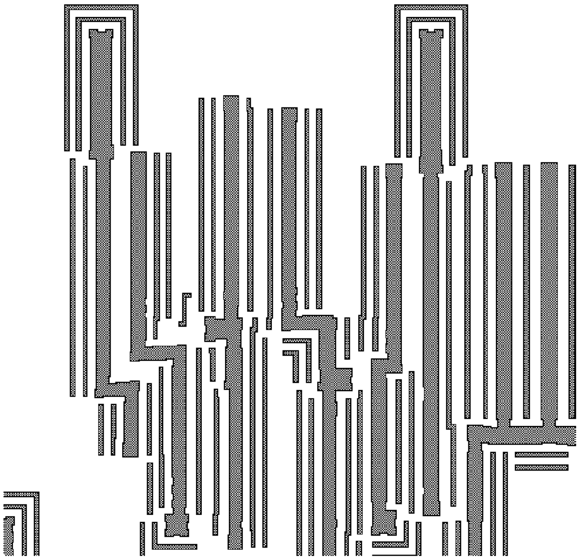


Figure 1.10: Model-based OPC on SRAF layout.

area.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter II, we first establish a framework for assessing the impact of process variation on circuit performance, product value and return on investment across various technologies. Elements of our framework include accurate device models and circuit simulation, along with Monte-Carlo analyses, to estimate parametric yields. We evaluate the merits of considering such previously unconsidered phenomena as correlations among process parameters. We also evaluate the impact of process variation with respect to such relevant metrics as *parametric yield at selling point*, amount of *required design guardbanding*, and *inferred process variation control*. Performance variation trends along with technology scaling in terms of the degree of control level in process variation and changes in number of critical paths is also presented.

Using the framework mentioned above, gate CD variation is identified as the most important parameter on which performance shows most sensitivity. To address this effect and to meet the stringent ITRS requirement of gate CD 3σ control while obtaining the lowest cost of ownership (CoO), Chapter III describes a novel minimum cost of correction (MinCorr) methodology. This approach determines the level of correction of each layout feature such that prescribed parametric yield is attained with minimum RET cost. This flow is implemented with model-based OPC explicitly driven by timing constraints. We apply a mathematical programming-based slack budgeting algorithm to determine OPC level for all polysilicon gate geometries. Designs using this methodology show up to 20% MEBES data volume reduction and 39% OPC runtime improvement.

Chapter IV uses state-of-art process control techniques to analyze the impact of systematic correction residual errors on a microprocessor's speedpath skew. A platform is created for diagnosing and improving OPC quality on gates with specific functionality such as critical gates or matching transistors. With the more accurate timing analysis we highlight the necessity of a *post-OPC verification* embedded design flow.

After this in Chapter V and Chapter VI, we show a framework to quantify the performance, manufacturability and mask cost impact of globally applying several common restrictive design rules (RDRs) to reduce pitch size induced process variations. At the end, a practical hybrid method for adapting restricted design rules and flexible design rules based on pattern matching are presented. In this way, we may greatly leverage learning from manufacture side to best use the design space for yield enhancement and performance improvement.

Finally, conclusions and directions for future work are given in Chapter VII.

CHAPTER II

Design Sensitivities to Variability

2.1 Introduction

Aggressive technology scaling has introduced new variation sources and made process variation control more difficult. As a result, semiconductor manufacturing equipment will be strained to maintain constant process variation levels in future technology nodes. Despite the relaxation of some 3σ tolerances, there are no known solutions for a number of near-term variability control requirements (according to the ITRS [2]). Moreover, observation of key markets that drive the semiconductor industry reveals the potentially large impact of variability on the value of semiconductor products. Semiconductor enterprises must be cognizant of the different risks and ROI opportunities from, e.g., an extra increment of T_{ox} or L_{eff} CD control, versus new design for value design technologies, versus revised performance targets for products, etc. In this chapter, we describe key elements of a framework that will allow the semiconductor industry to assess the impact of process variation on circuit performance, manufacturing cost, and product value in nanometer technologies (130nm through 70nm). Our framework is built on accurate circuit design models, statistical models of process variation, a combination of circuit simulators and analytical performance models, and application of Monte Carlo analyses to estimate parametric yields. We evaluate the merits of taking into account previously unconsidered phenomena such as correlations among process parameters. We also evaluate the impact of process variation with respect to such metrics as *parametric yield at selling point*, amount of *required design guardbanding*, and *inferred process control* for desired design guardbanding. Key contributions of our work include:

- a more comprehensive modeling of process variation according to the causing sources, with different handling of die-to-die (D2D) and within-die (WID) variations;
- accurate models of *correlations* of variation;
- realistic and quantified projection to future process nodes of the impact of variability on critical-path delays; and
- analysis of the sensitivity of performance variation to improved control of individual device parameters and variation sources, measured by change in the number of "sellable" chips produced, extent of guardbanding required to meet a given parametric yield target, and inferred process variation control for the desired design guardbanding.

Our experimental results yield surprising insights into the scaling of process variation impacts through the next two ITRS technology nodes, as well as the prioritization of various areas for future technology investment. The latter type of contribution, even if not fully achieved by this initial work, is required for principled allocation of R&D resources among multiple semiconductor supplier industries to solve the variability problem (cf. "shared red bricks" [22]) [1].

[1] Sources at a major semiconductor vendor indicated that substantial effort and capital has been invested for V_{th} control because of its huge impact on design performance. This anecdotal evidence

2.2 Taxonomy of Variation

Circuit variability refers to deviations of either process or circuit parameters (e.g., L_{eff}, V_{dd}, conductor thickness, etc.) from nominal values. It is introduced either during chip fabrication or due to circuit operation. Based on the inherent spatial scales of such variability, it is often characterized as either within-die or die-to-die variation. In this study, we consider the impact of both types of variations. The taxonomy of variability used in our framework is as follows.

- *D2D Variation.* D2D variation affects each element on a chip equally and adds a random effect across the wafer. This variation determines the nominal value of each parameter on the die with these values differing among chips across the wafer and from wafer to wafer. D2D variation is estimated to comprise approximately 50% of the total CD variance for today's technology generations [30]. It is mostly design-independent and is related to equipment properties, wafer placement, processing temperatures, etc. [31]. For simplicity, we only model D2D variations due to random effects (i.e., we ignore *systematic* D2D effects that can be predicted, modeled, and designed around during the design process).

- *WID Variation.* WID variation can be divided into two contributors, *systematic* and *random.*

  - *Systematic WID variation.* In contrast to D2D variation, systematic WID variation is layout-dependent and may cause chip malfunction. It is predictable in the sense that a given pattern yields the same characteristics on

supports our claim; that significant engineering effort and capital investment can be expended to reduce many sources of variability but we would like the investment to be directed to the key items.

all dies. Successful scaling of MOSFET technology to sub-100nm process geometries relies on compensation of systematic variation components at the design and reticle stages. Such corrections to systematic process variation can be applied to reduce the systematic WID variation but may not completely eliminate it. For example, the same optical proximity corrections applied to identical NAND gates may not result in identical physical dimensions due to the impact of different local feature densities.

– *Random WID variation.* Random WID variation is due to the inherent unpredictability of the semiconductor fabrication process. Fluctuations in channel doping, gate oxide thickness, and ILD permittivity are primarily due to random variation. This type of variation is likely to have spatial correlation, making nearby devices more similar than ones that are across the die from one another. As random phenomena cannot be compensated for and are difficult to minimize, this type of variability may eventually pose the most significant challenge to design of adequately yielding nanometer-scale MOSFET circuits.

2.3 Experimental Testbed and Methodology

In this section, we describe elements of our experimental testbed. The key components are

1. a parameterized scalable single critical path circuit model, and a multi-critical path model composed of a user-selectable number of independent single critical paths;

2. consideration of correlations among the parameter variations;

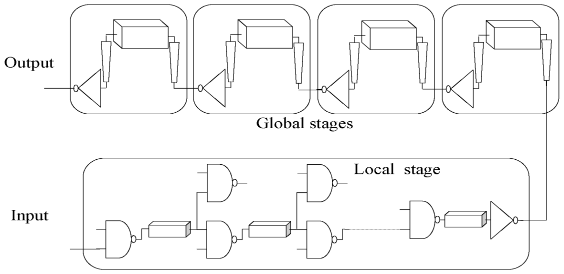


Figure 2.1: A single critical path structure for performance study.

Table 2.1: Parameter values and 3σ variations

Technology	180 nm		130 nm		100 nm		70 nm	
Device	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS
L_{eff}(nm)	100±16.7%	100±16.7%	65±16.7%	65±16.7%	45±16.7%	45±16.7%	28±16.7%	28±16.7%
T_{ox}(A)	22±4%	22±4%	16±4%	16±4%	13±%4	13±%4	10±4%	10±4%
$R_{dsw}(\Omega - \mu m)$	250±10%	450±10%	200±10%	400±10%	180±10%	300±10%	150±10%	280±10%
X_t (nm)	30		24		20		13	
V_{th}(V)	0.214±10%	-0.327±10%	0.232±10%	-0.273±10%	0.217±10%	-0.254±10%	0.169±10%	-0.218±11%
Interconnect	Local	Global	Local	Global	Local	Global	Local	Global
ε	3.5±3%		3.2±5%		2.8±5%		2.2±5%	
w(nm)	250±20%	525±20%	175±20%	335±20%	123±20%	237±20%	85±20%	160±20%
s(nm)	250±20%	525±20%	175±20%	335±20%	123±20%	237±20%	85±20%	160±20%
t(nm)	500±10%	1050±10%	280±10%	670±10%	197±10%	498±10%	145±10%	325±15%
h(nm)	500±15%	1050±15%	280±15%	670±15%	197±15%	498±15%	145±15%	325±15%
ρ (Ω-m)	2.2e-8±30%		2.2e-8±30%		2.2e-8±30%		2.2e-8±30%	
Rvia(Ω)	23±20%		25±20%2		27±20%		29±30%	
Length(μm)	62.5	5000	55.56	3333	50	2500	45.45	2000
Wn(μm)	1.26	15	1.008	9.75	0.756	6.75	0.479	4.2
Dynamic								
V_{dd}(V)	1.8±10%		1.2±10%		1.0±10%		0.9±10%	
Tr(ps)	160		125		110		87	
Temp (^{o}C)	25		25		25		25	

3. comprehensive and physically sensible handling of variation with respect to its underlying sources; and

4. use of detailed device modeling and circuit simulation within a Monte-Carlo methodology.

We start from the study of a single parameterized critical path, illustrated in Fig. 2.1. It is composed of l identical local stages and one long top-level buffered global interconnect. The parameter l is set to 10 at the 130nm technology node and is reduced by one in each subsequent generation to reflect aggressive pipelining techniques and other micro-architectural advancements. In each local stage, a 2-input NAND gate drives a short local line with length estimated by [2]. The NAND is sized to optimize the speed-power tradeoff (fanout=2), i.e., the knee of the delay vs.

sizing curve as in the Berkeley Advanced Chip Performance Calculator (BACPAC) [23]. The global line length remains constant at 10mm in all technology nodes, consistent with the 2001 ITRS projections of fixed die size for future microprocessors [2]. Optimal inverting repeaters are inserted at even intervals into the global line to minimize delay. Parasitic via resistance is considered. Overall, we closely follow the 2001 ITRS (high-performance MPU) critical path model [2]. For each local stage and the global line, we add two quiet parallel neighboring lines to provide a more realistic capacitance environment. Each line is modeled as a sufficiently long chain of L segments to capture the distributed RLC characteristics. We then combine n such identical single critical paths to capture the impact on chips with multiple critical paths. Input transition times to initial stages are set at 20% of the clock period. Further details of the critical path models, including line lengths, gate sizes, and signal transition times, are listed in Table 2.1. The nominal dimensions of interconnect are taken from [24]. Note that the critical path structure is meant to include sources of delay such as local computation and global communication that are important in generic paths of the today's very large scale integrated circuits. Hence, the path delay is not meant for cycle time estimation so the absolute value is not relevant.

Previous work assumes that variation sources are either independent of each other [3] [25] or perfectly correlated [25]. By contrast, our work recognizes several strong correlations. Specifics include:

- V_{th} is a function of T_{ox}, N_{ch}, L_{eff} and X_t, calculated from a delta-doping approximation and BSIM3v3 models. Here, X_t is the retrograde channel depth and N_{ch} is the effective channel doping [26].

- Corresponding parameters of NMOS and PMOS have a correlation coefficient of one, i.e., NMOS and PMOS in the same gate exhibit the same deviations from respective means. This pertains to parameters that are shared among the two device types, such as drawn channel length and oxide thickness, but is not applicable to steps such as channel doping that are independent for NMOS and PMOS.

- Assuming a fixed wire pitch, wire spacing variation is the negative of wire width variation. Metal thickness (T) and underlying interlevel dielectric (ILD) thickness (H) are negatively correlated with a correlation coefficient of -1 (this value stems from the relationship of trench etch depth in damascene processes as well as chemical-mechanical polishing effects which act to reduce the correlation between T and H).

- Spatially proximate devices and interconnections (e.g., in local stages) have similar variations.

- We model the spatial correlation among repeaters inserted along the global line by incorporating a distance-dependent correlation parameter. This correlation decays linearly with distance to a value of zero (implying complete independence) over a length scale of 1cm [27]. This value of 1cm is parameterizable.

- Our interconnect spatial correlation modeling is more involved. We divide the global line into $100\mu m$ segments. Interconnect parameters within each segment are perfectly correlated. We assume that correlation between segments decays linearly with separation. At a certain distance, this correlation equals zero. For interconnect width and space, we take this separation distance to be 5mm for all the technology nodes [28] while it is 2mm for metal thickness and ILD thickness [29]. In contrast to line width which is set by the dielectric etching process, ILD thickness and metal thickness is largely set by the CMP step in damascene processes. Numerous prior studies have investigated the concept of CMP planarization length; this relates to the distances over which features can be considered to be correlated due to pad deformation and other physical phenomena. This planarization length is typically found to be on the order of 2mm, motivating our choice of separation distance. This, in contrast to an interconnect model with perfect correlation, avoids the overestimation of interconnect variation.

- WID and D2D variation are assumed to be equal in magnitude [30]. Systematic WID variation equally affects all critical paths on the same die, while random WID variation adds random effects to the deterministic systematic WID variation.

We assume parameter variations to be normally distributed with mean and σ values derived from [2], [24], and industry sources. In [2], the allowed variability in physical gate length is fixed at 10%. The magnitude of the physical gate length is approximately half of the technology node, or the DRAM half-pitch. Translating this uncertainty to effective channel length, which is also a fraction of physical gate length due to source-drain extension (SDE) underdiffusion, we expect a 3σ for L_{eff} of greater than 10%. In this work, we approximate $L_{eff} = 0.6 \times L_{physical}$, leading to a 3σ process tolerance throughout the roadmap of 16.7%. Different approaches may be taken including an assumption that the SDE underdiffusion has a fundamental lower limit that pushes L_{eff} to be a smaller fraction of $L_{physical}$. This will result in either (1) larger uncertainty in L_{eff} or (2) less aggressive scaling of $L_{physical}$ to compensate.

Either of these alternatives can be readily investigated in our framework.

We model variabilities caused by three types of sources: systematic WID, random WID and random D2D, in our Monte-Carlo analysis based framework. A prespecified number (n) of nominally identical independent critical paths are considered (note that these critical paths are only independent with respect to WID variation; their D2D components are identical). Systematic WID variation, which is assumed to be due to layout pattern dependence, is applied across different critical paths. We expect different dies to have the same distribution of systematic WID variation. Therefore, systematic WID variation is used to shift the nominal value of the parameters for different critical paths, i.e., systematic WID variation is modeled by generating n samples from a Gaussian $N(0, \sigma_{SYS-WID})$ distribution *before* running the Monte-Carlo simulations.[2] Random WID variation is modeled as a Gaussian $N(0, \sigma_{RAN-WID})$ random variable. It observes spatial correlation at the die-scale as previously described. Random D2D variation is modeled as a Gaussian $N(0, \sigma_{RAN-D2D})$ random variable. All the variations observe the *w*-*s* and *t*-*h* correlations outlined previously. Hence, the value (X) for a given parameter for a device i in path j in the k^{th} Monte-Carlo run is given by Equation 2.1.

$$X = \mu + x^{j}_{SYS-WID} + x^{i,j,k}_{RAN-WID} + x^{k}_{RAN-D2D} \tag{2.1}$$

$$\forall\, 1 \geq j \geq n,\ 1 \geq k \geq m$$

where μ is the nominal value of the parameter, $x^{j}_{SYS-WID}$ is the systematic variation sample corresponding to path j and $x^{i,j,k}_{RAN-WID}$ ($x^{k}_{RAN-D2D}$) is the k^{th} sample for random WID (D2D) variation for device i in path j. Moreover, all variations are

[2] Since the significance of systematic WID variation lies in introducing mismatch in *pathdelays* rather than individual device delays, we do not consider systematic WID variation within a path.

Table 2.2: Comparison of RLC model with perfect correlations, spatial correlations and no correlations for 100 nm technology node

Delay (ps)	Mean	3σ	Normalized $\frac{3\sigma}{mean}$
0 correlation	1453.9	133.9	0.9808
Spatial Correlation	1452.5	136.4	1
Perfect Correlation	1454.3	139.5	1.0216

assumed to be independent. Therefore, the total variance of the parameter is given by Equation 2.2.

$$\sigma^2_{total} = \sigma^2_{SYS-WID} + \sigma^2_{RAN-WID} + \sigma^2_{RAN-D2D} \tag{2.2}$$

We perform circuit simulation for a single critical path in a projected 100nm technology with a distributed-lumped RLC interconnect model and all correlations included. Table 2.2 compares the delay distributions obtained using our Monte Carlo simulation methodology for RLC interconnect model with (1) perfect correlations (correlation=1), (2) no correlations (correlation=0) and (3) spatial correlations. As can be seen, simulations with perfect/no correlations in fact set the upper/lower bound for total delay variation and as such they either overestimate or underestimate. While the magnitude of the delay variation due to interconnect fluctuations is not large, this demonstrates the effect of more accurate and detailed modeling of process variability for the purpose of assessing its impact on circuit performance.

In contrast with the linear regression analysis used in [3], our studies use a Monte Carlo (MC) approach with 1000 trials where the variation sources all vary simultaneously. Each model of process variability, at each technology node, gives rise to 1000 sets of random parameter values within the single critical path model which we simulate using HSPICE. In the next section, we proceed to investigate the resulting delay distributions in the single critical path model, in an attempt to gauge (1) the true impact of variability on circuit performance, and (2) the true value of developing improved process control, e.g., 1nm tighter control on interconnect thickness. With knowledge of the above, simulations for multi-critical paths are then performed: the maximum delay is obtained for each die, and performance and yield analyses are performed on m such samples on the whole wafer.

2.4 Impact on Future Circuit Performance

To assess the impact of process variation on critical path delay we adopt two different metrics.

1. *Selling point parametric yield.* We assume target parametric yield to be 99.7%. This corresponds to the mean+3σ point on the delay distribution and is taken to be the *selling point* of the chip. We take the delay distributions for values drawn from Table 2.1 as the "baseline" results for all technologies. The selling point is calculated from the baseline distribution. The change in parametric yield at the selling point is then taken as a measure of impact of process variation.

2. *Guardbanding Analysis.* Guardbanding is the typical approach followed in industry to account for variability. A larger amount of guardbanding implies a more conservative design and hence is not preferred. The expected ("designed-for") value of performance is given by the mean of the delay distribution. Thus, the difference between the selling point and the mean gives the amount of guardbanding required. That is, $\frac{3\sigma}{mean}$ expressed as a percentage gives the required guardbanding.

Table 2.3: Trends of performance variation

Delay (ps)	130nm	100nm	70nm
Mean	1502	1453	1537
3σ/Mean (%)	30.03	28.16	26.75

We have conducted experiments that change the expected 3σ variation of all parameters listed in Table 2.1, but due to space constraints we report results only for L_{eff}, which is the most significant contributor to process variation impact. Thus, the 3σ variations listed in Table 2.1 for parameters other than L_{eff} are held fixed in the following experiments.

2.4.1 Studies for A Single Critical Path

Cumulative Effect of All Parameter Variations

We simulate a single critical path and measure delay with all the parameters varying with 3σ and mean values as specified in Table 2.1. This simulation result is taken as the "baseline" result for all the comparisons and analysis of a single critical path explained in the subsequent subsections. Table 2.3 and Fig. 2.2 shows the *baseline* delay variation trends with technology scaling. The $\frac{3\sigma}{mean}$ value of delay drops by 6.6% from 130nm to 100nm, and by 5.3% from 100nm to 70nm. Overall, it remains fairly constant with technology scaling. The slightly decreasing trend of delay variation for our *baseline* setup with technology scaling can be explained as follows.

Assuming a critical path is formed by N identical stages, among which either perfect correlation (correlation=1) or no correlation (correlation=0) exists, the $\frac{3\sigma}{mean}$ value of the total path delay is given by Equation 2.3.

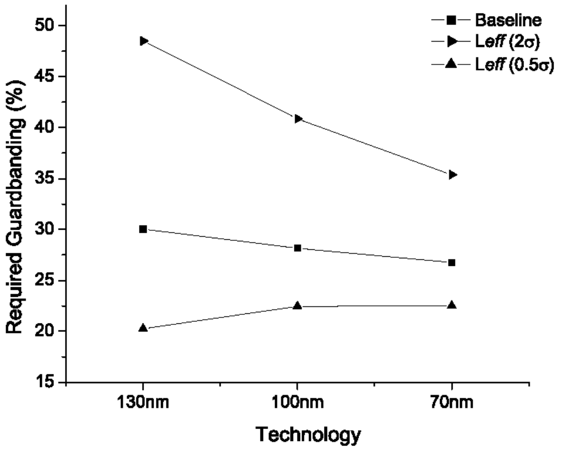


Figure 2.2: Effect of process control on required guardbanding to achieve 99.7% parametric yield.

$$(\frac{3\sigma}{mean})_{path} = \begin{cases} (\frac{3\sigma}{mean})_{stage}, & correlation = 1 \\ \\ \frac{1}{\sqrt{N}}(\frac{3\sigma}{mean})_{stage}, & correlation = 0 \end{cases} \tag{2.3}$$

Parameter N increases for global stages due to more aggressive buffering but decreases for local stages. The zero correlation assumption leads to a $(\frac{3\sigma}{mean})_{path}$ that is determined by both $\sqrt{N}$ and $(\frac{3\sigma}{mean})_{stage}$, while under the assumption of perfect correlation, it is only a function of $(\frac{3\sigma}{mean})_{stage}$.

Our analysis follows three steps. First, in order to investigate the trend of $(\frac{3\sigma}{mean})_{path}$, it is essential to know what trend $(\frac{3\sigma}{mean})_{stage}$ follows. Table 2.4 shows that the delay variation of a single local stage remains fairly constant though 130nm to 70nm technology node. Although a smaller average stage delay is expected for more advanced technologies, a constant trend for $(\frac{3\sigma}{mean})_{stage}$ is possible due to the ITRS expectations of a constant level of process variation achieved through advanced lithography tools. Second, the delay variations obtained for local stages in our testbed show that spatial correlation and perfect correlation assumptions have very close re-

Table 2.4: Trends of delay variation for a single local stage

Delay (ps)	130nm	100nm	70nm
Mean	92.15	82.53	77.04
3σ/Mean (%)	30.79	29.68	31.50

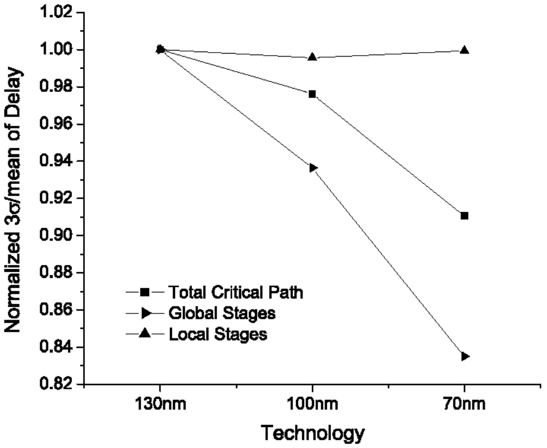


Figure 2.3: Trends of normalized required guardbanding for global stages, local stages and total critical path.

sults (difference within 4%). This implies that a decreasing N does not have an impact on path delay variation for local stages. Instead, the path delay shows a similar trend as obtained for a single stage - it remains fairly constant for the next two technology generations. Finally, the delay variation for global stages is shown to decrease at a reasonably fast rate due to the increased number of repeaters inserted (i.e., N is increased in Equation 2.3). This presents a dominating effect for future technologies and causes the trend of total delay variation to reduce, as shown in Fig. 2.3.

Sensitivity to Process Tolerance

To determine the sensitivity of performance to individual parameter tolerances, we changed the σ values from those in Table 2.1 to 0.5 and 2 times their original values. This was done for each parameter individually while maintaining the normal

Table 2.5: A comparison of changes in delay variation when the nominal σ for an individual parameter changes from 0.5X to 2X

Parameter	L$_{eff}$	T$_{ox}$	w
Increase in Delay $3\sigma/mean$	82.08%	3.96%	1.89%

σ for other parameters for each technology node. Among all the parameters listed in Table 2.1, delay variation is most sensitive to L$_{eff}$. Table 2.5 compares the respective changes in delay variation for the 100nm technology node when the nominal σ changes from 0.5X to 2X for L$_{eff}$, T$_{ox}$, and metal width w. Fig. 2.2 shows the impact on guardbanding of varying L$_{eff}$ control. Fig. 2.4 shows the impact on selling point yield. Loose L$_{eff}$ control (on the order of 2X the current levels) can cause a loss of up to 10.4% in yield. In Figure 2.4 it is interesting to note that the knee, or roll-off point, of selling point yield versus process control of the three investigated parameters all seem to be around the nominal variation level (equal to that in Table 2.1). This may imply that current levels of process control are near-optimal; however, slopes in the three curves clearly differ which also indicates a smaller ROI on enhanced L$_{eff}$ process control in future technologies. This last point is due to a rising lateral electric field in the device, which causes more pronounced velocity saturation effects. As seen in Table 2.1, the ratio of V_{dd} to L$_{eff}$ goes from 1.2/65nm in 130nm to 0.9/28nm in the 70nm process. This increase of 74% in average channel electric field dictates that the 70nm devices are more velocity-saturated, and their saturation drain currents are consequently less dependent on channel length. Another mechanism by which switching speeds (heavily dependent in digital circuits on saturation drain current) may become less sensitive to channel lengths is through use of strained silicon channels which improve carrier mobility and also lead to more velocity-saturated transistors.

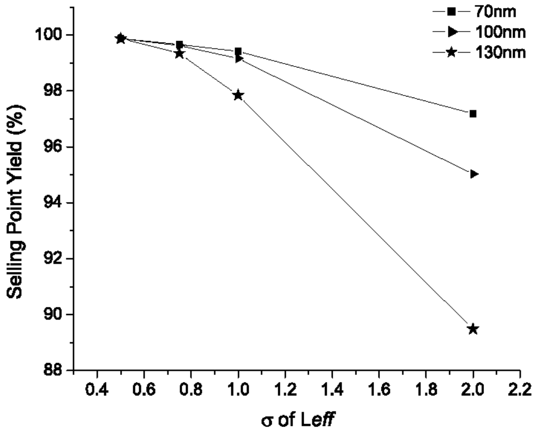


Figure 2.4: Effect of L_{eff} control on selling point parametric yield.

As described in the taxonomy of variations in Section 2.2, both D2D and WID variations play a key role in the overall circuit variability. Since techniques have been described to address or ameliorate the effects of either D2D or WID variations (e.g., [32], which targets within-die V_{th} variation by using adaptive body bias), it is instructive to analyze the relative roles played by each type of variation on circuit performance. Fig. 2.5 shows the sensitivity of the required guardbanding to various controls on uncertainty sources, e.g., WID, D2D, and the cumulative effect of both, for L_{eff}. As can be seen in Fig. 2.5, for one critical path, delay variation is most sensitive to random D2D variation. The sensitivity of the delay variation to CD fluctuations arising from random D2D, random WID, or cumulative effect decreases in future technologies. These results imply that greater benefit in terms of reducing the required design guardbanding can be achieved by focusing on reduction of random D2D variation rather than on random WID variation. For example, at the 70nm technology node, reducing random WID variation by half only yields a 1.9% smaller guardband, while the same reduction in random D2D variation achieves an improvement of 14.3%.

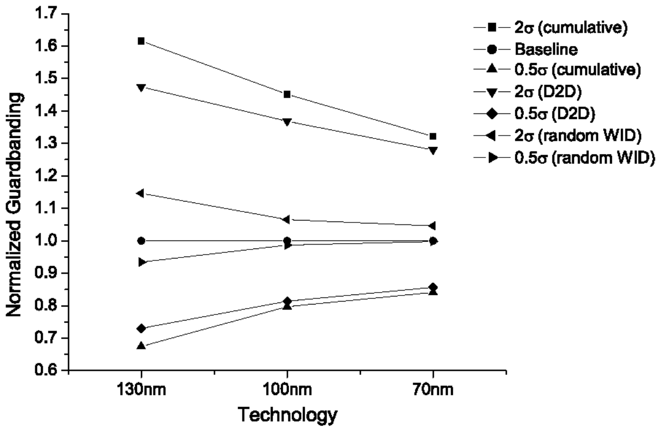


Figure 2.5: Effect of control over variability (L_{eff}) sources (cumulative, random D2D, and random WID) on normalized guardbanding required to achieve 99.7% parametric yield.

Impact of Technology Roadmap Deceleration

In this subsection, we consider scenarios in which no further improvements (beyond 130nm technology) are made in the control of a given process parameter, while control of other parameters scales according to Table 2.1. In other words, the absolute σ value for the given parameter (here L_{eff} as an example) is kept constant at its 130nm technology node value. While this *exact* scenario may be unlikely, we are generally seeking to explore the feasibility of a slowed technology roadmap in which large equipment and research expenditures may be reduced with comparatively small ramifications on circuit variability. Fig. 2.6 gives the “worst-case” impact of no further investments for control of L_{eff}. The results support our previous analysis: (1) it is the continually shrinking process variation that makes a decreasing variation trend possible with technology scaling; (2) without such shrinking, the required design guardbanding for the 70nm technology node would be 42.0% larger than the one shown in “baseline” case. It is worth mentioning that such analysis is pessimistic in that advanced lithography tools (e.g., 193nm or extreme ultraviolet, EUV) must

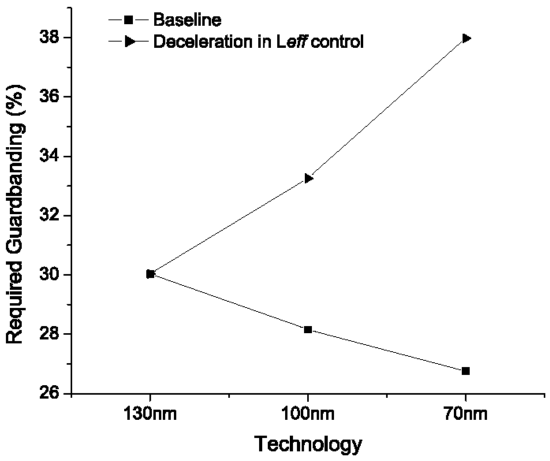


Figure 2.6: Impact of technology roadmap deceleration in L$_{eff}$ control on required guardbanding.

be implemented in any case to manufacture future technology generations, meaning that somewhat better variation control should be achievable with technology scaling alone. However the analysis provides a means to gauge the trade-off between the loss in performance and the savings in tool cost inherent in not expending *extra* effort on process variation control.

Inferred Tolerance of Process variation from Desired Guardbanding

Some process parameter controls involves new technologies (e.g., optical proximity correction (OPC), phase shift mask (PSM), etc.) and hence may incur very large costs. To ensure that ROI is maximized when moving from one process technology to the next, it is critical for process engineers to know what level of process control is necessary for adequate circuit variability performance. In our framework, we can infer from a desired level of circuit performance variability (which can be translated into guardbanding) to individual process parameter uncertainties. This is done by assuming a linear relationship between the magnitude of σ for an individual process parameter and the variation of total delay based on the simulation results for sensi-

Table 2.6: Tolerance of L_{eff} variation for a guardbanding budget of 30%

Technology node	130nm	100nm	70nm
Required σ	1.00X	1.15X	1.34X

tivity analysis [3]. For example, to achieve a guardbanding target of 30% (any value may be used), Table 2.6 gives the tolerance of L_{eff} variation in a number times the nominal σ from Table 2.1. It is noteworthy that for the same design guardbanding, the requirement for L_{eff} control increases from 130nm to 70nm, meaning that less effort on CD variation control may be applied. It indicates that to maintain 30% design guardbanding, levels of control dictated by the ITRS may be overly stringent. Again, this is driven primarily by the slowed voltage scaling predicted by the ITRS which yields device delays that are less sensitive to channel length. The importance of this analysis lies in the fact that the 2001 ITRS indicates a lack of known solutions to achieving the 10% physical gate length control requirement in 70nm technologies in 2006. If these requirements can indeed be relaxed, then more well-understood and cost-effective approaches may be applicable to maintain manufacturability at such device geometries.

2.4.2 Studies of Multi-Critical Paths

In this subsection, we describe the impact of multi-critical paths on circuit performance. We focus on the 100nm technology generation; the analysis can be easily extended to other technology generations.

Impact on Delay Distribution

Since systematic WID variation has the same impact across all dies, it does not affect yield when there is *one* dominant critical path on a die. However, when the

[3] The validation of this assumption is shown in the sections below.

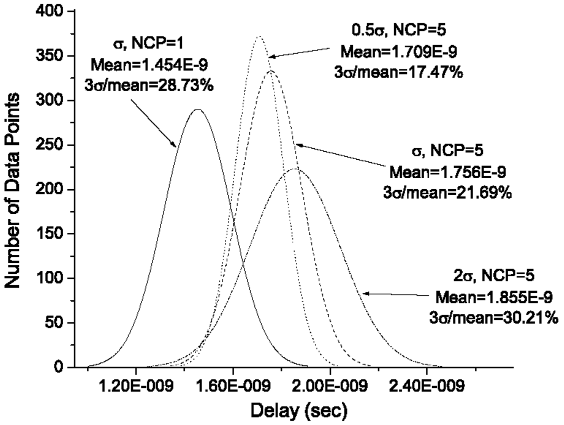


Figure 2.7: Comparison of different L$_{eff}$ variation control on five critical paths for 100nm technology node.

number of critical paths (NCP) increases, systematic WID variation exhibits a greater impact on chip performance, namely, it results in different delay distributions for different critical paths that have identical designed-for delays [33]. For NCP equals to five, we investigate the sensitivity of the delay distribution to CD variation control, and compare it with the "baseline" case for a single path, as shown in Fig. 2.7. Fig. 2.8 shows a shifting of the total delay mean when NCP increases, with varying levels of L$_{eff}$ process control. The mean value of total delay changes rapidly when the number of critical paths increases from 1 to 10, beyond which it saturates. Here again, a doubling of L$_{eff}$ variation shows more impact on delay variation than a similar reduction in its variation. This suggests that for more critical paths, the expected value of delay becomes more sensitive to CD variation control (note the vertical spread of the three data points at NCP = 50 vs. NCP = 5). These observations all indicate that, for a large number of critical paths, tighter control on process variation is desired. In future designs, we expect a larger NCP due to the push for lower power design. Design techniques such as dual-Vth, dual-Vdd, and more classical sizing techniques to reduce power all act to create larger numbers of critical and

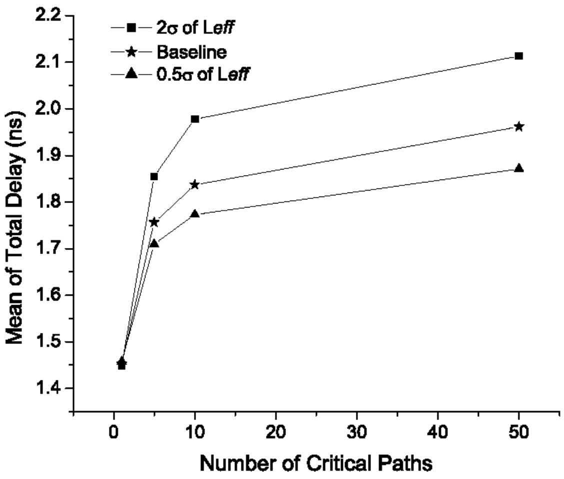


Figure 2.8: Delay mean value as a function of number of critical paths.

near-critical paths. As observed in [34], power-optimal solutions leveraging multiple supply and threshold voltages create an enormous amount of critical paths which will complicate timing verification and also hurt parametric yield. Further work will be required to identify stable design points that carefully trade-off between power and yield.

Sensitivity to Process Tolerance Analysis

To achieve 99.7% parametric yield, the required guardbanding as a function of L_{eff} σ is shown in Fig. 2.9. Under the same CD variation control, the required guardbanding becomes smaller as NCP increases, and its sensitivity to NCP falls off sharply for more than 10 paths. For example, in the “baseline” case ($\sigma = 1 \times \sigma_{normal}$), the required guardbanding drops by 48.9% when the NCP increases from 1 to 10, but by only 6.7% when NCP increases from 10 to 50. When NCP increases beyond 100, as we would expect in low-power designs, the required guardbanding will be slightly lower, but remains predictable from results obtained using fewer critical paths (e.g., $\simeq$ 10). We would eventually expect a lower bound on the required guardbanding with a continually increasing NCP. In contrast to the delay mean value, as NCP

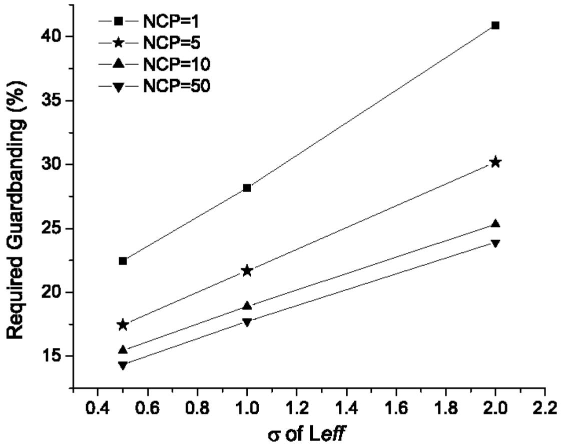


Figure 2.9: Impact of L_{eff} control on required guardbanding to achieve 99.7% parametric yield.

increases the design guardbanding becomes *less* sensitive to L_{eff} variation, which is shown by the decreasing line slopes in Fig. 2.9 for more critical paths. The linear dependence of the required guardbanding to the σ of L_{eff} also validates the linear assumption in the analysis for inferred variation tolerance stated earlier in Section 2.4.1.

With more critical paths, $\frac{3\sigma}{mean}$ of delay becomes smaller. However, due to the shifting in the average delay, the selling point delay (i.e., mean+3σ) becomes worse. To investigate the trend, we define the delay that gives 99.7% yield for one critical path with L_{eff} varying at normal σ as the selling point delay, and plot the parametric yield as a function of the number of critical paths under different process controls in Fig. 2.10. To reduce the yield loss caused by an increasing NCP, two options are to improve process control (incurring manufacturing costs) or to reduce the number of critical paths in circuit design (causing additional design effort and likely increasing power consumption).

While the above discussion has focused on the interaction of multiple critical paths and total variability, we would also like to investigate the relative contribution

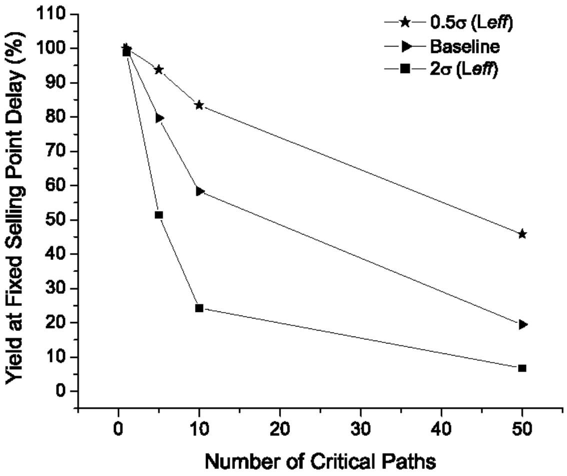


Figure 2.10: Parametric yield as a function of the number of critical paths and process variation control.

of systematic WID variation within the total variability budget. We perform MC simulations with NCP equals to 5 for both 100nm and 70nm technology nodes, along with NCP equals to 10 for the 100nm technology node only. These NCP values are chosen since design guardbanding becomes much less sensitive to process variation control for $NCP \geq 10$. Fig. 2.11 compares the sensitivity of the required guardbanding for 99.7% yield to different controls on L_{eff} variation. In this analysis, the change in L_{eff} variation is manifested in either systematic WID or the cumulative effect of *both* WID and D2D. The latter case corresponds to the previous analysis (e.g., Fig. 2.9). As can be seen in both cases, the required design guardbanding becomes less sensitive to CD control as NCP increases and technology scales. Also demonstrated is that better control of systematic WID CD variation is only effective in reducing the design guardband for smaller NCP.

In general, the following facts are observed: 1) for multiple critical paths, both the required design guardbanding and its sensitivity to CD control become smaller with technology scaling; 2) as NCP increases, a larger delay mean and a significantly

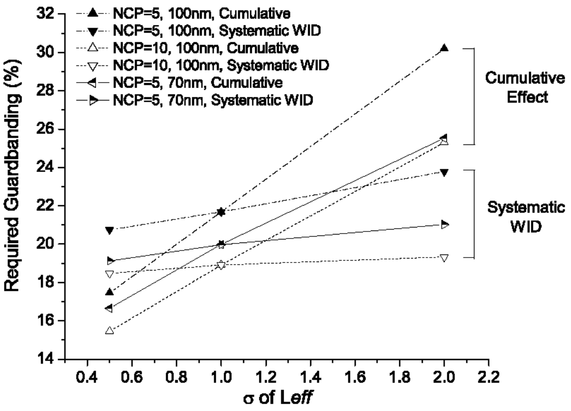


Figure 2.11: Sensitivity of required design guardbanding for 99.7% yield to variation with respect to its sources (simulations results for the same NCP and technology generation are shown in the same line styles).

smaller delay variation are expected; 3) increased selling point delay (mean+3σ) will cause large yield losses in ASIC designs unless the effect of CD variability is carefully modeled during the design process; 4) delay *mean* value is more sensitive to CD variation control for larger number of critical paths, while delay *variation* shows more sensitivity for smaller numbers of critical paths; and 5) both delay mean and delay variation are more sensitive to NCP under larger amounts of expected process variation, i.e., for the same level of CD control, these parameters vary widely for 1 <NCP< 10, beyond which the sensitivity is reduced. We may conclude that to improve yield at a fixed selling point delay, reducing NCP to $\leq$ 10 is the most effective way to achieve a smaller delay mean, and to create a situation where process control is most valuable.

With these results and recommendations in mind, we next explore the concept of designing integrated circuits with performance *distributions* in mind, rather than deterministic performance points. For instance, in a binned design methodology (as with high-end microprocessors) the value of each die depends on the achieved performance. Designing to maximize total value, considering both parametric yields and relative market prices, is an open area in the literature and is discussed in the next section.

2.5 Variation-Centric Physical Design

Conventional design is based on the goal of performance optimization, or design for performance (DFP). As process variation inevitably leads to performance distributions, it implies the possibility of *design for value* (DFV) methodologies to maximize the yield. In this section, we attempt to motivate the case for DFV in nanometer CMOS design. Performance is measured by critical path delay T. It is a function of design variables x_i and process parameters y_i, i.e.

$$T = f(x_1, ..., x_m, y_1, ...y_n) \tag{2.4}$$

Design for performance seeks to find values of x_i to minimize T, given the nominal values of y_i, and ignoring process variations, i.e.

$$y_i = y_{i\_nom}, \quad \text{Minimize } T \tag{2.5}$$

Alternatively, worst-case values of y_i may be used. This is representative of a corner-based approach where all parameters are very pessimistically taken at their 3σ points, again within a deterministic framework.

We define *value* to be the total dollars earned from the chip that is sold on the open market. A value function $v(f)$ gives the market value of the chip for some performance measure f (e.g. speed, power). Thus, the total value of a given process is obtained as:

$$Value = \Sigma v(f) \times yield(f) \tag{2.6}$$

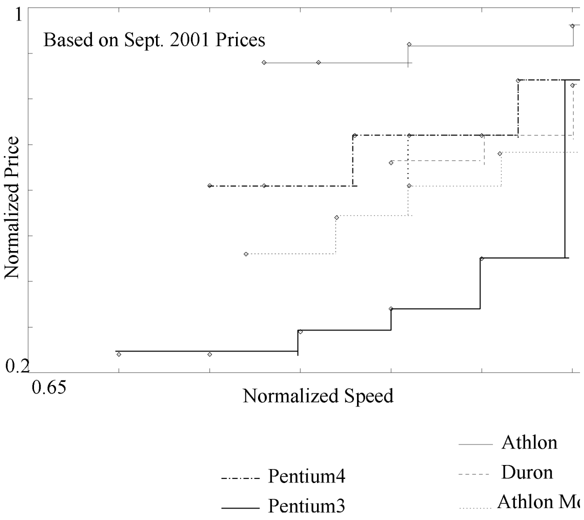


Figure 2.12: Normalized market price of recent μP products.

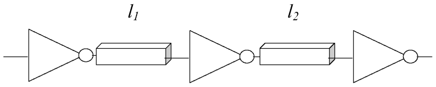


Figure 2.13: Simulation structure for DVP vs. DFV study.

where $yield(f)$ has the usual meaning. Examples of microprocessor unit (MPU) value functions are shown in Fig. 2.12. Design for value seeks to find values of x_i to maximize yield of $T < T_m$, given the variability distributions of parameters y_i, where T_m is the target delay, i.e.

$$y_i = N(\mu_i, \sigma_i), \quad \text{for } 1 < i < n, \text{ Maximize } P_T(T_m) \tag{2.7}$$

The two approaches of Equation 2.5 and Equation 2.7 may not be equivalent, as is demonstrated by the following example. To investigate the difference between design for performance and design for value and its impact on physical design, we conduct a MC simulation experiment. A simple example of a global line with a source, sink, and repeaters is constructed as shown in Fig. 2.13. It is set up using the baseline 130nm technology. An analytical delay model was developed for this case as SPICE is too computationally expensive for this experiment. The interconnect capacitance

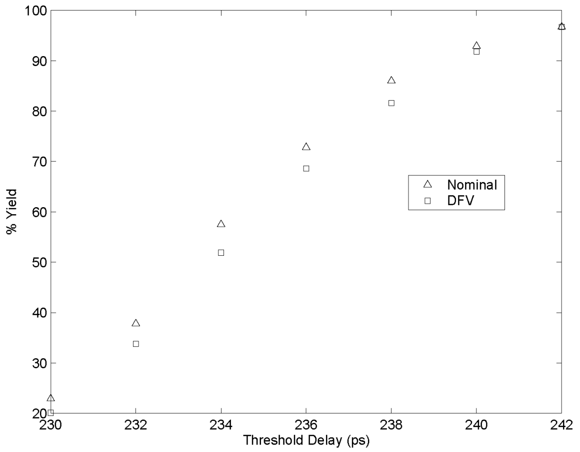


Figure 2.14: Difference in yield between DFP and DFV.

model is taken from [35] for the case of parallel lines between two ground planes. Device resistance is assumed to be 0.8V$_{dd}$/I$_{dsat0}$ while source/drain series resistance R_s is taken to be 0.1V$_{dd}$/I$_{dsat0}$ [36]. Inverter delay is calculated using models from [37]. The RLC interconnect delay model presented in [36] is used to calculate the inverter and interconnect delay. Normally distributed variation in L$_{eff}$ is considered as the critical dimension typically has the largest variation and impact on circuit performance. Nominal and σ values of various parameters are shown in Table 2.1. Correlation is assumed to decay linearly with distance and the correlation for distance greater than 10mm is taken to be 0. Repeater location (distance from source) is varied from 0 to 4mm in 0.1 mm steps. A nominal optimum repeater location is first calculated by sweeping the location for nominal process parameter values. Next, the DFV optimum is calculated for each given selling point or threshold delay by running 1000 Monte-Carlo simulations at each repeater location. The difference between yields of DFP and DFV optimizations is shown in Figure 2.14. The parametric yield difference ranges from 1.1% to 5.6% depending on the selling point delay. This value difference may be even more pronounced if more complex value functions (as the ones in Figure 2.12) are used for DFV optimization.

2.6 Conclusions

In this chapter, we have presented a new framework for assessing the impact of process variation on circuit performance, product value, and return on investment for alternative process improvements. We present our results in terms of new metrics such as guardbanding, parametric yield at selling point, and inferred variation tolerance. This framework follows a comprehensive taxonomy of variations, and can handle variation differently with respect to its sources. We use accurate models of correlations and Monte Carlo techniques based on circuit simulation. Our main conclusions are as follows.

- With technology scaling and ITRS mandated fixed levels of process variability, delay variation decreases, whether measured as the amount of guardbanding required to circumvent it, decrease in parametric yield that may need to be tolerated, or the inferred variation tolerance for a preset design guardbanding.

- For chips containing one dominant critical path, systematic WID variation does not affect yield, and design guardbanding is most sensitive to random D2D variation control.

- Performance is very sensitive to L_{eff} variation but the sensitivity reduces with technology scaling due to enhanced velocity saturation and a growing number of critical paths.

- Under the same level of process variation, a larger NCP results in a smaller delay variation but larger delay mean. Because of the shift in delay mean value, a larger selling point delay is expected.

- For the same NCP, looser control of CD variability leads to a larger required design guardbanding accompanied by a larger delay mean value, both of which show more sensitivity to relaxed process specifications than to tightened specs.
- The delay distribution shifts to higher means but tighter overall distributions as the number of critical paths increases but this effect saturates beyond approximately 10 critical paths.
- For ASIC designs, reducing NCP is the most effective way to achieve a smaller average delay.
- Variability impact can be restricted by innovative design and this may be preferable due to the very costly nature of process improvement techniques.

Our results suggest the potential utility of *Design for Value* (DFV) methodologies that take variability into account during design optimizations. For instance, one may argue in favor of *selling point optimization* rather than traditional nominal performance optimization. Also, there may be multiple selling points with some prespecified *value* associated with each selling point. The total design value is then given by $\Sigma v(f) * yield(f)$, for a given value function v of performance measure f, and given parametric yield distribution $yield(f)$. DFVD then seeks to find values of design parameters to maximize value, assuming normally distributed process parameters. This calls for research into such probabilistic optimizations, as well as efforts to quantify the potential value and costs associated with both manufacturing and design solutions to the process variability issue.

CHAPTER III

Performance-Driven OPC for Mask Cost Reduction

3.1 Introduction

Continued technology scaling in the subwavelength lithography regime results in printed features that are substantially smaller than the optical wavelength used to pattern them. For instance, 130 nm CMOS processes use 248 nm exposure tools, and the industry roadmap through the 45 nm technology node will use 193 nm (immersion) lithography. The International Technology Roadmap for Semiconductors (ITRS) [38] identifies aggressive microprocessor (MPU) gate lengths and highly controllable gate CD control as two critical issues for the continuation of Moore's Law cost and integration trajectories. To meet ITRS requirements (see Table 1.1), resolution enhancement techniques (RETs) such as optical proximity correction (OPC) and phase shift masks (PSM) are applied to an increasing number of mask layers and with increasing aggressiveness. The recent steep increase in mask costs and lithographic complexity due to these RET approaches has had a harmful impact on design starts and project risk across the semiconductor industry. Cost of ownership (COO) has become a key consideration in adoption of various lithography technologies.

3.1.1 OPC and Mask Cost

The increasing application of RETs makes mask data preparation (MDP) a serious bottleneck for the semiconductor industry: figure counts explode as dimensions shrink and RETs are used more heavily. Compared with the mask set cost in 0.35 μm, the cost at the 0.13 μm generation with extensive PSM implemented is four times larger [39]. Figure counts, corresponding to polygons as seen in the IC layout editor grow tremendously due to sub-resolution assist features and other proximity corrections. Increases in the fractured lay-out data volume lead to disproportionate increases in mask writing and inspection time. According to the 2005 ITRS [38], the maximum single-layer MEBES file size increases from 64GB in 130 nm to 216GB in 90 nm. Another observation concerns the relationship between design type and lithography costs, namely, that the total cost to produce low-volume parts is dominated by mask costs [6]. Half of all masks produced are used on less than 570 wafers (this translates roughly to production volumes of $\leq$ 100,000 parts). At such low usages, the high added costs of RETs can-not be completely amortized and the corresponding cost per die becomes very large. Thus, designers and manufacturers are jointly faced with determining how best to apply RETs to standard cell libraries to minimize mask cost.

In this work we focus on OPC, which is a major contributor to mask costs as well as design turnaround time (TAT). More than a 5X increase in data volume and several days of CPU runtime are common side effects of OPC insertion in current designs [40]. With respect to the cost breakdown shown in Figure 3.1, OPC affects mask data preparation (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 where run time is roughly proportional to to

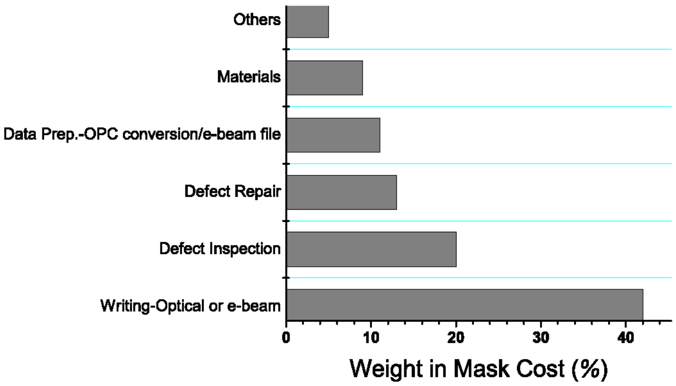


Figure 3.1: Relative contributions of various components of mask cost [4].

feature complexity, comprise the dominant approach to high-speed mask writing. In the standard mask data preparation 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 mask data preparation, 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 [41]. 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 (e.g., according to the 2005 ITRS [38], the maximum single-layer MEBES file size increases from 216GB in 90 nm to 729 GB in 65 nm) 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 [6] 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.

3.1.2 Design Function in the Design-Manufacturing Interface

A primary failure of current approaches to the design-manufacturing interface is the lack of communication across disciplines and/or tool sets. For example, it is well documented that mask writers do not differentiate among shapes being patterned - given this, gates in critical paths are given the same priority as pieces of a company logo and errors in either of these shapes will cause mask inspection tools to reject a mask. In this light, 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, such "overcorrection" leads to higher mask costs and larger runtimes.

3.1.3 A Performance-Driven OPC Methodology

In this work, we propose a performance-driven OPC methodology that is demonstrated to be highly implementable within the limitations of current industrial design flows. Contributions of our work include the following.

- *Quantified CD error tolerance.* We propose a mathematical programming based budgeting algorithm that outputs edge placement error tolerances (in nm) for layout features.

- *Integration within a commercial MDP flow.* We describe a practical flow implementable with commercial tools and validate the minimum cost of correction methodology.

- *Reduction of OPC overhead.* We measure OPC overhead in terms of additional MEBES features as well as runtime of the OPC insertion tool and show substantial improvements in both.

Table 3.1: Correspondence between the traditional gate sizing problem and the minimum cost of correction (to achieve a prescribed selling point delay with given yield) problem.

Gate Sizing		MinCorr
Area	$\equiv$	Cost of Correction
Nominal delay	$\equiv$	Delay $\mu + k\sigma$
Cycle Time	$\equiv$	Selling point delay
Die Area	$\equiv$	Total Cost of OPC

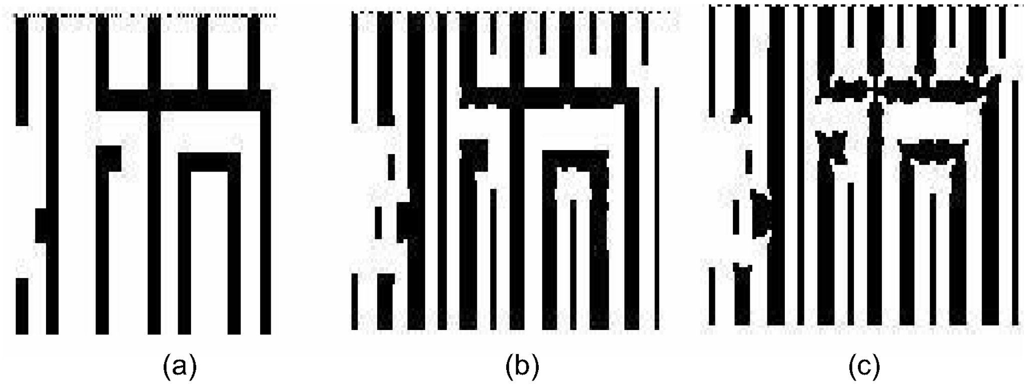


Figure 3.2: An example of three levels of OPC [42]. (a) No OPC, (b) Medium OPC, (c) Aggressive OPC.

3.2 General Cost of Correction Flow (MinCorr) Based on Sizing

We describe a generic yield closure flow which is very similar to traditional flows for timing closure. In this section, we describe the elements of such a flow.

In this generic *sizing based MinCorr* flow, we emphasize the striking similarity to conventional timing optimization flows. The key analogy - and assumption - is that there are discrete allowed “sizes” in the MinCorr problem that correspond to allowed levels of OPC aggressiveness (see Figure 3.2). Furthermore, for each instance in the design there is a cost and delay penalty associated with every level of correction. The mapping between traditional gate-sizing and the MinCorr problems is reproduced in Table 3.1. This flow involves construction of cost/yield aware libraries for each level of correction, and a commercial STA tool together with a selling point yield bonding algorithm which applies timing driven cost optimization. We acknowledge the following facts during the flow development process:

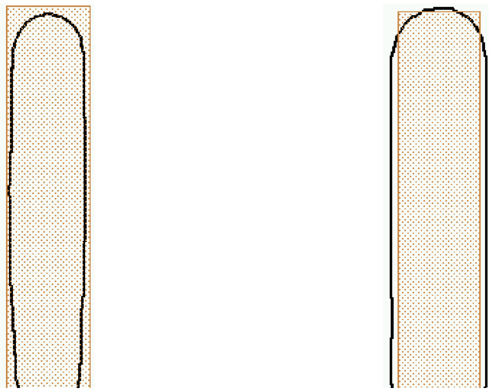
- We assume that different levels of OPC can be independently applied to any gate in the design. Corresponding to each level of correction, there is an effective channel length L_{eff} variation and an associated cost.

- Differentiate field-poly from gate-poly features. Field poly features do not impact performance and hence any delay-constrained MinCorr approach should not change the correction of field-poly. Moreover, quality metrics of field-poly are different from those of gate-poly (e.g., contact coverage). By recognizing these two types of poly features, we may avoid over estimating cost savings achieved with this approach.

- The mask writing time which dominates mask cost ([4]) is a linear function of figure count numbers [43]. These numbers, as proxies for mask cost implications, are extracted for the cells from post-OPC layout with commercial OPC insertion tool.

- OPC corrects the layout for pattern-dependent through-pitch CD variation. Such variations are predictable, for example, by lithography simulations.

With these facts, the *MinCorr* problem is summarized as: given a range of allowable corrections for each feature in the layout as well as the mask data volume and CD deviation associated with each level of correction, find the level of correction for each feature such that prescribed circuit performance is attained with minimum total



(a) EPE > 0 **(b) EPE < 0**

Figure 3.3: The signed edge placement error (EPE).

correction cost. Commercial OPC tools are driven by *edge placement errors* (EPEs), rather than critical dimensions (CDs). Thus, we specify a practical *MinCorr* with a practical implementation - *EPEMinCorr*. We can summarize the key contribution of EPEMinCorr as: *we devise a flow to pass design constraints on 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* (e.g., Figure 3.2 shows OPC layers driven by different EPE requirements). 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 3.3. Mask data volume is heavily dependent on the assigned EPE tolerance that the OPC insertion tool is asked to achieve. For example, Figure 3.4 shows the change in MEBES file size for cell with applied OPC as the EPE tolerance is varied. In this particular example, loosened

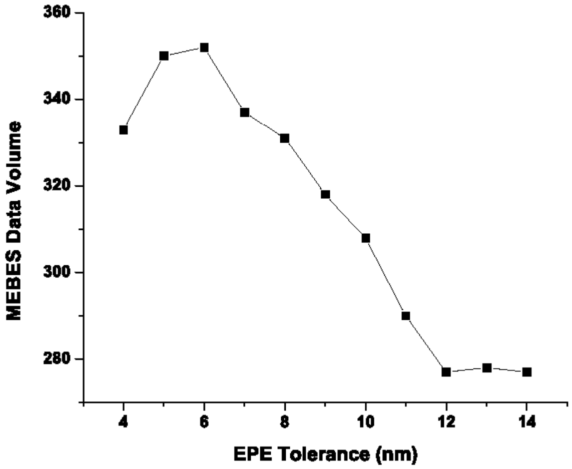


Figure 3.4: Mask data volume (kB) vs EPE tolerance for a NAND3X4 cell in TSMC 130nm technology.

EPE tolerances can reduce data volume by roughly 20% relative to tight control levels.

Since model-based OPC corrects for pattern-dependent CD variation, which is systematic and predictable, we assert that OPC actually determines *nominal timing.* 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 as mentioned above - 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 3.5 shows our EPEMinCorr flow. The quality of results generated by the flow are measured as MEBES data volume of fractured post-OPC insertion layout

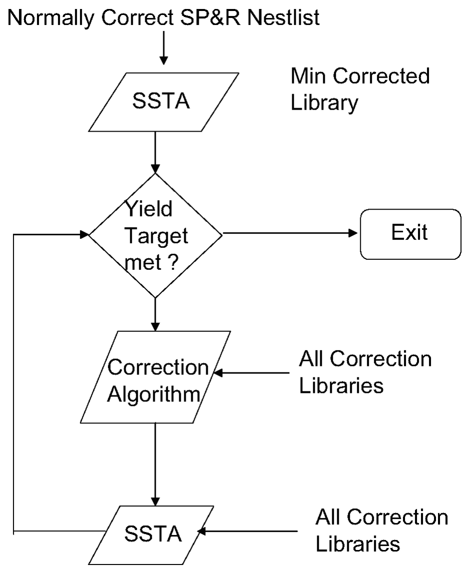


Figure 3.5: The EPEMinCorr flow to find quantified edge placement error tolerances for layout features and drive OPC with them.

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 the Figure 3.5 EPEMinCorr flow.

3.2.1 Slack Budgeting

The slack budgeting problem seeks to distribute slack at the primary inputs of combinational logic (i.e., sequential cell outputs) to various nodes in the design. One of the earliest and simplest approaches, the zero-slack algorithm (ZSA) [44], iteratively finds the minimum-slack timing path and distributes its slack equally among the nodes in the path. The MISA algorithm for slack budgeting proposed in [45] distributes slack iteratively to an independent set of nodes. As with ZSA, the objective is to maximize the total added incremental delay budget on timing arcs. A weighted version of MISA is also proposed in [45].

We observe:

- Neither MISA variant is guaranteed to provide optimal solutions.
- ZSA is much faster than MISA, and a weighted version of ZSA can also be formulated.
- While [46] formulates the budgeting problem as a convex programming problem, full-chip MISA or mathematical programming is, as far as we can determine, too CPU-intensive for inclusion in a practical flow.

We propose to approximate full-chip mathematical programming by iteratively solving a sequence of linear programs (LPs). In each iteration, slack is budgeted among the top k available paths. Once a budget is obtained for a node, this budget is retained as an upper bound for subsequent iterations. The process is repeated until all nodes have been assigned a slack budget or path slack is sufficiently large. The basic LP has the following form:

$$\begin{aligned} \text{Maximize} \sum_{i=1}^{n} C_i s_i \\ \sum_{j \in P_k} s_j \leq S_k \ \forall\ k \in \text{Current path list} \\ s_j \leq s_j^f \ \forall\ j \in F \end{aligned} \tag{3.1}$$

where C_i denotes the correction cost decrease per unit delay increase for cell i, and s_i is the slack allocated to cell i. The notation P_k is used to denote the k^{th} most critical path, and S_k is the slack of this path. Finally, F denotes the set of nodes with slacks fixed from previous iterations. An example sequence of LPs might be obtained by allowing k to take on the range from 1 to 100 in the first iteration, 101 to 200 in the second iteration, and so on.

We observe that when a budgeting formulation is adopted in place of a sizing formulation, the method of accounting for changes in next-stage input pin capacitance becomes an open question. To be conservative, we generate timing reports with pin input capacitances that correspond to the loosest tolerance (i.e., largest pin capacitance) but gate delays corresponding to the tightest achievable tolerance. C_i is obtained via a pre-built look-up table (similar to .lib format) containing the increase in data volume, mapped against delay change.

Our budgeting procedure yields positive delay budgets leading to positive EPE tolerances. Since EPE tolerance is a signed quantity (e.g., in Mentor Calibre, a common OPC insertion tool), negative EPE tolerances (corresponding to reduced gate length and faster delay) can also be obtained in a similar way based on hold-time or leakage power constraints. However, in this work we assume equal positive and negative EPE tolerances since we deal with purely combinational benchmarks and focus on timing rather than power.

3.2.2 Calculation of CD Tolerances

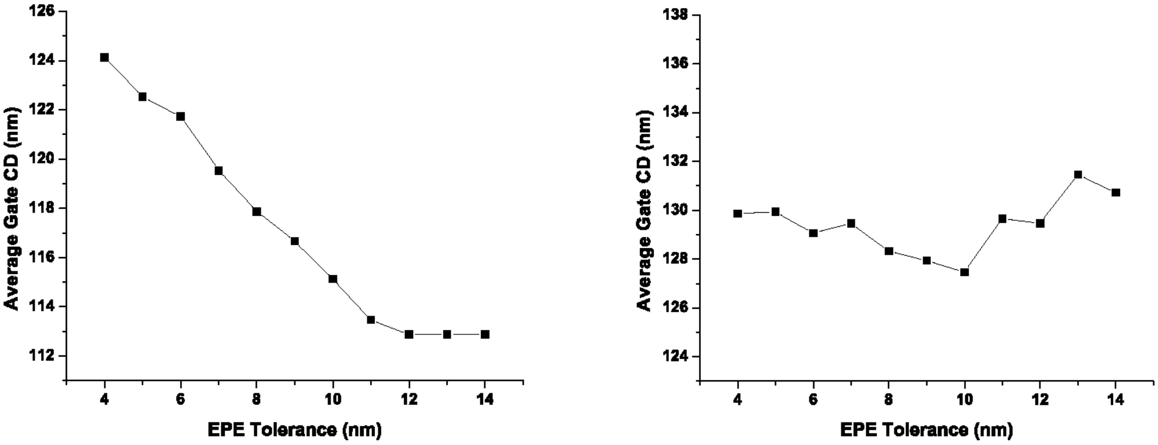
To map delay budgets found from the above linear programming based formulation to CD tolerances, we require characterization of a standard-cell library with varying gate lengths. Using such an augmented library, along with input slew and load capacitance values for every cell instance, we can map delay 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. Subtracting these budgeted gate lengths from nominal gate lengths yields the CD tolerance for every cell in the design.

3.2.3 Calculation of EPE Tolerances

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.

In most lithography processes, gates shrink along their entire width such that the printed gate length is always smaller than the drawn gate length, except at the corners of the critical gate feature. OPC typically biases the gate length such that corrected gate length is *larger* than the designer-drawn gate length. Thus, model-based OPC shifts edges *outward*, i.e., in the "positive" direction, until it meets the EPE tolerance specification. If the step size of each edge move is small enough, the EPE along the gate width will always be negative (since we are approaching the larger nominal gate length value starting from the smaller printed gate length value). As a result, actual printed gate length will almost always be smaller than the drawn gate length, leading to leakier but faster devices.

To achieve a more unbiased deviation from nominal, we exploit the behavior of the OPC tool by applying simple pre-biasing of gate features in an attempt to achieve EPE tolerances that are equal to CD tolerance. Specifically, we pre-bias each gate feature by its intended EPE tolerance. For instance, for a drawn gate length of 130 nm and EPE tolerance of 10 nm, the printed CD would typically lie between 110 nm and 130 nm (each edge shifts by 10 nm inward). If the gate length is biased by 10 nm so that the OPC tool views 140 nm as the target CD, the printed CD would



(a) Without pre-bias. **(b) With pre-bias.**

Figure 3.6: Comparison of average printed gate CD with and without pre-bias for the cell macro NAND3X4.

lie between 120 nm and 140 nm, which amounts to a ±10 nm CD tolerance. In this way, pre-biasing achieves CD tolerances equal to the EPE tolerance. An example of the average CD for a specific gate-poly with and without pre-biasing is shown in Figure 3.6. It is clear that pre-biasing achieves its goal of attaining average CDs that are very close to the target CD (130 nm in our case). Another point illustrated in Figure 3.6 is that the variation in CD (measured as the standard deviation of CD taken across all edge-fragments) grows as the EPE tolerance is relaxed. This is shown more clearly in Section 3.3.4.

3.2.4 Constrained OPC

We enforce the obtained EPE tolerances within a commercial OPC insertion flow. We use *Calibre* [47] as the OPC insertion tool; details of constraining the tool are described in the next section.

Table 3.2: Benchmark details.

Test Case	Source	Cell Count
c432	ISCAS85	337
c5315	ISCAS85	2093
c6288	ISCAS85	4523
c7552	ISCAS85	2775
alu128	Opencores	12403
r4_sova	Industry	34288

3.3 Experimental Setup and Results

In this section we describe our experiments and the results obtained in order to validate the EPEMinCorr methodology.

3.3.1 Test Cases

We use several combinational benchmarks drawn from ISCAS85 suite of benchmarks and Opencores [48]. These benchmark circuits are synthesized, placed and routed in a restricted TSMC 0.13 μm library containing a total of 32 cell macros with cell types of BUF, INV, NAND2, NAND3, NAND4, NOR2, NOR3, and NOR4. The test case characteristics are given in Table 3.2.

3.3.2 Library Characterization

We assume a total of EPE tolerance levels ranging from ±4 nm to ±14 nm. 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 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, standalone OPC is fairly representative of data volume changes with changing EPE tolerance. 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 3.4 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 4 nm). When slack is distributed to various nodes, we extract the load capacitances that are used to identify entries in the sensitivity table. Cost change is most sensitive to delay changes when the load capacitance is small (this typically indicates a small driver and subsequently small amount of data volume) and the sensitivity numbers are on the order of 1X to 10X MEBES features per ps delay change.

3.3.3 EPEMinCorr 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. As indicated in Figure 3.7, we first separate the entire poly layer into gate poly and field poly components. The field-poly tolerance is taken to be ±14 nm while gate-poly tolerance ranges from ±4 nm to ±14 nm. We tag the assigned EPE tolerance to cell names. In this way, we can track the EPE tolerance of each gate individually. We take 1 nm as our step size (step size is the minimum perturbation to an edge that model-based OPC can make, and smaller step sizes lead to better correction accuracy at the cost of runtime) when applying OPC to obtain very precise correction levels.

We set the iteration number to the minimum value beyond which adding mask

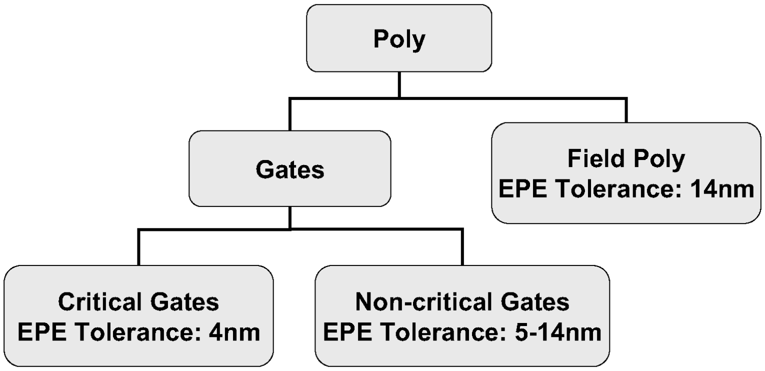


Figure 3.7: Summary of EPE assigment for OPC level control.

.

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.

3.3.4 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 (not true for the biggest benchmark r4_sova where more paths are needed as discussed below) as well as the load capacitance for each driving pin. As noted above, STA is run with a modified 134 nm (EPE tolerance tightest on gate poly and loosest on field poly) library with pin capacitances corresponding to 144 nm (loosest EPE tolerance) to remain conservative after slack budgeting. We use *CPLEX v8.1* [49] as the mathematical programming solver to solve the budgeting linear program. Two types of benchmarks are involved in our experiments: (i) large designs with a "wall" of critical paths, e.g., r4_sova in

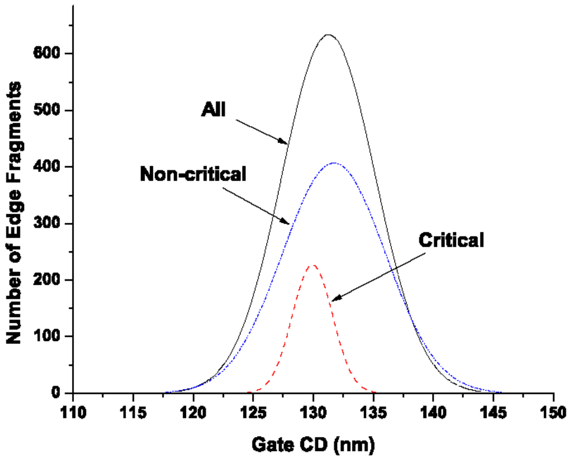


Figure 3.8: Gate CD distribution for c432. Gates with budgeted 4nm EPE tolerance are labeled critical gates while others are labeled as non-critical. The y-axis shows the number of fragments of gate edges with a given printed CD. .

Table 3.2; and (ii) circuits with fairly small sizes, e.g., benchmarks except r4_sova. For (ii), a single iteration is efficient to solve the budgeting problem; for (i) however, more iterations may be necessary because some paths which are potentially critical but are not reported due to the constraint of maximum number of critical paths may become top critical later on as they are not treated as optimization objects by the slack budgeting algorithm, resulting in performance degradation. One possible solution to this problem is to perform iterations to selectively include those paths that may cause performance degradation, as slack budgeting objects. Another simple but not as efficient option is to increase the constraint of maximum number of critical paths in the slack report. We deploy a hybrid way for r4_sova in our case, i.e., the constraint on the initial number of critical paths is increased from 500 to 10000, then in each iteration 5000 more paths that are potentially critical are included for slack budgeting. After 8 iterations the performance degradation due to the selective OPC is reduced to less than 1% (first iteration gives 4.3% performance degradation).

The extracted CD variation for test case $c432$ after EPEMinCorr OPC is shown

Table 3.3: Impact of EPEMinCorr optimization on cost and CD. All runtimes are based on a 2.4GHz Xeon machine with 2GB memory running Linux.

Testcase	Traditional OPC Flow				EPEMinCorr Flow							
	CD Distribution		OPC	Delay	Budget-	CD Distribution (nm)				OPC	Delay	Normalized
	All Gates (nm)		Runtime	(ns)	ing Run-	All Gates		Crit. Gates		Runtime	(ns)	MEBES
	mean	σ	(hr)		time (s)	mean	σ	mean	σ	(hr)		Volume
c432	130.9	1.55	0.2643	1.33	1	131.3	3.90	129.9	1.67	0.2047	1.33	0.87
c5315	130.2	1.83	1.261	1.94	3	131.7	4.70	129.7	1.89	1.180	1.94	0.82
c6288	129.7	1.52	3.275	5.21	9	131.4	4.45	129.7	1.27	2.697	5.21	0.86
c7552	129.6	1.65	1.856	1.59	4	132.0	4.77	130.1	1.99	1.428	1.59	0.81
alu128	130.4	1.63	13.89	3.28	11	131.5	4.93	130.8	2.04	9.215	3.28	0.80
r4_sova	130.1	1.98	38.65	8.19	29,648	131.9	5.00	130.0	1.75	23.32	8.26	0.80

in Figure 3.8. The distributions show that Calibre is able to enforce assigned tolerances very consistently. A tighter CD distribution for critical gates is achieved while non-critical gates (which can tolerate a larger deviation from nominal) have a more relaxed (and hence less expensive to implement) gate length distribution. Table 3.3 compares the runtime and data volume results for EPEMinCorr OPC and traditional OPC. For relatively small circuits, a single iteration of the budgeting approach ensures that there is no timing degradation going from the traditional to the EPEMinCorr flow, and the budgeting runtimes are negligibly small, ranging from 1s to 11s. For large designs especially those with a "wall" of critical paths, iterations may be required to avoid performance degradation and the sum of budgeting runtimes of each iteration may reach several hours (7 hours for r4_sova). The important result is the amount of mask cost reductions achieved whether measures as runtime of model-based OPC or fractured MEBES data volume. EPEMinCorr flow reduces MEBES data volume by 13%-20%. Such reductions directly translate to substantial mask-write time improvements. OPC runtimes are improved by 6%-39%. These percentage numbers translate to a huge absolute TAT savings. For instance, the EPEMinCorr flow saves 16.3 hours compared to the traditional OPC flow on a 34000 gate benchmark.

3.4 Conclusions and Future Work

We have proposed and implemented a practical means of reducing masks costs and the computational complexity of OPC insertion through formalized performance-driven OPC assignment. In particular we focus on the use of edge placement errors to drive OPC insertion tools and leverage EPEs as the mechanism to direct these tools to correct only to the levels required to meet timing specifications. An iterative linear programming based approach is used to perform slack budgeting in an efficient manner. This formulation results in a specific slack budget for each gate which is then mapped to allowable critical dimensions in the standard cell. Finally EPEs are generated from the CD budget and tags are placed on gates to indicate to the OPC insertion tool the appropriate level of correction. Our results on several benchmarks ranging from 300 to 34000 cells show up to 20% reductions in MEBES data volume which is frequently used as metric for RET complexity. Furthermore, the runtime of the OPC insertion tool is reduced by up to 39% - this is critical since running OPC tools at the full-chip level is an extremely time-consuming step during the physical verification stage of IC design.

In future technologies allowable CD tolerances may be set more by bounds on acceptable leakage power than by traditional delay uncertainty constraints. We plan to incorporate power constraints into our formulation. Moreover, we plan to extend the EPEMinCorr methodology for field poly features. Impact of field polysilicon shapes on performance comes from their overlap with contact layer. So field poly extensions to EPEMinCorr will have to evaluate error in terms of contact coverage area. Expensive masking layers include diffusion, contact, metal1 and metal2 besides polysilicon. The performance impact of OPC errors on these other layers can also be computed and consequently EPEMinCorr methodology extended.

Another direction of work is exploring other degrees of freedom in OPC besides EPE tolerance which have a strong effect on mask cost. Two such parameters are fragmentation and minimum jog length.

In a follow-up work of an industrial scale of application [50], a methodology similar to EPEMinCorr was used to optimize mask cost for a big design block. The resulting OPC'd layout went through dummy mask write at a mask shop. The authors reported 25% shot count reduction and up to 32% reduction in mask write time.

CHAPTER IV

Advanced Timing Analysis Based on Post-OPC Extraction of Critical Dimensions

4.1 Introduction

In contrast to the traditional design flow as illustrated in Figure 4.1, in this chapter we present a methodology for post-OPC embedded static timing analysis by extracting residual OPC errors from a state-of-the-art placed and routed full-chip microprocessor layout and for deriving actual (calibrated to silicon) L_{gate} values. The implementation of this automated flow is achieved through a combination of post-OPC layout back-annotation and selective extraction from the global circuit netlist. This approach improves upon the traditional design flow practices where ideal (drawn) L_{gate} values are employed, which leads to poor performance predictability of the as-fabricated design. When post-OPC L_{gate} values are used, timing analysis results indicate substantial differences in the order of speed path criticality, with the worst-case slack increasing by 36.4%. Pin slacks of critical instances are worsened on average and a larger number of critical cells are reported in terms of both total cell count as well as the number of cell types involved on critical paths.

These results point to the need for adoption of process-based simulation results within the traditional design flow so that design optimization is better targeted and effort is not wasted optimizing paths that are not actually critical in the fabricated part. In addition, this flow can be used by OPC engineers to locate critical layout patterns across the full-chip layout where OPC could not achieve the desired L_{gate} control. Once identified, a calibrated OPC algorithm could then be applied locally to attain large savings on full-chip OPC run time. This is particularly essential in designs with matching transistors (e.g., mixed signal designs, clock generation in microprocessors). Furthermore, this flow can easily identify cells either with high frequency of occurrence in critical paths or with large pin slacks, so that various optimizations such as well-calibrated localized OPC algorithms, multi-threshold voltage assignments, gate sizing, etc., can be applied directly to the full-chip layout to improve overall chip performance. While these optimizations are already applied prior to sign-off, a post-OPC timing analysis will provide more accurate CD data and allow for further improvement.

4.2 Overview of Methodology

CD distortions introduced by proximity effects have a significant impact on circuit performance. OPC is used to compensate for these CD distortions yet it becomes a major source of systematic variability itself. Analysis in [51] demonstrated that systematic intra-chip L_{gate} variability is the main cause of speed degradation for large circuits, especially those with a large number of critical paths and short logic depths. In addition, designers place significant emphasis on fixing worst-case speed paths that are identified in timing simulations based on ***ideal*** L_{gate} values (potentially corner-based) which may be quite misleading in the sub-wavelength lithography regime. Process simulations showing aerial images of printed features are used to check the effectiveness of OPC and can provide accurate L_{gate} predictions after a layout is complete, pointing to the possibility of post-OPC timing verification and OPC recalibration. This section describes a newly developed methodology as illustrated in Figure 4.2.

- *Process CD Simulation for Critical Gates.* The starting point of the flow is a post-OPC layout on which process CD simulations can theoretically be performed across the entire chip although this would be very time consuming. The OPC layers used for mask creation are generated based on carefully calibrated optical and resist models for the particular lithographic conditions. In our analysis we focus on the systematic poly gate variation introduced by RETs/OPC; the presented flow could easily be extended to include variation of metal layers as well. We perform a selective process CD simulation only for gates on critical paths as predicted in the full-chip timing analysis, tagging them with a *critical layer*. These tagged critical gates are then extracted along with the peripheral geometries within a certain distance (i.e., the optical diameter, beyond which geometries have no impact for the given optical lithography system) to account for optical interactions from adjacent cells. We do this on a path-by-path basis to facilitate future analysis and diagnosis. We perform aerial image simulations with Mentor Graphics Calibre RET tools for each of the critical paths, and for simplicity, the Si-based CDs are extracted and defined as the dimension at the center point for each transistor. Each process CD value is identified with the GDS coordinates of the corresponding transistor which are used to map process CDs back to the circuit netlist.

- *Library Re-characterization.* This step updates the cell timing library with Si-based process CDs by creating a *location-aware* SPICE netlist for each cell.

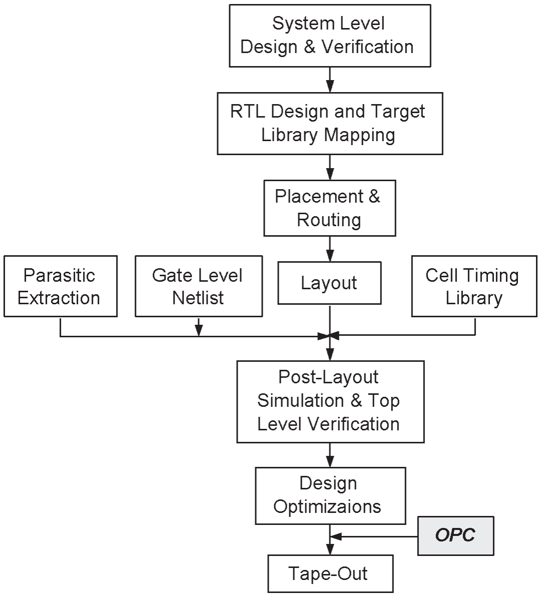

System Level Design & Verification
RTL Design and Target Library Mapping
Placement & Routing
Parasitic Extraction
Gate Level Netlist
Layout
Cell Timing Library
Post-Layout Simulation & Top Level Verification
Design Optimizaions
OPC
Tape-Out

Figure 4.1: Typical design flow in sub-micron VLSI design.

After the process CDs for critical gates are found and back-annotated, the original library is expanded with re-characterized cells. Each transistor is identified by its lower-left coordinates relative to the origin point of the cell. The origin point is determined by the cell's placement orientation and the relative coordinates are calculated with additional info on gate geometry. A special LVS is performed to extract the location of each transistor within a cell in the library. After that, the L_{gate} values in the SPICE netlists are modified with the extracted Si-based process CDs. Finally, SPICE simulations are performed for timing re-characterizations to account for the impact of systematic L_{gate} variations. We do not consider changes in cell parasitics due to gate CD variations which are expected to be second order compared to the drive current changes that are modeled. These re-characterized cells, corresponding to their locations and distinguished by cell names, are combined with a typical cell library for timing analysis.

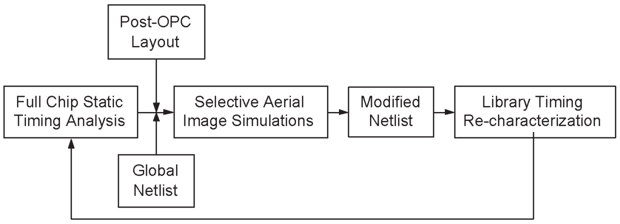

Post-OPC Layout
Full Chip Static Timing Analysis
Selective Aerial Image Simulations
Modified Netlist
Library Timing Re-characterization
Global Netlist

Figure 4.2: Methodology overview.

- *Static Timing Analysis with Process CDs.* The global chip netlist is modified to map to the expanded cell library. Then a full-chip timing analysis is performed using commercial tools and the top critical paths are reported. These critical paths may include cells that were not re-characterized in the previous step (i.e., they are on paths that the original pre-OPC timing analysis did not flag as critical) in which case re-charac-terization of newly critical cells will be necessary.

- *Results Comparison with Traditional Timing Analysis.* Timing analysis results based on process-simulated CDs are compared with those simulated in the traditional flow with *ideal* gate CDs at typical operating conditions. We choose to make comparisons at the typical process corner rather than the worst process corner to avoid the overly pessimistic worst-case process corner and also since post-OPC CD extractions are performed at best focus conditions, corresponding to a typical process. Therefore, our goal is to determine whether the systematic L_{gate} variations introduced by RET/OPC have a significant impact on typical corner performance.

4.3 Experiments and Results

We focus on a state-of-the-art microprocessor design in 90 nm technology, using a 193nm optical lithography system. The optical diameter outside of which neighboring geometries have no impact, is set as 4 μm based on experimental results. Due to newly discovered critical cells in each timing analysis with Si-based process CDs, three iterations of cell timing re-characterizations are performed with runtime of approximately three hours per iteration. In the final speed path report, under 2% of total critical cells have not been re-characterized using their extracted CDs; these cells appear only in paths with lower critical ordering. The total runtime for this flow is kept low due to the following: (1) only a small subset of all instances ($\sim$0.5%) are selected for post-OPC process CD simulations; (2) process CD simulations for each of the critical paths, as well as timing re-characterization for critical cells, is done in parallel; and (3) most cells on critical paths are simple, such as inverters and MUXes, so that little runtime for cell timing re-characterizations is required. In the experimental results below we outline the need for a post-OPC timing verification design flow with the potential for re-optimization or further OPC assignment. This flow can be supported with a fast and localized OPC assignment algorithm; for each design optimization requiring layout modification, the OPC layer will be changed although only slightly. With the majority of the layout geometries remaining the same, there is no need to regenerate OPC patterns for the bulk of features and thus incremental OPC capabilities are desirable.

4.3.1 Critical Paths Reordering

We define *critical paths* as paths with $slack \leq 0$. These paths are rank ordered with the most critical ones first and we assign path IDs according to this ordering (e.g., the most critical path is path ID 1). The experimental results show that in general the Si-based timing report lists more paths (2.7X) as critical than the ideal L_{gate} based timing report, with average path slacks worsened by 24.4%. Figure

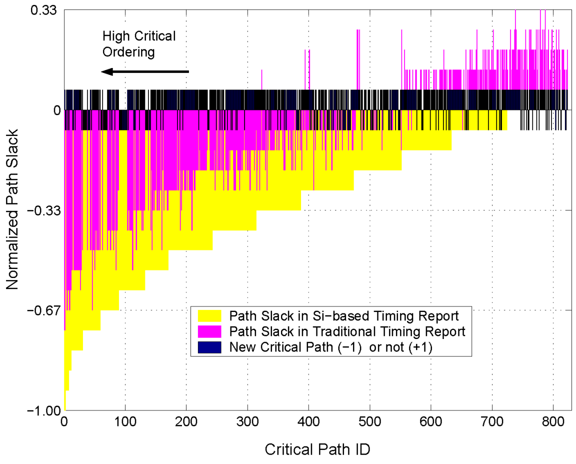


Figure 4.3: Critical path reordering.

4.3 shows the slack distributions for critical paths from the Si-based timing report and the traditional timing report, and also provides a binary indicator with value +1 meaning that the path appears in both of timing reports while a value of −1 indicates that the specified path only appears in the Si-based timing report. The y-axis relates normalized path slack where the slack of the most critical path in the Si-based timing report is the reference value. In general, we observe more negative slack paths in the Si-based timing report. A considerable number of paths with high critical ordering are not reported in the traditional timing analysis (indicated as −1). We now discuss these paths separately based on whether they appear in both reports (old paths) or only in the Si-based timing report (new paths).

- *Old Paths.* For paths existing in both timing reports, we examine the critical ordering difference. The sign (negative/positive) of the critical ordering difference indicates the shifting direction (more/less critical, respectively) of that critical path in the Si-based timing report compared to the traditional timing report, while the absolute value indicates how many paths it passes to achieve the new critical ordering. As shown in Figure 4.4 the range of critical

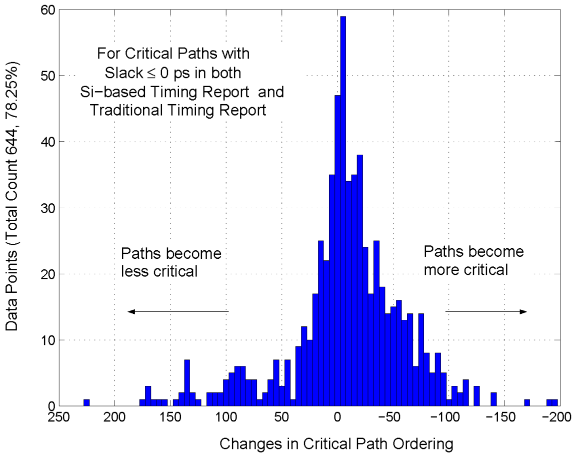


Figure 4.4: Reordering of critical paths.

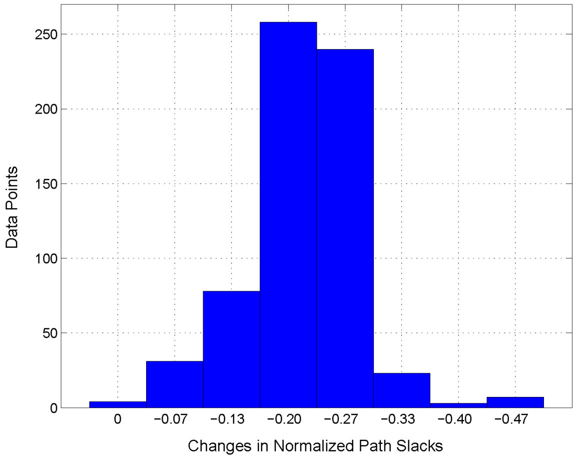


Figure 4.5: Worsened slacks for *old* critical paths.

ordering difference is between −195 and +224, implying that paths become up to 195 paths more critical and 224 paths less critical, respectively. Even for paths with positive changes (less critical overall) they typically become more critical in the sense that their absolute slack is often worse than that found in the traditional timing report. The path slack changes are shown in Figure 4.5, where the worst case slack is increased by 36.4%, and the average slack change is 0.22 with a maximum shift of 0.47 (again these values are normalized to the worst-case slack observed in the Si-based timing report).

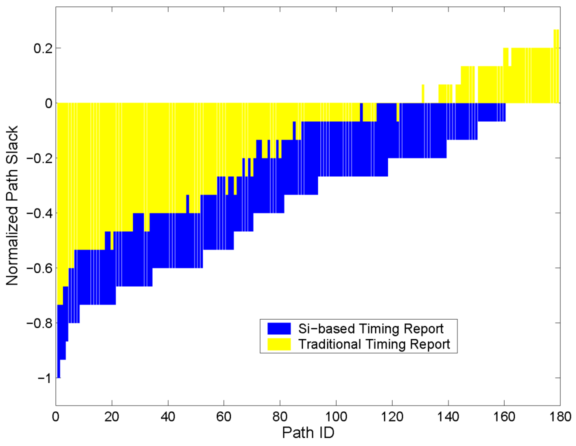


Figure 4.6: Slacks for *new* critical paths.

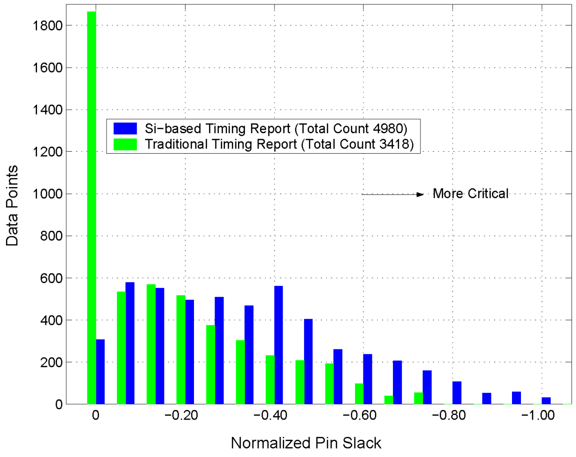


Figure 4.7: Pin slacks for critical cells.

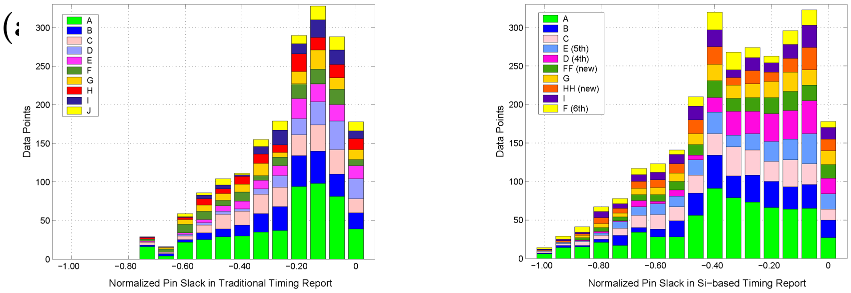
- *New Paths.* When post-OPC CDs are used, 21.8% new critical paths are identified. Some of these new paths are highly critical with large slack violations according to the post-OPC analysis (see Figure 4.6). Path slacks in new path cases are worsened by 0.22 (normalized) on average. As a result, using the traditional timing analysis as a guideline for design optimization will be very misleading in that certain paths will be not be considered for resizing, etc. although they will actually be critical post-fabrication.

4.3.2 Tracing for Critical Cells

To investigate whether specific cells consistently appear on critical paths, and thus are good candidates for cell re-design and/or close attention during OPC assignment, we define critical cells as instances with pin slack ≤ 0 and compare pin slacks and the top ten most frequently used critical cells under the two timing analysis schemes. As indicated in Figure 4.7, the distribution of pin slack for critical cells is wider in the Si-based timing report, with a shift to larger slack violations on average. More cells become critical in terms of the total number (increasing by 45.7%) as well as cell types (increasing by 12.7%). These results indicate that more design effort on optimization and cost will be involved. By identifying the most frequently used critical cells, more effort may be placed on optimizing the delay of these cells, and better CD control may be achieved by using a more OPC-friendly cell layout. We find that 53% and 50% of total critical cells in traditional and Si-based timing analysis, respectively, are made up of the top ten critical cell types. As shown in Figure 4.8 (a) and (b) we observe that the ordering of the top ten most frequently used critical cell types changes slightly while the average pin slack becomes 42.7% worse in the post-OPC analysis. From these results we conclude that library-level design optimizations should be made based on a Si-based analysis.

4.3.3 OPC-Driven Analysis

Model-based OPC corrects the layout on a point-by-point basis, considering all neighboring features as well as the complex interactions between the stepper and mask. By decomposing the edges of each feature into small fragments, the overall OPC quality can be improved through more fine-grained edge movements. Such movements for each fragment are constrained by neighboring geometries, and the



(a)
Ideal L_{gate} based timing report

(b)
Si-based timing report

Figure 4.8: Top ten most frequently used cells in ideal L_{gate} based timing report vs. in Si-based timing report.

use of a given prioritization scheme for each fragment may limit the convergence of the overall OPC correction. For instance, the number of iterations of edge movements required to converge below the OPC-specified residual error might vary from fragment to fragment. This issue can be dealt with by adjusting the priority scheme based on the identification of these problematic layout patterns. Our flow provides a way to compute the normalized OPC residuals for instances on critical paths and associate them with that cell's occurrence frequency. As shown in Figure 4.9 there are several cells with both large CD errors and a high frequency of occurrence. In these cases either customized OPC recipes can be created or cell layout patterns could be adjusted to improve the overall printability.

To further diagnose the impact of neighboring features, we examine layout patterns that cause gates to exhibit large CD variations after OPC is applied. Figure 4.10 shows an example of the impact of environment for three instances of a single (identical) gate polygon in terms of L_{gate} errors. The highlighted gate is an inverter and we focus on the P-transistor (bottom device). Two heuristic rules to enhance manufacturability are often applied: (1) use of a single pitch on the critical layer,

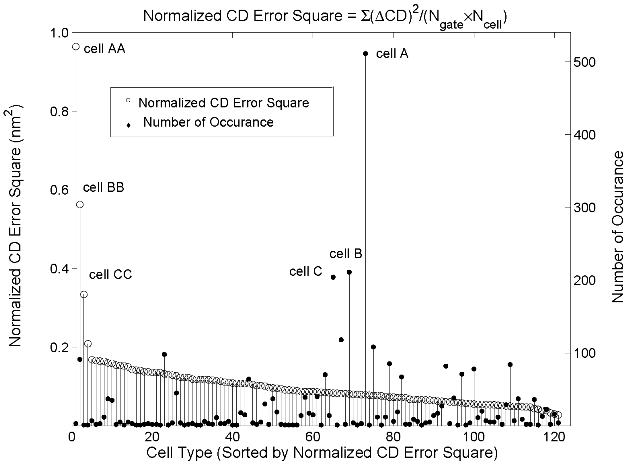
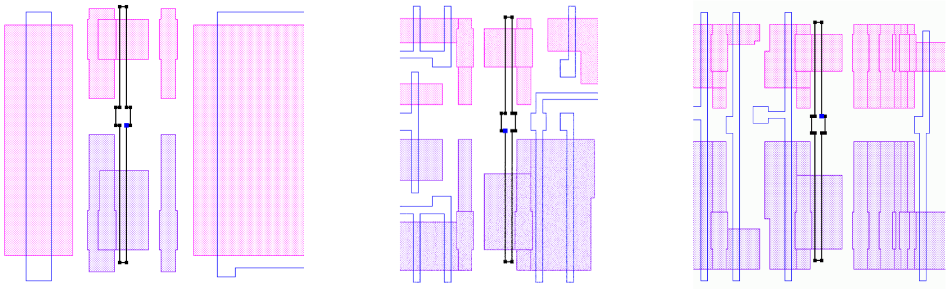


Figure 4.9: Normalized L_{gate} residual for cells on critical paths.



(a) **(b)** **(c)**

CD error: 7.33X **6X** **1X**

Figure 4.10: Impact of environment on L_{gate} residuals of the same gate. The error is extracted for the P-transistor only in the highlighted inverter and is normalized to (c).

and (2) avoidance of non-rectilinear shapes (e.g., L's or T's). This is reflected in Figure 4.10. The layout in (a) is undesirable due to the large decoupling capacitances nearby while (b) is also poor due to the many L shapes in neighboring poly. For the same inverter layout in (c) the neighborhood around the P-transistor leads to better printability, reducing the L_{gate} errors (normalized) by 7.33X compared to (a).

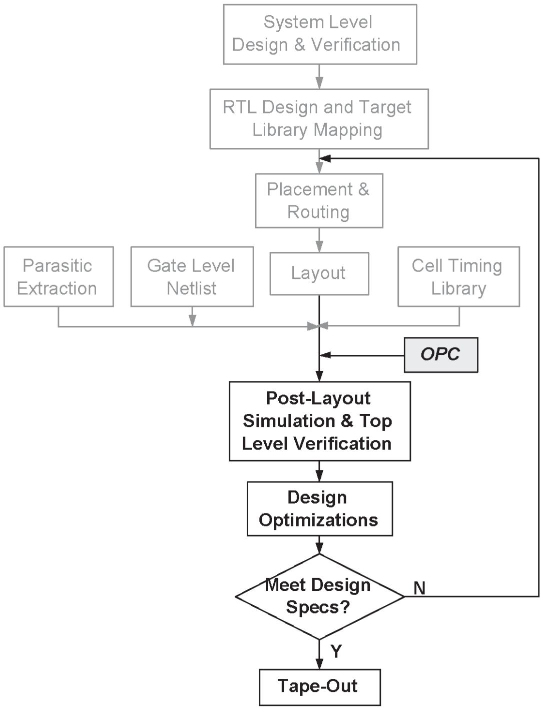


Figure 4.11: Design flow with post-OPC verification.

4.4 Results Summary

An automated flow for post-OPC performance verification is presented. Experimental results on a 90nm microprocessor show significant changes in the timing analysis compared to the traditional methodology of performing final timing analysis before OPC application. The number of critical paths increased by 170% while the worst-case slack violation increased by 36.4%. Among all critical paths found in the post-OPC flow, 21.8% were not reported in the traditional timing analysis. These changes demonstrate that traditional performance analyses are no longer valid in nanometer-scale designs that rely on complex resolution enhancement technologies.

A post-OPC performance verification design flow can enable process variation-aware design optimization as well as drive tradeoffs when significant variability is unavoidable. Using the framework presented in this chapter, one design flow based on post-OPC verification with only slight modifications to the traditional flow is given by Figure 4.11. The post-OPC process CD extraction is performed at the same stage when interconnect parasitics are extracted and back-annotated. In this way the large systematic gate CD variations due to RET/OPC are taken into account during static timing analysis to achieve more accurate performance predictions. The methodology described in this chapter is based on the ability to tag critical gates such as those on critical paths or matching gates so that specific corrections can be applied to these gates to achieve better CD control rather than attempting to reduce gate CD variation in all scenarios. This type of back-annotation can be easily extended to the metal and contact layers in order to enable RC extraction based on Si-based post-OPC dimensions. Integrating the OPC step into the design flow effectively will allow design-time optimizations to be aware of the manufacturing process and achieve improved performance and yields in the final as-fabricated design.

CHAPTER V

Toward a Methodology for Manufacturability-Driven Design Rule Exploration

5.1 Introduction

Although RETs have historically been a strictly post-layout procedure, they now need to become part of a cohesive design flow in which libraries and layouts are optimized directly based on conflicts discovered by the RET tool [20]. This “trickle-down” effect of RETs towards the design process is also manifested by more conservative design rules, particularly for the critical polysilicon layer. In particular, the ability to print very tight pitches as well as print a wide range of pitches in a given layer is very difficult for subwavelength lithographic systems. As a result, there is a trend towards limiting the range of allowed pitches in the polysilicon layer [21]. This type of restricted design rule (RDR) seeks to enforce a particular style of layout that is known to be highly manufacturable. As with any design rule, it is a tradeoff between manufacturability and performance, where performance can be measured as layout density, delay, power, etc. By nature, these RDRs seek to push the tradeoff more in favor of the manufacturing side, sacrificing performance in the process. Despite the move towards RDRs, there has been no comprehensive and systematic study of their

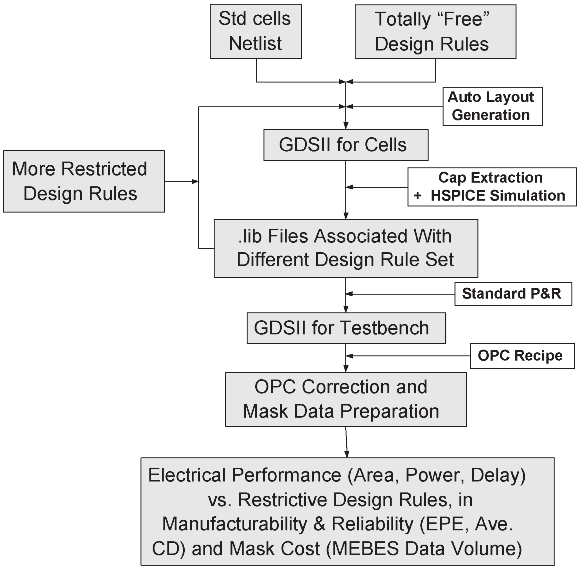

Std cells Netlist
Totally "Free" Design Rules
Auto Layout Generation
More Restricted Design Rules
GDSII for Cells
Cap Extraction + HSPICE Simulation
.lib Files Associated With Different Design Rule Set
Standard P&R
GDSII for Testbench
OPC Recipe
OPC Correction and Mask Data Preparation
Electrical Performance (Area, Power, Delay) vs. Restrictive Design Rules, in Manufacturability & Reliability (EPE, Ave. CD) and Mask Cost (MEBES Data Volume)

Figure 5.1: ASIC design flow targeting RDR evaluation.

expected impact on manufacturability and performance. This approach presents an analysis of various RDR sets applied within an ASIC design methodology. We seek to minimize mask costs, maintain circuit performance, and enhance feature printability and reliability.

5.2 RDR Evaluative Methodology

5.2.1 ASIC Design Flow Targeting RDR Evaluation

To evaluate the performance and manufacturability impact of restricted design rules, we set up the design flow shown in Figure 5.1. Initially we have a set of default design rules based on IBM 0.13 μm technology and a pruned standard cell netlist containing basic cell types such as BUF, INV, NAND, NOR, AND, OR, AOI, and OAI. We then create GDS representations for each cell with an automatic layout generation tool. After parasitic extraction, each cell is characterized for both timing and power performance to generate a .lib file. At this point we have the necessary infrastructure to proceed to synthesis/place and route (P&R).

The library generation process is repeated by altering the set of design rules through inclusion of a single candidate RDR, such as adding stricter requirements for poly gate spacing, minimum poly line end extension, etc. The goal of these added RDRs is to improve the final printability and reliability with as little performance impact as possible. We re-generate layouts and .lib files for a number of candidate RDR sets, then perform synthesis/P&R, and obtain timing, power, and area reports after back-annotation for several benchmark circuits. Note that the circuit topology is unchanged in all implementations of a given benchmark. That is, we do not re-synthesize the circuit with a new library but instead map the gate-level netlist to a new .lib and proceed with the back-end of the typical ASIC flow.

After circuits are placed and routed for each individual library, we perform OPC for each layout with a general but comprehensive model-based OPC recipe containing information such as the line end correction procedures, concave and convex corner correction instructions, etc.[1] The amount and impact of the applied RET is a function of the circuit layout which in turn depends on cell layout among other factors. Thus, we can evaluate how specific design rule changes impact both circuit performance (delay, area, power) and manufacturability/printability/mask cost as measured on MEBES data volume, histograms of resulting edge placement errors (EPE), etc. The next section contains more details about EPE and MEBES data volume. Specific EDA tools used within this overall flow are:

- Layout automatic generation - Prolific *Progenesis* [52];

- Physical capacitance extraction - Mentor Graphics *Calibre xRC* [53];

[1] All OPC-related results shown in this chapter are extracted based on simulations on layout test patterns with industry optical and process conditions.

- Timing and power characterization tool - Synopsys *HSPICE* and *Powerarc* [54];
- Synthesis/P&R - Synopsys *Design Compiler* [54] and Cadence *Silicon Ensemble* [55];
- Back-annotated timing simulation - Synopsys *PrimeTime* [54];
- OPC layer generation, EPE extraction, and mask data preparation (MDP) - Mentor Graphics *Calibre RET* [53].

This section discusses candidate design rules that can be altered in an attempt to improve printability or manufacturability. Modern design rule manuals have hundreds of entries; we examine just a handful of possible RDRs on the polysilicon layer, which is the most critical for transistor performance, in order to draw concise conclusions. In particular, spacing between features is one of the most important rule types that affects circuit manufacturability: the light field of a given feature is greatly affected by the location of neighbor features, leading to CD variations that can result in loss of parametric yield. Most of the design rules we investigate therefore deal with either intra-layer or inter-layer spacings. As another example, minimal polysilicon overlap of diffusion is a critical design rule as it ensures that the edges of a MOSFET maintain consistency in dimensions with the interior portion of the channel.

Our starting point is a default flexible design rule set within which all spacing rules are at their minimum values and bent gates or 45-degree routes of poly are allowed (i.e., bentgate is “on”). From this point we construct restricted design rule sets by first turning bentgate “off” and then investigating the following rule categories: increased minimum poly to poly spacing, increased minimum field poly to diffusion spacing,

Table 5.1: RDR default and modified values (note that the corresponding rule names appearing in all following figures are included in parentheses after values)

Rule name	Default(μm)	Modified(μm)	
Bentgate line width	"off" 0.12	"on", *baseline* 0.12 (bentgate)	"on" 0.14 (bent_w14)
Poly_poly space	0.20 (sp_20)	0.24 (sp_24)	0.28 (sp_28)
Poly_diffusion space	0.08	0.10 (pdsp_10)	0.12 (pdsp_12)
Poly end extension	0.28	0.34 (povg_34)	0.40 (povg_40)

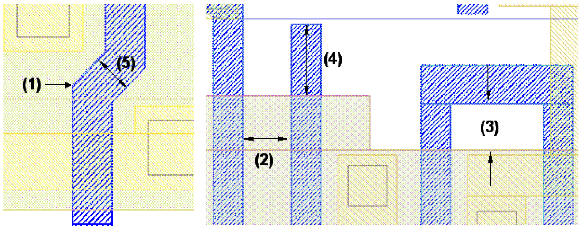


Figure 5.2: Layout illustrations of RDR candidates.

larger minimum poly line end extension beyond diffusion, and also turning bentgate back on while increasing the minimal allowable linewidth in a bent gate structure. Figure 5.2 depicts layouts corresponding to the RDR candidates we investigate:

- Bentgate "on" as *baseline* (Figure 5.2 (1));
  - Bentgate line width (Figure 5.2 (5)).
- Bentgate "off"
  - Poly to poly spacing (Figure 5.2 (2));
  - Poly to diffusion spacing (Figure 5.2 (3)); and
  - Poly end extension (Figure 5.2 (4)).

5.2.2 RDR Candidates

Once the form of the specific RDRs are decided, we then seek to find the range of values that the RDRs should take on so that we can expect printability improvements. For example, it is clear that poly to poly spacing cannot be set below the value in the default design rule set since that spacing has already been determined to be the minimum allowable that ensures decent printability. To create more conservative design rules, we want to examine the impact of larger poly to poly spacings. However, if this spacing becomes too large it can actually jeopardize manufacturability since many modern lithography systems are not adept at printing intermediate pitch values [20].

To investigate the range of poly pitches that print well using our (fixed) OPC recipe, we use edge placement errors (EPE) as a quantifying metric. EPE is a common measure of how closely a printed feature actually reflects the corresponding designed feature. Usually EPE has larger magnitude near the ends (along the width dimension) of a transistor gate; this implies that in small-width gates the impact of CD variability is relatively larger and that the edges of a device may exhibit substantial leakage currents since a smaller-than-nominal channel length leads to exponentially more subthreshold leakage through short-channel effects [56]. This also points to line end extension rules as a possible RDR. As indicated in Figure 5.3, with a more restrictive minimum poly to poly spacing rule the EPE distribution of a NAND2X2 (2-input NAND of size 2) without OPC shows a consistent left shift until it reaches approximately 0.70 μm at which point it then moves back to tighter distributions. In general, with advanced off-axis illumination approaches such as annular or quadrupole illumination there could be several pitch ranges where the optical diffraction results in poor printed images (in this chapter, this manifests as

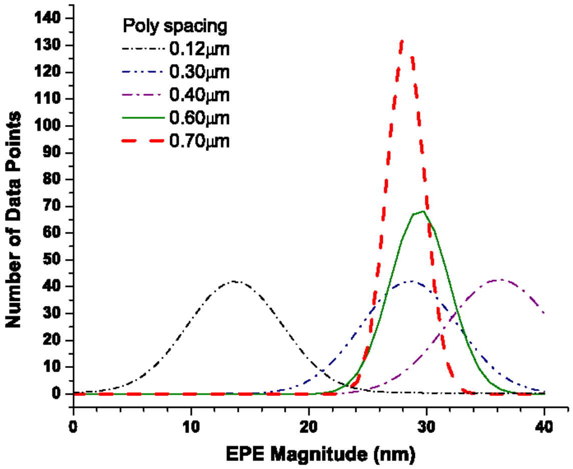


Figure 5.3: Impact of pitch on the EPE histogram of a NAND2X2 without OPC.

larger EPEs). These pitch regions, determined by the details of the entire lithography process, are sometimes referred to as the *forbidden pitch* ranges, and should be avoided by IC designers. As can be seen, the EPE (or CD) variation becomes smaller for isolated lines but the average value increases. This behavior can be attributed to the fact that the radius of influence of optical diffraction effects extends to approximately $0.6\mu m$ and any pitch above that prints similarly poorly [40]. In our study, we define 0.42 μm to 0.72 μm (equivalent to 0.30 μm to 0.60 μm poly spacing where poly width is set to its minimum value of 0.12 μm) as our forbidden pitch ranges. In our study, we investigate RDRs that take on the values shown in Table 5.1.

5.2.3 Evaluation Metrics for Manufacturability/Cost

As described above, EPEs are used as a measure of OPC effectiveness with a goal of zero EPE for all polygons forming transistor gates. However, an "edge" placement error does not provide complete insight to the actual critical dimension or CD - two edges (or EPEs) are needed to determine CD, indicating the need to localize each EPE and match it with the EPE value on the immediately opposing side of the polygon. Considering that each single transistor may actually have multiple CDs

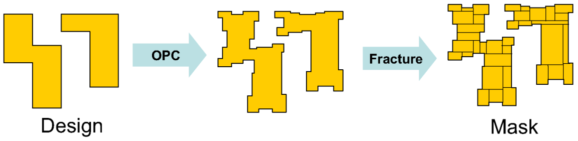


Figure 5.4: Mask data preparation (post OPC).

due to irregular printed image (i.e., transistor gate lengths can be non-uniform along the width dimension), we find an average CD for each transistor by calculating the gate and active overlap area with the simulated printed image and dividing it by the measured gate width. When CD is reported in the remainder of this chapter, it refers to the average gate-length calculated in this manner.

Moreover, we use the mask writer format (MEBES) data volume to evaluate the complexity of the resulting mask for the critical layer. We use this as an OPC or design-cost metric since GDSII files must be fractured into MEBES format (see Figure 5.4[2]) during mask data preparation and this step has become a serious bottleneck due to large figure counts from RETs. For out purposes, MEBES data volume reflects the complexity of an OPC layer which is impacted by the design rules used for that layer. In summary, EPEs and averaged CD variation are used as criteria for manufacturability while we use MEBES data volume to evaluate OPC cost.

5.3 Testbed and Experimental Results

We use IBM 0.13 μm CMOS technology in the following simulations and ISCAS85 circuits as benchmarks to evaluate our .lib files. The descriptions of these testcases are as follows:

- c7552 - a 32-bit adder/comparator, the largest circuit in ISCAS85;

[2] Figure courtesy M. Reiger, Synopsys Inc.

- c6288 - a 16×16 multiplier; and

- c5315 - a 9-bit ALU.

5.3.1 Impact of Defocus

Defocus in the lithography system is a key parameter that strongly affects the printability of fine resolution images since it determines the process window (defined as the range of exposure dose and defocus within which acceptable image tolerance is maintained). When the absolute amount of defocus exceeds a certain "best-focus" value, the printed features can go out of the CD variation tolerance, since only a limited depth of focus (DOF) is allowed in a lithography system. Four different defocus values, 0, 0.1 μm, 0.2 μm, and 0.3 μm are tested in our experiments. If not specifically mentioned, all designs are simulated at 0 defocus with a comprehensive model-based OPC recipe.

Impact of Defocus on Maximum EPE Levels

Figures 5.5 and 5.6 show the extracted maximum gate CD EPE tolerance (i.e., the maximum EPE observed for any gate in the specified design). As can be seen, with constant defocus, as poly spacing increases from 0.20 μm to 0.24 μm the maximum EPE tolerance either remains constant or decreases, demonstrating an impact on CD variation of this design rule. The maximum observed EPE nearly doubles as defocus rises to 0.3 μm, indicating that focus variation is a large contributor to CD variation as has been pointed out elsewhere [57]. Looking at the range of RDR sets, we first see that the default design rule set leads to very large EPEs, up to 40 nm for a 130 nm process. Furthermore, the simple removal of bent gates (shown as RDR set "sp_20") helps dramatically while further changes to the design rule set can also improve the worst-case EPE. The best RDR sets from these data sets are the "povg_34" set which increases the minimum poly overlap of active by 60 nm and the "sp_24" set which relaxes the poly-to-poly spacing by 20% relative to the baseline. We also note that the relaxation of some design rules (e.g., "pdsp_10") can actually worsen printability of some difficult features in a layout compared to the "sp_20" design rule set.

Impact of Defocus on CD Distribution

To assess the impact of focus variation on CD, we use aerial image calculation which models optical effects[3]. The intensity level for aerial image simulation is fixed at the value which gives best aerial image for the "isofocal spacing"[4] at best-focus. Separately, the isofocal spacing is computed to be 200 nm by defocus simulations of a simple test structure. This intensity level was maintained constant with defocus. We then extracted the averaged CDs and their variation from aerial image contours, as shown in Table 5.2. The 200 nm poly-spacing rule prints the best through-focus as it results in cell layouts with inter-device spacings closest to the isofocal spacing. This suggests that intelligent choice of the min-poly spacing which is cognizant of the isofocal spacing as defined by the process can improve defocus characteristics of the design.

Impact of Defocus on Functional Yield

In [2], the allowed variability in physical gate length is fixed at 10%. This translates to an average maximum allowable EPE of 5% on each edge of the gate. Note that it is possible for a printed gate to have larger EPE on both sides and still maintain a nominal L_{gate} (i.e., positive and negative EPEs may appear simultaneously

[3] We ignore resist effects in this analysis as Calibre models are calibrated at best-focus and may not yield accurate print image results for defocus conditions.

[4] The spacing for constant width that has nearly zero variation through a range of defocus levels.

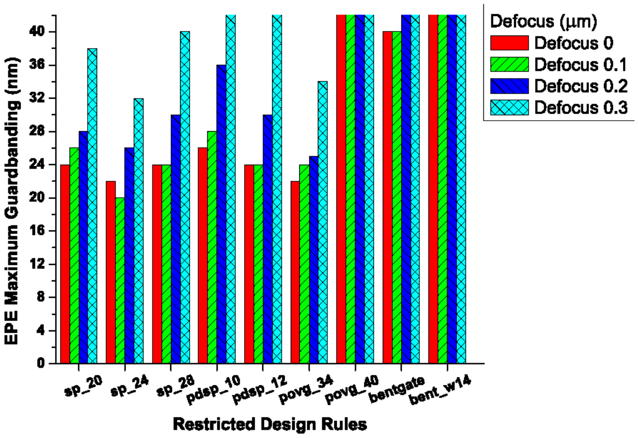


Figure 5.5: Impact of defocus on c6288.

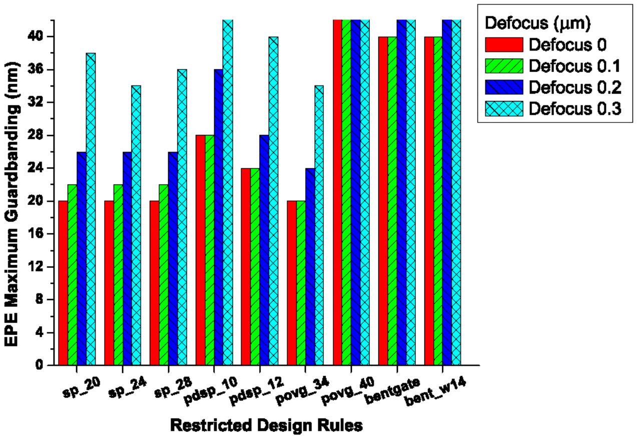


Figure 5.6: Impact of defocus on c7552.

and cancel the effect of each other) but this increases the possibility of functional failure in a relatively dense circuit. To examine the fraction of printed gates in our benchmark circuits that meet this ITRS requirement, we define functional yield to be the percentage of total gates that print with less than 5% EPE for all fragments of the gate.

As seen in Figure 5.7, for nearly every RDR set the functional yield is rather sensitive to focus variation. This is expected since printability gets markedly worse when features are out of focus. However, we find that the RDRs associated with increased poly line-end extensions (povg_∗) show dramatically less sensitivity of functional

Table 5.2: Impact of defocus on extracted CD mean and variation (unit: nm)

	Defocus							
	0		0.1(μm)		0.2(μm)		0.3(μm)	
RDR	Mean	σ	Mean	σ	Mean	σ	Mean	σ
sp_20	147.2	7.79	140.2	7.98	138.3	8.08	136.2	7.99
sp_24	147.0	7.79	141.0	7.91	138.1	7.94	136.1	9.54
sp_28	146.7	7.97	139.8	7.84	137.7	7.54	134.8	8.67
pdsp_10	147.1	8.23	140.3	8.16	138.2	8.36	135.8	9.03
pdsp_12	147.1	8.48	141.2	8.23	137.9	8.44	135.5	8.55
povg_34	140.2	8.27	138.9	8.13	138.8	8.90	136.1	9.10
povg_40	140.5	9.58.	142.5	9.25	139.2	9.03	136.4	33.7
bentgate	146.9	8.03	139.1	7.60	135.7	7.41	132.6	7.97
bent_w14	147.0	7.80	139.3	7.36	135.4	7.13	132.9	7.08

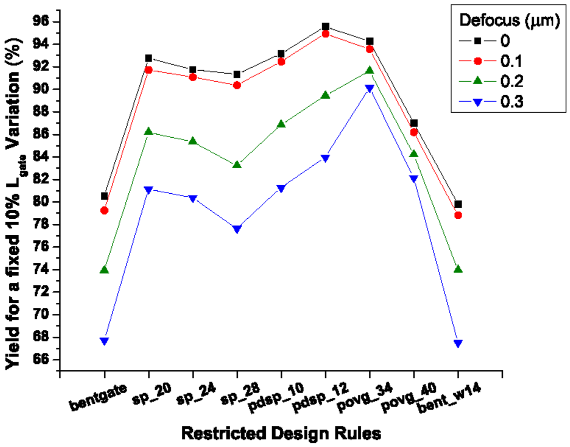


Figure 5.7: Functional yield for a fixed 10% L$_{gate}$ variation for c7552.

yield to defocus. This implies that design rule sets that include relaxed (larger), poly line-end extension rules may have larger process windows which reduce manufacturing overhead/cost. We observe from the figure that the use of bent gates with off-axis illumination (as we are using) produces a large number of gates with substantial (>5%) EPEs. Finally, we also see that the “pdsp_12” design rule set provides a very high percentage of gates within the stated ITRS specification indicating it has promise as an RDR.

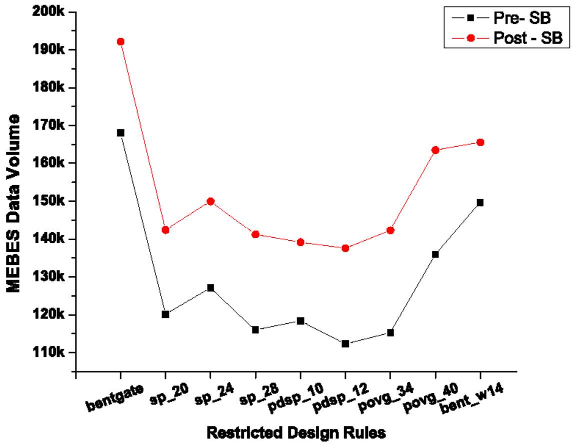


Figure 5.8: Impact of scattering bars on data volume for various RDRs for the c7552 circuit.

5.3.2 Scattering Bars

Isolated lines usually suffer more optical distortion effects than dense lines since lithography and RET recipes are not tuned or optimized for isolated lines. Although OPC corrects for the iso-dense bias at zero defocus, with non-zero defocus isolated lines tend to print narrower (or wider depending on the lithography system being used). Scattering bars (SBs), which are extremely narrow lines that do not actually print on the wafer, can modify the wavefront and reduce these distortions. However, liberal use of SBs adds considerable data volume in the MEBES format and places additional requirements on the resolution of the mask writing equipment. For the experiment of this section we modified our OPC recipe by adding scattering bars. These SBs are added whenever a poly line is fairly isolated; their impact is to make all poly lines in the design look similarly dense. Figure 5.8 shows the increase in data volume when SBs are inserted for our experimental setup. We observe a relatively consistent 15-20% increase in data volume when including SBs in the various RDR-based libraries. Insertion of scattering bars depends on the desired tradeoff between DOF margin and RET cost.

Table 5.3: Comparison of the single pitch library (SP) and the reduced default library (RDL)

RDR	3σ CD Uncertainty			Normalized Performance			
	Defocus (μm)						
	0.1	0.2	0.3	Delay	Area	Power	MEBES
RDL	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SP	0.91	0.79	0.75	1.05	1.10	1.01	0.75

5.3.3 Approach of the Single Pitch RDR

Modern processes are usually tuned to favor one particular pitch (e.g., in off-axis illumination the angle of illumination to the mask is optimized so that one pitch can be printed perfectly due to the diffraction of light). Although within a limited range the illumination distortion caused by pitch differences may be compensated with other techniques such as SBs, designers still must keep the forbidden pitch range in mind for better yield. A "single pitch, single orientation" rule, where orientation implies horizontal or vertical gate routes, is a highly desirable solution from a lithography perspective but it requires significant constraints in library design and P&R. For simplicity, the AOI and OAI cell types are excluded in this section. A larger pitch number is expected than the default value so that a contact can be inserted between two poly lines. We obtain a pseudo single pitch library in which 97.6% of the gate pitches are fixed at a single value, while the remaining 2.4% are among three other other values. This is due to limitations in the cell layout synthesis tools. We compare the results with the reduced "sp_20" library, where AOI and OAI cell types are excluded, and all RDRs are set at default except that bentgate is "off". With a scattering-bar OPC recipe only tuned at defocus $0.1\mu m$ for the single pitch library, this RDR shows good potential to reduce the 3σ L_{gate} uncertainty (may reach 24.60% as shown in Table 5.3). Moreover, the MEBES data volume can be 25% less with

Table 5.4: Summary of normalized performance and manufacturability results

Testcase	RDR	Delay	Area	Power	MEBES	Yield
c7552	bentgate	1	1	1	1	1
	sp_20	1.09	0.96	0.88	0.72	1.15
	sp_24	1.00	1.01	0.92	0.76	1.14
	sp_28	1.02	0.99	0.92	0.69	1.13
	pdsp_10	1.02	1.00	0.91	0.70	1.16
	pdsp_12	1.04	1.00	0.88	0.67	1.19
	povg_34	1.02	0.98	0.88	0.69	1.17
	povg_40	0.98	1.06	0.91	0.81	1.08
	bent_w14	1.05	0.95	0.94	0.89	0.99
c6288	bentgate	1	1	1	1	1
	sp_20	0.99	1.12	1.02	0.87	1.13
	sp_24	1.02	1.07	0.96	0.84	1.11
	sp_28	1.01	1.10	0.99	0.85	1.11
	pdsp_10	0.97	1.10	0.98	0.85	1.12
	pdsp_12	0.97	1.13	1.00	0.81	1.15
	povg_34	1.03	1.06	0.98	0.80	1.14
	povg_40	0.99	1.10	0.94	0.87	1.08
	bent_w14	0.96	1.05	1.03	0.99	1.00
c5315	bentgate	1	1	1	1	1
	sp_20	0.99	1.00	0.85	0.75	1.12
	sp_24	0.94	1.05	0.94	0.79	1.11
	sp_28	1.00	1.09	1.00	0.76	1.12
	pdsp_10	0.90	1.05	0.92	0.807	1.07
	pdsp_12	0.93	1.04	0.91	0.70	1.17
	povg_34	0.90	1.15	1.03	0.85	1.16
	povg_40	0.93	1.20	1.06	0.94	1.07
	bent_w14	0.94	1.04	1.04	1.00	1.00

some penalty on performance (less than 6% in delay and power and about 10% in area).

5.3.4 Experiment on Circuit Performance

While the above discussion has been targeted at the manufacturability improvements provided by various RDRs, we must simultaneously consider the performance penalties incurred. In this section we report on the timing, area, and power implications of the aforementioned RDRs for the three studied benchmarks. Table 5.4 summarizes the circuit performance, mask data volume, and parametric yield given a 10% CD variation tolerance budget for all RDRs considered in this work. Looking at all three benchmarks we first point out that the range of delay values is quite small over all RDRs (5-10% worst-case spread) while the area and power impact is somewhat larger (up to 20% spread in both). The minimum poly_diffusion spacing rule as 0.12 μm ("pdsp_12") appears to be the most favorable rule for low MEBES data volume and high yield with acceptable performance. In particular it is useful to compare the "sp_20" and "pdsp_12" design rules which differ only in the poly_diffusion spacing rule. The latter shows improvements in both data volume and yield with negligible performance penalties (including better delay in all three circuits). The two line end extension rules (shown as "povg_$*$") exhibit very similar characteristics and show excellent robustness to process defocus as mentioned earlier. The use of bent gates with minimum size may typically save area but at the expense of greatly increased data volume and substantial yield loss. As a result, it is now commonplace to see bent gates prohibited in modern design rule sets to improve manufacturability. All of the above indicates that there are good performance arguments to introduce RDRs in modern processes to reduce cost of ownership, without hurting yield and circuit performance.

CHAPTER VI

DRC Plus: Augmenting Standard DRC with Pattern Matching on 2D Geometries

6.1 Introduction

Design rule checks (DRC) are the industry workhorse for constraining design to ensure both physical and electrical manufacturability of VLSI circuits. For example, complex issues in resolution enhancement technology (RET) including k, NA, source frequency λ, source shape and off-axis illumination, are all summarized as DRC in the form of allowable geometric measurements such as minimum line width, minimum space, forbidden pitches, etc. Even complex interactions, such as that between aerial image, photoresist, and line-ends, are summarized as specifications on tip-to-tip spacing and tip-to-line spacing in DRC. The guarantee to circuit designers and layout engineers is well understood: follow these DRC, and subsequent processing including OPC, lithography, and process steps will manufacture the circuit to the drawn specification [58]. This abstraction is necessary to free designers to deal with other formidable challenges in circuit design, such as area, capacitance, robustness, and delay. However, as devices continue to shrink from generation to generation and manufacturing challenges become more formidable, the clean abstraction provided by traditional DRC begins to crumble in two ways.

First, while aggressive RET and OPC models are carefully tuned for simple 1D geometries of parallel lines and spaces, complex interactions of 2D geometries with various RET choices are difficult to measure, difficult to model accurately, and difficult to analyze. This is the motive behind many suggestions to impose rigid design rules which use highly regular structures for layout, and avoid these 2D geometries altogether [20; 59]. The drawback is the potential of restricting design flexibility so much, that designers cannot effectively optimize for other aforementioned challenges of circuit design. While this latter point can be debated, the concern is clear: there are problematic 2D geometries which are difficult for DRC to capture.

Second, for DRC to work well, the abstraction should be clear and concise, so that the distinction between what is DRC clean and DRC dirty is clear in the minds of layout engineers. Moving from generation to generation, however, the design rule manual itself has evolved into a complex tome containing some cryptic rules put in to handle exceptional cases as they occur [60]. This exception handling procedure has been used with mixed success. Although it typically resolves the issue at hand, quite often, it is the enforcement of some DRC rule that causes more complex 2D geometries to be generated; in effect, designers meet the letter of the law, as defined by the DRC implementation code, without understanding the *spirit of the rule.* This leads to even more exceptional cases being added to the DRC manual, further increasing its complexity. There needs to be a way to specify a 2D geometry to be avoided, which can be easily understood, and specifically targeted so as to minimize unwanted side effects.

The goal of DRC Plus is to resolve both of these issues, that is, to provide a way of marking specific 2D geometries as undesirable, which can be clearly documented, and easily implemented, without unwanted side effects. It does so by using fast, image-based pattern matching as a method for identifying specific undesirable 2D geometries [61]. The image pattern itself, becomes a clear entry in the design manual which says, "Avoid the 2D geometry pictured here." However, this is insufficient for the goal of minimizing unwanted side effects. As explained later, when comparing DRC Plus to pattern matching, a single image comparison turns out to be a poor metric for determining *good* layout from *bad* layout. Consequently, we apply a second, simple preferred DRC rule which is only applied to the matched 2D pattern at hand. In general then, a single DRC Plus rule consists of a two things: (1) a 2D pattern identifying a problematic 2D geometric configuration, and (2) a simple DRC rule identifying the desired fix only where the problematic configuration exists.

The remaining sections of this chapter describe the details of DRC Plus. Section 6.2 explains the structure of the DRC Plus implementation. Section 6.3 compares DRC Plus against other DFM techniques, including DRC, yield design rules (YRC), restrictive design rules (RDR), simulation based layout printability verification, and pattern matching. Section 6.4 explains the creation process for DRC Plus rules, and explains how to identify a 2D configuration as problematic. Section 6.5 presents run-time performance results of DRC Plus when applied to real design data. Conclusions and future work follow in Section 6.6.

6.2 DRC Plus

In Figure 6.1, various layouts with line-end spaces of -20nm (sub-nominal), +0nm (min-space), and +20 nm (relaxed) are shown in Columns 2, 3 and 4, respectively. On Row 2, a standard DRC rule has been defined with a minimum space requirement, and the hashed error marker in Column 2 correctly identifies the -20nm sub-nominal line-end space as a DRC violation. On Row 3, the same global DRC rule must be

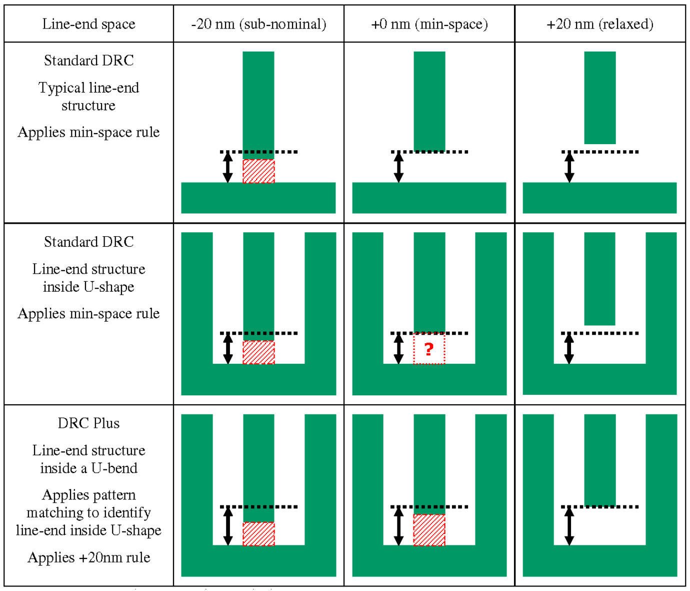


Figure 6.1: DRC Plus compared to standard DRC.

applied to a specific 2D situation where the line-end is surrounded on all sides by a U-shape, which again flags the -20nm sub-nominal line-end space in Column 2 as an error. However, because of the U-shape surrounding the line-end, the +0nm min-space line-end in Column 3 also exhibits manufacturing issues, as indicated by the "?"-mark, even though it is DRC clean. On Row 4, DRC Plus first identifies this "line-end into U-shape" configuration using pattern matching, and then applies a +20nm relaxed line-end space rule, specifically in this situation. In this way, both the -20nm sub-nominal space in Column 2 and the +0 nm min-space in Column 3 are correctly flagged as DRC Plus errors, which must be fixed in design.

The combination of a pattern match, followed by a preferred DRC rule to apply on those patterns constitutes a DRC Plus rule. For the line-end into U-shape example in Figure 6.1, the corresponding DRC Plus rule, to which we have given a hypothetical id "demo100" is shown in Figure 6.2. The rule has a description very similar to a simple DRC rule, except each rule is annotated with a layout clip to describe the specific 2D situation where the preferred rule is applied, and a match tolerance, explained later in Section 6.3.1. A collection of individual DRC Plus rules can be gathered together as a DRC Plus technology rule deck and design manual in a fashion similar to DRC.

The DRC Plus software developed by us is built on top of existing software tools as shown in Figure 6.3. Blocks in blue, the 2D Pattern Match Engine and DRC Engine, are software provided by vendors. For each DRC Plus rule in the deck, the 2D pattern is matched against the target layout using the 2D Pattern Match Engine [61]. This produces a set of match locations, represented as polygon markers. These match locations, in conjunction with the preferred DRC rules and the target layout are passed to the DRC engine. The resulting check results produced by the DRC engine are also the DRC Plus check results, so there is no difference in the output file format between standard DRC and DRC Plus. Consequently, the overhead for integrating DRC Plus into the standard DRC flow is minimal.

6.3 DRC Plus vs. Other DFM methods

In this section, we compare DRC Plus to other DFM methods, including DRC, preferred or yield design rules (YRC), regular or restrictive design rules (RDR), simulation based layout printability and scoring methods, and simple 2D pattern matching. Through these comparisons, in particular with standard DRC, we illustrate in more detail, the strength and weaknesses of DRC Plus as a whole.

6.3.1 DRC Plus vs. DRC

As a DFM technique, DRC Plus is most similar to an advanced version of DRC, as its name implies. It operates directly on the geometry of drawn designs without any simulation models; it results in the same pass/no-pass check result with an error marker denoting its location, and a simple description of the fix; and with a high-performance 2D-pattern matching engine, its run-time is comparable with that of a standard DRC rule. In addition, DRC Plus provides new functionality to the design DRC flow, by directly capturing and reporting marginal 2D situations to design in a simple, concise manner, as shown in the previous section. In doing so, it improves on the accuracy of design rule checks as a whole in predicting manufacturability. The upshot of all this is that DRC Plus has the potential to reduce the cost of manufacturing by finding potential yield detractors in design before OPC and mask tape-out, and by allowing layout designers to optimize right to the boundaries. These points can be illustrated with a simple simulation experiment in which the placement of 2 contacts are systematically varied in a diagonal corner-to-corner configuration. Each 2-contact pattern is then passed through the standard mask generation flow, including target sizing, model-based OPC, and MRC. The resulting mask pattern is then simulated with a calibrated model to determine final edge placement in resist. The simulation results are shown in Figure 6.4 (a).

In Figure 6.4 (a), the axes represent the (x,y) placement of the second contact relative to the first, and each point represents a particular 2-contact pattern. The shaded color represents the histogram of simulated edge placement error (EPE) in resist, ranging from no error in white, to poor in grey, to worst error in black. Poor EPE in resist indicates that both contact holes, by symmetry, print smaller than they should in resist, and this in turn means that they are both at risk of not forming a good electrical connection. To avoid this potential yield detractor, it would be prudent to apply a minimum space DRC using a square metric, as represented by the thick dashed line labeled *Yield DRC* in Figure 6.4 (a). Everything to the top and right of that line meets the minimum space Yield DRC, and consequently, should have no problems printing based on simulation. However, this Yield DRC rule ignores substantial white areas below and to the left of the line, which indicates that it is safe to push the contacts closer together when the contact pair is nearly vertical or nearly horizontal. If, instead, the Euclidean space DRC labeled *aggressive DRC* in Figure 6.4 (a) were adopted, we could use this extra physical design space when the contacts are nearly horizontal or vertical, however we would risk manufacturability problems when the contacts are placed in a diagonal configuration. In an effort to remain conservative, the choice would seem to favor the use of the so-called Yield DRC, sacrificing a small amount of design space. However, the fact is that in a vast majority of instances, minimum space contacts are placed vertically or horizontally from each other, so that this small amount of design space could impact the total area of design by several percent.

With DRC Plus, however, this choice is obviated: we can now use the minimum space aggressive DRC, and then employ pattern matching to capture the undesirable corner to corner configuration and enforce a much larger minimum space specification in such cases. This is illustrated in Figure 6.4 (b). The plot in Figure 6.4 (b) is the same as that in Figure 6.4 (a), but the Yield DRC has been replaced with a DRC Plus rule. The DRC Plus rule is a combination of a match target pattern with 2 contacts placed just over minimum space, 45° angle with respect to each other, and a preferred DRC rule with a larger minimum space, to be enforced only on matched patterns. The set of patterns that *match* form a region in design space, which in this particular example, is the diamond-shaped solid line *Match Region* in Figure 6.4 (b). The size of the Match Region is determined by the match tolerance parameter of the DRC Plus rule, which is defined as the maximum allowable percentage area difference, between the target pattern and the pattern under consideration, for a match to occur. In general, as the match tolerance becomes larger, the Match Region becomes larger,

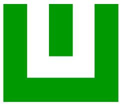
DRC Plus Rule ID	Pattern	Preferred Rule
demo100	 Match tolerance 2%	Line-end space ≥ min_space + 20 nm

Figure 6.2: An example of a DRC Plus rule.

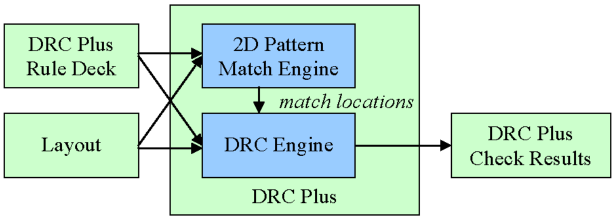
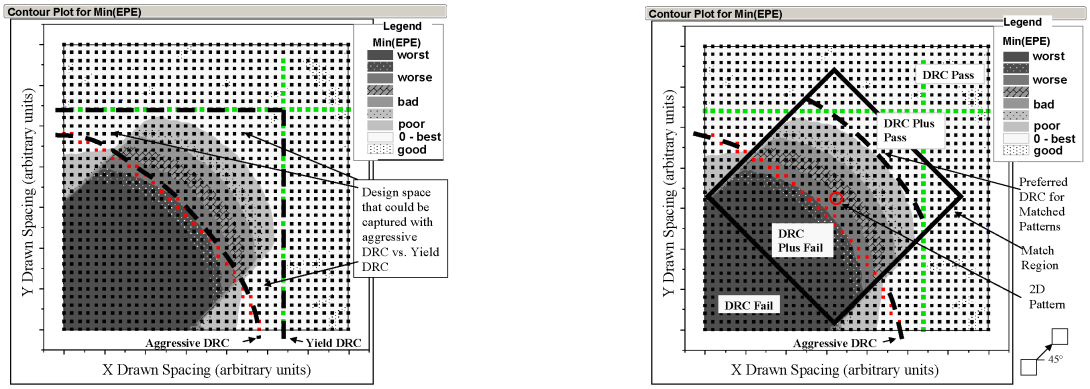


Figure 6.3: DRC Plus software block diagram.

the match becomes fuzzier, and more patterns are caught by the pattern matching engine.

In Figure 6.4 (b), the action of standard DRC interacts with DRC Plus as follows: for all designs that do not match the diagonal contact pattern, i.e., they lie outside the match region in Figure 6, the aggressive DRC rule is applied. For designs that match the diagonal contact pattern, i.e. lie inside the Match Region in Figure 6, the preferred DRC for that pattern is applied. As clearly shown in Figure 6.4 (b), the combination of DRC and DRC Plus allows us to apply a simple aggressive DRC under most conditions, while capturing the manufacturability issues of a particular 2D configuration. It is certainly possible for an ingenious DRC rule coder to duplicate the behavior of DRC Plus by implementing a 2D pattern match with DRC code. However, we assert that using DRCs in this way is difficult and error-prone, analogous to using garden shears to cut a piece of paper. DRCs are incredibly powerful for broadly restricting designs to generally manufacturable regions. The pattern matching used in DRC Plus is useful to enforce specific instances where a more yield-friendly set of DRC rules are needed. Finally as verification that simula-



(a) **(b)**

Figure 6.4: Simulated edge placement error (EPE) in resist of various contact placements: (a) standard DRC options; (b) aggressive standard DRC and a DRC Plus rule.

tions presented in Figure 6.4 (a) and 6.4 (b) are accurate, we include in Figure 6.5 an SEM image of an instance of the diagonal contact pattern at minimum space aggressive DRC with design geometry overlay. This image is made using design-based metrology automation tools [62], and clearly shows the small contact hole in resist, in relation to nearby contacts, at a corner of the process window. Ironically, these are redundant contacts, which probably would have been better left as a single contact. In summary, DRC Plus can be thought of as a refinement tool in the DRC toolbox to identify problematic layouts directly in design. DRCs are incredibly powerful for broadly restricting designs to generally acceptable regions as well as defining a clear, sharp pass/no-pass criterion. In contrast, DRC Plus applies pattern matching techniques to fine-tune this cutoff in specific instances. Insofar as one specific pattern can capture a specific manufacturability issue, including mask constraints, lithography issues, and process issues, DRC Plus can place that knowledge directly in the hands of layout designers with a simple pattern image, and enforce a preferred DRC rule specifically in those circumstances.

6.3.2 DRC Plus vs. Preferred or Yield Design Rules

Quite often, a technology rule deck contains a DRC deck that is enforced, and a *preferred* rule deck with less aggressive design rules which are recommended [63]. Conceptually, preferred rules would be applied in non-critical design areas. As a matter of practice, however, preferred rules are not consistently applied across design due to a lack of enforcement. Through the use of pattern matching, DRC Plus provides a mechanism for specifying situations when preferred rules must be applied.

6.3.3 DRC Plus vs. Regular Design Grid or Restrictive Design Rules (RDR)

There are suggestions that as technology shrinks continue, design rules imposing highly regular design grids for layout are needed [20; 59]. In effect, the goal is to completely disallow problematic patterns altogether. The drawback is, of course, the potential of restricting design flexibility so much that designers cannot effectively optimize for other challenges of circuit design, such as area or speed. Also, there is evidence that even on a regular grid, difficult-to-manufacture 2D situations may still occur. A simple example of this is shown in Figure 6.6, where the optical correction to tips competes with correction to the poly corners due to limited space on mask, which leads to bad convergence that can cause greater (30% to 60% more) line-end pullback than a normal poly tip-to-tip or tip-to-line configuration in single pitch, single orientation poly layout. In contrast DRC Plus is much more flexible and specific, requiring preferred design rules to be applied only to known problematic patterns, and retaining design freedom in all other situations.

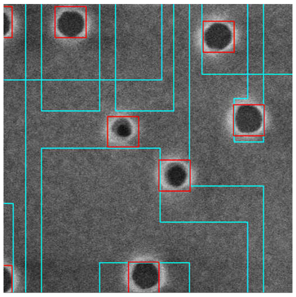


Figure 6.5: SEM image of redundant diagonal contact pair at minimum aggressive DRC rules.

6.3.4 DRC Plus vs. Simulation-based Layout Printability and Scoring Methods

A variety of simulation engines have been suggested to more accurately and directly predict manufacturability, rather than relying on design rules [64; 65]. Based on simulation of the manufacturing process technology, it is possible to predict with some certainty how a particular design will *print* on wafer. One drawback of this approach is the computational complexity of simulation, which often limits application of these tools to sub-blocks of the design. Another major drawback is the difficulty in obtaining an accurate simulation model early enough to significantly impact design, which often occurs in parallel with process development. Finally, the outputs of these simulation-based methods typically do not generate simple pass/no-pass criteria, nor do they offer simple suggestions on how to remedy the design. In contrast, DRC Plus may be as fast or faster than standard DRC, assuming an efficient pattern matching engine. The overhead of the pattern matching step is balanced against the complexity reduction in the DRC step, which is only applied to the matched patterns. Consequently, DRC Plus can easily be applied to the entire design. Since DRC Plus does not require complex simulation models it can be used early in design/process development phase. In addition, the output of DRC Plus is in the same format as DRC, providing a pass/no-pass result with a simple description of the problem and how to fix it. The additional knowledge provided by simulation-based tools may be captured through careful selection of the pattern and the preferred rules. In fact, we actively use such analysis tools in the creation of DRC Plus rules.

6.3.5 DRC Plus vs. Pattern Matching

A single pattern match cannot determine manufacturability or provide a straightforward pass/no-pass criterion. As an example, consider Row 3 of Figure 6.1 in Section 6.2. If the center image with 100nm line-end space is used as a pattern, pattern matching cannot distinguish between the 80nm line-end space, and the 120nm line-end space. From a match tolerance point of view, as described in Section 6.3, these two layouts have exactly the same area difference, but they behave very differently from a manufacturability point of view [61]. In DRC Plus, the additional application of a preferred DRC rule on pattern match regions is the means by which pattern match results may be tied directly to manufacturability and a pass/no-pass criterion.

6.4 Creating DRC Plus Rules

Like DRC, DRC Plus operates strictly on design geometries without any innate understanding of OPC, resolution enhancement techniques (RET), mask constraints, OPC, lithography, or process. Just as a minimum space rule is an integration of a multitude of technology-related issues, the components of a DRC Plus rule are a summary of these same manufacturability issues for the specified pattern. To this end, DRC Plus relies on analysis from other sources, including lithography simulators, layout scoring methods, OPC experience, RET experience and fab experience to identify undesirable 2D patterns and propose a preferred DRC for those situations. In particular, we rely heavily on an printability simulation engine for hotspot detection and simulation analysis [64; 65]. Once a pattern is identified, pattern matching is

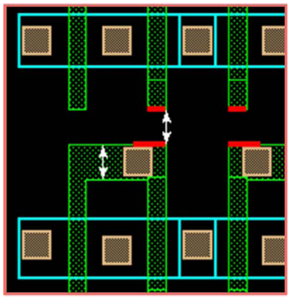
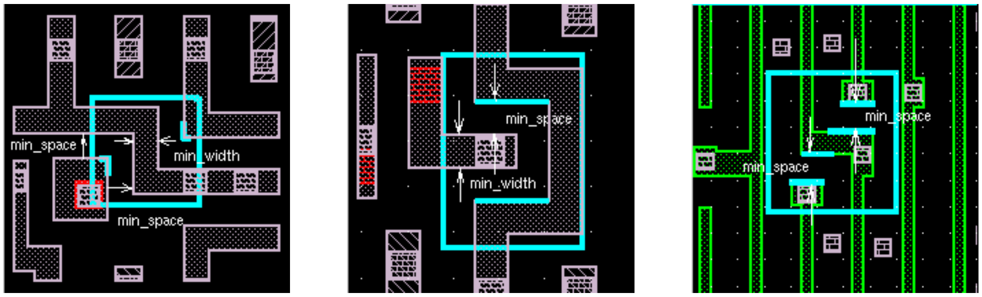


Figure 6.6: Problematic configuration of poly tip to poly corners in regular, single pitch, single orientation poly design.

applied to any design data available, full-chip if possible, to find similar patterns that may also be yield detractors. These are each passed through a lithography simulator to check for hotspots in order of increasing match tolerance. Through this process, a match tolerance threshold is identified beyond which either no hotspots are found or patterns become irrelevant.

At this point, based on matched patterns, causes of the hotspots are identified and preferred rules are checked to resolve the issue at hand, often drawing on existing DFM recommendations where such exists. A simulated design fix is then applied to the patterns, which passes the preferred rule, and it is verified through lithography simulation that the hotspot has been resolved. The pattern, match tolerance threshold, and preferred rule are then captured in a *DRC Plus rule*, and placed in the technology rule deck to be qualified in a process similar to DRC. The process described above is largely a manual process with some human judgment involved in each step, not unlike basic DRCs. Illustrations of some DRC Plus rules we have identified using this analysis procedure are shown in Figure 6.7. In each figure, a bright rectangle indicates the pattern, and the critical minimum space and minimum width are checked using the preferred DRC. Certainly automation of the rule creation procedure is desirable; how to achieve this goal is a topic for future research.



(a) (b) (c)

Figure 6.7: Illustrations of DRC Plus rules:(a) close convex corners causing middle line pinching; (b) small U-shape opening causing middle line pinching; (c) close landing pad causing bridging.

Table 6.1: DRC Plus runtime and memory usage

Layout	Total Runtime (one CPU)	Pattern Match	Memory Usage	Preferred DRC on Matched Areas
1 DRC Plus rule, M1 full chip	54 min	40 min	400 MB	15 min
15 Poly/M1 clips, full test chip, pattern match only	60 min	60 min	300 MB	N/A
3 M2 clips, full test chip, pattern match only	5 min	5 min	400 MB	N/A

6.5 Runtime Performance of DRC Plus

As DRC Plus adds an extra pattern match step to every preferred DRC rule, it is reasonable to investigate its impact on the physical verification runtime of design. In our experience, we find little impact for the most part due to the highly efficient 2D pattern matching engine used [61]. Table 6.1 reports typical runtimes and memory usage observed on a single CPU on one of our standard Linux computing clusters. Clearly the pattern matching engine is highly efficient in both performance and memory usage, and its resource usage is well within what is expected for full chip DRC runs. In addition, Row 2 shows that a pattern match operating over the entire chip compares favorably with the DRC engine, which only measures distances inside pattern match regions.

6.6 Summary and Future Work

In summary, we have shown that DRC Plus is a powerful tool for capturing manufacturability issues in specific 2D situations that are DRC clean. It does so in an elegant, easy-to-understand fashion, using a combination of fast pattern matching and enforcement of preferred or yield DRCs. Performance-wise, DRC Plus runtimes and memory usage is comparable to that of standard DRC. In many ways DRC Plus is simply an advanced extension of DRC, which is an advantage. For example, the output of DRC Plus is in the same format as DRC, making it is easy to integrate into the existing DRC review infrastructure. DRC Plus reduces the huge coding effort required for checking complex 2D situations and with fuzzy matching it allows further investigation into close but non-exact matching cases where printability may also be marginal. The main challenge to the use of DRC Plus is the creation of good high quality DRC Plus rules and future research into the automation of this process would be helpful. On the other hand, because the pattern itself limits the applicability of the DRC Plus rule, there is far less fear of causing unintended consequences than standard exhaustive DRC rules.

CHAPTER VII

Conclusions and Directions for Future Work

7.1 Conclusions

The aim of this work is to provide solutions to optimize tradeoffs among design, manufacturability, and cost of ownership posed by technology scaling and submicron lithography. These solutions may take the form of robust circuit designs, cost-effective resolution technologies, accurate modeling in design flow considering process variations, and design rules assessment. Though each of these approaches may be taken separately, the aim will be to make sure that any trade-offs with either other metrics or steps adopted in the current design flow are effective and justified.

With the framework established for assessing the impact of process variation on circuit performance, product value, and return on investment for alternative process improvements, we present results in terms of new metrics such as guardbanding, parametric yield at selling point, and inferred variation tolerance. This framework includes a comprehensive taxonomy of variations, and can handle variation differently with respect to its sources. We use accurate models of correlations and Monte Carlo techniques based on circuit simulation. Our main conclusions are the following.

1. With technology scaling and ITRS mandated fixed levels of process variability, delay variation decreases, whether measured as the amount of guardbanding

required to circumvent it, parametric yield loss, or the inferred variation tolerance for a preset design guardbanding.

2. For chips containing one dominant critical path, systematic WID variation does not affect yield, and design guardbanding is most sensitive to random D2D variation control.

3. Performance is very sensitive to L$_{eff}$ variation but the sensitivity reduces with technology scaling due to enhanced velocity saturation and a growing number of critical paths.

4. Under the same level of process variation, a larger NCP results in a smaller delay variation but larger delay mean. Because of the shift in delay mean value, a larger selling point delay is expected.

5. For the same NCP, looser control of CD variability leads to a larger required design guardbanding accompanied by a larger delay mean value, both of which show more sensitivity to relaxed process specifications than to tightened specs.

6. The delay distribution shifts to higher means but tighter overall distributions as the number of critical paths increases but this effect saturates beyond approximately 10 critical paths.

7. For ASIC designs, reducing NCP is the most effective way to achieve a smaller average delay.

8. Variability impact can be restricted by innovative design and this may be preferable due to the very costly nature of process improvement techniques.

To reduce the impact of gate CD 3σ variation on chip performance while also limiting masks costs and the computational complexity of OPC insertion, we have implemented a practical flow. In particular we focus on the use of edge placement errors (EPEs) to drive OPC insertion tools and leverage EPEs as the mechanism to direct these tools to correct only to the levels required to meet timing specifications. An iterative linear programming based approach is used to perform slack budgeting in an efficient manner. This formulation results in a specific slack budget for each gate that is then mapped to allowable critical dimensions in the standard cell. Finally EPEs are generated from the CD budget and tags are placed on gates to indicate to the OPC insertion tool the appropriate level of correction. Results on several benchmarks ranging from 300 to 34000 cells show up to 20% reduction in MEBES data volume which is frequently used as a metric for RET complexity. Furthermore, the runtime of the OPC insertion tool is reduced by up to 39% - this is critical since running OPC tools at the full-chip level is an extremely time-consuming step during the physical verification stage of IC design.

There has been unavoidable systematic gate CD variations induced from pitch size related OPC correction residues, which may cause chip performance degradation by alternating the top critical paths timing and ordering. We established an automated flow for post-OPC performance verification. Experimental results on a 90nm microprocessor show significant changes in post-OPC timing analysis compared to the traditional methodology of performing final timing analysis before OPC application. The number of critical paths increased by 170% while the worst-case slack violation increased by 36.4%. Among all critical paths found in the post-OPC flow, 21.8% were not reported in the traditional timing analysis. These changes demonstrate that traditional performance analyses are no longer valid in nanometer-scale designs that rely on complex resolution enhancement technologies. The methodology enables tagging critical gates such as those on critical paths or matching gates so that specific corrections can be applied to these gates to achieve better CD control rather than attempting to reduce gate CD variation in all scenarios. Integrating the OPC step into the design flow effectively will allow design-time optimizations to be aware of the manufacturing process and achieve improved performance and yield in the final as-fabricated design.

Restricted design rules (RDRs) have been suggested as a fundamental prevention mechanisms to greatly reduce lithography induced variation, at the expense of design margin. We created an automatic flow to systematically analyze the impact of several sets of typical RDRs on circuit performance, mask data volume, and parametric yield, given a 10% CD variation tolerance budget. The experimental results show that the range of delay values is quite small over all RDRs (5-10% worst-case spread) while the area and power impact is somewhat larger (up to 20% spread in both). The minimum poly_diffusion spacing rule of 0.12 μm ("pdsp_12") in a 130 nm technology appears to be the most favorable rule to achieve low MEBES data volume and high yield with acceptable performance. In particular it is useful to compare the "sp_20" and "pdsp_12" design rules which differ only in the poly_diffusion spacing rule. The latter shows improvements in both data volume and yield with negligible performance penalties (including better delay in all three studied circuits). Two rules that increase the minimum line end extension exhibit very similar characteristics and show excellent robustness to process defocus. The use of bent gates with minimum size often saves area at the expense of greatly increased data volume and substantial yield loss. As a result, it is now commonplace to see bent gates prohibited in modern design rule sets to improve manufacturability. All of the above indicates that there are good performance arguments to introduce small sets of RDRs in modern processes to reduce cost of ownership, with limited impact on yield and circuit performance.

Furthermore, for layers such as upper level metals that only have a second order impact on performance variation, it is not cost effective to deploy RDRs globally. To prevent yield detractors due to complex 2D pattern situations, a hybrid flow for augmenting the standard DRC procedure with pattern matching together with local RDR enforcement is established. This way, pattern dependent corner cases leading to yield problems particularly printing difficulties and significant effort in standard design rules enhancement, can be captured and avoided at an early stage (DRC Plus adopted design flow).

7.2 Suggestions for Future Work

The EPEMinCorr methodology shows potential for reducing cost of other layers besides gate poly. Also, in future technologies the EPEMinCorr methodology may be modified to be driven by bounds on acceptable leakage power rather than by traditional delay uncertainty constraints. Moreover, this methodology can be extended for field poly features, which impact performance due to their overlap with the contact layer. Expensive masking layers such as diffusion, contact, metal1 and metal2 may also be aided by this methodology.

Post-OPC process CD annotation can be extended to the metal and contact layers in order to enable RC extraction from Si-based post-OPC dimensions. Post-OPC extraction and timing analysis flow will enable design-time optimizations with more close-to-Si process feedback therefore may achieve improved performance and yields in the final as-fabricated design.

To reduce variability and detect yield detractor patterns, global RDRs on critical layers together with locally deployed hybrid RDRs for non-critical layers will greatly reduce design complexity for 45nm technology and beyond. “DRC Plus” may also be used to set allowable design topologies. The challenge would be to develop high quality DRC Plus rules.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] F. M. Schellenberg, “Sub-Wavelength Lithography Using OPC”, *Semiconductor Fabtech*, 9th Edition, 1999, pp. 205-209.

[2] International Technology Roadmap for Semiconductors, 2003 http://public.itrs.net/

[3] S. R. Nassif, “Delay Variability: Sources, Impacts and Trends”, *ISSCC*, 2000, pp. 368-369 .

[4] *Optical Lithography Cost of Ownership - Final Report* , http://www.sematech.org/docubse/document/4014atr.pdf

[5] W. Grobman, *Personal communication.*

[6] M.L. Rieger, J.P. Mayhew and S. Panchapakesan, “Layout design methodologies for sub-wavelength manufacturing”, *Proc. Design Automation Conference*, 2001, pp. 85-92.

[7] L. W. Liebmann, “Resolution Enhancement Techniques in Optical Lithography: It’s not just a Mask Problem”, *Proc. SPIE Vol. 4409*, 2001, pp. 23-32.

[8] M. D. Levenson, “Wavefront Engineering from 500 nm to 100 nm CD”, *Optical Microlithography X, Proc. SPIE Vol. 3051*, 1997, pp. 2-13.

[9] B. Lin, “Phase Shifting Masks Gain an Edge”, *IEEE Circuits and Devices*, 1993, pp. 28-35.

[10] S. R. Nassif, “Modeling and Forecasting of Manufacturing Variations”, *Proc. Fifth International Workshop on Statistical Metrology*, 2000, pp. 3-10.

[11] S. T. Ma, A. Keshavarzi, V. De, and J. R. Brews, “A Statistical Model for Extracting Geometric Sources of Transistor Performance Variation”, *IEEE Transactions on Electron Devices*, 51(1), 2004, pp. 36-41.

[12] M. Orshansky, L. Milor, and C. Hu, “Characterization of Spatial Intrafield Gate CD Variability, Its Impact on Circuit Performance, and Spatial Mask-Level Correction”, *IEEE Transactions on Semiconductor Manufacturing*, 17(1), 2004, pp. 2-11.

[13] C. Visweswariah, “Death, Taxes and Failing Chips”, *Proc. Design Automation Conference*, 2003, pp. 343-347.

[14] A. B. Agrawal, D. Blaauw, V. Zolotov, and S. Vrudhula, “Statistical timing analysis using bounds and selective enumeration”, *Proc. Design Automation Conference*, 2003, pp. 348-353.

[15] M. Orshansky and K. Keutzer, “A General Probabilistic Framework for Worst Case Timing Analysis”, *Proc. Design Automation Conference*, 2002, pp. 556-561.

[16] S. Postnikove and S. Hector, “ITRS CD Error Budgets: Proposed Simulation Study Methodology”, May 2003.

[17] L. Chen, L. Milor, C. Ouyang, W. Maly, and Y. Peng, “Analysis of the Impact of Proximity Correction Algorithms on Circuit Performance”, *IEEE Transactions on Semiconductor Manufacturing*, 12(3), 1999, pp. 313-322.

[18] B. Stine, D. Boning, J. Chung, D. Ciplickas, and J. Kibarian, “ Simulating the Impact of Poly-CD Wafer-Level and Die-Level Variation On Circuit Performance”, *Proc. Second International Workshop on Statistical Metrology*, 1997, pp. 24-27.

[19] P. Gupta and F.L. Heng, “Toward a Systematic-Variation Aware Timing Methodology”, *Proc. Design Automation Conference*, 2004, pp. 321-326.

[20] L.W. Liebmann, “Layout Impact of Resolution Enhancement Techniques: Impediment or Opportunity”, *Proc. ACM/IEEE Intl. Symp. on Physical Design*, 2003, pp. 110-117.

[21] L. Stok and J. Cohn, “There is Life in ASICs”, *Proc. ACM/IEEE Intl. Symp. on Physical Design*, 2003, pp. 48-50.

[22] A. B. Kahng, “The Road Ahead: Shared Red Bricks”, *IEEE Design and Test*, March 2002.

[23] D. Sylvester and K. Keutzer, “System-level modeling using BACPAC”, *International Workshop on System-level Interconnect Prediction*, 1999.

[24] “Berkeley Predictive Technology Model”, *http://www-device.eecs.berkeley.edu/ ptm*

[25] K. A. Bowman and J. D. Meindl, “Impact of Within-Die Parameter Fluctuations on Future Maximum Clock Frequency Distributions”, *CICC*, 2001, pp. 229-232.

[26] W. Zhang and Z. Yang, “A New Threshold Voltage Model for Deep-Submicron MOSFET’s with Nonuniform Substrate Dopings”, *Proc. Electron Devices Meeting*, 1997, pp. 39-41.

[27] D. Boning and S. Nassif, *Design of High-Performance μP Circuits, Chapter 6: Models of Process Variation in Device and Interconnect*, 1998, pp. 98-116.

[28] T. Park, T. Tugbawa, J. Yoon, D. Boning, J. Chung, R. Muralidhar, S. Hymes, Y. Gotkis, S. Alamgir, R. Walesa, L. Shumway, G. Wu, F. Zhang, R. Kistler and J. Hawkins, "Pattern and Process Dependencies in Copper Damascene Chemical Mechanical Polishing Processes", *VLSI Multilevel Interconnect Conference*, Santa Clara, CA, June 1998.

[29] S. Hymes, K. Smekalin, T. Brown, H. Yeung, M. Joffe, M. Banet, T. Park, T. Tugbawa, D. Boning, J. Nguyen, T. West and W. Sands, "Determination of the Planarization Distance for Copper CMP Process", *Materials Research Society 1999 Spring Meeting*, San Francisco, CA, April 1999.

[30] K. A. Bowman, Intel Corp., *Personal communication.*

[31] K. A. Bowman, S. G. Duvall and J. D. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution", *IEEE Int. Solid-State Circuits Conf.*, 2001, pp. 278-279.

[32] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan and V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage", *IEEE Int. Solid-State Circuits Conf.*, 2002, pp. 422-478.

[33] M. Orshansky, L. Milor, P. Chen, K. Keutzer and C. Hu, "Impact of Systematic Spatial Intra-Chip Gate Length Variability on Performance of High-Speed Digital Circuits", *IEEE Int. Conf. Computer-Aided Design*, 2000, pp. 62-67.

[34] A. Srivastava and D. Sylvester, "Minimizing Total Power by Simultaneous V_{dd}/V_{th} Assignment", *Proc. Asia-South Pacific Design Automation Conference*, 2003, pp. 400-403.

[35] S.-C. Wong, G.-Y. Lee and D.-J. Ma, "Modeling of Interconnect Capacitance, Delay and Crosstalk in VLSI", *IEEE Trans. on Semiconductor Manufacturing*, 40(1), 2000, pp. 108-111.

[36] K. Chen, C.Hu, P. Fangm M.R. Lin abd D.L. Wollesen, "Predicting CMOS Speed with Gate Oxide and Voltage Scaling and Interconnect Loading Effects", *IEEE Transactions on Electron Devices*, 44(11), 1997, pp.1951-1957.

[37] T. Sakurai and R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas", *IEEE J. of Solid State Circuits*, 25(2), 1990, pp. 584-594.

[38] International Technology Roadmap for Semiconductors, 2005. http://public.itrs.net/

[39] Chiang Yang, "Challenges of Mask Cost and CycleTime", *SEMATECH: Mask Supply Workshop*, Intel, 2001.

[40] P. Gupta, F.-L. Heng and M. Lavin, “Merits of Cellwise Model-Based OPC”, *Proc. SPIE International Symposium on Microlithography*, 2004.

[41] S. Murphy, Dupont Photomask, *SEMATECH: Mask Supply Workshop*, 2001.

[42] K. Wampler, ASML MaskTools, personal communication, March 2003.

[43] C. Spence, et al., “Mask Data Volume - Historical Perspective and Future Requirements”, *Proc. SPIE 22nd European Mask and Lithography Conference*, 2006, Volume 6281.

[44] R. Nair, C.L. Berman, P.S. Hauge and E.J. Yoffa, “Generation of Performance Constraints for Layout”, *IEEE Transactions on Computer Aided Design*, 8(8), 1989, pp. 860-874.

[45] C. Chen, E. Bozorgzadeh, A. Srivastava and M. Sarrafzadeh, “Budget Management with Applications”, *Algorithmica*, 2002, pp. 261-275.

[46] E. Bozorgzadeh, S. Ghiasi, A. Takahashi and M. Sarrafzadeh, “Optimal Integer Delay Budgeting on Directed Acyclic Graphs”, *DAC*, 2003.

[47] http://www.mentor.com

[48] http://www.opencores.org

[49] http://www.ilog.com

[50] Y. Zhang, R. Gray, O.S. Nakagawa, P. Gupta, H. Kamberian, G. Xiao, R. Cottle, and C. Progler, “Interaction and balance of mask write time and design RET strategies”, *Proc. SPIE Photomask*, Japan, 2005.

[51] M. Orshansky, L. Milor, P. Chen, K. Keutzer and C. Hu, “Impact of Spatial Intrachip Gate Length Variability on the Performance of high-Speed Digital Circuits”, *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, 2002, pp. 544-553.

[52] “ProGenesis User’s Manual”, *Prolific Inc.*, Newark, CA 94560.

[53] “Calibre xRC, RET, Mask Data Preparation User’s Manual”, *Mentor Graphics Corp.*, Wilsonviller, OR 97070.

[54] “HSPICE, Powerarc, Design Compiler, PrimeTime User’s Manual”, *Synopsys Inc.*, Mountain View, CA 94093.

[55] “Silicon Ensemble User’s Manual”, *Cadence*, San Jose, CA 95134.

[56] Y. Taur and T.H. Ning, “Fundamentals of Modern VLSI Devices”, *Cambridge University Press*, 1998, United Kingdom.

[57] S. Postnikov and S. Hector, "ITRS CD Error Budgets: Proposed Simulation Study Methodology", *International Technology Roadmap for Semiconductors*, May, 2003.

[58] L. Capodieci, "From Optical Proximity Correction to Lithography-Driven Physical Design (1996-2006): 10 years of Resolution Enhancement Technology and the roadmap enablers for the next decade", *Optical Microlithography XIX*, edited by Donis G. Flagello, Proc. of SPIE Vol. 6154, 615401, 2006.

[59] L. Pileggi, H. Schmit, A. J. Strojwas, et al., "Exploring Regular Fabrics To Optimize The Performance-Cost Trade-Off", *Proceedings of the ACM/IEEE DAC*, 2003.

[60] C. Webb, "Layout Rule Trends and Affect upon CPU Design", *Design and Process Integration for Microelectronic Manufacturing IV*, edited by Alfred K. K. Wong, Vivek K. Singh, Proc. of SPIE Vol. 6156, 615602, 2006.

[61] Eclair Pattern Matcher End User's Guide, Version 5.0, CommandCAD, Inc., March 2006.

[62] C. Tabery, L. Page, "Use of Design Pattern Layout for Automated Metrology Recipe Generation", *Proceedings of SPIE: Metrology, Inspection and Process Control for Microlithography XIX*, vol. 5752, pp. 1424-1434, 2005.

[63] J. Yang, E. Cohen, C. Tabery, N. Rodriguez, M. Craig, "An Up-stream Design Auto-fix Flow for Manufacturability Enhancement", *43rd Design Automation Conference*, 2006.

[64] J. A. Torres, N.C. Berglund, "Towards Manufacturability Closure: Process Variations and Layout Design", *IEEE EDPS Workshop*, 2005.

[65] J. A. Torres, "Litho-friendly design: a necessary complement to RET",*Microlithography World*, vol. 15, no. 2, pp. 10, 12-13A, May 2006.