UNIVERSITY OF CALIFORNIA, SAN DIEGO

Lithography-Driven Design for Manufacturing in Nanometer-Era VLSI

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical Engineering (Computer Engineering)

by

Chul-Hong Park

Committee in charge:

Professor Andrew B. Kahng, Chair
Professor Chung-Kuan Cheng
Professor Lawrence E. Larson
Professor Bill Lin
Professor Tajana Simunic Rosing

2008

The dissertation of Chul-Hong Park is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2008

To my wife, Jinyoung, without whose love, encouragement, sacrifices and belief, this thesis would not have been possible.

To my children, Sehyun and Junkyu, for their lovely smiles and sweet hugs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express sincere graditude to my advisor Prof. Andrew B. Kahng for his many valuable suggestions, persistent encouragement, support and guidance during the course of my doctoral studies. Many ideas in this dissertation have emerged and got refined from discussions with him. His enthusiasm, passion and integral view on research has made a deep impression on me, and continue to lead me toward new challenges. I am also grateful to Son Young Kahng for caring about my family and proofreading this thesis.

I certainly feel privileged and grateful to work with friends and collaborators in Professor Kahng's VLSICAD Lab. I would like to thank Kwangok Jeong, Prof. Puneet Gupta, Dr. Swamy Muddu, Kambiz Samadi, Dr. Puneet Sharma, Rasit Topaloglu, Dr. Qinke Wang, Dr. Xu Xu and Dr. Hailong Yao for their invaluable comments and excellent research ideas. I would like to thank our lab's administrator Virginia McIlwain, for all her help, care and encouragement in the past five years.

I would also like to thank my thesis committee members, Prof. Chung-Kuan Cheng, Prof. Lawrence E. Larson, Prof. Bill Lin and Prof. Tajana Simunic Rosing for taking their time to review and evaluate my research work. I am very grateful for their valuable feedback.

I wish to express my appreciation to Dr. Chang-Gyu Hwang, president of Samsung Electronics, and Dr. Jeong-Taek Kong, former vice president of CAE in Samsung Semiconductor, for granting the Samsung Scholorship. I would also like to thank two vice presidents of CAE, Mr. Moon-Hyun Yoo and Dr. Kyu-Myung Choi, for their care, advice and encouragement, which made this thesis all worthwhile.

Most of all, I wish to thank all my achievements to God, without His graceful guidance and countless blessing, I would not have been able to reach this point in my life. I am grateful to my parents, Kwanmo Park and Sunja Cho, parents-in-law, Sang Hur and Namhee Jeong, and the rest of my family for their endless love and encouragement. Without them this work would never have come into existence. I am indebted to my wife, Jinyoung for her sacrifices, belief and love. My two children, Sehyun and Junkyu, have been a great source of joy to me during my studies. I will never forget their lovely smiles and sweet hugs. This thesis is dedicated to them.

The material in this thesis is based on the following publications. Authors' names are listed in alphabetical order in all publications.

- Chapter II is based on the following publications:
  - P. Gupta, A. B. Kahng and C.-H. Park, "Detailed Placement for Enhanced Control of Resist and Etch CDs", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(12), 2007, pp. 2144 - 2157.
  - P. Gupta, A. B. Kahng and C.-H. Park, "Detailed Placement for Improved Depth of Focus and CD Control", *Proc. Asia and South Pacific Design Automation*, 2005, pp. 343 - 348.
  - P. Gupta, A. B. Kahng and C.-H. Park, "Enhanced Resist and Etch CD Control by Design Perturbation", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2005, pp. 59923P-1 – 59923P-11.
- Chapter III is based on the following publications:
  - A. B. Kahng, S. Muddu and C-.H. Park, "Auxiliary Pattern-Based OPC for Better Printability, Timing and Leakage Control", *Journal of Microlithography, Microfabrication and Microsystems*, 7(1), 2008, pp. 013002-1 – 013002-13.
  - A. B. Kahng and C.-H. Park, "Auxiliary Pattern for Cell-Based OPC", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63494S-1 – 63494S-10.
- Chapter IV is based on the following publications:

  - A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, "Lens Aberration Aware Placement for Timing Yield", submitted to *ACM Transactions on Design Automation of Electronic Systems*, 2008.
  - A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, "Lens Aberration-Aware Timing-Driven Placement", *Proc. Design Automation and Testing in Europe*, 2006, pp. 890 - 895.

- Chapter V is based on the following publications:
  - K. Jeong, A. B. Kahng, C.-H. Park and H. Yao, "Dose Map and Placement Co-Optimization for Timing Yield Enhancement and Leakage Power Reduction", *Proc. ACM/IEEE Design Automation Conference*, 2008, to appear.

- Chapter VI is based on the following publications:
  - A. B. Kahng, C.-H. Park and X. Xu, "Fast Dual Graph-Based Hotspot Detection", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2008, to appear.
  - A. B. Kahng, C.-H. Park and X. Xu, "Fast Dual Graph-Based Hotspot Detection", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63490H-1 – 63490H-8.

My coauthors (Prof. Andrew B. Kahng, Prof. Puneet Gupta, Dr. Swamy Muddu, Dr. Puneet Sharma, Dr. Qinke Wang, Dr. Xu Xu, Dr. Hailong Yao and Mr. Kwangok Jeong) have all kindly approved the inclusion of the aforementioned publications in my thesis.

VITA

1968	Born, Busan, South Korea
1992	B.S., Mathematics, Kyung Hee University, Seoul, South Korea
1994	M.S., Mathematics, Kyung Hee University, Seoul, South Korea
1994 – 2003	Senior Engineer, Samsung Semiconductor, R&D Center, Hwasung, South Korea
2007	C.Phil., Electrical Engineering (Computer Engineering), University of California, San Diego
2008	Ph.D., Electrical Engineering (Computer Engineering), University of California, San Diego

PUBLICATIONS

All papers coauthored with my advisor Prof. Andrew B. Kahng have authors listed in alphabetical order.

- A. B. Kahng, C.-H. Park and X. Xu, "Fast Dual Graph-Based Hotspot Detection", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2008, to appear.

- A. B. Kahng, S. Muddu and C-.H. Park, "Auxiliary Pattern-Based OPC for Better Printability, Timing and Leakage Control", *Journal of Microlithography, Microfabrication and Microsystems*, 7(1), 2008, pp. 013002-1 – 013002-13.

- K. Jeong, A. B. Kahng, C.-H. Park and H. Yao, "Dose Map and Placement Co-Optimization for Timing Yield Enhancement and Leakage Power Reduction", *Proc. ACM/IEEE Design Automation Conference*, 2008, to appear.

- P. Gupta, K. Jeong, A. B. Kahng and C.-H. Park, “Electrical Metrics for Lithographic Line-End Tapering”, *Proc. Photomask and Next-Generation Lithography Mask Technology*, 2008, pp. 70238A-1 – 70238A-12.

- A. B. Kahng, C.-H. Park, X. Xu and H. Yao, “Double Patterning Lithography Aware Intelligent Layout Decomposition”, *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2008, to appear.

- R. J. Greenway, K. Jeong, A. B. Kahng, R. S. Mackay, C.-H. Park and J. S. Petersen, “32nm 1-D Regular Pitch SRAM Bitcell Design for Interference-Assisted Lithography”, *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2008, to appear.

- P. Gupta, A. B. Kahng and C.-H. Park, “Detailed Placement for Enhanced Control of Resist and Etch CDs”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(12) 2007, pp. 2144 - 2157.

- P. Gupta, A. B. Kahng, C.-H. Park, K. Samadi and X. Xu, “Wafer Topography-Aware Optical Proximity Correction”, *IEEE Transactions on Computer-Aided Design*, 25(12), 2006, pp. 2747 - 2756.

- A. B. Kahng, C.-H. Park and X. Xu, “Fast Dual Graph-Based Hotspot Detection”, *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63490H-1 – 63490H-8.

- A. B. Kahng and C.-H. Park, “Auxiliary Pattern for Cell-Based OPC”, *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63494S-1 – 63494S-10.

- A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, “Lens Aberration-Aware Timing-Driven Placement”, *Proc. Design Automation and Testing in Europe*, 2006, pp. 890 - 895.

- P. Gupta, A. B. Kahng, S. Muddu, S. Nakagawa and C.-H. Park, “Modeling OPC Complexity for Design for Manufacturability”, *Proc. SPIE BACUS*

Symposium on Photomask Technology and Management, 2005, pp. 59921W-1 – 59921W-11.

- P. Gupta, A. B. Kahng and C.-H. Park, "Detailed Placement for Improved Depth of Focus and CD Control", *Proc. Asia and South Pacific Design Automation*, 2005, pp. 343 - 348.

- P. Gupta, A. B. Kahng and C.-H. Park, "Manufacturing-Aware Design Methodology for Assist Feature Correctness", *Proc. SPIE Conf. on Design and Process Integration for Microelectronic Manufacturing*, 2005, pp. 131 - 140.

- P. Gupta, A. B. Kahng and C.-H. Park, "Improving OPC Quality Via Interactions Within the Design-to-Manufacturing Flow", *Proc. Photomask and Next-Generation Lithography Mask Technology*, 2005, pp. 131 - 140.

- P. Gupta, A. B. Kahng, C.-H. Park, K. Samadi and X. Xu, "Wafer Topography-Aware Optical Proximity Correction for Better DOF Margin and CD Control", *Proc. Photomask and Next-Generation Lithography Mask Technology*, 2005, pp. 844 - 854.

- P. Gupta, A. B. Kahng and C.-H. Park, "Enhanced Resist and Etch CD Control by Design Perturbation", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2005, pp. 59923P-1 – 59923P-11.

- P. Gupta, A. B. Kahng, C.-H. Park, P. Sharma, D. Sylvester and J. Yang, "Joining the Design and Mask Flows for Better and Cheaper Masks", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2004, pp. 318 - 329.

ABSTRACT OF THE DISSERTATION

Lithography-Driven Design for Manufacturing in Nanometer-Era VLSI

by

Chul-Hong Park
Doctor of Philosophy in Electrical Engineering (Computer Engineering)
University of California, San Diego, 2008
Professor Andrew B. Kahng, Chair

Photolithography has been a key enabler of the aggressive IC technology scaling implicit in Moore's Law. As minimum feature sizes approach the physical limits of lithography and the manufacturing process, resolution enhancement techniques (RETs) dictate certain tradeoffs with various aspects of process and performance. This in turn has led to unpredictable design, unpredictable manufacturing, and low yield. As a result, close communication between designer and manufacturer has become essential to overcome the uncertainties of design and manufacturing.

The design for manufacturability (DFM) paradigm has emerged recently to improve communications at the design-manufacturing interface and to reduce manufacturing variability. DFM is a set of technologies and methodologies that both help the designer extract maximum value from silicon process technology and solve "unsolvable" manufacturing challenges. Traditional DFM techniques, which include design rule check (DRC) and optical proximity correction (OPC), have been successfully used until now. However, as the extent and complexity of lithography variations increase, traditional techniques are no longer adequate to accommodate the various lithography demands. This thesis focuses on ways to mitigate the impact of lithography variations on design by establishing new interfaces between design and manufacturing. The motivations for doing so are improved printability, timing and leakage as well as reduced design cost.

To improve printability, we propose a detailed placement perturbation technique for improved depth of focus and process window. Using a dynamic

programming (DP)-based method for the perturbation, the technique facilitates insertion of scattering bars and etch dummy features, reducing inter-cell forbidden pitches almost completely. We also propose a novel *auxiliary pattern-enabled* cell-based OPC which can improve the edge placement error over cell-based OPC. The technique improves runtime which has grown unacceptably in model-based OPC, while retaining its runtime advantage as well as timing and leakage optimization. The detailed placement framework is also available to allow opportunistic insertion of auxiliary pattern around cell instances in the design layout.

Aberration leads to linewidth variation which is fundamental to achieve timing performance and manufacturing yield. We describe an aberration-aware timing analysis flow that accounts for aberration-induced cell delay variations. We then propose an aberration-aware timing-driven global placement technique which utilizes the predictable slow and fast regions created on the chip due to aberration to improve cycle time. The use of the technique along with field blading achieves significant cycle time improvement.

DoseMapper technique adopted in advanced lithography equipments has been used to reduce the across-chip linewidth variation. We propose a novel method to enhance timing yield as well as reduce leakage power by combined dose map and placement optimizations. The new dose map is not determined to have the same critical dimension (CD) in all transistor gates, but optimized to have different linewidths. That is, for devices on setup timing-critical paths, a smaller than nominal CD will be desirable, since this creates a faster-switching transistor. On the other hand, for devices on hold timing-critical paths, a larger than nominal gate CD will be desirable, since this creates a less leaky transistor.

Last, the golden verification signoff tool using simulation-based approach represents a runtime-quality tradeoff that is high in quality, but also high in runtime. We are motivated to develop a low-runtime *pre-filter* that reduces the amount of layout area to be analyzed by the golden tool, without compromising the overall quality finding hotspots. We demonstrate a dual graph-based hotspot filtering technique that enables fast and accurate estimation.

I

Introduction

As optical lithography advances into the 45nm technology node and beyond, minimum feature size outpaces the introduction of advanced lithography hardware solutions. In particular, the linewidth tolerance required for manufacturability of poly and metal layers is extremely difficult to achieve due to various sources. The prominent sources of linewidth variation are defocus, exposure dose, lens aberration, pattern-dependent proximity, etching effects, etc. Resolution enhancement techniques (RETs) such as optical proximity correction (OPC) [21, 25, 95, 116], phase shift mask (PSM) [77, 81] and off-axis illumination (OAI) [69, 91] have been aimed at the major optical wave components, namely, direction, amplitude and phase. Combinations of these techniques can provide advantages of enhanced linewidth control and depth of focus (DOF) margin at minimum pitch. However, the adoption of RETs dictates certain side-effects, i.e., increasing mask writing time and difficulty of mask inspection. Strong OAI also causes a lower process margin at pitches beyond the optimum light angle. Avoiding these side-effects presents a host of new challenges for physical design automation [110].

Design for Manufacturing (DFM) techniques address the questions related to the exchange of information between design and manufacturing, and the use of this information for better printability and enhanced yield [42]. Design rules have been the usual medium of communication for the manufacturer to convey manufacturing limitation to design. However, as the complexity and extent of process variations have increased, rule checking is no longer sufficient. Perpetuating the design rule framework in the presence of RET and other manufacturing constraints results in a huge number of rule checks [80]. On the other hand, a wide range of equipment improvements continually afford opportunities to leverage design information for cost and turnaround time improvement. However, what has been missing is any connection with "design awareness". That is, new equipment is used solely to reduce linewidth variation, but misses the goal of optimizing device performance or parametric yield.

The focus of this thesis is on lithography-aware physical design techniques. To motivate the problems addressed in this thesis, we next present a brief overview of the IC (Integrated Circuit) physical design and manufacturing flow, resolution enhancement techniques and lithography-aware design methodologies.

I.A The IC Physical Design and Manufacturing Flow

The physical design step in IC design has taken an important role in DFM since that phase is the middle step between front-end logic design and the manufacturing process. Physical design converts a register transfer level (RTL) description of the operation of a digital circuit, into layout which will be realized as a reticle (or mask) for wafer printing. A simple schematic of the physical design and manufacturing flow is shown in Figure I.1. Physical design includes partitioning, global and detailed placement, global and detailed routing, and engineering change order (ECO) for placement and routing. Since the placement *locks down* layout in given placement sites, placement retains its prime importance in the DFM framework. In this thesis, we focus primarily on discussions of placement, layout generation, photolithography and etching processes, and describe various improvements in these areas.

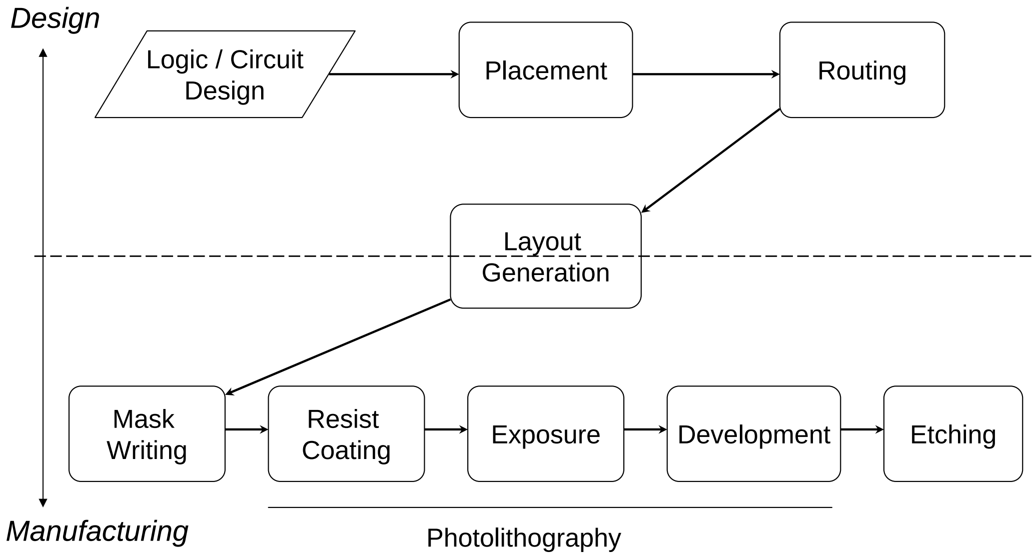


Figure I.1: Simple schematic of the physical design and manufacturing flow.

I.A.1 Placement

The placement takes a given synthesized circuit netlist from the logic and the circuit design step. Placement assigns exact locations for standard cells, macro blocks and I/O pads within the chip area. Standard cells are logic modules with a predetermined internal layout, which have generally the same height but various widths. The cells have fixed connections on the left and right side that abut with each other. A row consists of a number of placement sites, and hence the width of a cell is the sum of placement site widths (called *"sitewidth"* in this thesis). There are whitespaces in a row which are empty sites unoccupied by cells.

Placement assignment problems affect the performance and routing resource. The whitespace between cells is thus optimized with respect to performance demands (i.e., wirelength, timing) and congestion, with power and leakage as secondary objectives. Placement results are typically evaluated using the half-perimeter wirelength (HPWL) of the placed circuit hypergraph [71]. It is an exact estimator of the rectilinear Steiner minimum tree (RSMT) for 2 and 3-pin nets, and there is mounting evidence of its relationship to routing congestion [28].

Global placement determines the rough position of components and may produce a great deal of overlap of components. The circuit in the global placement phase can be clustered to reduce the size of complexity of the placement problem. The process of global placement and unclustering is iterated until the components close to well-spread. Then, legalization performs the removal of the overlapped components and assigns valid positions to all movable components. The detailed placement receives a set of cells which has undergone coarse placement and is performed by scanning through the rows of the substrate and selecting the left-most positioned vacant site of each row as a candidate site for placement. Detailed placement is comprised of three phases: (i) global moving, (ii) whitespace distribution and (iii) cell order polishing [63].

I.A.2 Layout Generation

Layout defines the physical structure of the integrated circuit in terms of planar geometric shapes which correspond to the patterns of metal, poly and other semiconductor layers. Since all design rules are implemented by the polygon shapes, the polygon represented by a GDSII format is the medium of communication between designer and manufacturer. Design rule checking (DRC) [10] involves sufficient margins to account for variability in semiconductor manufacturing processes [96]. A width rule specifies the minimum width that can be patterned at a given lithography machine. A spacing rule specifies the minimum distance between two adjacent objects, which has to address a pattern bridging margin in lithography. One other rule is a relationship between two layers, i.e., an enclosure rule of vias and contacts must be covered by a metal layer.

Primary RETs are also related to the polygon modification in the layout [107]. That is, OPC is the proactive distortion of polygon shapes to compensate for patterning inaccuracies. SRAF (Sub-Resolution Assist-Feature) inserts extremely narrow polygons into the layout to reduce the proximity effect between dense and isolated patterns. OAI enables a higher resolution and process margin depending on the pitch of the polygon.

The polygon data of the IC layout is processed into a form readable by mask writers. The IC design is written onto the masks during the making of masks. Mask writers can print features only with a specific set of sizes. Hence layout polygon data must be decomposed into smaller features that can be written individually by the mask writer. This transformation is generally referred to as *layout fracturing*. The MRC (Mask Rule Check) processes with the fractured polygon to ensure dimensions of the smallest features on the mask [70].

I.A.3 Photolithography Process

With the shrinking of VLSI feature size in the subwavelength regime, process variation has become a critical factor for performance, power and cost (i.e., yield). As a result, photolithography has been a key technology enabler of the aggressive IC technology scaling implicit in Moore's Law. Resultant light distortions create patterns on silicon that are substantially different from a GDSII layout. Although light distortions have traditionally not affected the design flow, the techniques used to control these distortions have a potential impact on the design flow that is as formidable as the recently addressed submicron transition.

Photolithography is the process of transferring circuit patterns on a layout to the surface of a silicon wafer. It uses light to transfer a geometric pattern from a photomask (or simply "mask") to a light-sensitive chemical (photoresist, or simply "resist") on the substrate. Lithography involves a complex series of resist coating, soft-bake, exposure and post-exposure bake (PEB) development. The complex series of chemical treatments engrave the exposure pattern into the material underneath the photoresist. In this thesis, we restrict this discussion to primary steps (i.e., resist coating, exposure and development) in the pattern transfer process.

Resist coating

Photoresist is a light-sensitive material which is classified into two groups, positive resists and negative resists. For positive resists, the resist is exposed with Ultraviolet (UV) light wherever the underlying material is to be removed.

Positive resist exposed to light becomes soluble to the photoresist developer and unexposed positive resists normally have very low solubility, i.e., the unexposed areas finally become IC circuit patterns. The mask thus contains an exact copy of the pattern which is to remain on the wafer. Negative resists behave in just the opposite manner. Exposure to the UV light causes the negative resist to become polymerized. The exposed negative resist thus have low solubility, while unexposed negative resist is soluble in development. Modern lithography process requires the short exposure times to correspond to the demand for high throughput in 300mm wafer production. Chemical amplified resist (CAR) increases the intrinsic sensitivity to UV light and enables the smaller k_1 for higher resolution patterning [55].

The wafer is covered with resist by spin coating. A viscous, liquid solution of photoresist is dispensed onto the center of the wafer through a resist nozzle. The resist is then subjected to centrifugal forces, and a thin and uniform resist layer adheres to the wafer. Most spin-coating processes are conducted at final spin speeds of 3000-7000 rpm (revolutions per minute) for a duration of 20-30 seconds. Resist coating is followed by a soft-bake, which improves the adhesion of the resist to the wafer and anneals the stresses introduced during the spin-coating. To improve the adhesion, antireflective coating (ARC) can be placed between the photoresist and substrate. The ARC is a polymer based liquid chemistry to suppress the reflection of the light on the surface of substrate, and hence improves CD control as well as adhesion [78].

Exposure and developement

After the soft-bake process, exposure and development steps are performed to create the circuit pattern. The modern lithography tool is a step-and-scan system which is a hybrid of scanner and stepper systems. Scanner projects a slit of light from the mask onto the wafer through the optical lens system. During the scanning operation, a small portion of wafer, called a field, is exposed under the mask. Multiple copies of the chip on the mask is printed onto the wafer. Then,

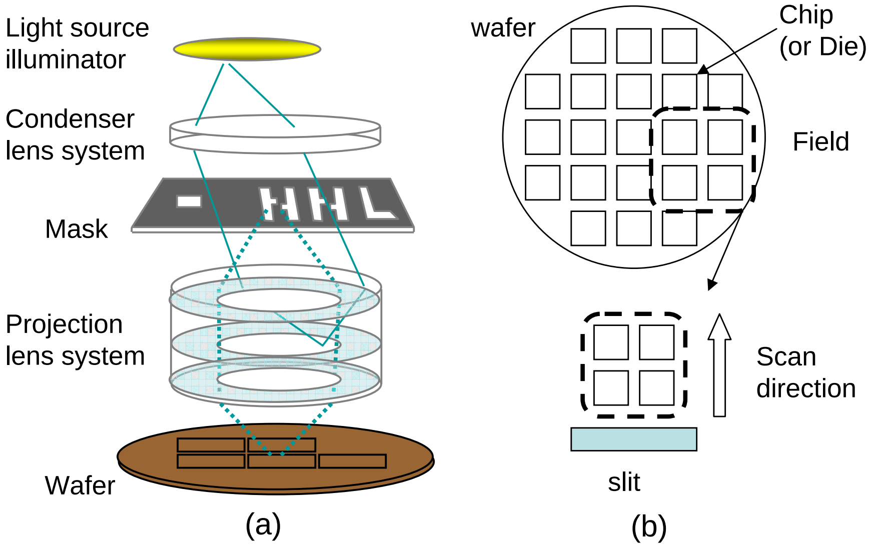


Figure I.2: Schematic of photolithography system: (a) a step-and-scan system and (b) exposure field scanned by slit.

the wafer is stepped to a new location and the scanning operation is repeated until all the chips in the wafer are exposed. A simple schematic of an optical lithography system with its main components – light source, a condenser lens, mask, projection (or objective), and the wafer – is shown in Figure I.2.

Optical lithography typically uses ultraviolet excimer lasers that have different wavelengths according to technology node, e.g., krypton fluoride laser (KrF: 248nm wavelength) has been used in 180nm - 130nm nodes and algon fluoride laser (ArF: 193nm wavelength) has been applied to 90nm - 32nm nodes. A condenser is a lens that serves to deliver uniform light with adequate intensity to the mask. The projection lens captures some portion of the diffraction order through the mask, and then delivers the image onto wafer. However, due to the finite size of lens and the higher diffraction orders of the light, pattern information generated by Fourier transform of the mask, are not captured. The loss of diffraction information leads to limited image quality and resolution. This resolution of optical projection lithography results in limited diffractions as described by the following Raleigh's equation:

$$R = k_1 \frac{\lambda}{NA} \tag{I.1}$$

- R is the minimum half pitch that can be resolved.

- λ is the exposure wavelength of the illumination source. ArF laser is used in modern lithography which is expected to extend to 32nm patterning.

- k_1 is a process dependent factor determined mainly by the resist capability, the tool control, the reticle pattern adjustments and the process control. Smaller k_1 values allow the printing of more dense patterns. It has a fundamental lower limit of 0.25. With the aim of double patterning lithography [14], the value tends to have the range of 0.18 - 0.28 for 32nm technology node.

- NA is the numerical aperture of the projection lens and equals n sinθ where n is the minimum index of refraction of the image medium (1.0 for air, 1.33 for pure water, and up to 1.56 for oils). The sine of the maximum half-angle of light can make it through a lens to the wafer [84].

NA of the projection lens is a measure of its ability to capture diffraction orders. Simple formulation of the diffraction can be expressed as the following equation.

$$sin(\theta) = \frac{o\lambda}{P} \tag{I.2}$$

where θ is the diffraction angle. o is the diffraction order of light and equals $0, \pm 1, \pm 2, \cdots$. P is the period (or pitch) between the patterns. Smaller P causes larger diffraction, as shown in F igure I.3. As a result, “sharp” transitions in the mask image, corresponding to higher diffraction orders, are not transferred to the wafer. To capture the higher diffraction orders, there are obviously two approaches: (1) high NA, and (2) high-refraction material. Research on lithographic optics design started in the late 1960s [113]. There are various configurations used

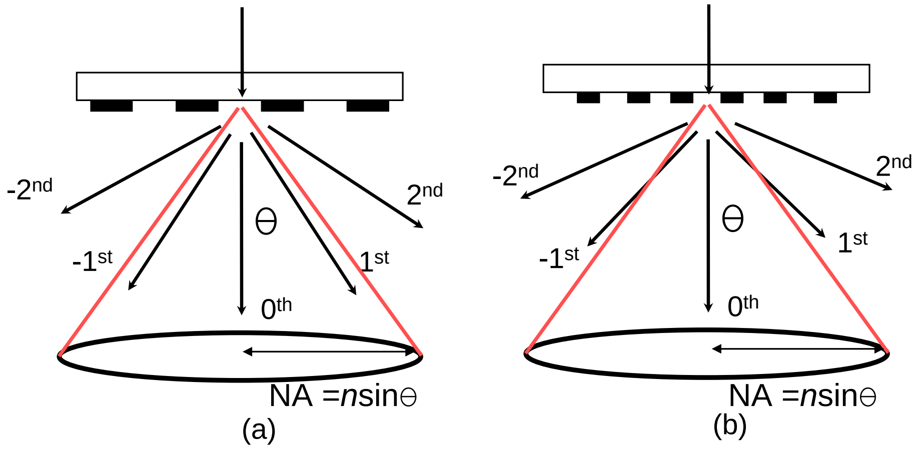


Figure I.3: Comparison of diffraction angle according to pattern pitches: (a) mask with large pitch and (b) mask with small pitch.

to explore the lens design space. The design possibilities are widened by a combination of reflective mirror surfaces and refractive lens elements. High NA system which has been improved by new lens elements approaches the fundamental limit (NA $\approx$ 0.93 in air). Research on high-refraction material started in the late 1990s. *Immersion lithography* enables numerical apertures greater than one, where the bottom of the lens is immersed in a high refractive index fluid such as water. The first system with 1.2 NA is now available, e.g., ASML TWINSCAN XT:1700i (193 nm immersion scanner) [1]. However, immersion lithography requires truly high-index fluids (NA $= 1.55 \sim 1.6$), with corresponding advances in high-index resists and optical materials, which will be a new challenge for the 32nm technology node.

The *depth of focus* (DOF) is one of the major measures to assess printing quality. The DOF is given by the expression

$$DOF = k_2 \tfrac{\lambda}{(NA)^2} \tag{I.3}$$

where k_2 is an empirically determined constant. Because of the inverse square dependency on NA, the depth of focus with high NA is extremely shallow. For this reason, planarization technique for wafer topography such as CMP (Chemical Mechanical Polishing) are required [118]. Ideally, the top of the wafer plane must coincide with the focal plane of the objective lens and this results in formation of the image at the best focus. If the wafer shifts vertically from the focal plane of the lens, then the aerial image is transferred out of focus. DOF is defined as the shift in the focus that results in tolerable deviation in the image from its intended dimension.

I.A.4 Etch Process

Etch process removes materials that are unprotected by a previous photolithography process step. These etch processes transform a single layer of semiconductor material into the poly, via, contact and interconnects that produce an integrated circuit. The etch process has gradually been issued due to intra-die CD variability of metal and poly, which results in leakage, timing and RC variability. In dry etch processes such as plasma, ion, and reactive ion etch (RIE), different consumptions of etchants with different pattern density lead to etch skew between dense and isolated patterns [76]. Etch proximity effect is determined by the complex physical, transport, and chemical interactions in an etch chamber. Moreover, etch proximity effect is heavily influenced by the actual layout of the integrated circuit. For example, all available etchants in areas with low density are consumed rapidly, and thus the etch rate then drops off significantly. In areas with high density of patterns, the etchants are not consumed as quickly. As a result, the proximity behavior of the photo process differs from that of the etch process [39]. For sub-90-nm processes, resist and etch effects can no longer be treated as a small perturbation on a purely optical OPC model. Hence, OPC models must account for such etch proximity-effects that occur due to the main-etch step and any additional etch steps.

I.B Resolution Enhancement Techniques

Optical Proximity Correction. OPC is the deliberate and proactive distortion of photomask shapes to compensate for systematic and stable patterning inaccuracies. OPC has proved to be a useful technique for matching photoresist edges to

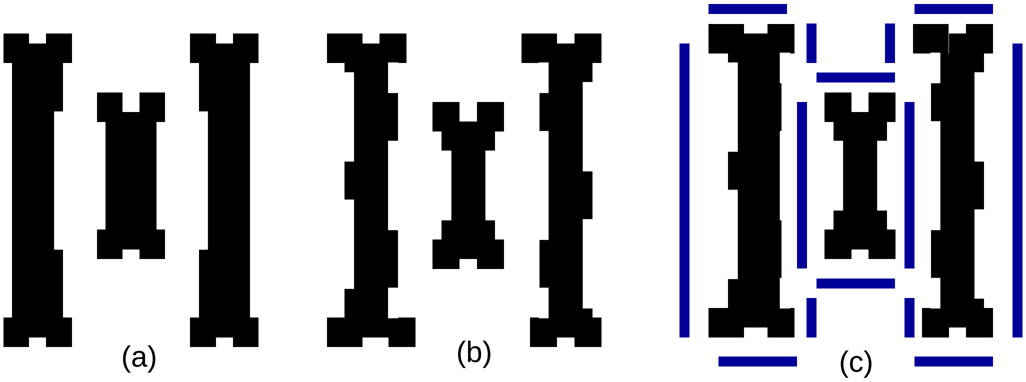


Figure I.4: OPC shapes with different OPC methods: (a) rule-based OPC, (b) model-based OPC and (c) SRAF OPC.

layout edges with essentially a layout sizing technique. This patterning distortion arises from (1) high-frequency light information that falls outside the lens and (2) coherent with light from one shape interacts with light in another shape on the electrical field. The three most common application for OPC are (1) linewidth, (2) pullback of line-ends, and (3) corner rounding corrections. There are various OPC methods used as follows.

- **Rule-based OPC:** Rule-based OPC modifies shapes by a set of "rules" [95, 96]. Typically, the rule includes pattern bias for linewidth control and hammerheads and serifs added to the ends and corners as shown in Figure I.4(a). The method is relatively straightforward and can be accomplished with a design rule checking (DRC) tool wich enables fast compensation for pattern distortion. However, as more complex and various proximity effects are involved, the method may not work correctly.

- **Model-based OPC:** Model-based OPC [115] which involves the fast aerial image simulation to compute the wafer results is done by fragmenting polygon edges into small segments, by adding small polygons or cutting the segments as shown in Figure I.4(b). The method enables highly accurate correction, but is more CPU intensive due to the requirement of scanning the entire layout [97].

- **SRAF OPC:** Rule- and model-based OPC have limitations in enhancing process margins with respect to depth of focus and exposure dose. The *SRAF* (Sub-Resolution Assist Feature) OPC technique combines pattern biasing with assist feature insertion to compensate for the deficiencies of bias OPC. SRAFs, which are extremely narrow lines that do not actually print on the wafer, modify the wavefront and allow the lens pupil to receive higher-order pattern information [21]. The SRAFs are placed adjacent to primary patterns, so that a relatively isolated primary line behaves more like a dense line as shown in Figure I.4(c). This works well for bringing the lithographic performance of isolated and dense lines into agreement. The SRAF OPC considerably improves a larger overlap of process window between dense and isolated patterns.

Phase Shift Mask. PSM utilizes the interference generated by phase differences. Photomask creates shifted and unshifted regions differentiated by etched quartz or light-blocking material, e.g., molybdenum/silicon (MoSi). Light passing through the two regions enhances interference effects in the image and increases contrast. The improved contrast allows increase of the resolution on the wafer. A brief discussion of the varieties of phase shift masks follows.

- **Alternating PSM (altPSM):** Each critical feature, which is a shape in the design, must be flanked by two phase shifters of opposing phases in order to create destructive interference between them [77]. Figure I.5(a) shows a diagram illustrating the distribution of an electric field and aerial image intensity in the altPSM. The reliance on destructive interference between two different phases (0^o and 180^o) creates the dark image having zero intensity and produces high-resolution patterns with large DOF margin.

- **Attenuated PSM (attPSM):** Due to the phase-assignment problem of altPSM, attPSM has been the most common application of PSM [81]. attPSM is composed of two layers making up the absorber, such as MoSi with 6% intensity transmittance, which is then processed in the same way as a standard chrome-on-glass mask. The interference of the light transmitted by the

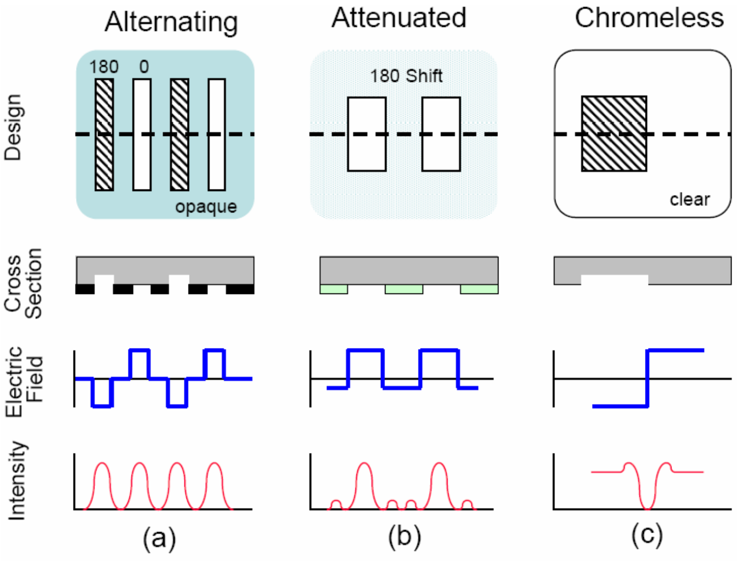


Figure I.5: Examples of phase shift mask: (a) alternating PSM, (b) attenuated PSM and (c) chromeless PSM.

attPSM material and that transmitted by the spaces (the quartz) produces a sharper transition from bright to dark at the edge in the resulting aerial image.

- **Chromeless PSM (cPSM):** cPSM can be described as a 100% transmission altPSM, i.e., there is no phase shifter material or chrome on the mask. cPSM uses the destructive interference between glass and etched quartz and generates extremely narrow lines [73]. However, mask manufacturing issues including etch uniformity and the loss of antireflective coating limits the application of cPSM to regular layers such as storage-poly and contact. Due to the technical hurdles, cPMS has not been adopted in industry.

Off-axis illumination. With a finite size of the lens, a large fraction of 0^{th} order diffraction reaches the objective lens, while a small fraction of 1^{st} order diffraction passes through the lens, which mean degrading of contact as shown in Figure I.6. In contrast, off-axis illumination (OAI) brings light to the mask at an oblique angle, and hence high-order diffraction enhances pattern quality [91, 114]. As the angle of diffraction through certain aperture shapes matches a given pitch, higher-

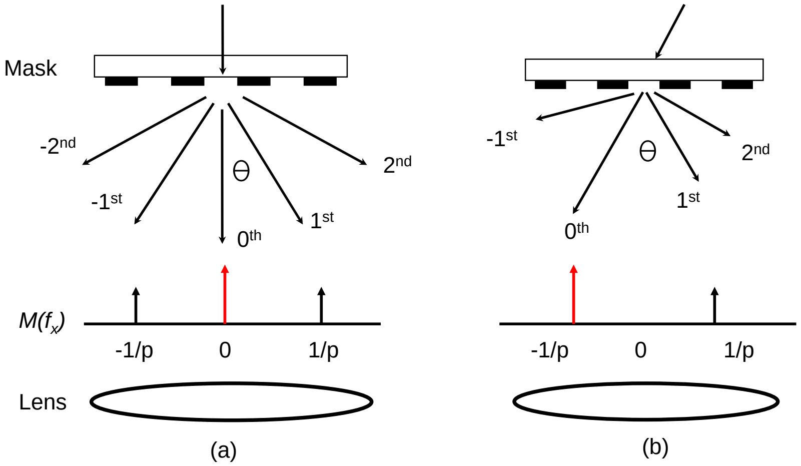


Figure I.6: Comparison of diffraction angles of (a) conventional illumination and (b) off-axis illumination.

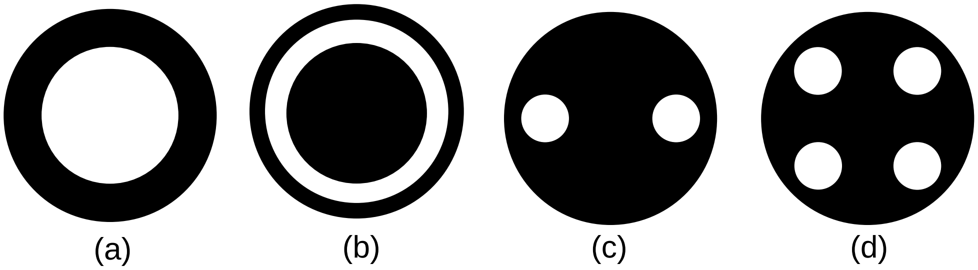


Figure I.7: Examples of aperture shapes for OAI: (a) circular, (b) annular, (c) dipole and (d) quadrupole apertures.

order pattern information can be projected on the pupil plane as determined by the numerical aperture (NA) of the illumination system. This technique enables certain pitches on the mask to obtain a higher resolution and extended focus margin. However, other pitches beyond the optimum angle will have a *lower* process margin compared to conventional illumination (i.e., with a circular aperture). Different types of apertures as shown in Figure I.7 must be optimized for the specific mask pattern being printed.

I.C Lithography-Driven Design Methodologies

To alleviate lithography-induced variability from the design side, techniques such as restricted design rules (RDRs) have also been proposed [79]. RDRs may set limitations on gate pitch and orientation, and hence reduce the need for aggressive OPC. However, layout restrictions lead to decreased designer freedom, which ultimately results in area increase. The increase in area resulting from layout restrictions limits the benefits of performance and cost gains obtained by process scaling. Lavin et al. [75] has explored how design restrictions could improve the manufacturability of design in a RDR, which called layout using gridded glyph geometry objects (*L3GO*). L3GO layouts have conventional overall structures (cells and layers). Shapes are defined by a coarse grid and typically a simple fraction of the minimum manufacturable feature pitch.

Gupta et al. [41] proposed an implementable flow that drives model-based OPC explicitly by timing constraints, with the objective of reducing mask data volume and OPC runtime. Mathematical programming based slack budgeting is used to determine edge placement error (EPE) tolerance budgets for all polysilicon gate geometries. These tolerances are then enforced by a commercial OPC tool to achieve MEBES data volume and OPC runtime reductions.

Kahng et al. [57] proposed the defocus-aware leakage estimation methodology which is comprised of two modules: (1) linewidth prediction, and (2) leakage calculation. The linewidth prediction module uses placement information of the design along with locations of devices within each cell in the cell library to compute pitches of all devices in the design. It then uses the Bossung table, which captures systematic variation of linewidth induced by defocus and pitch, to compute linewidths of all devices. The leakage calculation module computes leakage of all devices given their linewidths and line-ends.

Muddu [87] demonstrated the use of predictive linewidth models in fast and accurate leakage estimation and optimization. The through-focus systematic linewidth model is proposed to achieve accurate leakage estimation. A novel detailed placement perturbation approach that leverages systematic pitch and focus interactions is proposed to improve leakage in light of systematic linewidth variation. These two methods demonstrate the use of predictive models of variation in driving variation-aware design analysis and optimization.

Huang et al. [54] proposed OPC constrained maze routing (OPCCMR) to control the OPC complexity in metal and can be optimized by a multi-constrained shortest path (MCSP) problem. The Lagrangian relaxation method solves the MCSP by relaxing the space between patterns in higher density configuration. The Lagrangian method is again divided into two sub-problems such as Lagrangian Sub-problem (LSP) and Lagrangian Multiplier Problem (LMP). The LSP finds the shortest path for a set of Lagrangian multiplier and the objective of LMP maximizes the lower bound of MCSP. The netlist is rerouted to satisfy the constraints by using the OPC-friendly maze router.

To reduce lithography variation and optimized cost-aware design, new physical design and equipment-aware design methodologies must be adopted in future VLSI physical design. This thesis presents some techniques for constructing lithography-aware physical design techniques. We next review the specific topics that are addressed in this thesis.

I.D This Thesis

The focus of this thesis is on lithography-aware physical design techniques in nanometer-era VLSI. These techniques are categorized into three topics: (1) *placement-aware* DFM techniques are described in Chapter II and Chapter III, (2) *equipment-aware* DFM techniques are proposed in Chapter IV and Chapter V, and (3) *graph-based* lithography hotspot detection is discussed in Chapter VI. The key ideas and motivations in this thesis are summarized as follows.

- Detailed placement for enhanced control of resist and etch CDs. Sub-resolution assist feature (SRAF) and etch dummy insertion techniques have been absolutely essential for process window enhancement and CD control in photo and etch processes. However, as focus levels change during lithography manufacturing, CDs at a given “legal” pitch can fail to achieve manufacturing tolerances. Placed standard cell layouts may not have the ideal whitespace distribution to allow for optimal assist-feature insertion. At the same time, etch dummy features are used in the mask data preparation flow to reduce CD skew between resist and etch processes and improve the printability of layouts. However, etch dummy rules conflict with SRAF insertion because each of the two techniques require specific design rules. Chapter II of this thesis proposes a novel dynamic programming-based technique for *Assist-Feature Correctness* (AFCorr) and *Etch-dummy Correctness* (EtchCorr) in *detailed placement* of standard cell designs.

- Auxiliary pattern-based OPC for better printability, timing and leakage control. The most prominent OPC method, *model-based OPC*, alters the layout data of the photomask that enables drawn layout features to be accurately reproduced by lithography and etch processes onto the wafer. However, model-based OPC is computationally expensive and its runtime increases with technology scaling. The *cell-based OPC* approach improves runtime by performing OPC once per cell *definition* as opposed to once per cell *instantiation* in the layout. However, cell-based OPC does not comprehend inter-cell optical interactions that affect feature printability in a layout context. This leads to printability, and consequently, performance and leakage, degradation. We propose auxiliary patterns (AP) which are non-functional poly features that are added around a standard cell to “shield” it from optical proximity effects. Chapter III of this thesis proposes *auxiliary pattern-enabled* cell-based OPC to improve printability of cell-based OPC, while retaining its runtime advantage as well as timing and leakage optimization.

- Lens aberration aware placement for timing yield. Process variations due to lens aberrations are to a large extent systematic, and can be modeled for purposes of analyses and optimizations in the design phase. Traditionally, variations induced by lens aberrations have been considered random due to their small extent. However, as process margins reduce, and as improvements in reticle enhancement techniques control variations due to other sources with increased efficacy, lens aberration-induced variations gain importance. Chapter IV of this thesis proposes an aberration-aware timing analysis flow that accounts for aberration-induced cell delay variations. We then propose an aberration-aware timing-driven analytical placement approach that utilizes the predictable slow and fast regions created on the chip due to aberration to improve cycle time. We study the dependence of our improvement on chip size, as well as its use with field blading which allows partial reticle exposure.

- Dose map and placement co-optimization for timing yield enhancement and leakage power reduction. A wide range of equipment improvements continually afford opportunities to leverage design information for cost and turnaround time improvements. For example, ASML's DoseMapper technology [2] has been extensively used within the automatic process control context to improve global CD uniformity. The DoseMapper technique is used solely to reduce Across-Chip Linewidth Variation (ACLV) and Across-Wafer Linewidth Variation (AWLV) metrics for a given integrated circuit during the manufacturing process. However, to achieve optimum device performance (e.g., clock frequency) or parametric yield (e.g., total chip leakage power), not all transistor gate CD values should necessarily be the same. Chapter V of this thesis proposes to exploit the recent availability of fine-grain exposure dose control in the stepper to achieve both design-time (placement) and manufacturing-time (yield-aware dose mapping) optimizations of timing yield and leakage power. Our placement and dose map co-optimization can simultaneously improve both timing yield and leakage power of a given design. We formulate the placement-aware dose map optimization as a quadratic program, and solve it using an efficient quadratic programming solver.

- Fast dual graph based hotspot filtering. Lithography for mass production potentially suffers from decreased patterning fidelity. This results in the generation of many *hotspots*, which are actual device patterns with relatively large CD and image errors with respect to on-wafer targets. When these hotspots fall on locations that are critical to the electrical performance of a device, device performance and parametric yield can be significantly degraded. The golden verification signoff tool using simulation-based approach represents a runtime-quality tradeoff point that is high in quality, but also high in runtime. We are motivated to develop a low-runtime "*pre-filter*" that reduces the amount of layout area to be analyzed by the golden tool, without compromising the overall quality of hotspot-finding. Chapter VI of this thesis proposes a novel detection algorithm for hotspots induced by lithographic uncertainty. Our goal is to obtain a superset of actual hotspots, then our method can dramatically reduce the layout area processed by golden hotspot analysis.

II

Detailed Placement for Enhanced Control of Resist and Etch CDs

II.A Introduction

Across-chip linewidth variation (ACLV) induced by photolithography and etch processes has been a major barrier in ultra-deep submicron manufacturing. As a result, resolution enhancement techniques (RETs) such as optical proximity correction (OPC) [116], phase shift masks (PSM) [77], and off-axis illumination (OAI) are being pushed ever closer to fundamental resolution limits [42]. Combinations of these techniques can provide advantages for lithography manufacturing, e.g., OAI and OPC, together with sub-resolution assist feature (SRAF), achieve enhanced CD control and focus margin at minimum pitch.

However, when OAI is used, there will always be pitches for which the angle of illumination works with the angle of diffraction to produce a bad distribution of diffraction orders in the lens. These pitches are called *forbidden pitches* because of their lower printability, and designers should avoid such pitches in the layout. Forbidden pitches consist of Horizontal (H-) and Vertical (V-) forbidden pitches, depending on whether they are caused by interactions of poly geometries in the same cell row or in different cell rows, respectively. The resulting *forbidden pitch problem* for the manufacturing-critical poly layer must be solved before detailed routing. Since detailed routing works on fixed placement except some small placement ECOs as required, detailed routing “locks in” the poly layer layout. At the same time, we wish to address the forbidden pitch problem as late as possible, to avoid extra rework upon modification of the manufacturing recipe. We first describe a novel dynamic programming-based algorithm for assist-feature correctness (*AFCorr*), which uses flexibility in detailed placement to avoid all possible H- and V-forbidden pitches and the manufacturing uncertainty that they cause.

Etch-dummy features are introduced into the layout to reduce the CD distortion induced by etch proximity. The etch-dummies are placed outside the active layers so that leftmost and rightmost gates on active-layer regions are protected from ion scattering during the etch process. However, etch-dummy rules conflict with SRAF insertion because each of the two techniques requires specific spacings from poly. In such a regime, the assist-feature correct placement methodology must consider assist-feature and etch-dummy correction. In this chapter, we also present a novel SRAF-aware etch-dummy insertion method (*SAEDM*) which applies flexible etch-dummy rules according to the distance from active edge to leftmost (or rightmost) poly. As a result, the layout is more conducive to assist-feature insertion after etch-dummy features are inserted. Finally, we introduce a dynamic programming-based technique for etch-dummy correctness (*EtchCorr*) which can be combined with the SAEDM in detailed placement of standard cell designs [45, 46].

In this chapter, we present various analyses of lithographic printability within the context of the standard cell based design methodology. Our goal is to minimize CD variation and enhance feature printability and reliability. Our main contributions are as follows.

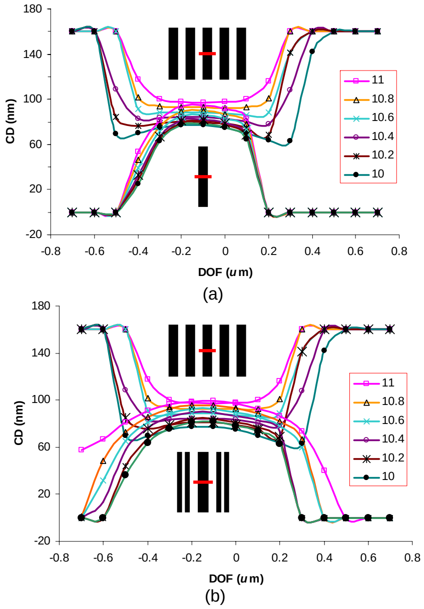
- We propose a novel post-detailed placement perturbation algorithm for assist-feature correctness (*AFCorr*). The dynamic programming based algorithm of AFCorr reduces the incidence of forbidden pitches by calculating H- and V-perturbation cost and finding an optimal perturbation of cell placements in a given row, subject to upper bounds on cell displacement. Particularly in conjunction with intelligent process-aware library layout, this technique achieves substantial improvements in depth of focus (DOF) margin and CD control.

- We present an SRAF-aware etch-dummy insertion method (*SAEDM*) which optimizes etch-dummy insertion to make the layout more conducive to assist-feature insertion.

- We describe the *Corr* post-detailed placement perturbation algorithm, which combines two techniques of etch-dummy correctness (*EtchCorr*) and AFCorr, removes forbidden pitches of resist CD and reduces the skew between resist and etch CDs, simultaneously. We test this method within a complete industrial flow and achieve up to 100% reduction in the number of cell border poly geometries having forbidden pitch violations.

- Various techniques that combine AFCorr, SAEDM and EtchCorr are validated with respect to wafer printability, database complexity, and device performance. The penalty in data size, OPC runtime and delay are within 3%, 4% and 6%, respectively, which is negligible compared to the large printability improvements and to the inherent “noise” in the relevant place-and-route tools.

The remainder of this chapter is organized as follows. In Section II.B, we review RET and its layout impact, focusing our discussion on strong OAI and OPC with SRAF. We then introduce a novel placement perturbation technique for assist-feature correctness. Evaluation flows to validate its impact on lithographic manufacturability and experimental results are described. In Section II.C, we describe etch-dummy insertion problems and the design perturbation algorithms of SAEDM and EtchCorr for better etch-dummy insertion. We evaluate various techniques that combine AFCorr, SAEDM and EtchCorr with respect to printability and design metrics. We conclude in Section II.D with directions for ongoing research.



(a)

(b)

Figure II.1: Comparison of Bossung plots between dense and isolated lines: (a) results of Bias OPC and (b) results of SRAF OPC.

II.B Assist-Feature Correctness

II.B.1 RET and Layout Impact

The extension of optical lithography beyond the quarter-micron regime has been enabled by a number of resolution enhancement techniques (RETs). These RETs address the available three degrees of freedom in lithography, namely, aperture, phase, and pattern uniformity [106]. However, the adoption of different RETs dictates certain tradeoffs with various aspects of process and performance [20].

Off-axis illumination (OAI) brings light to the mask at an oblique angle. As the angle of diffraction through certain aperture shapes matches a given pitch, higher-order pattern information can be projected on the pupil plane as determined by the numerical aperture (NA) of the illumination system. This technique enables certain pitches on the mask to obtain a higher resolution and extended focus margin. However, other pitches beyond the optimum angle will have a *lower* process margin compared to conventional illumination (i.e., with a circular aperture). Since strong OAI is an essential technique in current lithography, these other pitches should be forbidden, and their avoidance is a new challenge for physical design automation.

OPC is the deliberate and proactive distortion of photomask shapes to compensate for systematic and stable patterning inaccuracies. *Bias OPC*, the most common and straight-forward application of OPC, has proved to be a useful technique for matching photoresist edges to layout edges with essentially a layout sizing technique. However, bias OPC has limitations in enhancing process margins with respect to depth of focus and exposure dose. The Bossung plot[1] in Figure II.1 shows that bias OPC is not sufficient to reduce the CD difference between isolated and dense patterns with varying focus and exposure dose. The CD distortion in the isolated pattern is usually a problem since lithography and RET recipes are not tuned or optimized for isolated lines [98]. The *SRAF OPC* technique combines pattern biasing with assist-feature insertion to compensate for the deficiencies of bias OPC. SRAFs (or, scattering bars (SB)), which are extremely narrow lines that do not actually print on the wafer, modify the wavefront and allow the lens pupil to receive higher-order pattern information. The SRAFs are placed adjacent to primary patterns, such that a relatively isolated primary line behaves more like a dense line. This works well for bringing the lithographic performance of isolated and dense lines into agreement. The DOF margin of the isolated line as shown in Figure II.1(b) is considerably improved from that shown in Figure II.1(a), and a larger overlap of process window[2] between dense and isolated lines is achieved. The key observation is that the SRAF technique places more constraints on the spacing between patterns. SRAFs can be added whenever a poly line is sufficiently isolated, but certain minimum assist-to-poly and assist-to-assist spacings are required to prevent SRAFs from printing in the space [80].

[1] The Bossung plot shows multiple CD versus defocus curves at different exposure doses, and has been a useful tool to evaluate lithographic manufacturability. The common process window between dense and isolated patterns is an increasingly important requirement to maintain CD tolerances in the subwavelength lithography regime.

[2] Process window is defined as the range of exposure dose and defocus within which acceptable CD tolerance is maintained.

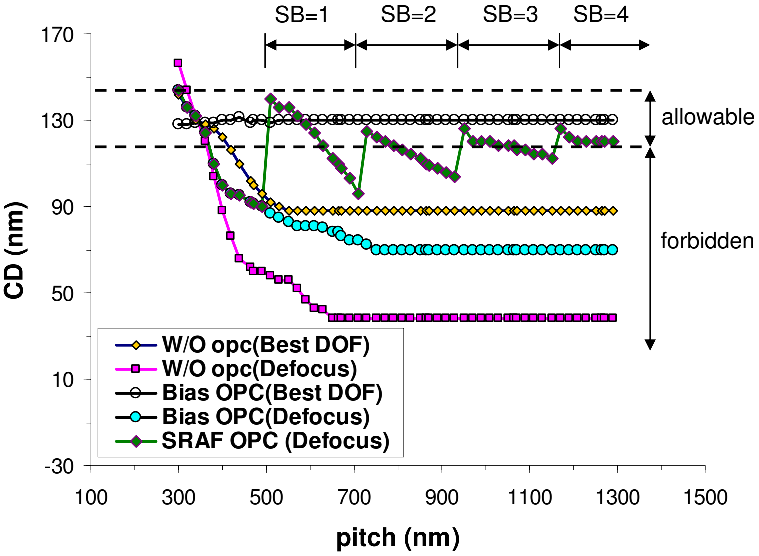


Figure II.2: Through-pitch proximity plots for 130nm technology for best focus without OPC, worst defocus without OPC, best focus with Bias OPC, worst defocus with Bias OPC, and worst defocus with SRAF OPC.

We now briefly review previous works related to forbidden pitches and their design implications. Socha et al. [114] observe that under more aggressive illumination schemes such as annular and quasar illumination, some optical phenomena become more prominent, most notably the forbidden pitch phenomenon. Shi et al. [111] give a theoretical analysis of pattern distortion in forbidden pitches, due to destructive light field interference. Although SRAFs are an effective method to collect high-order diffraction on the entrance pupil plane of a projection lens [99], Shi et al. report that incorrect SRAF placements around a given main feature can actually degrade the process latitude of that feature. A number of previous works have proposed techniques to control forbidden pitches using optimization of optical conditions such as NA and illuminator aperture shape of OAI [72,125,126]. All of these works using optimizations of NA and OAI have sought to enlarge the ranges of allowable pitches as shown in Figure II.2. However, approaches with process optimizations cannot completely remove the forbidden pitches, i.e., forbidden pitch always exists at any process condition.

Table II.1: SRAF rule table in 0.13μm and 0.09μm lithography.

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

II.B.2 SRAF Rule and Forbidden Pitch Extraction

Lack of space may prohibit insertion of a sufficient number of SRAFs, and as a result patterns may violate CD tolerance through defocus. *Forbidden pitches* are pitch values for which the tolerance of a given target CD is violated. *Allowable pitches* are all pitches other than forbidden pitches. In this section, we summarize the criteria for SRAF insertion and forbidden pitch extraction considering a worst-defocus model. Our SRAF insertion rule is initially generated based on the theoretical background given in [111]. Positioning of SRAFs is then adjusted based on OPC results. Large CD degradation through-pitch increases pattern bias as model-based OPC is applied, and this requires trimming of the SRAF rule to guarantee better process margin and prevent the SRAFs from printing.[3] After applying SRAF OPC to test patterns with the best-focus model, OPC'ed pitch patterns are simulated with the worst-defocus model which will be described in detail in Section II.B.4. This evaluation yields the forbidden pitches, considering maximum printability and manufacturability. The forbidden pitch rule is determined based on CD tolerance and worst defocus level, which are in turn dependent on requirements of device performance and yield. SRAF OPC restores printing when there is enough room for one scattering bar. But then larger pitches are forbidden

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

until there is enough room for two scattering bars. We thus can extract a set of forbidden pitches which will be demonstrated in Section II.B.4. In all of the work we report here, CD tolerance is assumed to be ±10% of minimum linewidth while the worst defocus level is assumed to be 0.5μm and 0.4μm for the 130nm and 90nm technology nodes, respectively. All of these results are summarized in Table II.1.

II.B.3 AFCorr Placement Algorithm

In this section, we describe the proposed AFCorr placement perturbation algorithm for assist-feature correction. Single orientation polysilicon geometries are becoming common for the current and future process generations. We consider the H-forbidden pitches within a cell row and the V-forbidden pitches between adjacent cell rows [43, 98]. In the present work, we treat the placement of a given cell row independently of all other rows, even though the cost function is calculated with respect to both H- and V-perturbations in order to avoid all forbidden pitches. Assuming that the spacings within the cell are assist-correct, then the only source of incorrect spacings between poly shapes for assist-feature insertion is cell placement. Adjacent cells within the same standard cell row as well as cells within adjacent cell rows which have shapes overlapping interact for this purpose. The vertical poly shapes (typically gates) at the left and right periphery of a cell which overlap with similar shapes in the neighboring cells in the row constitute the horizontal interaction. Similarly, horizontal poly shapes (typically field) at the top and bottom periphery of the cell which overlap with similar shapes in vertically adjacent cells (in adjacent rows) constitute the vertical interaction. In the following we describe the *single-row* AFCorr perturbation algorithm, by which we solve the 2D AFCorr problem one cell row at a time.

Let $C_{a,j}$ be a cell at the a^{th} position in the j^{th} row. To explain the interactions of border poly geometries, we adopt the following notations.

- **Horizontal polygon interaction:** Given a cell $C_{a,j}$, let $LP_{a,j}$ and $RP_{a,j}$ be the sets of valid poly geometries in the cell which are located closest to the left and right outlines of the cell, respectively. Only geometries with length

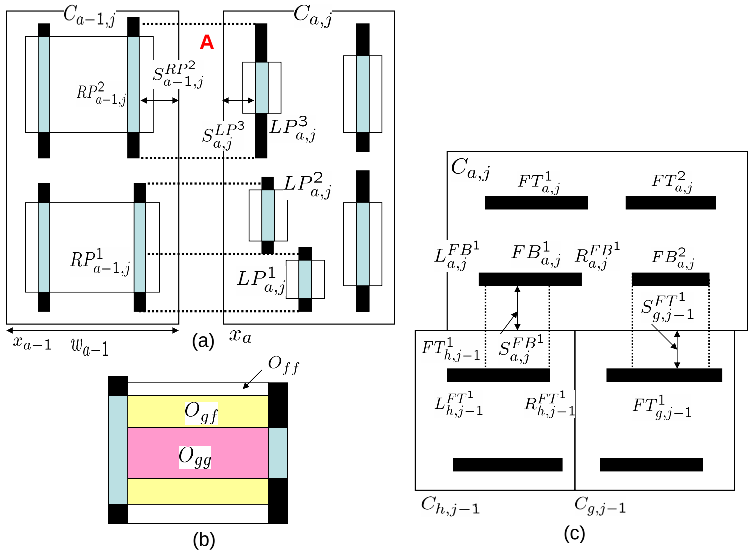


Figure II.3: (a) H-interactions of gate-to-gate, gate-to-field and field-to-gate, (b) overlapped area in the region A of (a), and (c) V-interactions of field-to-field polys.

larger than the minimum allowable length of SRAF features are considered. Define $s^{LP^i}_{a,j}$ to be the space between the left outline of the cell and the i^{th} left border poly geometry. O_{gg}, O_{ff} and O_{gf} correspond to the length of overlapped area in the cases of gate-to-gate, field-to-field and gate-to-field poly as shown in Figure II.3. In addition, c_{gg}, c_{ff}, and c_{gf} are proportionality factors which specify the relative importance of printability for gate and field poly[4]. Typically, gate poly geometries need to be better controlled through process as they have more direct impact on performance. Therefore, a typical order of importance is $c_{gg} \geq c_{fg} \geq c_{ff}$.

- **Vertical polygon interaction:** Given a cell $C_{a,j}$, let $FB_{a,j}$ and $FT_{a,j}$ be the sets of valid field poly geometries in the cell which are located closest to the bottom and top outlines of cell, respectively. Define $s^{FB^i}_{a,j}$ ($s^{FT^i}_{a,j}$) to be the space between the bottom (and top) outline of the cell and the i^{th} bottom (and top) border poly geometry. O_{ff} corresponds to the length of

[4] Gate is the overlap region of polysilicon and diffusion. Field poly represents the rest area of polysilicon except the gate.

field-to-field overlap between horizontal geometries in adjacent cell rows[5].

Assume an ordered set $AF = AF_1, \ldots, AF_m$ of spacings which are "assist-correct," i.e., if the spacing between two gate poly shapes belongs to the set AF, then the required number of assist-features can be inserted between the two poly geometries. For example in Figure II.2, the peaks of the CD correspond to AF_i. The acceptable CD tolerance range (e.g., 10%) results in a range of acceptable pitches starting at AF_i. AF is assumed to be sorted in increasing order. Note that the set AF may contain a number of spacings which correspond to varying SRAF widths. Let w_a denote the width of cell $C_{a,j}$ and let x_a denote its (leftmost) placement coordinate in the given standard cell row, where coordinates increase from left to right. In addition, let $\delta_{a,j}$ denote the placement perturbation of cell $C_{a,j}$ to adjust the spacing between cells. $\delta_{a,j}$ is positive if the cell is moved towards the right and negative otherwise. Then the **assist-correct placement perturbation problem** is:

$$\text{Minimize} \sum \mid \delta_{a,j} \mid$$

$$\delta_{a,j} + x_{a,j} - x_{a-1,j} - \delta_{a-1} - w_{a-1} + s_{a,j}^{LP^f} + s_{a-1,j}^{RP^g} \in AF$$

s.t. LP^f and RP^g overlap at horizontal cell row

$$s_{a,j}^{FB^m} + s_{h,j-1}^{FT^n} \in AF$$

s.t. FB^m and FT^n overlap at vertical cell row

The objective can be made aware of cells in critical paths by a weighting function. Since the available number of allowable spacings is very small, obtaining a completely assist-correct solution is usually not possible in a fixed cell row width context. Therefore, a more tractable objective is to minimize the expected CD error at a predetermined defocus level.

[5] Gates are typically laid out in a single orientation. We assume this orientation to be vertical in this work.

HCost(a,b,a-1,i) of Cell $C_{a,j}$

Input:

User-defined weight for overlapping field-polys: c_{ff}

User-defined weight for overlapping gate-polys: c_{gg}

User-defined weight for overlapping gate and field-polys: c_{gf}

Origin (left x coordinate) of cell $C_{a,j} = b$

Origin $C_{a-1,j} = i$

Width of cell $C_{a,j} = w_a$

Width of cell $C_{a-1,j} = w_{a-1}$

Output:

Value of $HCost(a, b, a-1, i)$: horizontal cost of placing cell C_a at placement site b when C_{a-1} is placed at site i.

Algorithm:

```
Case $a = 1$: $HCost(1, b, 0, i) = 0$
Case $a > 1$ Do
For every pair of left poly (LP) geometry in cell $C_{a,j}$
    and right poly (RP) geometry in cell $C_{a-1,j}$ that overlap{
  /* Let Hspace be the spacing between $LP$ and $RP$. Let $AF_l$ be
  the largest assist correct spacing smaller than Hspace. Let the CD
  degradation slope (delta CD/delta spacing) for $AF_l$ be $slope(l)$. */
    Split the vertical overlap between LP and RP into field-to-field
      $O_{ff}$, field-to-gate $O_{fg}$ and gate-to-gate $O_{gg}$ overlaps.
  /* Calculate overlap weight between $RP$ and $LP$. */
    $weight = slope(l) \times (Hspace - AF_l)$
        $\times(c_{ff}O_{ff} + c_{gf}O_{gf} + c_{gg}O_{gg})$
         s.t. $AF_{l+1} > Hspace \geq AF_l$,
    $HCost(a, b, a-1, i)$ += $weight$
  }
```

Figure II.4: Horizontal cost ($HCost$) calculation.

This "continuous" version of the problem is similar in nature to placement legalization of row-based placements but with manufacturability-based cost metrics instead of traditional wirelength metrics. Placement legalization has been previously solved in literature [59] using dynamic programming techniques. We solve this "continuous" version of the above problem with the following dynamic programming recurrence.

$$\begin{aligned}
Cost(1,b) = &\ \mid x_1 - b \mid \\
Cost(a,b) = &\ \lambda(a) \mid (x_a - b) \mid + Min_{i=x_{a-1}-SRCH}^{x_{a-1}+SRCH}\{Cost(a-1,i) \\
&+ \alpha HCost(a,b,a-1,i) + \beta VCost(a,b)\}
\end{aligned}$$

$Cost(a,b)$ is the cost of placing cell a at placement site number b. The cells and the placement sites are indexed from left to right in the standard cell row. α and β give the relative importance between $HCost$ and $VCost$. Typically, $HCost$ has more weight because $HCost$ is related to gate printability which determines device performance. $HCost$ is the measure of total expected CD degradation of vertical poly geometries at the worst defocus for the cell. It can be thought of as the weighted change in area of vertical poly geometries in the cell. Similarly $VCost$ is the measure of total expected CD degradation of horizontal poly geometries at the worst defocus.

Note the above memory-less cost structure which ensures that once the optimal solution up to cell i is obtained, it contains the optimal solution up to cell $i-1$. This optimal substructure is essential for dynamic programming. We restrict the perturbation of any cell to $\pm SRCH$ placement sites from its initial location. This helps contain the delay and runtime overheads of AFCorr placement post-processing. λ is a factor which decides the relative importance of preserving the initial placement and the final AFCorr benefit achieved for each given cell instance. In the current implementation, λ is directly proportional to the number of critical timing paths that pass through the given cell instance. $HCost$ and $VCost$ correspond to the printability deterioration under defocus conditions for

VCost(a,b) of Cell $C_{a,j}$

```
Input:
  $C_{a,j}$: $a^{th}$ cell in $j^{th}$ row
  User-defined weight for overlapping field-polys: $c_{ff}$
  Origin $x$ (left) coordinate $C_{a,j} = b$
Output:
  $VCost(a,b)$: vertical cost of placing cell $C_a$ at placement site $b$.
Algorithm:
  Case $j = 1$: $VCost(a,b) = 0$
  Case $j > 1$ Do
  For every pair of bottom poly geometry in cell $C_{a,j}(FB)$
    and top poly geometry in cell $C_{h,j-1}(FT)$ that overlap{
      Call the geometries $FB$, $FT$
    /* Let Vspace be the vertical spacing between $FT$ and $FB$.
    Let $AF_l$ be the largest assist correct spacing smaller than Vspace.
    Let $O_{ff}$ denote the field-to-field overlap lengths. */
      $weight = slope(l) \times c_{ff} O_{ff} \times (Vspace - AF_l)$
            s.t. $AF_{l+1} > Vspace \geq AF_l$,
      $VCost(a,b)$ += $weight$
    }
```

Figure II.5: Vertical cost ($VCost$) calculation.

the horizontal and vertical interactions respectively. $Cost(a,b)$ depends on the difference between the current nearest-neighbor spacing of the polys and the closest assist-feature correct spacing. The methods that we use to compute $HCost$ and $VCost$ are shown in Figures II.4 and II.5. $Slope(l)$ is defined as delta CD difference over delta pitch between AF_l and AF_{l+1}. Thus, perturbation cost is a function of *slope*, length and weight of overlapped polys, and space for SRAF insertion. Our algorithm takes a legal placement as an input, and outputs a legal placement with

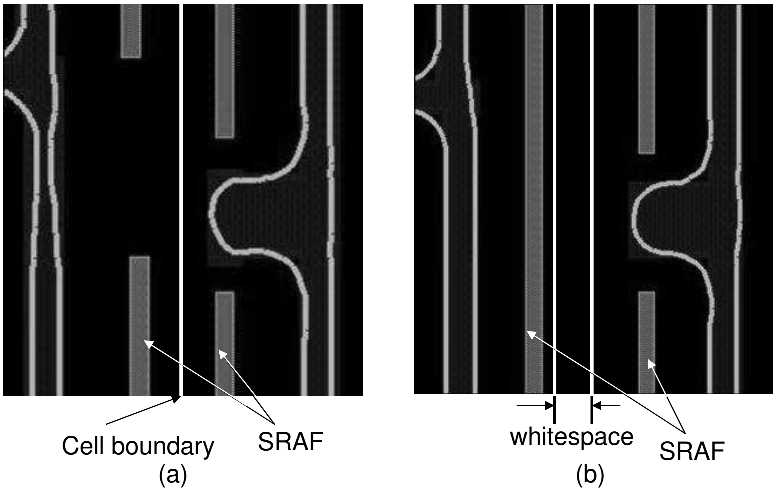


Figure II.6: (a) Cell placement before horizontal AFCorr and (b) cell placement after horizontal AFCorr.

better depth of focus properties. In addition, $VCost$ depends on the number of abutted cells, L and R, and the number of field-to-field poly interactions. The runtime of the AFCorr algorithm is $O(ncell \times SRCH)$, where $ncell$ is the total number of cells in the design.

Figure II.6 shows an example of a resist image profile with and without AFCorr technique. Horizontal-forbidden pitch is caused by interactions of poly geometries in the same row. After cell placement-perturbation in the horizontal direction, additional SRAFs can be inserted at increased whitespace between cells and thus enhance pattern printability. In addition, vertical forbidden pitch violation is caused by inter-cell row interactions. As seen in Figure II.7(a), there is not enough space between the vertically adjacent poly geometries (coming from cells in adjacent cell rows) which results in less SRAFs than needed. By moving the cell in the upper row leftwards, this violation can be removed and printability enhanced.

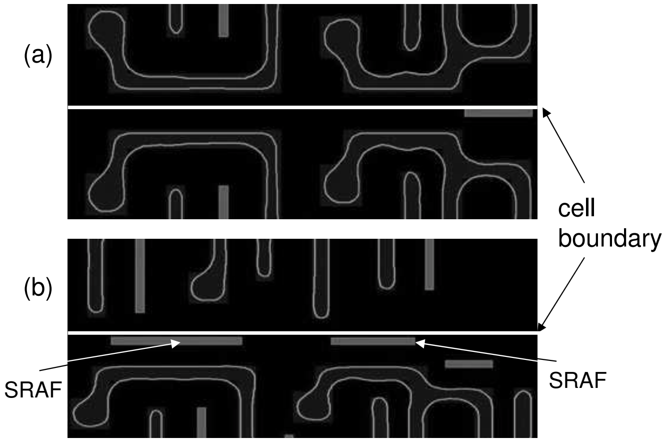


Figure II.7: (a) Cell placement before vertical AFCorr and (b) cell placement after vertical AFCorr.

II.B.4 Experimental Setup and Results

We synthesize the AES and ALU benchmark design from Opencores in Artisan TSMC 0.13μm and Artisan TSMC 0.09μm libraries using Synopsys *Design Compiler* (v 2003.06-SP1). AES synthesizes to 12993 cells and 10286 cells in 130nm and 90nm technologies, respectively. ALU synthesizes to 13279 cells and 8722 cells in 130nm and 90nm technologies, respectively. The synthesized netlists are placed with row utilization ranging from 50% to 90% using Cadence *SOC Encounter* (v 2004.10). All designs are trial routed before running timing analysis. On the lithography side, we use KLA-Tencor *Prolith* (v 9.1) to generate models for OPC. Mentor *Graphics Calibre* (v 9.3_5.12) is used for model-based OPC, SRAF OPC and optical rule checking (ORC). Photo simulation is performed with wavelength λ = 248nm and numerical aperture NA = 0.6 for 130nm, and λ = 193nm and NA = 0.75 for 90nm. An annular aperture with σ = 0.85/0.65 is used.

We use three printability quality metrics. *Forbidden Pitch Count* is the number of border poly geometries estimated as having greater than 10% CD error through-focus. *EPE Count* is the number of edge fragments on border poly geometries having greater than 10% edge placement error at the worst defocus level. *SB Count* is the total number of scattering bars or SRAFs inserted in the

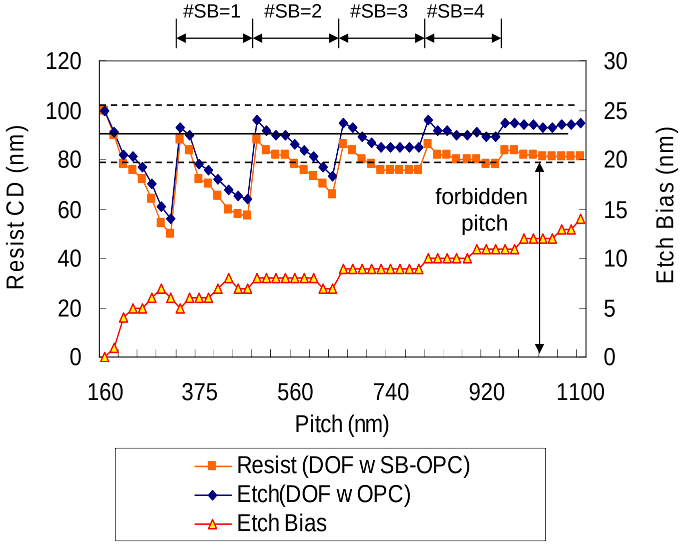


Figure II.8: Through-pitch proximity plots and etch skew for 90nm technology with worst defocus with SRAF OPC and worst defocus with etch OPC (left Y-axis), and etch bias (right Y-axis).

design. A higher number of SRAFs indicates less through-focus variation and is hence desirable. We use $c_{fg} = c_{gg} = c_{ff} = 0.33$, $\lambda(a) = sitewidth/10 \times$ (number of top 200 critical paths passing through cell a) and $SRCH = 20$.

We first evaluate lithography printability of AFCorr with H- and V- assist correction. Proximity plot with fixed linewidth for the 0.13μm RET is illustrated in Figure II.2. CD degradation increases through-pitch as the defocus level increases. Patterns in the pitches of over 0.4μm before OPC are outside the allowable tolerance range at the worst defocus of 0.5μm. After bias OPC, pitches up to 0.38μm are allowable for CD tolerance while all pitches larger than 0.38μm are forbidden. After evaluating SRAF OPC patterns with the worst defocus model, a set of forbidden pitches of 0.13μm technique is obtained: [0.37, 0.51), [0.635, 0.73), [0.82, 0.95), and [1.09, 1.17) microns. Forbidden pitches still remain after SRAF OPC even though SRAF insertion considerably reduces forbidden pitches in comparison to bias OPC. Proximity plot with SRAF OPC for 90nm technology is illustrated in Figure II.8. Resist CDs after SRAF OPC are evaluated with the

Table II.2: Summary of forbidden pitch results. Forbidden pitch counts change slightly based on different H- vs. V-weights.

	Utilization (%):	90		80		70		60		50	
	H:V weight	H F/P	V F/P	H F/P	V F/P	H F/P	V F/P	H F/P	V F/P	H F/P	V F/P
130nm	0.9:0.1	4002	92	290	21	2	5	0	0	0	0
	0.7:0.3	5234	60	533	15	5	2	1	0	0	0
	0.5:0.5	5878	54	573	14	10	1	2	0	0	0
90nm	0.9:0.1	4639	82	541	21	10	5	0	0	0	0
	0.7:0.3	5321	70	721	15	11	2	1	0	0	0
	0.5:0.5	6072	43	891	14	14	1	1	0	0	0

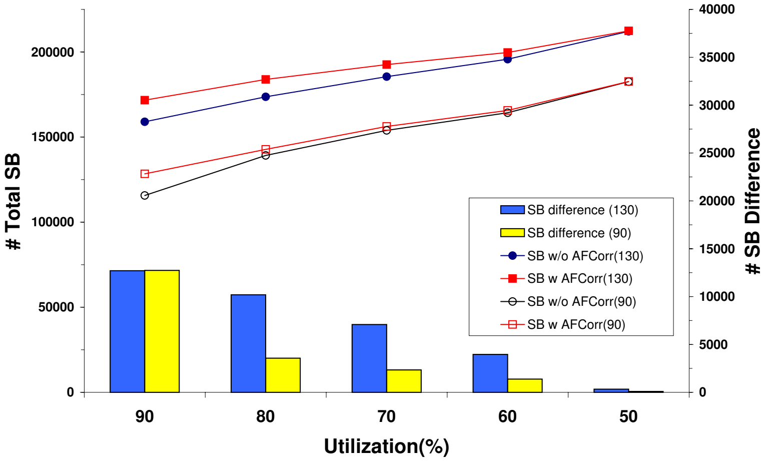


Figure II.9: Number of SRAFs with and without AFCorr for each of five different utilizations.

worst defocus model of 0.4μm. Resist CDs violate the allowable CD tolerance[6] as distance between SRAF and poly increases. A set of forbidden pitches of resist CD for 90nm RET is calculated: [0.3, 0.41), [0.45, 0.57), [0.64, 0.73), and [0.78, 0.89) microns. The generated SRAF rules may be summarized as shown in Table II.1. SRAF width and SRAF-to-pattern space are respectively 40nm and 120nm for 90nm technology.

Table II.2 shows the results of horizontal and vertical forbidden pitches with various H- vs. V-weights. Increasing the weight of HCost reduces the number of horizontal forbidden pitches while increasing the number of vertical forbidden pitches. H- and V-forbidden pitch counts are reduced by 94%-100% and 76%-100% for 130nm, and by 96%-100% and 87%-100% for 90nm, respectively. The design with 0.9 α for HCost and 0.1 β for VCost weights results in the highest reduction of total forbidden pitch counts and is chosen to evaluate SB count, runtime, etc. Figure II.9 shows that the total number of SRAFs increases as the utilization decreases, due to increased whitespace between cells. The benefit of AFCorr decreases with lower utilization because the design already has enough whitespace

[6] Allowable CD tolerance is assumed to be 10% of minimum linewidth in the worst defocus level.

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

	Utilization (%):	90		80		70		60		50	
	Flow:	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr	Typical	AFCorr
130nm	# Forbidden	20632	4094	3201	311	2011	7	1421	0	219	0
	# SB	158987	171691	173673	183860	185493	192578	195741	199704	212079	212412
	# EPE	4630	4721	5975	562	4276	15	1732	0	199	0
	Runtime (s)	7821	7902	7876	7934	7913	7973	7998	8013	8021	8121
	GDS (MB)	48.9	48.9	48.8	48.9	48.2	48.4	48.3	48.5	48.2	48.4
	Delay (ns)	4.2	4.6	4.5	4.7	4.5	4.6	4.6	4.9	5.2	5.4
90nm	# Forbidden	22121	4721	4821	562	3812	15	2001	0	321	0
	# SB	115652	128387	139182	147520	153904	156244	164264	165649	182572	182666
	# EPE	7523	1262	4813	532	2131	107	1329	59	163	5
	Runtime (s)	6211	6327	6322	6431	6482	6499	6521	6571	6672	6692
	GDS (MB)	43.1	43.3	43.2	43.3	43.2	43.3	43.7	43.8	44.6	44.8
	Delay (ns)	2.7	2.7	2.6	2.6	2.4	2.47	2.8	2.9	3.1	3.2

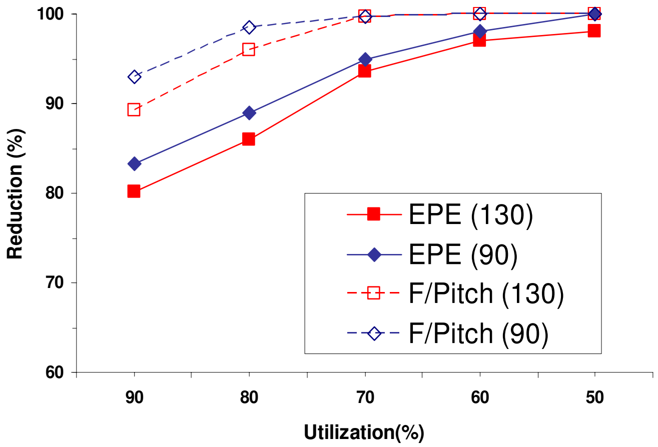


Figure II.10: Reductions of forbidden pitches with AFCorr methodology for each of five different utilizations.

for SRAF insertion. Because of the additional number of SRAFs inserted, there is a small increase in SRAF OPC runtime ($< 3.6\%$) and final data volume ($< 3\%$). Reductions of edge placement errors (EPE) and forbidden pitch are investigated for each utilization as shown in Figure II.10. Total Forbidden Pitch Count is reduced by 89%-100% in 130nm and 93%-100% in 90nm. EPE Count is reduced by 80%-98% in 130nm and 83%-100% in 90nm. In addition, SB Count improves by 0.1%-7.4% for 130nm and 0%-7.9% for 90nm. Note that these numbers are small as they correspond to the entire layout rather than just the border poly geometries. The change in estimated post-trial route circuit delay ranges from -7% to +11%. All of these results for AFCorr are summarized in Table II.3.

II.C Etch-Dummy Correctness

II.C.1 Etch-Dummy and Layout Impact

Etch-dummy features are inserted to reduce the CD difference between resist and etch processes for 90nm and below technology nodes. In dry etch processes such as plasma, ion, and reactive ion etch (RIE), different consumptions

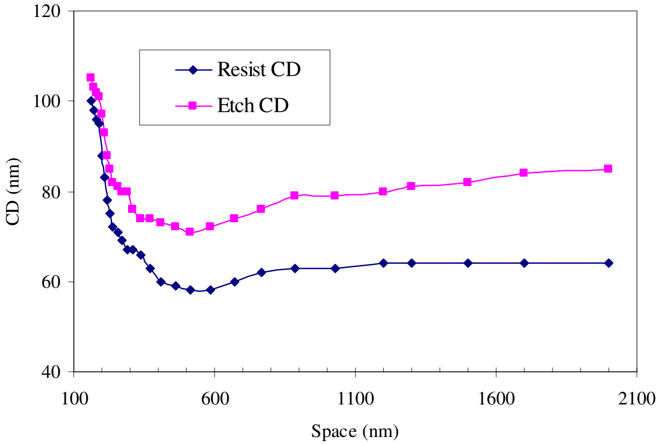


Figure II.11: Different proximity behaviors between photo and etching processes with pitch.

of etchants with different pattern density lead to etch skew between dense and isolated patterns. For example, all available etchants in areas with low density are consumed rapidly, and then the etch rate drops off significantly. In areas with high density of patterns, the etchants are not consumed as quickly. As a result, the proximity behavior of the photo process differs from that of etch process, as shown in Figure II.11. Therefore, there is CD skew between resist and etch process with varying pitch. In general, the etch skew of two processes increases as pitch increases. OPC is typically used to compensate for CD variation of the resist process, and then etch-dummies are inserted to reduce the CD skew between two processes.

When etch-dummies are placed adjacent to primary patterns, a relatively isolated primary line will behave more like a dense line, and thus the etch-dummies can reduce the etch skew. Moreover, the maximum relevant pitch is reduced through etch-dummy insertion. This is an important consideration with respect to model-based OPC, which calculates the proximity effect of all patterns within a given proximity range, such that larger proximity range increases OPC runtime. Granik [39] observes that the proximity range of the etch process is around 3μm, which prevents conventional model-based OPC from delivering a good OPC mask

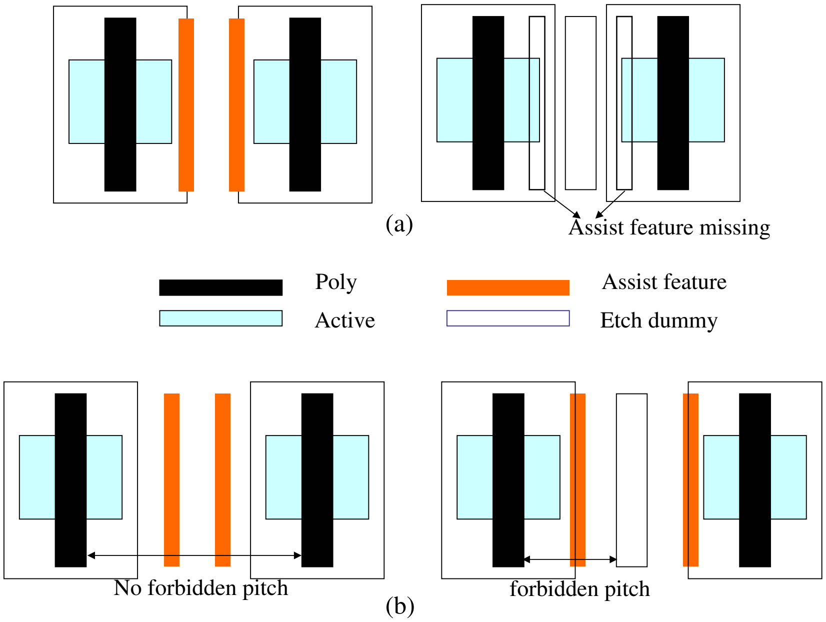


Figure II.12: Conflict between SRAF, etch-dummy rules: (a) assist-feature missing and (b) forbidden pitch occurrence.

within feasible turnaround time.

Etch-dummy correction problem. Given a layout, find an etch-dummy placement such that the following conditions are satisfied:

- Condition (1): Etch dummies are inserted between primary patterns with certain spacing to reduce etch skew between resist and etch processes.
- Condition (2): Etch dummies are placed outside of active-layer regions.

Thus, *Etch-dummy correction problem* is to determine perturbations to inter-cell spacings so as to insert the optimal number of etch-dummies. Forbidden pitch correction in the resist process is required after inserting etch-dummy because the etch-dummy cannot be placed too closely to primary patterns due to Condition (2). Etch-dummy insertion can make printability of resist process worse in certain pattern configurations. Figure II.12 shows examples such as (a) assist-features missing and (b) forbidden pitch occurrence. Assist-features can be missed due to lack of space between primary pattern and etch-dummy, even when

Table II.4: Comparison of etch-dummy rules between conventional etch-dummy method and SAEDM. Note that AS_l + AS_r = ES - ED_l.

Etch-dummy rules		Typical method		SAEDM	
	ES (X)	DS_l	DS_r	DS_l	DS_r
#ED = 0	$0 \leq X < ED_1$				
#ED = 1	$ED_1 \leq X < ED_2$	$(ES$ -$EW)/2$	$(ES$ - $EW)/2$	$AS_l + DAS$	$AS_r + DAS$
#ED = 2	$X \leq ED_2$	DAS	DAS	$AS_l + DAS$	$AS_r + DAS$

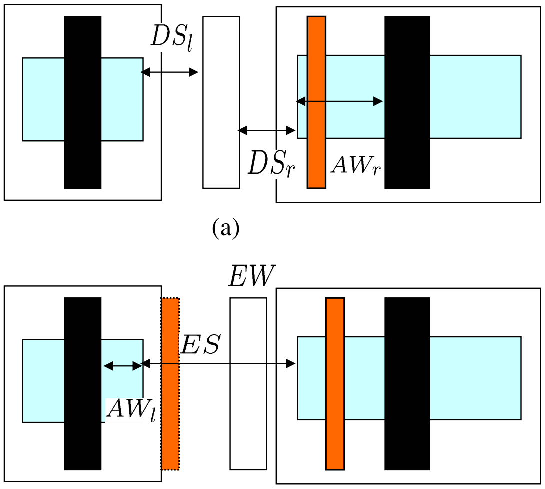
there is enough space to insert multiple SRAFs before etch-dummy insertion. New forbidden pitches for assist-features can occur in the spacing between poly and etch-dummy due to mismatch between rules for assist-feature and etch-dummy corrections. Therefore, we now propose a new *Corr problem* that combines both assist-feature and etch dummy insertion methods as follows.

Assist-feature and etch-dummy correction problem. Given a standard cell layout, determine perturbations to inter-cell spacings so as to simultaneously insert SRAFs in forbidden pitches and insert etch-dummies to reduce etch skew.

II.C.2 SRAF-Aware Etch-Dummy Generation

To reduce etch proximity, at most one etch-dummy for each active (or diffusion) geometry is needed since the etch skew depends on pattern-to-pattern spacing regardless of local pattern density [50], i.e., etch skew decreases as the spacing is reduced. SRAFs and etch-dummies are generated by rule-based methods with look-up tables (LUTs) since simulation tools are much slower than rule-based tools.

Typically, etch-dummy rules consist of etch dummy-to-active space (DAS), etch-dummy width (EW) and etch dummy-to-dummy space (DDS) with respective values of 120nm, 100nm and 200nm being typical for 90nm technology. Let ES denote the space between active geometry in the left and right cells as shown in Figure II.13. Let ED_1 and ED_2 denote the required spaces to insert one and two



(a)

(b)

Figure II.13: (a) Typical etch-dummy generation and (b) SRAF-aware etch-dummy generation.

etch-dummies in ES, respectively. For typical methods of etch-dummy insertion, minimum space rules for one and two etch-dummies are $ED_1 = 2*DAS+EW$ and $ED_2 = 2*DAS+2*EW+DDS$, respectively. The first etch-dummy in the typical etch-dummy rule is always placed at the center of the space between two active geometries, while the active-to-etch dummy space for the second etch-dummy is always according to the space rule, DAS. Once etch-dummies have been inserted for only etch proximity control, the spacing between poly and etch-dummy may not be appropriate for SRAF insertion. Figure II.13(a) shows an example where the left-hand side SRAF cannot be inserted due to lack of poly-to-etch dummy spacing.

Let AW_l and AW_r denote the distances between border polys and active geometries located at left- and right-cells, respectively. Let $AF = AF_1, \ldots, AF_m$ denote a set of "assist-correct" spacings. AF_j is the j^{th} member of the set of assist-feature correct spacings AF. Let AS_l and AS_r denote additional spacings needed for assist-correctness in the left- and right-cells, respectively. To avoid missing SRAFs and occurrence of forbidden pitches, we propose a new *SRAF-aware etch-dummy method* (SAEDM) considering active width (AW) during insertion of etch-dummy, as follows:

$$\begin{aligned} &\mathit{Minimize} && \text{index values of } j \text{ and } k \text{ in a set } AF \\ &s.t. && AS_l = AF_j - (AW_l + DAS) \text{ and } AS_r = AF_k - (AW_r + DAS), \\ & && \text{and } (AS_l + AS_r) \leq (ES - ED_1) \end{aligned} \tag{II.1}$$

SAEDM searches assist-correct spacing with minimum index values in a set AF, so that the sum of the additional spacings AS_l and AS_r corresponding to assist-correct spacings is less than $(ES - ED_1)$. Let DS_l and DS_r denote the left- and right-spaces from etch-dummy to border active geometries in left- and right-cells, respectively. Thus, new etch-dummy spaces of $DS_l = AS_l + DAS$ and $DS_r = AS_r + DAS$ are both assist-correct and etch dummy-correct. Note that the etch-dummy after SAEDM is no longer located at the center of an active-to-active space since DS_l differs from DS_r, as shown in Figure II.13(b). Table II.4 compares DS_l and DS_r values returned by the typical etch-dummy method and by SAEDM.

II.C.3 Corr Placement Algorithm

Assist-correct pitch rules are violated if there is not enough space to insert AS_l and AS_r. We now describe an etch-dummy correction *EtchCorr* placement perturbation algorithm using intelligent whitespace management. EtchCorr differs from AFCorr as follows: (1) EtchCorr is based on the active-to-cell outline spacing while AFCorr is poly-to-cell outline spacing. (2) EtchCorr calculates the virtual positions of etch-dummy in order to both insert SRAF in assist-correct spacing and etch-dummy in etch dummy-correct spacing. Let etch dummy-correct spacing (EDS) be inter-device spacing with etch skew less than 10% of minimum linewidth. Thus the etch dummy-correct perturbation problem is to minimize design perturbation to insert etch-dummies optimally and thus to reduce etch skew between resist and etch processes. However, as we discussed, a new design correction technique *Corr* which combines the two methods of assist-correct (AFCorr) and etch-

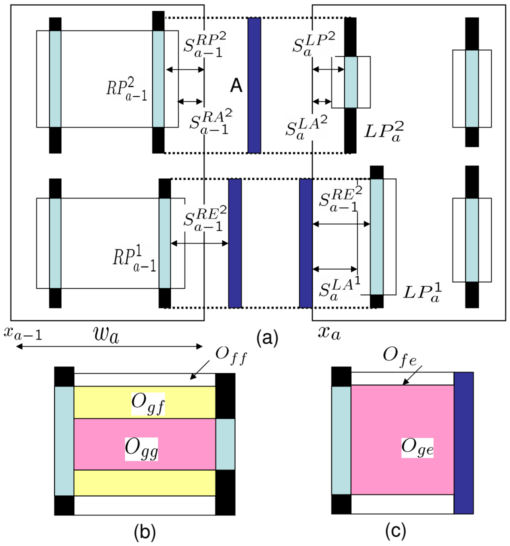


Figure II.14: The placement perturbation problem for assist and etch-dummy insertion: (a) multiple interactions of gate-to-dummy and field-to-dummy, (b) overlap area when there is no etch-dummy and (c) overlap area in presence of etch-dummy.

correct (EtchCorr) placements is required to avoid conflict between assist-feature and etch-dummy insertions.

In the following, we describe the single-row Corr perturbation algorithm. Let $s_a^{RP_i}$ and $s_a^{RA_j}$ respectively denote the spacing between the right outline of the cell and the i^{th} right border poly, and the spacing between the right outline of the cell and j^{th} active geometry. $s_a^{RE_i}$ is the spacing from right border poly to etch-dummy as shown in Figure II.14. Let δ denote a cell placement perturbation to adjust the spacing between cells. ES, the space between border actives, is $x_a - x_{a-1} - w_{a-1} + s_{a-1}^{RA_i} + s_a^{LA_i}$. Then the **Corr placement perturbation problem** is:

Minimize $\sum \mid \delta_i \mid$ such that

$$\begin{cases} \text{If } (ES < ED_1) \\ \delta_a + x_a - x_{a-1} - \delta_{a-1} - w_{a-1} + s_{a-1}^{RP^i} + s_a^{LP^i} \in AF \\ \delta_a + x_a - x_{a-1} - \delta_{a-1} - w_{a-1} + s_{a-1}^{RA^i} + s_a^{LA^i} \in EDS \\ \text{s.t. } -SRCH \leq \delta_{a-1} \text{ and } \delta_a \leq SRCH \\ \text{otherwise} \\ S_{a-1}^{RP^i} - S_{a-1}^{RA^i} + S_{a-1}^{RE^i} + \delta_{a-1} \text{ and } S_a^{LP^i} - S_a^{LA^i} + S_a^{LE^i} + \delta_a \in AF \\ S_{a-1}^{RE^i} + \delta_{a-1} \text{ and } S_a^{LE^i} + \delta_a \in EDS \\ \text{s.t. } -SRCH \leq \delta_{a-1} \text{ and } \delta_a \leq SRCH \end{cases}$$

The terms $AFCost$ and $EDCost$ denote assist-feature and etch-dummy costs, respectively. $AFCost$ depends on the difference between the current nearest-neighbor spacing of the polys and the closest assist-correct spacing. The methods of computing $AFCost$ and $EDCost$ are shown in Figure II.15[7]. Let $AFslope(j)$ be defined as ratio of resist CD degradation and change in pitch between AF_j and AF_{j+1}. $EDslope(j)$ is the ratio of etch CD degradation and poly-to-dummy space. ED_1 is the required space to insert one etch-dummy. The formulation is similar to the AFCorr when the space between border actives is not enough for a dummy insertion. However, $Corr$ perturbation problem calculates poly-to-dummy spacings instead of poly-to-poly spacings when there are etch-dummies between cells. O_{gg}, O_{ff} and O_{gf} respectively correspond to the length of overlap areas of gate-to-gate, field-to-field and gate-to-field poly as shown in Figure II.14. O_{ge} and O_{fe} correspond to the overlapped length of gate-to-dummy and field-to-dummy. In addition, c_{gg}, c_{ff}, and c_{gf} are proportionality factors which specify the relative

[7] The Figure shows only H-AFCost computation for simplicity. We do not include computation of the vertical EDCost, as the primary focus of etch-dummies is gate CD control.

Cost(a,b,a-1,i) of Cell C_a

Input:

User-defined weights for poly-to-poly overlap: c_{gg}, c_{ff}, c_{gf}

User-defined weights for poly-to-dummy overlap: c_{ge}, c_{fe}

Width of cell $C_a = w_a$

Output:

Value of $AFCost$ and $EDCost$: costs for corrections of assist-feature and etch-dummy of placing cell C_a at placement site b, respectively.

Algorithm:

```
/* Cost of placing cell $C_a$ at placement site ‘b’ when cell $C_{a-1}$ is placed at site ‘i’. */
Case $a = 1$: $AFCost(1, b) = EDCost(1, b) = 0$
Case $a > 1$ Do {
If $(AFspace < ED_1)$ {
  For every pair of left poly geometry in cell $C_a(LP)$
    and right poly geometry in cell $C_{a-1}(RP)$ that overlap{
    Call the geometries $LP$, $RP$
    Split the vertical overlap between LP and RP into field-to-field $O_{ff}$,
      field-to-gate $O_{fg}$ and gate-to-gate $O_{gg}$ overlaps.
    $AFweight = AFslope(j) \times (AFspace - AF_j)$
        $\times(c_{ff}O_{ff} + c_{gf}O_{gf} + c_{gg}O_{gg})$ s.t. $AF_{j+1} > AFspace \geq AF_j$
    $EDweight = EDslope(AFspace) \times(c_{ge}O_{ge} + c_{fe}O_{fe})$
  }}
Else {
  For every pair of pattern geometries in $C_a(LP)$, $C_{a-1}(RP)$ and dummy that overlap{
    Call the geometries $LP$, $RP$, and a dummy pattern
    Split the vertical overlap between poly and dummy into gate-to-dummy $c_{ge}$
      and poly-to-dummy $c_{fe}$ overlaps.
    $AFweight = AFslope(j) \times (AW_l + DS_l - AF_j) \times (c_{ge}O_{ge} + c_{fe}O_{fe})$
    $AFweight += AFslope(l) \times (AW_r + DS_r - AF_l) \times (c_{ge}O_{ge} + c_{fe}O_{fe})$
    $EDweight = (EDslope(AW_1 + DS_l) + EDslope(AW_r + DS_r)) \times (c_{ge}O_{ge} + c_{fe}O_{fe})$
  }}
    $AFCost(a, b, a-1, i) += AFweight$
    $EDCost(a, b, a-1, i) += EDweight$
  }
```

Figure II.15: The algorithm for $AFCost$ and $EDCost$ computations.

importance of printability for gate and field poly. W_1 and W_2 are user-defined weights for $AFCost$ and $EDCost$, respectively.

$$
\begin{aligned}
Cost(1,b) = &\ \mid x_1 - b \mid \\
Cost(a,b) = &\ \lambda(a) \mid (x_a - b) \mid + Min_{i=x_{a-1}-SRCH}^{x_{a-1}+SRCH}\{Cost(a-1,i) \\
&+ W_1 AFCost(a,b,a-1,i) + W_2 EDCost(a,b,a-1,i)\}
\end{aligned}
$$

II.C.4 Modified Design and Evaluation Flow

To account for new geometric constraints that arise out of SRAF OPC in physical design, we add forbidden pitch extraction and post-placement optimization into the current ASIC design methodology. Figure II.16 shows the modified design and evaluation flow in the regime of forbidden pitch restrictions. Of course, we must assume that the library cells themselves have been laid out with awareness of forbidden pitches, and indeed our experiments with commercial libraries confirm that there are no forbidden pitch violations in poly geometries within individual commercial standard cells. Our method solves forbidden pitch violations between placed cells. SRAF insertion rules to enhance DOF margin are determined based on best and worst focus models.[8]

The post-placement optimization is performed based on forbidden pitches and slopes of CD error within them. After AFCorr, we obtain a new placement which is more conducive to insertion of SRAFs, thus allowing a larger process window to be achieved. The two layouts generated by conventional and assist-correct flow undergo comprehensive SRAF OPC. The amount and impact of the applied RET is a function of the circuit layout. Thus we can evaluate how assist-correct placement impacts circuit performance and printability/manufacturability according to the metrics of SRAF insertions and edge placement errors (EPE). The following sections give more details of forbidden pitch extraction and design implementation.

[8] In general, the best focus is shifted from zero to about 0.1μm due to refraction in the resist. The worst defocus is the maximum allowable defocus corner for manufacturability in a lithography system. As the CD tolerance is +/-10%, the worst defocus model can be extracted by the Bossung plot in Fig II.1, i.e., worst defocus model is 0.5μm for 130nm technology.

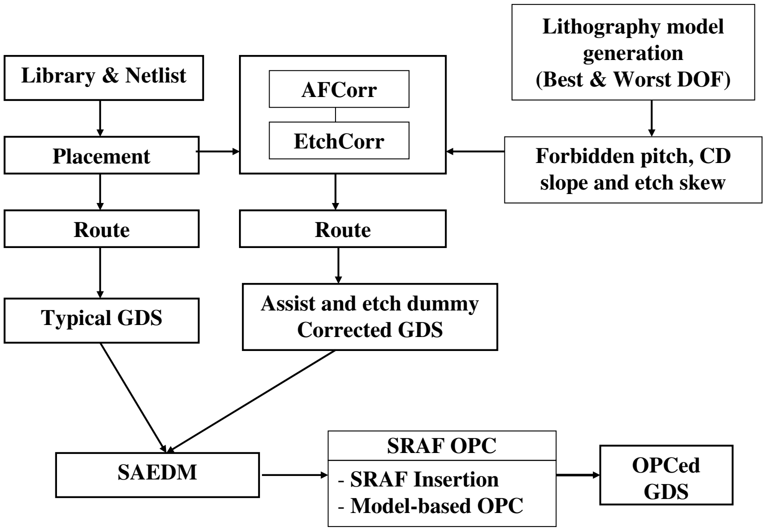

Library & Netlist
Placement
Route
Typical GDS
AFCorr
EtchCorr
Route
Assist and etch dummy Corrected GDS
Lithography model generation (Best & Worst DOF)
Forbidden pitch, CD slope and etch skew
SAEDM
SRAF OPC
- SRAF Insertion
- Model-based OPC
OPCed GDS

Figure II.16: The modified design and evaluation flow. Note the steps of forbidden pitch extraction, SAEDM and post-placement optimization that are added to the ASIC design flow.

Table II.5: Etch process conditions for the simulator in 90nm technique.

Stage	Etch time (sec)	Material	Vertical etch rate (sec)	Horizontal etch rate (sec)	Faceting Parameter Parameter
1	10	ArF Sumitomo	10.66	-0.6	0.5
		AZ BarLi-2	10.52	-0.7	0.0
		Si Nitride	10.28	-0.7	0.0
2	60	ArF Sumitomo	0.3	-0.12	0.5
		AZ BarLi-2	3.4	-0.2	0.0
		Si Nitride	30.4	-0.3	0.0
3	36	ArF Sumitomo	10.65	0.9	0.5
		AZ BarLi-2	0.25	1.0	0.0
		Si Nitride	0.0	1.5	0.0

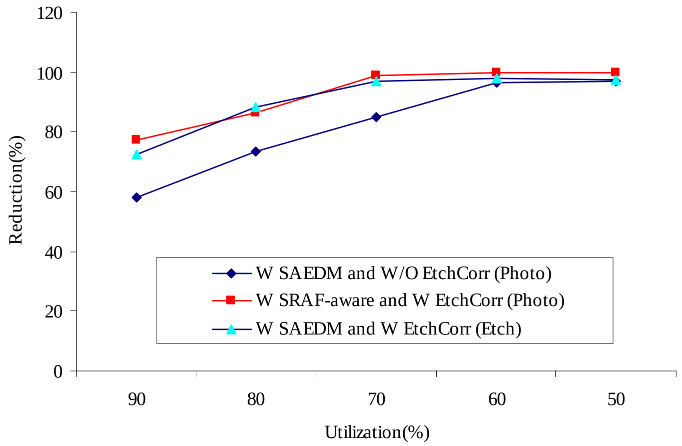


Figure II.17: Reductions of forbidden pitches with various etch-dummy insertion methodologies for each of five different utilizations.

II.C.5 Experimental Setup and Results

To account for new geometric constraints that arise from SRAF and etch-dummy in physical design, we extract forbidden pitch, CD slopes of resist and etch process with pitch, and CD skew induced by etch process. Post-placement optimization generates a new placement wherein the coordinates of cells have been adjusted to avoid the forbidden pitches and to reduce etch skew. The target etch process consists of three etch steps: (1) 10 second breakthrough etch step to get through the BARC (Bottom Anti-Reflective Coating), (2) 60 second main etch step, and (3) 36 second overetch step. The breakthrough and main etch steps in the model produce a fair amount of deposition, taking the resist profile to 100nm. The overetch step trims this back to the 90nm range. A set of etch parameters is shown in the Table II.5. We only consider the first breakthrough etch step to remove Si Nitride because the second etch, step to etch gate poly, does not impact CD variation with pitch [8].

We use the same benchmark designs as AFCorr and evaluate pattern printability with combinations of (1) SAEDM, (2) SAEDM+AFCorr and (3) SAEDM+AFCorr+EtchCorr (i.e., SAEDM+Corr). We generated SRAF rules with results in Table II.1. SRAF width and SRAF-to-pattern space are 40nm and 120nm, respectively. In addition, dummy-to-active space, etch-dummy width and etch

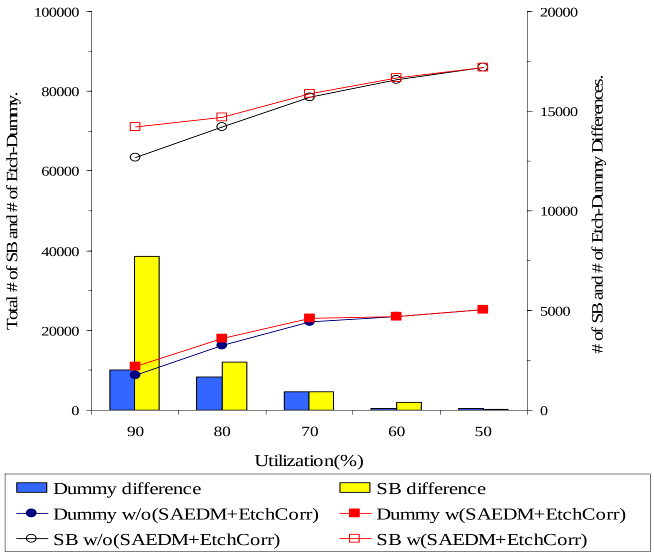


Figure II.18: Number of inserted SRAF and etch-dummy features with various etch-dummy insertion methodologies for each of five different utilizations.

dummy-to-dummy space correspond to 120nm, 100nm and 200nm respectively. However, the spacing between active and etch-dummy varies because SAEDM changes the space with the active width. Resist and etch CDs vary with location of the SRAF insertion, and resist CDs violate the allowable CD tolerance as distance between SRAF and poly increases. The trend of etch CD follows the variation of resist CD. The skew of resist and etch CDs continuously increases with pitch and is not saturated by 1.1μm as shown in Figure II.11.

After Corr placement perturbation, we obtain a new placement wherein the coordinates of cells minimize the occurrence of forbidden pitches of resist and etch processes. Total cost of Corr is calculated using specific weights of resist and etch costs (in the results reported, we use respective weights $W_1 = 0.9$ and $W_2 = 0.1$). Note that our post-placement perturbation problem reduces to the previously-studied AFCorr problem if $W_2 = 0$.

We evaluate the reduction of Forbidden Pitch Count with various etch-dummy insertion methodologies in resist and etch processes shown in Table II.6. After (1) SAEDM, Forbidden Pitch Count of photo process can be reduced by 57% - 94% with various utilizations because etch dummy-to-poly spacings be-

Table II.6: Forbidden pitch results with various etch-dummy insertion methodologies in photo (or resist) and etch processes.

	Utilization (%):	90	80	70	60	50
Photo	W/O SAEDM, W/O AFCorr, W/O EtchCorr	37433	31314	29216	26765	21282
	(1) W SAEDM, W/O AFCorr, and W/O EtchCorr	15743	8330	4423	2075	1198
	(2) W SAEDM, W AFCorr, and W/O EtchCorr	2432	822	23	0	0
	(3) W SAEDM, W AFCorr, and W EtchCorr	3566	1116	51	0	0
Etch	W/O SAEDM, W/O AFCorr, and W/O EtchCorr	15816	8812	4656	4345	3530
	(1) W SAEDM, W/O AFCorr, and W/O EtchCorr	16418	9729	5282	5002	4209
	(2) W SAEDM, W AFCorr, and W/O EtchCorr	5423	2221	172	109	92
	(3) W SAEDM, W AFCorr, and W EtchCorr	4321	1032	143	92	92

come assist-correct. However, Forbidden Pitch Count of the etch process may increase by up to 6% in certain layout configurations since the SAEDM increases the poly-to-etch dummy spacing. The Forbidden Pitch Counts of etch process in (2) SEADM+AFCorr and (3) SAEDM+Corr are respectively reduced by up to 64%-97% and 73%-98% across a range of utilizations as shown in Figure II.17. (3) SAEDM+Corr facilitates additional SRAF and dummy insertion by up to 10.8% and 18.6%, respectively. Figure II.18 shows that the total number of SRAFs and etch dummies increases as the utilization decreases. Note that these numbers are small as they correspond to the entire layout rather than just the border poly geometries. EPE Count is reduced by 91%-100% in resist process and 72%-98% in etch process. The change in estimated post-trial route circuit delay ranges from 3% to 5.8%. The increases of data size and OPC runtime overheads of Corr are within 3% and 4% respectively. The runtime of Corr placement perturbation is negligible ($\sim$ 5 minutes) compared to the runtime of OPC ($\sim$ 2.5 hours). All of these results for Corr are summarized in Table II.7.

Table II.7: Summary of SAEDM+Corr results. Runtime denotes the runtime of SRAF and etch-dummy insertion, as well as model-based OPC. The Corr perturbation runtime ranges from 4 to 5 minutes for all testcases. GDS size is the post-OPC data volume.

	Utilization(%):	90		80		70		60		50	
	Flow:	Typical	SAEDM+Corr	Typical	SAEDM+Corr	Typical	SAEDM+Corr	Typical	SAEDM+Corr	Typical	SAEDM+Corr
Photo	# EPE	42102	3723	32434	1243	29349	98	28721	13	23134	2
	# Forbidden	37433	3566	31314	1116	29216	51	26765	0	21282	0
	# SB	63349	71051	71101	73501	78513	79432	82820	83230	85991	86026
Etch	# EPE	17209	4812	9213	1200	4820	182	4821	109	3890	109
	# Forbidden	15816	4321	8812	1032	4656	143	4345	92	3530	92
	# Dummy	8876	10911	16240	17920	22088	23001	23390	23499	25237	25309
Other	Runtime (s)	6835	7011	7451	7535	7529	7632	7685	7698	7943	7944
	GDS (MB)	41.1	42.3	41.2	43.2	42.2	42.3	42.9	42.8	43.6	43.6
	Delay (ns)	2.478	2.305	2.458	2.602	2.522	2.47	2.867	3.176	3.113	3.046

II.D Conclusions

In this work, we have presented novel methods to optimize etch-dummy insertion rules and detailed standard cell placements for improved etch-dummy and assist-feature insertion. We obtain a practical and effective approach to achieve assist-feature compatibility in physical layouts. The *AFCorr*, as an approach to achieve assist-feature compatibility, leads to reduced CD variation and enhanced DOF margin. We also introduce a dynamic programming-based technique, *Corr*, to achieve etch-dummy insertion correctness in the detailed placement step of standard cell based chip implementation. Corr with SAEDM leads to reduced CD variation and increased insertion of assist-features and etch-dummies. For our testcases we have observed the following.

- In lithographic printability evaluation of AFCorr, H- and V-forbidden pitch counts for border poly geometries are reduced by 94%-100% and 76%-100% for 130nm, and by 96%-100% and 87%-100% for 90nm, respectively. For EPE count, the reductions of 80%-98% in 130nm and 83%-100% in 90nm are obtained. We also achieve up to 7.6% increase in the number of inserted scattering bars.

- In pattern printability evaluation, the forbidden pitch count of photo process between polysilicon shapes of neighboring cells is reduced by up to 54%-94%, 92%-100%, and 90%-100% for SAEDM, SAEDM+AFCorr and SAEDM+Corr, respectively. The forbidden pitch count of etch process of SEADM+AFCorr, and SAEDM+Corr is respectively reduced by up to 64%-97% and 73%-98% across a range of utilization. Corr with SAEDM facilitates additional SRAF and dummy insertion by up to 10.8% and 18.6%, respectively.

- In terms of impact on other design metrics, the increases of data size, OPC runtime and maximum delay overheads of Corr are within 3% and 4%, respectively. In addition, maximum delay overhead of 6% is within noise of the

P&R tools [58]. The runtime of Corr placement perturbation is negligible ($\sim$ 5 minutes) compared to the runtime of OPC ($\sim$ 2.5 hours).

We are currently engaged in further experimental validation and research. Our ongoing research is in the following directions.

- *Restricted design rules (RDRs).* It may be possible to derive forbidden pitches from a set of restricted design rules which allow only few pitches in the layout. With increasing adoption of RDRs, "legalization" of layouts with respect to these rules become an important task where a Corr-like methodology can be useful. Part of our ongoing work analyzes "correct-by-construction" standard cell layouts which are always EtchCorrect in any placement scenario. We intend to compare such an approach with EtchCorr placement perturbation in terms of design as well as manufacturability metrics.

- *Extension to other layers.* Placement affects shapes on diffusion, contact, metal1 and metal2 layers in addition to the polysilicon layer. The Corr cost function can be extended to include forbidden pitches from these other layers. In future technology generations, process window for local metal layers is becoming a big concern and again since placement determines most of the local metal layout, a Corr-like technique can help.

- *Preferential treatment of devices.* Certain devices and cells may be able to tolerate more process variation than others in the design. For instance, narrow devices typically have a smaller process window. We are investigating techniques to bias the AFCorr and EtchCorr solution in favor of such devices to reduce timing and power impact and increase overall parametric yield.

II.E Acknowledgments

This chapter is in part a reprint of:

- P. Gupta, A. B. Kahng and C.-H. Park, “Detailed Placement for Enhanced Control of Resist and Etch CDs”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(12), 2007, pp. 2144 - 2157.

- P. Gupta, A. B. Kahng and C.-H. Park, “Detailed Placement for Improved Depth of Focus and CD Control”, *Proc. Asia and South Pacific Design Automation*, 2005, pp. 343 - 348.

- P. Gupta, A. B. Kahng and C.-H. Park, “Enhanced Resist and Etch CD Control by Design Perturbation”, *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2005, pp. 59923P-1 – 59923P-11.

I would like to thank my coauthors Prof. Puneet Gupta and Prof. Andrew B. Kahng.

III

Auxiliary Pattern-Based OPC for Better Printability, Timing and Leakage Control

III.A Introduction

OPC is a key RET that enables fabrication of integrated circuit (IC) features using sub-wavelength optical lithography. OPC modifies the shapes of IC layout features to enable their printability in silicon. In sub-180nm technology nodes, OPC is performed by iterative modification of layout feature edges. The iterative correction is performed until the resulting simulated image matches the target layout. The correction process itself can be driven by simulation using model of lithography and wafer processing steps during fabrication. Specifically, the models used for OPC describe the relationship between pattern information and aerial image, and resist and etch process parameters [38]. However, this approach, which we refer to as model-based OPC (MBOPC), is computationally expensive because of its iterative nature. Since MBOPC relies on simulation, its runtime has grown unacceptably with each successive technology generation, and it has emerged as a major bottleneck for turnaround time (TAT) of IC data preparation and manufacturing.

To address the OPC runtime issue, a cell-based OPC (COPC) approach has been proposed. See, e.g., Gupta et al. and Wang et al. [47, 122]. The COPC approach runs OPC once per each cell definition (i.e., per "cell *master*") rather than once per placement or unique instantiation of each cell (i.e., per "cell *instance*"). In other words, in the COPC approach, master cell layouts in the standard cell library are corrected before the placement step, and then placement and routing steps of IC design flow are completed with the corrected master cells. Since COPC is performed *once* for all cell masters in the library, it achieves significant OPC runtime reduction over MBOPC, which is performed at the full-chip layout level for every design that uses the cells. Unfortunately, optical proximity effects (OPE) in lithography cause interaction between layout pattern geometries. Since the neighboring environment of a cell in a full-chip layout is completely different from the environment of an isolated cell, the COPC solution can be incorrect when instantiated in a full-chip layout. As a result, there can be a significant discrepancy in printed feature critical dimension (CD) between COPC and MBOPC solutions.

In this work, we devise a novel *auxiliary pattern* (AP) technique which *shields* poly patterns near the cell outline from the proximity effect of neighboring cells. Consequently, COPC with AP achieves the same printability as MBOPC, but without any runtime overhead. APs inserted at the cell boundary reduce the difference between OPC impact of a cell in an isolated and a layout context. This effectively allows the *substitution* of an OPC'ed cell with APs directly in the layout. Auxiliary patterns are vertical (V-AP) and/or horizontal (H-AP) non-functional (dummy) poly lines [60]. *V-AP features are located within the same cell row and print on the wafer. H-AP features are located in the overlap region between cell rows; their width is comparable to that of sub-resolution assist features (SRAFs) and they do not print on the wafer.*

Optimization of RET by elongating design features and adding trim to SRAF features was recently proposed by Wallace et al. [121]. In this technique, line-ends facing a gap are elongated and SRAFs are added between them. To reduce the gap between SRAF and the features, the SRAF patterns are trimmed and included on the photomask. This increases the contrast in the line-end, thereby improving line-end shortening. This technique does not have any layout area impact, but it does not improve SRAF continuity at the boundary between different cells. Garg et al. [36] recently proposed a technique for insertion of dummy poly lines in empty spaces between poly gates within cell layouts. The dummy poly features are added as extensions to existing poly lines. The insertion of dummy poly improves the regularity of poly and enables tuning of the OPC recipe for improved process window. However, this technique has a $5-11\%$ cell area impact, which can translate to a design level area increase.

In contrast to the SRAF-based and dummy poly based approaches presented in recent literature, our approach seeks to minimize the difference between cell-based OPC and model-based OPC solutions by inserting dummy poly lines (auxiliary patterns, or APs) on all sides of a cell instance. To facilitate insertion of AP for some cell instances in the design placement (i.e., layout), it is helpful to perturb cell locations for some types of AP as detailed in Section III.C.3. *Indeed, to maximize the total amount of AP insertion in all cells in the design, we perturb detailed placement of standard cells using a a dynamic programming (DP)-based approach.* This allows opportunistic instantiation of AP around cell instances, depending on availability of free space in the layout. Note that placement perturbation does not increase the design area; it merely re-adjusts cell locations amidst the whitespace to allow AP insertion. We achieve 100% AP insertability in placements with row utilization less than 70%. In designs with row utilizations of 80% and 90%, the insertability of AP decreases to 98% and 80% respectively (due to lack of whitespace for AP insertion in cells). The movement of cells in the detailed placement may result in potential design level timing impact. To minimize the impact of dynamic programming-based approach on design timing, we perform timing-aware modification of cell placement. In our approach, we do not perturb the locations of timing-critical cells nor, consequently, the routes connected to them.

Apart from runtime improvement, AP-based OPC can be used to enhance the accuracy of post-litho timing and leakage analysis. Lithography simulation-based design analysis, optimization and signoff is becoming a necessity in the sub-100nm technology nodes [127]. However, performing chip-level lithography simulation is computationally expensive. Furthermore, two instances of the same standard cell will print differently based on their respective placement neighborhoods. This necessitates the creation of multiple variants for each cell in order to perform post-litho timing and leakage analysis. Ideally, it is preferable to use a *single* aerial image of every standard cell for post-litho analysis at a specific process condition. AP-based OPC allows this by shielding cells from their neighbors. Cao et al. [19] recently proposed a methodology for standard cell characterization considering litho-induced systematic variations. The objective of their work is to enable efficient post-litho analysis by running litho-aware characterization. To minimize the difference between isolated and placement context of a standard cell, vertical dummy poly patterns are inserted at the cell boundary. Our approach differs from that of Cao et al. in two main aspects: (1) we perform dummy poly insertion on *all sides of a cell* to shield OPE, and (2) we perform opportunistic, timing-aware insertion of AP at the *full-chip level by perturbing detailed placement.* In other words, we use detailed placement to maximize the insertion of AP in cell instantiations.

In our approach, vertical (V)-APs are designed to print on the wafer in order to shield the proximity effectively. V-AP width and spacing can be adjusted depending on the extent of OPE at the placement level. We also insert non-printing H-APs to shield OPE between cell rows. Placement of H-APs reduces line-end pullback, resulting in improvement of the curvature of poly litho contours. The curvature of poly around a line-end extends deep into the device region (i.e., poly over diffusion) as illustrated in Figure III.1. A decrease in the extent of line-end pullback improves performance and leakage variability.

The primary objective of our work is to reduce OPC turnaround time without any impact on timing and leakage of the design. Our main contributions are as follows.

- We propose a novel approach for application of COPC to designs, based on the insertion of APs. APs minimize CD difference between COPC and

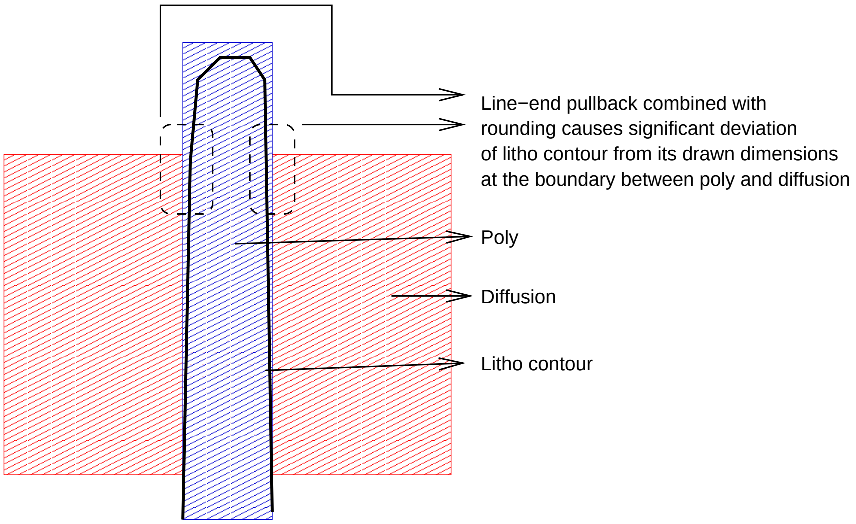

Line−end pullback combined with rounding causes significant deviation of litho contour from its drawn dimensions at the boundary between poly and diffusion
Poly
Diffusion
Litho contour

Figure III.1: Line-end pull back combined with rounding can impact device corner linewidth significantly.

MBOPC. Consequently, AP-enabled COPC achieves significant reduction in runtime compared to MBOPC. We demonstrate that COPC with V-AP and a combination of V- and H-AP can achieve better edge placement error (EPE) than conventional COPC.

- AP insertion might not be feasible on all instantiations of a standard cell in the design. To enable AP insertion in all cell instantiations in the layout with no area penalty, *we propose an opportunistic, DP-based methodology for perturbation of detailed placement to allow AP insertion.* The perturbation of placement maximizes the opportunity for AP insertion. However, detailed placement changes can potentially lead to change in design timing. We minimize the timing impact by introducing timing-awareness in our DP-based perturbation approach. All cell instances on critical paths are marked as fixed and are not moved during placement perturbations. This ensures that all routes connected to these critical instances (and their cell delays) do not change during subsequent engineering change order (ECO) routing steps[1].

[1] ECO steps are executed by the place-and-route tool to perform minor modifications to design

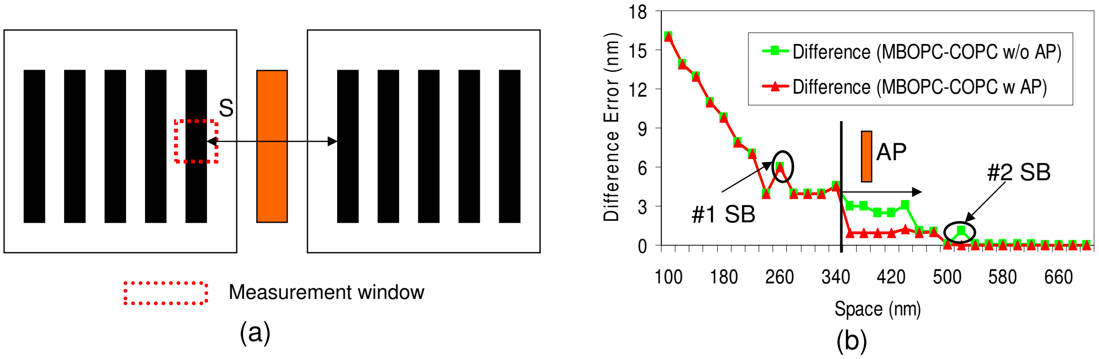


Figure III.2: CD impact of AP on linewidth: maximum CD differences between COPC and MBOPC are 3nm without AP and 1nm with AP. The width of vertical-AP is as large as the minimum linewidth of a feature on the poly layer.

- Using a litho-aware characterization methodology, we demonstrate an average improvement of 65% and 42%, respectively, in leakage and timing variability of AP-based OPC. At the cell layout level, we also show that timing and leakage behavior of AP-based OPC is comparable to that of MBOPC. Since full-chip analysis of post-litho timing and leakage power is not feasible, we compare EPE of all poly features between AP-based OPC and MBOPC. We show that AP-based OPC achieves EPE comparable to that of MBOPC at the full-chip level.

This chapter is organized as follows. We evaluate CD impact of AP in terms of linewidth, line-end and contact poly in Section III.B. In Section III.C, we discuss AP generation, printability impact and a placement perturbation method for improving feasibility of AP insertion. In Section III.D, we discuss details of the litho-aware timing and leakage characterization flow. We use the flow to demonstrate improvement in timing and leakage variability of AP-based OPC over COPC. We discuss our experimental setup and results in Section III.E. In Section III.F, we summarize our contributions.

layout.

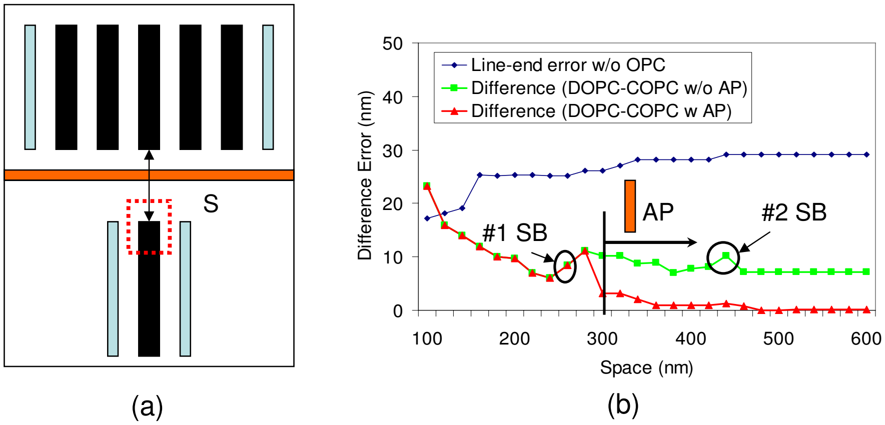


Figure III.3: CD impact of AP on line-end: maximum CD differences between COPC and MBOPC are 10nm without AP and 3nm with AP. The width of horizontal-AP is as small as that of a sub-resolution assist feature (SRAF).

III.B CD Impact of Auxiliary Pattern

The key role of the auxiliary pattern technique is to shield poly patterns near the cell outline from proximity effects of neighboring cells. We devise three test structures to evaluate CD impact of AP in terms of linewidth, line-end and contact poly[2]. Each test structure has two test cells which consist of linewidth of 0.1μm, pitch of 0.3μm and line length of 2.0μm. For simulation of CD impact, vector aerial image simulation is performed with wavelength $\lambda = 193$nm and NA $= 0.7$ for 90nm. Annular illumination with $\sigma = 0.85/0.57$ is used. For SB (Scattering Bar) insertion rules, SB width $= 0.04\mu$m, SB-to-poly spacing $= 0.12\mu$m and SB-to-SB spacing $= 0.12\mu$m are used.

Figure III.2(a) shows a test pattern structure to evaluate the CD impact of AP on linewidth. We use AP width of 0.1μm, AP-to-poly space of 0.13μm and AP-to-AP space of 0.14μm. AP can be inserted as long as the space between border poly is greater than 0.36μm. This spacing is determined by the minimum design rule. Figure III.2(b) shows the CD impact of AP on linewidth. In this plot,

[2] Contact poly defines the overlapped area of poly and contact which may also be called “contact coverage”.

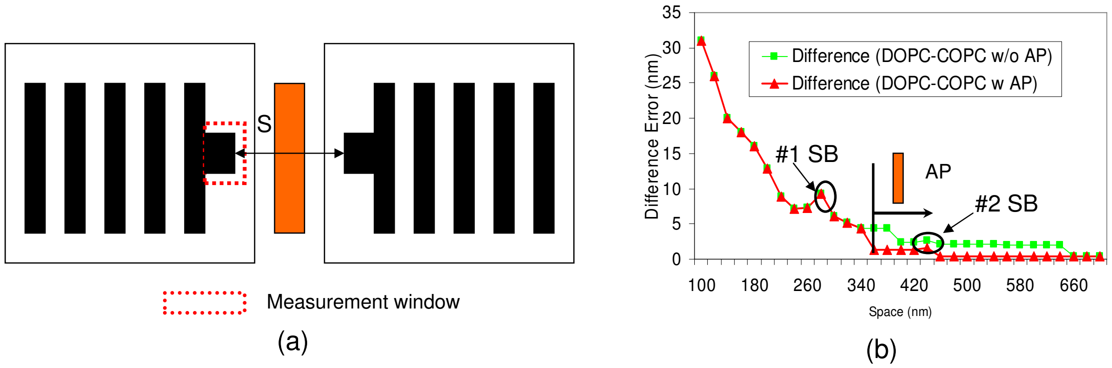


Figure III.4: CD impact of AP on contact poly: maximum CD differences between COPC and MBOPC are 5nm without AP and 1nm with AP.

the x-axis indicates the space between border poly of two adjacent cells and the y-axis indicates CD difference between MBOPC and COPC, measured in terms of CD. The maximum CD difference between MBOPC and COPC without AP is 3nm, while the maximum difference with AP is only 1nm.

The proximity shield effect of AP with respect to line-end shortening is shown in Figure III.3(b). We use horizontal AP width of 0.04μm for proximity shielding. The minimum space between line-end poly for insertion of APs is 0.3μm. The maximum CD difference between MBOPC and COPC without AP is 10nm, while the maximum difference with AP is about 3nm. The CD difference thus is reduced by up to 75% with AP insertion. We also evaluate the effectiveness of AP with respect to contact poly which is the closest geometry to neighboring cells. The minimum space between contact poly for insertion of AP is 0.36μm, as shown in Figure III.4(b). The maximum CD difference between MBOPC and COPC without AP is 5nm, while the maximum difference with AP is about 1.5nm. Consequently, COPC with AP achieves the same printability as MBOPC with respect to line patterning issues.

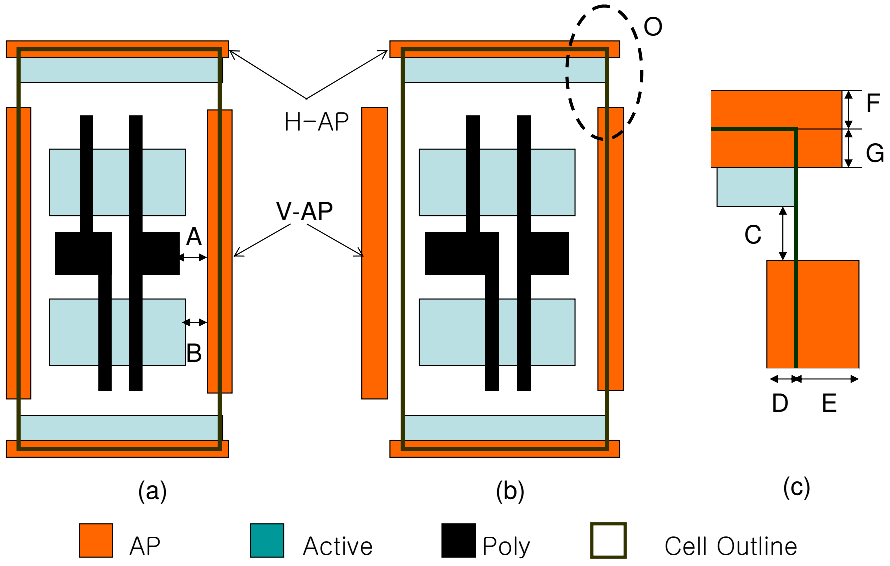


Figure III.5: Examples of standard cell layouts with APs: (a) Type-1 V-AP, (b) Type-2 V-AP and (c) an enlargement of the region O of (b).

III.C AP Methodology

In this section, we discuss details of AP generation, placement perturbation, and a modified design flow to enable AP-based OPC.

III.C.1 AP Generation

Auxiliary patterns overcome the deficiencies of the COPC approach for standard cell layouts. AP features consist of vertical (V-AP) and/or horizontal (H-AP) dummy poly as shown in Figure III.5(a) and Figure III.5(b). V-AP features are located within the same cell row as the standard cell, while H-AP features are located in the overlap region between adjacent cell rows. Devices in the layout are typically laid out vertically (assuming horizontal cell rows). Since the impact of lithography on gate CD is more interesting from a designer's perspective, patterns laid out vertically at cell boundaries within the same cell row should be shielded from proximity effects for maximum value and accuracy of cell-based OPC. Thus, the width of V-AP is as large as the minimum linewidth of a feature on the poly

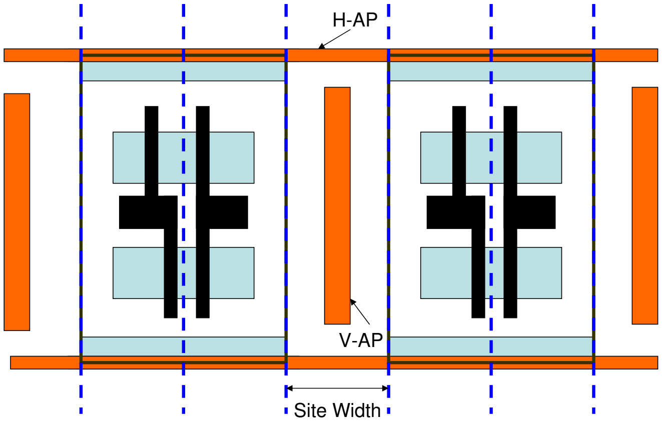


Figure III.6: An example of a standard cell layout with Type-3 V-AP.

layer. On the other hand, the width of horizontal-AP is as small as the width of a sub-resolution assist feature (SRAF). H-AP differs from the SRAF technique in that the location of SRAFs depends on the distance between poly lines, while the AP is located exactly at the cell boundary. In general, there is an active layer at the boundary between different cell rows, and hence the H-AP must not be allowed to print on the wafer. There are three types of V-AP according to the location of insertion.

Type-1 V-AP. Figure III.5(a) illustrates a Type-1 V-AP located at the center of (i.e., centered about) the cell outline, such that the left width (D in Figure III.5(c)) is the same as the right width of cell outline to right edge of V-AP (E in Figure III.5(c)). Spaces A and B respectively define the space between border poly and AP, and the space between active-layer geometry and V-AP. Rule A typically means the minimum design rule of poly-to-poly space. However, to insert at least one SRAF between border poly and AP, rule A can be the poly-to-poly spacing for inserting one SRAF. Since A and B in a typical standard cell are smaller than the required minimum spacing, it is desirable for the pattern geometries of each standard cell to be modifiable to permit the instantiation of cells with a Type-1 V-AP.

Type-2 V-AP. Type-2 V-AP locations satisfy both A and B of minimum design rules as shown in Figure III.5(b). Width D is different from width E. The Type-2

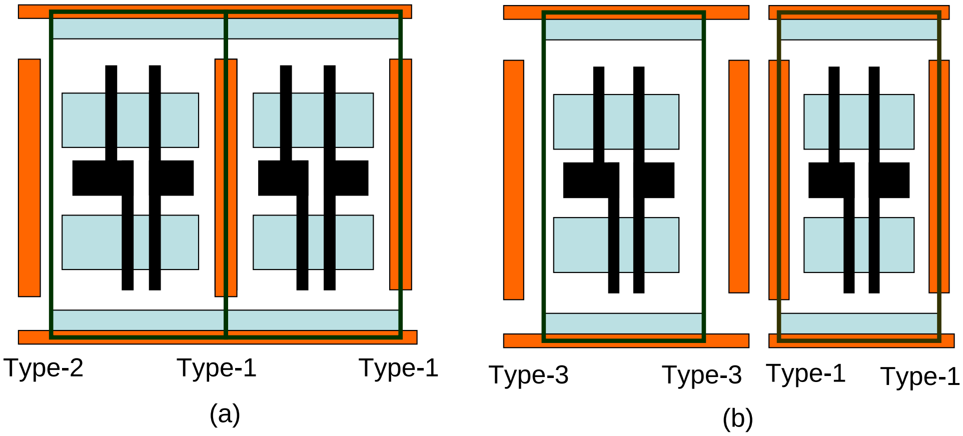


Figure III.7: Standard cell layouts constructed by combinations of the three types of APs: (a) a two-cell layout with Type-1 and Type-2 V-APs and (b) a two-cell layout with a combination of Type-1 and Type-3 V-APs.

V-AP can also be placed outside the cell outline. In Figure III.5(c), which is an enlargement of the region O of Figure III.5(b), C is the space between V-AP and the active layer, and is the same as the minimum space between the poly line-end and the active layer. The width from cell outline to the bottom edge, G, of the H-AP is the same as the width between cell outline and the top edge, F, of AP.

Type-3 V-AP. Figure III.6 illustrates a Type-3 V-AP that is placed at the center of the placement site. Since placing the Type-3 V-AP at the center of the site achieves enough space between poly and AP, the Type-3 V-AP can maintain minimum space rules such as poly-to-poly and poly-to-active spacing while simultaneously minimizing the area penalty.

Various auxiliary patterns can be constructed by combinations of the above three types of APs. Figure III.7 shows two examples: (a) a two-cell placement with a combination of cells with Type-1 and Type-2 APs; and (b) a two-cell placement with a combination of cells with Type-1 and Type-3 V-APs. Thus, in the application of the AP technique, all combinations of all possible types of AP are feasible and can be considered. In addition, Figure III.7(a) and Figure III.7(b) show APs completely overlapped or having certain required spacing to each other, respectively.

III.C.2 Area Penalty with AP

In this section, we discuss the area impact of AP insertion and its trend with technology scaling (i.e., design rule shrinkage). Standard cells with AP can increase cell area in the layout. For Type-1 V-AP, APs are located at the center of the cell outline and hence, the area penalty is equal to the width of AP. In the case of Type-2 V-AP, the area penalty in a cell depends on the spacing between border poly and cell outline, and the spacing between the border active layer geometry and the cell outline. In this case, the penalty is equal to 2× the sum of the AP width and the spacing to satisfy both A and B of minimum design rules. The layout of Type-3 V-AP depends on the width of the placement site. The penalty with Type-3 V-AP is the sum of the placement site width and the AP width.

The proximity shield effect of AP is affected by the shrinkage of CMOS design rules. The decrease of feature pitch due to technology scaling affects the optical proximity between layout features. This has implications for OPC and consequently AP insertion. An increase in the number of features within a fixed optical interaction region results in an increase in the proximity effects between them. This may necessitate insertion of increased number of AP at cell boundaries to shield from proximity of neighbors. However, the optical interaction radius scales with technology. The optical proximity range (OPR) depends on the optical wavelength, the numerical aperture (NA), and the coherence of the illumination source. To pattern features with smaller dimensions, NA is increased every technology node. The NA of lithography equipment used in the 65nm node is higher than that of the 90nm node (65nm NA = 0.9-1.2; 90nm NA = 0.7-0.85). The OPR, which determines the number of neighbor causing CD variation of border poly, decreases with higher NA [124]. For example, OPR decreases 21% as NA increases from 0.75 to 0.95, for a given set of illumination settings. Effectively, the scaling in OPR is somewhat slower than design rule scaling[3]. AP-to-border poly spacing thus needs to be increased compared to that of the 90nm node. On the

[3] We assume that design rule scaling from 90nm to 65nm is in the range 25% – 30%. In addition, design rule scaling to 45nm is supposed to be 50%.

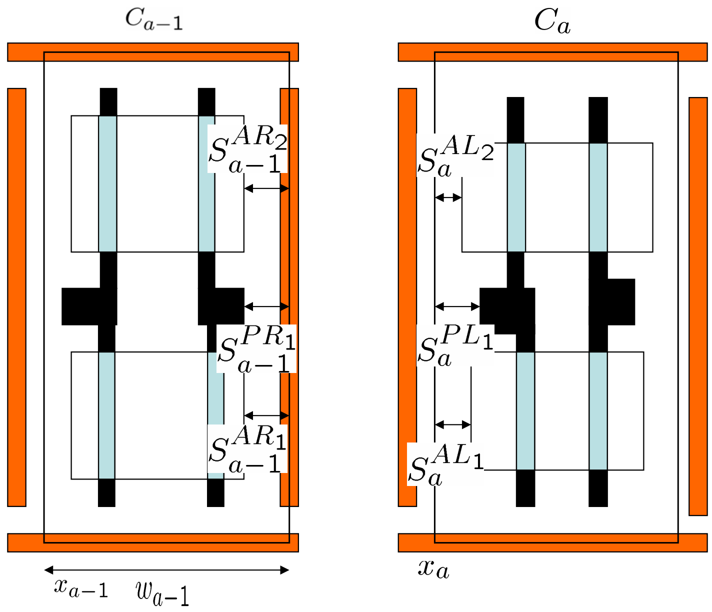


Figure III.8: An example of algorithm for post-placement optimization.

other hand, OPR scaling from 90nm to 45nm is only 37%, which is much slower than design rule scaling. However, most standard cell libraries at 45nm have a dummy poly between border poly and cell outline for reducing interaction from neighboring cells. As AP is placed between standard cells that have the dummy poly, AP may shield the proximity effect without an increase of AP-to-border poly spacing. We believe that the area penalty induced by AP is not significant even with design rule scaling. Furthermore, chip size does not change using our intelligent placement optimization, which we describe next.

III.C.3 Post-Placement Perturbation for Improved AP Insertion

The presence of an AP in close proximity to another AP corresponding to a different cell may violate minimum spacing rules for some placement configurations. This may inhibit insertion of AP for cells in such configurations. Hence, we propose to insert AP at the design level by perturbing the detailed placement. These perturbations do not increase chip size since they simply take advantage of (by repartitioning) existing whitespace of the standard cell placement. In this section, we describe a new detailed-placement perturbation algorithm using various types of AP. This approach extends the algorithm presented by Gupta et al. [43] to handle all three types of AP.

Define S_a^{AL} to be the space between the left outline of the cell and the active geometry, and S_{a-1}^{AR} to be the space between the right outline of the cell and active layer. Similarly, let S_a^{PL} be the space between the left outline of the cell and the poly, and S_{a-1}^{PR} be the space between the right outline of the cell and poly layer. S_a^L and S_{a-1}^R are defined as follows.

$$\begin{aligned} S_{a-1}^R &= \min\ \{(S_{a-1}^{AR_1}, ..., S_{a-1}^{AR_n}), (S_{a-1}^{PR_1}, ..., S_{a-1}^{PR_n})\} \\ S_a^L &= \min\ \{(S_a^{AL_1}, ..., S_a^{AL_n}), (S_a^{PL_1}, ..., S_a^{PL_n})\} \end{aligned} \qquad \text{(III.1)}$$

Assume a set $AS = AS_1, \ldots, AS_m$ of spacings which are "AP-correct", i.e., if the spacing of boundary shapes between cells belongs to the set AS, then the required number of APs can be inserted between cells. For example, AS_1 and AS_2 are the required spacings for one AP and two APs, respectively. Figure III.8 shows an example portion of the input for our post-placement optimization algorithm. Let W_a denote the width of the cell C_a and let x_a and x_a^i denote the (leftmost) placement coordinates of the original standard cell and the modified standard cell with Type-i AP, respectively. Let δ denote a placement perturbation by which the modified standard cell will have an AP-correct spacing. Then the **AP-correct placement perturbation problem** may be formulated as:

$$\text{Minimize} \sum \mid \delta_i \mid$$

$$\text{Subject to } \delta_a + x_a^i - \delta_{a-1} - x_{a-1}^i - W_{a-1} + S_a^L + S_{a-1}^R \in AS$$

Our objective is to minimize total placement perturbation from the original cell-location and area penalty. We solve for the perturbed placement locations of the cells using a dynamic programming recurrence. We solve this "continuous" version of the above problem with the following dynamic programming recurrence.

$$Cost(1,b) = \mid x_1^i - b \mid$$

$$Cost(a,b) = \lambda(a) \mid (x_a^i - b) \mid +$$

$$Min_{j=x_{a-1}^i-SRCH}^{x_{a-1}^i+SRCH}\{Cost(a-1,j) + APCost(a,b,a-1,j)\}$$

APCost(a,b,a-1,j) of Cell C_a

Input:

Origin x (left) coordinate and length of cell C_a = b

Origin x (left) coordinate and length of cell C_{a-1} = j

Width of cell $C_a = w_a$

Width of cell $C_{a-1} = w_{a-1}$

Output:

Value of $APCost$

Algorithm:

```
Case $a = 1$ : $APCost(1,b) = 0$
Case $a > 1$ Do
/* For three AP types for left and right outline,
   calculate weight according to boundary geometries. */
        $space$= $x_a^i$ - $x_{a-1}^i$ - $W_{a-1}$ + $S_a^L$ + $S_{a-1}^R$
         if ($space \neq AS$) $weight = \infty$
              else $weight = space$
       $APCost(a,b,a-1,j)$ += $weight$
```

Figure III.9: *APCost* calculation.

$Cost(1,b)$ is the cost of placing the first cell of each standard cell row at placement site number b. $Cost(a,b)$ is the cost of placing cell a at placement site number b. The cells and the placement sites are indexed from left to right in the standard cell row. We restrict the perturbation of any cell to $SRCH$ placement sites from its initial location for timing-driven placement. $APCost$ is the measure of total expected CD degradation of the vertically oriented poly geometries closest to the cell boundary at the worst defocus value for the cell. *APCost* depends on the space between border polys. If the space is smaller than the required spacing for one AP, *APCost* is infinite since it causes overlap between APs. The method of computing *APCost* is shown in Figure III.9.

The modified cell placement corresponding to a feasible set of AP insertions can then be incorporated into a modified standard cell GDSII. Cell definition in DEF (Design Exchange Format) is changed according to the standard cell GDSII used during post-placement optimization. For example, NAND2X2_T1_T3 is a new cell definition in DEF with Type-1 V-AP at left outline and Type-3 V-AP at right outline of NAND2X2. Thus, the proposed placement optimization can modify the standard cell placement and is consistent with the set of available APs for each cell.

III.C.4 Modified Design Flow

Figure III.10 shows the flow sequence for AP generation and placement perturbation of instances. A standard cell layout is input to an AP generation step, and then to an SRAF insertion step. The resulting layout is input to OPC insertion step, which results in a set of OPC'ed standard cell layouts corresponding to the master cells. These OPC'ed cell layouts will be instantiated within the final layout according to the results of post-placement optimization. The AP-correct placement takes the OPC'ed standard cell layout as an input. A final cell-based OPC layout is generated from the modified AP-correct placement and the OPC'ed standard cell layouts.

III.D Cell Characterization Considering Litho Effects

AP-based OPC achieves substantial reduction in edge placement error (EPE) over COPC at any given focus condition. To demonstrate the timing and

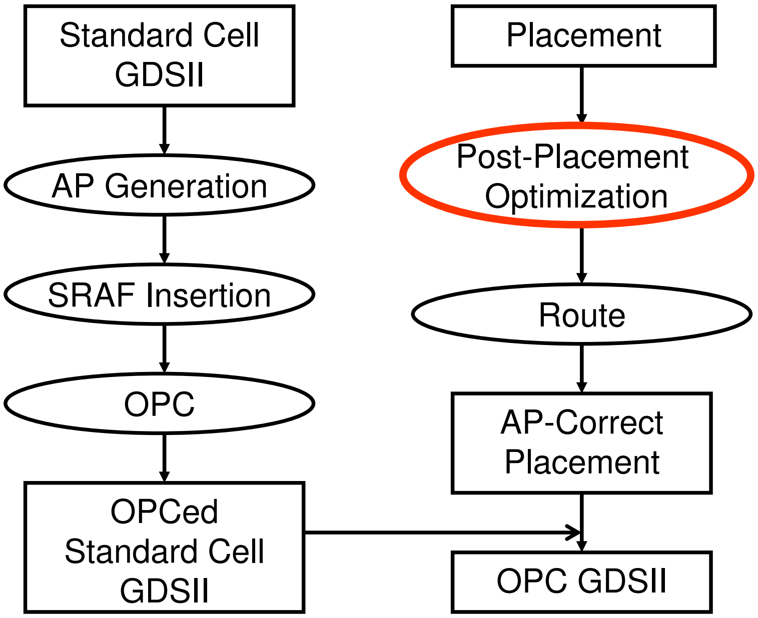


Figure III.10: Block diagram of a system for AP generation and placement perturbation of layout objects.

leakage impact of AP-based OPC and COPC, we perform lithography-aware cell characterization. In the rest of this section, we discuss details of this flow.

A significant fraction of across-chip linewidth variation (ACLV) is caused by linewidth change depending on poly line pitch, poly line shape (corners, jogs etc.) and their orientations. Printed poly shape varies as a function of focus, exposure dose and layout parameters within the process window. In addition to linewidth (i.e., gate CD), field poly length, gate width and contact enclosure may also change [48]. However, these do not affect electrical parameters (i.e., delay and leakage) significantly. Delay is partially determined by saturation current and decreases linearly with decrease in linewidth. Subthreshold leakage increases exponentially with decrease in linewidth. Since linewidth is the smallest dimension related to devices, its variation translates to significant performance and leakage variability. Consequently, we focus only on characterization of gate CD impact in our litho-aware analysis.

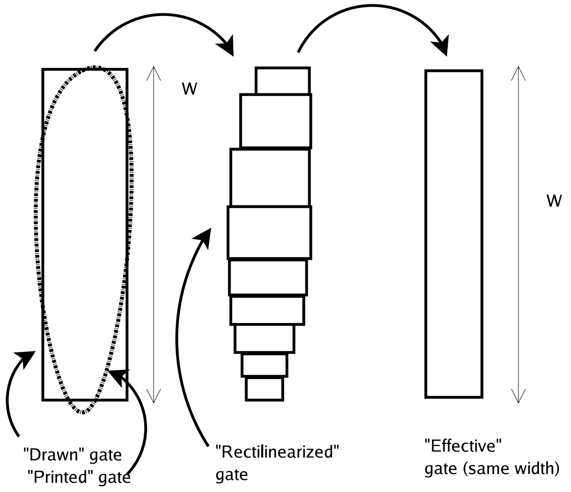


Figure III.11: Calculation of L_{avg} for timing and leakage from non-uniform geometry device.

III.D.1 Average Gate CD Computation

SPICE simulations can be performed to characterize timing and leakage profiles of a standard cell using printed gate CD. However, existing device models for SPICE can only handle rectangular transistors while printed devices have non-rectangular geometry. The post-litho timing analysis flow presented by Yang et al. [127] considers CD at the center of the device and uses it as a representative value for the entire device. However, this is not accurate, since I_{on} and I_{off} of a device depend on its CD profile. To account for the gate CD profile using existing device models, we compute the *average* gate length for each device. I_{on} and I_{off} have different sensitivities to the same gate CD profile. Hence, we compute L_{avg} differently for timing and leakage. To compute L_{avg} of non-uniform geometry devices, we use the method outlined by Heng et al. [52]. Their basic flow proposed in the chapter takes in a gate shape contour (from lithography simulation) and performs rectilinearization. In this step, the non-uniform geometry is divided into multiple small rectangles with different W and L as shown in Figure III.11. Separately, lookup tables for device I_{on} and I_{off} are created for different W and L combinations from SPICE simulations. I_{on} and I_{off} of the non-uniform geometry

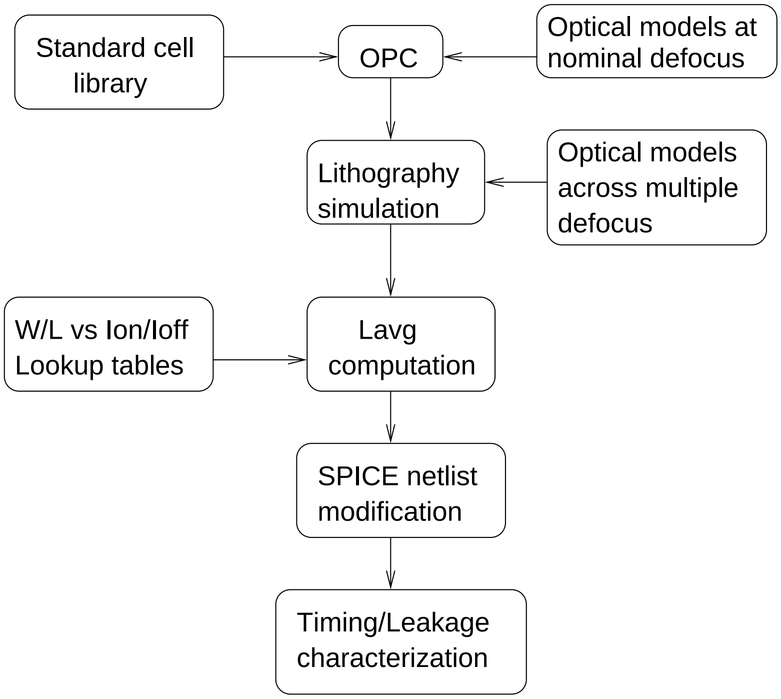

Standard cell library
OPC
Optical models at nominal defocus
Lithography simulation
Optical models across multiple defocus
W/L vs Ion/Ioff Lookup tables
Lavg computation
SPICE netlist modification
Timing/Leakage characterization

Figure III.12: Litho-aware standard cell characterization flow.

device are computed by summing up the corresponding values for each rectilinear (small) device from the lookup tables. L_{avg} of the actual printed gate contour is the gate length of a rectangle of the same gate width that yields the same on- or off- current (done by reverse lookup in the I_{on}/I_{off} table). This methodology yields $L_{avg,timing}$ and $L_{avg,leakage}$ corresponding to timing and leakage, respectively, and accounts for the non-uniformity in gate CD along the width of the gate.

III.D.2 Litho-Aware Cell Characterization

The values of L_{avg} computed for each device in a standard cell are now used for accurate post-litho timing and leakage characterization. Standard cell SPICE netlists specify device names and their width and length (W/L) only. Positional information of devices is absent in the SPICE netlist. To associate printed CD of devices to their names, we run Layout-Versus-Schematic (LVS) on standard cell layouts to obtain their locations. Using LVS information, we update SPICE netlists with L_{avg} gate lengths computed from rectilinearization of printed gate shapes. We create two versions of the SPICE netlist: one for timing characterization (updated with $L_{avg,timing}$) and the other for leakage characterization (updated with $L_{avg,leakage}$). The complete litho variation-aware cell characterization flow is summarized in Figure III.12.

III.E Experimental Results

In this section, we describe our experimental setup to (1) compare the printabilities (in terms of EPE count) of MBOPC, COPC and AP-based OPC; (2) demonstrate improvement in timing and leakage variability of AP-based OPC over COPC; and (3) demonstrate comparable timing and leakage variabilities of AP-based OPC and MBOPC.

III.E.1 Experimental Setup

To compare MBOPC and AP-based OPC, we first prepare two designs (AES and ALU) from `opencores.org` for application of OPC. The circuits are synthesized using Synopsys *Design Compiler* (v 2003.06-SP1) [12] with tight timing constraints and a set of 50 most frequently used cells in the Artisan TSMC 90nm library. AES and ALU are synthesized to 11553 and 8572 cells respectively. The synthesized netlists are then placed with row utilization ranging from 50% to 90%. On the lithography side, Mentor Graphics *Calibre* (v 9.3 5.11) [9] is used for model-based OPC, assist-feature insertion and optical rule checking (ORC). Vector aerial image simulation is performed with wavelength λ = 193nm and NA = 0.7 for 90nm. Annular illumination with $\sigma = 0.85/0.57$ is used. Our OPC setup conforms to those used in industry-strength recipes. To evaluate EPE for each type of OPC, we first perform MBOPC on the entire design using the setup described above. For AP-based OPC, we implement the flow described in Section III.C.4.

To compare the timing and leakage variabilities of different OPC types at the cell-level, we compare isolated and layout contexts of standard cells. The isolated context refers to the stand-alone version of the cell and the layout context refers to the standard cell in a placement context. The layout context is constructed by placing copies of a given standard cell on all its four sides, to simulate OPE inside the center cell. We then perform: (1) cell-based OPC without AP (denoted as COPC(WO)), (2) AP-based OPC with vertical-only AP (denoted as COPC(V)), (3) AP-based OPC with horizontal and vertical AP (denoted as COPC(HV)), and (4) model-based OPC (MBOPC) on both versions of all chosen standard cells. We then perform lithography simulation at nominal and at 100nm defocus. We then execute the litho-aware characterization flow described in Section III.D.

At the design level, comparison of timing and leakage variabilities from different types of OPC is not straightforward. To evaluate the necessity for performing chip-level post-lithography timing and leakage power analysis, we first evaluate gate poly EPE. AP-based OPC can be used as replacement for MBOPC without incurring performance degradation (due to CD variation), while achieving significant savings in OPC runtime. The litho quality achieved by MBOPC is an upper bound on that achieved by AP-based OPC measured in terms of EPE. This OPC runtime versus CD tradeoff can be utilized in a design-aware fashion to minimize design performance and power impact while improving OPC runtime. For instance, MBOPC can be applied to all timing-critical features and AP-based OPC can be applied to all nontiming-critical features. To explore this runtime versus performance impact tradeoff, we perform MBOPC and AP-based OPC on different fraction of cells. The choice of cell instances for performing MBOPC is determined by their timing criticality. The total OPC runtime is the sum of MBOPC runtime on all timing-critical cell instances and the runtime of MBOPC for individual masters that are instantiated in the design.

To run MBOPC on timing-critical cells in the design, we first perform timing analysis on the design to identify cell instances on paths with slacks within 10%, 20%, 30% and 40% of clock cycle time. We then create a cover layer on all timing-critical cells in the design layout and run MBOPC only on the identified cells. OPC on the entire layout is completed by substituting AP OPC'ed cells into the layout. The flow discussed above creates a "timing criticality-aware" OPC

Table III.1: AP insertion error for five different row utilizations across different post-placement optimizations. “Typical” corresponds to the original placement. “T3” and “All” represent AP-correct placements with Type-3 AP and all types of AP, respectively.

Utilization (%)	90			80			70			60			50		
Flow	Typical	T3	All	Typical	T3	All	Typical	T3	All	Typical	T3	All	Typical	T3	All
AES	9115	2512	1925	3199	68	55	3166	0	0	1873	0	0	1589	0	0
ALU	5613	3099	2542	2085	219	179	1670	0	0	727	0	0	813	0	0

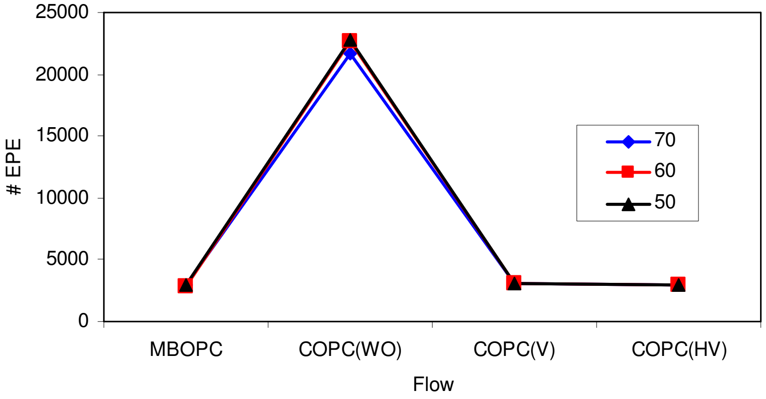


Figure III.13: EPE count of gate with various OPC methods for each of three different utilizations. COPC(WO) is cell-based OPC without AP. COPC(V) is cell-based OPC with only V-AP. COPC(HV) is cell-based OPC with H- and V-APs.

solution of the layout. We refer to this solution as *hybrid OPC*. To compare this with a pure AP-based OPC solution, we substitute AP OPC'ed masters for all cells in the design to create an AP OPC'ed GDS. We then perform ORC to evaluate gate EPE.

III.E.2 Experimental Results

We evaluate the quality of AP-based OPC by comparing it with MBOPC. The criteria chosen for evaluation are (1) AP insertion error, (2) OPC metrics (EPE, OPC runtime, filesize), and (3) leakage and timing spread. AP insertion error is defined as the number of vertical edges of standard cells in which AP cannot be inserted even after post-placement optimization. Table III.1 shows the AP insertion error for five different utilizations and for three different placement contexts: (1) typical cell placement, (2) optimized cell placement with only Type-3 AP, and (3) optimized cell placement with all combinations of AP. For row utilizations that are $<$ 70%, post-placement optimizations can achieve 100% AP applicability without increasing chip size. Post-placement optimization with all combinations of AP can reduce AP insertion error over optimization with Type-3

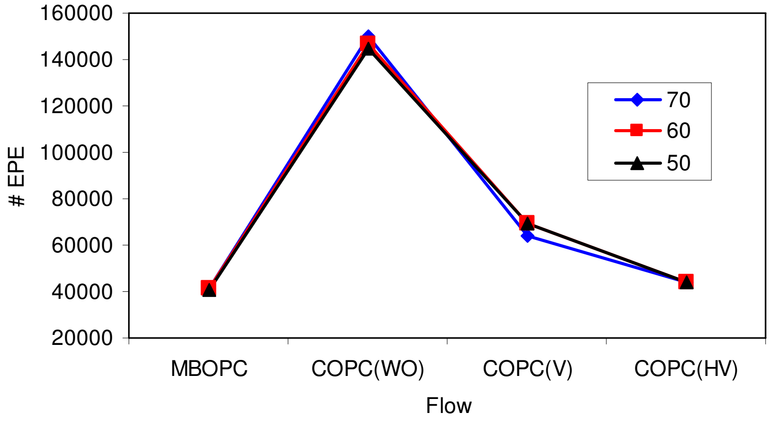


Figure III.14: EPE count of poly lines of AES design for three different row utilizations.

AP by an average of 20% for utilizations greater than 70%.

To evaluate the impact of AP-based OPC on printability, we perform ORC on gate and field poly and measure EPE count. For this study, we perform ORC to flag all layout edge fragments with error greater than 10% of drawn CD at the worst defocus condition. Figure III.13 shows the EPE count of gates of the ALU design with various OPC methods. EPE count of two AP-based OPC methods match that of MBOPC within 3%. Figure III.14 shows the EPE count of poly lines of AES testcase. EPE count of OPC with only V-AP is 35% more than that of MBOPC. This is because of poly line-end shortening due to OPE between cell rows. However, EPE count of AP-based OPC with H- and V-APs match that of MBOPC within 6%. This also corresponds to an average improvement of 68% over COPC without AP (COPC(WO)). We compare the average CD difference of devices near cell outline for three cases of COPC with MBOPC. The average CD differences for (a) COPC with no placement optimization, (b) COPC with placement optimization and (c) COPC with placement optimization and AP over MBOPC are 7.2nm, 2.5nm and 1.2nm respectively. Figure III.15 shows the actual layouts with various OPC methods.

OPC runtimes for MBOPC, COPC(WO), COPC(V) and COPC(HV) are summarized in Table III.2. OPC runtime denotes the runtime of assist-feature

Table III.2: Printability (in terms of EPE), OPC/ORC runtime and post-OPC GDSII file size for different types of OPC. COPC(HV) improves EPE over COPC(WO) by an average of 68%. Poly EPE count of COPC(HV) matches that of MBOPC within 6%.

Design	Utilization (%)	Flow	# EPE (Gate)	# EPE (Poly)	GDSII size (MB)	OPC Runtime (sec)	ORC Runtime (sec)
AES	70	MBOPC	6972	41365	3826	7932	943
		COPC(WO)	37682	150300	741	144	543
		COPC(V)	7528	63942	776	168	598
		COPC(HV)	7240	44110	789	192	621
	60	MBOPC	6988	41043	3823	7943	940
		COPC(WO)	36649	146574	743	144	547
		COPC(V)	7522	69198	780	168	599
		COPC(HV)	7290	44023	799	192	641
	50	MBOPC	6974	40636	3811	7943	938
		COPC(WO)	36496	144382	740	144	547
		COPC(V)	7509	69198	786	168	602
		COPC(HV)	7217	44012	799	192	641
ALU	70	MBOPC	2895	30029	3213	4109	772
		COPC(WO)	21675	86926	721	120	364
		COPC(V)	3076	39988	745	136	394
		COPC(HV)	2947	31323	774	160	410
	60	MBOPC	2827	29751	3221	4109	777
		COPC(WO)	22711	92740	722	120	373
		COPC(V)	3092	39481	744	136	389
		COPC(HV)	2964	31101	776	160	410
	50	MBOPC	2949	29446	3222	4121	778
		COPC(WO)	22823	91946	703	120	376
		COPC(V)	3036	45012	742	136	399
		COPC(HV)	2981	31323	776	160	411

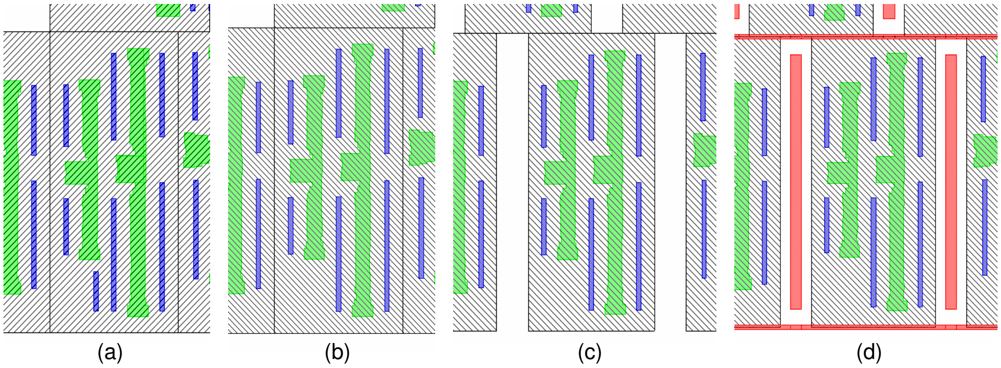


Figure III.15: Layouts with various OPC methods: (a) MBOPC, (b) COPC (WO), (c) COPC(V) and (d) COPC(HV). Red, blue and green colors represent AP, SBAR and OPC geometries, respectively.

insertion, MBOPC and AP insertion (in case of AP-based OPC). The AES and ALU designs use 48 and 40 standard cell definitions, respectively. From the table, we observe that COPC (WO, V and HV) improves the runtime by an order of magnitude versus MBOPC. The improvement will be more apparent as the design size increases. From the table, we can observe that COPC (WO, V and HV) runtimes are comparable between AES and ALU testcases. But we can clearly see the sharp rise in MBOPC runtime as the number of instances increases from 8572 (ALU) to 11553 (AES). COPC(HV) reduces runtime over MBOPC by 42X and by 25X for AES and ALU, respectively. We can also observe reduction of GDSII file size and ORC runtimes. COPC maintains the original cell hierarchy, thereby reducing GDSII file size and ORC runtime over MBOPC.

Table III.3 shows the percentage spread in leakage and timing of eight standard cells at nominal defocus. Leakage spread at any given focus condition is the percentage change in cell leakage power between the isolated and the layout context of the cell. Timing spread is computed as the percentage change in the rise delay of the cell output pin at a fixed load capacitance and slew condition[4]. From the table, we observe that COPC(HV) improves leakage variability over

[4] In our experiments, we measured delay values at a load capacitance of $6.5pF$ and a transition time of $140ps$.

Cell	% Leakage var			% Timing var		
	WO	HV	MB	WO	HV	MB
and3x1	8.17	1.99	2.96	4.09	0.59	0.67
invx2	9.67	2.55	2.32	2.48	0.42	1.34
mx2x1	0.35	1.63	2.86	4.31	1.24	2.85
nand2bx1	7.06	2.21	3.38	0.51	1.51	1.13
nand2x2	8.48	0.64	1.69	0.82	0.71	1.23
nor2x2	10.65	1.58	2.05	0.66	0.66	0.44
nor4x2	10.20	1.03	2.14	1.22	0.37	0.27
xor2x1	1.77	1.66	1.55	0.65	1.07	0.72

Table III.3: Comparison of leakage and timing spread of standard cells between WO (i.e., COPC(WO)), HV (i.e., COPC(HV)) and MB (i.e., MBOPC). COPC(HV) improves leakage variability over COPC(WO) in the range 1% − 92% and timing variability in the range 1% − 85%.

COPC(WO) by an average of 65%. COPC(HV) improves timing variability over COPC(WO) by an average of 42%. Another important trend apparent from the results is that the leakage and timing spread of COPC(HV) and MBOPC are comparable to within one percentage point for all of the cells.

Table III.4 shows the comparison between gate EPE count between hybrid OPC solution and a pure AP-based OPC solution for AES testcase with 70% row utilization. The number of timing-critical cells in the design, based on different timing slack criteria, is also shown. Gate EPE count is the number of edge fragments on border poly geometries that have greater than 3nm EPE at the best focus level. From the table, we can observe that hybrid OPC runtime increases in proportion to the number of cells for which MBOPC is applied. For AP-based OPC, the total gate EPE count and runtime are independent of the number of timing-critical cells and are 1479 and 326 seconds, respectively. From the gate EPE count trend, we can observe that MBOPC and COPC(HV) achieve similar EPE on the gate poly (and, consequently, similar CD control). Gate EPE count for

Timing slack (as a % of cycle time)	# cells within timing slack	Hybrid OPC runtime (s)	Gate EPE count (hybrid OPC)
10	5640	36930	1471
20	6369	42588	1472
30	9165	60837	1480
40	9221	61329	1480

Table III.4: Gate EPE count for hybrid OPC for different fractions of timing-critical cells in AES testcase implemented in TSMC 90nm technology. Hybrid OPC runtime is proportional to number of timing-critical cells.

hybrid OPC as well as COPC(HV) are within 0.1% of each other. This eliminates the need for design level post-litho timing and leakage power analysis.

III.F Conclusions

We have proposed a novel auxiliary pattern (AP) based cell OPC method that has the OPC TAT advantages of COPC and printability performance comparable to that of MBOPC. Using a timing-aware DP-based method, that perturbs detailed cell placements, we have demonstrated a method for opportunistic insertion of AP at the full-chip level to maximize the benefits of AP-based OPC. Our AP-based OPC approach has shown a factor of 42X reduction in OPC runtime compared to MBOPC. The runtime advantage will be substantially higher for larger designs. Printability analysis of AP-based OPC has shown that V-AP and V/H-AP can match gate EPE count of MBOPC within 3%. This is an improvement of 68%, on average, over cell-based OPC without APs. Our post-placement optimization method can achieve 100% AP applicability in designs with utilization less than 70%. For designs with utilization greater than 70%, we can achieve up to 80% AP applicability. Our proposed DP-based perturbation approach is timing-aware; it does not modify the placement (and consequently routing) of timing-critical cells in the design, thereby preserving timing. Using a litho-aware timing and leakage analysis flow, we have demonstrated 65% and 42% reductions in timing and leakage variabilities respectively, over cell-based OPC. Further, the spread in leakage and timing match those of MBOPC within 1%. This demonstrates that adoption of AP-based OPC does not degrade design performance and power. AP-based OPC can be adopted in an industrial flow with significant runtime savings without any performance degradation.

III.G Acknowledgments

This chapter is in part a reprint of:

- A. B. Kahng, S. Muddu and C-.H. Park, "Auxiliary Pattern-Based OPC for Better Printability, Timing and Leakage Control", *Journal of Microlithography, Microfabrication and Microsystems*, 2008, to appear.
- A. B. Kahng and C.-H. Park, "Auxiliary Pattern for Cell-Based OPC", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63494S-1 – 63494S-10.

I would like to thank my coauthors Dr. Swamy Muddu and Prof. Andrew B. Kahng.

IV

Lens Aberration Aware Placement for Timing Yield

IV.A Introduction

Aberrations [78] can be described as the departure from ideal imaging induced by an imperfect lens system, as shown in Figure IV.1. Aberrations cause optical path differences among the rays, resulting in wavefront deviation from a reference sphere at the exit pupil; this induces blur and distortion of images. Undesirable imaging artifacts from aberration are uncorrectable and, indeed, are sometimes exacerbated through use of resolution enhancement techniques (RETs) such as phase-shift mask and off-axis illumination [17]. The effects of lens aberrations on lithographic imaging [37, 117] include shifts in the image position, image asymmetry, reduction of the process window, and the appearance of undesirable imaging artifacts. *Zernike's coefficients* capture the deviation from ideal imaging and may be used during lithography simulation to predict the impact of lens aberration on critical dimension (CD) [78,103]. CD variation caused by lens aberration is relatively small compared to that caused by defocus and pattern proximity. However, most CD error caused by proximity can be corrected by RETs. Thus, lens aberration has turned out to be a major source of residual errors in across-field linewidth variation (AFLV) [34].

Recent studies of lens aberration control have focused on measurement systems [33,112] and pattern sensitivity of aberration [123], as well as lens mounting systems to compensate for the aberration [86]. However, despite these efforts, the impact of lens aberration on CD will be an ever-present barrier to manufacturing yield as minimum design rules are pushed ever closer to fundamental resolution limits. From the design perspective, variations in CD affect the delays, slews, input capacitances and leakage of a given logic cell. We also observe that the maximum difference in delays of all timing arcs in a cell (delay skew) increases significantly with lens aberration, as different MOS devices in the layout are affected differently by aberration.

Progler et al. [102] studied the impact of lens aberration on statistical timing behavior and observed that certain aberration coefficients are associated with large timing error. Orshansky et al. [92] found that spatial gate CD variation leads to a large variation in the raw speed of CMOS logic. Misleading timing results are obtained, which lead to slower and/or malfunctioning circuits because the simulation of a circuit's behavior has ignored the spatial CD information. The systematic variability of gate CD caused by lens aberration can be modeled in order to achieve better performance by way of accurate timing analysis at all stages of physical implementation [93, 94]. However, more accurate analysis of gate delay impact is required as the scaling of lithographic features makes the impacts of lens aberrations even more complex.

In this chapter, we use lens aberration-aware global placement for timing improvement. It is worth discussing why we use global placement, as opposed to OPC or detailed placement (for example), as the appropriate 'knob' for this compensation and optimization. First, lens aberration can slightly change according to lens heating and lens contamination. OPC is very sensitive to changes of lens aberration because it embodies direct perturbations of mask shape; on the other hand, placement does not directly change any mask shape, but only rearranges cell instances according to fast and slow regions in a lens field. Thus, a placement-based solution is less sensitive to variations of aberration parameters than an OPC-based solution. Second, lens aberration globally changes cell characteristics within a lens field. That is, cells at the range of a few microns may use the same Zernike's coefficients, while cells at the range of a millimeter must use different Zernike's coefficients. On the other hand, detailed placement is effective for compensation of micron-ranged variability (i.e., proximity effects of resist and photo processes). [44] proposed a detailed placement technique to avoid forbidden pitch between cells, which is an optimization on the length scale of $0.5 \sim 2$ μm. The detailed placement approach may also increase wirelength significantly, and thus has no clear advantage (in terms of convergent flow, etc.) over global placement. We believe that the use of a global placer to minimize total wirelength with respect to global bins can more efficiently handle the lens aberration problem, as compared to a detailed placement approach.

In the following, we first describe a novel aberration-aware static timing analysis flow that integrates (i) results of lithography simulation to measure CD across the lens field, (ii) SPICE simulation-based library performance characterization that captures variant CD combinations in library cell instances, and (iii) placement information. We also propose an aberration-aware timing-driven analytical placement framework that utilizes the aberration-aware timing analysis flow to minimize clock cycle time and avoid hold-time violations, without significantly increasing total wirelength. The placer is driven by models that capture the impact of lens position on timing arc delays in cells, and by weighted-wirelength models. Essentially, we preferentially place cells that are setup-time (resp. hold-time) critical at lens field locations where aberrations cause the cell delay to decrease (resp. increase).

The contributions of our work are as follows.

- Using industry OPC recipes, aberration parameters, and design testcases, we show that the variation in timing due to lens aberration can be significant. Over the cells in a 90nm foundry library, we observe cell delay (averaged over all timing arcs) to change by $2\% - 8\%$. The maximum difference in delays over all timing arcs of a cell (*delay skew*) increases significantly.

- We develop a novel aberration-aware timing analysis flow that affords more

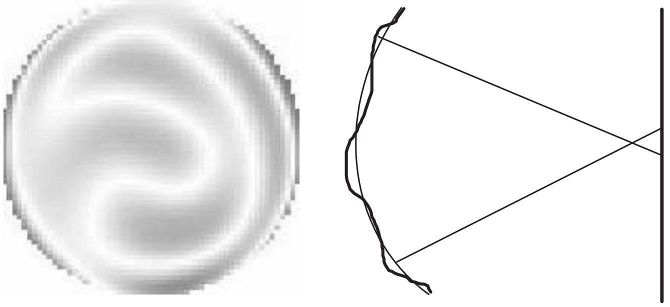


Figure IV.1: An imperfect lens system.

accurate timing analysis, taking into account the position of the chip in the lens field. It also considers the increase in delay skew caused by aberration.

- We propose a novel aberration-aware, timing-driven analytical placer that considers the impact of lens aberrations on timing to minimize clock period and avoid hold-time violations without significant total wirelength increase. Averaged over our two testcases, worst-case cycle time and total negative slack respectively reduce by $\sim$ 4.749% (116ps) and $\sim$ 7.535% at the cost of $\sim$ 1.341% increase in wirelength, with no hold-time violations – a very substantial performance improvement.

The remainder of this chapter is organized as follows. In Section IV.B, we describe lens aberration and study its impact on CD and gate delay. Section IV.C proposes a novel aberration-aware timing analysis and an accompanying flow. Section IV.D describes our aberration-aware analytical placement formulation and implementation details. Test designs, experimental conditions and experimental results are described in Section IV.E. We conclude in Section IV.F with directions for ongoing research.

IV.B Design Impact of Lens Aberration

In this section we briefly describe how lens aberration impacts CD and consequently circuit delay.

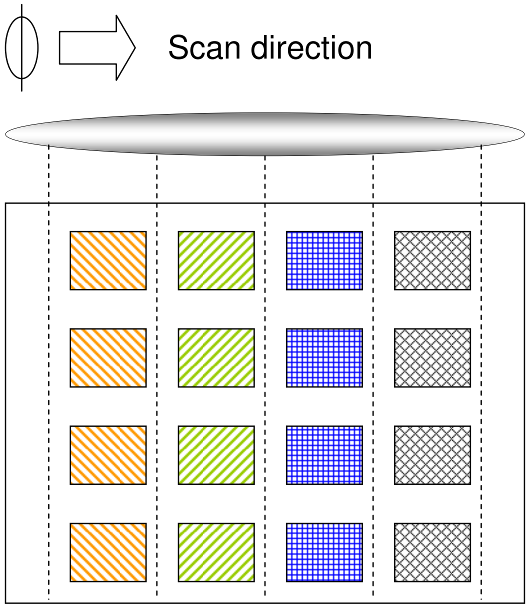


Figure IV.2: Different CD qualities of chips in a reticle due to aberration across the lens field.

IV.B.1 CD Impact of Lens Aberration

Several manufacturing process steps are involved in the transfer of the pattern on the mask to the photoresist, and then to the wafer. Lens aberration comes into play when the photoresist is exposed to light during lithography. Broadly speaking, a lithography setup includes one or more illumination sources, a mask, several lenses, and photoresist applied to the wafer. Modern lithography systems use step-and-scan to expose small portions of the wafer at a time, and then shift to the next region. The portion of the wafer that gets exposed in a step is called the *lens field*, or simply *field*. In each step, the photoresist is exposed to light through a slit that is scanned from one side of the field to the other [124].

Lens aberration parameters (Zernike's coefficients), which capture the divergence from ideal behavior of light, change as the slit translates horizontally. Hence, the CD error induced by lens aberration varies along the horizontal direction but stays constant along the vertical direction. While the variation in CD along the horizontal direction is continuous, it is reasonable to discretize it and

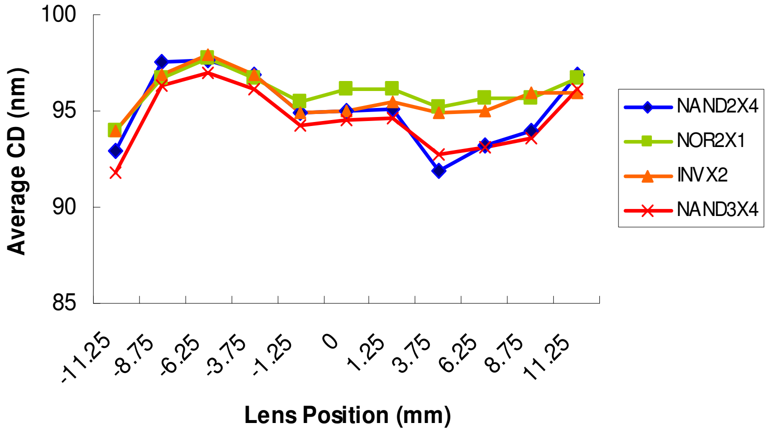


Figure IV.3: Average gate CD varies across the lens field; the range of this variation for the NAND2X4 cell is 8nm.

assume it to remain constant over small regions as shown in Figure IV.2. Based on industry-supplied Zernike's coefficients at multiple locations in the lens field, we run lithography simulation on some frequently-used standard cells from a 90nm foundry library, and study the impact on CD. Figure IV.3 shows average CD variation of devices in BUFX4, INVX2, NAND2X4 and NOR2X1 cell instances as their position within the lens field is varied. For example, average gate CD variation of NAND2X4 at 100nm worst defocus is up to 8nm across the entire lens field. In addition, we investigate the *CD skew* (maximum difference in CD over all devices in a cell) of different cells. Large CD skew can unbalance the timing arcs of a cell, as we discuss in greater detail in Section IV.C. Figure IV.4 shows the CD skew for NAND2X4 as its position in the lens field is changed. It is evident from these studies that the aberration impact on CD error is large across the lens field, and must be modeled to reduce guardbanding and overdesign.

IV.B.2 Delay Impact of Lens Aberration

Variations in CD directly and indirectly affect circuit delay. At the device level, increase in gate CD causes an approximately linear decrease in saturation on-current of the device, which partially determines delay. Since lens aberration

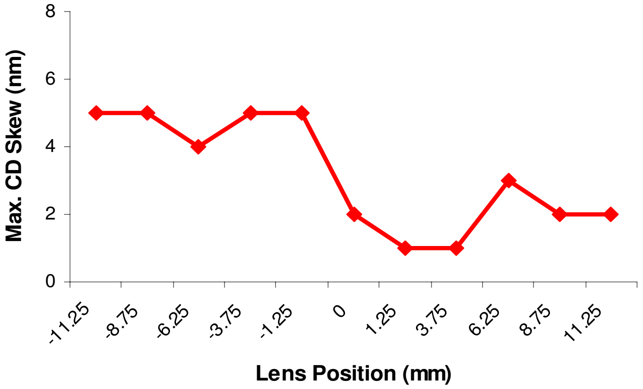


Figure IV.4: Maximum CD skew among all gates in NAND2X4 cell.

affects different devices in a cell differently, each of the cell's timing arcs can be affected differently. Most standard cells are designed such that the maximum difference in delays of timing arcs (*delay skew*) is small [61]. Due to lens aberration, however, this delay skew can increase - i.e., arcs that are governed by larger-than-nominal CDs will be slowed down, while those governed by smaller-than-nominal CDs will be sped up. Figure IV.5 shows how the delay, averaged over all timing arcs, changes for four cell masters as the cell instance location is varied from the lens center. Figure IV.6 shows the aberration-induced increase in delay skew with respect to the delay skew of the nominal (or drawn) cell as the location of cell NAND2X4 is varied in the field.[1]

CD variations also cause variations in cell input capacitance and output slews (transition times). Input capacitance affects the loading of fanin cells and consequently their delays; interconnect delays are also affected. Similarly, slews affect the output slews and delays of cells in the fanout cone. Again, to avoid unnecessary guardbanding, the performance analysis flow (library model characterization, timing/SI analysis, etc.) must comprehend these systematic variations.

[1] In the figure, the increase is always over 40% because in computing nominal delay skew, library characterization applies an equal CD error to all devices at worst-case process conditions. To compute aberration-induced delay skew, however, lithography simulation is performed at the worst-case process corner and all devices get different CD errors.

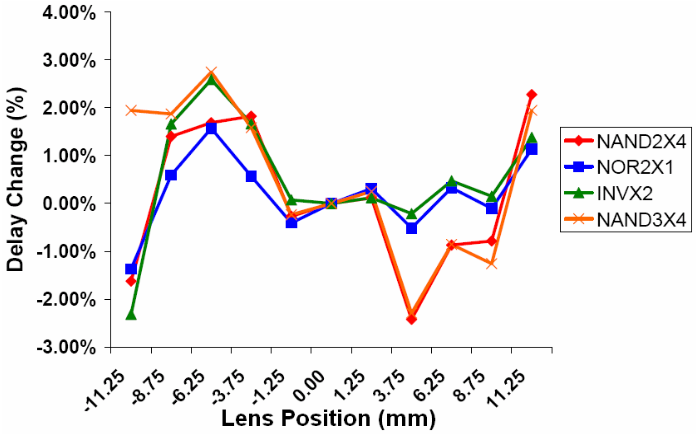


Figure IV.5: Change in average delay with lens position, with respect to the center of the lens.

IV.C Aberration-Aware Timing Analysis

In this section we describe our aberration-aware timing analysis flow. While the flow is complete and self-contained, it is at the same time designed for, and will be used by, the analytical placement framework described in Section IV.D. Our aberration-aware timing analysis flow involves two main steps: (1) constructing timing libraries of all standard cells for different locations in the lens field; and (2) using placement information of the design to compute the location of all cell instances in the lens field, then using this location information to look up appropriate models in the timing library for use with off-the-shelf static timing analysis (STA) tools.

Before describing our analysis flow, we describe two alternative flows and our reasons for not using them. In the first alternative flow, variants of each cell are created such that the CD of all devices in the cell is different for each variant, but the same for all devices in a given variant. A timing library can be created using SPICE models for all the variants. Since all devices in a cell variant have the same CD, we call this library a *cell-level* granularity library. To perform timing analysis on a placed design, lithography simulation is performed to obtain CDs of all devices in all cells. For each cell, the CDs of its devices can be averaged,

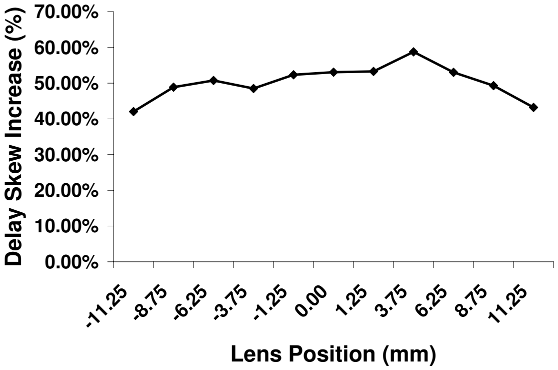


Figure IV.6: Percentage increase in delay skew (maximum difference in delays of all timing arcs) of the NAND2X4, relative to the maximum delay skew of nominal (or drawn) cell, cell as lens position is changed.

and the closest-matching available cell variant in the timing library then fed to off-the-shelf STA. However, as CD skews can be large, averaging of device CDs can introduce inaccuracy in the estimated impact of aberration. In other words, the effect of non-uniform CDs is non-uniformity in timing arc delays, rather than average increase or decrease in the delays of all timing arcs. Our experiments have found that the cell-level library-based approach is very inaccurate compared to the approach that we adopt.

The second alternative flow creates *a priori* variants for each cell master, such that there is one variant for every possible assignment of CDs to devices. This means that given any assignment of CDs to devices, an exactly matching, pre-characterized cell variant can be found. After lithography simulation provides CDs of all devices in all cells, a correctly matching variant can be picked for use in timing analysis. Though this flow is very accurate, it requires a very large number of cell variants (exponential in the number of devices in the cell); this is infeasible with respect to both characterization time and library size.

In our proposed flow, variants are created for each cell for different lens field locations. Figure IV.7 illustrates our timing library construction flow. We begin with standard cell GDSIIs and use Mentor Graphics *Calibre* (v 9.3_5.11) [9] for sub-resolution assist feature (SRAF) generation and model-based OPC. We use

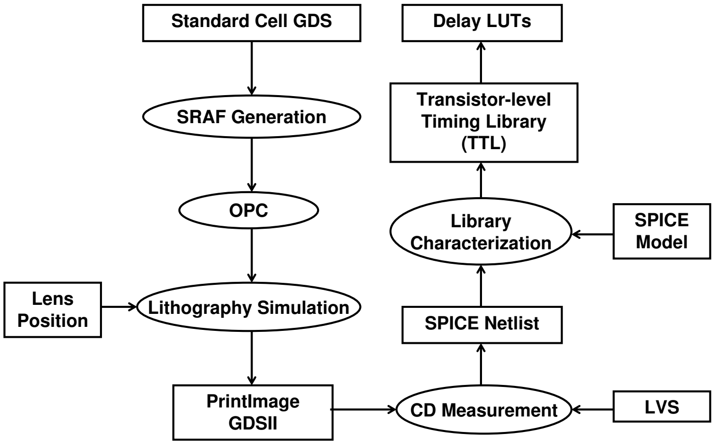


Figure IV.7: Aberration-aware timing analysis and its flow.

Zernike's coefficients for eight sampling positions in the lens field (data provided by a major chip maker), and compute the other coefficients at 19 different locations with 1.5mm stepsize on the field using linear interpolation. Using the post-OPC standard cell GDSIIs and Zernike's coefficients, we perform lithography simulation at 19 different field locations with wavelength $\lambda = 193$nm, numerical aperture NA $= 0.75$, and annular aperture $\sigma = 0.75/0.50$. After lithography simulation, we have 19 *PrintImage* GDSII results for each standard cell; we then measure the CD of each of the MOS devices in each GDSII result.

Figure IV.8(a) shows the *PrintImage* contour generated by Mentor Graphics *PrintImage* (v 9.3_5.11) [9] for one device.[2] To measure the CD of the PrintImage contours, we first take an intersection with the active layer to obtain the contour of the gate. Contours are rectilinearized and split into rectangles in a staircasing fashion. The lengths of all rectangles are then averaged with rectangle widths as weights to compute the CD of the gate (i.e., $CD_{gate} = \sum^n l_i \times w_i / \sum^n w_i$ where n is the number of rectangles into which the contour is split, and l_i and w_i are the length and width of the i^{th} rectangle).

The measured CDs are then used to alter SPICE netlists of standard cells, preparatory to running library characterization. A complication arises be-

[2] Mentor Graphics *PrintImage* produces rectilinear contours; our approach, however, is generic enough to be used for arbitrary polygonal contours.

cause GDSII typically does not have device names, while SPICE netlists only reference devices by device names. We solve this problem by applying LVS (layout vs. schematic) to obtain a mapping between device locations and device names. After modifying the SPICE netlists, we run Cadence SignalStorm (v 4.1) [3] to perform library characterization. Since lens aberrations affect different devices in a cell differently, the altered SPICE netlists may no longer have equal CD for all devices. We call our characterized library a *transistor-level timing library* (TTL); it accurately captures the delay skew induced by CD skew while incurring manageable added complexity of characterization effort and library size. The choice of the number of field locations to use depends on the extent and rate of change of aberration-induced CD. A larger number of field locations improves the accuracy but also increases the number of cell variants in the cell library.

Our test library contains 50 combinational cells. For each we create 19 variants corresponding to 19 field locations. Library characterization requires approximately 6 hours (wall time) running on 18 CPUs ranging from Intel *Xeon* 1.4GHz to AMD *Opteron* 2.2GHz. We do not create variants for the 13 sequential cells in our library due to large CPU time (estimated at 60 hours on our machines) required by their characterization. We note that while the characterization time can be significant, it is a one-time task for each process.

IV.D Aberration-Aware Timing-Driven Placer

Because of lens aberrations, a cell placed at different locations within the reticle will exhibit varying performance characteristics. In order to improve timing yield after manufacturing, we propose a lens aberration aware timing-driven placement formulation that minimizes total timing-weighted delays of cells in conjunction with common timing-driven placement objectives such as minimizing total timing-weighted wirelength. We implement our method based on a general analytical placement framework and describe implementation details in this section.

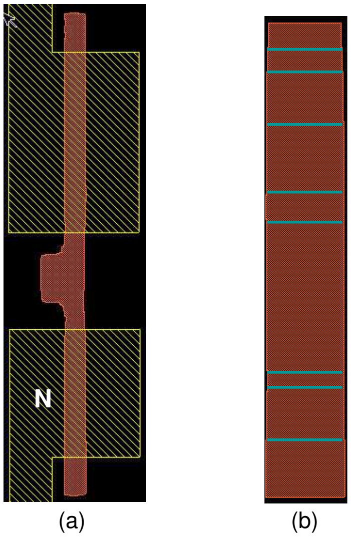


Figure IV.8: Polygon generation for CD measurement: (a) result of PrintImage simulation of an inverter and (b) rectilinearized polygon representation of a gate device in the region N of (a).

IV.D.1 Introduction of Analytical Placement

Analytical placement methods have recently received increased attention from both academia and industry [31, 32, 53, 64, 88, 120]. Specifically, recent work implements *APlace*, a general analytic placement framework [62, 64–66], which has high solution quality and strong extensibility. Here we briefly introduce the APlace analytic placement framework, upon which we build our proposed aberration-aware timing-driven placement method.

APlace formulates global placement as a *constrained nonlinear optimization problem*: the layout area is uniformly divided into global bins and APlace minimizes total half-perimeter wirelength (HPWL) while maintaining equalized cell area in each global bin (i.e., uniform density). A formal problem formulation is as follows:

$$\begin{aligned} &min \quad HPWL(\mathbf{x},\mathbf{y}) \\ &s.t. \quad\ \ D_g(\mathbf{x},\mathbf{y}) = D \quad \text{for each global cell } g \end{aligned} \tag{IV.1}$$

where $(\mathbf{x},\mathbf{y})$ is the vector of center coordinates of cells, $HPWL(\mathbf{x},\mathbf{y})$ is the total

HPWL of the current placement, $D_g(\mathbf{x},\mathbf{y})$ is a density function that equals the total cell area in a global bin g, and D is the average cell area over all global bins.

APlace applies smooth approximations of the HPWL and density functions and solves the constrained optimization problem in Equation (IV.1) using the simple *quadratic penalty method.* For example, the placer solves a sequence of unconstrained minimization problems of the form

$$min \quad HPWL(\mathbf{x},\mathbf{y}) + \tfrac{1}{2\mu}\textstyle\sum_g (D_g(\mathbf{x},\mathbf{y}) - D)^2 \tag{IV.2}$$

for a sequence of values $\mu = \mu_k \rightarrow 0$, with the solution of each unconstrained problem being used as an initial guess for the next one. A *Conjugate Gradient* (CG) solver is employed to optimize the objective function in Equation (IV.2). The conjugate gradient method is quite useful in finding an unconstrained minimum of a high-dimensional function. Also, the memory required is only linear in the problem size, which makes the approach adaptable to large-scale placement problems.

IV.D.2 Aberration-Aware Placement Formulation

We now propose a novel aberration-aware timing-driven placement objective for improved timing yield after manufacturing, and describe its integration into the analytical placement framework. We perform aberration-aware timing-driven placement by optimizing a hybrid placement objective. Besides the typical objective of minimizing total timing-weighted net wirelength, we also minimize the sum of timing-weighted delays of timing-critical cells. The aberration-aware timing-driven placement formulation is as follows:

$$\begin{aligned} min \quad & WWL(\mathbf{x},\mathbf{y}) + W_a \sum_v w(v) \cdot g_{t_v}(x_v) \\ s.t. \quad & D_g(\mathbf{x},\mathbf{y}) = D \text{ for each global bin } g \\ & \quad \text{and } g_{t_v}(x_v) = MAX\{g^1 t_v(x_v), \cdots, g^n t_v(x_v)\} \end{aligned} \tag{IV.3}$$

where $WWL(\mathbf{x},\mathbf{y})$ is the sum of timing-weighted net HPWL of the current placement and W_a is the weight for the aberration-aware timing-driven objective function terms, which is the sum of timing-weighted delays of timing-critical cells.[3] In the formulation, $g_{t_v}(x_v)$ is the delay function, obtained from the TTL timing library described above, for cell instance v's timing model t_v; it is a function of v's horizontal position x_v in the chip. In the situation where there are multiple copies ($n > 1$) of chips in the reticle, we let $g^i t_v(x_v)$ be the delay function for the i^{th} chip[4], and we consider the maximum delay of cell instance v over all copies so that the performance of the slowest chips is improved. We note that this is a pessimistic approximation of cells' delays, since not all timing-critical cells may exhibit their maximum delays on the same chip copy. However, we do not consider this pessimism to be significant, since the impact of aberration on delays of all cells is similar, and a chip copy that has large delay for one cell likely has large delays for other cells as well. For example, cells except INV1 and INV2 (which have only one isolated line) have similar cell delay behavior, as we saw previously in Figure IV.5. This is because all cells share similar parameters of pitch, width and design rules, and because linewidth variation due to lens aberration is not a function of cell type, but rather a function of pattern geometry. We thus believe that our delay upper-bounding is not significantly pessimistic.

Our problem formulation applies to the single lens as well as multiple chips on a wafer. A modern fab may employ multiple lithography lens systems. For high-volume, cutting-edge designs such as microprocessors, it is already common practice to have stepper-specific masks. Stepper-specific masks are tuned according to the stepper "signature" as part of the RET/mask data preparation flow. Our methodology brings aberrations upstream in the design and is easily adoptable when stepper-specific masks are used. Further, recent studies of lens

[3] We divide the objective into two parts since we consider the aberration-induced variation in only the cell delay. Aberration-induced CD variation of the gates affects timing yield, while aberration-induced CD variation of wires may be neglected in comparison to the impact of HPWL in wire delay. Note that larger CD of a wire increases the capacitance, but decreases the resistance, and vice versa.

[4] The critical-path delay of the i^{th} copy of the chip depends on the horizontal position of that copy in the reticle. We assume that the chip size can be determined using any initial placement optimization, and that the horizontal position of a copy of the chip can then be obtained from a reticle floorplan. Our aberration-aware placement optimization thus incorporates the chip size and the reticle floorplan.

aberration enable quick measurement of Zernike's parameters to capture lens aberrations [33, 112]. Even when identical masks are used on multiple steppers, it is preferable and common practice to use steppers from the same manufacturer to reduce stepper-to-stepper variations. Steppers from the same manufacturer have very similar aberrations and our methodology can use the Zernike's coefficients from any one stepper to optimize the design. It is also possible to extract the systematic aberration components from a database of aberration measurements of various lithography systems, and then generate a lens aberration map incorporating "universal" Zernike's parameters which can be applied to our aberration-aware placement flow. All of these scenarios leverage the basic design optimization that is proposed in this chapter.

As with traditional net weighting methods, we assign timing weights to cells based on timing criticality and path sharing. First, a cell along a timing-critical path should receive a heavy weight. Second, a cell with many timing-critical paths passing through should have a large weight as well. Therefore, we assign to cell v the weight $w(v)$, given as

$$w(v) = \textstyle\sum_{v\in\pi}(D_s(slack_s(\pi), T_s) \cdot D_h(slack_h(\pi), T_h) - 1) \tag{IV.4}$$

where

$$D_s(slack_s(\pi), T_s) = \begin{cases} (1 - s/T)^{\delta} & s \leq 0 \\ 1 & s \geq 0 \end{cases} \tag{IV.5}$$

and

$$D_h(slack_h(\pi), T_h) = \begin{cases} (1 + s/T)^{\delta} & s \leq 0 \\ 1 & s \geq 0 \end{cases} \tag{IV.6}$$

Here, δ is the criticality exponent, and u is the expected improvement of the longest (or shortest) path delay after this timing-driven iteration. T is $T_s = (1 - u) \cdot \max_{\pi}\{delay(\pi)\}$ for setup-critical paths or $T_h = (1 + u) \cdot \min_{\pi}\{delay(\pi)\}$ for hold-critical paths. Additionally, $slack_s(\pi) = T_s - delay(\pi)$ is the slack of a setup-critical path π, while $slack_h(\pi) = delay(\pi) - T_h$ is the slack of a hold-critical path

π. In Equation (IV.4), we compute a weight for each timing-critical path based on its slack, and obtain the timing weight of a cell by summing up the weights of timing-critical paths passing through it.

For timing-driven edge weights, existing approaches can be broadly divided into two classes, *path-based* and *net-based*. The path-based approach is based on mathematical programming techniques, and can maintain an accurate timing view during optimization. But, its drawback is relative high complexity. We use a net-weighting based approach which assigns weight to nets based on their timing criticality [63, 74, 85].[5] The basic idea is that a timing-critical net should receive a heavy weight, and an edge with many paths passing through it should have a heavy weight as well. We thus assign to edge e the weight $w(e)$, given as

$$w(e) = 1 + \textstyle\sum_{e \in \pi}(D_s(slack_s(\pi), T_s) \cdot D_h(slack_h(\pi), T_h) - 1) \tag{IV.7}$$

where $D_s(slack_s(\pi), T_s)$ and $D_h(slack_h(\pi), T_h)$ have the same formulations as in Equations (IV.5) and (IV.6). Note that the balance of timing weights between wire and cell is determined by the weight W_a for consideration of the aberration-aware timing-driven objective. Note also that the constant 1 in Equation (IV.7) means that the weight for non-timing critical nets is 1 (whereas timing-critical nets will have a weight > 1). Equation (IV.4) does not require the constant 1 because the total wirelength of non-timing critical nets is optimized at the same time besides the timing-related objectives.

IV.D.3 Placement Flow

Our aberration-aware timing-driven placement and evaluation flow is shown in Figure IV.9. In addition to the design netlist, we also input the delay functions of cell models, which represent how the delays of given cell models change with their horizontal position in the chip.

The timing-driven process in our placer may include several iterations. As shown in Figure IV.9, during each iteration, we send the intermediate placement

[5] Note that in the timing analysis step, we use commercial extraction tools for accurate wire delay estimation.

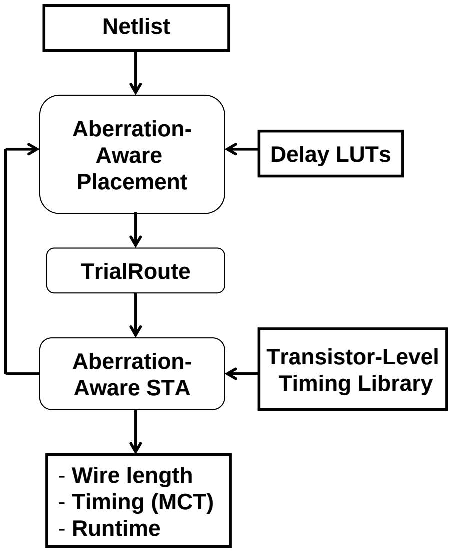


Figure IV.9: Aberration-aware timing-driven placement and evaluation flow.

to *TrialRoute* (Cadence *SOC Encounter* v 2004.10) [4] to perform a fast global and detailed routing, and extract RC parasitics.[6] We then change the type of each cell in the netlist according to its horizontal position within the lens field and use Synopsys *PrimeTime* (v W-2004.12-SP2) [12] to perform accurate aberration-aware static timing analysis (STA) with the transistor-level timing libraries (TTLs) described in Section IV.C. The resulting critical paths are imported into the placer to decide timing weights for nets and cells. The total timing-weighted cell delay is then minimized using the Conjugate Gradient solver, together with the timing-weighted wirelength objective, and subject to density constraints.

IV.D.4 Implementation Details

As mentioned above, for each master cell, we create 19 different variants according to 19 lens field locations. Through the recticle floorplan, we can extract the position of the i^{th} chip in the field and (in the timing analysis) instantiate timing model variants corresponding to the actual position of each instance of the

[6] Separately, we have verified that TrialRoute results give the same conclusions as final detailed routing results. We use TrialRoute because of runtime constraints for our large number of experiments.

given master cell. Thus, there is no need to create variants for different copies. In Equation (IV.3), $g^it_v(x_v)$ is the delay function of the i^{th} chip, which is generated using (interpolation of) the position – specific delay model variants.

We compute the weight of the aberration-aware objective W_a in Equation (IV.3) according to the x-gradients derived from the wirelength and delay terms, so that the scaled gradients of delay functions are comparable to the wirelength gradients, i.e.,

$$W_a = \alpha \cdot (\sum_v |\frac{\partial WWL}{\partial x_v}|) \ / \ (\sum_v |\frac{\partial g_{t_v}}{\partial x_v}|) \qquad \text{(IV.8)}$$

The delay ratio α decides the ratio of the delay gradients to the wirelength gradients, and must be carefully tuned according to the impact of reduced cell delay and increased net wirelength on design performance.

We derive the delay of a cell at a specific horizontal field position by averaging the rise and fall delays of all timing arcs with zero wire load, according to the transistor-level timing libraries. Thus, the delay functions represent how gate delays vary with horizontal locations and gate CDs. Due to simulation limits, delay functions have accurate values only at discrete horizontal coordinates, and consequently are expressed as look-up tables (LUTs). We obtain delay at continuous positions using linear interpolation and compute gradients accordingly.

A smoothing technique [40] can be applied to smooth the delay curves. To reduce the effect of local minima, we use a local search method with search space smoothing technique. The smoothing technique transforms the given problem into a series of problem instances with different terrain structures. Initially, a simplified instance with a smooth terrain surface is solved using the local search algorithm [40]. Then, the solution of the problem instance is then taken as the initial solution for the next problem instance that has a slightly more complicated search space. The problem is again solved using the same algorithm. The above procedure is repeated until the final problem instance having the original search space is solved. Given a normalized delay function, a smooth function is a pre-

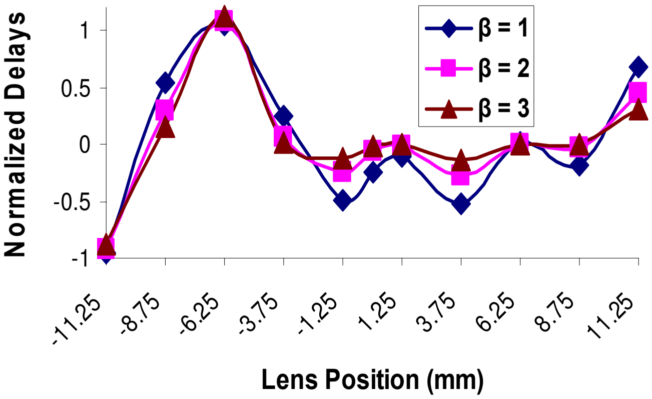


Figure IV.10: Delay curves of NOR2X1 with a variety of smoothing factors (β's).

defined smoothing factor $\beta \geq 1$ as follows:

$$g' = \begin{cases} \overline{g} + (g - \overline{g})^{\beta} & if\ g \geq \overline{g} \\ \overline{g} - (\overline{g} - g)^{\beta} & if\ g \leq \overline{g} \end{cases} \tag{IV.9}$$

where $\overline{g}$ is the average value of the delay function. Figure IV.10 shows delay curves with a variety of smoothing factors β for NOR2X1. A delay function generated from a larger β exhibits a smoother curve, while a delay function generated from a smaller β exhibits a more rugged curve.

IV.E Experimental Setup and Results

In this section, we empirically test our aberration-aware placement approach on two designs within a standard design flow using commercial design automation tools. We assess the impact on timing, wirelength, and runtime.

Experimental setup. We use two designs from OpenCores [11] as our testcases. The circuits are synthesized using Synopsys *Design Compiler* (v W-2004.12-SP3) [12] with tight timing constraints and a set of 63 most commonly used standard cells (50 combinational, 13 sequential) from Artisan TSMC 90nm library, then floorplanned in Cadence *SOC Encounter* (v 2004.10) [4]. The design characteristics

Table IV.1: Design characteristics of two benchmark circuits.

Design	**Utilization (%)**	**Chip Size (mm)**	**#Cells**	**#Nets**
AES	60	0.50	17304	17465
JPEG	60	1.41	118321	125036

are summarized in Table IV.1. The experimental flow is shown in Figure IV.9. The inputs for each design include synthesized netlists, floorplan, timing constraints, aberration-aware timing libraries, delay look-up tables derived from the libraries for convenience of the placer, and physical libraries in LEF format. The placer executes iteratively with STA to improve and converge on timing.

We evaluate the following three timing-driven placers.

- *TradPl_TD*: Analytical timing-driven placer, APlace, with the traditional (or standard) STA during the placement optimization. This is the traditional timing-driven analytical placer.
- *APlace_TD*: Timing-driven APlace with aberration-aware STA. Aberration-aware STA accounts for aberration-induced cell delay changes, and therefore computes more accurate timing slacks which are used in the timing-driven placer objective function.
- *AberrPl_TD*: Aberration-aware timing-driven placer, with timing-driven wirelength and aberration objectives, and aberration-aware STA. This improves upon *APlace_TD* by explicitly accounting for aberration-induced cell delay changes in the placement objective function.

We use aberration-aware STA to compare the three placers for circuit delay.

We expect larger chips to benefit more from our aberration-aware placement technique since they will have larger CD and delay variation induced by an imperfect lens system across the layout region. However, our testcases are not sufficiently large to witness the effect of lens aberration that may be observed in

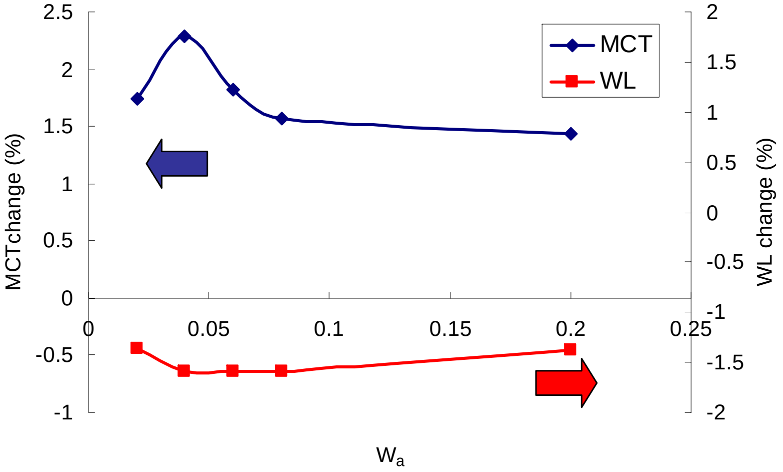


Figure IV.11: MCT change of AberrPl_TD according to the weight of the aberration-aware objective W_a for testcase AES.

real-world systems on chip. Hence, in our studies we scale the aberration map, which captures the impact of aberration at every chip location, along the horizontal direction to mimic the aberration that is observed in larger modern designs.

We perform three sets of experiments to evaluate the performance improvement under different die size and field size scenarios: (1) when there is only one copy of the chip in the lens field; (2) when there are multiple copies (the number of which is determined by a scaling factor, with a variety of scaling factors), and (3) when field blading is performed for partial reticle exposure. We compute timing weights with criticality exponent $\delta = 4$ and expected improvement $u = 10\%$. Note that we only perform APlace_TD and AberrPl_TD for experiments (2) and (3) since the result of TradPl_TD is always the same as the result of (1). Figure IV.11 shows the minimum cycle time (MCT) change of AberrPl_TD according to the weight of the aberration-aware objective W_a for circuit AES. With $W_a = 0.04$, MCT and trial-routed wirelength are optimized and we use this value of W_a in our experiments. In general, MCT improvement results in increase of wirelength.

After each placement, we perform global and detailed routing, RC extraction, and finally aberration-aware timing analysis using Synopsys *PrimeTime*. MCT of the slowest chip in the reticle is reported by aberration-aware STA to

Table IV.2: Comparison of traditional timing-driven placement (TradPl_TD) versus APlace_TD placement or AberrPl_TD placement for testcases AES and JPEG.

Design	**Method**	**Place**		**TrialRoute**		**STA**	
		HPWL	**CPU**	**WL**	**#Vias**	**MCT**	**TNS**
		(e9 um)	**(s)**	**(e5 um)**	**(e5)**	**(ns)**	**(ns)**
AES	TradPl_TD	1.1699	1432	6.521	1.2521	1.8491	156.3829
	APlace_TD	1.1803	1457	6.541	1.2531	1.8013	150.8231
	Impr. (%)	**-0.8919**	**-1.7458**	**-0.3067**	**-0.0743**	**2.5850**	**3.5552**
	AberrPl_TD	1.1922	1471	6.645	1.2542	1.7443	144.9321
	Impr. (%)	**-1.9090**	**-2.7235**	**-1.9016**	**-0.1629**	**5.6668**	**7.3223**
JPEG	TradPl_TD	6.2880	23598	3.717	6.1762	2.9252	213.4321
	APlace_TD	6.3312	23791	3.743	6.1874	2.8875	206.3124
	Impr. (%)	**-0.6871**	**-0.8179**	**-0.6995**	**-0.1809**	**1.2888**	**3.3357**
	AberrPl_TD	6.3932	24139	3.780	6.1938	2.7751	196.8943
	Impr. (%)	**-1.6731**	**-2.2926**	**-1.6949**	**-0.2846**	**5.1313**	**7.7484**

measure performance of timing-driven placements. We also report HPWL and runtime for placement, and routed wirelength and the number of vias after routing. All experiments are conducted on Linux machines with 2.4GHz CPU and 4GB memory.

Experimental results. Table IV.2 summarizes the results of TradPl_TD, APlace_TD, and AberrPl_TD on our two testcases, AES and JPEG, when there is one die in a reticle. In comparison to TradPl_TD, APlace_TD reduces MCT by 2.585% (48ps) with 0.892% HPWL increase and 0.307% increase of trial-routed wirelength for AES, and reduces MCT by 1.289% (38ps) with 0.687% HPWL increase and 0.7% increase of trial-routed wirelength for JPEG. Our aberration-aware placer (AberrPl_TD), in comparison to traditional timing-driven placement (TradPl_TD), reduces MCT by 5.667% (105ps) with 1.909% HPWL increase and 1.902% increase of trial-routed wirelength for AES, and reduces MCT by 5.13% (150ps) with 1.673% HPWL increase and 1.695% increase of trial-routed wirelength for JPEG. More-

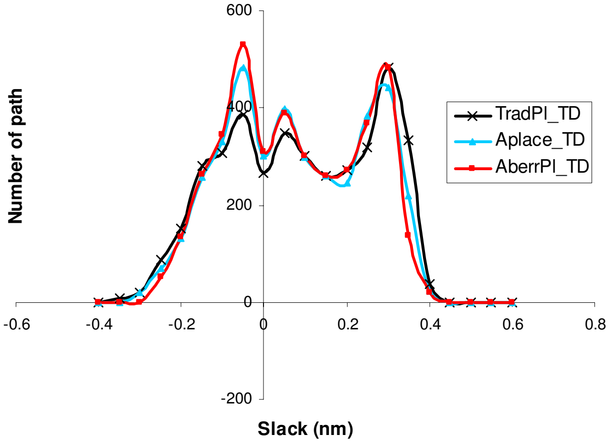


Figure IV.12: Slack distributions of TradPl_TD, APlace_TD and AberrPl_TD for testcase AES.

over, Aberr_TD, in comparison to TradPl_TD, reduces total negative slack (TNS) by 7.322% for AES, and by 7.748% for JPEG. Figure IV.12 shows the slack distributions of TradPl_TD, APlace_TD and AberrPl_TD for AES.

Impact of scaling. Our second set of experiments evaluates the effect of chip size on performance improvement obtained with our aberration-aware placement method. We perform AberrPl_TD with a variety of scaling factors, such that the number of die copies within the reticle is 1x1, 2x2, 4x4, 6x6, and 8x8. The results for circuits AES and JPEG are presented in Table IV.3 and Table IV.4, respectively. We report the improvement of the slowest chips among the multiple copies of chips. Comparing with APlace_TD, we see that MCT of AberrPl_TD with the scaling factor improves by $2.731 - 3.164\%$ $(50 - 57\text{ps})$ for AES and by $2.884 - 3.892\%$ $(85 - 112\text{ps})$ for JPEG. Trial-routed wirelength increases by $1.434 - 1.59\%$ for AES and by $0.668 - 0.989\%$ for JPEG, which is negligible compared to the significant MCT improvement.

Figure IV.13 shows the MCT and trial-routed wirelength improvement as a function of the scaling factor. We observe that the performance improvement obtained gradually decreases as the number of copies in the field increases. However, larger chip size may not always achieve better timing improvement compared

Table IV.3: Results of aberration-aware placement (AberrPl_TD) with a variety of scaling factors for testcase AES.

Copies	**Method**	**Place**		**TrialRoute**		**AberrSTA**
		HPWL (e9 um)	**CPU (s)**	**WL (e5 um)**	**#vias (e5)**	**MCT (ns)**
1	APlace_TD	1.1803	1457	6.541	1.2531	1.8013
	AberrPl_TD	1.1922	1471	6.645	1.2542	1.7443
	Imp (%)	**-1.0081**	**-0.9609**	**-1.5900**	**-0.0886**	**3.1636**
2	APlace_TD	1.1814	1469	6.548	1.2531	1.8212
	AberrPl_TD	1.1923	1486	6.651	1.2545	1.7651
	Imp (%)	**-0.9210**	**-1.1572**	**-1.5730**	**-0.1085**	**3.0812**
4	APlace_TD	1.1813	1478	6.555	1.2531	1.8461
	AberrPl_TD	1.1927	1491	6.657	1.2544	1.7942
	Imp (%)	**-0.9677**	**-0.8796**	**-1.5561**	**-0.1037**	**2.8093**
6	APlace_TD	1.1814	1482	6.556	1.2532	1.8483
	AberrPl_TD	1.1926	1499	6.651	1.2545	1.7974
	Imp (%)	**-0.9503**	**-1.1471**	**-1.4490**	**-0.1061**	**2.7558**
8	APlace_TD	1.1814	1487	6.555	1.2532	1.8500
	AberrPl_TD	1.1929	1502	6.649	1.2545	1.7995
	Imp (%)	**-0.9759**	**-1.0087**	**-1.4340**	**-0.1021**	**2.7310**

to the smaller chip size with aberration-aware placement. For example, suppose that there are two regions of the field which make gate CDs small (i.e., gate delay fast) due to aberration, as shown in Figure IV.14. In the case of 1x1 copy, aberration-aware placement will attempt to place timing-critical cells in these two regions to improve the gate delay. However, due to the limited size of the regions, not all timing-critical cells in a timing-critical path can be accommodated in one region. As a result, the separation of cells from a timing-critical path into two regions increases the wirelength, and consequently delay, of the timing-critical path. In the case of 2x2 copies, all timing-critical cells can be placed in one region or

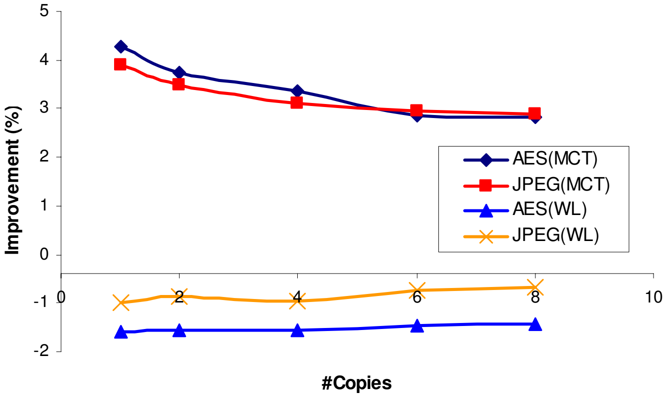


Figure IV.13: Routed wirelength (WL) and MCT of AberrPl_TD as functions of the scaling factor for testcases AES and JPEG.

the neighborhood of the region. As a result, aberration-aware placement does not significantly affect wirelength, and 2x2 copies could have smaller delay than 1x1 copy.

Impact of blading. A third set of experiments validates the proposed method when used in conjunction with lens field blading which allows partial reticle exposure. Balasinski [15] proposed a multilayer mask technology which relies on sharing the reticle space between multiple layers of the same design. Based on the concept, which cuts out parts of the lens field, we propose a new *blading column technique* (BCT) to further optimize MCT in conjuction with our aberration-aware placement. The technique avoids the use of those portions of the aberration map that induce a large, positive gate delay variation.

In our experiments we assume that there are four die copies in the field as shown in Figure IV.15. BCT allows any two dies to be exposed, thereby only partially using the reticle. For example, if we blade columns (2, 4) at the first exposure, only columns (1, 3) in the lens field are exposed for the chips in reticle columns 1 and 3. Chips in columns (2, 4) can be exposed in a second exposure after moving the wafer stage to use columns (1, 3) in the lens field again. Note that we use only some columns for exposure of all chips, selectively blading the columns that have aberration that is unfavorable to chip performance. Unfortunately, BCT

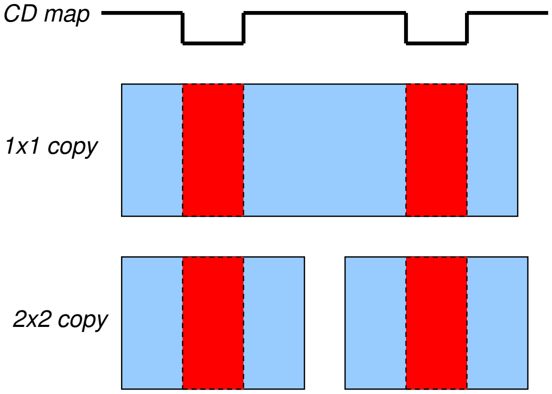


Figure IV.14: Example showing non-monotonicity of achievable MCT versus chip size. Red color represents fast cell delay regions in the lens aberration map. It is possible to achieve better MCT even with smaller chip size (e.g., 2x2 copies per field instead of 1x1 copy per field).

requires two exposure passes and thus the throughput is halved. When not all columns are used, our aberration-aware placement performs timing optimization for only the columns in the lens field that are used. We evaluate the use of our technique with blading by considering several blading schemes and assessing the impact on chip performance, HPWL, and trial-routed wirelength.

We assume that there are four columns, where column numbers increase from left to right, in the reticle with 4x4 die copies of chip. The results are summarized in Table IV.5 and Table IV.6. Three comparisons in MCT improvement are presented: (1) blading versus no blading (**Impr.1**), (2) blading column of AberrPl_TD versus blading column of APlace_TD (**Impr.2**) and (3) blading column of AberrPl_TD versus no blading of APlace_TD (**Impr.3**). The results of APlace_TD and AberrPl_TD show the performance improvements obtained using BCT. We observe that for testcase AES, APlace_TD and AberrPl_TD can respectively reduce MCT by 1.724% and 1.556% with 0.003% and 0.107% HPWL increase, and 0.031% and 0.03% increase in trial-routed wirelength. With (1, 2) blading columns for the JPEG testcase, APlace_TD and AberrPl_TD can respectively reduce MCT by 0.784% and 2.131%, with 0.0162% and 0.0187% HPWL increase, and 0.0534%

Table IV.4: Results of aberration-aware placement (AberrPl_TD) with a variety of scaling factors for circuit JPEG.

Copies	**Method**	**Place**		**TrialRoute**		**AberrSTA**
		HPWL (e9 um)	**CPU (s)**	**WL (e5 um)**	**#vias (e5)**	**MCT (ns)**
1	APlace_TD	6.3312	23791	3.743	6.1874	2.8875
	AberrPl_TD	6.3932	24139	3.780	6.1938	2.7751
	Imp (%)	**-0.9792**	**-1.4627**	**-0.9885**	**-0.1036**	**3.8918**
2	APlace_TD	6.3340	23801	3.746	6.1881	2.9009
	AberrPl_TD	6.3988	24211	3.778	6.1940	2.8002
	Imp (%)	**-1.0236**	**-1.7226**	**-0.8542**	**-0.0952**	**3.4710**
4	APlace_TD	6.3381	23821	3.745	6.1891	2.9309
	AberrPl_TD	6.3918	24203	3.781	6.1943	2.8396
	Imp (%)	**-0.8474**	**-1.5396**	**-0.9612**	**-0.0835**	**3.1176**
6	APlace_TD	6.3379	23801	3.744	6.1892	2.9210
	AberrPl_TD	6.4021	24298	3.772	6.1944	2.8613
	Imp (%)	**-1.0124**	**-2.0881**	**-0.7479**	**-0.0837**	**2.9549**
8	APlace_TD	6.3379	23802	3.745	6.1881	2.9380
	AberrPl_TD	6.4020	24299	3.77	6.1942	2.8532
	Imp (%)	**-1.0104**	**-2.0881**	**-0.6676**	**-0.0989**	**2.8835**

and 0.1322% increase in trial-routed wirelength. The absolute MCT improvements achieved with BCT and AberrPl_TD foR AES and JPEG are 28ps and 61ps, respectively.

We also compare MCT improvements of blading for AberrPl_TD versus corresponding improvements for APlace_TD (**Impr.2**), and MCT improvements for AberrPl_TD with blading versus corresponding improvement for APlace_TD with no blading (**Impr.3**)[7]. For the AES testcase, AberrPl_TD (**Impr.2**) can

[7] There are no entries in **Impr.2** and **Impr.3** of APlace_TD as shown in Tables IV.5 and IV.6 since we compare APlace_TD with AberrPl_TD and record the improvements in the AberrPl_TD column.

Table IV.5: Results of timing-driven APlace (APlace_TD) and aberration-aware placements (AberrPl_TD) with a variety of blading columns for testcase AES. Three comparisons in MCT improvement are presented.

Method	**Blading Col.**	**Place**		**TrialRoute**		**AberrSTA**	**Impr.1**	**Impr.2**	**Impr.3**
		HPWL	**CPU**	**WL**	**#vias**	**MCT**	**MCT**	**MCT**	**MCT**
		(e9 um)	**(s)**	**(e5 um)**	**(e5)**	**(ns)**	**(%)**	**(%)**	**(%)**
APlace_TD	No Blading	1.18128	1478	6.555	1.2531	1.8461	—	—	—
	2,4	1.18131	1479	6.556	1.2534	1.8260	1.0846	—	—
	1,3	1.18122	1477	6.557	1.2535	1.8268	1.0422	—	—
	3,4	1.18131	1479	6.556	1.2537	1.8461	0.0000	—	—
	1,2	1.18131	1481	6.556	1.2537	1.8171	1.5710	—	—
	1,4	1.18131	1480	6.557	1.2536	1.8142	1.7241	—	—
	2,3	1.18131	1479	6.556	1.2535	1.8260	1.0846	—	—
AberrPl_TD	No Blading	1.1927	1491	6.657	1.2544	1.7942	—	2.8093	—
	2,4	1.1929	1492	6.658	1.2549	1.7731	1.1746	2.8978	3.9509
	1,3	1.1929	1493	6.659	1.2550	1.7721	1.2329	2.9966	4.0075
	3,4	1.1929	1491	6.657	1.2550	1.7728	1.1931	3.9688	3.9688
	1,2	1.1931	1493	6.659	1.2553	1.7663	1.5560	2.7944	4.3215
	1,4	1.1930	1490	6.658	1.2551	1.7678	1.4691	2.5572	4.2371
	2,3	1.1930	1489	6.658	1.2550	1.7731	1.1746	2.8978	3.9509

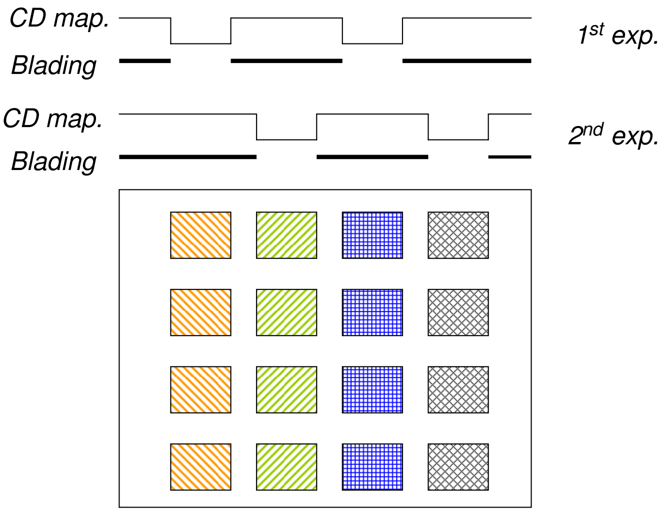


Figure IV.15: Illustration of *blading column technique*. For the first exposure, columns 1 and 3 in a lens field are used for chips in reticle columns 1 and 3, while columns 2 and 4 are bladed.

reduce MCTs by 2.557 − 3.969% (i.e., 46 − 73ps), with 0.9676 − 1.074% (resp. 1.541 − 1.556%) increase in half-perimeter (resp. trial-routed) wirelength. For the JPEG testcase, AberrPl_TD (**Impr.2**) can reduce MCTs by 2.811 − 4.434% (i.e., 82−129ps), with 0.848−0.868% (resp. 0.9072−1.0152%) increase in half-perimeter (resp. trial-routed) wirelength. **Impr.3** shows the maximum improvement of AberrPl_TD with the blading column technique. For AES, AberrPl_TD with (1, 2) blading reduces MCT by 4.322%, (i.e., 80ps) with 1.00% (resp. 1.587%) increase in half-perimeter (resp. trial-routed) wirelength. For JPEG, AberrPl_TD with (1, 2) blading reduce MCT by 5.182% (i.e., 152ps) with 0.866% (resp. 1.095%) increase in half-perimeter (resp. trial-routed) wirelength. Averaged over our two testcases, worst-case cycle time and total negative slack respectively reduce by ∼ 4.749% (i.e., 116ps) and ∼ 7.535%.

We consider the observed MCT improvements (i.e., 80 − 152ps) achieved by our aberration-aware placement and the blading column technique to be quite significant. Such MCT reductions can tremendously improve parametric yield and quicken timing closure. The penalties of HPWL, trial-routed wirelength, and the

Table IV.6: Results of timing-driven APlace (APlace_TD) and aberration-aware placements (AberrPl_TD) with a variety of blading columns for circuit JPEG. Three comparisons in MCT improvement are presented.

Method	**Blading Col.**	**Place**		**TrialRoute**		**AberrSTA**	**Impr.1**	**Impr.2**	**Impr.3**
		HPWL	**CPU**	**WL**	**#vias**	**MCT**	**MCT**	**MCT**	**MCT**
		(e9 um)	**(s)**	**(e5 um)**	**(e5)**	**(ns)**	**(%)**	**(%)**	**(%)**
APlace_TD	No Blading	6.3381	23821	3.745	6.1891	2.9309	—	—	—
	2,4	6.3374	23828	3.746	6.1892	2.9237	0.2428	—	—
	1,3	6.3387	23832	3.744	6.1893	2.9164	0.4951	—	—
	3,4	6.3390	23824	3.743	6.1892	2.9240	0.2356	—	—
	1,2	6.3391	23824	3.747	6.1893	2.9080	0.7835	—	—
	1,4	6.3391	23826	3.746	6.1892	2.9121	0.6455	—	—
	2,3	6.3383	23824	3.748	6.1893	2.9138	0.5866	—	—
AberrPl_TD	No Blading	6.3918	24203	3.781	6.1943	2.8396	—	3.1176	—
	2,4	6.3924	24213	3.783	6.1943	2.8292	0.3632	3.2345	3.4695
	1,3	6.3924	24211	3.782	6.1943	2.8274	0.4236	3.0480	3.5280
	3,4	6.3918	24206	3.781	6.1944	2.8281	0.4066	3.2837	3.5115
	1,2	6.3930	24214	3.786	6.1944	2.7789	2.1312	4.4337	5.1824
	1,4	6.3931	24217	3.781	6.1943	2.8076	1.1263	3.5865	4.2088
	2,3	6.3921	24211	3.782	6.1943	2.8318	0.2716	2.8107	3.3808

number of vias are less than 1.5%, and are practically negligible. We note that the concept of stepper-specific place-and-route has long been attractive to high-end, high-volume custom products - e.g., Dr. N. Sherwani of Intel posed exactly this challenge to the physical design community at the 1999 International Symposium on Physical Design (ISPD). With the future of process module costs, inherent equipment variabilities, and exclusivity of fabless-foundry tie-ups all being unclear today, we believe that it will be important to have stepper-specific layout flows available going forward.

IV.F Conclusions

We have proposed an accurate aberration-aware timing analysis flow and a novel aberration-aware timing-driven placement technique, *AberrPl*, as a practical and effective approach to improve timing yield after manufacturing. We implement our method based on a general analytical placement framework and test it within a standard industry flow using leading-edge tools. We also study the dependence of our improvement on chip size, and when the technique is used along with field blading which allows partial reticle exposure. Averaged over our two testcases, worst-case cycle time and total negative slack respectively reduce by $\sim$ 4.749% (116ps) and $\sim$ 7.535% at the cost of $\sim$ 1.341% increase in wirelength, with hold-time violations.

The benefits of AberrPl_TD are expected to increase in future technology nodes. We are currently engaged in further experimental validation and research. Our ongoing research is in the following directions.

- The proposed aberration-aware placement approach aims at improving performance of all design copies in the reticle field and hence is limited by the slowest ones. However, for many designs, chips of slower speeds can also be sold, albeit at a lower value (speed binning). We plan to improve our approach so that the total value of all chips is maximized.

- We also wish to enhance our placer to comprehend leakage constraints, since leakage is increasingly starting to determine yield and is exponentially affected by CD.

- We are researching the possibility of an aberration-aware OPC method which applies different OPC models for devices at different lens positions, instead of the simple OPC method with average Zernike's coefficients across the reticle, to improve pattern printability and lithographic process window. While we noted at the outset that global placement seems to be a more appropriate knob than OPC for compensation of lens aberration, we wish to pursue a clear confirmation or refutation of this intuition.

- For chip manufacturing, a modern fab can employ multiple lithography lens systems. Different lenses will have different lens aberrations. For very high-volume production of a chip (ASIC or microprocessor), it may be the case that multiple systems are used simultaneously. We plan to improve our placement engine to achieve "generic" aberration-aware placements that improve parametric yield in light of the systematic lens aberrations of all the lenses.

- A modern reticle may contain multiple chips (especially for ASICs), even outside the shuttle context. Different chips are located at different points in the reticle field. We are developing an aberration-aware placement to address such multiple concurrent instances of design optimization.

- Restricted design rules have been receiving increased attention from industry (e.g., relaxed pitch helps to reduce CD asymmetry caused by coma aberration). We intend to evaluate such approaches with AberrPl in terms of design and manufacturability metrics.

IV.G Acknowledgments

This chapter is in part a reprint of:

- A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, “Lens Aberration Aware Placement for Timing Yield”, submitted to *ACM Transactions on Design Automation of Electronic Systems*, 2008.
- A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, “Lens Aberration-Aware Timing-Driven Placement”, *Proc. Design Automation and Testing in Europe*, 2006, pp. 890 - 895.

I would like to thank my coauthors Dr. Puneet Sharma, Dr. Qinke Wang, and Prof. Andrew B. Kahng.

V

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

V.A Introduction

As noted above, critical dimension (CD) variation is a dominant factor in the variation of delay and leakage current of transistor gates in integrated circuits. With advanced manufacturing processes, CD variation is worsening due to a variety of systematic variation sources at both within-die and reticle- or wafer-scale; the latter sources include radial bias of spin-on photoresist thickness, etcher bias, reticle bending, uniformity of wafer starting materials, etc. A statistical leakage minimization method is proposed in [16], which obtains significant reduction in total leakage by simultaneously varying the threshold voltage, gate sizes and gate lengths. [49] proposed to apply gate-length (CD) biasing only on the devices in non-critical paths for leakage power control without negative effects on timing.

A recent technology from ASML, called *DoseMapper* [56, 129], allows for optimization of ACLV (Across-Chip Linewidth Variation) and AWLV (Across-

Wafer Linewidth Variation)[1] using an exposure dose (or, simply, dose) correction scheme. DoseMapper in the ASML tool parlance exercises two degrees of control, *Unicom-XL* and *Dosicom* [101], which respectively change dose profiles along the lens slit and the scan directions of the step-and-scan exposure tool.

Today, the DoseMapper technique is used solely (albeit very effectively – e.g., [108]) to reduce ACLV or AWLV metrics for a given integrated circuit during the manufacturing process. However, to achieve optimum device performance (e.g., clock frequency) or parametric yield (e.g., total chip leakage power), not all transistor gate CD values should necessarily be the same. For devices on setup timing-critical paths in a given design, a larger than nominal dose (causing a smaller than nominal gate CD) will be desirable, since this creates a faster-switching transistor. On the other hand, for devices that are on hold timing-critical paths, or in general that are not setup-critical, a smaller than nominal dose (causing a larger than nominal gate CD) will be desirable, since this creates a slower-switching and less leaky transistor. What has been missing, up to now, is any connection of such "design awareness" – that is, the knowledge of which transistors in the integrated-circuit product are setup or hold timing-critical – with the calculation of the DoseMapper solution.[2]

In this chapter, we propose a novel method to enhance timing yield as well as reduce leakage power by combined dose map and placement optimizations. The contributions of our work are as follows.

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

- A new design- and equipment-aware dose map optimization (*DMopt*) method that uses dose to modulate gate poly CD across the exposure field, so as to optimize a function of delay and leakage power of the circuit.

[1] ACLV is primarily caused by the mask and scanner, while AWLV is affected by the track and etcher [109].

[2] Optimization of gate CDs according to setup or hold timing (non-)criticality has been used by [49]. What we propose below uses a coarser knob (i.e., the dose map) for design-aware CD control, but has the advantage of not requiring any change to the mask or OPC flows.

- A new dose map-aware placement optimization (*dosePl*) heuristic that considers systematic CD changes at different areas within a given dose map, and seeks to optimize circuit timing yield by selectively re-placing critical and near-critical cell instances based on golden extraction and timing analysis results.

Note that two distinct optimizations are possible, i.e., the place-ment-aware dose map optimization (*DMopt*) and the dose map-aware placement optimization (*dosePl*). While this chapter focuses mainly on *DMopt*, *dosePl* (discussed in the Appendix) is also attempted on a placement-aware timing and leakage optimized dose map. This chapter is organized as follows. Section V.B introduces fundamentals of the DoseMapper concept. Section V.C describes details of the design-aware dose map optimization. Section V.D discusses the overall optimization flow. Experimental results are presented in Section V.E.

V.B DoseMapper Fundamentals

Figure V.1 illustrates the intrafield DoseMapper concept. In Figure V.1, the slit exposure correction is performed by Unicom-XL. The actuator is a variable-profile gray filter inserted in the light path. The default filter has a second-order (quadratic) profile, and ASML [1] recommends use of a quadratic slit profile to model data in the slit direction. It is also possible to obtain a customized profile; lithography systems with Unicom XL (e.g., the XT:1700i machine) support a slit profile up to 8^{th} order in the dose recipe. Additionally, a maximum gradient constraint of 1% per mm at mask scale in the slit direction is applied in the ASML tool's CD Analyzer to calculate the dose recipe; this limits the correction range for higher-order corrections. Overall, a correction range of ±5% can be obtained with Unicom-XL for the full field size of 26mm in the X-direction.

Scan exposure correction is realized by means of Dosicom, which changes the dose profile along the scan direction. The dose generally varies only gradually during scanning, but the dose profile can contain higher-order corrections depend-

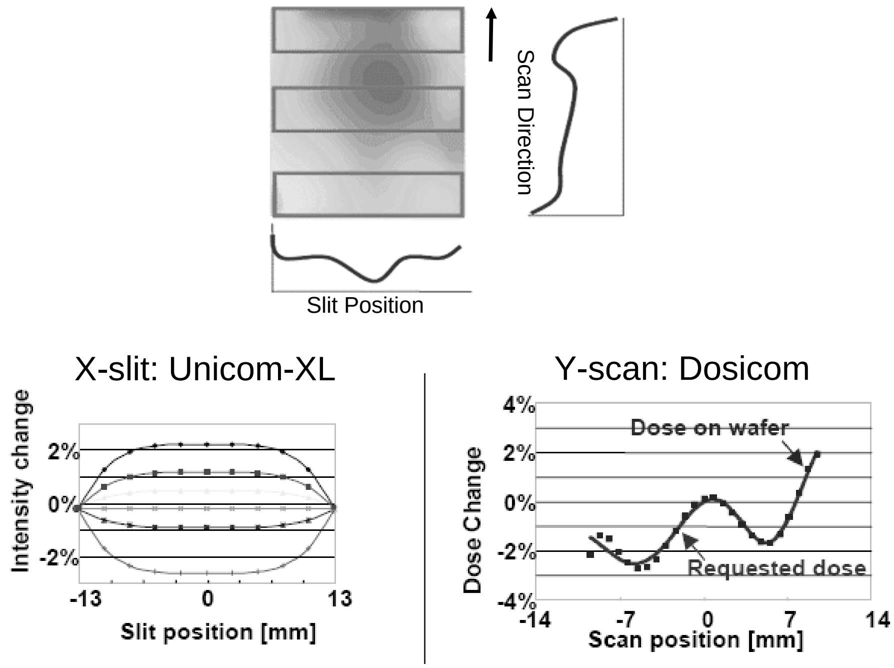


Figure V.1: Unicom-XL and Dosicom, which change dose profiles in slit- and scan-directions, respectively. Source: [2].

ing on the exposure settings. The *dose set*, $D_{set}(x)$, is used to model parameters for a dose recipe formed of Legendre polynomials (Legendre functions of the first kind) as

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

where L_n are Legendre coefficients and P_n are Legendre polynomials. Up to eight Legendre coefficients can be supported. The correction range for the scan direction is $\pm 5\%$ (10% full range) of the nominal energy of the laser. When the requested X-slit and Y-scan profiles are sent to the lithography system, they are converted to system actuator settings (one Unicom-XL shift for all fields, and a dose offset and pulse energy profile per field).

Dose sensitivity is the relation between dose and critical dimension, measured as CD [nm] per percentage [%] change in dose. Increasing dose decreases CD as shown in Figure V.2, i.e., the dose sensitivity has negative value. To calculate the dose sensitivity ($\triangle CD/\triangle E$, [nm/%]), a focus-exposure matrix (FEM) must

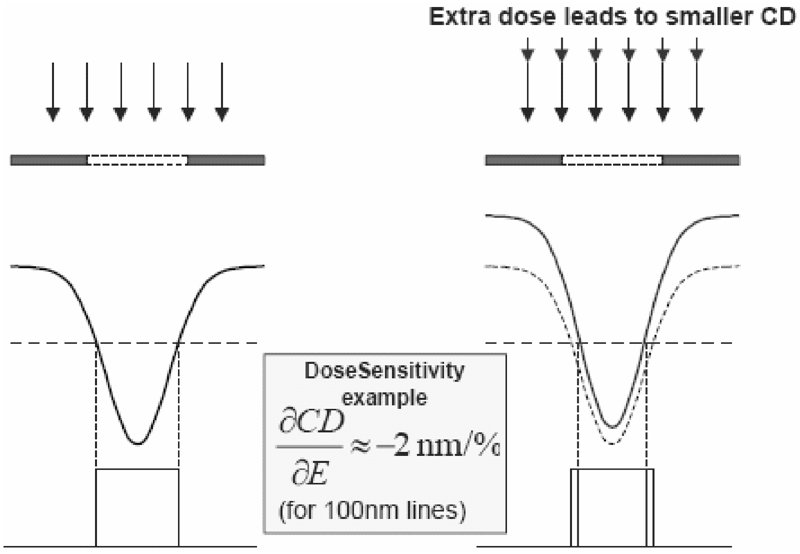

Extra dose leads to smaller CD
DoseSensitivity example
$\frac{\partial CD}{\partial E} \approx -2$ nm/%
(for 100nm lines)

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

be exposed on a product wafer for each product layer using standard production settings (e.g., reticle (6% attPSM), resist and illumination settings).

V.C Dose Map Optimization for Improved Delay and Leakage

V.C.1 Dose Map Optimization Problem

The design-aware dose map problem, for the objective of timing yield and leakage power, can be stated as follows. *Given placement P with timing analysis results, determine the dose map to improve timing yield as well as reduce total device leakage.*

In the following, for simplicity of exposition we assume that the reticle area taken by a single copy of the integrated circuit is the same as the area of the exposure field. In practice, the exposure field will contain one or more copies of the integrated circuit(s) being manufactured. It is simple to extend the proposed algorithms to the case where the exposure field contains multiple copies of the integrated circuit(s) being manufactured; smoothness or gradient constraints are scaled, and multiple copies of the dose map solution are tiled horizontally and vertically.

For the dose map optimization problem, we partition the exposure field into a set of rectangular grids $R = |r_{i,j}|_{M\times N}$ where the (uniform) width and height of rectangular grid $r_{i,j}$ are both less than or equal to a user-specified parameter G. G controls the granularity of the dissected rectangular grids: a smaller value of G corresponds to a larger number of rectangular grids, along with a more precisely specified new dose map and better timing yield improvement. However, G cannot be too small, because of current DoseMapper equipment limitations. G can be determined so as to balance between DoseMapper equipment constraints and timing yield improvement.

V.C.2 Circuit Delay Calculation

A typical dose sensitivity D_s at $\leq$ 90nm is -2nm/% [108]; we assume this value below in our experimental evaluations. Gate length changes linearly with dose tuning, i.e, $\Delta L_p = D_s \times d_{i,j}$, where ΔL_p is the gate length change of gate p and $d_{i,j}$ is a percentage value which specifies the relative change of dose in the rectangular grid $r_{i,j}$ wherein the gate is located.

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

When the gate length changes in a small range, the effects of the change

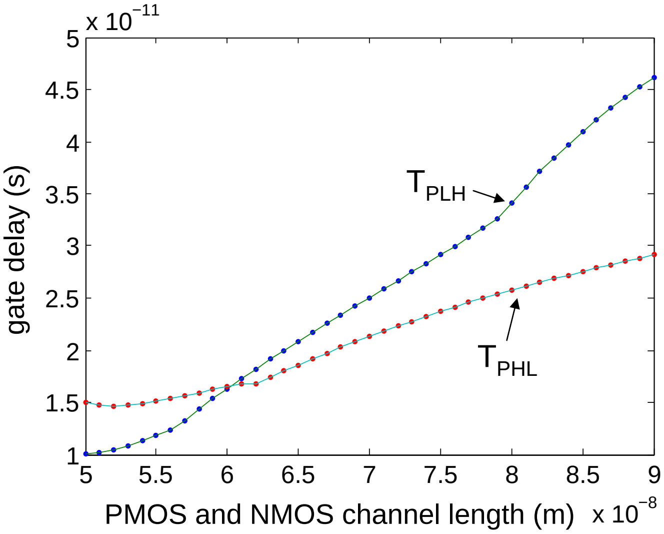


Figure V.3: Delay of an inverter vs. gate length.

on other topologically adjacent gates are typically small.[3] Hence, we assume that the gate delay increases linearly as the gate length increases. Since gate length increases linearly when the dose on the gate varies, there is a linear relationship between the change of gate delay and the change of dose on the gate, i.e, $\Delta t_p = t'_p - t_p = A_p \times \Delta L_p = A_p \times D_s \times d(r(p))$. Here, t_p and t'_p are the delay of gate p before and after the percentage dose change $d(r(p))$ in the rectangular grid $r(p)$ where gate p is located, ΔL_p is the change in gate length of gate p, and A_p is a fitted parameter that is dependent on input slew and load capacitance of each gate. In other words, for each distinct standard cell, and for each combination of input slew and load capacitance, a different value of A_p is obtained from preprocessing of Liberty nonlinear delay model tables. Total runtime of this procedure for an entire production standard cell library is less than an hour on a single processor.

For circuit delay calculation, without loss of generality we consider a combinational circuit with n gates as in [22]. Sequential circuits may be addressed similarly, e.g., by 'unrolling' them, using standard techniques, to combinational

[3] We recognize that off-path loading, slew propagation, and crosstalk timing windows can all change, and will be eventually accounted for precisely by golden signoff analysis. However, we assume in our optimization framework – as is fairly standard in the sizing literature – that these effects are negligible, and we validate our results with golden signoff analysis.

circuits that traverse from primary inputs and sequential cell outputs, to sequential cell inputs and primary outputs. For a given combinational circuit, we add to the corresponding circuit graph one fictitious source node, which connects to all primary inputs, and one fictitious sink node, which connects from all primary outputs. Nodes are indexed by a reverse topological ordering of the circuit graph, with the source and sink nodes indexed as $n + 1$ and 0, respectively.

V.C.3 Leakage Power Quadratic Approximation

For simplicity, we do not include dose-dependent change of wire delay in our problem formulation; note that a dose map optimization on the transistor gate layer of the IC will not affect wire pattern, and thus will not affect golden wire parasitics. In our implementation, wire delay is obtained from golden static timing analysis reports and added in between gates.

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

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

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

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

In our optimization, we assume that the change of leakage power of a gate is a quadratic function of gate length change[4], i.e, $\Delta P(\Delta L_p) = \alpha\times\Delta L_p+\beta\times(\Delta L_p)^2$

[4] We recognize that leakage power is exponential in gate length. We use a quadratic approximation to facilitate the problem formulation and solution method.

for gate p. Assume that the original dose in the chip area is uniform. The goal of the design-aware dose map optimization (*DMopt*) is to tune the dose map to adjust the channel lengths of the gates and thereby reduce a weighted sum of circuit delay and total leakage power, subject to upper and lower bounds on delta dose values per grid, and a dose map smoothness bound to reflect the fact that exposure dose must change gradually between adjacent grids.

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

$$\Delta P_{leakage} = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{p \in r_{i,j}} \alpha_p \times D_s \times d_{i,j} + \beta_p \times D_s^2 \times d_{i,j}^2 \qquad \text{(V.5)}$$

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

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

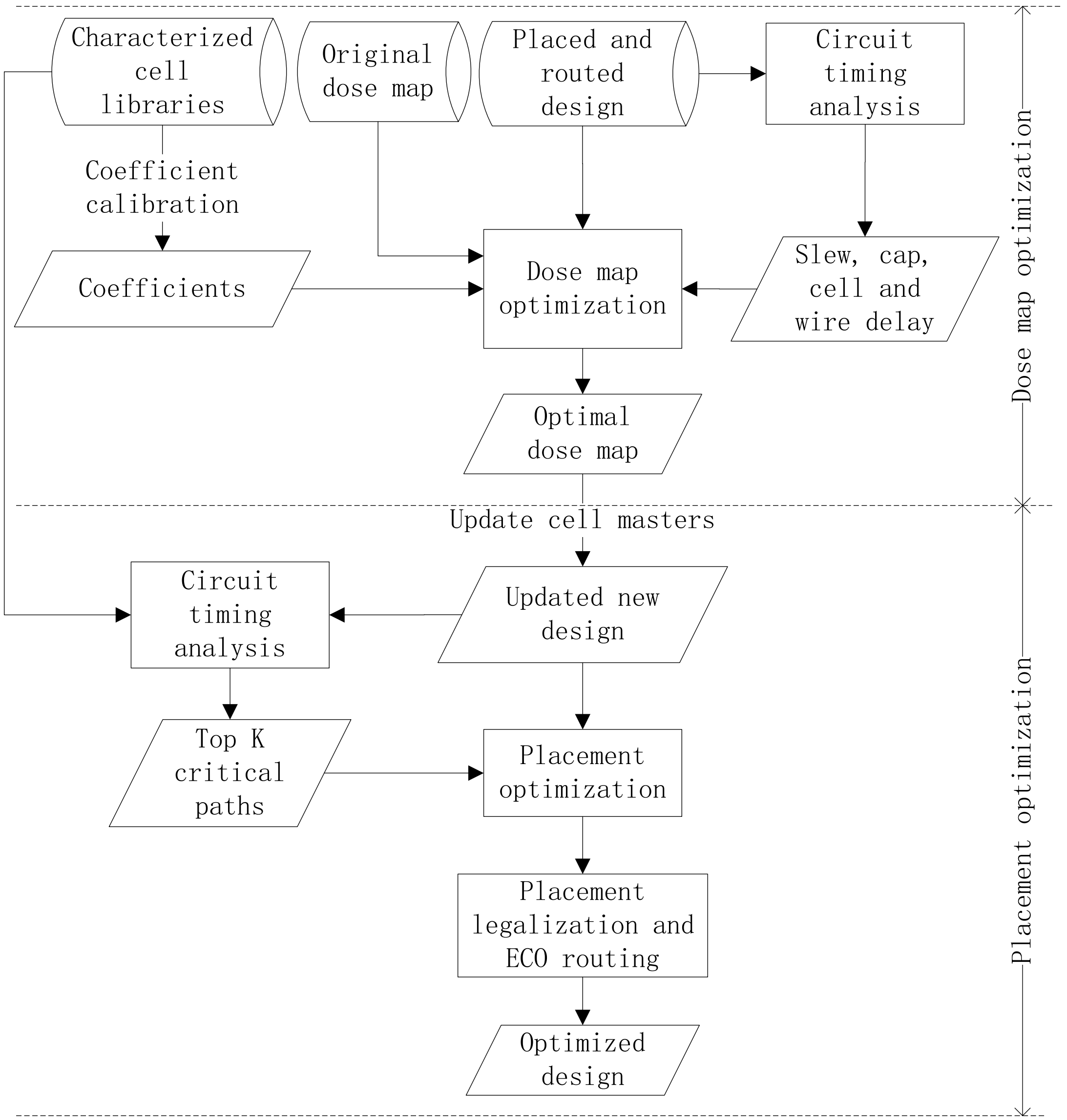


Figure V.4: Flow of the timing and leakage power optimization with integrated *DMopt* (top half) and *dosePl* (bottom half; details given in the Appendix).

V.D Timing and Leakage Power Optimization Flow

V.D.1 Overall Optimization Flow

Figure V.4 shows the whole flow integrating *DMopt* together with *dosePl* (discussed in the Appendix) for timing and leakage optimization. Note that the timing and leakage optimization flow is carried out after V_{th} and V_{dd} assignment processes. For the timing and leakage related dose map optimization problem, the input consists of the original dose map, the characterized standard cell timing libraries (or, other timing models that comprehend the impact of dose on transistor gate length) for different gate lengths, and the circuit with placement and routing information. By "placement and routing information", we also include implicit information that is necessary for timing and power analyses, notably, extracted wiring parasitics. With the nominal gate-length cell timing and power libraries, and the circuit itself with its placement, routing and parasitics data, timing analysis can be performed to generate the input slews and output load capacitances of all the cells. With the input slews and output load capacitances of all the cells, the original dose map, and characterized cell libraries of different gate lengths, our dose map optimization is executed to determine doses that adjust gate lengths of the cells for timing and leakage optimization, subject to dose map constraints. Finally, the optimal dose map is output.

According to the optimal dose map, the cell instances in different grids of the dose map will have different gate lengths as well as different cell masters in the characterized cell libraries, i.e., the design will be updated according to the dose map. With the characterized cell libraries, timing analysis is performed on the new design with the updated cell masters to identify the top K (e.g., $K = 10000$) critical paths for the complementary *dosePl* (see the Appendix) process to optimize. The *dosePl* process is based on a cell swapping strategy, which may introduce an illegal placement result. Therefore, a legalization process is invoked to legalize the swapped cells. ECO routing is then executed for the affected wires to refine the

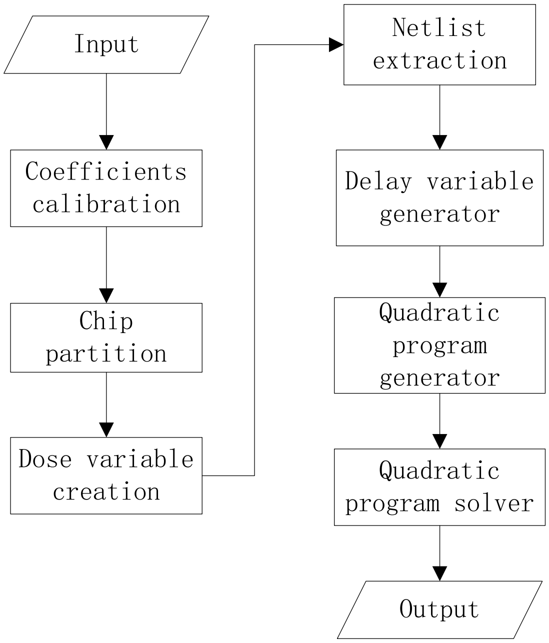


Figure V.5: Detailed view of design-aware dose map optimization flow.

design with optimized timing yield.

V.D.2 Summary of the Dose Map Optimization Flow

The dose map optimization in Figure V.5 is summarized as follows. The input consists of the original dose map, the characterized cell libraries of different gate lengths, and the input slews and output capacitances of all the cells in the circuit. From the characterized cell libraries of different gate lengths, the coefficients in the linear function of delay and the quadratic function of leakage power on gate length are calibrated. As noted above, when gate delay calculation in the cell libraries adopts a lookup table method, where the entries are indexed by input slews and output capacitances, the coefficients of the delay functions may be calibrated for each entry in each delay table. Then, according to the input slew and output capacitance values that were obtained for each cell in the previous step, the coefficients associated with the nearest entry (or, entries with interpolation) in the table are applied to calculate the delay of the cell.

The exposure field is then partitioned into rectangular grids. For each grid, a variable $d_{i,j}$ represents the percentage amount of dose change in the grid. Maximum circuit delay is captured using variables a_p that represent the arrival time at the output of each cell. When all the variables are obtained, a quadratic programming problem instance is generated by introducing the dose map correction range constraints, dose map smoothness constraints, and the delay constraints, as well as the objective of minimizing the weighted sum of circuit delay and total leakage power of all the cells. Finally, a quadratic programming solver solves the problem and finds the optimal dose change to each grid with respect to the original dose map; this yields the optimized dose map.

Table V.1: Characteristics of designs implemented in Artisan TSMC 90nm.

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

V.E Experimental Setup and Results

To assess the effectiveness of the proposed dose map optimization algorithm, we first sweep the dose change from -5% to $+5\%$ for all the rectangular grids in industrial testcase AES (shown in Table V.1) and perform timing analysis using Synopsys PrimeTime (v Z-2006.12) [13] and leakage power estimation using Cadence *SOC Encounter* (v 06.10) [4]. The timing analysis and leakage power estimation are based on pre-characterized cell libraries with gate length variants. Delay and leakage power results are given in Table V.2. “MCT” refers to minimum cycle time and “$P_{leakage}$” refers to the total leakage power of all the cells. Table V.2 shows that timing yield improvement can be obtained at the cost of leakage power increase, whereas leakage power reduction can be obtained at the cost of timing yield degradation. Uniform dose change in all the rectangular grids cannot obtain timing yield improvement without leakage power increase. However, our proposed dose map optimization algorithm can obtain substantial timing yield improvement with little or no increase in total leakage power.

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

Dose change	$d_{i,j} = 0$	$d_{i,j} = +0.5$	$d_{i,j} = +1$	$d_{i,j} = +1.5$	$d_{i,j} = +2$	$d_{i,j} = +2.5$	$d_{i,j} = +3$	$d_{i,j} = +3.5$	$d_{i,j} = +4$	$d_{i,j} = +4.5$	$d_{i,j} = +5$
MCT (ns)	1.990	1.971	1.950	1.932	1.905	1.893	1.868	1.845	1.818	1.791	1.758
imp. (%)	–	0.964	2.029	2.915	4.257	4.906	6.161	7.302	8.652	10.012	11.661
$P_{leakage}$ (uW)	2430.214	2546.756	2678.096	2824.598	2994.978	3180.969	3404.057	3654.222	3939.749	4253.778	4619.039
imp. (%)	–	-4.796	-10.200	-16.228	-23.239	-30.893	-40.072	-50.366	-62.115	-75.037	-90.067
Dose change	$d_{i,j} = 0$	$d_{i,j} =$ -0.5	$d_{i,j} =$ -1	$d_{i,j} =$ -1.5	$d_{i,j} =$ -2	$d_{i,j} =$ -2.5	$d_{i,j} =$ -3	$d_{i,j} =$ -3.5	$d_{i,j} =$ -4	$d_{i,j} =$ -4.5	$d_{i,j} =$ -5
MCT (ns)	1.990	2.011	2.031	2.057	2.078	2.093	2.115	2.135	2.155	2.172	2.188
imp. (%)	–	-1.031	-2.076	-3.359	-4.401	-5.155	-6.296	-7.257	-8.283	-9.142	-9.949
$P_{leakage}$ (uW)	2430.214	2324.525	2225.130	2135.234	2054.458	1980.457	1914.474	1850.809	1796.545	1746.507	1699.788
imp. (%)	–	4.349	8.439	12.138	15.462	18.507	21.222	23.842	26.075	28.134	30.056

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

AES	Nom L_{gate}	$\lambda = 1.0$		$\lambda = 1.1$		$\lambda = 1.2$		$\lambda = 1.3$		$\lambda = 1.4$		$\lambda = 1.5$	
		DMopt	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)
MCT (ns)	1.990	1.825	8.310	1.819	8.620	1.819	8.622	1.815	8.776	1.814	8.839	1.805	9.327
$P_{leakage}$ (uW)	2430.2	2350.8	3.270	2370.4	2.462	2382.2	1.975	2396.7	1.381	2424.9	0.217	2433.6	-0.138
Runtime (s)	–	232.667		239.900		310.786		270.844		127.617		141.948	

JPEG	Nom L_{gate}	$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.4$		$\lambda = 1.6$	
		DMopt	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)	*DMopt*	imp.(%)
MCT (ns)	2.906	2.776	4.480	2.726	6.172	2.703	6.968	2.687	7.536	2.673	8.007	2.670	8.124
$P_{leakage}$ (uW)	4354.2	3817.1	12.336	4009.5	7.916	4096.4	5.920	4182.1	3.953	4276.0	1.796	4343.2	0.252
Runtime (s)	–	1961.216		1670.229		2108.800		1946.394		1895.096		1835.088	

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

V.F Conclusions

We have proposed a novel method to improve the timing yield of the circuit as well as to reduce total leakage power, using design-aware dose map and dose map-aware placement optimization. Our discussion has focused mainly on the placement-aware dose map optimization. As discussed in the Appendix, we have also explored dose map-aware placement optimization on a placement-aware timing and leakage optimized dose map. Our proposed method is based on the fact that the exposure dose in the exposure field can change the gate/transistor lengths of the cells in the circuit, which is useful for optimization of gate delay and gate leakage power. Experimental results show more than 8% improvement in minimum cycle time of the circuit at no cost of leakage power increase. Our ongoing work includes the testing of the proposed dose map-placement co-optimization system on more testcases, especially on larger industrial 65nm designs.

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

V.G Appendix: Dose Map-Aware Placement

After a placement-specific dose map has been calculated, it is natural to ask whether a dose map-specific placement can further improve the result. In this appendix, we describe a simple cell swapping-based dose map-aware placement (*dosePl*) optimization. The *dosePl* problem can be stated as follows. *Given the original placement result and a timing and leakage-aware dose map, determine cell pairs which can be swapped to achieve maximum timing yield improvement.*

Cell-swapping based optimization. The basic idea behind the cell swapping-based optimization method is to swap cells on timing-critical paths (referred to as *critical cells* hereafter) to high-dose regions and non-critical cells to low-dose regions, to further enhance the circuit performance. We define the *bounding box of a cell* as the bounding box of all the cell's fanin cells and all of its fanout cells, as well as the cell itself. Our intuition is that moving a cell within its bounding box has lower likelihood of increasing total wire length or timing delay than moving it outside the bounding box. Thus, we seek pairs of cells $cell_l$ with bounding box b_l and $cell_m$ with bounding box b_m in different dose regions, such that $cell_l$ is in b_m and $cell_m$ is in b_l. With this restriction, we filter out candidate cell swaps that are too disruptive to wirelength and timing.

Additional heuristics to avoid wirelength increase. When two cells satisfy the condition that they are located in each other's bounding boxes, it is still possible for total wirelength to increase when the cells are swapped. We thus adopt the following heuristics to further filter out unpromising cell pairs.

(1) Distance between the two cells to be swapped. When the distance between two cells is very large, the impact of cell swapping on total wirelength is potentially large. Therefore, we avoid considering swaps of cells that are far apart.[7]

(2) HPWL-based (half-perimeter wire length) wire length comparison. We may also filter cell swaps by computing updated HPWL-based wirelength estimates;

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

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

AES	Nom L_{gate}	$\lambda = 1.0$				$\lambda = 1.2$				$\lambda = 1.5$			
		DMopt	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)
MCT (ns)	1.990	1.825	8.310	1.811	9.002	1.819	8.622	1.807	9.194	1.805	9.327	1.799	9.617
$P_{leakage}$ (uW)	2430.2	2350.8	3.270	2353.0	3.180	2382.2	1.975	2384.4	1.887	2433.6	-0.138	2435.1	-0.200
Runtime (s)	–	232.667		3.392		310.786		3.317		141.948		4.902	

JPEG	Nom L_{gate}	$\lambda = 0.6$				$\lambda = 1.2$				$\lambda = 1.6$			
		DMopt	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)	*DMopt*	imp.(%)	*dosePl*	imp.(%)
MCT (ns)	2.906	2.776	4.480	2.761	4.989	2.687	7.536	2.668	8.179	2.670	8.124	2.647	8.910
$P_{leakage}$ (uW)	4354.2	3817.1	12.336	3817.1	12.336	4182.1	3.953	4182.1	3.953	4343.2	0.252	4343.3	0.251
Runtime (s)	–	1961.216		173.849		1946.394		131.202		1835.088		86.832	

only if the estimated wirelength increase for all incident nets is below a predefined threshold (e.g., 20% in our experiments reported below) will the cell swap be attempted.

On the number of swaps and cell priority. For a given critical path, several cell swaps may suffice to reduce the path delay, and further cell swapping will introduce unnecessary wirelength and leakage increase. So, an upper bound on the number of cells swapped for each critical path is specified in our heuristic's implementation. The priority for a critical cell during swapping is decided according to the following two factors.

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

(2) Slack of critical paths. The larger the total path delay (= smaller slack) of a given critical path, the more important it is to swap cells on the path to achieve cell delay improvement. Therefore, higher priority is assigned to cells on paths with greater timing criticality.

Based on the above two heuristic factors, critical cells are assigned weights as calculated in Equation (V.6) where C_l is the critical paths where $cell_l$ is located. In our implementation, cells are processed path by path (obtained from golden timing analysis), in order from most timing-critical to least critical. Therefore, cells on more critical paths always have higher priorities than cells on less critical paths. Cells in the same critical path are sorted in non-increasing order according to their weights.

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

Pseudocode of the cell swapping heuristic: The pseudocode of our cell swapping heuristic is presented as Algorithm 1. The cell swapping process is based on the critical paths, which are first sorted in non-increasing order according to their total path delays. Cells of a given path are then swapped. Since it is not necessary

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

```
Find cells in top $K$ critical paths by golden timing analysis;
Compute weights for critical cells as in Equation (V.6);
Sort critical paths in non-decreasing order according to their slacks;
for $k = 1$ to $K$ do
  Sort the cells in critical path $c_l$ in non-increasing order according to their weights;
  for all cell $cell_l \in$ critical path $c_k$ do
    if # swapped cells in path $c_k$ $n(c_k) > \gamma_1$ then break; end if
    Compute bounding box $b_l$ of cell $cell_l$ in path $c_k$;
    Get the set of rectangular grids $R$ intersected with $b_l$;
    Sort the grids in $R$ in non-increasing order according to the dose $d(r)$ in each grid $r$;
    Set $flag \leftarrow$ false;
    for all $r \in R$ do
      if $d(r) < d(r(cell_l))$ then break; end if
      Sort the non-critical cells $NC$ in grid $r$ in non-decreasing order by Manhattan distance from $cell_l$;
      for all $cell_m \in NC$ do
        if $dis(cell_l, cell_m) > \gamma_2$ then break; end if
        if $cell_l \in b_m$ and $cell_m \in b_l$ and $\Delta HPWL(cell_l) < \gamma_3$ and $\Delta HPWL(cell_m) < \gamma_3$ then
          Swap $(cell_l, cell_m)$;
          Update the number of swapped cells $n(c_s)$ for all critical paths $c_s$ such that $cell_l \in c_s$;
          Set $flag \leftarrow$ true;
          break;
        end if
      end for
      if $flag =$ true then break; end if
    end for
  end for
end for
```

to swap all the cells in a critical path to improve the path's timing, the number of cells swapped for each path is recorded and the swapping process for a path is terminated when the number of swapped cells reaches a user-defined parameter γ_1 (in our experiments, up to $0.2 \times$ the cell count on the path). The swapping process checks the bounding box constraint, the dose constraint, and the distance between candidate swapping pairs, and computes HPWL-based wirelength increase when the pair is swapped. If a candidate pair passes all the checks, it is swapped and the corresponding critical paths are updated to record the increased number of swapped pairs. The cell swapping process continues until all critical paths are considered. When the swapping process finishes, the perturbed placement is legalized and routed by a standard placement tool's engineering change order (ECO) placement and routing functionality. After final ECO routing, golden timing analysis is performed with updated parasitics to evaluate the circuit delay improvement.

Experimental results

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

Figure V.6 shows the slack profiles of design AES, including (i) the original design, (ii) the design after dose map optimization ($\lambda = 1.5$, dose correction range is $\pm 5\%$, 20×50 rectangular grids), and (iii) the design when all the gates in the top 10000 critical paths are enforced using maximum possible dose (i.e., +5% on the original dose). The purpose of enforcing the maximum possible exposure dose on the critical gates is to find out the optimization headroom left after the *DMopt* process. From Figure V.6, the worst slack of the original design is improved significantly by dose map optimization process. But, this lessens opportunity for the following placement process. On the one hand, the difference between the worst slacks of the dose-optimized design and the biased design ("best" design)

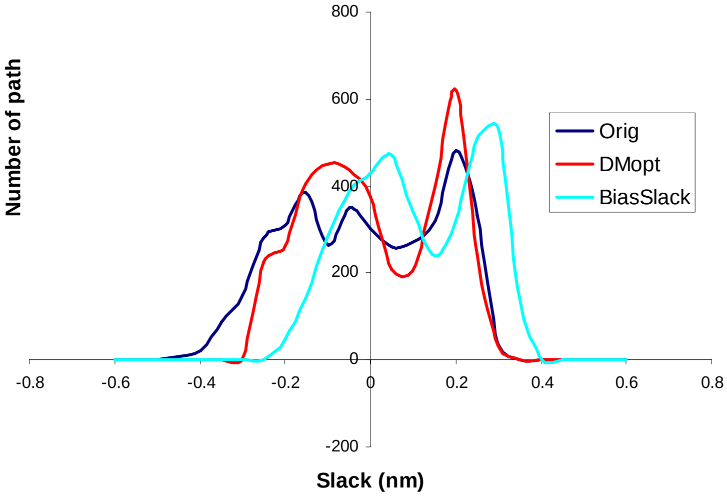


Figure V.6: Slack profiles of design AES before *DMopt* and after *DMopt*, and the biased design wherein all gates in the top 10000 critical paths are given maximum possible exposure dose (+5%).

is quite small (less than 0.05ns); on the other hand, in the dose map-optimized design, the number of critical paths, whose slack values are quite near the worst slack value, is large. To further improve timing, the placement process must swap many cells in order to address all the critical paths with slack values near the worst case. This can introduce many wire detours and make other non-critical paths become critical. In light of the relatively small opportunity left for *dosePl* process, the observed improvement confirms the effectiveness of the cell swapping based algorithm.[8]

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

V.H Acknowledgments

This chapter is in part a reprint of:

- K. Jeong, A. B. Kahng, C.-H. Park and H. Yao, "Dose Map and Placement Co-Optimization for Timing Yield Enhancement and Leakage Power Reduction", *Proc. ACM/IEEE Design Automation Conference*, 2008, to appear.

I would like to thank my coauthors Kwangok Jeong, Dr. Hailong Yao, and Prof. Andrew B. Kahng.

VI

Fast Dual Graph Based Hotspot Detection

VI.A Introduction

For optical lithography, manufacturability is roughly defined by the k_1 factor from the Rayleigh equation [82]. Beyond the 45nm CMOS technology node, even with use of a high-end optical exposure system such as immersion lithography with higher numerical aperture (NA), it is necessary to have a k_1 factor lower than 0.25. The primary risk posed by lower k_1 is the likely degradation of patterning fidelity and its impact on VLSI circuits performance and variability. Lower k_1 could decrease patterning fidelity and result in generation of many *hotspots*, i.e., actual device patterns which have relatively large CD and image errors with respect to on-wafer targets. Hotspots include a variety of pattern deformations, e.g., line-end pullback (shortening), corner rounding, necking, and bridging [119]. Pullback is shrinkage of geometries due to overdose at narrow line-ends. Necking is a reduction in linewidth that is induced by a hammerhead or neighboring wide line. We separate hotspots into *open faults* for necking and shortening, and *bridging faults* for corner rounding and bridging.

Under ultra-low k_1 conditions ($k_1 < 0.3$), in particular, many hotspots may arise anywhere. Hotspots can form under a variety of conditions such as the original design being unfriendly to the RET that is applied to the chip, pattern combinations unanticipated by rule-based OPC, or inaccuracies in model-based OPC. When these hotspots occur at locations that are critical to the electrical performance of a device, they can reduce the yield and performance of the device. It is therefore necessary to detect hotspots earlier in the layout design flow [26,51,90].

Park et al. [98] proposed a detection method for critical patterns (hotspots), using a design rule check (DRC) tool. The approach is a rule-based detection which generates lookup tables with line and space parameters. However, for more complex patterns, the number of layout pattern parameters required to enable detection increases. As a result, the speed advantage of the rule-based approach is reduced. On the other hand, the simulation-based approach has occupied the mainstream and has been able to detect hotspots accurately [27, 105]. Furthermore, software solutions running on customized hardware platforms have been developed so that aerial image simulation can be carried out quickly [18]. However, hotspots can change according to process conditions. Achieving required accuracy of hotspot detection strongly depends on qualified optical and process models. Model generation corresponding to process variation represents a significant overhead in terms of validation, measurement and parameter calibration.

For hotspot detection, there is typically only one golden physical verification signoff tool in the design flow, and even though analogous tools may be *qualified* for other junctures in the flow, there is little point in trying to replace the golden signoff tool. Rather, the golden signoff tool represents a runtime-quality tradeoff point that is high in quality, but also high in runtime. The objective of our work is to develop a low-runtime *pre-filter* that reduces the amount of layout area to be analyzed by the golden tool, without compromising the overall quality of hotspot-finding.

In this chapter, we describe a novel detection algorithm for hotspots induced by lithographic uncertainty (i.e., process window). Our approach utilizes a layout-derived graph which reflects pattern-related CD variation. Our goal is to detect a set of potential lithographic hotspots within minutes without degrading accuracy (i.e., without missing any actual hotspots). The key intuition behind our approach is that CD variation is generally the result of “bad” patterns or effects. A hotspot corresponds to a location with relatively large CD variation caused by several spatially proximate “bad” patterns. We assume that this effect is accumulative and the net effect of several spatially proximate “bad” patterns can be represented by the total weight of one merged face in the graph (or, merged node in the dual graph).

The main steps for bridging hotspot detection are as follows.

- **Layout Graph Construction.** Given a layout L, the *layout graph* $G = (V,\ E_c \cup E_p)$ consists of nodes V, corner edges E_c and proximity edges E_p. A face in the layout graph includes several close features and the edges between them. Edge weight can be calculated from a traditional 2-D model or a lookup table.

- **Graph Planarization.** For any two crossing edges in G, we delete the edge with smaller weight.

- **Three-Level Hotspot Detection.** (1) *Edge-level* detection finds the hotspot caused by two close features or “L-shaped” features; (2) *face-level* detection finds the pattern-related hotspots which span several close features; and (3) *merged-face-level* detection finds hotspots with more complex patterns. That is, we construct the dual graph G^D of the layout graph G (which correspond to faces of G) and sort the dual nodes according to their weights. We merge the sorted dual nodes that share the same feature, in sequence.

A local pattern density based hotspot filter is used to reduce the number of falsely detected bridging hotspots. We use normalized image log slope (NILS) to evaluate quality of pattern and susceptibility of the hotspot pattern to focus and exposure errors. The mask error enhancement factor (MEEF) for dense pitches is in general higher than for isolated pitches. Higher MEEF causes lower NILS [100,128], hence the lower NILS is with dense pitches, the higher the probability of a bridging hotspot [67]. As a result, the use of a pattern density filter may improve detection accuracy for bridging hotspots[1]. Necking hotspot detection is done by comparing the total weight of each dual node with a given threshold value.

The remainder of this chapter is organized as follows. In Section VI.B, we describe the problem formulation, our dual graph based hotspot detection algorithm, and implementation details. Section VI.C presents our evaluation flow and experimental results. We conclude in Section VI.D with directions for ongoing and future research.

VI.B Dual Graph Based Hotspot Detection

Recall that *hotspots* are the locations in the design where the magnitude of edge displacement is exceptionally large. In other words, hotspots are printed features whose CD variations are greater than a given threshold value.

VI.B.1 Problem Formulation

We formulate the fast hotspot detection problem as follows. **Hotspot detection problem**

Given: Layout L, and threshold of CD variation which defines a hotspot.
Detect: Hotspots which may result in large CD variation.
To Minimize: The number of undetected hotspots and falsely detected hotspots.

The basic function for detection depends on process variations (i.e., defocus and exposure) and pattern parameters (i.e., width and space). To reduce the number of process conditions for hotspot validation, the effects of pattern complexity must be comprehended. Figure VI.1 shows patterns with three different complexities: (a) one wide metal line, (b) two wide metal lines, and (c) four wide metal lines. Figure VI.2 shows that different pattern complexities lead to differ-

[1] While a dense pattern has higher likelihood of a hotspot, the local wiring density filter cannot filter out all hotspots since the hotspot is also a complex function of the distance between two features, overlapped projection length, the widths of the two lines, etc.

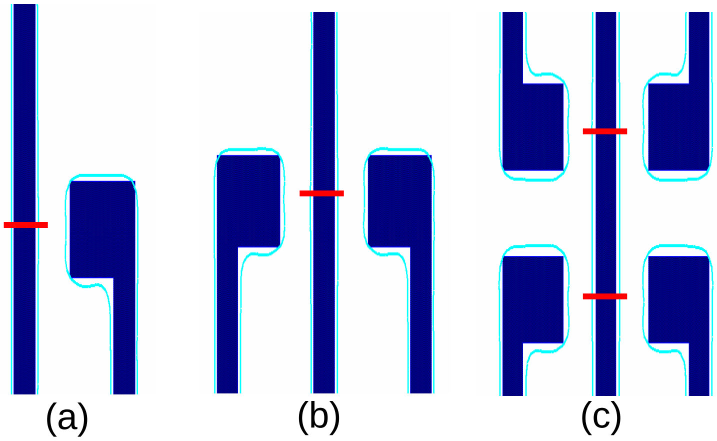


Figure VI.1: Test patterns to evaluate CD variation induced by pattern complexity. Red lines show CD measurement location. Blue contours represent the simulation results at worst-case defocus. Two CD values are averaged in the case of (c).

ent CD variations. Three process conditions, C-1, C-2 and C-3 are evaluated, respectively corresponding to NA $= 0.85$ and $\sigma = 0.96/0.76$; NA $= 0.75$ and $\sigma = 0.75/0.55$; and NA $= 0.75$ and $\sigma = 0.75/0.45$. In the figure, the labels a, b and c respectively correspond to the patterns (a), (b) and (c) shown in Figure VI.1. Two CD values are averaged in the case of (c). The CD variation may also be affected by different process conditions. However, the patterns with more complex configuration, e.g., (c) in Figure VI.1, have larger CD variation than the patterns with simple complexity, e.g., (a) in Figure VI.1, at all process conditions. Therefore, our key observation is that the higher the pattern complexity, the higher the probability of a hotspot.

We propose to use a cost derived from our graph based approach, which is described in detail in Section VI.B.2, to represent the pattern complexity and hence the hotspot probability. For bridging hotspots, Figure VI.2 shows that this cost can track the hotspot probability well according to the change of pattern complexity. Figure VI.3 plots average cost versus CD for necking hotspots. For this example, features with printed CD smaller than 80nm are viewed as necking hotspots. We can see that our proposed cost has good correlation with printed CD. The graph-derived cost is thus closely related to hotspot probability.

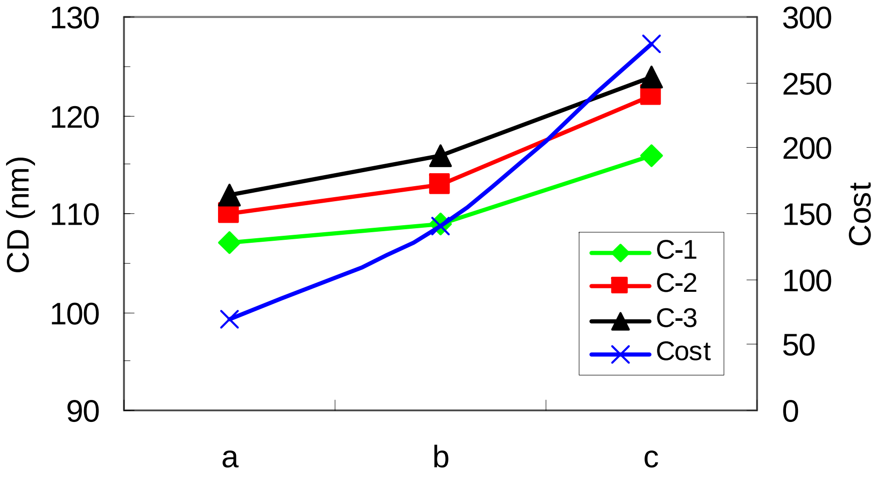


Figure VI.2: Evaluation of CD variation and cost in the graph-based approach with various test patterns and process conditions. C-1: NA = 0.85 and $\sigma = 0.96/0.76$. C-2: NA = 0.75 and $\sigma = 0.75/0.55$. C-3: NA = 0.75 and $\sigma = 0.75/0.45$.

We now propose a new graph-based hotspot detection method which is very fast and accurate. We group pattern-induced bridging type CD variations into three cases.

1. **Corner induced CD variation.** As shown in Figure VI.4(a), two orthogonal connected features form a corner which may lead to corner rounding CD variations.

2. **Proximity induced CD variation.** As shown in Figures VI.4(b) and (c), two close features may lead to bridging CD variations.

3. **Line end induced CD variation.** One line end may lead to bridging CD variations. This effect can be treated as a special case of proximity induced CD variation.

Pattern-induced necking (open) type CD variations can also be grouped into three cases.

1. **Wide line induced variation.** As shown in Figure VI.5(a), one wide line tapering to a thin line may lead to necking CD variations.

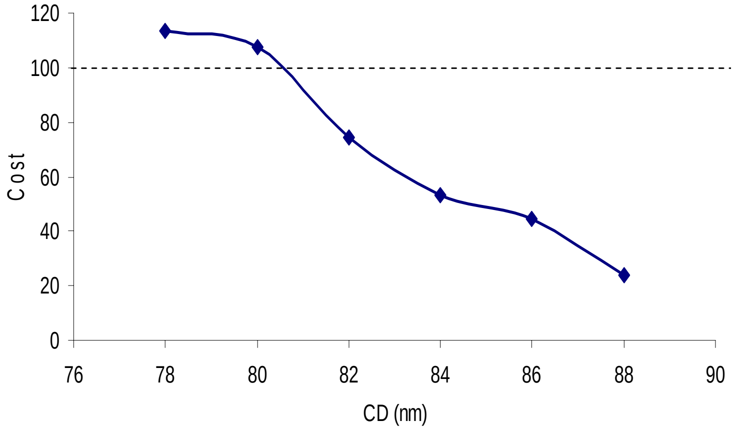


Figure VI.3: A plot of average cost vs. CD.

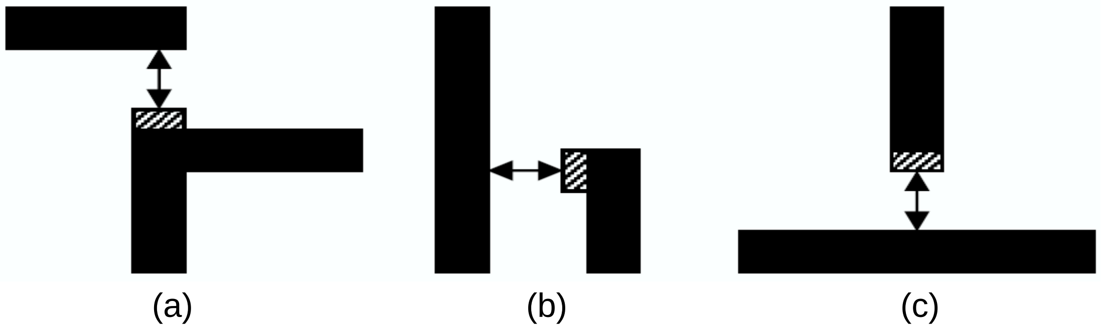


Figure VI.4: Effects leading to bridging hotspots: (a) corner-induced CD variation, (b) proximity-induced CD variation, and (c) line end-induced CD variation.

2. **Line end induced CD variation.** As shown in Figure VI.5(b), the line end of one of two parallel features may lead to necking CD variations.

3. **Wide line proximity induced CD variation.** As shown in Figure VI.5(c), one wide line close to a thin line may lead to necking CD variations.

In lithography, a given hotspot may be the result of a single effect as shown in Figure VI.6(a), or the combination of several effects in an accumulative way as shown in Figures VI.6(b) and VI.6(c). The accumulative property of hotspots makes detection and filtering very difficult. In our approach, we try to capture this accumulative effect with an iterative merging process. Recall the three steps of our proposed bridging hotspot detection flow:

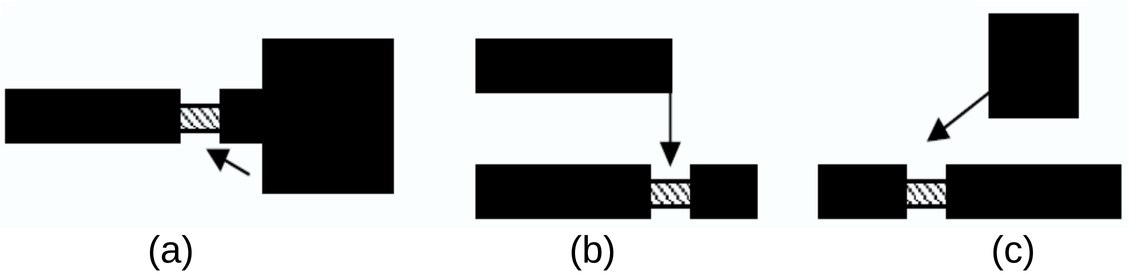


Figure VI.5: Effects leading to necking (open) hotspots: (a) wide line-induced variation, (b) line end-induced CD variation and (c) wide line proximity-induced CD variation.

- Layout Graph Construction. For given layout L, construct the layout graph $G = (V, E_c \cup E_p)$ with nodes V, corner edges E_c and proximity edges E_p.
- Graph Planarization. For any two crossing edges, delete the one with smaller weight.
- Three-Level Hotspot Detection. Perform edge-level (Figure VI.6(a)), face-level (Figure VI.6(b)) and merged-face-level (Figure VI.6(c)) detection to find hotspots with complex patterns.

For the necking hotspot detection, the total weight of each node is compared to the threshold value.

Intuitively, the single effects of "bad" patterns are represented by the weight of one edge and the accumulative effect of several closely-related effects is represented by the total weight of a merged face which includes several connected edges.

We present each step in detail in the following sections.

VI.B.2 Layout Graph Construction for Bridging Hotspots

To quickly detect the hotspots, the first step of our algorithm is to build a *layout graph* which reflects the pattern-related CD variation. As shown in Figure VII.5, given a layout L, the layout graph $G = (V, E_c \cup E_p)$ consists of nodes V, corner edges E_c and proximity edges E_p.

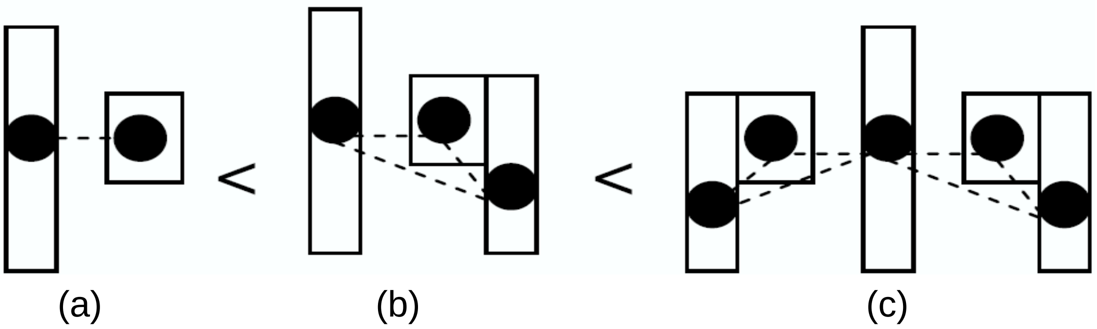


Figure VI.6: Effects leading to hotspots: (a) edge-level, (b) face-level and (c) merged face-level.

1. Every horizontal or vertical line is divided into line segments whose length is smaller than a threshold value l_0. For each line segment, create a node $v \in V$ located in the middle of the line segment.

2. For two orthogonal connected lines, connect two corresponding nodes with a corner edge $e \in E_c$ whose weight is a constant w_c.

3. Create a proximity edge $e \in E_p$ between two closely proximate lines having the same direction, where the weight of the edge is a function of the separation distance, overlapped projection length and the widths of the two lines. Since the line end effect and all necking effects are special proximity induced effects, we use the proximity edges for these effects with different weighting functions.

Figure VI.8(a) shows an example of a layout graph for the layout. The layout graph has 9 nodes representing 9 lines, 3 corner edges (dashed edges) and 10 proximity edges (solid edges).

One crucial issue is the edge weighting scheme. We propose both closed-form formula based and lookup table based weighting schemes. In the closed-form formula scheme, we assume that the weights of corner edges are a constant c. As shown in Figure VI.9, the weights of the proximate edges are given by $\frac{w_1 \times w_2 \times f(l)}{d \times d}$, where w_1 and w_2 are the widths of the two features, and d is the distance between the two features. $f(l)$ is function of the length of the overlapped projection l, where

```
Input: Layout $L$, $\epsilon$
Output: $G = (V, E_c \cup E_p)$
For all line
    If (length $> l_0$)
        Divide the line into segments of length $< l_0$
For all line segments create a node $v$
For any two orthogonal connected line segments
    connect the two nodes with a corner edge $e \in E_c$
For any two closely proximate line segments
    connect the two nodes with a proximity edge $e \in E_p$
```

Figure VI.7: Layout graph construction.

$f(l) = 100$ if l is between -50nm and 300nm and $f(l) = 0$ otherwise, as shown in Figure VI.10. This means that the hotspot can occur within a particular distance between corners of two features. Empirically, the wire-bridging occurs near the line end due to OPC correction. Therefore, we use a simple model in which the proximity effects only exist when there is small overlap between two lines. In addition, proximity effects are intuitively more obvious for larger width features and smaller distance. In a lookup table based weighting scheme, the weights of the proximate edges are determined by feature widths, spacing, and length of the overlapped projection. Although the lookup table based weighting scheme is more accurate, it also brings overhead in parameter tuning. In this chapter, we use only the closed-form formula based weighting approach.

VI.B.3 Layout Graph Construction for Necking Hotspots

The layout graph G construction for necking hotspots is shown as follows.

1. Every horizontal or vertical line is assigned a node $v \in V$ located in the middle of the line.

2. For any wide line crossing a thin line, connect two corresponding nodes with

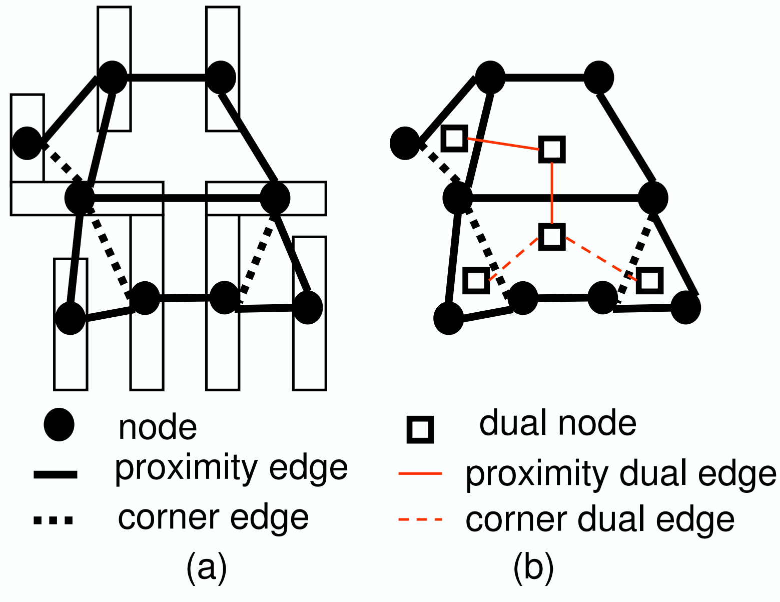


Figure VI.8: Example of (a) layout graph and corresponding (b) dual graph.

an edge $e \in E$ whose weight is a constant w_0.

3. For any two parallel thin lines next to each other, create an edge $e \in E$ whose weight is a constant w_1.

4. For any wide line close to a thin line, connect two corresponding nodes with an edge $e \in E$ whose weight is a constant w_2.

5. For any two nodes v_1 and v_2 which connect to the same node v, connect v_1 and v_2 with a zero-weight edge $e \in E$.

The purpose of the last step is to ensure that there is always a face containing two neighboring edges of a given node.

VI.B.4 Dual Graph Generation

The next step is to convert the layout graph $G = (V, E_c \cup E_p)$ into its *dual graph* $G^D = (V^D, E_c^D \cup E_p^D)$. One fact is that the dual graph G^D exists if G is a *planar* graph, i.e., there is no crossing edge. Therefore, as shown in Figure VI.11,

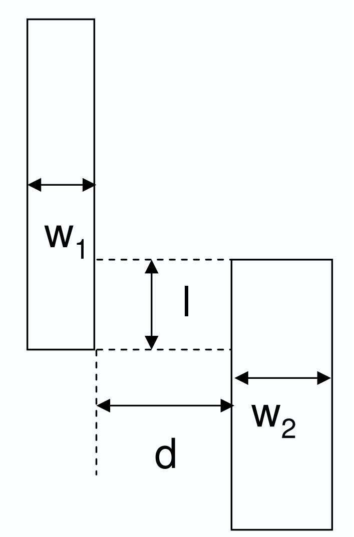


Figure VI.9: Weights of the proximate edges.

we must delete the edge with smaller weight for any two crossing edges in G. In practice, the impact of deleted edges is negligible since a deleted edge has smaller weight, which implies smaller CD variation effects. Also, the number of deleted edges is relatively small since the edges are always added between neighboring or touching features and it is unusual to have crossing edges. For our testcases, the percentage number of deleted non-zero-weight edges is below 0.1%.

The dual graph G^D of the layout graph G is constructed by representing every face f of G with a dual node n whose weight is equal to the sum of the edge weights of f. An edge e which belongs to faces f_1 and f_2 in G is represented by a dual edge $e^d = \{n_1, n_2\}$ in G^D having the same weight as e.

We calculate the total edge weight for each node in Lines 2-4 of the figure. A node is selected as a candidate hotspot if its total weight is greater than a threshold value ϵ. While we used the total edge weight of a face (i.e., a dual node in the dual graph) for bridging hotspots, since bridging hotspots are related to two or more features, for necking hotspots we just use the total edge weight of a node since the necking hotspots are located in one feature (node). Hence, we do not need to construct the dual graph for necking hotspots. The time complexity is dominated by the graph construction, which is $O(n)$.

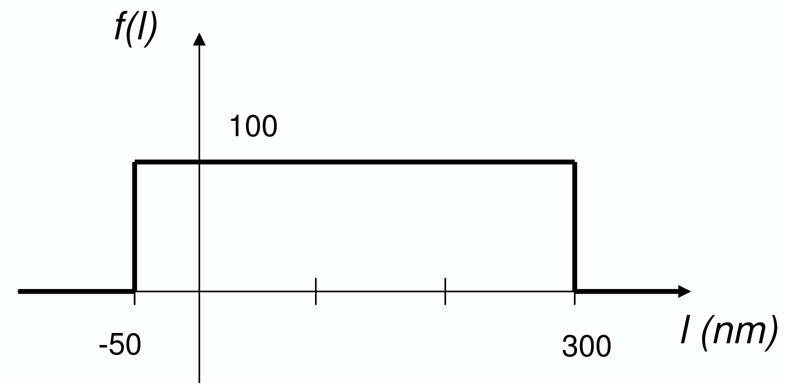


Figure VI.10: Function of the length of the overlapped projection l.

```
Input: $G = (V, E_c \cup E_p)$
Output: Dual graph $G^d$
For any two crossing edges in $G$
    Delete the edge with smaller weight
Represent each face in G with a dual node
Represent each edge in $G$ with a dual edge
```

Figure VI.11: Dual graph generation.

VI.B.5 Three-Level Hotspot Detection

The intuition behind our hotspot detection method is that a hotspot is the result of the combination of several proximate "bad" patterns. With the assumption that the CD variation effect is cumulative, the effect can be reflected by the dual node weight, i.e., the total edge weight of one face in the layout graph. However, a hotspot may also relate to the lines of several faces. Therefore, we need to consider dual node merging to capture all possible hotspots. Our proposed iterative dual node merging heuristic is shown in Figure VI.12. The heuristic starts with the layout graph G construction in Line 1. We perform edge-level detection in Lines 2-4. We then delete one edge with smaller weight for any pair of crossing edges to make G a planar graph, and construct the dual graph G^D from G. In Lines 7-9, we perform face-level hotspot detection. Finally, we perform merged-face-

```
Input: Layout $L$, $\epsilon_0$, $\epsilon_1$, $\epsilon_2$, $d_0$
Output: A list of hotspots in $L$
Construct layout graph $G$ from $L$, $S \leftarrow \emptyset$
For all edges $e$ whose weight $> \epsilon_0$
    // edge-level detection
    $S \leftarrow S \cup \{e\}$
    Delete $e$ from $G$
Perform graph planarization to delete one of crossing edges
Construct dual graph $G^D$ from $G$
For all dual nodes $n$ whose weight $> \epsilon_1$
    // face-level detection
    $S \leftarrow S \cup \{n\}$
    Delete $n$ from $G^D$
While ($\exists$ dual nodes that can be merged)
    // merged-face-level detection
    Sort all dual nodes according to weight
    Sequentially, merge each node with all spatially
    adjacent dual nodes
    If (weight of the merged node $n_c > \epsilon_2$)
        $S \leftarrow S \cup \{n_c\}$
        Delete $n_c$ from $G^D$
For all hotspots in $S$
    If (local wiring density $< d_0$)
        Remove it from $S$
```

Figure VI.12: Iterative dual node merging heuristic for bridging hotspots.

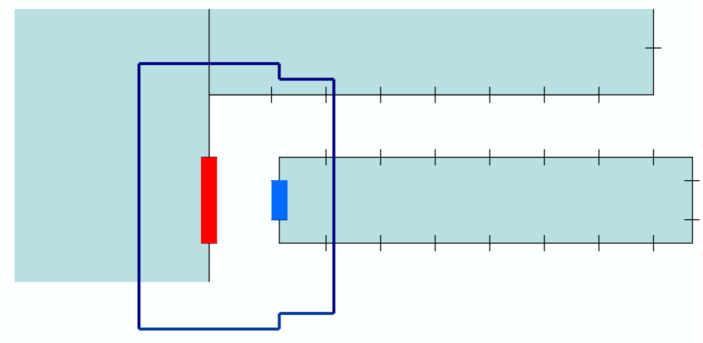


Figure VI.13: Example of a layer marking hotspot patterns.

level detection by sorting dual nodes according to their weights and sequentially merging spatially adjacent dual nodes (i.e., the dual nodes connected with dual edges). Intuitively, nodes with larger weight represent locations with higher CD variation. The weight of the merged node is equal to the sum of dual node weights minus the dual edge weight[2]. The purpose of deleting found hotspots in Line 4, 9 and 15 is to eliminate redundant hotspot detection. Since the simulation window (e.g., $4 \times 4\mu$m) for each hotspot is large enough to cover any possible neighboring hotspots, it is not necessary to include any spatially close hotspots in the final hotspot set. A local wiring density based hotspot filter is used to reduce the number of falsely detected hotspots. In this filter, we first find the center of a given hotspot, and then the local wiring density is the density within the box of 1μm $\times$ 1μm around the center. After the hotspot detection, a bounding box which covers all features in the edge/faces is drawn as the hotspot marker.

The time complexity for graph construction is $O(n)$, where n is the number of features. The edge-level hotspot detection (Lines 2-4) is $O(n)$. Dual graph generation time is $O(m\ log\ m)$, where m is the number of edges. The face-level hotspot generation (Lines 7-9) time is $O(k)$, where k is the number of faces. The merged face-level hotspot generation (Lines 10-15) runtime is $O(k\ log\ k)$. The density-based filter time is linear with respect to the number of detected hotspots.

[2] In our current implementation, we use a simple model which assumes the accumulative effect can be represented by the sum of weights. More complicated models such as weighted-sum may lead to better solution quality, but, will require more parameters to be extracted and tuned.

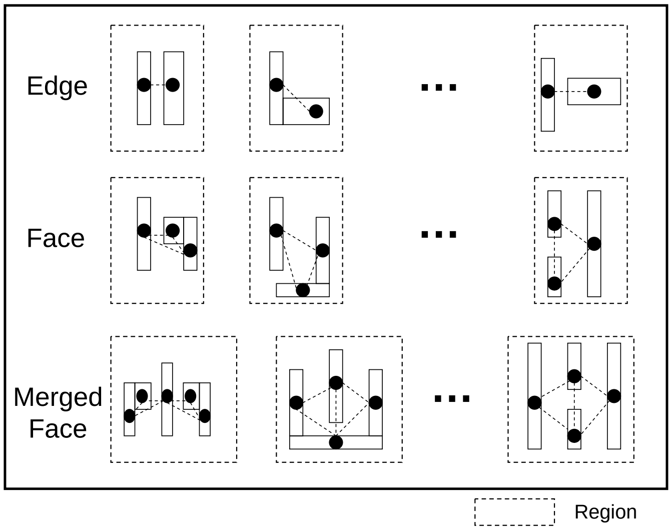


Figure VI.14: Illustration of a test layout used to calibrate our model.

VI.B.6 Model Parameter Extraction

One key issue with our proposed method is how to determine the model parameters. We use the following steps to build the model.

1. Create a test layout as illustrated in Figure VI.14 that includes a set of isolated regions. Within each region, instantiate a pattern with varying linewidth, line space and overlap length.

2. Construct the dual graph for the test layout. Each isolated region is marked as edge, face or merged face and the total weight for each region is calculated.

3. Run golden simulation tools on the test layout to detect all hotspots.

4. ϵ_0 is chosen as the smallest weight among all of the edge regions which have one or more hotspots. The values of ϵ_1 and ϵ_2 can be similarly determined. d_0 is the smallest wiring density of all regions having hotspots.

VI.C Experimental Results

We empirically test our approach on several real designs within a standard industry flow using leading-edge tools. We measure the number of truly detected hotspots and falsely detected hotspots, relative to area and runtime.

VI.C.1 Experimental Setup

For the evaluation according to various process conditions as shown in Table VI.1, we use one benchmark design in our experiments which is the ALU core with 8.7K instances from Artisan libraries in a 90nm technology using Synopsys *Design Compiler* (v 2003.06-SP1) [12]. The chip size is 335μm × 285μm. The synthesized netlist is placed with row utilization of 70% using Cadence *SOC Encounter* (v 3.3) [4]. The netlist of the design is from *OpenCores* [11]. For the validation according to various metal layers as shown in Table VI.2, we use five benchmark designs from major chip makers in a 65nm technology. We have implemented our proposed iterative dual node merging heuristic in C++.

On the lithography side, *CalibreOPC* and *CalibreORC* from Mentor Graphics *Calibre* (v 9.3_5.11) [9] are used for model-based OPC and optical rule check (ORC), respectively. For a 90nm design, simulation is performed with wavelength λ = 193, numerical aperture NA = 0.75, and annular aperture σ = 0.75/0.50. We use 0μm DOF model and 0.35 aerial image threshold for OPC, then evaluate the OPC'ed layer under the various values of DOF and threshold. For five 65nm designs, simulations are performed with wavelength λ = 193, numerical aperture NA = 0.85, and annular aperture σ = 0.96/0.76. OPC and ORC are performed with 0μm DOF model and 0.28 aerial image threshold, and with 0.1μm DOF model and 0.30 aerial image threshold, respectively.

VI.C.2 Experimental Results

We use a layout sizing technique to mark the hotspots and compare simulation-based detection with our dual graph based detection. Simulation-based detection makes fragments out of a given pattern and decides whether each fragment is a hotspot, based on magnitude of edge displacement. As a result, there may be several marked layers at a line-end and/or at a corner of a pattern. On the other hand, our graph-based detection marks all patterns affecting hotspots,

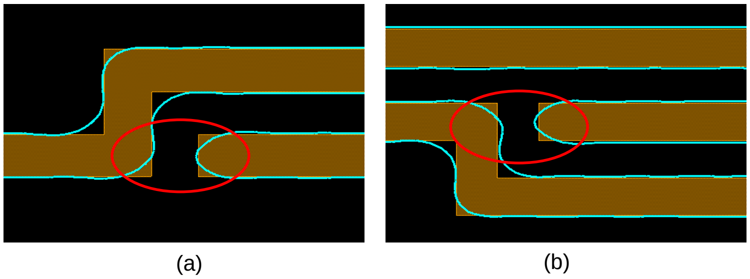


Figure VI.15: Results of hotspot detection: comparison of (a) no hotspot patterns versus (b) hotspot patterns.

and hence it is difficult to compare hotspots of simulation-based and graph-based methods. We size all layers marked as a hotspot after ORC by $0.5\mu m$[3]. Then all sized layers are merged into one layer which includes hotspots and is used for comparison. Figure VI.13 shows an example of a hotspot marking layer which is the result of merging of two layers in the ORC result.

We also notice that bridging hotspots depend on the local pattern density. To reduce the number of falsely detected bridging hotspots, a filter based on local pattern density has been used. Figure VI.15 shows the results of bridging hotspot detection; (a) no hotspot pattern and (b) hotspot pattern. The pattern in a denser region is a bridging hotspot while the same pattern in a sparse region is not a bridging hotspot. The results of the dual graph-based method match the simulation-based method. In addition, we show an example of necking hotspot detection in Figure VI.16. The necking hotspot causes a reduction in the linewidth due to combined effect of pullback and corner-rounding at the wide line. The dual graph-based method can thus detect hotspots induced by pattern density and wide lines, which cannot be achieved by the rule-based approach.

For a 90nm design, we evaluated the hotspot detection under four different process conditions. The number of hotspots can increase with higher threshold (exposure dose) and defocus. The values of ϵ_0, ϵ_1, ϵ_2 and d_0 in Figure VI.12 are chosen according to different conditions as shown in Table VI.1. We can see that only the value of ϵ_2 need change if only ET is changed, while we need to change

[3] The sizing amount is reasonable when we consider the proximity range of $0.6\mu m$ at 90nm.

Table VI.1: Result of bridging hotspot detection for 90nm testcases. Runtime of our method includes sum of graph generation, hotspot detection and hotspot report. 100% hotspot detection is achieved with few falsely detected hotspots.

Simulation	# Hotspots				Runtime (sec)		Parameters			
Condition	ORC	Detected	Falsely Detected	Rel. Area(%)	ORC	Dual Graph	ϵ_0	ϵ_1	ϵ_2	d_0
Defocus(μm)=0.1, ET=0.36	17	17	13	0.5	690	1.37	0.7	0.36	0.67	0.17
Defocus(μm)=0.1, ET=0.37	21	21	22	0.72	690	1.52	0.7	0.36	0.64	0.17
Defocus(μm)=0.1, ET=0.38	25	25	46	1.19	690	2.32	0.7	0.36	0.625	0.17
Defocus(μm)=0.2, ET=0.38	152	152	1291	24.18	690	4.38	0.65	0.34	0.537	0.14
Average	53.75	53.75	343	6.65	690	2.4				

Table VI.2: Comparison of hotspot detection efficiency of ORC and our proposed method at 65nm node. Filtered ORC + DG represents the total runtime, which is the sum of runtimes for dual-graph based detection (DG) and ORC for the hotspot area filtered by our DG.

Testcase	# instance	Layer	Bridging HS				Open HS				Runtime (sec)		
			ORC	Det.	False	Rel. Area (%)	ORC	Det.	False	Rel. Area (%)	ORC	DG	Filtered ORC + DG
1	5M	M2	8	8	328	0.13	6	6	7	0.005	50823	125	194
		M3	23	23	1127	0.46	440	440	158	0.23	50912	328	679
		M4	0	0	0	0	29	29	3	0.01	40431	19	23
2	6M	M2	4	4	697	0.28	4	4	4	0.003	51840	91	238
		M3	1	1	106	0.04	114	114	61	0.07	51807	85	142
		M4	0	0	0	0	44	44	8	0.02	45644	20	29
3	4M	M3	0	0	0	0	4	4	2	0.002	42761	5	6
4	42K	M2	10	10	125	14.6	0	0	0	0	964	2	143
5	27K	M2	1	1	14	5.8	0	0	0	0	674	1	40
Average			5.22	5.22	266.44	2.368	71.22	71.22	27.0	0.038	37317.3	75.1	166

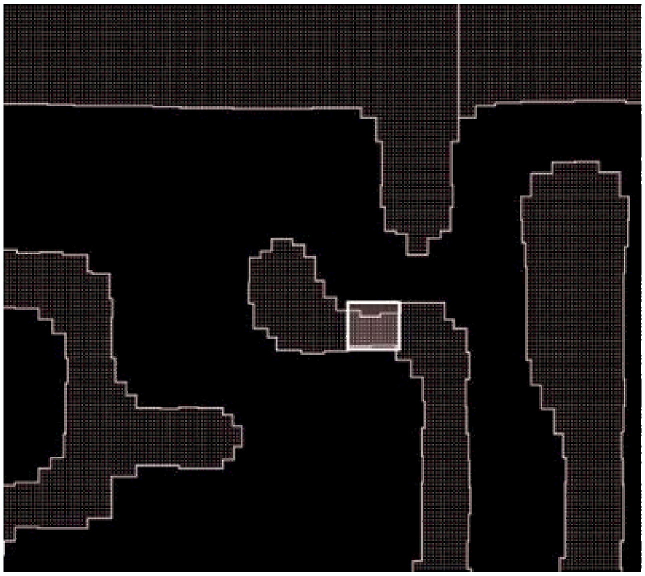


Figure VI.16: Example of open hotspot detection.

all parameters if defocus is changed. Our proposed hotspot detection method achieves good accuracy (100% of hotspots are detected) with smaller false detection overhead. Our approach can also track hotspots well according to changes in process condition. Relative area *(Rel. area)* as shown in Table VI.1 is formulated as #hotspot * (simulation window) / (chip area), with simulation window again being $4 \times 4\mu$m. Table VI.1 shows that 100% hotspot detection is achieved with a small number of falsely detected hotspots. The *Rel. areas* are from 0.5% to 24.2%. Average runtime (including graph generation, hotspot detection and hotspot report) for four test cases is more than 287× faster compared to the ORC tool.

For five 65nm designs, we detect 5 bridging and 71 open hotspots on average, which perfectly matches the results of golden commercial tools. The parameter values of ϵ_0, ϵ_1, ϵ_2 and d_0 are 0.55, 0.83, 0.91 and 0.18, respectively. For all designs, we can use the same parameters for the hotspot detection if lithography conditions are not changed. For necking hotspots, we set the weight of each feature as 0.3, and ϵ as 0.9. *Filtered ORC + DG* represents the total runtime which is the sum of runtime for our dual-graph based detection (DG) and ORC runtime for filtered hotspot area[4]. Average runtime of our method is more than 496× faster than the commercial tool. Average runtime of *Filtered ORC + DG* is also more than 224× faster than the commercial tool. The *Rel. areas* of bridging and open faults are between 0.04% ∼ 14.6% and 0.002% ∼ 0.23% of total areas, respectively. We thus can fillter out between 85.4% and 99.9% of areas for hotspot re-checking. The results are summarized in Table VI.2.

VI.D Conclusions

With the continued shrinkage of minimum feature sizes, hotspots, i.e., printed image with large CD variation, present an important threat for manufacturing yield. Therefore, it becomes more and more important to quickly and accurately detect the hotspots in a layout. In current design flows, there is only one golden physical verification signoff tool and little point in trying to replace the golden signoff tool. However, the golden signoff tool is high in quality, but also high in runtime. This chapter has presented a low-runtime *pre-filter* that reduces the amount of layout area to be analyzed by the golden tool, without compromising the overall quality of hotspot-finding. Specifically, we have described a novel, fast, dual graph based lithographic hotspot detection algorithm without significant accuracy degradation.

For four testcases in 90nm technology, our method can detect all bridging hotspots with average runtime reduction of more than 287× compared to the leading commercial tool. Layout area checked by the commercial golden tool is reduced by amounts ranging from 75% to 99.5%. For five benchmark designs in 65nm technology, we achieve that 100% detection of bridging and open hotspots with few falsely detected hotspots. Average runtimes of our method represent more than 496× reduction compared to the commercial tool.

Our ongoing work includes incorporation of the fast hotspot detection

[4] The commercial tool ORC must confirm whether patterns detected by our DG are truly detected hotspots or falsely detected hotspots. The total runtime includes our dual-graph based detection and ORC re-checking over the hotspot area filtered by our detection approach.

engine within detailed routing to improve yield. We also plan to explore the idea of a "*corner density-based filter*" which may reduce falsely detected hotspots since more vertices within a given proximity radius implies the higher probability of a hotspot. This would be a complement to our currently proposed graph-based detection approach.

VI.E Acknowledgments

This chapter is in part a reprint of:

- A. B. Kahng, C.-H. Park and X. Xu, "Fast Dual Graph-Based Hotspot Detection", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2008, to appear.
- A. B. Kahng, C.-H. Park and X. Xu, "Fast Dual Graph-Based Hotspot Detection", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63490H-1 – 63490H-8.

I would like to thank my coauthors Dr. Xu Xu and Prof. Andrew B. Kahng.

VII

Conclusions and Future Work

VII.A Conclusions

Modern lithography tools can image a complex chip pattern with billions of pixels, within an exposure time of a fraction of a second. However, Moore's Law continues to drive higher performance with smaller circuit features. Aggressive technology scaling has introduced new variation sources and made process variation control more difficult. In particular, while photolithography has been a key enabler of the aggressive IC technology scaling implicit in Moore's Law, minimum feature sizes have outpaced the introduction of advanced lithography hardware solutions.

Currently, the topic of lithography-aware design for manufacturability (DFM) attracts a tremendous amount of interest. Low k_1-factor lithography drives many new process-design interactions that must be comprehended early in the development process to ensure rapid yield ramp-up and acceptable steady-state yield. The purpose of this thesis has been to improve design robustness and performance in light of increased variations, and to reduce manufacturing complexity and cost through the deep integration of design and manufacturing.

In this thesis, we have addressed key problems in the design-to-manufacturing interface, and proposed approaches to improve printability, timing and leakage.

- Sub-resolution assist features (SRAFs) provide an absolutely essential technique for critical dimension (CD) control and process window enhancement in subwavelength lithography. The adoption of off-axis illumination (OAI) and SRAF techniques to enhance resolution at minimum pitch worsens printability of patterns at other pitches. Further, etch micro-loading increases the skew between resist and etch CDs. Etch dummy features have been introduced into the layout to reduce the CD distortion induced by the etch proximity. The etch dummies are placed at the outside of active layers so that leftmost and rightmost gates on active-layer regions are protected from ion scattering during the etch process. However, etch-dummy rules conflict with SRAF insertion because each of the two techniques requires specific spacings from poly. In Chapter II of this thesis, we present three novel methodologies, SAEDM, AFCorr and EtchCorr, to account for interactions of poly geometries between standard cells. We obtain a practical and effective approach to achieve assist-feature compatibility in physical layouts.

- The runtime of model-based optical proximity correction (OPC) tools has grown unacceptably with each successive technology generation, and has emerged as one of the major bottlenecks for turnaround time (TAT) of IC data preparation and manufacturing. In Chapter III of this thesis, we present *auxiliary pattern-enabled* cell-based OPC which can minimize the CD differences between cell-based OPC and model-based OPC. AP-based OPC improves the edge placement error over cell-based OPC by 68%. To enable effective insertion of AP in cell instances at a full-chip layout level, we propose a dynamic programming based method for perturbation of detailed placement. Our approach modifies the detailed placement to allow opportunistic insertion of AP around cell instances in the design layout. By perturbing placement, we achieve 100% AP applicability in designs with placement utilization less than 70%. AP-based OPC substantially reduces leakage and timing variability compared to conventional cell-based OPC, to a level essentially matching that of model-based OPC.

- Lens aberration is the departure of the performance of an optical system from the ideal predictions of paraxial optics. Aberration leads to linewidth variation which is fundamental to achieve timing performance and manufacturing yield. Our studies with lithography and SPICE simulations show significant CD and delay impacts of lens aberration on timing-critical cells. In Chapter IV of this thesis, we present a novel aberration-aware timing-driven analytical placement technique, *AberrPl_TD*, which minimizes the sum of timing-weighted delays of timing-critical cells. We implement our method based on a general analytical placement framework and test it within a complete industrial flow. We evaluate our technique on two testcases, AES and JPEG implemented in 90nm technology. The proposed technique reduces cycle time by 4.322% (80ps) at the cost of 1.587% increase in trial-routed wirelength for AES. On JPEG, we observe a cycle time reduction of 5.182% (132ps) at the cost of 1.095% increase in trial-routed wirelength.

- In current technologies, delay and leakage power reduction continue to be among the most critical design concerns. ASML's DoseMapper technique has been used to reduce ACLV and AWLV metrics during the manufacturing process. However, to achieve optimum device performance or parametric yield, not all transistor gate CD values should necessarily be the same. In Chapter V of this thesis, we propose to exploit the recent availability of fine-grain exposure dose control in the stepper to achieve both design-time (placement) and manufacturing-time (yield-aware dose mapping) optimizations of timing yield and leakage power. Our placement and dose map co-optimization can simultaneously improve both timing yield and leakage power of a given design. We first formulate the placement-aware dose map optimization as a quadratic program, and solve it using an efficient quadratic programming solver. The complementary dose map-aware placement optimization is performed using an efficient cell swapping heuristic. Experimental results are promising: with typical 90nm stepper (ASML DoseMapper) parameters, we achieve more than 8% improvement in minimum cycle time of the circuit without any leakage power degradation.

- Hotspots can be formed under a variety of conditions, notably the original design being unfriendly to the RET that is applied. Hotspots negatively affect the device performance and parametric yield. In Chapter V of this thesis, we present a novel approach to filtering and detection of hotspots induced by lithographic uncertainty. Our goal is to rapidly detect all lithographic hotspots without significant accuracy degradation. In other words, we seek a filtering method: as long as there are no “false negatives”, i.e., we reliably obtain a superset of actual hotspots, then our method can dramatically reduce the layout area processed by time-consuming golden hotspot analysis. Our hotspot detection algorithm includes layout graph construction, graph planarization, three-level bridging hotspot detection, and necking hotspot detection. Experimental results show that our method is promising: for benchmark designs in 90nm and 65nm technologies, 100% of bridging and open hotspots are detected with few falsely detected hotspots. The runtime of our method is up to 496× faster compared with the commercial tool.

VII.B Future Work

As scaling advances into sub-32nm technologies, the prospects for lithography techniques such as extreme ultraviolet (EUV) and immersion ArF (IArF) remain unclear. An EUV imaging system is composed of mirrors coated with multilayer structures designed to have high reflectivity at 13.5nm wavelength. As a result, there are many significant technical problems for implementing EUV lithography in terms of mask-blank fabrication, high-output power source, resist material, etc. IArF requires truly high-refractive index fluids (to enable $\mathrm{NA} = 1.55 \sim 1.6$), and concurrent advances in high-index resists and optical materials.

Lithography-aware DFM at 32nm spans not only geometric operations with limited electrical interactions (e.g., OPC), but also (1) methods for layout and manufacturing handoff for novel patterning approaches, as well as (2) quantification of design and cost tradeoffs inherent in various forms of layout regularity. This begins with an understanding of patterning options. An important option is double patterning lithography (DPL) [30], which involves partitioning of dense circuit patterns into two separate layers so that decreased pattern density can improve resolution and depth of focus (DOF). One other option is hybrid optical maskless lithography (HOMA) [35] which is also a double exposure method using maskless interference lithography (IL) and projection technology (PL). A key challenge to deployment of DPL and HOMA is the conversion of existing layout designs to specific forms of layout for improved manufacturability. We are currently engaged in further studies of new layout styles and designs for equipment. Our ongoing research is in the following directions.

- DPL layout decomposition must satisfy the following requirement: two features must be assigned opposite *colors* (corresponding to mask exposures) if their spacing is less than the *minimum coloring spacing*. However, there exist pattern configurations for which features within this minimum coloring spacing cannot all be assigned different colors [14,104]. We are developing an intelligent layout decomposition algorithm that includes graph construction, conflict cycle detection, and smart node splitting processes.

- HOMA requires both the regular patterns on layout grid and the trimming patterns to make actual circuit design. However, the second exposure for the trim patterns significantly affects the linewidth variation of the regular pattern. We intend to investigate a new framework of regular layout to minimize an area penalty as well as a linewidth impact during trimming regular patterns.

- With increased back-end-of-the-line (BEOL) resistivities, and DPL and/or HOMA for critical layers, new interfaces and synergies between design and lithography processes must be developed around overlay and alignment. Possibilities include misalignment-tolerant layout styles, as well as design-driven alignment targets. We also wish to enhance our design methodologies to comprehend overlay constraints, since overlay increases leakage and timing variability and exponentially affects yield.

- A wide range of equipment improvements, such as DoseMapper and adaptive process control (APC), continually afford opportunities to leverage design information for cost and turnaround time improvements. For example, maskless lithography [89], where the layout data is directly written onto a wafer, require fast data throughputs, e.g., on the order of tens of terabits per second. To overcome the throughput problem, we are investigating new compression techniques for layout data.

In the remainder of this chapter, we introduce ongoing research for DPL, which is expected to be an essential technology at the design-manufacturing interface for 32nm and beyond.

VII.B.1 Layout Decomposition for DPL

A key issue in DPL from the design point of view is the decomposition of the layout for multiple exposure steps [29]. This recalls strong alternating-phase shift mask (altPSM) coloring issues and automatic phase conflict detection and resolution methods [24]. DPL layout decomposition must satisfy the following requirement: two features must be assigned opposite *colors* (corresponding to mask exposures) if their spacing is less than the *minimum coloring spacing*. However, there exist pattern configurations for which features within this minimum coloring spacing cannot all be assigned different colors [14, 104]. In such cases, at least one feature must be *split* into two or more parts. The pattern splitting increases manufacturing cost and complexity due to (1) generation of excessive line-ends, which causes yield loss due to overlay error in double-exposure, as well as line-end shortening under defocus; and (2) resulting requirements for tight overlay control, possibly beyond currently envisioned capabilities. Other risks include line edge (CD) errors due to overlay error, and interference mismatch between different masks. Therefore, a key optimization goal is to reduce the total cost of layout decomposition, considering the above-mentioned aspects.

We formulate the optimization of DPL layout decomposition using integer linear programming (ILP). A pre-processing step fractures layout features into small pieces according to vertex coordinates of neighboring features. From the fractured polygon pieces, we optimize polygon splitting with a process-aware cost function that avoids small jogging line-ends, maximizes overlap at dividing points of polygons, and preferentially makes splits at landing pads, junctions and long runs [29]. A layout partitioning heuristic helps achieve scalability for large layouts. We present an overall layout decomposition method which includes graph construction, conflict cycle detection, and smart node splitting processes.

VII.B.2 DPL Layout Decomposition Flow

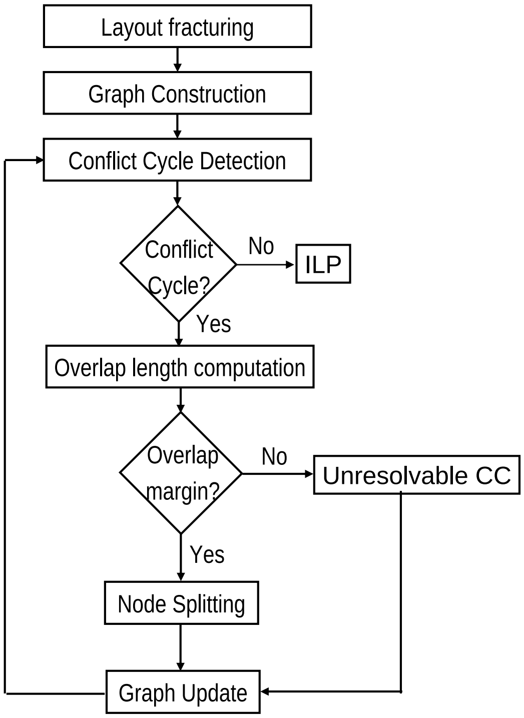


Figure VII.1: Overall DPL layout decomposition flow.

Figure VII.1 shows the overall flow for DPL layout decomposition. Given a layout, the polygonal layout features are first fractured into a set of non-overlapping rectangles using the minimum-sliver fracturing algorithm of [68]. The minimum-sliver fracturing minimizes the number of small rectangles and helps simplify downstream operations. Next, a *conflict graph* is constructed over the rectangular features according to the given minimum coloring spacing, t. Each node in the graph represents a rectangular feature; an edge exists between two nodes if the corresponding features do not touch each other, and the distance between the features is less than t.

We cast DPL layout decomposition as a problem of modifying the conflict graph by decomposing selected layout feature nodes (thus adding new nodes and inducing new edges) so that the graph can be properly 2-colored. To this end, the key is the removal of *conflict cycles* (CCs), which are the odd-length - and hence not 2-colorable - cycles in the conflict graph. We use a breadth first search (BFS) based conflict cycle detection algorithm to find the conflict cycles in the graph.[1]

When a conflict cycle is found, *smart node splitting* is applied to find the best layout feature to split, considering the maximization of overlap lengths. If the maximum possible overlap length is less than a given required overlap margin, then the layout has an *unresolvable conflict cycle* (uCC) which must be flagged to the designer for layout modification. Otherwise, the layout feature will be split into smaller features to remove the conflict cycle. (Note that the node splitting process is necessary to remove a conflict cycle with overlap length greater than the given overlap margin, and that there is no other way to remove a conflict cycle other than by costly layout modification.) The graph is then updated, and the conflict cycle detection and node splitting processes are iterated until no conflict cycle remains in the graph.

After the iterative conflict cycle detection and node splitting process, ILP-based coloring is performed on the final conflict cycle-free graph to find an optimal coloring solution, considering minimization of the number of cuts (or line-

[1] Depth first search (DFS) based cycle detection may also be used. We find that BFS-based conflict cycle detection is more efficient for conflict cycles with fewer edges (≤ 10), whereas DFS-based detection is more efficient for conflict cycles with more edges (> 10). Since most conflict cycles in the layouts we have studied have fewer than 7 edges, we adopt BFS-based conflict cycle detection.

ends) and design rule violations. Finally, a post-processing phase reports minimum overlap length for all pairs of touching features (= adjacent split parts of an original layout feature, which have been assigned different mask colors), and any design rule violations in the final mask solution.

Figure VII.2 shows an example of the DPL-based coloring according to the layout decomposition flow: (a) input layout, (b) fractured layout and conflict graph, (c) conflict cycle removal by node splitting and (d) DPL coloring with ILP.

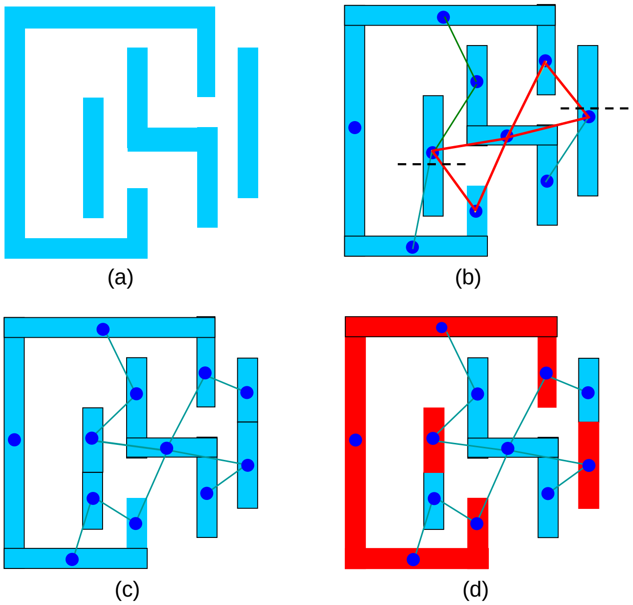


Figure VII.2: Example of graph and layout coloring according to the DPL flow: (a) input layout, (b) fractured layout and conflict graph, (c) conflict cycle removal, and (d) ILP-based DPL coloring.

VII.B.3 The DPL Color Assignment Problem

We formulate the color assignment problem as follows.

Fracturing and color assignment problem

Given: Layout L, and maximum distance between two features (i.e., polygons), t, at which the color assignment is constrained.

Find: A fracture of L and a color assignment of fractured features to minimize the total cost.

Subject to: For any two non-touching fractured features (nodes) n_i and n_j with distance $0 < d_{i,j} < t$, assign different colors. For any two touching features with $d_{i,j} = 0$ and different colors, there is a cost $c_{i,j}$.

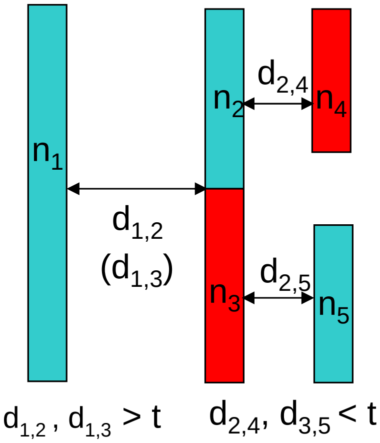


Figure VII.3: Example of color assignment problem: feature n_2 (resp. n_3) is assigned a different color from n_4 (resp. n_5), since $d_{2,4} < t$ (resp. $d_{3,5} < t$).

Figure VII.3 illustrates the color assignment problem. Feature n_2 (respectively, n_3) is assigned a different color from n_4 (resp. n_5), because $d_{2,4} < t$ (resp. $d_{3,5} < t$). Since $d_{1,2} > t$ and $d_{1,3} > t$, there is no need for the pairs of features n_1 and n_2, and n_1 and n_3, to be assigned different colors. Note that when two touching fractured features, e.g., n_2 and n_3 in the figure, are assigned different colors, the two features raise the manufacturing cost (that is, risk) due to overlay error. We should maximize the overlap between the respective mask layouts of n_2 and n_3 in this case, as we now discuss.

We have seen that there may exist pattern configurations for which features within the minimum coloring spacing cannot all be assigned different colors, and that In such cases, we must split at least one feature into two parts. However, this causes pinching under worst process conditions of defocus, exposure dose variation and misalignment. Thus, two line-ends at a dividing point must be sufficiently overlapped. At the same time, the extended features that address the overlap requirement must also satisfy the spacing rules of DPL: the spacing between patterns at the dividing point must be greater than the minimum coloring spacing. Figure VII.4 shows how two choices of dividing point lead to different minimum spacings after layout decomposition. In the figure, each layout decomposition can remove the conflict cycle. However, when extending patterns for overlay margin, the dividing point in Figure VII.4(a) causes violation of the minimum coloring spacing, t. The dividing point in Figure VII.4(b) maintains the minimum coloring spacing even with line-end extension for overlay margin.

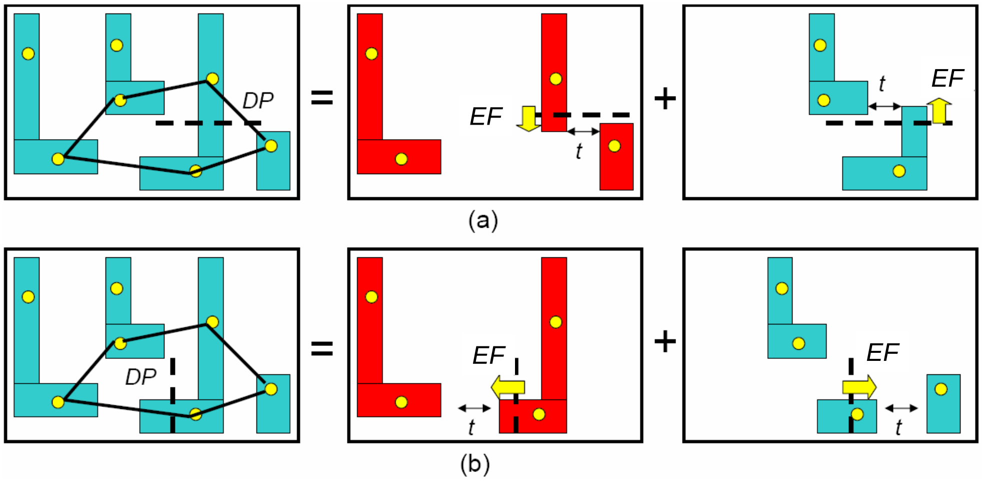


Figure VII.4: Two examples of dividing points (DPs): (a) extended features (EFs) at the dividing point cause a coloring violation, and (b) extended features at the dividing point maintain the minimum coloring spacing rule.

Fracturing and Conflict Graph Construction

Given a layout L, a rectangular layout L_R is obtained by fracturing layout polygons into rectangles. We fracture into rectangles [68] so that distance computation and other feature operations (e.g., feature splitting) become easier. Our layout decomposition process begins with construction of a conflict graph based on the fractured layout. As illustrated in Figure VII.5, given a (post-fracturing) rectangular layout L_R, the layout graph $G = (V, E)$ is constructed by:

- representing each feature (i.e., rectangle) by a node n; and
- for any two non-touching features within distance t, connecting the two corresponding nodes with an edge e.

For non-touching features that are adjacent in the graph, either the two features belong to different original polygonal layout features, or there do not exist other features between the features (i.e., no features entirely block the two non-touching features). In Figure VII.5, we have edge set $\{E_{1,3},\ E_{3,5},\ E_{5,6}\}$. There is no edge between n_2 and n_4 since node n_3 blocks these two nodes.

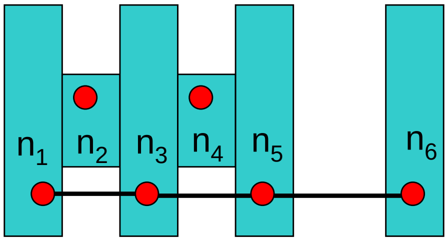


Figure VII.5: Example of conflict graph construction: every (rectangle) feature is represented by a node, and no features entirely block two non-touching features that are adjacent in the graph.

Conflict Cycle Detection

To detect pattern configurations in which non-touching features cannot be assigned different colors, we find odd-length cycles in the conflict graph.

Definition 1: A **conflict cycle** is a cycle in the conflict graph which contains an odd number of *edges*.

Given a conflict graph as in Figure VII.2(b), we apply a breadth first search (BFS) technique, given in Algorithms 2 and 3, to detect conflict cycle.

Algorithm 2 Conflict cycle detection algorithm.

Input: Conflict cycle graph G.

Output: Report the nodes in one conflict cycle if there are any conflict cycles.

```
Set distance $d_i \leftarrow -\infty$ for each node $n_i \in G$;
Make a queue $Q$ and enqueue node $n_0 \in G$ into $Q$;
Set distance $d_0 \leftarrow 0$ for $n_0$;
while $Q$ is not empty do
    Dequeue the first node $n_j$ in $Q$;
    for all nodes $n_k$ adjacent to $n_j$ do
        if $d_k \geq 0$ then
            if $d_k = d_j$ then
                Report a conflict cycle as given in Algorithm 3;
                return;
            end if
        else
            Set $d_k \leftarrow d_j + 1$;
            Enqueue $n_k$ into $Q$;
        end if
    end for
end while
```

Time complexity of the conflict cycle detection algorithm is $O(V+E)$, where V and E are respectively the number of nodes and edges in the conflict graph G.

Conflict cycle detection and conflict cycle removal (cf. the node splitting process in Section VII.B.3) processes are carried out in an iterative manner. Each time a conflict cycle is detected, the conflict cycle removal process is invoked to remove the conflict cycle. Further rounds of detection and removal are performed until the graph G is conflict cycle-free. As described in the experimental results below, total runtime for the whole process, including the conflict cycle detection, conflict cycle removal and min-cost color assignment, is reasonable: less than 8 minutes for layouts of more than 110K features (422K rectangles after fracturing).[2]

Smart node splitting

Node splitting is applied to nodes in conflict cycles so that we may eventually obtain a graph without any conflict cycles. Each time a conflict cycle is detected, we compute the maximum possible overlap length over all possible node splits that remove the conflict cycle (recall Figure VII.4). If this maximum achievable overlap length is greater than the required overlap margin, the node splitting process is carried out to split one node into two nodes and eliminate the conflict cycle. The conflict graph is then updated with newly generated nodes, and another iteration with BFS-based conflict cycle detection begins. The time complexity of smart node splitting is $O(C)$, where C is the number of nodes in the conflict cycle.

Definition 2: The *node projection* $P_{i,j}$ from node n_i to node n_j is a set of points on n_j that have distance to node n_i less than t.

Fact 1: In the conflict cycle graph, node projections between each pair of nodes that are adjacent in the conflict graph are non-empty.

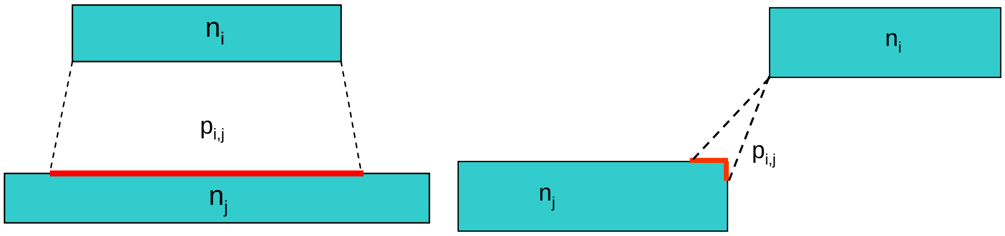
Figure VII.6 shows two examples of node projections.

Definition 3: A horizontal (vertical) *merged projection* $m_h(P)$ ($m_v(P)$) for a

[2] This runtime includes all stages: layout partitioning, all rounds of conflict cycle detection and removal, ILP-based color assignment, etc. The total runtime of BFS-based conflict cycle detection (Algorithm 2) across all conflict cycle detection rounds is less than 4 seconds.

Algorithm 3 Conflict cycle reporting algorithm.

```
Input: Conflict cycle graph $G$ with marked distances and nodes $n_j$ and $n_k$
    in the detected conflict cycle in Algorithm 2.
Output: Report the nodes in the detected conflict cycle in a double-linked list,
    where edges exist between adjacent nodes in the list.
Make a map $F$ to store the father node for each node;
Set $F(n_j) \leftarrow$ NULL, $F(n_k) \leftarrow$ NULL;
Make a queue $Q'$ and enqueue nodes $n_j$ and $n_k$ into $Q'$;
while $Q'$ is not empty do
  Dequeue the first node $n_r$ in $Q'$;
  for all nodes $n_s$ adjacent to $n_r$ do
    if $d_s + 1 = d_r$ then
      if $n_s$ is visited then
        Make a double-linked list $L$;
        Push $n_s$ into $L$;
        Push $n_r$ to the back of $L$;
        Set $n_f \leftarrow F(n_r)$;
        while $n_f \neq$ NULL do
          Push $n_f$ to the back of $L$;
          Set $n_f \leftarrow F(n_f)$;
        end while
        Set $n_f \leftarrow F(n_s)$;
        while $n_f \neq$ NULL do
          Push $n_f$ to the front of $L$;
          Set $n_f \leftarrow F(n_f)$;
        end while
        return $L$;
      end if
      Set $F(n_s) \leftarrow n_r$;
      Mark $n_s$ as visited;
      Enqueue $n_s$ into $Q'$;
    end if
  end for
end while
```



(a) n_i is aligned to n_j.

(b) n_i is not aligned to n_j.

Figure VII.6: Node projection examples.

given projection P on node n is the union of P and all the projections on n that horizontally (vertically) overlap with P.

Definition 4: Node projections are *separable* if they are disjoint.

Definition 5: The *overlap length* of a newly generated node for the corresponding dividing point is the length that the node can be extended across the dividing point without introducing new edges in the conflict cycle graph.

Rule-based node splitting: Given a conflict cycle and a node n_i in the cycle, if (i) the horizontal (vertical) merged projections $m_h(P_{j,i})$ ($m_v(P_{j,i})$) and $m_h(P_{k,i})$ ($m_v(P_{k,i})$) corresponding to adjacent nodes n_j and n_k are separable, (ii) the resulting overlap lengths of the horizontal (vertical) splitting are not less than the given overlap margin, and (iii) there are no design rule violations after splitting, then node n_i can be horizontally (vertically) split into two nodes to remove the conflict cycle. The dividing point may be chosen in between the merged projections such that no merged projections are cut, and no violations of overlap margin or design rules occur.

Figure VII.7 shows an example of rule-based node splitting. In the figure, assume there is a conflict cycle between nodes n_i, n_j and n_k, and the horizontal merged projections $m_h(P_{j,i}) = P_{j,i} \cup P_{l,i}$ and $m_h(P_{k,i}) = P_{k,i}$ on node n_i corresponding to nodes n_j and n_k are separable with overlap lengths not less than the given overlap margin. Hence, node n_i can be split into two new nodes at the dividing point shown, with corresponding overlap lengths of $o_{i,j}$, $o_{i,k}$ and $o_{i,l}$. The dividing

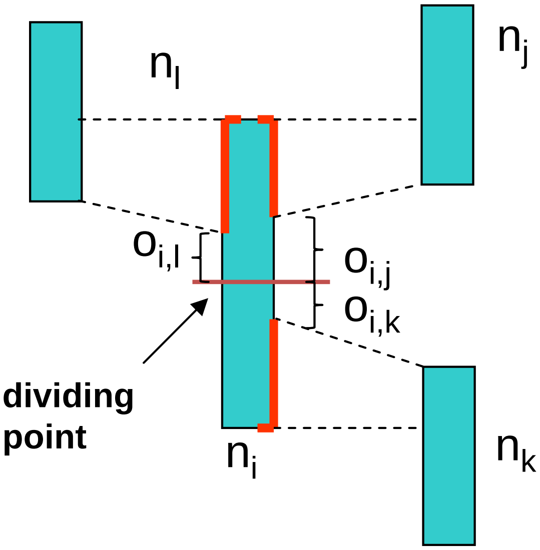


Figure VII.7: Example of rule-based node splitting: $o_{i,j}$, $o_{i,k}$ and $o_{i,l}$ are overlap lengths.

point is in between the lower point of $P_{l,i}$ and the upper point of $P_{k,i}$[3]. Generally, the position of the dividing point is chosen so as to maximize the smaller overlap length. A more detailed illustration of overlap length is given in Figure VII.8, where the two touching features n_4 and n_5 are assigned different colors, and thus the overlap between n_4 and n_5 is required to be larger than the given overlap margin to guarantee successful manufacturing. The overlap lengths of the touching features n_4 and n_5 are denoted as $o_{4,5}$ and $o_{5,4}$, respectively. When computing overlap length, two features of the same color cannot be extended such that the distance between them is less than the minimum coloring spacing t (e.g., in the figure, n_4 cannot be extended to touch the projection of feature n_7).

Of course, not all conflict cycles can be eliminated by the node splitting method. DPL layout decomposition fails when pattern features within the color spacing lower bound cannot be assigned different colors. Such a failure, which corresponds to a uCC, has two cases: (a) there is no dividing point to remove the conflict cycle among all of rectangles which have nonzero overlap length, and (b)

[3] After merging the projections $P_{j,i}$ and $P_{l,i}$ horizontally, the lower point of the merged projection is the same as that of projection $P_{l,i}$. Since there is no projections that horizontally overlap with projection $P_{k,i}$, the horizontal merged projection of $P_{k,i}$ is equal to $P_{k,i}$

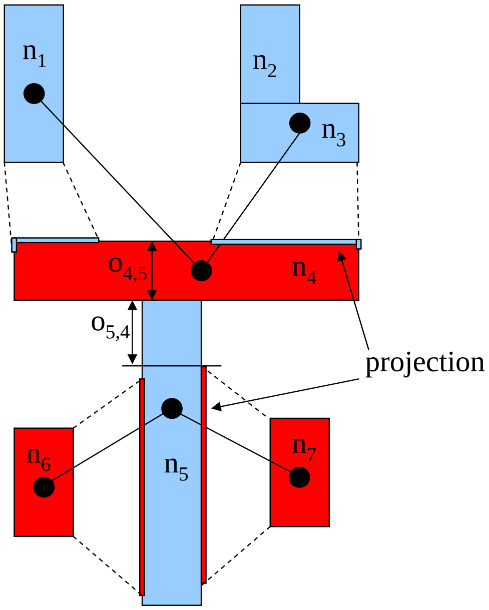


Figure VII.8: Example of overlap length calculation: $o_{4,5}$ and $o_{5,4}$ are the overlap lengths for n_4 and n_5, respectively.

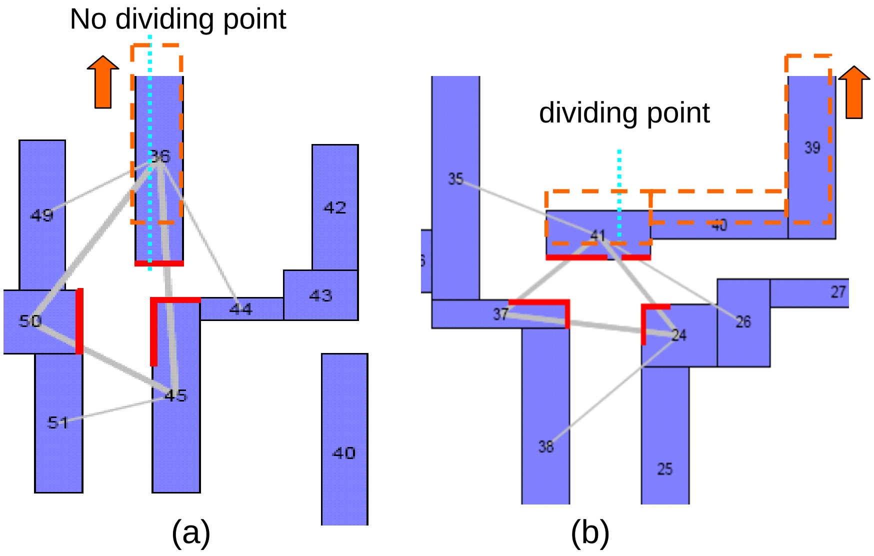


Figure VII.9: Examples of unresolvable conflict cycle (uCC): (a) uCC with zero overlap length, and (b) uCC with non-zero overlap length (less than the overlap margin).

there is a dividing point to remove the conflict cycle, but the overlap length is less than the overlap margin. Figure VII.9 illustrates these two types of uCC. If we divide the rectangle in the center as shown in Figure VII.9(a), the size of the rectangle violates the minimum design rule (CD). Removal of the conflict cycle may be achieved by layout pertubation which increases the spacing to neighboring patterns to be $> t$ (as shown in orange color). In Figure VII.9(b), the overlap length at the dividing point is less than the required overlap margin. A fix by layout perturbation is similarly available. We observe that the space required to increase overlap length is less than that required to remove the conflict cycle, i.e., the pattern can be split after a smaller perturbation.

In summary, node splitting handles conflict cycles in two ways: (i) splitting a node in the conflict cycle, or (ii) reporting an unresolvable conflict cycle for layout optimization to eliminate. Whenever a conflict cycle is detected in the conflict cycle graph, it is eliminated using one of these ways. By construction, the rule-based node splitting does not cause any violation of design rules or overlap margin for the newly generated nodes. On the other hand, if any feature split in the conflict cycle will result in a violation, then an unresolvable conflict cycle is reported. Hence, the iterative conflict cycle detection and node splitting process will terminate without any conflict cycle in the graph, and we have:

Fact 2: The iterative conflict cycle detection and node splitting method can obtain a conflict cycle graph which is 2-colorable.

Min-cost color assignment problem formulation

Finally, the minimum-cost color assignment problem is formulated as follows.

Min-cost color assignment problem

Given: A list of rectangles R which is color assignable, and maximum distance between two features, t, at which the color assignment is constrained.

Find: A color assignment of rectangles to minimize the total cost.

Subject to: Any two non-touching rectangles with $0 < d(i,j) \leq t_{ij}$ must be assigned different colors.

As stated previously, for any two touching rectangles with $d(i,j) = 0$, there is a corresponding cost $c_{i,j}$ (if they are assigned different colors). We cast this as an integer linear program (ILP):

$$\textbf{Minimize: } \sum c_{i,j} \times y_{i,j}$$

Subject to:

$$x_i + x_j = 1 \tag{VII.1}$$

$$x_i - x_j \leq y_{i,j} \tag{VII.2}$$

$$x_j - x_i \leq y_{i,j} \tag{VII.3}$$

where x_i and x_j are binary variables (0/1) for the colors of rectangles r_i and r_j, and $y_{i,j}$ is a binary variable for any pair of touching rectangles r_i and r_j. Constraint (VII.1) specifies that non-touching rectangles r_i and r_j within distance t should be assigned different colors. Constraints (VII.2) and (VII.3) are used for evaluating the cost when touching rectangles r_i and r_j are assigned different colors. The cost for touching rectangles is defined as.

$$c_{i,j} = \alpha \cdot f(w_{i,j})/(f(l_i) \cdot f(l_j)) + \beta \tag{VII.4}$$

where $w_{i,j}$ is the width of the rectangle edge between rectangle r_i and r_j; l_i and l_j are lengths of the rectangle edges of r_i and r_j which are opposite to the touching edge; and α and β are user-defined parameters for tuning of the optimization objective. Function f is defined as.

$$f(x) = \begin{cases} FS_{min} & \forall x \geq FS_{min} \\ x & \forall x < FS_{min} \end{cases} \tag{VII.5}$$

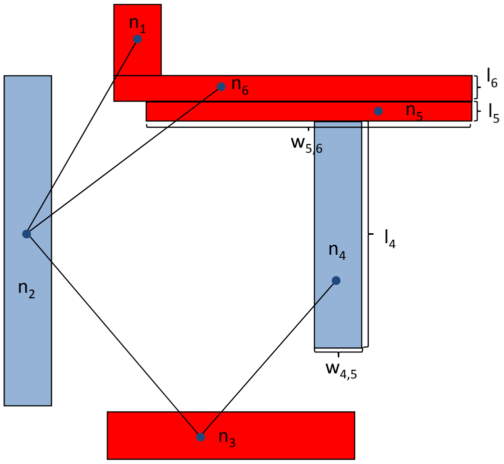


Figure VII.10: Example of cost function: $l_5 < l_6 < FS_{min}$; $c_{4,5} = \alpha/l_5 + \beta$; $c_{5,6} = \alpha \cdot FS_{min}/(l_5 \cdot l_6) + \beta$; $c_{5,6} > c_{4,5}$.

where FS_{min} is minimum feature size, below which a design rule violation occurs. Our ILP problem formulation seeks to minimize design rule violations and the number of cuts on the layout polygons.

Minimizing design rule violations. During the layout fracturing process, small rectangles may be generated due to specific polygonal layout features (e.g., n_5 and n_6 in Figure VII.10). According to Equation (VII.4), it is easy to understand that higher costs will be assigned to pairs of touching rectangles of smaller sizes. By minimizing the total cost, the ILP-based problem formulation aims to minimize the design rule violations in the final layout.

Minimizing the number of cuts. Cuts are susceptible to line end shortening effects. Therefore, the number of cuts should be minimized to improve the quality of the decomposed layouts. In Equation (VII.4), by setting the second term (β) to be greater than the first term, cut minimization can be given higher priority relative to consideration of design rules.

Figure VII.10 illustrates the cost function computation, where:

- $l_5 < l_6 < FS_{min}$;
- $c_{4,5} = \alpha \cdot f(w_{4,5})/(f(l_4) + f(l_5)) + \beta = \alpha/l_5 + \beta$;

Table VII.1: Parameters of the testcases: The minimum spacing between features and the minimum line width are 140nm and 100nm, which are scaled down by 0.4 to be 56nm and 40nm, respectively.

Design	#Cells	#Polygons	#Rectangles	min. spacing (nm)		min. width (nm)	
				Before	×0.4 Scaling	Before	×0.4 Scaling
AES	17304	90394	362380	140	56	100	40
TOP-A	12320	110760	422300	140	56	100	40
TOP-B	42700	385910	1460690	140	56	100	40

- $c_{5,6} = \alpha \cdot f(w_{5,6})/(f(l_5) + f(l_6)) + \beta = \alpha \cdot FS_{min}/(l_5 \cdot l_6) + \beta$;

- $c_{5,6} > c_{4,5}$;

From this computation, we get $c_{5,6} > c_{4,5}$. Given such cost on the touching rectangles, the ILP solver can output the color assignment results summarized in Figure VII.10. Since l_5 (the length of rectangle n_5) is less than minimum feature size FS_{min}, the design rule will be violated if n_5 is assigned a different color than n_6. In this way, the cost function supports the minimization of design rule violations.

VII.B.4 Experiments

Our layout decomposition system is implemented in C++. We use one real-world design (AES) implemented using *Artisan* 90nm libraries using Synopsys *Design Compiler* (v 2003.06-SP1) [12]. Because real-world synthesized netlists do not use all of the available standard-cell masters, we also run experiments with two artificial designs (TOP-A and TOP-B) that instantiate more than 600 different types of cell masters from the same library. The testcases are placed with row utilizations of 70% and 90% using Cadence *SOC Encounter* (v 3.3) [4]. Table VII.1 shows the parameters of the testcases. The minimum spacing between features in the 90nm library-based layout is 140nm, with minimum feature size of 100nm. To obtain experimental results that reflect future designs with smaller feature sizes, we scale down the GDS layout by a factor of 0.4, which results in 56nm minimum spacing and 40nm minimum feature size.

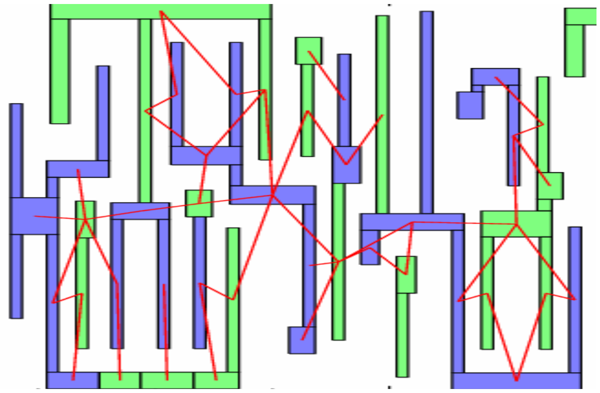


Figure VII.11: Example of DPL layout decomposition in the poly layer.

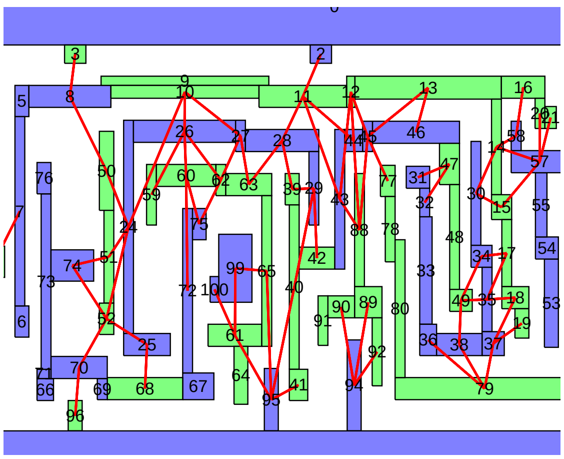


Figure VII.12: Example of DPL layout decomposition in the M1 layer.

We sweep the color spacing lower bound as well as placement utilization, and evaluate solution quality according to various metrics, including number of conflict cycles, number of unresolvable conflict cycles, and overlap length. Table VII.2 shows the results from our layout decomposition system. In Table VII.2, "t(nm)" refers to the minimum coloring spacing; the column under "#CCs" gives the number of detected conflict cycles; "#uCCs" represents the number of unresolvable conflict cycles; and "#Cuts" refers to the number of touching rectangle pairs with different colors. (Note that we do not consider the cuts on unresolvable conflict cycles. Thus, as the number of unresolvable conflict cycles increases, the reported total number of cuts may decrease.) The minimum 3 overlap length values for all the cuts in the final decomposed layout are reported, and "num" represents the number of touching rectangle pairs with the given overlap length value. We also verify whether there exist any design rule violations in the final decomposed layout.

From the experimental results, we see that, as expected, the CC and uCC values increase as the minimum coloring spacing t increases. Also, AES has a smaller number of conflict cycles and unresolvable conflict cycles relative to TOP-A and TOP-B. This is because AES uses fewer types of cell masters, and these types of cell masters have fewer inherent unresolvable conflict cycles; on the other hand, while TOP uses many different cell masters, including some which contain more conflicts and unresolved conflict cycles. Tracking the CC and uCC metrics across 70% and 90% placement utilizations, we can infer that unresolvable conflict cycles mainly exist within each cell instance rather than between cell instances (i.e., there is only a small impact from the different utilizations). Figures VII.11 and VII.12 show small examples of our layout decomposition solutions, where all features are correctly decomposed with respect to the pre-specified overlap margin.

In our preliminary experiments, the overlap margin is set to 8nm, i.e., when a conflict cycle cannot be removed by node splitting with overlap length greater than 8nm, an unresolvable conflict cycle is reported. As noted above, the number of uCCs increases with t. We observe that it is costly to make layout perturbations in a post-layout processing loop, and such that layout modifications

Table VII.2: Experimental results of layout decomposition system in AES, TOP-A and TOP-B. t(nm) is the minimum coloring spacing. #CCs is the number of conflict cycles detected. #uCCs is the number of unresolvable conflict cycles.

Design (Util.(%))	t(nm)	#CCs	#uCCs	#Cuts	Minimum 3 overlap lengths						Runtime (s)
					$1st$ (nm)	num	$2nd$ (nm)	num	$3rd$ (nm)	num	
AES (70%)	58	0	0	29	44	4	56	1	60	1	353.5
	59	0	0	33	56	1	67	16	98	4	353.2
	60	0	0	33	56	1	65	16	97	4	354.3
	61	62	1	132	18	57	32	1	37	7	346.6
TOP-A (70%)	58	162	0	2535	13	100	21	11	28	10	465.7
	59	222	60	3225	9	100	24	10	28	21	467.5
	60	222	160	3158	21	10	26	21	30	20	470.1
	61	3181	1290	9930	11	271	12	20	13	80	479.1
TOP (70%)	58	562	0	8861	13	350	22	32	28	34	4791.2
	59	781	210	11277	9	350	24	34	29	73	4565.7
	60	781	560	11033	22	34	26	73	30	70	4538.8
	61	10174	4987	34536	10	374	12	85	14	27	4548.1
AES (90%)	58	0	0	29	44	9	56	1	60	2	449.1
	56	0	0	33	56	1	67	10	98	4	444.4
	60	0	0	33	56	1	65	10	98	2	443.2
	61	58	1	132	18	57	32	1	38	7	316.6
TOP-A (90%)	58	166	0	2533	13	100	21	11	28	7	528.0
	59	227	60	3224	9	100	24	7	28	20	528.8
	60	227	160	3157	21	7	26	20	30	20	466.5
	61	3187	1291	9914	11	269	12	100	13	100	479.2
TOP (90%)	58	574	0	8906	13	350	22	27	28	29	4865.2
	59	794	210	11316	9	350	24	29	29	75	4662.4
	60	794	560	11067	22	29	26	75	30	70	4558.4
	61	10179	4990	34539	10	377	12	88	14	27	4550.3

can be avoided by decreasing the value of t (e.g., $t = 60$ for AES and $t = 58$ for TOP) via smaller k_1 in lithography. From the columns under "Minimum 3 overlap lengths", we can see that all the overlap lengths in the final mask decomposition are greater than the pre-specified overlap margin (8nm margin in 45nm node [7]), which supports the effectiveness of our layout decomposition system.

VII.C Acknowledgments

This chapter is in part a reprint of:

- A. B. Kahng, C.-H. Park, X. Xu and H. Yao, "Double Patterning Lithography Aware Intelligent Layout Decomposition", *Proc. SPIE BACUS Symposium on Photomask Technology and Management*, 2008, to appear.

I would like to thank my coauthors Dr. Xu Xu, Dr. Hailong Yao and Prof. Andrew B. Kahng.

Bibliography

[1] “ASML”,
http://www.asml.com/.

[2] “ASML DoseMapper”,
http://wps2a.semi.org/cms/groups/public/documents/membersonly/van_schoot_presentation.pdf .

[3] “Cadence SignalStorm”,
http://www.cadence.com/datasheets/6256_LibChar_TP_v2.pdf/.

[4] “Cadence SOC Encounter”,
http://www.cadence.com/products/digital_c/soc_encounter/.

[5] “ILOG CPLEX”,
http://www.ilog.com/products/cplex/.

[6] “Insightful S-Plus”,
http://www.insightful.com/.

[7] “International Technology Roadmap for Semiconductors”,
http://public.itrs.net/.

[8] “KLA-Tencor ProLith”,
http://www.kla-tencor.com/.

[9] “Mentor Graphics Calibre RET”,
http://www.mentor.com/products/ic_nanometer_design/mask_syn/index.cfm/.

[10] “Mentor Graphics Calibre DRC”,
http://www.mentor.com/calibre/.

[11] “OpenAccess”,
http://openeda.si2.org/.

[12] "Synopsys Design Compiler", http://www.synopsys.com/products/logic/design_compiler.html/.

[13] "Synopsys PrimeTime", http://www.synopsys.com/products/analysis/ptsi_ds.html/.

[14] E. Bailey, A. Tritchkov, J.-W. Park, L. Hong, V. Wiaux, E. Hendrickx, S. Verhaegen, P. Xie and J. Versluijs, "Double Pattern EDA Solutions for 32nm HP and Beyond", *Proc. SPIE Design for Manufacturability through Design-Process Integration*, 2007, pp. 65211K-1 – 65211K-12.

[15] A. Balasinski, "Multi-Layer and Multi-Product Masks: Cost Reduction Methodology", *Proc. BACUS Symposium on Photomask Technology and Management*, 2004, pp. 351-359.

[16] S. Bhardwaj, Y. Cao and S. Vrudhula, "Statistical Leakage Minimization Through Joint Selection of Gate Sizes, Gate Lengths and Threshold Voltage", *Proc. Asia and South Pacific Design Automation Conference*, 2006, pp. 953-958.

[17] T. A. Brunner, "Impact of Lens Aberrations on Optical Lithography", *IBM Journal of Research and Development*, http://www.research.ibm.com/journal/rd/411/brunner.html, 1997.

[18] Y. Cao, Y.-W. Lu, L. Chen and J. Ye, "Optimized Hardware and Software for Fast, Full Chip Simulation", *Proc. SPIE Conference on Optical Microlithography*, 2005, pp. 407-414.

[19] K. Cao, S. Dobre and J. Hu, "Standard Cell Characterization Considering Lithography Induced Variations", *Proc. ACM/IEEE Design Automation Conference*, 2006, pp. 801-804.

[20] L. Capodieci, P. Gupta, A. B. Kahng, D. Sylvester and J. Yang, "Toward a Methodology for Manufacturability Driven Design Rule Exploration", *Proc. ACM/IEEE Design Automation Conference*, 2004, pp. 311-316.

[21] J. F. Chen, T. Laidig, K. Wampler and R. Caldwell, "Optical Proximity Correction for Intermediate-Pitch Features using Sub-Resolution Scattering Bars", *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, 15(6), 1997, pp. 2426-2433.

[22] C.-P. Chen, C. C. N. Chu and M. D. F. Wong, "Fast and Exact Simultaneous Gate and Wire Sizing by Lagrangian Relaxation", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 18(7), 1999, pp. 1014-1025.

[23] Y.-S. Cheon, P.-H. Ho, A. B. Kahng, S. Reda and Q. Wang, "Power-Aware Placement", *Proc. ACM/IEEE Design Automation Conference*, 2005, pp. 795-800.

[24] C. Chiang, A. B. Kahng, S. Sinha, X. Xu and A. Zelikovsky, "Bright-Field AAPSM Conflict Detection and Correction", *Proc. Design Automation and Testing in Europe*, 2005, pp. 908-913.

[25] N. Cobb and A. Zakhor, "Experimental Results on Optical Proximity Correction and Variable Threshold Model", *Proc. SPIE Conference on Optical Microlithography*, 1997, pp. 458-468.

[26] M. Cote and P. Hurat, "Standard Cell Printability Grading and Hot Spot Detection", *Proc. International Symposium on Quality Electronic Design*, 2005, pp. 264-269.

[27] M. Cote and P. Hurat, "Layout Printability Optimization Using a Silicon Simulation Methodology", *Proc. International Symposium on Quality Electronic Design*, 2004, pp. 159-164.

[28] C. Chu and Y.-C. Wong, "FLUTE: Fast Lookup Table Based Rectilinear Steiner Minimal Tree Algorithm for VLSI Design", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 27(1), 2008, pp. 70-83.

[29] M. Drapeau, V. Wiaux, E. Hendrickx, S. Verhaegen and T. Machida, "Double Patterning Design Split Implementation and Validation for the 32nm Node", *Proc. SPIE Conference on Design for Manufacturability through Design-Process Integration*, 2007, pp. 652109-1 – 652109-15.

[30] M. Dusa et al., "Pitch doubling through dual-patterning lithography challenges in integration and litho budgets", *Proc. SPIE Conference on Optical Microlithography*, 2007, pp. 65200G-1 - 65200G-10.

[31] H. Eisenmann and F. M. Johannes, "Generic Global Placement and Floorplanning", *Proc. ACM/IEEE Design Automation Conference*, 1998, pp. 269-274.

[32] H. Etawil, S. Areibi and A. Vannelli, "Attractor-Repeller Approach For Global Placement", *Proc. International Conference on Computer Aided Design*, 1999, pp. 20-24.

[33] N. Farrar, A. Smith, D. Busath and D. Taitano, "In-Situ Measurement of Lens Aberrations", *Proc. SPIE on Optical Microlithography*, 2001, pp. 18-29.

[34] D. G. Flagello, H. Laan, J. Schoot, I. Bouchoms and B. Geha, "Understanding Systematic and Random CD variations using Predictive Modeling Techniques", *Proc. SPIE on Optical Microlithography*, 1999, pp. 162-175.

[35] M. Fritze, T. M. Bloomstein, B. Tyrrell, T. H. Fedynyshyn, N. N. Efremow, D. E. Hardy, S. Cann, D. Lennon, S. Spector and M. Rothschild, "Hybrid Optical Maskless Lithography: Scaling beyond The 45nm Node", *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, 23(6), 2005, pp. 2743 - 2748.

[36] M. Garg, A. Kumar, I. van Wingerden and L. Le Cam, "Litho-Driven Layouts for Reducing Performance Variability", *Proc. IEEE International Symposium on Circuits and Systems*, 2005, pp. 3551-3554.

[37] J. Gortych and D. Williamson, "Effects of Higher-Order Aberrations on the Process Window", *Proc. SPIE on Optical Microlithography*, 1991, pp. 368-381.

[38] Y. Granik, N. B. Cobb and T. Do, "Universal Process Modeling With VTRE for OPC", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 377-394.

[39] Y. Granik, "Correction for Etch Proximity: New Models and Applications", *Proc. SPIE Conference on Optical Microlithography*, 2001, pp. 98-112.

[40] J. Gu and X. Huang, "Efficient Local Search With Search Space Smoothing: A Case Study of the Traveling Salesman Problem (TSP)", *IEEE Transactions on Systems, Man, and Cybernetics* 24(5), 1994, pp. 728-735.

[41] P. Gupta, A. B. Kahng, D. Sylvester and J. Yang, "A Cost-Driven Lithographic Correction Methodology Based on Off-the-Shelf Sizing Tools," *Proc. ACM/IEEE Design Automation Conference*, 2003, pp. 16-21.

[42] P. Gupta and A. B. Kahng, "Manufacturing-Aware Physical Design", *Proc. International Conference on Computer Aided Design*, 2003, pp. 681-687.

[43] P. Gupta, A. B. Kahng and C.-H. Park, "Detailed Placement for Improved Depth of Focus and CD Control", *Proc. Asia and South Pacific Design Automation Conference*, 2005, pp. 343-348.

[44] P. Gupta, A. B. Kahng and C.-H. Park, "Detailed Placement for Enhanced Control of Resist and Etch CDs", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(12) 2007, pp. 2144 - 2157.

[45] P. Gupta, A. B. Kahng and C.-H. Park, "Manufacturing-Aware Design Methodology for Assist Feature Correctness", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2005, pp. 131-140.

[46] P. Gupta, A. B. Kahng and C.-H. Park, "Enhanced Resist and Etch CD Control by Design Perturbation", *Proc. BACUS Symposium on Photomask Technology and Management*, 2005, pp. 59923P-1 – 59923P-11.

[47] P. Gupta, F.-L. Heng and M. Lavin, "Merits of Cellwise Model-Based OPC", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2004, pp. 182-189.

[48] P. Gupta, A. B. Kahng, S. Nakagawa, S. Shah and P. Sharma, "Lithography Simulation-Based Full-Chip Design Analyses", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2006, pp. 61560T-1 – 61560T-8.

[49] P. Gupta, A. B. Kahng, P. Sharma and D. Sylvester, "Gate-Length Biasing for Runtime-Leakage Control", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(8), 2006, pp. 1475-1485.

[50] K. Hashimoto, T. Kuji, S. Tokutome, T. Kotani, S. Tanaka and S. Inoue, "A Tandem Process Proximity Correction Method", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 1070-1081.

[51] K. Hashimoto, S. Usui, S. Nojima, S. Tanaka, E. Yamanaka and S. Inoue, "Hot Spot Management in Ultra-low k1 Lithography", *Proc. SPIE Conference on Optical Microlithography*, 2005, pp. 1207-1219.

[52] F.-L. Heng, J.-F. Lee and P. Gupta, "Toward Through-Process Layout Quality Metrics", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2005, pp. 161-167.

[53] B. Hu and M. Marek-Sadowska, "FAR: Fixed-Points Addition and Relaxation Based Placement", *Proc. ACM International Symposium on Physical Design*, 2002, pp. 161-166.

[54] L.-D. Huang and M. D. F. Wong, "Optical Proximity Correction (OPC)-Friendly Maze Routing", *Proc. ACM/IEEE Design Automation Conference*, 2004, 186-191.

[55] H. Ito and C. G. Willson, "Chemical Amplification in the Design of Dry Developing Resist Materials", *Technical Papers of SPE Regional Technical Conference on Photopolymers*, 1982, pp. 331-353.

[56] N. Jeewakhan et al., “Application of DoseMapper for 65-nm Gate CD Control: Strategies and Results”, *Proc. BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63490G-1 – 63490G-11.

[57] A. B. Kahng, S. Muddu and P. Sharma, “Defocus-Aware Leakage Estimation and Control,” *Proc. International Symposium on Low Power Electronics and Design*, 2005, pp. 263-268.

[58] A. B. Kahng and S. Mantik, “Measurement of Inherent Noise in EDA Tools”, *Proc. International Symposium on Quality Electronic Design*, 2002, pp. 206-211.

[59] A. B. Kahng, I. Markov and S. Reda, “On Legalization of Row-Based Placements”, *Proc. ACM Great Lakes Symposium on VLSI*, 2004, pp. 214-219.

[60] A. B. Kahng and C.-H. Park, “Auxiliary Pattern for Cell-Based OPC”, *Proc. BACUS Conference on Photomask Technology and Management*, 2006, pp. 6349S-1 – 6348S-10.

[61] A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, “Lens Aberration Aware Timing-Driven Placement”, *Proc. ACM/IEEE Design Automation and Testing in Europe*, 2006, pp. 890-895.

[62] A. B. Kahng and Q. Wang, “Implementation and Extensibility of an Analytic Placer”, *Proc. ACM International Symposium on Physical Design*, 2004, pp. 18-25.

[63] A. B. Kahng and Q. Wang, “An Analytic Placer for Mixed-Size Placement and Timing-Driven Placement”, *Proc. International Conference on Computer Aided Design*, 2004, pp. 565-572.

[64] A. B. Kahng and Q. Wang, “Implementation and Extensibility of an Analytic Placer”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 24(5), 2005, pp. 734-747.

[65] A. B. Kahng, S. Reda and Q. Wang, “APlace: A General Analytic Placement Framework”, *Proc. ACM International Symposium on Physical Design*, 2005, pp. 233-235.

[66] A. B. Kahng, S. Reda and Q. Wang, “Architecture and Details of a High Quality, Large-Scale Analytical Placer”, *Proc. International Conference on Computer Aided Design*, 2005, pp. 890-897.

[67] A. B. Kahng, C.-H. Park and X. Xu, “Fast Dual-Graph Based Hot-Spot Detection”, *Proc. BACUS Symposium on Photomask Technology and Management*, 2006, pp. 63490H-1 – 63490H-8.

[68] A. B. Kahng, X. Xu and A. Zelikovsky, “Fast Yield-Driven Fracture for Variable Shaped-Beam Mask Writing”, *Proc. SPIE Conference on Photomask and Next-Generation Lithography Mask Technology*, 2006, pp. 62832R-1 – 62832R-9.

[69] K. Kamon, T. Miyamoto, Y. Myoi, M. Fujinaga, H. Nagata and M. Tanaka, “Photolithography System Using Modified Illumination”, *Japanese Journal of Applied Physics*, 32(1A), 1993, pp. 239-243.

[70] K. Kato, K. Nishizawa and T. Inoue, “Advanced Mask Rule Check (MRC) Tool”, *Proc. Photomask and Next-Generation Lithography Mask Technology*, 2006, pp. 62830O-1 – 62830O-11.

[71] A. Kennings and I. Markov, “Analytical Minimization of Half-Perimeter Wirelength”, *Proc. Asia and South Pacific Design Automation Conference*, 2000, pp. 179-184.

[72] K. Kim, Y. Choi, R. Socha and D. Flagello, “Optimization of Process Condition to Balance MEF and OPC for Alternating PSM”, *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 240-246.

[73] M. Kling, N. Cave, B. J. Falch, C. C. Fu, K. Green, K. Lucas, B. J. Roman, A. J. Reich, J. Sturtevant, R. Tian, D. Russell, L. Karklin and Y. Wang, “Practicing Extension of 248-nm DUV Optical Lithography Using Trim-Mask PSM”, *Proc. SPIE Conference on Optical Microlithography*, 1999, pp. 10-17.

[74] T. Kong, “A Novel Net Weighting Algorithm for Timing-Driven Placement”, *Proc. International Conference on Computer Aided Design* , 2002, pp. 10-14.

[75] M. Lavin and F.-L. Heng and G. Northrop, “Backend CAD Flows for “Restrictive Design Rules”, *Proc. International Conference on Computer Aided Design*, 2004, pp. 739-746.

[76] N. Layadi, J. I. Colonell and J. Lee, “An Introduction to Plasma Etching for VLSI Circuit Technology”, *Bell Labs Technical Journal*, 1999, pp. 155-171.

[77] M. D. Levenson, N. Viswanathan and R. Sympson, “Improving Resolution in Photolithography With a Phase-Shifting Mask”, *IEEE Transactions on Electron Devices*, 29(12), 1982, pp. 1812-1846.

[78] H. J. Levinson, *Principles of Lithography*, SPIE Press, 2001.

[79] L. Liebmann, A. Barish, Z. Baum, H. Bonges, S. Bukofsky, C. Fonseca, S. Halle, G. Northrop, S. Runyon and L. Sigal, “High Performance Circuit

Design for the RET-enabled 65-nm Technology Node", *Proceedings of the SPIE Conference on Optical Microlithograhpy*, 2004, 20-29.

[80] L.W. Liebmann, "Layout Impact of Resolution Enhancement Techniques: Impediment or Opportunity?", *Proc. ACM International Symposium on Physical Design*, 2003, pp. 110-117.

[81] B. J. Lin, "The Attenuated Phase-Shifting Mask", *Solid State Technology*, 1992, pp. 43-47.

[82] B. J. Lin, "The k_3 Coefficient in Nonparaxial Lambda / NA Scaling Equations for Resolution, Depth of Focus, and Immersion Lithography", *Journal of Microlithography, Microfabrication, and Microsystems*, 1(1), 2002, pp. 7-12.

[83] C. A. Mack, *Field Guide to Optical Lithography*, SPIE Press, 2006.

[84] C. A. Mack, *Fundamental Principles of Optical Lithography*, John Wiley and Sons, Ltd., 2007.

[85] A. Marquardt, V. Betz and J. Rose, "Timing-Driven Placement for FPGAs", *Proc. ACM Symposium on FPGAs*, 2000, pp. 203-213.

[86] T. Matsuyama, Y. Shibazaki, Y. Ohmura and T. Suzuki, "High NA and Low Residual Aberration Projection Lens for DUV Scanner", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 687-695.

[87] S. Muddu, "Predictive Modeling of Integrated Circuit Manufacturing Variation", *Ph.D. Thesis*, UC San Diego ECE Department, 2008.

[88] W. Naylor et al., "Non-Linear Optimization System and Method for Wire Length and Delay Optimization for an Automatic Electric Circuit Placer", *US Patent* 6301693, Oct. 2001.

[89] B. Nikolic, B. Wild, V. Dai, Y. A. Shroff, B. Warlick, A. Zakhor and W. G. Oldham, "Layout Decompression Chip for Maskless Lithography", *Proc. SPIE Conference on Emerging Lithographic Technologies*, 2004, pp. 1092-1099.

[90] S. Nojima, S. Mimotogi, M. Itoh, O. Ikenaga, S. Hasebe, K. Hashimoto, S. Inoue, M. Goto and I. Mori, "Flexible Mask Specifications", *Proc. BACUS Symposium on Photomask Technology*, 2002, pp. 187-196.

[91] T. Ogawa, M. Uematsu, T. Ishimaru and M. Kimura, "The Effective Light Source Optimization With the Modified Beam for the Depth-of-Focus Enhancements", *Proc. SPIE Conference on Optical Microlithography*, 1994, pp. 19-30.

[92] M. Orshansky, L. Milor, L. Nguyen, G. Hill, Y. Peng and C. Hu, "Intra-Field Gate CD Variability and Its Impact on Circuit Performance", *IEDM Technical Digest*, 1999, pp. 479-482.

[93] 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*, 21(5), 2002, pp. 544-553.

[94] 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.

[95] O. Otto, J. Garofalo, K. K. Low, C. M. Yuan, R. Henderson, C. Pierrat, R. Kostelak, S. Vaidya and P. K. Vasudev, "Automated Optical Proximity Correction: A Rules-Based Approach", *Proc. SPIE Conference on Optical Microlithography*, 1994, pp. 278-293.

[96] J.-S. Park, C.-H. Park, S.-U. Rhie, Y.-H. Kim, M.-H. Yoo, J.-T. Kong, H.-W. Kim and S.-I. Yoo, "An Efficient Rule-Based OPC Approach Using a DRC Tool for 0.18um ASIC", *Proc. International Symposium on Quality Electronic Design*, 2000, pp. 82-88.

[97] C.-H. Park, T.-K. Kim, H.-J. Lee, J.-T. Kong and S.-H. Lee, "An Automatic Gate CD Control for a Full Chip Scale SRAM Device", *Proc. BACUS Symposium on Photomask Technology and Management*, 1998, pp. 350-357

[98] C.-H. Park, Y.-H. Kim, J.-S. Park, K. Kim, M.-H. Yoo and J.-T. Kong, "A Systematic Approach to Correct Critical Patterns Induced by the Lithography Process at the Full-Chip Level", *Proc. SPIE Conference on Optical Microlithography*, 1999, pp. 622-629.

[99] J. Petersen, "Analytical Description of Anti-Scattering and Scattering Bar Assist Features", *Proc. SPIE Conference on Optical Microlithography*, 2000, pp. 77-89.

[100] M. V. Plat, K. B. Nguyen, C. A. Spence, C. F. Lyons and A. Wilkison, "Impact of Optical Enhancement Techniques on the Mask Error Enhancement Function (MEEF)", *Proc. SPIE Conference on Optical Microlithography*, 2000, pp. 206-214.

[101] I. Pollentier et al., "In-Line Lithography Cluster Monitoring and Control Using Integrated Scatterometry", *Proc. SPIE Conference on Data Analysis and Modeling for Process Control*, 2004, pp. 105-115.

[102] C. J. Progler, A. Borna, D. Blaauw and P. Sixt, "Impact of Lithography Variability on Statistical Timing Behavior", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2004, pp. 101-110.

[103] C. J. Progler and A. K. Wong, "Zernike Coefficients: Are They Really Enough?", *Proc. SPIE Conference on Optical Microlithography*, 2000, pp. 40-52.

[104] J. Rubinstein and A. R. Neureuther, "Post-Decomposition Assessment of Double Patterning Layout", *Proc. SPIE Conference on Optical Microlithography*, 2008, pp. 69240O-1 – 69240O-12.

[105] E. Sahouria, Y. Granik, N. Cobb and O. Toublan, "Full-Chip Process Simulation for Silicon DRC", *International Conference on Modeling and Simulation of Mircosystems*, 2000, pp. 32-35.

[106] F. M. Schellenberg, L. Capodieci and R. Socha, "Adoption of OPC and the Impact on Design and Layout", *Proc. ACM/IEEE Design Automation Conference*, 2001, pp. 89-92.

[107] F. M. Schellenberg and L. Capodieci, "Impact of RET on Physical Layouts", *Proc. ACM International Symposium on Physical Design*, 2001, pp. 52-55.

[108] J. B. van Schoot, O. Noordman, P. Vanoppen, F. Blok, D. Yim, C.-H. Park, B.-H. Cho, T. Theeuwes and Y.-H. Min, "CD Uniformity Improvement by Active Scanner Corrections", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 304-314.

[109] R. Seltmann et al., "ACLV-Analysis in Production and Its Impact on Product Performance", *Proc. SPIE Conference on Optical Microlithography*, 2003, pp. 530-540.

[110] N. Seong, H. Kim, H. Cho, J. Moon and S. Lee, "Measurement of Pitch Dependency of Overlay Errors under OAI by Using an Electric CD Measurement Technique", *Proc. SPIE Conference on Optical Microlithography*, 1999, pp. 1170-1174.

[111] X. Shi, S. Hsu, F. Chen, M. Hsu, R. Socha and M. Dusa, "Understanding the Forbidden Pitch Phenomenon and Assist Feature Placement", *Proc. SPIE Conference on Metrology, Inspection, and Process Control for Microlithography*, 2002, pp. 985-996.

[112] Y. Shiode, S. Okada, H. Takamori, H. Matusda and S. Fujiwara, "Method of Zernike Coefficients Extraction for Optics Aberration Measurement", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 1453-1464.

[113] R. N. Singh, A. E. Rosenbluth, G. L.-T. Chiu and J. S. Wilczynski, "High-Numerical-Aperture Optical Designs", *IBM Journal of Research and Development*, 1997, pp. 39-48.

[114] R. Socha, M. Dusa, L. Capodieci, J. Finders, F. Chen, D. Flagello and K. Cummings, "Forbidden Pitches for 130nm Lithography and Below", *Proc. SPIE Conference on Optical Microlithography*, 2000, pp. 1140-1155.

[115] J. Stirniman and M. Rieger, "Optimizing Proximity Correction for Wafer Fabrication Processes", *Proc. BACUS Symposium on Photomask Technology and Management*, 1994, pp. 239-246.

[116] J. Stirniman and M. Rieger, "Fast Proximity Correction With Zone Sampling", *Proc. SPIE Conference on Optical / Laser Microlithography*, 1994, pp. 294-301.

[117] K. K. H. Toh and A. Neureuther, "Identifying and Monitoring Effects of Lens Aberrations in Projection Printing", *Proc. SPIE Conference on Optical Microlithography*, 1987, pp. 202-209.

[118] T. Tugbawa, T. Park, D. Boning, T. Pan, P. Li, S. Hymes, T. Brown and L. Camilletti, "A Mathematical Model of Pattern Dependence in Cu CMP Process", *Proc. International Chemical-Mechanical Polishing Symposium*, 1999, pp. 605-615.

[119] S. Vipul, C.B. Keshav, K.G. Surnanth and P. R. Suresh, "Transistor Flaring in Deep Submicron - Design Considerations", *Proc. International Conference on VLSI Design*, 2002. pp. 2-8.

[120] N. Viswanathan and C. C.-N. Chu, "FastPlace: Efficient Analytical Placement Using Cell Shifting, Iterative Local Refinement and a Hybrid Net Model", *Proc. ACM International Symposium on Physical Design*, 2004, pp. 26-33.

[121] C. H. Wallace and C.-H. Jang, "Sub-Resolution Assist Features for Photolithography With Trim Ends", *US Patent Application* 20070128525, June, 2007.

[122] X. Wang, M. Pilloff, H. Tang and C. Wu, "Exploiting Hierarchical Structure to Enhance Cell-Based RET With Localized OPC Reconfiguration", *Proc. SPIE Conference on Design and Process Integration for Microelectronic Manufacturing*, 2005, pp. 361-367.

[123] A. K. Wong, "Theoretical Discussion on Reduced Aberration Sensitivity of Enhanced Alternating Phase-Shifting Masks", *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 395-368.

[124] A. K. Wong, *Resolution Enhancement Techniques in Optical Lithography*, SPIE Press, 2001.

[125] A. K. Wong, R. Ferguson, S. Mansfield, A. Molless, D. Samuels, R. Schuster and A. Thomas, “Level-Specific Lithography Optimization for 1-Gb DRAM”, *IEEE Transactions on Semiconductor Manufacturing*, 13(1), 2000, pp. 76-87.

[126] J. Word, S. Zhu and J. Sturtevant, “Assist Feature OPC Implementation for the 130nm Technology Node With KrF and No Forbidden Pitches”, *Proc. SPIE Conference on Optical Microlithography*, 2002, pp. 1139-1147.

[127] J. Yang, L. Capodieci and D. Sylvester, “Advanced Timing Analysis Based on Post-OPC Extraction of Critical Dimensions”, *Proc. ACM/IEEE Design Automation Conference*, 2005, pp. 359-364.

[128] K. Yeh and W. Loong, “Simulations of Mask Error Enhancement Factor in 193nm Immersion Lithography”, *Japanese Journal of Applied Physics*, 45(4A), 2006, pp. 2481-2496.

[129] G. Zhang et al., “65nm Node Gate Pattern Using Attenuated Phase Shift Mask With Off-Axis Illumination and Sub-Resolution Assist Features”, *Proc. SPIE Conference on Optical Microlithography*, 2005, pp. 760-772.